HAL
open science

Nouvelles stratégies pour l'étude des facteurs génétiques impliqués dans le cancer du sein familial

Juliette Coignard

To cite this version:

Juliette Coignard. Nouvelles stratégies pour l'étude des facteurs génétiques impliqués dans le cancer du sein familial. Santé publique et épidémiologie. Université Paris Saclay (COmUE), 2019. Français. NNT : 2019SACLS426 . tel-03005974

HAL Id: tel-03005974

https://theses.hal.science/tel-03005974v1

Submitted on 15 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université PARIS-SACLAY

NNT : 2019SACLS426

Thèse de doctorat

Nouvelles stratégies pour l'étude des facteurs génétiques impliqués dans le cancer du sein familial.

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École Doctorale n°570 Santé Publique (EDSP)
Spécialité de doctorat: Santé publique - génétique statistique

Thèse présentée et soutenue à Paris, le 14 novembre 2019, par

Juliette Coignard

Composition du Jury :

Jean Bouyer Directeur de Recherche, Université Paris Sud	Président
Emmanuelle Génin Directeur de Recherche, Université de Bretagne Occidentale	Rapporteur
David Goldgar Research Professor, University of Utah	Rapporteur
Sylvie Mazoyer Directeur de Recherche, Université de Lyon	Examinateur
Antoine de Pauw Conseiller en génétique, Institut Curie	Examinateur
Stefan Michiels Chargé de Recherche, Université Paris Saclay	Examinateur
Nadine Andrieu Directeur de Recherche, Inserm	Directeur de thèse
Antonis Antoniou Professor, University of Cambridge	Co-Directeur de thèse

Remerciements

Ce travail n'aurait jamais pu voir le jour sans l'aide et le soutien de toutes les personnes qui m'ont entourée au cours de ces 4 dernières années, qui m'ont permis de grandir et de m'épanouir tant scientifiquement et professionnellement que personnellement. J'ai eu la chance de ne recevoir que de la bienveillance de la part des personnes que j'ai côtoyées, et ça, ça n'a pas de prix !

Tout d'abord je tiens à remercier toutes les filles, et les quelques garçons, du Boul'Mich de m'avoir permis de travailler dans les meilleures conditions. Amivi, Anne-Laure, David, Diana, Dorothée, Ève, Fabienne, Gaëlle, Juana, Julie, Marie, Marie-Gabrielle, Marilou, Max, Nadine, Noura, Sarah, Séverine, Om, Yue… Merci à la PIGE, et particulièrement à Dorothée et Séverine de m'avoir aidée à m'en sortir avec les tonnes de questionnaires de GENESIS et de m'avoir aidée à percer les mystères des données manquantes et des incohérences. Merci à Fabienne pour nos nombreuses discussions, et pour m'avoir écoutée râler maintes et maintes fois sur les imputations et sur les SNPs. Merci à mes boul'michtonneuses préférées, Dorothée et Anne-Laure. On passe plus des ¾ de notre temps au travail alors c'est vraiment agréable d'y avoir des amies. Bon, Anne-Laure tu nous as lâchées bien trop tôt, et pour l'Australie en plus…

Je souhaite remercier très particulièrement Marie-Gabrielle, une collègue de bureau tout à fait surprenante au soutien sans faille sans qui cette dernière année aurait pu être chaotique… bien plus chaotique ! Merci pour tes histoires à dormir debout. Merci pour ton bureau qui ressemble au sac de Mary Poppins. Sauf que tu as déteint sur moi et, maintenant, le mien ressemble au tien, et pour ça je ne te remercie pas. Merci d'apporter les chouquettes pour me motiver à arriver à 9h pétante. Merci pour ta précieuse aide sur R et ta magie sur Word et sur Excel. Tu n'as pas réussi à me faire aimer ces logiciels mais tu as réussi à ce qu'ils ne me rendent pas folle, c'est déjà ça ! Merci d'être un Bescherelle ambulant et d'avoir lu ce mémoire de thèse et de l'avoir relu et encore relu et encore et encore relu… à la recherche de la dernière faute récalcitrante, du moindre espace inutile et des tournures de phrases un peu douteuses. Merci d'être restée plus d'une fois pour me tenir compagnie et me soutenir durant les journées de travail qui s'éternisaient. Merci pour tes fous rires sans aucun bruit où tu

donnes la drôle impression d'être en plein bug. Merci d'avoir accepté mon volume sonore pas tout à fait adapté à tes oreilles supersoniques. Merci pour tout et merci d'avoir été ma collègue de bureau pendant ces 4 dernières années.

Je remercie aussi l'équipe de Strangeways à Cambridge de m'avoir accueillie pendant 6 mois. Merci à Antonis Antoniou d'avoir accepté de codiriger cette thèse, de m'avoir permis de travailler sur des données internationales et de participer à des meetings aux quatre coins du monde. Merci également à Dan et Jonathan d'avoir répondu à la multitude de questions que je leur ai posées par mail tout au long de cette thèse. Un très grand merci à Michael de m'avoir aidé à de très nombreuses reprises et sans qui ce projet aurait difficilement pu aboutir. Mes deux séjours à Cambridge n'auraient pas été aussi agréables sans mes deux extraordinaires proprios, Goska et Rob. Je vous remercie de tout mon cœur d'avoir été si accueillants, gentils et motivants. Grâce à vous j'ai réussi à atteindre mon objectif des 15 km et j'en suis très fière… Bon, depuis je ne courre plus, mais ça on n'est pas obligé de le retenir !

Ma thèse m'a aussi donné l'opportunité de découvrir les joies de l'enseignement. Et dans ce domaine, je remercie très sincèrement Jean Bouyer, Faroudy Boufassa et tous les étudiants du M1 Santé Publique à qui j'ai donné des TD de biostats. Merci à vous de m'avoir donné la chance de me découvrir un très fort attrait pour l'enseignement. Moi qui ne supportais pas les oraux il y a peu de temps… c'est fou mais j'ai vraiment pris mon pied à animer ces TDs. Merci à Jean et Faroudy d'avoir été si accueillants et rassurants, vous y êtes pour beaucoup !

Je remercie également ma famille et mes amis d'avoir été à mes côtés. Merci à mon petit chat de me supporter depuis 10 ans déjà et de m'avoir accompagnée tout le long de cette thèse. Merci de m'avoir permis de me déconnecter à chaque fois que je rentrais à la maison et d'oublier le travail et ses contrariétés même après avoir passé ma journée à me plaindre à Fabienne… Merci d'avoir organisé avec moi le plus beau des mariages au cours de cette dernière année de thèse, un projet qui me paraissait bien trop fou au départ mais qui a finalement été l'une des meilleures décisions pour me faire positiver lorsque mes SNPs me rendaient folles. Je te remercie d'être aujourd'hui le meilleur des maris. Merci à mon papou, à ma mamou, à mon petit frère et à ma Sylvie que j'aime de tout mon cœur. Merci à ma belle-mère et mon beau-frère préférés. Et merci à tous les copains, et à ma team parfaite, ma Grosse d'amour, mon Fab, ma Marjojo, mon Mika et mon Antho. Merci d'être une bande de copains

en or sans qui ces 4 dernières années de vie auraient été bien moins drôles, bien moins remplies et surtout bien moins intéressantes.

Merci à mes rapporteurs, David Goldgar et Emmanuelle Génin, à mes examinateurs, Sylvie Mazoyer, Stefan Michiels et Antoine de Pauw et au président de mon jury de thèse, Jean Bouyer, de m'avoir fait l'honneur d'accepter d'évaluer mon travail.

Il est important pour moi de remercier également toutes ces personnes qui n'ont jamais compris que je n'étais plus à l'école… ne vous inquiétez pas, si tout se passe bien j'y retourne l'année prochaine, et pour de vrai cette fois !

Et j'ai bien sûr gardé le meilleur pour la fin… si je devais remercier une seule personne ce serait bien sûr Nadine, ma directrice de thèse, et quelle directrice de thèse ! Lors de mes recherches pour trouver LA thèse idéale, la chose la plus importante pour moi était d'avoir un(e) directeur(rice) de thèse avec des qualités humaines en plus de ses qualités scientifiques. Eh bien, on peut dire que je ne me suis pas trompée… Si je devais décrire Nadine en un mot, j'utiliserais le mot « humain ». Avec ses qualités, beaucoup de qualités, mais aussi ses défauts. Nadine, l'humain, qui n'hésite pas à te dire quand c'est parfait mais aussi quand c'est complètement à chier. Nadine, l'humain, qui te laisse t'exprimer et défendre ton avis et te fait déculpabiliser face à tes erreurs. Nadine, l'humain, qui te permet de percevoir tes propres qualités mais qui accepte aussi tes défauts. Nadine, l'humain, avec qui tu peux avoir des discussions à n'en plus finir, qu'elles soient scientifiques ou non. Je souhaitais un directeur de thèse avec des qualités humaines et j'ai été servie. Je tiens à te remercier de tout mon cœur de m'avoir permis de m'épanouir en tant que personne à part entière pendant ces 4 années de ma vie. Merci de m'avoir fait confiance. Merci de m'avoir mise en valeur. Merci de m'avoir toujours tirée vers le haut. Merci d'avoir toujours été franche et transparente avec moi. Merci de m'avoir obligée à penser et non à croire. Et merci de me dire d'arrêter de râler et de m'avoir répété qu'« il est indispensable d'être optimiste pour ne pas rester couchée sous sa couette ». Je pense que tu m'as trop mal habituée et qu'il sera vraiment difficile pour mon prochain chef d'arriver à ton niveau. Merci Nadine.

TABLE DES MATIÈRES

ÉTAT DE L'ART LE SEIN ET LE CANCER DU SEIN. 12

LE SEIN 13
I. ANATOMIE DU SEIN 13
II. RÔLE DES HORMONES DANS LE DÉVELOPPEMENT DES GLANDES MAMMAIRES 16
LE CANCER DU SEIN 21
I. PHYSIOPATHOLOGIE 21
1. Les carcinomes canalaires ou lobulaires *in situ* 22
2. Les carcinomes canalaires ou lobulaires infiltrants 23
II. ÉVALUATION PRONOSTIQUE DES CANCERS DU SEIN 23
1. Stade 23
2. Grade 24
3. Hétérogénéité moléculaire 25
III. ÉPIDÉMIOLOGIE DESCRIPTIVE DU CANCER DU SEIN 27
IV. LE DÉPISTAGE 29
1. Le dépistage organisé 29
2. Familles à haut risque de cancer du sein 29
V. LES FACTEURS DE RISQUE DU CANCER DU SEIN 33
1. Les facteurs environnementaux et de mode de vie 33
2. Les facteurs génétiques 41
3. Les gènes *BRCA1* et *BRCA2* 47

PREMIÈRE PARTIE FACTEURS DE RISQUE GÉNÉTIQUES SPÉCIFIQUES À UN SCHÉMA ENVIRONNEMENTAL PARTICULIER CHEZ LES FEMMES À HAUT RISQUE DE CANCER DU SEIN ET NON PORTEUSES D'UNE MUTATION DANS LES GÈNES *BRCA1* OU *BRCA2*. 54

INTRODUCTION 55
DONNÉES 58
I. LA POPULATION D'ÉTUDE : GENESIS 58
1. Critères d'inclusion 58
2. Données collectées 59
3. Description de la population d'étude 59
II. LES DONNÉES GÉNOTYPIQUES 60
1. La puce iCOGS 60
2. Voies biologiques d'intérêt 63
III. LES DONNÉES ENVIRONNEMENTALES 64
1. La censure 65
2. Variables gynéco-obstétriques 65
3. Variables liées aux expositions aux radiations 69
MÉTHODES 71
I. IMPUTATION DES DONNÉES MANQUANTES 71

1. Imputation simple 71
2. Imputation multiple 72
3. Imputation des SNPs non génotypés 74
II. La régression logistique 82
1. Description 82
2. Outils d'analyse 84
3. Facteurs confondants 86
4. Tests multiples 86
5. Stratégie pour définir des scores de risque 87
6. Test d'hétérogénéité 88
7. Test de permutations 89
Résultats 90
I. Description de la population 90
II. Facteurs non génétiques 93
1. Facteurs confondants 93
2. Facteurs gynéco-obstétriques 94
3. Expositions aux radiations 100
III. Facteurs génotypiques 104
1. Imputation des SNPs non génotypés 104
2. Analyse des SNPs 106
Discussion 112

DEUXIÈME PARTIE. FACTEURS GÉNÉTIQUES MODIFICATEURS DU RISQUE DE CANCER DU SEIN CHEZ LES FEMMES PORTEUSES D'UNE MUTATION DANS LES GÈNES *BRCA1* OU *BRCA2*. 117

Introduction 118
Données 119
I. La population d'étude 119
1. Consortium BCAC 119
2. Consortium CIMBA 120
II. Les données génotypiques 122
1. Description de la puce OncoArray 123
2. Contrôle qualité 123
Méthodes 125
I. Imputation des génotypes manquants 125
1. Paramètres de l'imputation 125
2. Contrôle qualité 125
3. Imputation jointe des régions d'intérêt 126
II. Analyse *case-only* 130
1. Sélection des sujets 130
2. Sélection des SNPs : Hypothèse d'indépendance 134
3. Les méthodes statistiques 135
III. Stratégie d'analyse 140
1. Les SNPs de prédisposition au cancer du sein déjà connus 140
2. Les potentiels nouveaux SNPs modificateurs 141
3. Calcul des risques de cancer du sein associés aux SNPs 144
4. Analyses de cartographie fine et prédiction *in silico* 144
Résultats 147
I. Variation génétique géographique 147
1. Variabilité génétique par pays 147
2. Variabilité génétique restante 149

II. RÉSULTATS DES ANALYSES *CASE-ONLY* 156
1. Les potentiels nouveaux SNPs modificateurs 156
2. Les SNPs de prédisposition au cancer du sein déjà connus 169
DISCUSSION 177

CONCLUSIONS ET PERSPECTIVES 183

RÉFÉRENCES BIBLIOGRAPHIQUES 185

ANNEXES 198

Annexe 1 – Définition d'un SNP 199
Annexe 2 - Questionnaire épidémiologique de l'étude GENESIS 200
Annexe 3 - Gènes impliqués dans les voies de signalisation des hormones 233
Annexe 4 - Gènes intervenant dans la réparation de l'ADN et le cycle cellulaire 260
Annexe 5 - Définition du principe de déséquilibre de liaison et d'haplotypes 276
Annexe 6 - Risque de cancer du sein associé aux facteurs gynéco-obstétriques – Analyses stratifiées sur l'année de naissance. 277
Annexe 7 - Études participant au consortium BCAC 280
Annexe 8 - Études participant au consortium CIMBA 282

VALORISATIONS SCIENTIFIQUES 284

ARTICLE SOUMIS 286

Table Des Illustrations

Figure 1 Anatomie du sein a) chez l'homme et b) chez la femme. 14
Figure 2 Anatomie de la glande mammaire. 15
Figure 3 D'après la Biologie de la lactation de Jack Martinet : Schéma général de l'évolution de la glande mammaire depuis l'embryon jusqu'à la fin de la première lactation. 16
Figure 4 Localisation de la tumeur selon le type de cancer du sein – 1. canalaire ou 2. lobulaire – et son niveau d'infiltration – a. in situ ou b. infiltrant. 21
Figure 5 Les différents types d'infiltration tumorale 22
Figure 6 Classement des cancers en fonction du nombre de sujets atteints en 2015 en France métropolitaine par localisation et par sexe[34] 28
Figure 7 Carte de France des différents sites de consultations d'oncogénétique (INCa) 31
Figure 8 Parcours global des cas index et des apparentés en oncogénétique (INCa) 32
Encadré 1 Calcul du r^2 79
Figure 9 Répartition des scores de risque définis à partir des facteurs gynéco-obstétriques 99
Figure 10 Répartition des scores de risque en fonction des expositions aux radiations au thorax 103
Figure 11 Distribution des taux de concordance des 487 fragments imputés 104
Figure 12 Distribution des taux de concordance des 139 fragments ré-imputés 105
Figure 13 Distribution de l'âge au diagnostic des cas de BCAC et CIMBA 133
Figure 14 Étapes d'analyse des SNPs de prédisposition au cancer du sein déjà connus dans la population générale 141
Figure 15 Étapes d'analyse pour la recherche de potentiels nouveaux SNPs modificateurs du risque de cancer du sein chez les porteurs d'une mutation BRCA1/2. 143
Figure 16 Projection des deux premières composantes principales selon le statut (a) ou le pays d'origine (b). 148
Figure 17 Projection des deux premières composantes principales par pays. 152
Figure 18 QQ-plot obtenu selon les différents ajustements. 154
Figure 19 Distribution du facteur d'inflation selon l'ajustement 155
Figure 20 Analyse d'hétérogénéité par pays des nouveaux SNPs identifiés chez les porteuses d'une mutation de BRCA1 161
Figure 21 Analyse d'hétérogénéité par pays des nouveaux SNPs identifiés chez les porteuses d'une mutation de BRCA2. 161
Figure 22 Analyse de sensibilité excluant chaque pays un à un pour les nouveaux SNPs identifiés chez porteuses d'une mutation de BRCA1. 162
Figure 23 Analyse de sensibilité excluant chaque pays un à un pour les nouveaux SNPs identifiés chez porteuses d'une mutation de BRCA2. 162
Figure 24 Comparaison avant et après ré-imputation a) des P-values et b) des ORs. 175
Figure 25 Histogramme des différences observées avant et après imputation des a) p-values et b) ORs 176

Table Des Tableaux

Tableau 1 Calcul du grade du cancer du sein (source - INCa) 24
Tableau 2 SNPs associés au risque de cancer du sein chez les femmes porteuses d'une mutation de BRCA1, pour les cancers du sein tous sous-types confondus, spécifiques aux tumeurs RE^- et spécifiques aux tumeurs RE^+. 51
Tableau 3 SNPs associés au risque de cancer du sein chez les femmes porteuses d'une mutation de BRCA2, pour les cancers du sein tous sous-types confondus, spécifiques aux tumeurs RE^- et spécifiques aux tumeurs RE^+. 52
Tableau 4 Traitements médicamenteux considérés comme des THS 69
Tableau 5 Taux de concordance obtenus après imputation du fragment de 52 672 577 pb à 57 672 577 pb du chromosome 7 avec un k_hap variant de 300 à 800 79
Tableau 6 Exemple de transformation des données imputées d'un SNP avec A et a comme allèle majeur et mineur respectivement. 85
Tableau 7 Description de la population de l'étude GENESIS 91
Tableau 8 Âge moyen à la censure selon l'année de naissance 92
Tableau 9 Répartition de l'âge à la censure en classe (≥ 60 ans, entre 50 et 59 ans et < 50 ans) selon l'année de naissance. 93
Tableau 10 Risque de cancer du sein associé à l'année de naissance, l'âge à la censure et au niveau d'éducation. 93
Tableau 11 Risque de cancer du sein associé aux facteurs gynéco-obstétriques 96
Tableau 12 Facteurs gynéco-obstétriques retenus dans le modèle complet : risque de cancer du sein associé avant et après imputation multiple. 98
Tableau 13 Répartition des cas et des témoins selon le score de risque établi à partir des facteurs gynéco-obstétriques 100
Tableau 14 Effet de l'exposition aux radiations au thorax au cours de la vie sur le risque de cancer du sein en fonction du nombre d'expositions, de l'âge à la première exposition et de la première grossesse menée à terme. 101
Tableau 15 Facteurs associés aux expositions aux radiations au thorax retenus dans le modèle complet : risque de cancer du sein associé avant et après imputation multiple. . 102
Tableau 16 Répartition des cas et des témoins selon le score de risque établi à partir des expositions aux radiations au thorax 103
Tableau 17 Top SNPs associés au risque de cancer du sein. 107
Tableau 18 Top SNPs associés au cancer du sein dans chacun des groupes de score de risque relatifs aux expositions aux radiations au thorax 109
Tableau 19 Top SNPs associés au cancer du sein dans chacun des groupes de score de risque relatifs aux expositions aux radiations – résultats dans chaque groupe 110
Tableau 20 Descriptif des types de tumeurs et des statuts des récepteurs aux œstrogènes (RE) et à la progestérone (RP) des femmes de BCAC 120
Tableau 21 Descriptif des types histologiques des tumeurs des femmes de BCAC. 120
Tableau 22 Descriptif des types de cancer du sein et du statut des récepteurs aux œstrogènes (RE) et à la progestérone (RP) des femmes de CIMBA par mutation 121

Tableau 23 Descriptif des types histologiques des tumeurs des femmes de CIMBA par mutation 122
Tableau 24 Critères d'inclusion des SNPs selon le Δr^2 entre BCAC et CIMBA en fonction du r^2 126
Tableau 25 Nombre de cas par étude et par pays. 127
Tableau 26 Design de l'analyse *Case-only*. 130
Tableau 27 Les différentes étapes de sélection (a) des témoins et (b) des cas 133
Tableau 28 Valeurs propres associées aux 15 composantes principales calculées 151
Tableau 29 Nouveaux SNPs modificateurs du risque de cancer du sein chez les femmes porteuses d'une mutation de BRCA1 158
Tableau 30 Nouveaux SNPs modificateurs du risque de cancer du sein chez les femmes porteuses d'une mutation de BRCA2 159
Tableau 31 SNPs les plus significatifs parmi les différents signaux indépendants de l'analyse CCVs pour les femmes porteuses d'une mutation de BRCA1. 164
Tableau 32 Prédiction d'INQUISIT des potentiels gènes cibles des CCVs trouvés dans l'analyse BRCA1 165
Tableau 33 SNPs les plus significatifs parmi les différents signaux indépendants de l'analyse CCVs chez les femmes porteuses d'une mutation de BRCA2. 167
Tableau 34 Prédiction d'INQUISIT des potentiels gènes cibles des CCVs trouvés dans l'analyse chez les femmes porteuses d'une mutation de BRCA2 168
Tableau 35 SNPs connus dans la population générale montrant une association dans l'analyse BRCA1 171
Tableau 36 SNPs connus dans la population générale montrant une association dans l'analyse BRCA2 172
Tableau 37 Qualité d'imputation des SNPs associés de façon significative avec le statut BRCA1 ou BRCA2 dans les analyses *case-only* réalisées sur les données imputées séparément. 173
Tableau 38 Distribution des SNPs après ré-imputation en fonction de la p-value et de la fréquence allélique (MAF). 173

Table Des Tableaux Supplémentaires

Les tableaux supplémentaires sont disponibles à l'adresse suivante :
https://zenodo.org/record/3518854

Tableau supplémentaire 1	SNPs associés au cancer du sein dans la population entière de GENESIS
Tableau supplémentaire 2	SNPs associés au cancer du sein dans la population totale de GENESIS et leurs ORs dans chacun des groupes formés selon les expositions aux radiations et selon les facteurs gynéco-obstétriques.
Tableau supplémentaire 3	SNPs associés au cancer du sein dans le groupe des femmes non exposées aux radiations.
Tableau supplémentaire 4	SNPs associés au cancer du sein dans le groupe des femmes non exposées aux radiations et leurs ORs dans chacun des groupes formés selon les expositions aux radiations.
Tableau supplémentaire 5	SNPs associés au cancer du sein dans le groupe des femmes à score de risque modéré selon les expositions aux radiations
Tableau supplémentaire 6	SNPs associés au cancer du sein dans le groupe des femmes à score de risque modéré selon les expositions aux radiations et leurs ORs dans chacun des groupes formés selon les expositions aux radiations.
Tableau supplémentaire 7	SNPs associés au cancer du sein dans le groupe des femmes à score de risque élevé selon les expositions aux radiations.
Tableau supplémentaire 8	SNPs associés au cancer du sein dans le groupe des femmes à score de risque élevé selon les expositions aux radiations et leurs ORs dans chacun des groupes formés.
Tableau supplémentaire 9	SNPs des gènes de la régulation des hormones sexuelles associés au cancer du sein des femmes à score de risque faible selon les facteurs gynéco-obstétriques.
Tableau supplémentaire 10	SNPs des gènes de la régulation des hormones sexuelles associés au cancer du sein des femmes à score de risque faible selon les facteurs gynéco-obstétriques et leurs ORs dans chacun des groupes formés.
Tableau supplémentaire 11	SNPs des gènes de la régulation des hormones sexuelles associés au cancer du sein des femmes à score de risque modéré selon les facteurs gynéco-obstétriques.
Tableau supplémentaire 12	SNPs des gènes de la régulation des hormones sexuelles associés au cancer du sein des femmes à score de risque modéré selon les facteurs gynéco-obstétriques et leurs ORs dans chacun des groupes formés.

Tableau supplémentaire 13	SNPs des gènes de la réparation de l'ADN associés au cancer du sein des femmes à score de risque faible selon les facteurs gynéco-obstétriques.
Tableau supplémentaire 14	SNPs des gènes de la réparation de l'ADN associés au cancer du sein des femmes à score de risque faible selon les facteurs gynéco-obstétriques et leurs ORs dans chacun des groupes formés.
Tableau supplémentaire 15	SNPs des gènes de la réparation de l'ADN associés au cancer du sein des femmes à score de risque modéré selon les facteurs gynéco-obstétriques.
Tableau supplémentaire 16	SNPs des gènes de la réparation de l'ADN associés au cancer du sein des femmes à score de risque modéré selon les facteurs gynéco-obstétriques et leurs ORs dans chacun des groupes formés.
Tableau supplémentaire 17	SNPs des gènes de la réparation de l'ADN associés au cancer du sein des femmes à score de risque élevé selon les facteurs gynéco-obstétriques.
Tableau supplémentaire 18	SNPs des gènes de la réparation de l'ADN associés au cancer du sein des femmes à score de risque élevé selon les facteurs gynéco-obstétriques et leurs ORs dans chacun des groupes formés.
Tableau supplémentaire 19	SNPs associés de façon significative ($p < 10^{-8}$) avec les mutations dans le gène *BRCA1* dans l'analyse *control-only*.
Tableau supplémentaire 20	SNPs associés de façon significative ($p < 10^{-8}$) avec les mutations dans le gène *BRCA2* dans l'analyse *control-only*.
Tableau supplémentaire 21	SNPs associés de façon significative ($p < 10^{-8}$) avec les mutations dans le gène *BRCA1* dans l'analyse *case-only*.
Tableau supplémentaire 22	SNPs associés de façon significative ($p < 10^{-8}$) avec les mutations dans le gène *BRCA2* dans l'analyse *case-only*.
Tableau supplémentaire 23	Potentiel SNPs causaux (CCVs) des régions trouvées associées dans l'analyse BRCA1.
Tableau supplémentaire 24	Potentiel SNPs causaux (CCVs) des régions trouvées associées dans l'analyse BRCA2.
Tableau supplémentaire 25	SNPs trouvés associés au cancer du sein tous sous-types dans la population générale : résultats obtenus dans l'analyse *case-only* chez les sujets porteurs d'une mutation *BRCA1*.
Tableau supplémentaire 26	SNPs trouvés associés au cancer du sein de type ER^- dans la population générale : résultats obtenus dans l'analyse *case-only*, restreinte aux cas de BCAC ayant une tumeur ER^-, chez les sujets porteurs d'une mutation *BRCA1*.
Tableau supplémentaire 27	SNPs trouvés associés au cancer du sein tous sous-types dans la population générale : résultats obtenus dans l'analyse *case-only* chez les sujets porteurs d'une mutation *BRCA2*.

État De L'art

Le sein et le cancer du sein

LE SEIN

I. Anatomie du sein

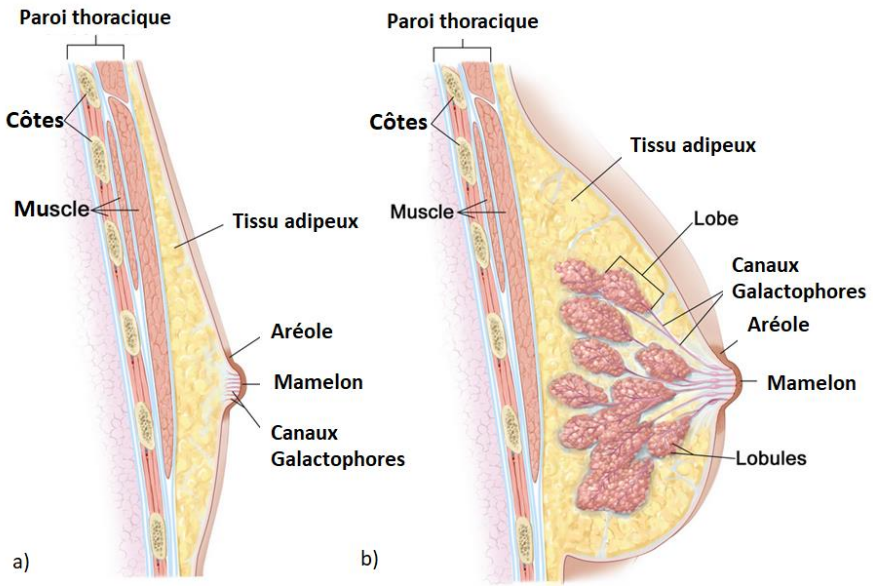
Les seins occupent la partie antéro-supérieure du thorax en avant des muscles pectoraux. Ils s'étendent en hauteur jusqu'à la clavicule et en largeur de l'aisselle au milieu du sternum environ. Ils ne contiennent pas de muscle et sont soutenus par des ligaments. Le sein correspond à l'organe contenant la glande mammaire, glande exocrine qui se développe au cours de la vie chez la femme. La glande mammaire est une masse de densité variable, organisée en une vingtaine de lobes. Les lobes sont séparés et maintenus par du tissu conjonctif et adipeux (Figure 1). Chaque lobe est composé de 20 à 40 lobules mammaires et chaque lobule contient de 10 à 100 alvéoles. Les lobes sont desservis par des canaux galactophores (ou lactifères) qui se rejoignent au niveau du mamelon. Le rôle des lobules est de produire le lait qui en période d'allaitement sera acheminé vers le mamelon grâce aux canaux.

Les seins sont des organes uniques au regard du développement postnatal qu'ils subissent. La prolifération cellulaire y est très importante pendant des périodes de temps très courtes au cours de la vie.

À la naissance, un réseau de canaux galactophores est présent à l'état rudimentaire chez les individus des deux sexes à partir de l'âge embryonnaire mais seules les femmes, sous l'influence hormonale, développent la partie glandulaire (Figure 1).

En période pré-pubertaire se produit une **augmentation lente mais régulière** (au même rythme que celui de la croissance) de la ramification des canaux galactophores et de la formation des lobules à partir du tissu conjonctif[1,2] (Figure 2).

Figure 1 - Anatomie du sein a) chez l'homme et b) chez la femme.

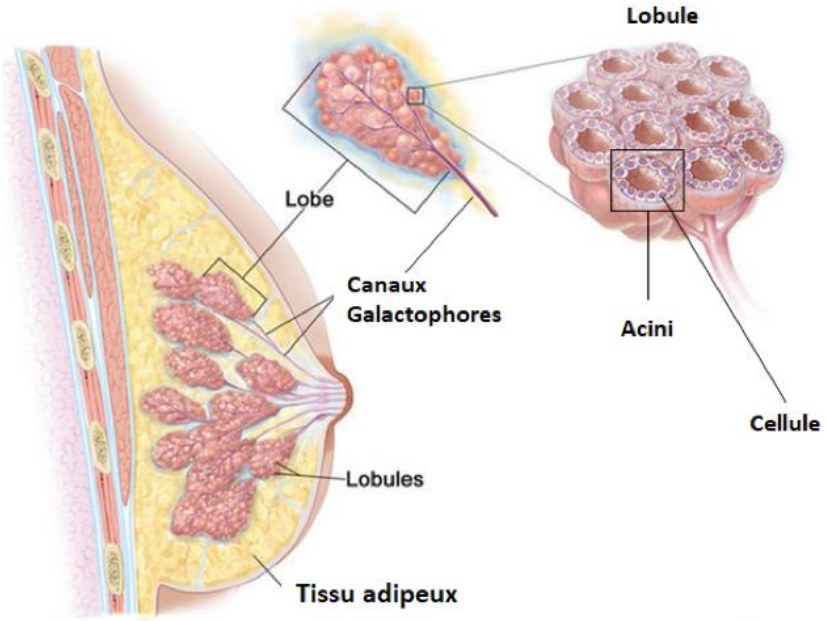


©2018 Terese Winslow. All rights reserved.

La puberté génère des modifications morphologiques significatives. Le **volume mammaire croît par augmentation du tissu mammaire** et du tissu graisseux périphérique. L'aréole mammaire s'élargit et se pigmente d'une couleur rosée. Elle présente une dizaine de petites saillies nommées tubercules de Morgagni. À la surface et en son centre, une saillie cylindrique brunâtre se forme et constitue le mamelon. Les **sécrétions hormonales liées aux premiers cycles ovulatoires stimulent la croissance et la multiplication des canaux galactophores**. Des ébauches d'acini se forment également aux extrémités de chaque canal (Figure 2). Ces ébauches glandulaires, appelées « bourgeons d'attente », n'ont qu'une activité minime qui se développera lors de la première grossesse.

Pendant les cycles menstruels ayant lieu **avant la première grossesse**, les variations hormonales permettent la préparation des glandes mammaires à une potentielle grossesse grâce à l'**augmentation de la prolifération des canaux**. À chaque fin de cycle, ces canaux involuent et la glande mammaire revient à un état proche de son état initial.

Figure 2 - Anatomie de la glande mammaire.



©2018 Terese Winslow. All rights reserved.

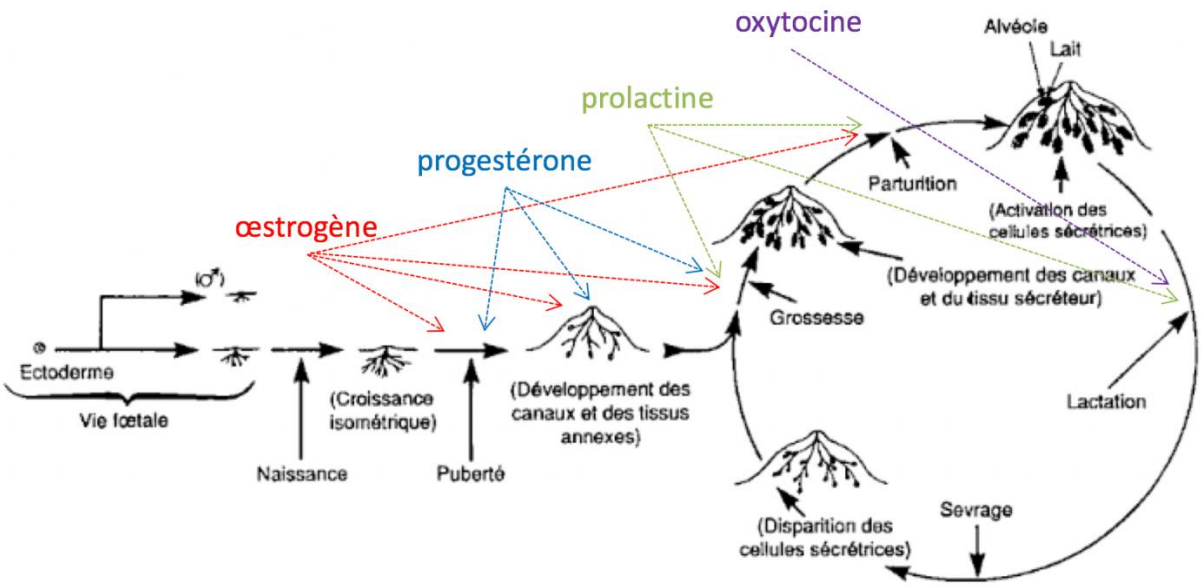
La **dernière étape du développement des glandes mammaires** a lieu au cours de la **première grossesse menée à terme**. Celle-ci permet de préparer le sein à l'allaitement du nouveau-né grâce à la formation des structures adéquates. Une **importante activité mitotique** se met alors en place permettant **la division et le développement des canaux galactophores** ainsi qu'un développement remarquable des ébauches d'acini déjà présentes. Ces canaux et glandes nouvellement créés occupent une grande majorité de la glande mammaire, entrainant une réduction importante des tissus conjonctifs et adipeux.

Pendant ce temps, l'aréole se pigmente un peu plus et prend un aspect grenu. Celui-ci résulte de la saillie des tubercules de Morgagni, qu'on appelle alors tubercules de Montgoméry.

L'activité des acini se manifeste à la fin de la grossesse avec la sécrétion d'un premier liquide jaunâtre et opaque, appelé le colostrum. Dans les premiers jours suivant l'accouchement, la glande mammaire continue à sécréter le colostrum qui est ensuite progressivement remplacé par la sécrétion lactée.

À la ménopause, la glande mammaire s'atrophie mais le volume du sein ne diminue pas toujours, l'atrophie étant compensée par l'augmentation des tissus graisseux.

Figure 3 - Schéma général de l'évolution de la glande mammaire depuis l'embryon jusqu'à la fin de la première lactation,
d'après la Biologie de la lactation de Jack Martinet.



Le cycle mammogénèse-lactogénèse-galactopoïese se reproduit à chaque lactation.

II. Rôle des hormones dans le développement des glandes mammaires

L'implication des hormones dans le développement des glandes mammaires, leur différenciation mais également leur tumorigénèse a été mise en lumière en 1896 lorsque Beatson[3] a découvert l'effet palliatif de l'ablation des ovaires chez une patiente atteinte d'un cancer du sein.

Le développement des glandes mammaires est dépendant des hormones, substances messagères de l'organisme assurant la transmission d'informations nécessaires à la régulation des fonctions organiques et des étapes du métabolisme. Les principales hormones intervenant au cours des différentes étapes du développement sont les hormones de la reproduction (œstrogènes, progestérone[4,5,6], prolactine[7] et ocytocine[8]) (Figure 3) ainsi que les hormones lutéinisantes (l'hormone folliculo-stimulante (FSH) et l'hormone lutéinisante (LH)), l'hormone de croissance (GH), les hormones thyroïdiennes et les glucocorticoïdes[8]. Ces hormones agissent de concert avec des facteurs de croissance (« Epidemial Growth Factor »

EGF ou « Insuline-like Growth Factor » IGF) et des constituants de l'épithélium et du mésenchyme[9].

Pendant la **période prénatale**, les glandes mammaires **ne requièrent pas d'hormones** pour se développer même si le fœtus est exposé aux hormones placentaires (notamment les œstrogènes, la progestérone et la prolactine). La présence de **prolactine** dans le liquide amniotique provoque lors de la naissance, la sécrétion d'un liquide appelé « lait de sorcière[a] », chez les filles comme chez les garçons.

Comme pendant la période prénatale, **avant la puberté** le développement des glandes mammaires est **indépendant des hormones** spécifiques de la reproduction[10]. Il suit la croissance de tous les autres organes sous l'influence de la **GH** et des **hormones thyroïdiennes**. Ces dernières possèdent des récepteurs dans les tissus mammaires permettant de stimuler l'expression de récepteurs de différents facteurs de croissance tels que l'**IGF-1** ou l'**EGF**. La GH a, quant à elle, un effet proliératif sur les cellules mammaires par le biais de l'IFG-1. Chez le fœtus mâle, le tissu mammaire subit une nécrose partielle qui empêche le développement de la glande mammaire à l'âge adulte. Ce phénomène est provoqué par la production de testostérone par les testicules, ce qui induit l'apoptose des bourgeons terminaux[11]. Cette nécrose est soutenue par l'absence de progestérone chez les hommes[12–14].

À partir de la puberté, les principaux développements de la glande mammaire ont lieu au cours des premiers cycles sexuels. L'hormone de libération des gonadotrophines hypophysaires (**GnRH**) sécrétée par l'hypothalamus induit la production des hormones lutéinisantes (appelées aussi hormones gonadotropes), la **FSH** et la **LH**, par l'hypophyse[15]. Elle sécrète de façon progressive une quantité de plus en plus importante de FSH et de LH entre l'âge de 9 et 12 ans, jusqu'à atteindre un seuil suffisant entre l'âge de 11 et 15 ans pour déclencher les cycles sexuels[16]. Ces hormones sont alors acheminées vers les ovaires par le sang et entrainent la production d'œstrogènes et de progestérone. Ces hormones sont à l'origine des cycles sexuels qui sont divisés en deux phases.

La première phase, appelée phase de prolifération, est sous le contrôle des **œstrogènes**. L'action des œstrogènes se fait par le biais des **récepteurs aux œstrogènes (ER)** existant

[a] Autrefois, la croyance populaire faisait de ce lait la boisson préférée des « sorcières ». Les sages-femmes devaient alors le tirer le plus rapidement possible afin de préserver le bébé de la folie.

sous 2 formes – alpha et beta (respectivement codées par les gènes *ESR1* et *ESR2*)[17] – et présents sur la paroi des cellules. La fixation des œstrogènes sur ces récepteurs va leur permettre d'être dirigés vers le noyau et de se fixer sur des séquences spécifiques de l'ADN, les éléments de réponses aux œstrogènes (ERE). Ces séquences d'ADN se trouvent au niveau du promoteur de gènes cibles impliqués dans la prolifération, le développement et la différenciation mammaire. Cette fixation permet d'activer ou d'inhiber l'expression de ces gènes. Ils entrainent alors la prolifération des cellules épithéliales qui forment un réseau de canaux à partir du mamelon et se terminent par des bulbes. La présence d'œstrogènes entraine également la multiplication des cellules adipeuses[16]. Le site d'activité mitotique et de prolifération cellulaire intense se situe dans la région antérieure des bourgeons terminaux qui envahissent le tissu adipeux[14,18]. L'**hormone de croissance GH** participe avec les œstrogènes à la croissance des canaux galactophores[19,20].

La deuxième phase du cycle sexuel, la phase lutéale, est contrôlée par la production de **progestérone** dont l'action se fait par le biais des **récepteurs à la progestérone (PR)**, existant également sous deux formes, *PRA* et *PRB*, codées par le gène *PGR* (deux isoformes distinctes)[21]. L'action de cette hormone est encore controversée. En effet, les récepteurs à la progestérone s'expriment en réponse à l'activation des récepteurs aux œstrogènes, ce qui ne permet pas d'étudier les deux hormones de façon indépendante. Par conséquent, les effets spécifiques de la progestérone sont difficiles à déterminer[22]. Elle semble assurer la différenciation cellulaire et la maturation des lobules mammaires à partir desquels sera sécrété le lait maternel lors de l'allaitement[16,23]. À la fin de chaque cycle, la diminution brutale du taux de progestérone entraine une régression des canaux galactophores et par conséquent le volume des seins diminue légèrement.

Pendant la grossesse, des **œstrogènes** et de la **progestérone** sont sécrétés par le corps jaune ovarien issu de l'ovulation, mais également par le placenta pendant du premier tiers de la grossesse. À l'approche de l'accouchement, les taux d'œstrogènes et de progestérone sont respectivement 30 et 10 fois plus élevés qu'en temps normal[16]. Les œstrogènes préparent les glandes mammaires à l'allaitement en stimulant le développement du système canalaire. La progestérone active de façon très importante la différenciation lobulaire.
Au cours de la grossesse, une autre hormone intervient dans le développement des glandes mammaires : la **prolactine**. Cette dernière est l'hormone lactogène essentielle. Sécrétée par

l'hypophyse, elle reste à un niveau relativement bas pendant les deux premiers tiers de la grossesse et augmente ensuite de façon progressive jusqu'à atteindre des taux très élevés au moment de l'accouchement[14].
Parallèlement à son effet activateur de la différenciation lobulaire, la progestérone a aussi un effet inhibiteur. En effet, la sécrétion lactée permise par la prolactine est freinée par la progestérone, en particulier en s'opposant à l'augmentation de la synthèse **des récepteurs à la prolactine**. Au niveau hypophysaire, la progestérone inhibe la sécrétion de la prolactine, alors qu'au niveau mammaire, elle empêche l'action de la prolactine sur la production de certaines protéines composant le lactose.

À la fin de la grossesse, le taux d'œstrogènes dans les glandes mammaires permet un pic de production de prolactine par l'hypophyse.Simultanément, la production de progestérone diminue brutalement par involution du placenta et du corps jaune, ce qui participe au déclenchement de l'accouchement et de la lactogénèse.

Les hormones du métabolisme général (l'**insuline**, les glucocorticoïdes comme le **cortisol** mais aussi les hormones thyroïdiennes comme la **thyroxine**) interviennent également dans le développement de la glande mammaire pendant la grossesse, mais leur rôle, bien qu'essentiel, est moins bien caractérisé[24].

Pendant les **sept premiers jours après l'accouchement**, la **prolactine** permet la sécrétion d'une quantité très importante de lait. Cette sécrétion requiert également la présence d'autres hormones essentielles (GH, cortisol, thyroxine et insuline) qui fournissent les acides aminés, les acides gras, le glucose et le calcium nécessaires à la formation du lait[25]. Ensuite, la quantité de prolactine diminue pour revenir au taux précédant la grossesse. Au début de la tétée, le nouveau-né ne reçoit pas de lait, mais cela provoque un reflex au niveau du système nerveux de la mère qui entraine la production de prolactine et permet la libération du lait. De ce fait, en absence de tétée, il n'y a pas de production de prolactine et les glandes mammaires perdent leur capacité à produire du lait au bout d'environ une semaine.
Le lait est sécrété dans les glandes mammaires grâce à l'action de la prolactine mais il est nécessaire de l'acheminer des alvéoles jusqu'aux tétons. C'est le rôle de **l'ocytocine (OXT)**. Cette hormone est libérée grâce à plusieurs stimuli pré-allaitement (pleurs, agitation du nouveau-né ou préparation de la mère à l'allaitement)[26,27]. Elle est alors transportée par le

sang du cerveau jusqu'aux seins et entraine la contraction des cellules mammaires permettant l'éjection du lait[16].

Lors de la ménopause, l'activité ovarienne cesse. La diminution du taux d'hormones sexuelles entraine alors une réduction du volume des seins. Cette réduction est due à la diminution des glandes mammaires, à l'atrophie du tissu sécréteur et des canaux galactophores. Après la ménopause, les seins sont essentiellement constitués de tissus adipeux.

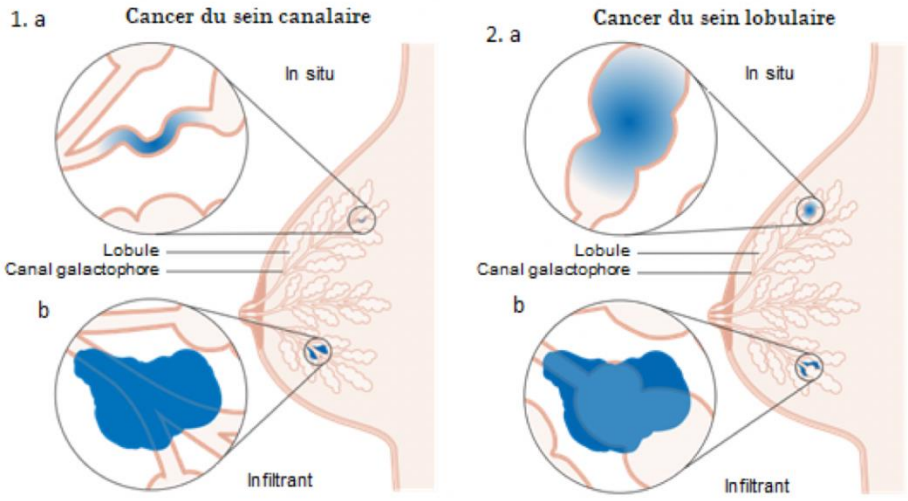
Les glandes mammaires subissent donc **tout au long de la vie des changements physiologiques** qui diffèrent **en fonction du stade de développement** de la femme et **des épisodes reproductifs.** Cela entraine de **fortes variations d'exposition des glandes mammaires aux différentes hormones impliquées dans la reproduction** et par conséquent de très nombreuses phases de **prolifération cellulaire** suivies de phases de **différenciation cellulaire** et de **maturation.**

Le cancer du sein

I. Physiopathologie

Le cancer du sein correspond à une tumeur maligne localisée dans les glandes mammaires. Plusieurs types de cancer du sein se distinguent selon leur niveau d'infiltration tumorale, avec les carcinomes non-infiltrants (ou *in situ*) et les carcinomes infiltrants (ou invasifs) mais également selon leur type histologique, avec les cancers canalaires qui se développent dans les canaux galactophores et les cancers lobulaires qui se développent à partir d'un lobule mammaire (Figure 4).

Figure 4 - Localisation de la tumeur selon le type de cancer du sein – 1. canalaire ou 2. lobulaire – et son niveau d'infiltration – a. in situ ou b. infiltrant.



1. Les carcinomes canalaires ou lobulaires *in situ*

On parle de cancer du sein ***in situ*** lorsque les cellules tumorales prolifèrent à l'intérieur des canaux ou des lobules (Figure 5). La membrane basale n'ayant pas été rompue, les cellules cancéreuses n'ont pas infiltré le tissu voisin. Il n'y a donc théoriquement pas de risque d'envahissement ganglionaire.

Figure 5 - Les différents types d'infiltration tumorale

Le cancer canalaire *in situ* (CCIS) est le plus fréquent. Il représente 15 à 20 % des cancers du sein. Parmi les cancers *in situ*, 80 à 90 % sont d'origine canalaire (Figure 4 – 1.a).

Le cancer lobulaire *in situ* (CLIS) est beaucoup plus rare, représentant seulement 0,5 % des cancers du sein et 10 à 15 % des cancers *in situ* (Figure 4 – 1.b). D'un point de vue clinique, les cancers lobulaires *in situ* sont très souvent considérés comme des facteurs de risque de cancer du sein infiltrant, multipliant par 8 le risque par rapport à celui estimé dans la population générale. Contrairement aux autres types de cancers du sein, leur prise en charge ne requiert pas de traitement mais nécessite une surveillance accrue car la fréquence de survenue d'un cancer infiltrant 10 à 25 ans après le diagnostic d'un CLIS est de 20 à 30 %.

2. Les carcinomes canalaires ou lobulaires infiltrants

On parle de cancer du sein infiltrant lorsque les cellules cancéreuses ont dépassé la membrane basale et infiltré le tissu voisin (Figure 4). Plus de 75 % des cancers infiltrants sont d'origine canalaire (Figure 4 – 2.a). Les cancers lobulaires infiltrants (Figure 4 – 2.b) sont plus rares (5 à 10 %). D'autres formes encore plus rares, telles que le carcinome médullaire, mucineux, tubuleux ou encore papillaire ne représentent qu'1 ou 2 % des cancers infiltrants.

Les tumeurs du sein présentent donc une grande hétérogénéité qui existe également, dans certains cas, à l'intérieur d'une même tumeur. On parle alors d'hétérogénéité intra-tumorale.

II. Évaluation pronostique des cancers du sein

1. Stade

Le stade d'un cancer – qui correspond à son étendue dans la glande mammaire – est déterminé grâce aux examens de diagnostic. L'examen clinique réalisé avant tout traitement permet de définir le stade dit « pré-thérapeutique » du cancer. Après la chirurgie, un examen des tumeurs prélevées est réalisé et permet de définir le stade dit « anatomopathologique ». Trois critères sont pris en compte pour évaluer le stade du cancer du sein : la taille et l'infiltration de la tumeur, l'envahissement ou non des ganglions lymphatiques et la présence ou non de métastases. On parle de classification TNM pour « Tumor, Nodes, Metastasis », avec la classification cTNM pour l'examen clinique et pTNM pour l'examen post-chirurgical :

- La **taille et l'infiltration de la tumeur** donnent une indication sur le degré d'évolution de la maladie. En effet, lorsque les cellules cancéreuses apparaissent, elles forment d'abord une tumeur au niveau des canaux ou des lobules (*in situ*) puis cette tumeur devient progressivement infiltrante en traversant la membrane basale du canal ou du lobule.

- Le **nombre de ganglions envahis** par les cellules cancéreuses et leur localisation indiquent également le degré de propagation du cancer. Les ganglions lymphatiques sont les premiers touchés lorsque les cellules cancéreuses s'échappent des glandes mammaires. Si des ganglions sont atteints, cela signifie que la maladie a commencé à se disséminer.

- La **présence de métastases** indique également le degré de propagation du cancer. Les cellules cancéreuses peuvent envahir d'autres organes, le plus souvent le foie, les os et les poumons dans le cas du cancer du sein, et y développer des métastases.

2. Grade

Chaque cancer agit avec une agressivité différente. Cette agressivité est illustrée par le grade qui est défini lors de l'examen anatomopathologique. Trois paramètres sont alors évalués et permettent de définir le grade (Tableau 1) : l'**architecture tumorale** (moins les cellules cancéreuses ressemblent aux cellules mammaires normales et plus elles sont indifférenciées, plus elles sont agressives), la **forme du noyau** et le **nombre de cellules en division** (plus une cellule cancéreuse se divise vite, plus le risque de propagation du cancer est important).

Tableau 1 - Calcul du grade du cancer du sein (source - INCa)

Critère	Note	
	1	**3**
Architecture	La tumeur contient beaucoup de structures bien formées.	La tumeur contient peu ou pas du tout de structures bien formées.
Noyau	Les noyaux de la tumeur sont petits et uniformes.	Les noyaux de la tumeur sont gros et leur taille et leur forme varient.
Activité mitotique	Les cellules de la tumeur se divisent lentement = faible nombre de mitoses.	Les cellules de la tumeur se divisent rapidement = important nombre de mitoses.

Le grade prend une valeur de I à III, avec une agressivité croissante. Ce grade dépend de la note obtenue pour chacun des critères (Tableau 1). Un score final est calculé en ajoutant la note obtenue pour les trois paramètres. Un score de 3 à 5 correspond à un grade de stade I alors qu'un score entre 8 et 9 correspond à une tumeur très agressive de grade III.

3. Hétérogénéité moléculaire

Des **récepteurs aux œstrogènes**, **à la progestérone** et aux **facteurs de croissance** sont présents dans la cellule mammaire normale à des taux très faibles et sont impliqués dans de nombreux processus physiologiques, notamment la fonction de reproduction. Ces récepteurs sont des marqueurs majeurs dans les cellules tumorales. Leur niveau d'expression varie d'une tumeur à l'autre. Ces récepteurs peuvent être sur- ou sous-exprimés dans les cellules de la tumeur altérant alors l'activité des protéines ligands. Le niveau d'expression de ces récepteurs est également un facteur pronostique et prédictif de la maladie, avec un taux de survie plus ou moins élevé selon les récepteurs exprimés.

a. Récepteurs aux œstrogènes

Environ 70 % des cancers du sein expriment les **récepteurs aux œstrogènes**. On parle alors de cancer du sein positif aux récepteurs aux œstrogènes (RE^+). Le niveau d'expression de ces récepteurs y est beaucoup plus important que dans les glandes mammaires normales, où le taux d'expression est très faible[28]. Au bout de cinq ans, les femmes avec une tumeur RE^+ ont une survie globale de 92 % et une survie sans maladie de 74 % alors que celles ayant une tumeur RE^- ont une survie de 82 % et 66 % respectivement[29]. La présence de ces récepteurs augmente la probabilité de réussite d'un traitement par hormonothérapie. En 1998, l'Early Breast Cancer Trialists' Collaborative Group montre, à partir d'une étude portant sur 37 000 femmes, une réduction de 47 % de la récurrence du cancer et de 26 % de la mortalité, réduction liée à la prise de tamoxifène chez des femmes avec une tumeur RE^+[30]. L'utilisation du tamoxifène n'a pas de bénéfice pour les tumeurs RE^-.

b. Récepteurs à la progestérone

Comme pour les récepteurs aux œstrogènes, le taux de **récepteurs à la progestérone** est faible dans les glandes mammaires normales[28]. Plus de la moitié des tumeurs RE^+ sont également positives aux récepteurs à la progestérone (RP^+). L'expression des récepteurs RP est dépendante des œstrogènes. C'est pourquoi les tumeurs RE^-/RP^+ sont très rares. Il est donc difficile d'analyser le rôle pronostique de RP[31]. Bardou *et al.*[32] ont montré en 2004 un

taux de réussite plus important de l'hormonothérapie pour les tumeurs RE^+/RP^+ par rapport aux tumeurs RE^+/RP^-. Cependant, Dowsett *et al.*[33] ne mettent pas en évidence de différence du bénéfice du tamoxifène entre les individus RE^+/RP^+ et ceux RE^+/RP^-.

c. Récepteurs aux facteurs de croissances : HER2

La protéine **HER2** fait partie d'une famille de récepteurs transmembranaires impliqués dans la modulation de la prolifération et de la survie cellulaire normale. Environ 20 % des cancers du sein expriment ce récepteur et sont associés à un mauvais pronostic. La majorité des tumeurs $HER2^+$ sont résistantes aux hormonothérapies.

Plus récemment, les avancées technologiques de la génomique ont permis de proposer une classification des tumeurs en fonction de leur profil moléculaire. Les cancers du sein sont classés en 4 sous-types qui diffèrent au niveau du pronostic et de la réponse aux traitements. Ce classement est principalement basé sur l'expression des récepteurs ER, PR et HER2.

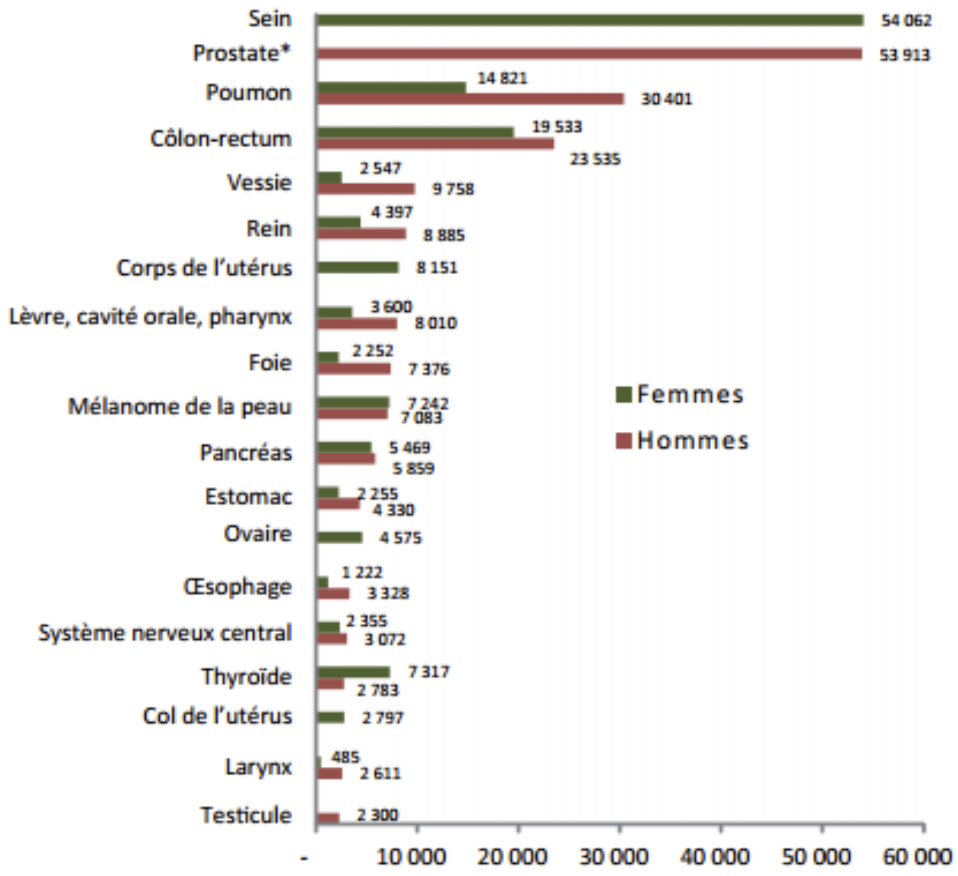
- Les tumeurs de type ***basal-like*** représentent environ 15 % des carcinomes canalaires invasifs. Elles correspondent aux tumeurs triples négatives n'exprimant ni les récepteurs hormonaux ni la protéine HER2. Ce sont des tumeurs de haut grade (grade III) avec une activité mitotique élevée et donc associées à un mauvais pronostic.
- Les tumeurs de sous-type **luminal A** expriment les récepteurs RE avec une surexpression du gène *ESR1* et des gènes régulés par les œstrogènes. Elles représentent environ 40 % des carcinomes canalaires invasifs. Elles sont le plus généralement de bas grade avec une activité mitotique faible et donc associées à un bon pronostic.
- Les tumeurs de sous-type **luminal B** ont les mêmes caractéristiques que celles de sous-type luminal A mais sont associées à un pronostic moins favorable. Elles représentent environ 20 % des tumeurs.
- Enfin, les tumeurs de type ***HER2-like*** sur-expriment le gène *HER2*. Ce sont des tumeurs agressives associées à un grade élevé et de mauvais pronostic. L'origine de cette hétérogénéité tumorale est encore largement inconnue.

III. Épidémiologie descriptive du cancer du sein

Le cancer du sein est le **cancer le plus fréquent chez la femme** (Figure 6) en France[34] avec 54 000 nouveaux cas par an, mais également dans le monde entier[35] avec 1,7 million de nouveaux cas par an. C'est le **deuxième cancer le plus fréquent** dans la population mondiale après le cancer de la prostate[34].

Il représente plus d'un tiers de l'ensemble des nouveaux cas de cancer chez la femme. Entre 1980 et 2005, le nombre de cancers du sein diagnostiqués a augmenté de 138 %. L'allongement de la durée de vie ainsi que la progression du dépistage organisé peuvent expliquer cette forte augmentation mais l'évolution des facteurs environnementaux et comportementaux y participe également. Cependant, depuis 2005, le nombre de cas observés chaque année a tendance à diminuer.

Figure 6 - Classement des cancers en fonction du nombre de sujets atteints en 2015 en France métropolitaine par localisation et par sexe[34]



Le cancer du sein est associé à un bon pronostic et la survie tend à s'améliorer avec une survie standardisée à 5 ans de 80 % pour les cas diagnostiqués entre 1989 et 1993 et de 87 % pour ceux diagnostiqués entre 2005 et 2010[34]. Cette augmentation de la survie peut être expliquée par une plus grande précocité du diagnostic permettant de prendre en charge les malades à un stade moins agressif de la maladie ainsi que par les différents progrès thérapeutiques. Cette maladie reste tout de même la première cause de décès par cancer chez les femmes en 2015, avec plus de 11 000 décès en France.

Un cancer du sein peut également survenir chez les hommes, mais cela reste exceptionnel, avec une incidence de 1 pour 100 000 et moins de 1 % de tous les cancers du sein.

IV. Le dépistage

1. Le dépistage organisé

Le programme national de **dépistage organisé** des cancers du sein a été mis en place en France au début de l'année 2004. Ce programme cible les femmes de 50 à 74 ans de la **population générale**, c'est-à-dire sans histoire familiale de cancer du sein ni facteurs de risque particuliers (antécédents personnels de cancer du sein ou image anormale lors d'une mammographie). À partir de 50 ans, ces femmes sont invitées à effectuer une mammographie de contrôle tous les 2 ans ainsi qu'un examen clinique des seins (palpation par un professionnel de la santé). Ce dépistage a pour but de réduire la mortalité liée au cancer du sein mais également d'améliorer l'information et la qualité des soins apportés aux femmes atteintes, en garantissant un accès au dépistage identique sur l'ensemble du territoire.

Pour la période 2013-2014, un total de 36 889 cancers (invasifs et *in situ*, à l'exception des carcinomes lobulaires *in situ*) ont été enregistrés dans le cadre du dépistage organisé soit 7,4 cancers diagnostiqués pour 1 000 femmes dépistées. En 2016, 50,7 % de la population cible a participé au programme, avec une légère augmentation au cours de la période 2015-2016.

2. Familles à haut risque de cancer du sein

Les femmes présentant un risque de développer un cancer du sein plus élevé que celui de la population générale ne sont pas concernées par le dépistage organisé. Ces femmes sont **à risque élevé de cancer du sein** à cause d'antécédents personnels de cancer du sein, d'une image anormale lors de la dernière mammographie de contrôle ou d'un diagnostic histologique d'hyperplasie atypique du sein[b] ou **à très haut risque de cancer du sein** à cause d'une mutation constitutionnelle sur un gène de prédisposition au cancer du sein (voir le paragraphe page 41). Elles ont alors un suivi spécifique et une surveillance accrue commençant avant l'âge prévu par le dépistage organisé.

[b] L'hyperplasie est une affection bénigne du sein due à la prolifération de cellules qui tapissent les canaux ou les lobules du sein.

Les femmes ayant des **antécédents personnels de cancer du sein** sont classées dans la catégorie à haut risque et les modalités de suivi diffèrent en fonction de leurs antécédents. Pour les femmes ayant eu un cancer du sein ou un carcinome canalaire *in situ*, le suivi consiste en un examen clinique tous les six mois pendant les 2 ans suivant la fin du traitement. Ce suivi est ensuite réalisé tous les ans. Pour les femmes ayant un diagnostic d'hyperplasie atypique, une mammographie annuelle est préconisée pendant 10 ans. Après cette période de 10 ans, elles sont redirigées vers le dépistage organisé.

Les femmes ayant des **antécédents de cancer du sein ou de l'ovaire chez des femmes de leur famille** (mère, sœur(s) ou fille(s)), ou de **cancer du sein chez des hommes de leur famille**, ont un risque potentiellement très élevé de cancer du sein. On parle de **syndrome seins-ovaires** (ou **HBOC** en anglais, pour « Hereditary Breast and Ovarian Cancer »[36]). Cette histoire familiale de cancers gynécologiques suggère la présence d'une altération génétique constitutionnelle (mutation) favorisant le développement de la maladie. Ces femmes peuvent être orientées vers une **consultation d'oncogénétique** pour envisager la recherche de mutations.

L'oncogénétique est une discipline qui s'est organisée dans les années 90. Les premières consultations d'oncogénétique ont été organisées en France sous l'impulsion du Groupe Génétique et Cancer (GGC) qui a été créé en 1991 au sein de la Fédération nationale des Centres de lutte contre le cancer (FNCLCC) connue sous le nom du groupement de coopération sanitaire de moyens UNICANCER depuis 2010. En 2003, le Plan Cancer national a permis de renforcer et structurer cette discipline entrainant alors une amélioration de l'accès aux consultations et aux tests génétiques. Aujourd'hui, en France, le dispositif national d'oncogénétique s'organise autour de 147 sites de consultations (Figure 7) se répartissant dans 104 villes (France métropolitaine et départements d'outre-mer) et 25 laboratoires prennent en charge la réalisation des tests génétiques. La mise en place, l'organisation et la structuration de ces consultations d'oncogénétique est différente et spécifique à chaque pays.

Ce dispositif permet d'identifier les personnes génétiquement prédisposées au cancer du sein, qu'il s'agisse de personnes atteintes de la maladie ou de membres qui n'ont pas développé la maladie au moment de la consultation (apparentés non atteints). Selon le parcours type (Figure 8) la première personne reçue en consultation d'oncogénétique est le cas ayant,

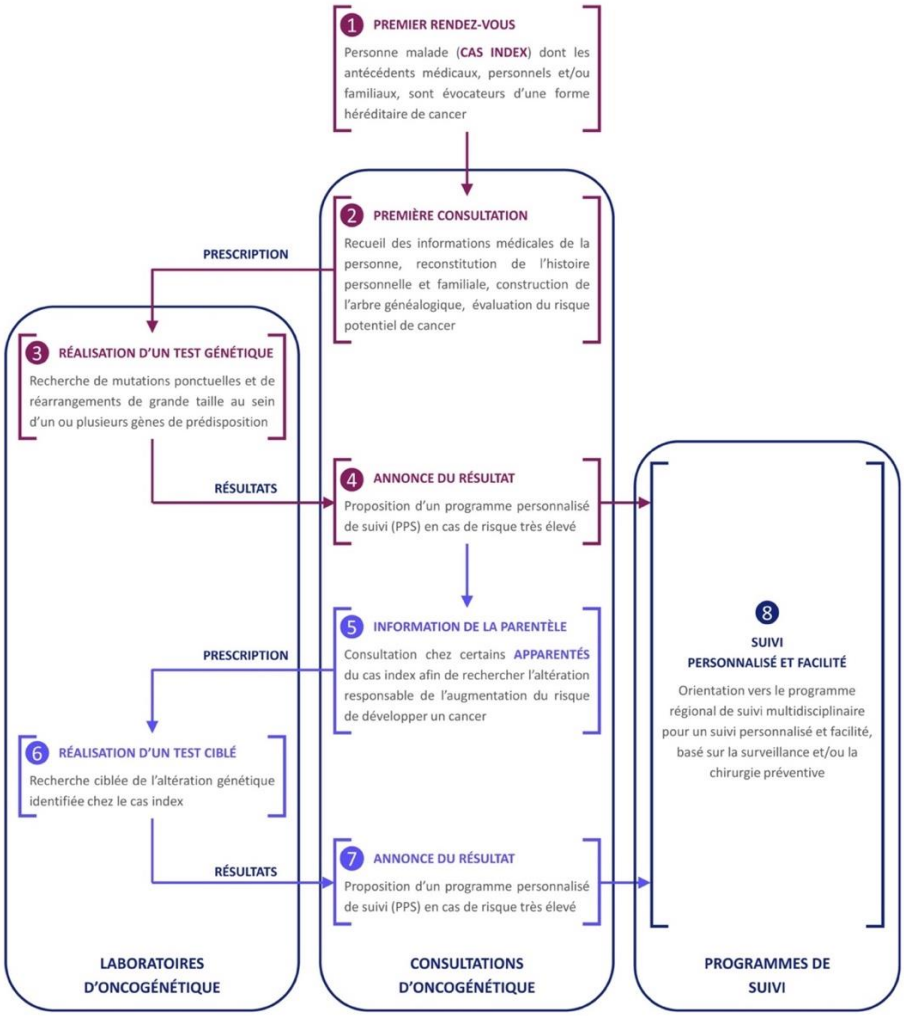
d'après le schéma familial, la plus forte probabilité d'être porteuse de l'altération génétique au sein d'une famille. On parle de **cas index**.

Figure 7 - Carte de France des différents sites de consultations d'oncogénétique (INCa)

La prise en charge et le suivi des femmes présentant une prédisposition génétique au cancer du sein et/ou de l'ovaire est en constante évolution suivant les avancées de la recherche sur les

facteurs de risque du cancer du sein et plus particulièrement des facteurs génétiques. Un certain nombre de facteurs génétiques sont aujourd'hui parfaitement identifiés comme intervenant dans le développement du cancer du sein. C'est le cas des gènes *BRCA1* et *BRCA2* même si tous les mécanismes ne sont pas encore bien compris. D'autres gènes sont connus comme intervenant dans l'étiologie du cancer du sein depuis longtemps et certains d'entre eux ont récemment été ajoutés au panel de gènes testés lors des consultations d'oncogénétique[37]. C'est le cas des gènes *PALB2* en 2015 ou *TP53* en 2018 (voir paragraphe « 2. Les facteurs génétiques », page 41).

Figure 8 - Parcours global des cas index et des apparentés en oncogénétique (INCa)



Cependant, **plus de 50 % des cas de cancer du sein des femmes HBOC restent inexpliqués. La recherche de facteurs génétiques et non génétiques** est donc nécessaire pour **permettre une prise en charge adaptée de ces patientes et de leur famille.**

V. Les facteurs de risque du cancer du sein

Le cancer du sein est une **maladie multifactorielle** dont le premier facteur de risque est l'âge. Le risque de cancer du sein augmente également avec certains **facteurs d'origine gynéco-obstétrique** (âge à la puberté, nombre d'enfants, âge à la première grossesse, allaitement, etc.) ou **liés au mode de vie** (utilisation de traitements hormonaux, surpoids, consommation d'alcool, tabagisme, etc.) ou encore à **l'environnement** (exposition aux rayonnements ionisants).

1. Les facteurs environnementaux et de mode de vie

Dans ce manuscrit, nous emploierons le terme « facteurs environnementaux » au sens large, pour tous les facteurs non génétiques, que ce soit les facteurs liés au mode de vie, les facteurs gynéco-obstétriques, la densité mammaire ou les expositions aux radiations.

a. Variables démographiques

L'**âge** est l'un des facteurs de risque de cancer du sein les plus importants. Le risque de développer cette maladie augmente avec l'âge, avec environ 10 % des cas âgés de moins de 35 ans et 20 % des cas âgés de moins de 50 ans. Selon le rapport du SEER[38,39], le risque individuel est proche de 1/1 500 pour les femmes de moins de 30 ans et de 1/12 pour les femmes de 70 ans dans la population générale. Près de 50 % des cas de cancer du sein sont diagnostiqués entre 50 et 69 ans. L'âge médian au diagnostic est de 63 ans.

b. Facteurs liés au mode de vie

- *Alcool et tabac*

Le lien entre la **consommation d'alcool** et le risque de cancer du sein est connu depuis le début des années 80. En 2012, une méta-analyse réalisée par Seitz *et al*.[40] sur les résultats de 113 études montre une association significative, bien que faible, de la consommation d'alcool

avec le cancer du sein (risque relatif (RR) combiné = 1,04 ; $IC_{95\%}$ = [1,01-1,07]). Une forte consommation, définie par la consommation de 3 verres d'alcool ou plus par jour, est associée à une augmentation du 40 à 50 % du risque de cancer du sein[40,41]. D'autres études cas-témoins ont également montré une augmentation du risque chez les femmes ayant commencé à consommer de l'alcool avant l'âge de 30 ans[42,43].

Au contraire, l'association entre la **consommation de tabac** et le cancer du sein reste controversée. Aucune association n'a été trouvée dans l'étude collaborative de Hamajima *et al.*[44] incluant plus de 58 000 cas et 95 000 témoins de cancer du sein. Une étude de cohorte récente impliquant 50 884 femmes n'a également mis en évidence aucun lien entre le cancer du sein et la consommation active ou passive de tabac. Cependant, cette étude montre une augmentation du risque de 17 % ($IC_{95\%}$ = [1,00 %-1,36 %]) chez les femmes confrontées au tabagisme passif dans l'enfance et de 16 % ($IC_{95\%}$ = [1,01-1,32]) pour celles qui l'ont été in utero[45]. D'autres études ont mis en évidence une augmentation du risque associé à une consommation de tabac jeune, avant l'apparition des premières règles ou la première grossesse menée à terme[46,47].

Alimentation, activité physique et indice de masse corporelle (IMC)

Bien que le lien entre l'**alimentation** et le développement d'un cancer du sein reste incertain, certaines études ont montré une augmentation ou une diminution du risque selon le type de régime alimentaire. Une étude réalisée à partir des données de la cohorte française E3N a montré qu'un régime alimentaire de type occidental incluant l'alcool était associé à une légère augmentation du risque de cancer du sein (hasard ratio (HR) = 1,20 ; $IC_{95\%}$ = [1,03-1,38])[48]. À l'inverse, cette étude montre une légère diminution du risque associé à un régime alimentaire de type méditerranéen (HR = 0,85 ; $IC_{95\%}$ = [0,75-0,95])[48]. Trichopoulou *et al.*[49] montrent également, par le biais de l'étude Greek EPIC, une légère diminution du risque associé au régime méditerranéen chez les femmes ménopausées (HR = 0,78, $IC_{95\%}$ = [0,62-0,98]). Cependant, aucun lien n'a été mis en évidence entre le risque de cancer du sein et la consommation de fruits et de légumes, ni celles de vitamines, de minéraux ou d'oligo-éléments[50–52].

L'**activité sportive** régulière est associée à une diminution du risque de cancer du sein avec un risque relatif de 0,88 ($IC_{95\%}$ = [0,95-0,91]) par rapport aux femmes sans activité sportive[47]. Cette diminution du risque est retrouvée que ce soit pour une activité sportive professionnelle ou non professionnelle. La méta-analyse réalisée par Neil-Sztramko *et al.*[53] en 2017 montre une diminution significative du risque de cancer du sein chez les femmes ménopausées avec une forte activité sportive et un indice de masse corporelle (IMC = $\frac{poids\ (kg)}{taille^2(cm)}$) inférieur ou égal à 25 kg/cm^2 (RR = 0,87, $IC_{95\%}$ = [0,81-0,93]). Cette diminution de risque n'est pas significative pour les femmes en surpoids (IMC > 31 kg/cm^2, RR = 0,93, $IC_{95\%}$ = [0,76-1,13]).

En effet, le poids, et plus précisément l'**IMC**, est également impliqué dans le développement du cancer du sein et le sens de l'association semble dépendre du statut ménopausique. L'association est négative avant la ménopause (RR = 0,98 par unité d'IMC, $IC_{95\%}$ = [0,97-0,99])[54]. Ce résultat peut s'expliquer par le fait que chez les femmes non ménopausées, un surpoids (IMC > 31 kg/cm^2) diminue le nombre d'ovulations et le taux d'œstrogènes sanguins, facteurs augmentant le risque de cancer du sein. Après la ménopause, un surpoids augmente le risque de cancer du sein (RR = 1,03 par unité d'IMC, $IC_{95\%}$ = [1,02-1,04]). Cette augmentation peut être expliquée par une production plus importante d'hormones, notamment d'œstrogènes, dans les tissus adipeux.

Ces facteurs liés au mode de vie ont un faible effet sur le cancer du sein en comparaison avec d'autres types de cancer comme le cancer du poumon dont le risque est multiplié par 15 par la consommation de tabac ou le cancer de l'œsophage dont le risque est multiplié par 40 par la consommation de tabac et d'alcool.

c. Facteurs gynéco-obstétriques

Beaucoup d'études épidémiologiques montrent une association de facteurs gynéco-obstétriques avec le risque de cancer du sein mais certains d'entre eux restent controversés.

- *Âge aux premières règles et ménopause*

L'**âge aux premières règles** est un facteur de risque de cancer du sein, avec un risque 2,2 fois supérieur ($IC_{95\%}$ = [1,2-4,0]) pour les femmes ayant eu leurs premières règles entre 10 et 11

ans, comparé à celles les ayant eu à 12 ans et après[55]. Le risque global de cancer du sein diminue de 3 % (RR = 0,97 ; $IC_{95\%}$ = [0,93–0,99]) par année supplémentaire[56,57]. Ce risque semble varier selon le type de tumeur. En effet, une étude « *case-only* » sur 35 568 cas de cancer du sein invasif montre qu'un âge précoce aux premières règles est moins fréquent chez les sous-types de cancer du sein RP^- que ceux RP^+ (OR = 0,88, $IC_{95\%}$ = [0,81–0,95])[58]. Selon Clavel-Chapelon *et al*.[56], cette association varie selon le statut ménopausique de la femme au moment du diagnostic de cancer du sein. Cette étude de cohorte réalisée sur 91 260 femmes, dont 1 718 cas de cancer du sein, montre une diminution du risque de 7 % par année (RR = 0,93 ; $IC_{95\%}$ = [0,87-0,99]) pour les femmes diagnostiquées avant leur ménopause et de 3 % par année (RR = 0,97 ; $IC_{95\%}$ = [0,93-1,01]) pour celles diagnostiquées après leur ménopause.

Une **ménopause** précoce est associée à une diminution du risque de cancer du sein. Le Collaborative Group on Hormonal Factors in Breast Cancer[57] a estimé le risque de cancer du sein à 1,43 ($IC_{95\%}$ = [1,33-1,52]) chez les femmes non ménopausées par rapport aux femmes ménopausées. Il montre également que la diminution du risque commence au cours de la périménopause[c]. Le risque des femmes en périménopause, comparé à celui des femmes ménopausées, est estimé à 1,24. La diminution progressive du risque au cours de cette période transitoire est corrélée à la diminution du taux d'hormones sexuelles, et plus particulièrement d'œstradiol, dans le sang. L'étude montre que le risque relatif de cancer du sein est augmenté de 1,029 ($IC_{95\%}$ = [1,025-1,032]) par an. Comme pour l'âge aux premières règles, il existe également une hétérogénéité de l'effet de la ménopause selon le sous-type moléculaire ($RE^{+/-}$) et selon le type histologique (lobulaire ou canalaire)[57].

Les premières règles et la ménopause marquent respectivement le **début et la fin de l'activité ovarienne,** et ainsi **la production des différentes hormones impliquées dans la reproduction.** C'est pourquoi, **plus le temps entre ces deux évènements est important, plus la fenêtre d'exposition aux hormones sexuelles est longue** et donc **plus le risque de cancer du sein est élevé**[57].

[c] Périménopause = période de transition hormonale précédant l'arrêt complet des règles, pouvant débuter jusqu'à une dizaine d'années avant la ménopause

Contraceptifs oraux

Le lien entre la prise de **contraceptifs oraux** et le cancer du sein est controversé. En 1996, Le collaborative Group on Hormonal Factor in Breast Cancer a réalisé une méta-analyse regroupant les données de plus de 53 000 femmes atteintes de cancer du sein et 100 000 femmes non atteintes. Cette méta-analyse a permis de montrer une association significative, bien que faible, entre le cancer du sein et l'utilisation de contraceptifs[59]. En 2017, Lina *et al*.[60] montrent une augmentation du risque de 1,20 ($IC_{95\%}$ = [1,14-1,26]) pour les femmes utilisant des contraceptifs hormonaux. Ce risque augmente de 1,09 ($IC_{95\%}$ = [0,96-1,23]) pour une utilisation cumulée inférieure à 1 an, à 1,38 ($IC_{95\%}$ = [1,26-1,51]) si cette utilisation dépasse 10 ans. Même après 5 ans d'arrêt de contraception hormonale, l'augmentation du risque semble persister.
Cependant, cette association n'est pas mise en évidence dans d'autres études comme celles de Marchbanks *et al*.[61] ou Hannaford *et al*.[62].

Grossesses

Dans la population générale, le risque de cancer du sein d'une femme nullipare est plus élevé que celui d'une femme ayant eu au moins un enfant, avec un risque avoisinant 1,5[63]. Parmi les femmes primi- ou multipares, le risque est d'autant plus important que la **première grossesse menée à terme** a lieu à un âge tardif[56], avec une augmentation de 3 % par an (RR = 1,03, $IC_{95\%}$ = [1,01-1,04]).

Une femme ayant eu sa première grossesse menée à terme avant l'âge de 20 ans a un risque diminué de 50 % par rapport aux femmes nullipares et de 33 % par rapport à celles ayant eu leur première grossesse menée à terme à 35 ans ou plus[64]. La première grossesse menée à terme correspond à la dernière étape du développement des glandes mammaires (voir paragraphe « I. Anatomie du sein », page 13). Avant celle-ci, les cellules ne sont pas totalement différenciées et donc beaucoup plus sensibles aux agents carcinogènes. Cette première grossesse menée à terme permet une différenciation cellulaire totale diminuant ainsi le risque de cancer du sein. La sensibilité mammaire aux carcinogènes est donc maximale entre la puberté et la première grossesse menée à terme : plus cette fenêtre est longue, plus le risque serait élevé.

Ce risque est diminué de 8 % (RR = 0,92, $IC_{95\%}$ = [0,88-0,96]) par grossesse menée à terme[56,65]. Cependant, la diminution du risque de cancer du sein associée à une grossesse menée à terme est précédée par une **augmentation transitoire** de ce risque. Elle est provoquée par une production extrêmement importante d'hormones sexuelles et dure environ 5 à 10 ans après la grossesse menée à terme[66,67]. Elle est ensuite suivie par une diminution durable du risque. La très récente analyse de Nichols *et al*. (2018)[68] réalisée sur les données de 15 études prospectives confirme ce résultat en montrant une augmentation du risque de cancer du sein de 1,80 ($IC_{95\%}$ = [1,63-1,99]) pendant les 5 premières années après la naissance de l'enfant. Au bout de 34 ans après la grossesse, un effet protecteur est mis en évidence avec un HR égal à 0,77 ($IC_{95\%}$ = [0,67-0,88]). Les auteurs montrent que la transition d'une association positive à une association négative à lieu environ 24 années après la naissance de l'enfant.

- *Allaitement*

L'**allaitement** est un facteur protecteur de cancer du sein. Une hypothèse expliquant cet effet protecteur est la diminution du nombre de cycles menstruels due au retardement de l'ovulation provoqué par l'allaitement. En plus des 7 % de diminution de risque de cancer du sein associé à une GMT, the collaborative Group of Reproductive Risk Factor a estimé une diminution de 4,3 % ($IC_{95\%}$ = [2,9-5,8]) pour chaque période d'allaitement de 12 mois[65]. Une méta-analyse effectuée sur 27 études montre une diminution du risque d'autant plus importante que la durée d'allaitement cumulée est importante[69].

- *Interruption de grossesse*

Les **interruptions de grossesses**, qu'elles soient volontaires (IVG) ou spontanées (fausse-couches), ne sont pas considérées comme des facteurs de risque de cancer du sein. Peu d'études montrent une association et les associations mises en évidence ne vont pas toutes dans le même sens. Ainsi, une méta-analyse, réalisée en 1996 et regroupant les résultats de 23 études, met en évidence une augmentation du risque lié à une interruption de grossesse (OR = 1,3, $IC_{95\%}$ = [1,2-1,4])[70]. À l'inverse, l'étude serbe de Ilic *et al*.[71], réalisée sur 191 cas et 191 témoins, montre qu'une interruption de grossesse entrainerait une diminution du risque de cancer du sein, tant pour les avortements spontanés (OR = 0,47, $IC_{95\%}$ = [0,25–0,90]) que pour les IVG (OR = 0,31, $IC_{95\%}$ = [0,10-0,98]).

Traitement hormonal substitutif

La prise de **traitements hormonaux substitutifs** (THS) après la ménopause augmenterait le risque de développer un cancer du sein. Cette augmentation du risque est estimée à 1,7 ($IC_{95\%}$ = [1,3-2,2]) par Li *et al.*[72] et varie selon le type histologique de la tumeur avec un risque relatif estimé à 2,7 ($IC_{95\%}$ = [1,7-4,3]) pour les cancers lobulaires invasifs et à 1,5 ($IC_{95\%}$ = [1,1-2,0]) pour les cancers canalaires invasifs. Plus la prise de THS est longue et plus le risque est important[59,72]. Cette augmentation de risque semble diminuer lorsque les femmes cessent d'utiliser des THS jusqu'à devenir nulle 5 ans après l'arrêt[73].

Différence par sous-types

Le rôle des facteurs gynéco-obstétriques sur le risque de cancer du sein est dû aux **expositions prolongées aux hormones sexuelles** (plus particulièrement aux œstrogènes), ce qui explique que ces facteurs soient **plus souvent associés aux sous-types exprimant les récepteurs hormonaux.**

En effet, les quelques études qui se sont intéressées à ces facteurs par sous-type ont montré qu'ils étaient plus prévalents dans les tumeurs RE^+ que dans celles RE^-. Par exemple, Yang *et al.*[58] ont montré dans leur analyse *case-only* que l'effet de la parité et de l'âge à la première grossesse menée à terme était également dépendant du sous-type moléculaire. Au début de l'année 2019, l'étude de Fotner *et al.*[68] sur les données des *Nurses' Health Studies* confirme cette hypothèse et montre que la parité est associée à une diminution du risque de cancer du sein RE^+ mais qu'elle n'a pas d'effet sur les cancers du sein RE^-. Les auteurs montrent également que l'effet de l'allaitement n'est pas le même selon le sous-type de cancer du sein. Le risque de cancer du sein de type luminal B semble diminuer à partir de 3 enfants, que les femmes aient allaité ou non. Par contre, pour les cancers du sein de type luminal A, cet effet protecteur de la parité (≥ 3 enfants) n'est retrouvé que pour les femmes ayant allaité.

d. Autres facteurs

Densité mammaire

La **densité mammaire** mesure le pourcentage de tissus épithéliaux et conjonctifs contenus dans le sein. Cette caractéristique est associée à une importante augmentation du risque de cancer du sein. D'après la classification BI-RADS[74], il existe 4 catégories selon la densité glandulaire :

- les seins graisseux homogènes contenant presque uniquement du tissu adipeux (< 25 % de tissu glandulaire) qui représentent 5 à 10 % des seins après 50 ans (D1) ;
- les seins graisseux hétérogènes avec quelques densités glandulaires dispersées (25 à 50 % de tissu glandulaire) qui représentent 50 % des femmes à 50 ans (D2) ;
- les seins denses hétérogènes avec une forte composante glandulaire (51 à 75 %) qui représentent 34 à 40 % des seins après 50 ans (D3) ;
- et enfin les seins denses homogènes (> 75 % de tissu glandulaire) qui représentent 5 à 10 % des seins après 50 ans (D4)[75].

De nombreuses études ont établi que le risque de cancer du sein augmente progressivement avec l'augmentation de la densité mammaire et que ce risque est 3 à 6 fois plus élevé chez les femmes aux seins denses homogènes que chez les femmes qui ont peu ou pas de densité mammaire (seins graisseux homogènes)[76,77]. Lorsqu'on compare les femmes aux seins denses à celles ayant des seins graisseux hétérogènes (qui représentent plus de 50 % de la population de plus de 50 ans), le risque relatif est de 1,2 à 1,5 pour les femmes aux seins denses (D3) et de 2,1 à 2,3 pour celles aux seins extrêmement denses (D4)[78,79].

Expositions aux radiations médicales

L'effet carcinogène et la sensibilité des tissus mammaires aux radiations ionisantes sont bien connus. D'après l'United Nations Scientific Committee on the Effects of Atomic Radiation, la dose moyenne de radiation naturelle reçue (rayons cosmiques, éléments radioactifs, etc.) est d'environ 2,4 millisieverts (mSv)[80] .Malheureusement, il n'est pas possible d'estimer l'effet de ces radiations naturelles sur le risque de cancer car la quantité exacte de radiations reçue par chaque personne est difficile à mesurer et il n'existe pas de groupes d'individus non exposés. Cependant, les études réalisées sur les populations japonaises d'Hiroshima ou de

Nagasaki, populations ayant subi une très forte exposition aux radiations lors des bombardements nucléaires, ont permis de mettre en évidence une association avec le risque de cancer du sein[81,82].

Cette association a été confirmée dans les études s'intéressant aux individus exposés à de fortes doses de radiations pour traiter des maladies bénignes, par exemple la tuberculose[83–85], la mastite postnatale aigüe[86] ou encore les hémangiomes[87]. Ces études ont permis d'estimer un risque relatif de cancer du sein entre 1,4 et 2,2 pour ces individus surexposés aux radiations. L'étude de Ronckers *et al*.[88] a également montré qu'il y avait un effet dose avec un excès de risque de cancer du sein proportionnel à la dose de radiation.

L'exposition à de faibles doses de radiations dans le cadre diagnostic ou du dépistage semble également augmenter le risque de cancer du sein, en particulier chez les femmes à haut risque. La méta-analyse de Jansen-van der Weide MC *et al*.[89] montre que le nombre d'expositions diagnostiques et l'âge à la première exposition sont associés à une augmentation du risque de cancer du sein, avec un OR estimé à 1,80 ($IC_{95\%}$ = [1,1-3,0]) pour un nombre d'expositions supérieur ou égal à 5 et à 2,0 ($IC_{95\%}$ = [1,3-3,1]) lorsque l'exposition a lieu avant l'âge de 20 ans. En 2012, Pijpe *et al*.[90] montrent grâce à leur étude de cohorte rétrospective GENE-RAD-RISK que l'effet augmente avec la dose de radiation reçue avant 30 ans, avec un OR égal à 1,63 ($IC_{95\%}$ = [0,96-2,77]) pour une dose inférieure à 0,0020 Gy et un OR de 3,84 ($IC_{95\%}$ = [1,67-8,79]) lorsque la dose dépasse 0,0174 Gy.

2. Les facteurs génétiques

a. Histoire familiale

Parmi les femmes atteintes de cancer du sein, 10 à 15 % font partie d'une famille à haut risque avec un(e) ou plusieurs apparenté(e)s ayant développé un cancer du sein. L'**histoire familiale** représente le facteur de risque le plus important, avec un risque qui augmente avec le nombre d'individus apparentés atteints. Dans une méta-analyse regroupant les résultats de 38 études réalisées en 1997, Pharoah *et al*.[91] montrent que le risque de cancer du sein chez une femme augmente de 2,1 ($IC_{95\%}$ = [2,0-2,2]) lorsqu'un cancer du sein est diagnostiqué chez au moins une apparentée au premier degré. L'étude menée par le Collaborative Group on

Hormonal Factors in Breast Cancer en 2001 sur 58 209 femmes atteintes et 101 986 non atteintes de cancer du sein a permis d'estimer de façon précise les risques relatifs en fonction du nombre d'apparentées atteintes. Elle montre qu'il y a une augmentation significative du risque de cancer du sein en fonction du nombre d'apparentées au premier degré (mère, sœur ou fille) atteintes[92]. Le risque de développer un cancer du sein pour une femme ayant une apparentée au premier degré atteinte est estimé à 1,80 ($IC_{99\%}$ = [1,69-1,91]) et il augmente à 2,93 ($IC_{99\%}$ = [2,36-3,64]) et 3,90 ($IC_{99\%}$ = [2,03-7,49]) avec respectivement 2 et 3 apparentées au premier degré atteintes. Cette étude montre également que l'effet de l'histoire familiale diffère avec l'âge. En effet, les femmes ayant une apparentée au premier degré atteinte d'un cancer du sein ont un risque, comparé à celui des femmes sans antécédents familiaux de cancer du sein, multiplié par 2,14 ($IC_{99\%}$ = [1,92-2,38) de développer un cancer du sein avant 50 ans et de 1,65 ($IC_{99\%}$ = [1,53-1,78]) après 50 ans. Cet effet de l'âge est similaire quel que soit le nombre d'apparentées atteintes. Le risque de ces femmes ayant des apparentées atteintes d'un cancer du sein varie également selon l'âge auquel les apparentées ont développé leur cancer, avec une augmentation du risque associé à un jeune âge au diagnostic de cancer du sein chez les apparentées. Selon Easton *et al.*[93], le risque relatif est estimé à 5,7 lorsque le cas et ses apparentées atteintes ont un âge inférieur à 40 ans alors qu'il n'est plus que de 1,4 si elles ont un âge supérieur à 60 ans.

Cette augmentation du risque relatif associée à des antécédents familiaux de cancer du sein révèle la présence de **prédispositions génétiques** associées au développement de la maladie. On parle alors d**e cancer du sein familial** et cette forme représente environ **5 à 10 % de la totalité des cas de cancer du sein**.

À partir des années 80 et du développement de l'informatique, les chercheurs ont voulu mettre en évidence la composante génétique qui pourrait expliquer ces agrégations familiales de cancer du sein[94–96]. Ils ont mis en évidence la transmission d'un gène dit alors « majeur ». Puis le développement des outils moléculaires a permis en 1990 à Hall *et al.*[97] de localiser un premier gène sur le chromosome 17 qu'ils nomment *BRCA1* (pour BReast CAncer 1) et à Wooster *et al.*[98] en 1994 d'identifier, grâce à une analyse de liaison génétique sur génome entier sur 15 familles, un deuxième gène sur le chromosome 13, qu'ils nomment *BRCA2* (pour BReast CAncer 2) et qui n'est pas lié au locus *BRCA1*. Ces deux gènes seront décrits plus en détail dans la suite de ce manuscrit (voir paragraphe « 3. Les gènes *BRCA1* et *BRCA2* », page 47). L'identification de ces deux gènes a été le point de départ d'une

recherche intensive de gènes de prédisposition au cancer du sein. Vingt-cinq ans après, les technologies de séquençage de l'ADN ont énormément évolué et permettent aujourd'hui d'utiliser des outils de séquençage à haut débit (ou NGS pour « Next Generation Sequencing ») très rapides et à faible coût. Ces outils peuvent être utilisés également dans le cadre des tests de diagnostic moléculaire afin de tester en parallèle de *BRCA1* et *BRCA2* d'autres gènes associés à un risque de cancer du sein plus modéré.

b. Panel de gènes sein-ovaire

Les consultations d'oncogénétique orientées vers la recherche de prédispositions au cancer du sein chez les individus à haut risque ont débuté avec la découverte des gènes *BRCA1* et *BRCA2*.

Depuis l'identification de *BRCA1* et *BRCA2*, d'autres gènes ont été associés au risque de développer un cancer du sein, avec des risques et des pénétrances plus faibles que ceux des gènes *BRCA1* et *BRCA2*. À l'instar de ces deux derniers gènes, certains de ceux identifiés au cours de ces 25 dernières années[99,100] sont depuis peu testés lors des consultations d'oncogénétique. En France, cela concerne le gène *PALB2* depuis 2015 et les gènes *CDH1*, *PTEN* et *TP53* depuis l'été 2018[37], d'après les recommandations du Groupe Génétique et Cancer (GGC). Ces gènes font partie d'un panel de gènes appelé le **panel « sein-ovaire »**, ciblant également les gènes de prédisposition au cancer de l'ovaire (*RAD51C*, *RAD51D*, *MLH1*, *MDH2*, *APCAM*, *MSH6* et *PMS2*).

c. Gènes non actionnables

Il n'y a pas de consensus entre les différents pays et le nombre de gènes testés diffère d'un pays à l'autre. Le consortium ENIGMA a réalisé une étude en 2018 dont le but était de décrire et comparer les panels de gènes utilisés dans chaque pays[101]. Au total, les pratiques de 61 centres répartis dans 20 pays ont été analysées. Le consortium s'est intéressé à 16 gènes – *ATM*, *BARD1*, *BRIP1*, *CHEK2*, *MRE11A*, *NBN*, *NF1*, *PALB2*, *PTEN*, *RAD50*, *RAD51C*, *RAD51D*, *STK11*, *TP53* et *MEN1* – tous associés au risque de développer un cancer du sein[102–112]. La présence d'une mutation dans l'ensemble des gènes est recherchée par plus de 50 % des centres mais seulement 6 gènes (*PALB2*, *TP53*, *PTEN*, *CHEK2*, *ATM* et *BRIP1*) sont testés régulièrement (c'est-à-dire pour plus de 50 % des patients).

d. Polymorphismes génétiques et études GWAS

Un polymorphisme génétique, communément appelé **SNP** (pour *single nucleotide polymorphism*) est une variation de l'ADN retrouvée fréquemment dans la population, c'est-à-dire à une fréquence supérieure à 1 % (Annexe 1, page 199). Un certain nombre de SNPs ont été associés au risque de développer un cancer du sein mais avec des risques relatifs estimés beaucoup plus faibles que ceux associés aux gènes de prédisposition cités précedemment, de l'ordre de 1,2.

Les premiers SNPs associés au cancer du sein ont été mis en évidence par des études « gènes candidats » de type cas-témoins analysant les variations génétiques présentes dans des gènes d'intérêt pouvant influencer le risque de cancer du fait de leur fonction biologique, comme les gènes intervenant dans la réparation de l'ADN ou dans le contrôle du cycle cellulaire[113–122]. Cependant, les SNPs mis en évidence dans ces études n'étaient, la plupart du temps, pas répliquées dans d'autres études[123–127] et ces études n'avaient, bien souvent, pas la puissance nécessaire pour détecter des risques relatifs inférieurs à 1,5 du fait de leur faible taille (plusieurs dizaines à quelques milliers d'individus)[128–130].

Différents groupes se sont regroupés en 2005 sous le nom du « **Breast Cancer Association Consortium** » (BCAC)[131] pour permettre la réalisation d'études d'association plus puissantes. Cette collaboration a permis d'établir une base de données de plus de 30 000 cas de cancers du sein et 30 000 témoins provenant d'une vingtaine d'études différentes. Un de leurs premiers travaux avait pour but de répliquer 16 SNPs trouvés associés dans les études participantes. Les analyses d'association réalisées à partir de populations de 12 000 à 31 000 femmes, selon le SNP analysé, ont permis de répliquer 5 de ces SNPs dont un SNP dans le gène *CASP8*[131].

En 2007, grâce au développement et à l'amélioration des puces à ADN permettant de génotyper un nombre toujours plus important de SNPs simultanément et grâce à la diminution du coût du génotypage, Easton *et al.*[132] et Hunter *et al.*[133] mettent en place deux études pangénomiques (« *Genome-wide Association Studies* » *GWAS*) sur le cancer du sein. Ces études réalisées sur le génome entier ont permis d'identifier de façon agnostique de nombreux SNPs associés au risque de cancer du sein dans la population générale. Ceux-ci sont localisés dans des régions codantes ou introniques (*FGFR2*[132,133], *LOC643714*[132], *MAP3K1*[132],

LSP1[132], *CASC16*[134], etc.) mais également dans des régions intergéniques (locus 8q[132], 2q35[134], 5p12[135], etc.).

En 2013, le consortium BCAC et ses équivalents pour les cancers de l'ovaire (OCAC) et de la prostate (PRACTICAL), se sont regroupés pour former le consortium « Collaborative Oncological Gene-environment Study » (COGS) dans le but d'améliorer la compréhension de la prédisposition génétique des cancers hormono-dépendants[136]. Ils développent alors une nouvelle puce à ADN, la puce iCOGS contenant environ 200 000 SNPs. Les analyses effectuées par le consortium BCAC ont permis d'identifier plus de 40 nouveaux SNPs associés au risque de cancer du sein dans la population générale[137], localisés dans des gènes comme *PEX14*, *CDCA7*, *FOXQ1* ou encore *RAD51L1*. Quatre autres SNPs spécifiquement associés aux cancers du sein RE$^-$ ont également été mis en évidence[138]. Les SNPs non génotypés de la puce iCOGS ont ensuite été imputés et 15 nouvelles associations ont été trouvées[139].

En 2016, un nouveau consortium, le consortium OncoArray, regroupant les consortia ayant participé au projet COGS mais également ceux sur les cancers du poumon et du colon, est créé[140]. Dans ce projet, dont le but est toujours d'améliorer la compréhension de la prédisposition génétique de ces différents cancers, une nouvelle puce portant le même nom a été élaborée. La puce OncoArray contient plus de 530 000 SNPs. Grâce à cette dernière, plus de 65 nouveaux SNPs associés au risque de cancer du sein dans la population générale[141] et 10 SNPs spécifiques aux tumeurs RE$^-$[142] ont été découverts.

À ce jour, **158 SNPs ont été trouvés associés en population générale** et **22** chez les personnes atteintes d'un **cancer du sein RE$^-$.** Tous ces SNPs sont **associés à de faibles effets sur le cancer du sein,** avec des OR variant de **0,7 à 1,5.**

e. Score de risque polygénique (PRS)

Les risques associés à chaque SNP sont très faibles et n'ont aucune utilité individuellement pour prédire le risque de cancer du sein. Cependant, l'effet combiné de plusieurs SNPs indépendants pourrait permettre de discriminer des femmes à risque plus ou moins élevé dans

la population. Cet effet combiné est calculé par un **score de risque polygénique** (PRS) qui fait l'hypothèse d'un effet additif des SNPs composant le polygène.

Pour chaque individu i, un PRS_i est de la forme suivante :

$$PRS_i = \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k \ldots + \beta_n x_n$$

avec β_k le log(OR) par allèle à risque pour le SNP k et x_k le nombre d'allèles à risque (0, 1, 2).

En 2015, Mavaddat *et al.*[143] ont testé le rôle prédictif d'un PRS composé de 77 SNPs qui ont été trouvés associés au risque de cancer du sein dans la population générale par le consortium COGS ou par des études antérieures. Ils montrent que ce PRS, qui suit une distribution normale, est associé à un OR de 1,65 ($IC_{95\%}$ = [1,58–1,72]) par unité d'écart type sur la distribution. Il permet de bien discriminer les femmes situées aux extrémités de cette distribution, c'est-à-dire les femmes ayant le moins d'allèles à risque pour les SNPs d'intérêt et celles en ayant le plus. L'OR des femmes avec le nombre d'allèles à risque le plus faible et faisant partie du premier percentile de la distribution est estimé à 0,32 ($IC_{95\%}$ = [0,25–0,40]), alors que celui des femmes avec le plus d'allèles à risque et se trouvant dans le dernier percentile est estimé à 3,36 ($IC_{95\%}$ = [2,95–3,83]). Le risque absolu de développer un cancer du sein à l'âge de 80 ans est de 3,5 % ($IC_{95\%}$ = [2,6 %–4,4 %]) pour les femmes du premier percentile et de 29 % ($IC_{95\%}$ = [24,9 %–33,5 %]) pour celles du dernier percentile.

f. Facteurs génétiques restant à identifier

Un grand nombre de facteurs génétiques associés au risque familial de cancer du sein ont été mis en évidence. Environ 15 à 20 % de ce risque familial sont aujourd'hui expliqués par la présence d'une mutation dans les gènes *BRCA1* ou *BRCA2* et 10 % par une mutation d'autres gènes associés à des pénétrances plus faibles. De plus, Michailidou *et al.*[141] montrent que les SNPs identifiés jusqu'alors expliquent autour de 18 % du risque familial. Environ **50 % du risque familial de cancer du sein** reste donc à expliquer.

Une des stratégies possibles pour trouver des SNPs encore inconnus est d'augmenter la couverture des puces à ADN en augmentant le nombre de SNPs génotypés et également en augmentant la taille de la population. Cela est en cours de réalisation par le biais du projet Confluence[d] (puce à ADN *GSA* (*Global Screening Array*)).

Une autre stratégie, que j'ai proposée dans le cadre de ma thèse, consiste à faire l'hypothèse que **l'effet de certains SNPs impliqués dans le développement d'un cancer du sein diffère selon les expositions environnementales** des individus. Cette stratégie consiste à analyser les SNPs en fonction des profils d'exposition aux facteurs environnementaux. Je me suis alors essentiellement intéressée aux expositions aux hormones sexuelles et aux expositions médicales aux radiations ionisantes. Ce travail a fait l'objet de la **première partie de ma thèse** intitulée « **Facteurs de risque génétiques spécifiques à un schéma environnemental particulier chez les femmes non porteuses d'une mutation dans les gènes *BRCA1* ou *BRCA2*** ».

3. Les gènes *BRCA1* et *BRCA2*

Les gènes ***BRCA1*** et ***BRCA2*** sont les deux premiers gènes trouvés associés au risque familial de développer un cancer du sein. Ce sont des gènes suppresseurs de tumeurs. Ils s'étendent sur environ 80 kilobases (kb). Les protéines BRCA1 et BRCA2 composées respectivement de 1 863 et 3 418 acides aminés, régulent de façon négative la prolifération cellulaire. Elles interviennent dans la même voie biologique assurant le maintien de l'intégrité du génome mais chacune d'entre elles intervient à des stades différents[144]. Une variation de séquence ou « variant » dans l'un des deux gènes codant pour ces protéines peut entrainer la production d'une protéine moins efficace, voire inefficace, et cela participe au développement incontrôlé de la tumeur et par conséquent à l'apparition d'un cancer. Plus de 20 000 mutations différentes dans les gènes *BRCA1* et *BRCA2* sont recensées dans la base de données globale de variants *BRCA Exchange* mise en place via le projet *BRCA Challenge*[145]. Ce projet a été initié par *the Global Alliance for Genomics and Health* (GA4GH) dans le but d'agréger les données relatives aux gènes *BRCA1* et *BRCA2* et de soutenir des projets collaboratifs à l'échelle internationale. Parmi ces variants, plus de 6 000 ont été validés par des experts provenant principalement du consortium ENIGMA[146], et plus de 3 500 sont considérés

[d] Les détails du projet Confluence sont disponibles sur https://dceg.cancer.gov/research/cancer-types/breast-cancer/confluence-project, consulté le 23 octobre 2019.

comme pathogènes. Ces variants pathogènes ou « mutations » augmentent le risque de cancer du sein en affectant la production et/ou l'efficacité de la protéine.

a. Caractéristiques des tumeurs

Il existe une hétérogénéité moléculaire entre les tumeurs des femmes de la population générale et celles porteuses d'une mutation de *BRCA1* ou *BRCA2*.

Les tumeurs BRCA1 sont plus souvent des tumeurs médullaires associées à un haut grade et une activité mitotique plus élevée que celles des femmes non mutées dans ce gène[147,148]. Environ 90 % de ces tumeurs n'expriment pas les récepteurs aux œstrogènes[149] et plus de 50 % d'entre elles n'expriment pas non plus les récepteurs à la progestérone et à la protéine HER2 (tumeurs triple négatives).

Au contraire, les tumeurs BRCA2 sont beaucoup plus hétérogènes. Elles ressemblent plus à celles de la population générale avec plus de 70 % de tumeurs RE^+ contre environ 20 % pour les tumeurs BRCA1[149]. Cependant, les tumeurs BRCA2 sont plus souvent associées à des grades élevés que les tumeurs des femmes de la population générale.

b. Estimation du risque de cancer du sein

D'après les dernières estimations de Kuchenbaecker *et al.*[150], le risque cumulé à 70 ans de développer un cancer du sein est de 72 % ($IC_{95\%}$ = [65 %-79 %]) pour les porteurs d'une mutation *BRCA1* et de 69 % ($IC_{95\%}$ = [61 %-77 %]) pour les porteurs d'une mutation *BRCA2*. Lorsqu'une mutation dans ces gènes est détectée au cours d'un test génétique, des recommandations et un suivi adapté sont proposés aux femmes. La surveillance mammaire consiste en une IRM et une mammographie annuelle de 30 à 65 ans et seulement une mammographie annuelle au-delà de 65 ans. La mastectomie et l'annexectomie prophylactiques sont proposées et discutées selon l'âge de la femme porteuse d'une mutation, le type de mutation et son histoire familiale[37].

Cependant, parmi les personnes porteuses d'une mutation dans les gènes *BRCA1* ou *BRCA2*, il existe également une forte hétérogénéité quant à leur risque de développer un cancer du sein. Comme pour la population générale, le risque de cancer du sein chez les porteurs d'une mutation *BRCA1*/2 augmente avec l'âge. Kuchenbaecker *et al.*[150] montrent que le risque

cumulé pour les femmes porteuses d'une mutation *BRCA1* est de 4 % ($IC_{95\%}$ = [2 %–7 %]) entre 21 et 30 ans alors qu'il augmente à 72 % ($IC_{95\%}$ = [65 %–79 %]) à plus de 70 ans. Pour celles porteuses d'une mutation dans le gène *BRCA2*, ce risque est le même que les femmes porteuses d'une mutation de *BRCA1* entre 21 et 30 ans et augmente jusqu'à 69 % pour les femmes âgées de plus de 70 ans. Cependant, le pic d'incidence de cancer du sein est entre 41 et 50 ans pour les femmes porteuses d'une mutation *BRCA1* et décalé de 10 ans, entre 51 et 60 ans, pour celles porteuses d'une mutation *BRCA2*[150]. Cette étude montre également une variation du risque selon le nombre d'apparentés au premier degré atteints d'un cancer du sein. Pour une femme de 40 ans porteuse d'une mutation de *BRCA1*, le risque cumulé est estimé à 16 % lorsqu'il n'y a aucun cancer du sein chez ses apparentées, alors qu'il est de 31 % à partir de 2 apparentées atteintes. Pour les femmes porteuses d'une mutation de *BRCA2*, ce risque cumulé passe de 5 % à 14 % respectivement. Ces variations appuient l'hypothèse que des facteurs génétiques et environnementaux au sens large modifient le risque de cancer du sein chez ces personnes.

La **recherche de ces facteurs modificateurs est nécessaire pour une prédiction du risque plus précise** qui permettrait une **amélioration des recommandations et du suivi** des femmes porteuses d'une mutation dans les gènes *BRCA1* ou *BRCA2*. De nombreux travaux ont été publiés sur les facteurs environnementaux modificateurs du risque de cancer du sein dans cette population[151–153]. Certains facteurs semblent avoir un effet similaire à celui dans la population générale, comme la prise de contraceptifs oraux[154], d'autres comme la parité semble avoir un effet différent[155]. Encore plus nombreux sont ceux publiés sur les facteurs génétiques modificateurs.

c. Facteurs génétiques modificateurs

Le *Consortium of Investigators of Modifiers of BRCA1/2* (CIMBA)[156] a été mis en place en 2005 pour trouver des facteurs génétiques modificateurs du risque de cancer du sein chez les individus porteurs d'une mutation *BRCA1/2* et évaluer leurs effets. Des études *GWAS* effectuées spécifiquement sur les femmes porteuses d'une mutation *BRCA1/2* ont été conduites et ont permis de mettre en évidence des SNPs modificateurs dans 3 régions chromosomique ou *loci* : en 1q32 (rs2363956 et rs8170)[157] et 19p13 (rs2290854)[158] pour les

femmes porteuses d'une mutation de *BRCA1* et en 6p24 (rs9348512)[159] pour les femmes porteuses d'une mutation de *BRCA2*.

Toutes les autres régions identifiées dans cette population l'ont été en analysant les régions qui avaient déjà été trouvées associées au cancer du sein dans la population générale. Les SNPs impliqués dans la prédisposition au cancer du sein, tous types moléculaires confondus, et aux cancers du sein RE^{+} dans la population générale sont plus souvent trouvés associés chez les femmes porteuses d'une mutation *BRCA2* dont la majorité des tumeurs sont RE^{+}[149]. Au contraire, les SNPs associés chez les femmes porteuses d'une mutation de *BRCA1* sont plus souvent ceux trouvés associés aux cancers du sein RE^{-} dans la population générale.

Trente-trois SNPs ont été associés au cancer du sein (25 tous sous-types confondus[160–167], 2 spécifiques aux tumeurs RE^{-}[164] et 6 spécifiques aux tumeurs RE^{+}[164]) chez les femmes porteuses d'une mutation *BRCA1* et 30 SNPs (23 tous sous-types confondus[160,161,168,162,169,164], 4 spécifiques aux tumeurs RE^{-}[164] et 3 spécifiques aux tumeurs RE^{+}[164]) chez les femmes porteuses d'une mutation de *BRCA2* (Tableau 2 et Tableau 3).

Tableau 2 - SNPs associés au risque de cancer du sein chez les femmes porteuses d'une mutation de BRCA1, pour les cancers du sein tous sous-types confondus, spécifiques aux tumeurs RE⁻ et spécifiques aux tumeurs RE⁺.

Locus	SNP	Référence
Tout type de tumeurs		
1q32	rs2363956	Antoniou 2010
1q32.1b	rs4245739	Kuchenbaecker 2014
2q35	rs1338704	Antoniou 2009
5p15.33	rs10069690	Bojesen 2013, Kuchenbaecker 2014
5p15.33	rs2242652	Bojesen 2013, Couch 2016
5p15.33	rs2736108	Bojesen 2013, Kuchenbaecker 2014
5p15.33	rs7725218	Kuchenbaecker 2014
6q25.1	rs2046210	Antoniou 2011, Kuchenbaecker 2014
6q25.1	rs3757318	Kuchenbaecker 2014
6q25.1	rs9397435	Antoniou 2011
11p15.5	rs3817198	Kuchenbaecker 2014
11q22.3	rs228595	Hamdi 2017
12p11.22	rs10771399	Antoniou 2012, Kuchenbaecker 2014
12q22	rs17356907	Kuchenbaecker 2014
16q12.1a	rs3803662	Antoniou 2008, Kuchenbaecker 2014
16q12.1b	rs17817449	Kuchenbaecker 2014
19p13	rs2290854	Couch 2013
19p13	rs61494113	Lawrenson 2016
19p13	rs67397200	Lawrenson 2016
19p13.11	rs56069439	Couch 2016
19p13.11	rs8170	Antoniou 2010, Kuchenbaecker 2014, Couch 2016
21q21.1	rs2823093	Kuchenbaecker 2014
22q12.2	rs132390	Kuchenbaecker 2014
22q13.1	rs6001930	Kuchenbaecker 2014

Locus	SNP	Référence
Tumeurs RE⁻		
6p25.3	rs11242675	Kuchenbaecker 2014
10q26.12	rs2981579	Kuchenbaecker 2014
Tumeurs RE⁺		
2q35	rs13387042	Kuchenbaecker 2014
3p26.1	rs6762644	Kuchenbaecker 2014
9p21.3	rs1011970	Kuchenbaecker 2014
10q22.3	rs704010	Kuchenbaecker 2014
10q26.12	rs2981579	Kuchenbaecker 2014
12q24.21	rs1292011	Kuchenbaecker 2014

Tableau 3 - SNPs associés au risque de cancer du sein chez les femmes porteuses d'une mutation de BRCA2, pour les cancers du sein tous sous-types confondus, spécifiques aux tumeurs RE⁻ et spécifiques aux tumeurs RE⁺.

Locus	SNP	Référence
Tout type de tumeurs		
1p11.2	rs11249433	Antoniou 2011
2q35	rs1338704	Antoniou 2009
3p24.1	rs4973768	Antoniou 2010 Kuchenbaecker 2014
5p12	rs10941679	Antoniou 2010, Kuchenbaecker 2014
5p15.33	rs10069690	Kuchenbaecker 2014
5p15.33	rs2736108	Kuchenbaecker 2014
5q11.2	rs889312	Antoniou 2008
6p23	rs204247	Kuchenbaecker 2014
6p24.3	rs9348512	Gaudet 2013, Kuchenbaecker 2014
6q14	rs17530068	Kuchenbaecker 2014
6q25.1	rs3757318	Kuchenbaecker 2014
6q25.1	rs9397435	Antoniou 2011
9p21.3	rs1011970	Antoniou 2012
9q31.2	rs865686	Antoniou 2012
10q21	rs10995190	Antoniou 2012
10q26.13	rs2981582	Antoniou 2008
11p15.5	rs3817198	Antoniou 2009, Kuchenbaecker 2014
11q13.3	rs554219	Kuchenbaecker 2014
11q24.3	rs11820646	Kuchenbaecker 2014
12p11.22	rs10771399	Kuchenbaecker 2014
12q24	rs12922011	Antoniou 2012
16q12.1	rs3803662	Antoniou 2008, Kuchenbaecker 2014
16q12.1b	rs17817449	Kuchenbaecker 2014

Locus	SNP	Référence
Tumeurs RE⁻		
2p24.1	rs12710696	Kuchenbaecker 2014
6q25.1	rs2046210	Kuchenbaecker 2014
8q24.21	rs11780156	Kuchenbaecker 2014
10q25.2	rs7904519	Kuchenbaecker 2014
Tumeurs RE⁺		
19q13.31	rs3760982	Kuchenbaecker 2014
21q21.1	rs2823093	Kuchenbaecker 2014
22q12.2	rs132390	Kuchenbaecker 2014

Par ailleurs, l'évaluation du PRS construit avec les SNPs trouvés en population générale, nommé ici « PRS_{BCAC} », sur la population de femmes mutées dans les gènes *BRCA1/2* a montré une efficacité moindre (Kuchenbaekcer *et al*., 2017)[170]. Le risque associé au PRS_{BCAC} pour les femmes porteuses d'une mutation dans le gène *BRCA1* est estimé à 1,14 ($IC_{95\%}$ = [1,11–1,17]) par unité d'écart type et à 1,22 ($IC_{95\%}$ = [1,17–1,28])[170] pour les femmes porteuses d'une mutation dans le gène *BRCA2* alors qu'il a été estimé à 1,65 dans la population générale[143]. Par conséquent, alors que les risques associés aux percentiles extrêmes de la distribution du PRS sont de 0,32 et 3,36 dans la population générale, ils ne sont que de 0,76 et 1,82 pour les femmes porteuses d'une mutation de *BRCA1* et de 0,80 et 1,51 pour celles porteuses d'une mutation de *BRCA2*.

Ces résultats suggèrent que la **variabilité dans l'estimation du risque de cancer du sein dans la population *BRCA1/2*** pourrait être **expliquée par des facteurs génétiques spécifiques aux femmes mutées** qui n'ont pas encore été identifiés.

Cependant, même s'il y aura toujours de nouveaux sujets porteurs d'une mutation de *BRCA1* ou *BRCA2*, il sera **difficile d'augmenter suffisamment la taille des études pour mettre évidence de nouveaux SNPs**. Il est donc **indispensable de développer de nouvelles stratégies pour trouver de nouveaux facteurs de risque génétiques modificateurs**. Ceci est l'objet de **la deuxième partie de ma thèse** intitulée « **Facteurs génétiques modificateurs du risque de cancer du sein chez les femmes porteuses d'une mutation dans les gènes *BRCA1* ou *BRCA2*** ».

Première Partie

Facteurs de risque génétiques spécifiques à un schéma environnemental particulier chez les femmes à haut risque de cancer du sein non porteuses d'une mutation dans les gènes *BRCA1* ou *BRCA2*

Introduction

Le facteur de risque majeur de cancer du sein est d'avoir des antécédents familiaux de cette maladie. Environ 20 % du risque familial de cancer du sein est attribué à une mutation fortement pénétrante dans les gènes *BRCA1*, *BRCA2* ou *PALB2*. Comme nous l'avons vu précédemment, il a été montré que l'incidence du cancer du sein augmente également chez les patientes porteuses d'une mutation dans d'autres gènes responsables de syndromes familiaux comme *TP53*, *PTEN*, *STK11* et *CDH1* ou dans des gènes à pénétrance plus modérée (c'est-à-dire associés à un risque relatif de l'ordre de 2 à 3) comme *ATM* ou *CHEK2*. Ces gènes expliqueraient environ 10 % supplémentaires du risque familial.

Les études réalisées pour identifier d'autres facteurs génétiques responsables des 70 % du risque familial restant ont été menées soit dans des familles à haut risque en utilisant des analyses de liaison génétique ou de nouvelles technologies de séquençage à haut débit, soit dans la population générale en utilisant des études d'association de type cas-témoins.
Tandis que les analyses de liaison sont restées infructueuses[171], la recherche de variants rares, grâce aux technologies de séquençage à haut débit de l'exome, a permis d'identifier des variants probablement délétères (mutations) dans les gènes *XRCC2*[172] et *RINT1*[173]. Cependant, il semble peu probable que ces nouveaux gènes confèrent des risques de cancer aussi élevés que les gènes *BRCA1*, *BRCA2* ou *PALB2*, et la fréquence de leur mutation pourrait être extrêmement faible.
Les études GWAS cas-témoins, quant à elles, ont permis d'identifier environ 170 nouveaux SNPs[141,142], associés à des fréquences supérieures à 1 %. L'effet des SNPs fréquents dans la population générale est faible – au plus associé à des risques relatifs de l'ordre de 1,2 – et ils expliqueraient seulement 18 % du risque familial[141,174]. Ces SNPs ont permis cependant la construction de polygènes avec la création d'un score de risque appelé PRS pour la population générale[141–143]. Ces PRS améliorent sensiblement le pouvoir prédictif des modèles développés[143,175,176] mais un peu plus de 50 % des cancers du sein familiaux restent inexpliqués. Il reste donc très probablement de nombreux facteurs génétiques impliqués dans ce risque familial à découvrir.

Une des stratégies pour trouver de nouveaux facteurs génétiques vise à mettre en place des analyses GWAS suffisamment puissantes pour mettre en évidence des SNPs associés à des

risques encore plus faibles que ceux précédemment trouvés afin d'améliorer les PRS existants pour les sujets de la population générale mais également pour les porteurs d'une mutation dans les gènes *BRCA1/2*. Cependant, les données disponibles au niveau mondial ont déjà été analysées et leur possibilité d'extension est limitée. Il faut donc penser à de nouvelles stratégies d'analyse pour espérer identifier ces nouveaux facteurs génétiques. Dans le cadre de ma thèse, j'ai donc proposé un **nouveau design d'étude du génome-entier** qui est présenté en **deuxième partie de ce mémoire**.

Une deuxième stratégie consiste à considérer qu'une partie de ce risque familial inexpliqué est la résultante d'effets joints entre des facteurs génétiques et des facteurs environnementaux au sens large. L'étude des interactions gènes-environnement (GxE) est un domaine de recherche actif depuis plusieurs années. Depuis les années 2000, la majorité des études GxE ont été faites sur des gènes candidats (82 % sur des SNPs candidats, 12 % sur des GWAS et 6 % sur les deux) selon Simonds *et al.*[177], ce qui signifie que les facteurs environnementaux ont été pris en compte *a posteriori*, c'est-à-dire après l'identification des SNPs. Cependant, de nombreuses études, comme celle de Milne *et al*. ou celle de Warren *et al*., n'ont montré aucune interaction significative avec cette stratégie[178–182]. D'autres études, comme celle de Nickels *et al*., ont trouvé des interactions GxE significatives avec la consommation d'alcool, l'IMC ou la parité[183–185]. La majorité de ces études[178–180,183] ont porté sur des SNPs déjà trouvés associés au cancer du sein et ne permettent donc pas de trouver de nouveaux SNPs dont l'association ne serait significative qu'après stratification de la population.
Au contraire, en 2014 Schoeps et *al.*[185] ont étudié l'interaction de 71 527 SNPs localisés dans des régions non déjà associées avec le cancer du sein et 10 facteurs environnementaux. Ils montrent que le SNP rs10483028 a un effet différentiel selon l'IMC. De la même manière, en 2018 Dierssen-Sotos *et al.*[186] ont étudié l'interaction entre les SNPs présents dans les gènes intervenant dans les voies biologiques impliquées dans la reproduction et les facteurs de risque reproductifs. Cette stratégie leur a permis de mettre en évidence un nouveau SNP, rs2220712, qui diminue l'effet protecteur de la parité sur le cancer du sein. À ma connaissance, aucune autre étude n'a recherché des interactions GxE avec des approches de type GWIS (*Genome Wide Interaction Study*)[177]. De plus, ces études s'intéressent aux interactions GxE en prenant en compte les facteurs environnementaux un à un.

Ceci m'a amené à proposer dans **cette première partie** de mon travail de thèse **une nouvelle approche d'analyse prenant en compte non pas chaque facteur un à un mais un**

« profil » d'expositions aux facteurs environnementaux. Le but est alors de mettre en évidence de nouveaux SNPs, c'est-à-dire non trouvés associés au cancer du sein dans la population générale[141,142] et qui seraient associés au risque de cancer du sein dans un contexte environnemental spécifique.

Données

I. La population d'étude : GENESIS

Ces travaux ont été effectués sur les données de l'étude française nationale GENESIS[187] dont l'objectif est d'identifier et de caractériser de nouveaux gènes de prédisposition au cancer du sein à partir de familles à « haut risque » de cancer du sein n'ayant pas de mutation identifiée dans les gènes *BRCA1* et *BRCA2*. Cette étude implique tous les centres de génétique clinique du Groupe Génétique et Cancer (GGC) d'Unicancer. L'investigation et la coordination de cette étude ont été réalisées par la Plateforme d'Investigation en Génétique et en Épidémiologie (PIGE) de l'équipe « Épidémiologie Génétique des Cancers », Inserm U900, Institut Curie, dans laquelle je travaille. L'étude GENESIS a été initiée et mise en place en collaboration avec D. Stoppa-Lyonnet (Institut Curie, Paris) et O. Sinilnikova (Centre Léon Bérard, Lyon). L'inclusion des participantes a débuté en février 2007 et a duré 6 ans, jusqu'en décembre 2013.

1. Critères d'inclusion

Le but de l'étude GENESIS étant de recruter des femmes à haut risque familial de cancer du sein, celles-ci ont été incluses par le biais de consultations d'oncogénétique. L'inclusion des participantes a été réalisée de manière rétrospective, pour les femmes dont la recherche de mutation *BRCA1/2* a eu lieu entre 2003 et 2007, et prospective jusqu'en décembre 2013. Toutes les femmes avec un carcinome infiltrant, lobulaire ou canalaire, ou un carcinome *in situ* canalaire, sans mutation détectée dans les gènes *BRCA1* ou *BRCA2* et ayant au moins une sœur également atteinte d'un cancer du sein étaient éligibles. Toutes les sœurs atteintes du cas index ont été invitées à participer à l'étude. Deux types de témoins ont été incluses : des femmes non apparentées au cas index et les sœurs non atteintes. Il a été demandé aux cas index d'inviter une amie et/ou une collègue sans antécédent de cancer du sein ou d'un autre organe, ayant le même âge que le cas index à l'interview, plus ou moins 3 ans. Les sœurs non atteintes ainsi que les parents et les frères ont également été inclus.

2. Données collectées

Tous les cas de cancer du sein (cas index et sœurs atteintes) ainsi que les témoins ont complété un questionnaire épidémiologique d'une trentaine de pages comprenant des questions sur leur environnement, leur style de vie et leur antécédents familiaux de cancer (Annexe 2, page 200). Les questions portaient sur des données démographiques comme l'âge ou le lieu de naissance, la consommation d'alcool et de tabac, des informations relatives à la vie reproductive (grossesses, contraception, ménopause, etc.), les expositions professionnelles et/ou médicales aux radiations ionisantes, les antécédents médicaux personnels et familiaux (apparentés aux premier et second degrés), etc. Les mammographies réalisées au moment du diagnostic ou 2 à 3 ans avant celui-ci pour les cas et les plus récentes pour les témoins ont été numérisées. Un échantillon sanguin a également été collecté pour tous les cas et témoins de l'étude afin que chaque femme soit génotypée avec la puce à ADN iCOGS.

3. Description de la population d'étude

L'étude GENESIS comprend au total 1 713 cas index, 824 sœurs atteintes de cancer du sein, 598 sœurs et 1 410 témoins indemnes. 98 % d'entre elles ont complété le questionnaire épidémiologique, 97 % ont fourni un échantillon sanguin et 68 % des femmes ayant passé une mammographie l'ont fournie. L'âge moyen des cas index est de 59,0 ans (sd = 9,4) à l'interview et de 49,7 ans (sd = 9,4) au diagnostic et l'année de naissance médiane est 1950. L'intervalle moyen entre le diagnostic et l'interview est donc de 9,1 ans (sd = 7,4). L'âge moyen des témoins à l'interview est de 55,7 ans (sd = 9,3). 93,6 % des cas index et sœurs atteintes recrutées sont des cas prévalents. Ces femmes appartiennent à des familles à haut risque de cancer du sein avec un âge moyen au diagnostic plus jeune que celui des femmes de la population générale (49,7 vs. 63 ans[e]).

Parmi ces femmes, nous avons inclus dans ce travail seulement celles d'origine caucasienne afin d'éviter une stratification de la population. En effet, il existe une très grande diversité génétique entre les populations d'origine européenne, asiatique ou africaine et les facteurs associés aux maladies multifactorielles comme le cancer du sein peuvent être très différents

[e] D'après les chiffres donnés par l'Institut Curie : https://curie.fr/dossier-pedagogique/le-cancer-du-sein, consulté le 23 octobre 2019.

entre ces populations. De plus, plus de 95 % des femmes de GENESIS sont d'origine caucasienne, avec seulement 39 femmes avec des origines ethniques mixtes, 43 femmes d'origine non caucasienne et 32 dont l'origine est inconnue.
L'étude cas-témoin sur le rôle des facteurs environnementaux a ainsi été réalisée sur 1 591 cas index et 1 381 témoins.

II. Les données génotypiques

Les données génotypiques des femmes de l'étude GENESIS ont été obtenues par génotypage avec la puce à ADN iCOGS. Parmi les cas index et les témoins d'origine européenne incluses dans l'étude GENESIS, 438 n'ont pas été génotypées (329 cas index et 109 témoins) faute de budget. Les analyses génétiques ont donc été faites sur 1 262 femmes atteintes d'un cancer du sein (cas index) et 1 272 témoins. Parmi les femmes atteintes, le délai entre le diagnostic et l'interview est inférieur ou égal à 5 ans pour 609 (48,3 %) d'entre elles, entre 5 et 10 ans pour 411 (32,6 %) et supérieur à 10 ans pour les 242 (19,2 %) femmes restantes.

1. La puce iCOGS

La puce à ADN iCOGS a été développée par le consortium « Collaborative Oncological Gene-environment Study » COGS[136] qui a été mis en place en 2013 pour améliorer la compréhension de la prédisposition génétique de trois cancers hormono-dépendants (cancer du sein, de l'ovaire et de la prostate). Les principaux objectifs de ce projet étaient d'identifier des variants génétiques associés à l'un de ces 3 cancers, de déterminer le risque associé à chacun de ces variants et d'évaluer d'éventuelles interactions entre les facteurs de risque génétiques et « environnementaux » et/ou de style de vie.

a. Composition de la puce

La puce Illumina est composée de 220 123 SNPs. Une partie de ces SNPs ont été identifiés comme étant associés aux cancers dans des études *GWAS,* mais d'autres variants sont plus « exploratoires »[136]. En effet, la puce comporte aussi des SNPs présents dans des régions génétiques impliquées dans des cancers autres que les cancers du sein, de l'ovaire ou de la

prostate, ou présents dans des gènes responsables de mécanismes cellulaires importants, ainsi que les SNPs d'intérêt de chaque consortium (par exemple des variants rares qui ne sont pas détectés par *GWAS,* ou des SNPs associés à des caractéristiques phénotypiques, comme l'âge aux premières règles ou l'IMC)[f].

b. Contrôle qualité

Le génotypage des femmes de GENESIS a été réalisé par la plateforme de séquençage Genome Quebec[g]. Nous avons ensuite récupéré les données brutes et un contrôle qualité des SNPs a été réalisé par Christine Lonjou (Inserm U900, Institut Curie) à l'aide du logiciel PLINK[188]. Certains SNPs ont été exclus des analyses après les étapes de contrôle qualité suivantes :

- *Équilibre d'Hardy-Weinberg*

La première étape du contrôle qualité consiste à vérifier que les SNPs respectent l'équilibre d'Hardy-Weinberg. C'est un principe fondamental de la génétique des populations qui a été démontré par le mathématicien anglais Hardy et le médecin allemand Weinberg en 1908[189]. Selon ce principe et sous les conditions de population à l'équilibre (c'est-à-dire sans dérive génétique), les fréquences alléliques dans une population restent constantes au cours des générations. Dans le cas d'un locus bi-allélique A/a, les fréquences des génotypes AA, Aa et aa sont respectivement p^2, $2pq$ et q^2 (avec p la fréquence de l'allèle A et q (= 1-p) celle de l'allèle a). L'équilibre d'Hardy-Weinberg implique que $p^2 + 2pq + q^2 = 1$. Dans la population de témoins, la fréquence des allèles de chaque SNP a donc été calculée afin d'étudier les SNPs qui respectent ce principe pour ne pas conclure à tort à une association (artefact possible dans une population qui n'est pas à l'équilibre). On part du principe qu'un écart à la panmixie[h] trop important résulte plus probablement d'une erreur de génotypage que d'un réel déséquilibre lié à une dérive génétique. Pour le vérifier, la structure des fréquences génotypiques obtenues à partir des données observées est comparée à celle des fréquences estimées selon l'équilibre d'Hardy-Weinberg, grâce à un test du Chi² de Pearson. Dans un

[f] La sélection détaillée des SNPs de la puce iCOGS est disponible via le lien : https://ccge.medschl.cam.ac.uk/files/2014/03/iCOGS_detailed_lists_ALL1.pdf, consulté le 23 octobre 2019.
[g] http://www.genomequebec.com, consulté le 23 octobre 2019.
[h] Principe qui considère que les sujets sont répartis de manière homogène au sein de la population et se reproduisent aléatoirement.

premier temps, les génotypes observés (O) sont dénombrés afin d'en déduire les fréquences alléliques :

$$p = \frac{2 * O(AA) + O(Aa)}{2 * (\, O(Aa + O(AA) + O(aa))} \quad et \quad q = 1 - p$$

Si l'équilibre d'Hardy-Weinberg est respecté, la fréquence attendue (E) de chacun des génotypes est :

$$E(AA) = p^2$$
$$E(Aa) = 2qp$$
$$E(aa) = q^2$$

Le Chi² à 1 ddl s'écrit : $\sum \frac{(O-E)^2}{E}$. La taille de notre population d'étude n'étant pas très grande, nous avons décidé d'utiliser un seuil d'exclusion élevé. Tous les SNPs ayant un test significatif au seuil $\alpha = 0{,}001$ dans la population témoin de GENESIS ont été exclus.

- *« Call-rate »*

Le taux de données non manquantes par SNP, appelé « call rate », a été calculé afin d'exclure les SNPs pour lesquels un nombre trop important de participantes avaient un génotype manquant. Le seuil de « call rate » a été fixé à 90 %, entrainant l'exclusion des SNPs ayant plus de 10 % de données manquantes, SNPs qui ne seront donc pas analysés.

- *Fréquence de l'allèle mineur (MAF)*

Notre population d'étude étant de petite taille, les SNPs rares ont peu de chance d'y être correctement représentés. Cet artéfact peut mener à de faux-positifs. Nous avons donc décidé de ne pas analyser les SNPs rares et ceux dont la fréquence pour l'allèle mineur est inférieure à 0,05 ont été exclus.

Au total, 197 025 SNPs de la puce iCOGS ont été conservés après le contrôle qualité et ont été utilisés pour l'imputation des SNPs non génotypés.

2. Voies biologiques d'intérêt

Le but de ce projet est d'étudier l'effet de facteurs environnementaux sur l'association entre les SNPs et le risque de cancer du sein. Pour cela, nous avons décidé de nous intéresser dans un premier temps à deux familles de mécanismes biologiques particuliers : les voies biologiques impliquées dans la vie reproductive et celles impliquées dans la réponse cellulaire à l'exposition aux radiations ionisantes. Le choix de ces deux familles de mécanismes a été motivé par leur importance biologique dans l'étiologie du cancer du sein. La recherche des voies biologiques intervenant dans ces mécanismes a été réalisée grâce à la base de données KEGG[190]. Cette base de données regroupe des informations moléculaires qui sont utilisées dans le but de comprendre les systèmes biologiques à l'échelle de la cellule, de l'organisme ou de l'écosystème. Elle m'a permis de générer la liste des gènes intervenant dans les voies de signalisation d'intérêt. Les analyses d'association ont ensuite été réalisées sur tous les SNPs localisés dans ces gènes.

a. Régulation et action des hormones impliquées dans la reproduction

Comme expliqué dans la partie État de l'art, les hormones de la reproduction jouent un rôle très important dans le développement des glandes mammaires. Les gènes intervenant dans leur synthèse, leur régulation et leur action peuvent donc jouer un rôle dans le développement d'un cancer du sein. Les étapes de signalisation de ces hormones sont très dépendantes des différentes étapes de la vie reproductive (menstruation, grossesses, allaitement, etc.). L'impact d'une variation génétique dans les gènes intervenant dans ces voies de signalisation peut donc fortement varier d'une femme à l'autre. C'est pourquoi nous avons décidé d'étudier particulièrement les gènes intervenant dans les voies de signalisation biologique des hormones de la reproduction.

Au total, la base de données KEGG recense 14 voies de signalisation biologiques intervenant dans la régulation hormonale : la voie de signalisation de l'hormone GnRH (impliquant les hormones hypophysaires FSH et LH), les voies de régulation et d'action des hormones stéroïdes ovariennes (œstrogènes et progestérone), la voie de signalisation de la prolactine, les voies de synthèse et d'action des hormones thyroïdienne, la voie de signalisation de l'ocytocine, les voies de signalisation et de resistance de l'insuline, la voie de synthèse et de

sécrétion du cortisol et la voie de la tyrosine kinase EGFR. Ces voies comptent 722 gènes répartis sur tout le génome (Annexe 3, page 233) et la puce iCOGS cible 5 972 SNPs localisés dans ces gènes.

b. Réparation de l'ADN

Au cours de la vie d'une cellule, le patrimoine génétique contenu dans notre ADN peut être modifié par le biais d'erreurs ou d'agression externes. Ces modifications passent la plupart du temps inaperçues car elles sont corrigées par des mécanismes réparateurs. Ces mécanismes sont pris en charge par des gènes, oncogènes ou suppresseurs de tumeurs, dont le rôle est de réparer l'ADN et de contrôler la multiplication cellulaire. Une variation de l'ADN peut avoir lieu dans l'un de ces gènes, ce qui pourrait modifier sa fonction et interférer dans la réparation. Une des agressions externes potentiellement mutagène est l'exposition aux radiations ionisantes. La combinaison des expositions aux radiations et des variations génétiques déjà présentes dans les gènes responsables du maintien de l'intégrité du génome pourraient avoir un effet encore plus important sur le risque de développer un cancer du sein. Les interactions entre les variations génétiques dans ces gènes et les expositions aux radiations ont donc été analysées.

À partir de la base de données KEGG, 8 catégories de gènes intervenant dans les mécanismes de maintien du génome ont été établies : le cycle cellulaire, la réplication de l'ADN, la réparation de l'ADN double brin incluant la recombinaison homologue, la jonction d'extrémités non homologues, la réparation par excision de nucléotides, la réparation de mésappariement, la voie de signalisation PI3/AKT et la voie de signalisation de l'anémie de Fanconi. Au total, ces mécanismes regroupent 607 gènes (Annexe 4, page 260) qui contiennent 6 970 SNPs de la puce iCOGS.

III. Les données environnementales

Les informations relatives aux facteurs environnementaux ont été collectées grâce au questionnaire épidémiologique auquel chaque femme de GENESIS a répondu.

La cohérence des informations communiquées par les femmes a été vérifiée avec l'aide de la PIGE afin de générer les variables utilisées par la suite dans les analyses. Toutes les questions

qui ont été mal ou non renseignées ont été codées comme données manquantes avant d'être imputées.

1. La censure

Seuls les évènements ayant eu lieu avant le cancer du sein sont pris en compte afin de pouvoir déterminer les facteurs étiologiques de la maladie. 93,6 % des cas de GENESIS étant des cas prévalents, les informations renseignées dans le questionnaire portent sur les périodes pré- et post-diagnostic du cancer. C'est pourquoi les informations ont été censurées à l'âge au diagnostic pour les cas index, pour considérer uniquement les informations relatives à la période pré-diagnostic. Pour les témoins, nous avons pris en compte toutes les informations renseignées jusqu'à la date de l'interview. Dans les analyses qui ont suivi, nous avons donc pris l'âge au diagnostic au premier cancer du sein comme âge à la censure pour les cas et l'âge à l'interview pour les témoins.

Il faut noter que le diagnostic d'un cancer peut avoir lieu en même temps qu'un autre évènement indépendant, comme une consultation chez le gynécologue pour une grossesse ou une radiographie. Pour s'assurer de ne prendre en compte que les évènements qui ont eu lieu avant le développement du cancer, nous avons décidé de ne considérer que les évènements ayant eu lieu un an et plus avant la censure. Nous avons appliqué cette règle aux cas et aux témoins.

2. Variables gynéco-obstétriques

Le questionnaire épidémiologique de l'étude GENESIS contient 10 questions relatives à la vie reproductive des femmes (Annexe 2, page 200). J'ai généré 34 variables à partir de ces questions.

a. Variables relatives aux menstruations

Il est demandé aux femmes de renseigner leur âge aux premières règles ou, dans le cas où elles ne s'en souviennent plus, d'indiquer une classe d'âge parmi les catégories suivantes : « avant 12 ans », « entre 12 et 14 ans » et « après 15 ans ». J'ai harmonisé les réponses de

toutes les femmes et les ai classées selon ces 3 catégories. Deux catégories supplémentaires ont été générées pour les femmes ayant répondu qu'elles n'avaient jamais eu de règles d'une part et pour les données manquantes d'autre part.
La variable traitant de la périodicité des règles a également été recodée en 5 catégories : « régulière avec des cycles de 25 à 31 jours », « régulière avec des cycles de 24 jours et moins », « régulière avec des cycles de 32 jours et plus », « irrégulières », « inconnues ».

b. Variables relatives à la contraception

Une question regroupe toutes les informations relatives à la prise de contraceptifs. Les contraceptifs pris en compte ici sont les contraceptifs hormonaux regroupant la pilule contraceptive mais également le patch, l'implant sous-cutané ou le stérilet aux hormones. Les femmes ayant eu recours à cette forme de contraception devaient renseigner leur âge à la première utilisation et la durée totale d'utilisation en mois au cours de leur vie avant le diagnostic du cancer. Celles ayant eu une ou plusieurs grossesses devaient décrire leur utilisation au cours de 3 périodes de vie : avant leur première grossesse, entre la première et la dernière grossesse et après la dernière grossesse. J'ai généré 3 variables à partir de ces informations :

- une variable binaire contraceptif (oui/non) ;
- une variable catégorielle pour l'âge de la première utilisation (≤ 20 ans, > 20 ans, jamais) ;
- une variable catégorielle pour la durée totale d'utilisation en mois (≤ 5 ans, > 5 ans, jamais).

Pour la durée d'utilisation, j'ai dans un premier temps généré une variable continue en additionnant les durées d'utilisation pour chaque période. Cependant, nous avons des données incomplètes pour certaines femmes. Par exemple, nous connaissons la durée d'utilisation de contraceptifs avant la première grossesse mais pas celle au cours des deux autres périodes. Une première possibilité était de considérer les données incomplètes comme des données manquantes et d'ignorer la partie de l'information connue (dans l'exemple, la durée avant la première grossesse). Cependant, cette stratégie nous faisait perdre de l'information. Nous avons donc décidé de conserver cette information connue et de considérer la durée obtenue comme une durée minimum d'exposition. Cela concerne 266 cas et 221 témoins.

c. Variables relatives aux grossesses

Il est demandé aux femmes combien de fois elles ont été enceintes (quel que soit le type de grossesse) et le nombre d'enfants nés vivants. Chaque grossesse devait être décrite par l'âge au début et à la fin de la grossesse, la durée de la grossesse en mois, la date de l'accouchement et le nombre d'enfants. Ces grossesses sont rangées en deux catégories : celles qui ont été menées à terme, ayant abouti à un enfant vivant ou mort-né, et celles ayant été interrompues par une fausse couche, une interruption de grossesse volontaire (IVG) ou thérapeutique (ITG) ou une grossesse extra-utérine (GEU).

J'ai utilisé ces données pour générer 3 variables décrivant les grossesses menées à terme :

- une variable binaire grossesse menée à terme (oui/non) ;
- une variable catégorielle représentant l'âge à la première grossesse menée à terme (≤ 20 ans, [20 ans – 25 ans[, [25 ans – 30 ans[, ≥ 30 ans et nullipare) ;
- une variable catégorielle représentant le nombre de grossesses menées à terme (0, 1, 2, ≥ 3).

J'ai ensuite généré 2 variables pour décrire les grossesses interrompues, fausses couches et tous types d'avortements confondus :

- une variable binaire interruption de grossesse (oui/non) ;
- une variable catégorielle par le nombre d'interruptions de grossesse (0, 1, 2, ≥ 3).

En ce qui concerne le nombre d'occurrences de chacun de ces évènements, j'ai conservé toutes les informations connues et considéré ce nombre comme un nombre minimum d'occurrences d'évènements. Cela concerne peu de femmes, 3 cas index pour le nombre de grossesses menées à terme et 4 cas index pour le nombre de grossesses interrompues.

La durée d'allaitement devait également être renseignée pour chacune des grossesses menées à terme. Cela m'a permis de générer deux variables :

- une variable binaire allaitement (oui/non) ;
- une variable catégorielle donnant la durée d'allaitement en mois (jamais, ≤ 10 mois, > 10 mois).

Cette variable correspond au nombre minimum de mois d'allaitement, dans le cas où les informations étaient incomplètes. Cela concerne 6 femmes (3 cas et 3 témoins).

d. Variables relatives à la ménopause

Les informations concernant la ménopause ont été recueillies à l'aide de plusieurs questions concernant la ménopause en soi, une éventuelle hystérectomie ou ovariectomie et la prise d'un traitement hormonal substitutif. Les variables générées à partir du questionnaire sont :

- une variable binaire ménopause (oui/non) ;
- une variable catégorielle pour l'âge à la ménopause (< 40 ans, [40 ans – 55 ans] ou > 55 ans) ;
- une variable catégorielle pour le type de ménopause (naturelle ou ovariectomie).

Dans le questionnaire, il était demandé aux femmes d'indiquer leur statut ménopausique, avec la définition classique de la ménopause : « sans règles depuis plus d'une année ». Cette définition inclut l'arrêt des règles provoqué par une hystérectomie, une chimiothérapie ou la prise de pilule contraceptive, évènements n'impliquant pas forcément l'arrêt définitif de l'activité ovarienne et donc la ménopause d'un point de vue physiologique. Une ovariectomie entraine une ménopause dans le cas où elle est bilatérale (ablation des deux ovaires). Cependant certaines femmes ayant subi l'ablation d'un seul ovaire ont déclaré une ménopause par ovariectomie. Le statut de ces femmes vis-à-vis de la ménopause est alors considéré comme inconnu.

En ce qui concerne la prise d'un traitement hormonal substitutif (THS), les réponses des femmes ont également été vérifiées. Un THS est constitué d'œstrogènes seuls ou d'une combinaison d'œstrogènes et de progestérone et vise à réduire les symptômes de la ménopause. Cependant certaines femmes ont déclaré des médicaments qui ne sont pas considérés comme des THS (par exemple, Stediril, Ovanon ou Adel qui sont des contraceptifs oraux, ou encore Yam proactif ou Evestrel qui sont des phytoestrogènes). Nous avons donc pris en compte dans l'analyse seulement les THS inclus dans la liste présentée dans le Tableau 4.

Tableau 4 - Traitements médicamenteux considérés comme des THS

Traitements hormonaux substitutifs		
Œstrogènes	Estrapatch	Trisequens
Estraderm (25, 50 ou 100 mg)	Aérodiol	Duova
Œstrogel	Femsept	Successia
Synapause (2 ou 4 mg)	Thais	Avadène
Estreva (pilule ou gel)	Oromone	Climodiène
Menorest	Climara	Naemis
Œsclim	Psysiogine	Angeliq
Dermestril	Estrofem	Femseptcombi
Provames 1 ou 2	Cycladiene	**Autre**
Lutestral	**Œstro-progestatif**	Evista
Ovestin	Climaston	Optruma
Œstrodose	Kilogest	Danatrol
Prémarin	Livial	
Progynova	Divina	
Systen	Activelle	
Provames	Climene	
Delidose	Femoston Conti	

e. Indice de Masse Corporelle

La taille et le poids des femmes sont aussi renseignés dans le questionnaire, ce qui m'a permis de calculer l'IMC ($= \frac{Taille\ (cm)^2}{Poids\ (kg)}$). Cette variable a ensuite été discrétisée en 3 catégories comme définies par l'Organisation Mondiale de la Santé (OMS) : corpulence normale avec un IMC compris entre 18,5 et 25, insuffisance pondérale avec un IMC inférieur à 18,5 et état de surpoids avec un IMC supérieur ou égal à 25[i]. Une catégorie supplémentaire a également été générée pour les femmes dont l'information est inconnue (donnée manquante).

3. Variables liées aux expositions aux radiations

Le questionnaire de GENESIS contient 9 questions relatives aux expositions aux radiations, expositions professionnelles ou médicales (Annexe 2, page 200). Cependant, ici nous nous sommes intéressés seulement aux radiations ionisantes à la poitrine. Au total, les 5 variables suivantes ont été générées et étudiées par Maximiliano Guerra (professeur invité, Université de Juiz de Fora, Brésil) en 2017 au cours de son année sabbatique dans l'équipe (article en préparation) :

[i] Organisation Mondiale de la Santé : Obésité et Surpoids (https://www.who.int/fr/news-room/fact-sheets/detail/obesity-and-overweight, consulté le 23 octobre 2019).

- Une variable binaire exposition aux radiations oui/non ;
- Une variable catégorielle pour le nombre d'expositions (0, 1-3, 4-9, 10 ou plus) ;
- Une variable catégorielle pour l'âge à la première exposition (après 20 ans, entre 15 et 19 ans, avant 15 ans) ;
- Une variable catégorielle pour l'exposition aux radiations en fonction de la première grossesse menée à terme (après ou avant) ;
- Une variable catégorielle pour le nombre d'expositions en fonction de la première grossesse menée à terme (moins de 5 expositions après la première grossesse menée à terme, plus de 5 expositions après la première grossesse menée à terme, moins de 5 expositions avant la première grossesse menée à terme et plus de 5 expositions avant la première grossesse menée à terme).

Méthodes

La recherche de nouveaux facteurs génétiques a été réalisée en intégrant des données sur les expositions environnementales (expositions aux facteurs de la reproduction et expositions aux radiations ionisantes). Les données manquantes concernant ces expositions ou les SNPs non génotypés ont dû être imputées. Pour ce faire, plusieurs stratégies d'imputation ont été employées.

I. Imputation des données manquantes

Les analyses statistiques ont été réalisées sur la population complète, sans exclure les sujets ayant des données manquantes. Dans notre population d'étude 657 sujets ont une donnée manquante dans au moins une des variables gynéco-obstétriques d'intérêt et 757 sujets pour les variables liées aux radiations. Leur exclusion diminuerait de façon drastique (plus de 20 %) la taille de la population et donc la puissance statistique. De plus, d'après la littérature, la stratégie du *case-complete* est acceptable seulement lorsque le processus des données manquantes est un processus appelé *Missing Complety At Random* (MCAR)[191]. Cela signifie qu'il n'existe aucune différence entre les individus ayant et n'ayant pas de données manquantes. Même si cela n'est pas vérifiable en pratique, il est très peu probable que les données manquantes des variables d'intérêt soient MCAR. Dans ce cas, les résultats obtenus à partir du sous-ensemble de sujets sans aucune donnée manquante seraient biaisées. J'ai donc créé une classe supplémentaire pour les données manquantes. Ainsi, tous les sujets de l'étude sont pris en compte dans les analyses. Ces données manquantes ont été, ou non, imputées par la suite en fonction du type de données et des analyses effectuées.

1. Imputation simple

Les premières analyses ont été réalisées en créant une catégorie supplémentaire pour les données manquantes. Cependant, il existe presque 10 % de données manquantes pour la variable ménopause. J'ai donc appliqué une procédure d'imputation simple pour cette

variable. Cette méthode consiste à remplacer chaque valeur manquante par une prédiction obtenue à partir des autres observations (moyenne, médiane, quantile, etc.).

Deux stratégies pouvaient être utilisées pour imputer ces données manquantes :

i. l'imputation de l'âge médian à la ménopause des témoins ;
ii. l'imputation de l'âge à la ménopause chez les témoins aux percentiles extrêmes de la distribution (1 % et 99 %). Cela m'a permis de définir l'âge où 99 % des témoins ne sont pas ménopausées (age_{min}) et celui où 99 % d'entre elles le sont (age_{max}) et, ainsi, d'imputer les données manquantes de la ménopause avec une erreur *a priori* de 1 %. On considère ainsi que toutes les femmes dont le statut ménopausique est inconnu sont ménopausées si elles ont un âge à la censure supérieur ou égal à age_{max} et ne le sont pas si elles ont un âge inférieur ou égal à age_{min}. Toutes celles qui ont un âge à la censure compris entre age_{min} et age_{max} exclus, conservent un statut manquant pour la variable ménopause.

J'ai décidé d'utiliser la deuxième stratégie pour traiter les données manquantes. Cette stratégie aboutit à une imputation simple partielle avec une partie des femmes qui n'ont pas été classées et conservent alors leur catégorie « donnée manquante ». Après cette imputation simple, il ne reste plus que 2,9 % des femmes avec une donnée manquante pour cette variable ménopause, soit plus de 3 fois moins qu'avant imputation.

2. Imputation multiple

Après avoir sélectionné les variables associées au cancer du sein dans notre population, l'étape suivante était de calculer un score de risque pour chaque femme. Pour cela il était nécessaire que chacune des femmes ait une valeur pour chaque variable prise en compte. Il a alors fallu imputer les données manquantes via une imputation multiple, à l'aide du logiciel Stata 14[192].

L'imputation multiple, proposée par Rubin en 1978[193], consiste à remplacer les valeurs manquantes par plusieurs valeurs estimées. Ces estimations sont basées sur un modèle qui permet d'utiliser toute l'information disponible dans la base de données et ainsi de préserver les relations entre les variables. Le processus d'imputation est effectué plusieurs fois, produisant plusieurs sets imputés pour un même sujet (d'où le terme « imputation multiple »).

Des analyses statistiques standards sont alors réalisées sur chacune des imputations puis les résultats obtenus pour un même sujet sont combinés pour produire une analyse globale. À la différence de l'imputation simple, l'estimation de plusieurs valeurs imputées pour un même sujet permet de prendre en compte la variabilité naturelle des données manquantes ainsi que l'incertitude provoquée par leur estimation.

L'imputation multiple se découpe en trois phases :

- La phase d'imputation où les données manquantes sont estimées M fois.

Deux algorithmes principaux peuvent alors être utilisés pour estimer les données manquantes : le premier est basé sur l'hypothèse d'une distribution multivariée normale[194] et le deuxième est basé sur l'imputation multiple par équations chainées[195].
Le premier algorithme fait l'hypothèse que toutes les variables incluses dans le modèle utilisé pour l'imputation suivent une distribution jointe. Cela n'est pas toujours vérifié, notamment lorsque le modèle inclut des variables binaires et catégorielles. Aucune de nos variables n'étant des variables continues, ce modèle n'était pas adapté.
Nous avons donc utilisé le modèle d'imputation par équations chainées (commande MI IMPUTE CHAINED[192] dans Stata). Cette méthode ne fait pas l'hypothèse d'une distribution multivariée normale et permet de déterminer une distribution spécifique pour chaque variable conditionnellement aux autres variables incluses dans le modèle d'imputation. Concrètement, ce modèle permet de ramener un problème multivarié à k dimensions (avec k, le nombre de variables à imputer), à k problèmes univariés successifs en conditionnant à chaque fois une variable imputée aux valeurs observées des autres variables. Les estimations des données manquantes sont alors définies grâce à un tirage au sort dans les distributions plausibles des données manquantes en fonction des données observées. Ce tirage au sort s'effectue grâce à un échantillonnage de Gibbs[196].

- La phase d'analyse séparée

Chaque base de données résultant des M imputations est analysée séparément en utilisant les méthodes classiques d'analyse (une régression logistique dans notre cas). Les résultats obtenus entre les M régressions ne varieront qu'en fonction des valeurs imputées des données manquantes.

- La phase d'analyse combinée

Les résultats des régressions individuelles effectuées sur les M bases de données sont alors combinés. Dans notre cas, les paramètres que l'on cherche à estimer sont les coefficients β associés à chaque variable et leur variance. Pour cela, le paramètre β est estimé en faisant la moyenne des paramètres obtenus dans les M imputations. La variance de chaque β est estimée par la somme combinée des variances intra- et inter-imputation.

Le nombre d'imputations M varie généralement entre 3 et 10 pour des raisons de temps de calcul. Cependant, il n'y a pas réellement de consensus sur ce paramètre et plus ce nombre M est élevé, plus les résultats obtenus seront précis[197]. Compte-tenu du nombre non négligeable de données manquantes, surtout pour les variables relatives aux expositions aux radiations, nous avons décidé d'imputer 100 jeux de données (M = 100).

3. Imputation des SNPs non génotypés

Plus de 80 millions de SNPs répartis sur le génome entier sont aujourd'hui répertoriés dans la base de données publique dbSNP[198]. Pour des questions de coût et de temps et malgré les avancées technologiques, il est impossible de génotyper tous ces SNPs[j]. La puce iCOGS ne contient qu'une infime partie (moins de 0,01 %) de l'ensemble des SNPs connus. Toutefois, à partir des SNPs génotypés, il est possible de prédire la valeur des SNPs non génotypés grâce au déséquilibre de liaison (DL) entre les SNPs d'une même région (Annexe 5, page 276). Les SNPs de la puce ont été sélectionnés pour leur capacité à marquer un groupe de SNPs ou même une région entière grâce à ce déséquilibre de liaison. On parle de marqueurs. Ces derniers sont en DL avec un nombre important d'autres SNPs localisés sur le même haplotype, ce qui nous permet alors de prédire la valeur des SNPs non génotypés de façon assez précise grâce à leur imputation à l'aide d'une base de données de génomes entiers de référence, comme ceux présents dans le panel de référence *1000Génomes*[199]. Cependant, la puce iCOGS ne contient pas de squelette de SNPs « marqueurs » tout le long du génome. Ce squelette, appelé *backbone SNP* en anglais et présent sur certaines puces (par exemple sur la puce OncoArray d'Illumina[140]), représente un ensemble de marqueurs localisés à équidistance et répartis sur tout le génome. Sans ce squelette, la qualité de l'imputation ne sera pas égale

[j] À l'heure actuelle, la puce à ADN la plus grande est la puce *Illumina Omni 5M array*. Elle contient environ 5 millions de SNPs.

entre les différentes régions du génome car certaines régions sont moins bien représentées que d'autres sur la puce. Cela a été pris en compte par la suite.

L'imputation des sujets, cas et témoins, de GENESIS a été effectuée conjointement. La première étape a été de construire les haplotypes avec le logiciel SHAPEIT[200] pour ensuite inférer les SNPs non génotypés grâce à l'imputation avec le logiciel IMPUTE2[201] à partir du panel de référence *1000Genomes*[199].

a. Pré-haplotypage avec le logiciel SHAPEIT

Pour pouvoir prédire la valeur des SNPs non génotypés, il est nécessaire de connaître la combinaison allélique de chaque sujet sur un même chromosome (haplotype) (Annexe 5, page 276). Cette combinaison allélique s'appelle la phase. Les données génotypiques que l'on obtient grâce à une puce à ADN comme la puce iCOGS ne contiennent pas cette information de phase. Une première étape de pré-haplotypage (ou pré-phasage) est donc nécessaire pour reconstruire cette information à partir de l'ensemble des génotypes d'une population étudiée. Le pré-haplotypage des sujets de GENESIS a été réalisé avec le logiciel SHAPEIT[200,202]. L'algorithme de SHAPEIT est basé sur un modèle de Markov Caché (HMM) utilisant un échantillonnage de Gibbs. Cette étape de pré-haplotypage permet de construire les haplotypes des femmes de GENESIS à partir des données de génotypage de la puce iCOGS et de la carte de recombinaison génétique issue de l'étude HapMap2[203]. Cette carte donne le taux de recombinaison entre des positions génétiques d'un chromosome donné. SHAPEIT définit dans un premier temps tous les haplotypes possibles pour chaque femme à partir des données génotypiques disponibles et des taux de recombinaison. Puis, selon le principe d'un échantillonnage de Gibbs, un très grand nombre d'itérations vont permettre de converger vers l'haplotype le plus vraisemblable pour chacune des femmes. À chaque itération, l'haplotype d'une femme de la population d'étude est défini à partir des paramètres de recombinaison de HapMap2 et des haplotypes de l'ensemble des femmes de l'étude à l'étape précédente (itération n-1). L'haplotype final dépend de l'haplotype de départ qui est attribué au hasard pour chaque sujet parmi tous les haplotypes possibles. Cet haplotype final n'est donc pas forcément l'haplotype le plus probable sachant les données génotypiques de la femme, mais il fait partie des haplotypes les plus probables.

Cette première étape d'haplotypage pourrait se faire simultanément à l'imputation grâce au logiciel IMPUTE2. En effet, le panel de référence utilisé par ce logiciel pour l'imputation est « phasé » (c'est-à-dire que les haplotypes des sujets sont disponibles) et permet d'imputer les génotypes non phasés de la population d'étude. Cependant, le temps de calcul serait beaucoup trop important étant donné la taille de nos données et celle des données de référence. Le panel de référence *1000Genomes* contenait au départ le génome entier de 1 000 sujets différents mais aujourd'hui il contient plus de 2 500 génomes entiers. Le fait d'avoir un panel de référence de plus en plus grand augmente la puissance et la qualité de l'imputation mais également la complexité computationnelle et le temps de calcul. L'utilisation préliminaire de SHAPEIT pour phaser les génotypes de GENESIS s'accompagne donc d'une baisse de précision dans l'inférence des génotypes effectuée par la suite, mais cette baisse de précision est négligeable en regard du temps de calcul économisé.

b. Imputation avec le logiciel IMPUTE2

Description

L'imputation des SNPs non génotypés est réalisée avec IMPUTE2[201,204] à partir des haplotypes prédits par SHAPEIT. Un panel d'haplotypes de référence et la carte de recombinaison génétique de chaque chromosome sont nécessaires pour cette imputation. Les déséquilibres de liaison entre les SNPs non génotypés et les SNPs génotypés sont calculés à partir d'un panel de référence. Ce déséquilibre de liaison permet alors d'estimer une probabilité pour chacun des 3 génotypes possibles (AA, Aa ou aa) pour chaque SNP imputé.

Panel de référence

Le panel de référence utilisé est celui qui a été développé dans le cadre de la Phase 3 du projet *1000Genomes*[199]. Le but de ce projet était de répertorier la majorité des variants génétiques ayant une fréquence supérieure à 1 % à partir du séquençage du génome de 2 504 sujets. La mise en place de cette ressource donnant l'accès à la variabilité génétique humaine dans le monde entier a été possible grâce au développement des technologies de séquençage dont le coût a diminué de façon drastique. La base de données de référence *1000Genome* est composée des haplotypes de 2 504 sujets de différentes origines ethniques, avec 503 sujets

d'origine européenne, 347 d'origine américaine, 489 d'origine d'Asie du sud, 504 d'origine d'Asie de l'est, et 661 d'origine africaine. Ce panel capture un large éventail de la diversité de la génétique humaine.

Avant d'effectuer l'imputation, il a été nécessaire de décider si une partie ou la totalité du panel devait être utilisé. En effet, il existe de très fortes variations de fréquence allélique selon l'origine ethnique des sujets. Certains variants sont fréquents chez les européens et quasiment absents chez les sujets d'origine africaine et inversement. Notre population d'étude étant restreinte aux femmes d'origine européenne, il semblait à première vue plus approprié d'utiliser un panel de référence composé seulement de sujets de la même origine afin de ne pas biaiser l'imputation. Cependant, les populations de *1000Genomes* ont une histoire démographique complexe avec beaucoup de migrations et de mélanges ethniques. Pour la plupart des SNPs, les sujets d'origine européenne du panel de référence seront plus proches des femmes de notre étude et permettront la meilleure prédiction des SNPs à imputer. En revanche, pour certains SNPs plus rares, des sujets d'origines ethniques différentes pourront apporter de l'information supplémentaire et permettront une meilleure prédiction des SNPs non génotypés. En effet, des sujets d'origine ethnique différente peuvent partager des segments génomiques d'ancêtres communs récents et IMPUTE2 peut alors utiliser ces haplotypes communs pour améliorer la précision de l'imputation. C'est pourquoi les développeurs et utilisateurs d'IMPUTE2 préconisent d'utiliser le panel de référence complet. Les imputations ont donc été réalisées en suivant ces recommandations. Pour les mêmes raisons, nous avons utilisé tous les SNPs présents dans la base de données de référence quelle que soit leur fréquence chez les sujets européens du panel de référence.

Parmi les 5 008 (2 x 2 504) haplotypes disponibles dans la base de données de référence, IMPUTE2 sélectionne ceux à utiliser pour l'imputation. Il compare le génotype en cours d'imputation à celui de tous les sujets de *1000Genomes* et sélectionne les k sujets de référence ayant le génotype le plus proche de celui dont l'imputation est en cours. C'est à partir de ces k génotypes de référence qu'il prédit la valeur des SNPs non génotypés. Il a donc fallu également déterminer le nombre d'haplotypes de référence k_hap qui devait être utilisé par IMPUTE2. Par défaut, IMPUTE2 utilise les 500 haplotypes de référence les plus proches (k_hap = 500) de l'haplotype à imputer. Afin de déterminer la valeur k_hap la plus adaptée à nos données, j'ai réalisé un test sur un fragment du chromosome 7 allant de la position

52 672 577 pb à 57 672 577 pb. Ce fragment a été imputé plusieurs fois, en faisant varier la valeur k_hap de 300 à 800 avec un incrément de 50.
Pour déterminer la valeur à utiliser pour nos imputations, j'ai utilisé le fait qu'en plus d'imputer les SNPs non génotypés, IMPUTE2 prédit également la valeur de tous les SNPs génotypés dans le but de vérifier la qualité de l'imputation (validation croisée). Cette validation croisée consiste à masquer les valeurs des SNPs génotypés une à une et à les imputer en se basant sur les haplotypes de référence et les autres SNPs génotypés. Les génotypes observés sont alors comparés aux valeurs imputées pour chaque SNP. On considère qu'une imputation est de bonne qualité si le taux de concordance entre les valeurs observées et celles prédites est supérieur ou égal à 0,95. Les résultats sont alors résumés dans des tables de concordance (Tableau 5). Celles-ci présentent le taux de concordance entre la valeur observée et celle imputée en fonction de la qualité d'imputation des SNPs. En effet, IMPUTE2 calcule un score pour chaque SNP noté r^2 reflétant la qualité de l'imputation (à ne pas confondre avec le r^2 représentant le déséquilibre de liaison entre deux SNPs). Le calcul de ce score est détaillé dans l'Encadré 1 page 79.

J'ai utilisé les résultats de cette validation croisée pour définir la valeur k_hap associée à une qualité d'imputation optimale. À partir des tables de concordance obtenues après chacune des imputations faites sur le même fragment du chromosome 7 (Tableau 5), on remarque que le pourcentage de concordance est équivalent pour les 11 valeurs k_hap testées (entre 0,919 % et 0,920 %) pour les SNPs associés à une forte qualité après imputation (c'est-à-dire un r^2 compris entre 0,9 et 1,0). Pour les SNPs associés à une faible qualité, ce pourcentage de concordance est à 0 pour les 11 imputations.
Comme attendu, la plus grande variation concerne les SNPs associés à une qualité moyenne (r^2 compris entre 0,4 et 0,7), dont la valeur fluctue d'une imputation à l'autre. Cette variation reste faible, de 1 % à 5 % et les imputations ayant la meilleure précision sont celles avec une valeur k_hap située entre 450 et 600.

Nous avons donc décidé de conserver la valeur conseillée par IMPUTE2, soit k_hap = 500. L'imputation des génotypes des femmes de GENESIS a donc été effectuée grâce aux 500 sujets du panel de référence ayant le génotype le plus proche.

Encadré 1 – Calcul du r^2

Calcul du score de qualité d'imputation r^2

Le score r^2 est une estimation de la qualité de l'imputation pour chaque SNP. Il correspond à l'estimation du coefficient de corrélation entre la valeur du SNP selon l'imputation et sa valeur réelle. Ce score permet de déterminer la confiance que nous pouvons avoir dans les résultats obtenus par la suite dans les analyses d'association.

Notons Y le SNP, X le dosage estimé après imputation et i les sujets.

Alors $\boldsymbol{\mu_x = \mu_y}$ et donc $\boldsymbol{E(X - \mu_x)(Y - \mu_y) = E(X - \mu_x)\{(X - \mu_x) + (Y - X)\}}$.

Puis $\boldsymbol{E(X - \mu_x)(Y - \mu_y) = E(X - \mu_x)^2 + E\{(Y - X)(X - \mu_x)\}}$.

Pour chaque sujet, X prend la valeur x_i qui est une valeur constante donc

$\boldsymbol{E\{(Y - X)(X - \mu_x)\} = (E(y_i) - x_i)(x_i - \mu_x) = 0}$ comme $\boldsymbol{E(y_i) = x_i}$

$\boldsymbol{E(X - \mu_x)^2 = \sum_i (x_i - \mu_x)/N}$

On a $\boldsymbol{E(Y - \mu_Y)^2 = \sum_i \{\mathrm{p}(y_i = 0)\mu_y{}^2 + \mathrm{p}(y_i = 1)(1 - \mu_y{}^2) + \mathrm{p}(y_i = 2)(2 - \mu_y{}^2)\}/N}$

Donc $\boldsymbol{r^2 = \rho^2 = \frac{\{E(X - \mu_x)(Y - \mu_y)\}^2}{E\{(X - \mu_x)^2\}E\{(Y - \mu_y)^2\}} = \frac{E(X - \mu_x)^2}{E(Y - \mu_y)^2}}$ avec ρ^2 l'estimation de r^2.

Tableau 5 - Taux de concordance obtenus après imputation du fragment de 52 672 577 pb à 57 672 577 pb du chromosome 7 avec un k_hap variant de 300 à 800

k_hap = 300

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	2765	40,7
[0,5-0,6]	20890	48,8
[0,6-0,7]	20655	56,5
[0,7-0,8]	23320	61,7
[0,8-0,9]	34280	71,2
[0,9-1,0]	779256	92,1

k_hap = 350

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	2911	39,6
[0,5-0,6]	21401	48,5
[0,6-0,7]	21200	55,8
[0,7-0,8]	23704	62,1
[0,8-0,9]	35168	70,9
[0,9-1,0]	776782	92,1

Intervalle : intervalles de r^2 c'est-à-dire la qualité l'imputation d'un SNP ; #Génotypes : nombre de SNPs associés à cet intervalle de confiance par IMPUTE2 ; %Concordance : taux de SNPs bien imputés parmi ceux présents dans cet intervalle.

k_hap = 400

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	3025	39,0
[0,5-0,6]	21885	48,5
[0,6-0,7]	21531	55,2
[0,7-0,8]	24253	61,9
[0,8-0,9]	36176	71,1
[0,9-1,0]	774296	92,1

k_hap = 450

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	3227	39,3
[0,5-0,6]	22236	48,2
[0,6-0,7]	21793	55,7
[0,7-0,8]	24657	61,6
[0,8-0,9]	36771	71,0
[0,9-1,0]	772482	92,0

k_hap = 500

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	3251	38,7
[0,5-0,6]	22660	48,5
[0,6-0,7]	22012	55,0
[0,7-0,8]	25102	61,8
[0,8-0,9]	37151	71,1
[0,9-1,0]	770990	92,0

k_hap = 550

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	3298	38,7
[0,5-0,6]	23100	48,4
[0,6-0,7]	22267	54,9
[0,7-0,8]	25334	61,6
[0,8-0,9]	37841	71,1
[0,9-1,0]	769326	92,0

k_hap = 600

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	3395	38,3
[0,5-0,6]	23392	48,4
[0,6-0,7]	22566	55,1
[0,7-0,8]	25910	61,8
[0,8-0,9]	38281	71,1
[0,9-1,0]	767622	92,0

k_hap = 650

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	3476	36,9
[0,5-0,6]	23615	48,2
[0,6-0,7]	22871	55,3
[0,7-0,8]	26188	61,7
[0,8-0,9]	38726	71,1
[0,9-1,0]	766290	92,0

Intervalle : intervalles de r^2 c'est-à-dire la qualité l'imputation d'un SNP ; #Génotypes : nombre de SNPs associés à cet intervalle de confiance par IMPUTE2 ; %Concordance : taux de SNPs bien imputés parmi ceux présents dans cet intervalle.

k_hap = 700

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	3515	38,2
[0,5-0,6]	23864	48,4
[0,6-0,7]	22941	55,2
[0,7-0,8]	26627	61,6
[0,8-0,9]	39344	71,0
[0,9-1,0]	764875	92,0

k_hap = 750

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	3579	39,3
[0,5-0,6]	24061	48,4
[0,6-0,7]	23192	55,1
[0,7-0,8]	26901	61,7
[0,8-0,9]	39664	70,9
[0,9-1,0]	763769	92,0

k_hap = 800

Intervalle r^2	#Génotypes	%Concordance
[0,0-0,1]	0	0,0
[0,1-0,2]	0	0,0
[0,2-0,3]	0	0,0
[0,3-0,4]	0	0,0
[0,4-0,5]	3651	40,0
[0,5-0,6]	24459	48,7
[0,6-0,7]	23392	54,7
[0,7-0,8]	27134	61,5
[0,8-0,9]	40012	71,0
[0,9-1,0]	762518	91,9

Intervalle : intervalles de r^2 c'est-à-dire la qualité l'imputation d'un SNP ; #Génotypes : nombre de SNPs associés à cet intervalle de confiance par IMPUTE2 ; %Concordance : taux de SNPs bien imputés parmi ceux présents dans cet intervalle.

Contrôle qualité des SNPs imputés

Contrôle de la qualité par fragment imputé

IMPUTE2 permet d'imputer des régions d'une longueur de 5 Mb maximum. Pour imputer le génome entier, il a donc été nécessaire de fragmenter le génome en 484 régions de 5 Mb. Une fois les imputations effectuées, j'ai vérifié que la précision de la prédiction était suffisante pour réaliser les études d'association. Pour cela, j'ai utilisé les tables de concordance générées par IMPUTE2 pour chaque fragment imputé. Les fragments ayant un taux de concordance total inférieur à 0,95 % ont été ré-imputés en augmentant la taille de la région. En effet, IMPUTE2 autorise l'imputation de régions de plus de 5 Mb en utilisant l'option *allow_large_region*, avec en contrepartie une augmentation du temps d'exécution.

Contrôle de la qualité par SNP

Plus les SNPs imputés ont un déséquilibre de liaison élevé avec les SNPs génotypés, plus leur qualité d'imputation r^2 est bonne. Dans les GWAS, le seuil habituellement utilisé pour exclure les SNPs selon leur qualité d'imputation est de 0,3. Ici, on s'attend à ce que la qualité d'imputation de la puce iCOGS soit plus faible que pour les puces GWAS du fait de l'absence du squelette de SNPs marqueurs le long du génome. C'est pourquoi nous avons augmenté le seuil et sélectionné pour la suite des analyses seulement les SNPs avec un r^2 supérieur ou égal à 0,5.

Fréquence de l'allèle mineur

Nous avons également exclu les SNPs dont l'allèle mineur était associé à une fréquence inférieure à 5 % dans la population témoin. En effet, les analyses ont été réalisées sur environ 2 500 sujets et elles ont été stratifiées selon les profils d'exposition aux facteurs environnementaux considérés. La puissance de ces analyses est donc faible du fait de la taille de notre population. Nous avons alors exclu les SNPs rares pour diminuer le nombre de résultats faux positifs.

II. La régression logistique

1. Description

Le but de la régression logistique est de caractériser les relations entre une variable à expliquer (dite variable dépendante) et une autre (régression logistique simple ou univariée) ou plusieurs autres variables dites « explicatives » prises en compte simultanément (régression logistique multiple ou multivariée). Ce modèle permet donc de relier la variable dépendante Y à des variables explicatives X_i. La régression logistique s'effectue sur une variable à expliquer dichotomique (ici, atteinte ou non d'un cancer du sein). Les variables explicatives peuvent être quantitatives, qualitatives ou binaires.

La fonction logistique est de la forme :

$$f(x) = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

Le modèle de régression logistique permettant d'étudier l'association entre le cancer du sein (M+) et des expositions (X_i) est donc :

$$P(M^+ | X) = \frac{1}{(1 + e^{-(\alpha+\beta_1 X_1+\cdots+\beta_n X_n)})}$$

avec α le risque de base et β_i le risque associé au facteur X_i et avec l'odds ratio (OR) égal à $e^{\beta i}$ (avec i allant de 1 à n).

La régression logistique permet de modéliser la probabilité que l'évènement M^+ (ici, le cancer du sein) ait lieu. Une *p-value* est calculée à partir de ce modèle. Cette valeur permet d'estimer le taux maximal d'erreur que l'on peut faire lorsque l'on rejette l'hypothèse nulle à tort, c'est-à-dire lorsque l'on conclut qu'il y a une association significative entre la variable Y et la variable X. En général, le seuil de signification *a priori* est fixé à 0,05. À partir d'une régression logistique, nous pouvons calculer l'odds ratio (OR) qui mesure une association entre l'exposition et l'évènement. Cependant, lorsque l'évènement étudié est un évènement rare (dont la prévalence est inférieure à 5 %) on considère que l'OR est une bonne estimation du risque relatif (RR).

Sachant que $logit(Z) = \ln\left(\frac{Z}{1-Z}\right)$ (avec Z une variable quelconque), le modèle logistique peut être simplifié par :

$$logit\big(P(M^+|X)\big) = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n$$

Le RR peut alors être estimé en calculant l'OR à partir de ce modèle avec $OR = e^{\beta}$.

2. Outils d'analyse

Trois logiciels ont été utilisés pour réaliser les analyses de régression logistique des facteurs étudiés :

- le logiciel R[205] avec la fonction GLM (*Generalized Linear Model*) pour les facteurs gynéco-obstétriques ;
- le logiciel STATA[192] et la fonction *logistic* pour les expositions aux radiations
- le programme *logitRegress*[k] développé par Jonathan Tyrer au *Centre for Cancer Genetic Epidemiology* à l'Université de Cambridge pour les SNPs.

Le programme *logitRegress* permet de réaliser une régression logistique à partir des résultats de l'imputation. Il requiert deux fichiers en entrée : les données imputées au format IMPUTE2 et un fichier contenant la liste des sujets, leur statut vis-à-vis du cancer du sein et éventuellement des variables supplémentaires d'intérêt sur lesquelles le modèle doit être ajusté (ici, l'âge à la censure). Pour chaque SNP et pour chaque sujet, le fichier IMPUTE2 nous donne la probabilité des 3 génotypes AA, Aa et aa, avec A l'allèle le plus fréquent, ou allèle majeur, et a l'allèle le plus rare, ou allèle mineur. Les données ne peuvent pas être utilisées dans ce format pour effectuer la régression logistique. Deux stratégies peuvent être utilisées pour les transformer :

- La première stratégie (Tableau 6) consiste à attribuer à un sujet le génotype le plus probable selon les résultats de l'imputation. Dans le cas de figure où, pour un SNP donné, une des trois probabilités est très élevée par rapport aux deux autres, il est facile d'attribuer un génotype à un sujet. Mais très souvent, nous nous retrouvons face à deux voire trois probabilités avec des valeurs assez proches et le fait de choisir le génotype associé à la probabilité la plus élevée n'est pas adapté. Pour contourner ce problème, il est nécessaire de définir un seuil de probabilité du génotype le plus probable en dessous duquel les SNPs seront exclus de l'analyse. Avec un seuil de probabilité fixé à 0,9, tous les SNPs n'ayant pas de génotype associé à une probabilité supérieure ou égale à 0,9 sont exclus. Les autres se verront attribuer le génotype le plus probable. Cette stratégie entraine l'exclusion de SNPs et donc la perte d'information, ce qui est amplifié par l'attribution d'un génotype unique.

[k] Programme est similaire au programme mlogit (https://ccge.medschl.cam.ac.uk/software/mlogit, consulté le 23 octobre 2019) qui n'est pas disponible en accès ouvert pour le moment.

- La deuxième stratégie (Tableau 6) consiste à transformer en dosage les données brutes imputées. Ce dosage correspond à une valeur unique calculée pour chaque SNP et pour chaque sujet à partir des probabilités des 3 génotypes. Cette valeur est calculée comme suit :

Dosage de l'allèle a = 0xP(AA) + 1xP(Aa) + 2xP(aa)

Un dosage proche de 0 signifie que le génotype homozygote AA est le plus probable alors qu'un dosage proche de 2 signifie que le génotype homozygote aa est le plus probable. Avec cette stratégie aucun SNP n'est exclu et les probabilités pour chaque génotype inféré par l'imputation seront prises en compte dans les analyses d'association. Le programme *logitRegress* utilise cette deuxième stratégie.

Tableau 6 - Exemple de transformation des données imputées d'un SNP avec A et a comme allèle majeur et mineur respectivement.

Sujets	Format IMPUTE2	Stratégie 1	Stratégie 2
1	**AA** 0,9 **Aa** 0,1 **aa** 0,0	Génotype = AA	Dosage = 0,1 (0x0,9 + 1x0,1 + 2x0,0)
2	**AA** 0,40 **Aa** 0,45 **aa** 0,15	Génotype = Aa	Dosage = 0,75 (0x0,40 + 1x0,45 + 2x0,15)

Le sujet 1 a une probabilité associée au génotype homozygote pour l'allèle majeur très élevée ; les deux stratégies donnent un résultat satisfaisant. Le sujet 2 a des probabilités quasiment identiques pour le génotype homozygote pour l'allèle majeur ou celui hétérozygote ; le résultat de la stratégie 1 est donc beaucoup moins acceptable. Ce SNP aurait été considéré comme manquant par la stratégie 1.

Pour chaque SNP, *logitRegress* réalise un test de rapport de vraisemblances. Le fichier de sortie contient le paramètre β qui mesure l'association entre le dosage du SNP et le cancer du sein. Il contient également la statistique de test du rapport de vraisemblance (LRT) à partir de laquelle nous pouvons calculer la *p-value* associée à l'association.

3. Facteurs confondants

L'utilisation d'un modèle de régression logistique permet de prendre en considération des facteurs confondants. Pour qu'un facteur soit considéré comme confondant, il faut qu'il y ait une association entre ce facteur, l'événement (ici, le cancer du sein) et l'exposition étudiée. Un facteur confondant non pris en compte peut amener à conclure à une association à tort ou à ne pas conclure. Ce type de biais peut être évité en ajustant le modèle avec le ou les facteur(s) confondant(s).

Les deux principaux facteurs que nous avons supposés comme confondants sont l'année de naissance et l'âge à la censure. En effet, lors de la mise en place du design de l'étude, il a été décidé que les témoins devaient avoir le même âge que les cas à l'interview plus ou moins 3 ans. Cependant, du fait que les cas sont à 63,4 % des cas prévalents, les cas et les témoins sont censurés à des périodes calendaires différentes. De ce fait, nous avons ajusté les analyses sur l'âge à la censure. L'âge au diagnostic des cas étant systématiquement plus jeune que l'âge à l'interview des témoins, cela a généré des décalages dans les années de naissance entre les cas et les témoins pouvant induire de potentiels biais (effet cohorte) si les facteurs étudiés varient au cours du temps. En effet, la vie reproductive des femmes a changé au cours des générations. Selon l'INSEE, le nombre moyen d'enfants chez les femmes nées en 1940 était de presque 3 alors qu'il est de 1,5 pour la génération de femmes nées en 1970[206]. De même, l'âge au premier enfant n'a cessé d'augmenter et il a atteint 28,5 ans en 2015 alors qu'il était autour de 24,0 ans en 1974[207]. La prise de contraceptifs ou de l'allaitement ont également changé au cours du temps. Cet effet cohorte est aussi retrouvé pour les expositions aux radiations car les connaissances vis-à-vis des effets délétères des radiations sur la santé ont entrainé une diminution des presciptions d'examens radiographiques mais également parce que les appareils radiographiques ont évolué et les doses délivrées ont diminué avec le temps[208].

4. Tests multiples

Le seuil de risque de première espèce habituellement utilisé pour les tests statistiques est $\alpha = 0{,}05$. Un test est donc déclaré statistiquement significatif lorsque la *p-value* associée est inférieure au seuil de 0,05. Ce seuil α est le risque d'erreur de type 1, c'est-à-dire la

proportion de résultats faux positifs que l'on accepte. C'est le seuil que nous avons utilisé pour les études d'association réalisées sur les facteurs environnementaux.
Cependant, ce seuil est trop élevé pour les analyses réalisées sur les facteurs génétiques. En effet, plus de 200 000 SNPs ont été analysés au total. Dans les cas où chacun de ces SNPs n'est pas en déséquilibre de liaison, le nombre de faux positifs à l'issue de ces 200 000 tests sera alors de 0,05x200 000 = 10 000. Ce seuil de signification a donc été corrigé en utilisant la méthode de Bonferroni, qui revient à rapporter le niveau α habituel au nombre de tests effectués. Notre approche étant une approche de type « gène candidat », nous avons défini ce seuil en rapportant le seuil α au nombre de gènes testés (n = 722), soit un seuil égal $\alpha^* = 7{,}0.10^{-5}$.

5. Stratégie pour définir des scores de risque

L'objectif est de regrouper les femmes selon leur profil de risque vis-à-vis des facteurs environnementaux. Pour ce faire, nous avons envisagé d'utiliser plusieurs méthodes de *clustering*. La première stratégie consiste à grouper les femmes selon leur probabilité *a posteriori* de développer un cancer du sein selon les facteurs environnementaux d'intérêt. Cette stratégie nécessite en amont de sélectionner les variables à intégrer dans un modèle multivarié pour la prédiction et de veiller à ce que ce modèle soit parcimonieux et que les variables choisies ne soient pas interdépendantes.

L'analyse de chaque facteur de risque d'intérêt m'a permis de construire un modèle de régression logistique multivarié parcimonieux contenant les facteurs influençant le risque de cancer du sein dans la population d'étude. La régression logistique estime le risque $\log(OR_i) = \beta_i$ associé aux différentes catégories de chaque facteur, selon la formule suivante :

$$logit\big(P(M^+|X)\big) = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n$$

À partir de ce modèle, le risque β_i associé à chaque facteur est utilisé pour calculer un risque global β_j d'être atteint pour chaque sujet j. Seuls les β_i associés aux facteurs de la reproduction ou aux expositions aux radiations ont été utilisés pour calculer ce d'être atteint β_j global.

À partir de β_j, nous avons calculé un score de risque SR_j pour chaque sujet j grâce à la formule :

$$SR_j = \frac{\exp(\beta_j)}{1 + \exp(\beta_j)}$$

Les femmes de GENESIS ont ensuite été classées selon leur score de risque SR_j (d'un score de risque faible à élevé). Ces catégories ont ensuite été utilisées pour stratifier l'étude des facteurs génétiques.

6. Test d'hétérogénéité

Un test d'hétérogénéité a été réalisé pour chaque SNP trouvé associé au risque de cancer du sein pour l'un des groupes de score de risque. Cette étape m'a permis d'évaluer si les risques associés à chaque SNP étaient spécifiques à chaque groupe ou si les différences d'association observées étaient seulement dues à un manque de puissance statistique. En effet, pour réaliser les analyses génétiques, la population d'étude a été divisée en 3 groupes. Cette stratification a diminué drastiquement la taille des échantillons d'analyse et donc la puissance de détection des associations.

Pour tester l'hétérogénéité entre chaque groupe, j'ai réalisé un test de rapport de vraisemblance comparant les modèles de régression avec et sans interaction entre le SNP X et la catégorie de score de risque C_{SR} :

$$Modèle\ 1 : Y = X$$
$$Modèle\ 2 : Y = X + X * C_{SR}$$

Le seuil de signification a été fixé à $\alpha = 5\ \%$. Un test significatif met en évidence une hétérogénéité et donc une association spécifique avec la catégorie concernée.

7. Test de permutations

Afin de valider nos résultats, j'ai également effectué un test de permutations permettant de calculer la valeur exacte de la *p-value* du test d'hétérogénéité et ainsi de vérifier la spécificité des SNPs identifiés. Dans un test d'hypothèse classique comme ceux que nous avons faits précédemment, les hypothèses émises se basent sur des distributions théoriques. Les tests de permutations sont des approches robustes, basées sur l'aléatoire et le ré-échantillonnage. Leurs résultats ne se basent pas sur une distribution théorique mais sur une distribution empirique construite à partir de nos données. Pour cela, n jeux de données sont tirés au sort. Un test d'hétérogénéité est alors réalisé n fois. L'ensemble des statistiques de test t0 obtenues permet alors de former la distribution empirique. La *p-value* exacte (p_{exacte}) est ensuite estimée en calculant la proportion des statistiques de test qui se situent au-delà de la valeur absolue de la statistique de test t0 calculée sur les données observées. Elle correspond à l'aire sous la courbe au-delà des valeurs t0 et -t0.

En pratique, j'ai réalisé un tirage au sort pour répartir les sujets dans les 3 groupes de score de risque de façon aléatoire. Nous avons attribué aléatoirement un groupe à chaque femme à chaque tirage au sort selon le pourcentage observé de cas et de témoins dans chacun des groupes de risque. Ce tirage au sort, répété 10 000 fois, m'a permis de construire aléatoirement 10 000 répartitions différentes. J'ai ensuite effectué un test d'hétérogénéité pour les 10 000 répartitions pour chaque SNP (ANOVA). Les déviances générées ont alors été comparées à celle observée afin de calculer la p_{exacte} du test d'hétérogénéité pour chaque SNP.

Résultats

I. Description de la population

Les caractéristiques de la population sont décrites dans le Tableau 7. Les 2 972 femmes, dont 85 % ont été génotypées, se répartissent en 1 381 témoins et 1 591 cas.

Dans la population entière, l'âge moyen à l'inclusion est de 57,5 ans (de 19 à 90 ans). Bien qu'il était demandé aux cas index d'inviter des témoins ayant le même âge qu'elles +/- 3 ans (design de l'étude), on remarque que les témoins sont en moyenne plus jeunes que les cas à l'inclusion (55,8 vs 59,0 ans).

Par ailleurs, 93,6 % des cas sont des cas prévalents avec un délai moyen entre l'inclusion et le diagnostic de 9,1 ans (sd = 7,4). Les cas ont été censurés au diagnostic et les témoins à l'inclusion afin de prendre en considération uniquement les expositions ayant eu lieu avant le cancer du sein. L'âge moyen à la censure est de 49,8 ans pour les cas et de 55,8 ans pour les témoins.

Cette différence dans la répartition de l'âge à la censure entre les cas et les témoins persiste lorsque l'on regarde par année de naissance (Tableau 8). Par ailleurs, certaines classes d'âge sont très peu, voire pas du tout, représentées dans certaines classes d'année de naissance (Tableau 8).

Tableau 7 - Description de la population de l'étude GENESIS

Caractéristiques	Témoins (n = 1 381)		Cas (n = 1 591)		Total (n = 2 972)	
	Nb	Proportion	Nb	Proportion	Nb	Proportion
Génotypé avec la puce iCOGS						
Oui	1 272	0,93	1 262	0,80	2 534	0,85
Non	109	0,07	329	0,20	438	0,15
Âge à l'inclusion						
Moyenne	55,8		59,0		57,5	
Déviation standard	0,27		0,23		0,18	
Intervalle de confiance	55,3-56,3		58,5-59,4		57,1-57,8	
Min et Max	19-83		32-90		19-90	
Âge au diagnostic/censure						
Moyenne	55,8		49,8		52,6	
Déviation standard	0,27		0,23		0,23	
Intervalle de confiance	55,3-56,3		49,4-50,2		52,2-52,9	
Min et Max	19-83		20-80		19-83	
Délai entre le diagnostic et l'inclusion						
Moyenne			9,2			
Déviation standard			0,18			
Intervalle de confiance			9,0-9,3			
Min et Max			0-48			
Année de naissance						
< 1950	538	0,38	777	0,48	1 315	0,44
[1950-1960]	478	0,36	543	0,34	1 021	0,34
]1960-1970]	295	0,21	258	0,17	553	0,19
> 1970	70	0,05	13	0,01	83	0,03
Niveau d'éducation						
Élevé/Intermédiaire	924	0,67	799	0,50	1 723	0,58
Primaire	444	0,32	734	0,46	1 178	0,40
Pas d'étude	13	0,01	58	0,04	71	0,02
Âge aux premières règles						
< 12 ans	676	0,48	767	0,48	1 443	0,48
[12-15] ans	530	0,38	597	0,37	1 127	0,37
> 15 ans	166	0,13	213	0,13	379	0,12
Inconnu	9	0,01	14	0,008	23	0,007
Grossesses menées à terme						
0	188	0,14	188	0,12	376	0,13
1	172	0,12	273	0,17	445	0,15
2	599	0,43	669	0,42	1 268	0,42
≥ 3	420	0,30	460	0,29	880	0,29
Inconnu	2	0,01	1	0,00	3	0,01
Allaitement						
Jamais	610	0,44	733	0,46	1 343	0,45
≤ 10 mois	655	0,47	719	0,45	1 374	0,46
> 10 mois	103	0,08	115	0,07	218	0,08
Inconnu	13	0,01	24	0,02	37	0,01
Interruption induite de grossesse						
0	1 097	0,79	1 234	0,78	2 331	0,78
1	218	0,16	273	0,18	491	0,17
2	51	0,04	57	0,03	108	0,04
≥ 3	13	0,01	24	0,01	37	0,01
Inconnu	2	0,001	3	0,00	5	0,001

Caractéristiques	Témoins (n = 1 381)		Cas (n = 1 591)		Total (n = 2 972)	
	Nb	Proportion	Nb	Proportion	Nb	Proportion
Interruption spontanée de grossesse						
0	1 100	0,80	1 288	0,81	2 388	0,80
1	209	0,15	228	0,14	437	0,15
2	52	0,04	57	0,04	109	0,04
≥ 3	18	0,01	14	0,01	32	0,01
Inconnu	2	0,00	4	0,00	6	0,002
Contraceptifs hormonaux						
≤ 5 ans	247	0,18	331	0,21	578	0,20
]5-10] ans	190	0,14	223	0,14	413	0,14
> 10 ans	477	0,35	447	0,28	924	0,31
Jamais	449	0,32	571	0,36	1 020	0,34
Inconnu	18	0,01	19	0,01	37	0,01
Statut ménopausique						
Non ménopausée	488	0,35	969	0,61	1 457	0,49
Ménopausée	867	0,63	562	0,35	1 429	0,48
Inconnu	26	0,02	60	0,04	86	0,03
Indice de Masse Corporelle						
[18,5-25]	698	0,59	1 059	0,67	1 957	0,66
< 18,5	32	0,03	68	0,04	100	0,03
> 25	451	0,38	461	0,29	912	0,31
Inconnu	0	0,00	3	0,00	3	0,00
Traitements hormonaux substitutifs						
Jamais	923	0,67	1 272	0,80	2 195	0,73
≤ 4 ans	114	0,08	88	0,06	202	0,07
> 4 ans	261	0,19	145	0,09	406	0,14
Inconnu	83	0,06	86	0,05	196	0,06
Expositions thoraciques aux rayons X						
Jamais	239	0,17	208	0,13	447	0,15
Oui	1 104	0,80	1 299	0,82	2 403	0,81
Inconnu	38	0,03	40	0,05	122	0,04

Tableau 8 - Âge moyen à la censure selon l'année de naissance

Année de naissance	Témoins	Cas
≤ 1945	n = 301 68,2 (sd = 4,1) [20-83]	n = 500 57,1 (sd = 9,2) [28-67]
[1946-1959]	n = 715 57,0 (sd = 4,3) [47-67]	n = 820 48,6 (sd = 6,6) [28-64]
≥ 1960	n = 365 43,4 (sd = 5,8) [19-52]	n = 271 40,3 (sd = 4,6) [24-50]

Tableau 9 – Répartition de l'âge à la censure en classe (≥60 ans, entre 50 et 59 ans et < 50 ans) selon l'année de naissance.

Année de naissance	Statut	Âge à la censure		
		≥ 60	**[50-59]**	**< 50**
≤ 1945	Témoin	301	0	0
	Cas	232	160	108
[1946-1959]	Témoin	242	445	28
	Cas	26	350	444
≥ 1960	Témoins	0	29	336
	Cas	0	1	270

II. Facteurs non génétiques

Nous avons dans un premier temps étudié l'association entre les facteurs environnementaux (facteurs gynéco-obstétriques et expositions aux radiations) et le risque de cancer du sein à l'aide de différents modèles de régression logistique prenant en compte les facteurs confondants spécifiques à chaque variable.

1. Facteurs confondants

Les facteurs confondants pris en compte pour toutes les variables sont l'année de naissance, l'âge à la censure et le niveau d'étude (Tableau 10). Leur association est liée au design de l'étude GENESIS.

Tableau 10 - Risque de cancer du sein associé à l'année de naissance, l'âge à la censure et au niveau d'éducation.

	Témoins	**Cas**	**OR**	**$IC_{95\%}$**	**p-value**
Âge à la censure					
< 50 ans	364	822	1		
[50-59[ans	474	511	0,47	[0,40-0,57]	< 0,001
≥ 60 ans	543	258	0,21	[0,17-0,25]	< 0,001
Année de naissance					
≤ 1945	301	500	1		
]1946-1959]	715	820	0,69	[0,58-0,82]	< 0,001
> 1960	365	271	0,45	[0,36-0,55]	< 0,001

	Témoins	Cas	OR	$IC_{95\%}$	p-value
Niveau d'éducation					
Élevé/Intermédiaire	924	799	1		
Primaire	444	734	1,17	[1,13-1,22]	< 0,001
Pas d'études	13	58	1,42	[1,27-1,60]	< 0,001

Les variables étudiées sont fortement corrélées à l'âge, que ce soit celles relatives à la vie reproductive ou celles relatives aux expositions aux radiations. Pour prendre ce phénomène en compte, nous avons alors la possibilité d'ajuster sur la variable en continu ou sur celle discrétisée en 3 classes (≥ 60, entre 50 et 59 et < 50 ans). Cependant, 95 % des femmes nées après 1960 sont âgées de moins de 50 ans et aucune (cas et témoin confondus) n'est âgée de plus de 60 ans. De plus, aucun témoin né avant 1945 n'est âgé de moins de 60 ans. Cette répartition impose un ajustement sur l'âge à la censure en continu.

2. Facteurs gynéco-obstétriques

Aucune association n'a été trouvée avec le cancer du sein pour l'âge aux premières règles, la périodicité des cycles menstruel, l'âge à la première grossesse menée à terme, l'allaitement, les grossesses interrompues ni pour les traitements hormonaux substitutifs (Tableau 11). Seuls l'utilisation de contraceptifs hormonaux, le nombre de grossesses menées à terme, le statut ménopausique et l'IMC ont été trouvés associés au cancer du sein (Tableau 11).

La prise de contraceptifs hormonaux est associée à une diminution du risque de cancer du sein avec un OR égal à 0,57 ($IC_{95\%}$ = [0,44-0,73]). Cette diminution de risque semble d'autant plus importante que l'âge à la première utilisation est précoce, avec un OR égal à 1,96 ($IC_{95\%}$ = [1,54–2,51]) pour une première utilisation après l'âge de 20 ans comparé à une première utilisation avant l'âge de 20 ans.

Le fait d'avoir eu au moins une grossesse menée à terme est associé au risque de cancer du sein dans la population des femmes de GENESIS (OR = 1,34, p = 0,03). Par ailleurs, la première grossesse menée à terme est associée à une augmentation du risque (OR = 1,52, $IC_{95\%}$ = [1,05–2,51]) alors que le fait d'avoir eu plus d'une grossesse menée à terme n'est pas significatif.

La ménopause est associée à une diminution du risque de cancer du sein avec une estimation ponctuelle de RR égale à 0,63 ($IC_{95\%}$ = [0,46–0,85]). Bien que non significative, une ménopause tardive (après 55 ans) est associée à une estimation ponctuelle du risque de cancer du sein de 1,59 ($IC_{95\%}$ = [0,93–2,72]) et un âge jeune (avant 40 ans) à la ménopause à une estimation ponctuelle du risque de cancer du sein de 0,46 ($IC_{95\%}$ = [0,16–1,30]).

L'IMC est également un facteur de risque de cancer du sein dans notre population, avec une insuffisance pondérale (IMC < 18,5) qui est associée à une augmentation du risque (OR = 1,72, $IC_{95\%}$ = [1,01–2,72]).

Les analyses ont également été stratifiées par année de naissance pour vérifier que ces facteurs n'étaient pas impactés par un potentiel « effet cohorte ». Aucune hétérogénéité n'a été mise en évidence entre les classes d'âge (Annexe 6, page 277). Bien que les associations ne soient pas retrouvées de façon significative dans toutes les classes, les estimations ponctuelles des ORs vont dans le même sens et les intervalles de confiance obtenus dans les classes d'année de naissance contiennent celui estimé dans l'analyse globale. Les estimations obtenues dans l'analyse globale ont donc été utilisées par la suite.

Tableau 11 - Risque de cancer du sein associé aux facteurs gynéco-obstétriques

Facteurs gynéco-obstétriques	Témoins	Cas	OR*	$IC_{95\%}$	p-value
Âge aux premières règles					
[12-15] ans	676	767	1		
< 12 ans	530	597	0,94	[0,78-1,14]	0,55
> 15 ans	166	213	1,06	[0,79-1,40]	0,70
Inconnu	9	14	1,65	[0,58-4,63]	0,34
Période[1]					
[25-31] jours	871	963	1		
< 25 jours	114	111	0,94	[0,67-1,31]	0,70
> 31 jours	66	69	1,07	[0,70-1,62]	0,76
Irrégulières	251	335	1,10	[0,88-1,41]	0,37
Inconnu	79	108	1,03	[0,71-1,50]	0,86
Utilisation de contraceptifs hormonaux[2]					
Non	263	388	1		
Oui	998	1 077	0,57	[0,44-0,73]	**< 0,001**
Inconnu	120	126	0,59	[0,41-0,87]	**< 0,001**
Âge à la première utilisation[2]					
≤ 20 ans	464	402	1		
> 20 ans	527	664	1,96	[1,54-2,51]	**< 0,001**
Durée d'utilisation[2]					
≤ 5 ans	276	354	1		
> 5 ans	722	723	0,82	[0,64-1,04]	0,10
Grossesses menées à terme[3]					
Non	188	188	1		
Oui	1 191	1 402	1,34	[1,02-1,77]	**0,03**
Inconnu	2	1	NA	NA	NA
Nombre de grossesses menées à terme[3]					
0	188	188	1		
1	172	273	1,68	[1,18-2,40]	**0,004**
2	599	669	1,29	[0,96-1,74]	0,09
≥ 3	420	460	1,27	[0,93-1,73]	0,12
Âge à la première grossesse menée à terme[3]					
≤ 20 ans	183	283	1		
]20-25] ans	562	634	0,98	[0,74-1,29]	0,89
]25-30] ans	334	344	0,93	[0,68-1,29]	0,69
> 30 ans	112	139	1,43	[0,96-2,13]	0,08

*Analyses ajustées sur l'âge à la censure en continu, l'année de naissance (≤ 1945, [1946-1959], ≥ 1960) et le niveau d'éducation (élevé/intermédiaire, primaire et pas d'études).

1. * + la prise de contraceptifs hormonaux (oui/non).

2. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans), le statut ménopausique (oui/non) et le nombre de grossesses menées à terme (0, 1, 2, ≥ 3).

3. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans), le statut ménopausique (oui/non) et la prise de contraceptifs hormonaux (oui/non).

Facteurs gynéco-obstétriques	Témoins	Cas	OR*	$IC_{95\%}$	p-value
Allaitement[4]					
Non	609	733	1		
Oui	764	845	0,90	[0,73-1,11]	0,34
Inconnu	8	13	3,69	[1,09-12,5]	0,03
Durée d'allaitement[4]					
Jamais	610	733	1		
≤ 10 mois	655	719	0,89	[0,72-1,10]	0,28
> 10 mois	103	115	1,02	[0,69-1,49]	0,92
Inconnu	13	24	2,08	[0,83-5,18]	0,11
Grossesses interrompues[5]					
Non	874	1 003	1		
Oui	505	585	1,09	[0,90-1,32]	0,58
Inconnu	2	3	NA	NA	NA
Nombre de grossesses interrompues[5]					
0	874	1 003	1		
1	330	392	1,13	[0,91-1,41]	0,25
2	133	139	1,03	[0,75-1,41]	0,87
≥ 3	42	54	0,95	[0,56-1,61]	0,86
Statut ménopausique[6]					
Non	488	969	1		
Oui	867	562	0,63	[0,46-0,85]	**0,002**
Inconnu	26	60	1,26	[0,71-2,24]	0,43
Type de ménopause[6]					
Naturelle	823	525	1		
Ovariectomie	43	37	0,99	[0,50-1,68]	0,99
Âge à la ménopause[6]					
[40-55] ans	700	466	1		
< 40 ans	15	10	0,46	[0,16-1,30]	0,14
> 55 ans	45	34	1,59	[0,93-2,72]	0,08
Inconnu	133	112	1,30	[0,93-1,81]	0,12
Traitements hormonaux substitutifs[7]					
Non	929	1 294	1		
Oui	391	230	0,92	[0,71-1,20]	0,54
Inconnu	61	67	1,42	[0,90-2,26]	0,13
IMC[8]					
[18,5-25]	898	1 059	1		
< 18,5	32	68	1,72	[1,01-2,92]	**0,04**
>25	451	461	0,99	[0,81-1,21]	0,92
Inconnu	0	3	NA	NA	NA

4. * + 3 + le nombre de grossesses menées à terme (0, 1, 2, ≥ 3).
5. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans), le statut ménopausique (oui/non), le nombre de grossesses menées à terme (0, 1, 2, ≥ 3) et l'âge à la première grossesse (< 20 ans, [20-25[, [25-30[, ≥ 30 ans, nullipare).
6. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans), et le nombre de grossesses menées à terme (0, 1, 2, ≥ 3).
7. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans), statut ménopausique (oui/non) et la prise de contraceptifs hormonaux (oui/non).
8. * + le statut ménopausique (oui/non).

Pour construire le modèle final, j'ai donc retenu l'utilisation de contraceptifs hormonaux, le nombre de grossesses menées à terme et l'IMC, trouvés associés au cancer du sein. En ce qui concerne la ménopause, seul le fait d'être ménopausée est associé de façon significative avec le cancer du sein (OR = 0,63, $IC_{95\%}$ = [0,46-0,85]) comparé aux femmes non ménopausées. Cependant, compte-tenu des estimations ponctuelles associées à l'âge à la ménopause, j'ai décidé de prendre l'âge à la ménopause dans le modèle final. Ce modèle a été ajusté sur l'âge à la censure en continu, l'année de naissance en classe (≤ 1945, [1946-1959] et ≥ 1960), le niveau d'éducation (élevé/intermédiaire, primaire et pas d'études), l'âge aux premières règles ([12-15], < 12 ans, > 15 ans) et l'âge à la première grossesse menée à terme (< 20 ans, [20-25[, [25-30[, ≥ 30ans, nullipare). Les résultats avant et après imputation multiple des données manquantes sont présentés dans le Tableau 12.

Tableau 12 – Facteurs gynéco-obstétriques retenus dans le modèle complet : risque de cancer du sein associé avant et après imputation multiple.

	Avant imputation multiple				**Après imputation multiple**			
	β	**OR**	**$IC_{95\%}$**	**p-value**	**β**	**OR**	**$IC_{95\%}$**	**p-value**
Utilisation de contraceptifs hormonaux*								
Non	0	1			0	1		
Oui	-0,5561	0,57	[0,46-0,73]	< 0,001	-0,5349	0,58	[1,45-0,75]	< 0,001
Inconnue	-0,5232	0,59	[0,40-0,86]	0,007				
Grossesses menées à terme*								
0	0	1			0	1		
1	0,4847	1,62	[1,04-2,52]	0,03	0,5235	1,68	[1,18-2,41]	0,004
2	0,2626	1,30	[0,89-1,89]	0,17	0,2798	1,32	[0,97-1,80]	0,07
≥ 3	0,2735	1,31	[0,91-1,91]	0,15	0,2732	1,31	[0,94-1,82]	0,10
Inconnu	NA	NA	NA	NA				
Âge à la ménopause*								
[40-55] ans	0	1			0	1		
< 40 ans	-0,7617	0,46	[0,16-1,30]	0,14	-0,5315	0,58	[0,20-1,65]	0,31
> 55 ans	0,4032	1,49	[0,87-2,57]	0,14	0,2940	1,34	[0,80-2,23]	0,26
Non ménopausée	0,3893	1,47	[1,09-2,98]	0,01	0,4306	1,53	[1,13-2,07]	0,005
Inconnu	0,2382	1,27	[0,97-1,77]	0,16				
IMC*								
[18,5-25]	0	1			0	1		
< 18,5	0,4959	1,64	[0,96-2,80]	0,06	0,5006	1,64	[0,96-2,81]	0,06
> 25	-0,0163	0,98	[0,80-1,20]	0,96	-0,0083	0,99	[0,81-1,21]	0,93
Inconnu	NA	NA	NA	NA				

*Ajusté sur l'âge à la censure en continu, l'année de naissance en 3 classes (≤ 1945, [1946-1959] , ≥ 1960), le niveau d'éducation (élevé/intermédiaire, primaire et pas d'études), l'âge aux premières règles ([12-15], < 12 ans, > 15 ans) et l'âge à la première grossesse (< 20 ans, [20-25[, [25-30[, ≥ 30 ans, nullipare).

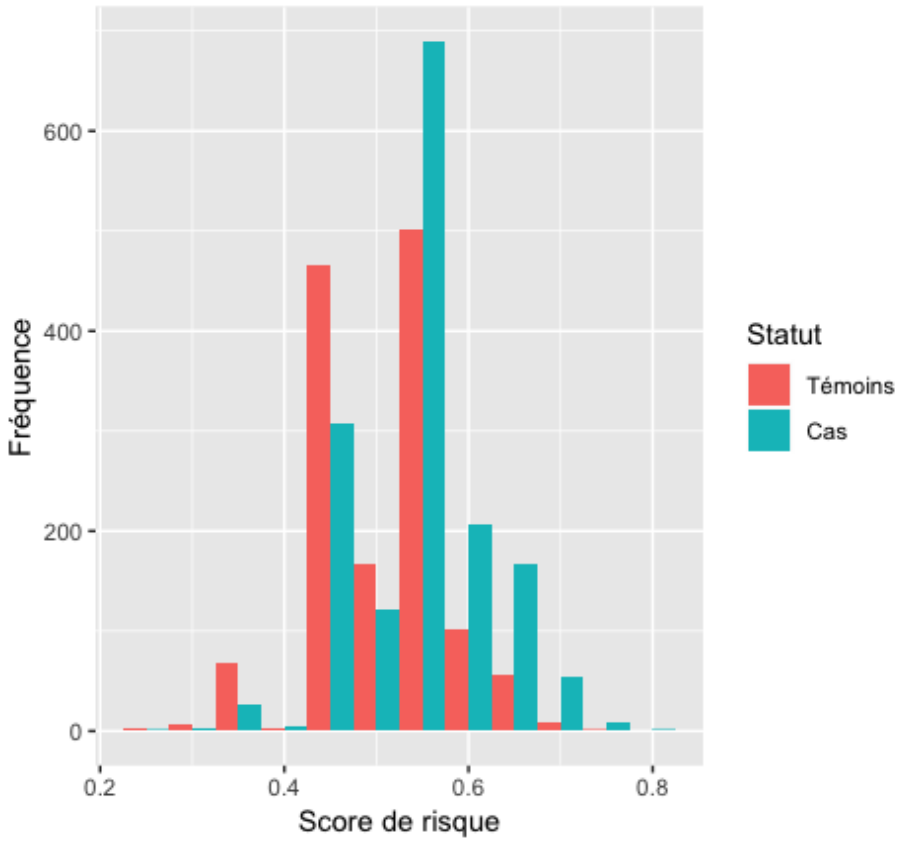
Pour mémoire, le but est de stratifier les analyses génétiques sur des groupes de scores de risque construits à partir des variables gynéco-obstétriques. Les estimations obtenues pour les 4 variables du modèle complet ont été utilisées pour calculer un score de risque à partir duquel les groupes seront définis. Pour chaque sujet j, un β_j global est alors calculé en additionnant les β_x associés à chaque variable X :

$$\beta_j = \beta_{contra}*X_{contra} + \beta_{nbGMT}*X_{nbGMT} + \beta_{âgeménop}*\beta_{âgeménop} + \beta_{IMC}*X_{IMC}$$

Un score de risque SR_j est calculé à partir de ce β_j grâce à la formule :

$$SR_j = \frac{e^{\beta i}}{1 + e^{\beta i}}$$

Figure 9 - Répartition des scores de risque définis à partir des facteurs gynéco-obstétriques



La moyenne des scores de risque définis à partir des facteurs gynéco-obstétriques est de 0,50 [0,25-0,74] chez les témoins et de 0,54 [0,25-0,81] chez les cas (Figure 9). J'ai utilisé ce score de risque pour classer les femmes en 3 groupes, correspondant à un score de risque « faible », « moyen » et « élevé » de cancer du sein à partir des facteurs gynéco-obstétriques. Pour cela j'ai discrétisé le score de risque en 3 classes grâce à la méthode des quantiles qui vise une équipartition des sujets par classe. J'ai utilisé la population entière pour définir ces classes de score de risque (faible : score ≤ 0,55 ; moyen : 0,55 < score < 0,61 ; élevé : score ≥ 0,61) afin de créer des groupes où les cas et les témoins sont appariés sur leur score.

Tableau 13 - Répartition des cas et des témoins selon le score de risque établi à partir des facteurs gynéco-obstétriques

Statut	**Score de risque**		
	Faible	**Moyen**	**Élevé**
Témoin	589 (43 %)	312 (23 %)	480 (35 %)
Cas	380 (24 %)	351 (23 %)	860 (54 %)

Le groupe à faible score de risque contient presque 2 fois plus de témoins (43 %) que de cas (23 %) alors que la répartition cas-témoin s'inverse dans le groupe à score de risque élevé (Tableau 13).

3. Expositions aux radiations

Les variables liées aux expositions aux radiations ionisantes au thorax ont été générées et analysées par Maximilianno Guerra (Tableau 14).

Ces analyses montrent que le nombre d'expositions aux radiations est fortement associé au risque de cancer du sein avec un OR égal à 2,05 ($IC_{95\%}$ = [1,55–2,73]). Le risque augmente de façon significative avec le nombre d'expositions aux radiations avec un OR de 1,70, 2,52 et 2,37 associés respectivement à 1-3, 4-9 et ≥ 10 expositions thoraciques (Tableau 14). L'OR augmente de 1,03 à chaque exposition supplémentaire ($IC_{95\%}$ = [1,01-1,04]).

Tableau 14 – Effet de l'exposition aux radiations au thorax au cours de la vie sur le risque de cancer du sein en fonction du nombre d'expositions, de l'âge à la première exposition et de la première grossesse menée à terme.

Caractéristiques	Nombre		OR	$IC_{95\%}$	p-value
	Cas	Témoins			
Exposition aux radiations ionisantes*					
0	208	239	1		
≥ 1	1 304	1 104	2,05	1,55-2,73	< 0,001
Inconnu	40	20	2,65	1,26-5,57	0,01
Nombre d'expositions*					
0	208	239	1		
1-3	392	390	1,70	1,23-2,34	0,001
4-9	251	200	2,52	1,76-3,61	< 0,001
≥ 10	263	215	2,37	1,64-3,43	< 0,001
Nombre inconnu	438	319	2,13	1,54-2,96	< 0,001
Âge à la première exposition (années)*					
≥ 20	485	490	1		
15-19	288	222	1,07	0,80-1,44	0,63
< 15	290	219	1,18	0,88-1,58	0,27
Âge inconnu	281	193	1,17	0,86-1,58	0,31
Selon la première grossesse menée à terme (GMT)*					
Après la 1re GMT	268	232	1		
Avant la 1re GMT	825	725	0,86	0,65-1,14	0,29
Inconnu	251	167	1,01	0,71-1,44	0,97
Nombre d'expositions selon la première GMT*					
Après 1re GMT et ≤ 5	186	178	1		
Après 1re et > 5	56	43	1,77	0,98-3,18	0,06
Avant 1re et ≤ 5	442	412	0,95	0,67-1,36	0,78
Avant 1re et > 5	272	215	1,12	0,77-1,61	0,56
Inconnu	388	276	1,08	0,76-1,52	0,67

*Ajusté sur l'âge à la censure (en continu), l'année de naissance (≤ 1945, [1946-1959], ≥ 1960), le niveau d'éducation (élevé, primaire, pas d'étude), l'exposition à au moins une mammographie (oui/non), l'IMC (< 18,5, [18,5-25], > 25), la consommation de tabac (oui/non) et les antécédents familiaux de cancers du sein (0 "aucun ou seulement une sœur pour les cas", 1 "apparentée(s) au 1er degré", 2 "apparentée(s) au 2nd degré").

Bien que non significatif, nous avons décidé de prendre également en compte l'âge à la première exposition aux radiations. En effet, la littérature a montré que les effets des radiations ionisantes étaient plus importants quand l'exposition avait eu lieu pendant la puberté c'est-à-dire pendant le développement des glandes mammaires composées de cellules encore immatures[209]. Une exposition avant l'âge de 15 ans est associée à une augmentation de 18 % du risque de cancer du sein dans la population de GENESIS (Tableau 14).

Le modèle complet contient donc le nombre d'expositions et l'âge à la première exposition ionisante et a été ajusté sur l'âge à la censure (en continu), l'année de naissance (≤ 1945, 1946-1959, ≥ 1960), le niveau d'éducation (élevé, primaire, pas d'étude), l'exposition à au moins une mammographie (oui/non), l'IMC (< 18,5, [18,5-25], > 25), la consommation de

tabac (oui/non) et les antécédents familiaux de cancers du sein (0 : aucun ou seule une sœur pour les cas, 1 : au 1er degré, 2 : au 2nd degré).

Tableau 15 – Facteurs associés aux expositions aux radiations au thorax retenus dans le modèle complet : risque de cancer du sein associé avant et après imputation multiple.

	Avant imputation				**Après imputation**			
	β	**OR**	**$IC_{95\%}$**	***p-value***	**β**	**OR**	**$IC_{95\%}$**	***p-value***
Nombre d'expositions aux radiations à la poitrine*								
0	0	1			0	1		
1	0,5145	1,67	[1,20-2,33]	0,002	0,4753	1,60	[1,15-2,23]	0,005
2	0,9121	2,49	[1,69-3,66]	$< 10^{-6}$	0,8146	2,25	[1,59-3,18]	$< 10^{-6}$
≥ 3	0,8393	2,31	[1,54-3,47]	$< 10^{-6}$	0,9828	2,67	[1,79-3,97]	$< 10^{-6}$
Inconnu	0,6462	1,91	[1,22-2,97]	0,004				
Âge à la première exposition*								
≥ 20	0	1			0	1		
[15-19]	-0,0230	0,97	[0,71-1,32]	0,86	-0,0143	0,98	[0,73-1,32]	0,92
< 15	0,1019	1,10	[0,82-1,49]	0,50	0,0744	1,07	[0,80-1,45]	0,62
Inconnu	0,1691	1,18	[0,77-1,82]	0,44				

*Ajusté sur l'âge à la censure en continu, l'année de naissance (≤ 1945, [1946-1959], ≥ 1960), le niveau d'éducation (élevé, primaire, pas d'étude), l'exposition à au moins une mammographie (oui/non), l'IMC (< 18,5, [18,5-25], > 25), la consommation de tabac (oui/non) et les antécédents familiaux de cancer du sein (0 "aucun ou seulement une sœur pour les cas", 1 "apparentée(s) au 1er degré", 2 "apparentée(s) au 2nd degré").

Comme pour les facteurs gynéco-obstétriques, les estimations obtenues pour ces 2 variables ont été utilisées pour calculer un score de risque pour chaque femme selon l'exposition aux radiations (Figure 10). Ce score de risque se répartit de 0,50 à 0,74 dans la population de GENESIS, avec une moyenne de 0,64 pour les témoins et 0,65 pour les cas (Figure 10), 447 sujets (239 témoins et 208 cas) ont un score de 0,5, correspondant au score des sujets qui n'ont jamais été exposés aux radiations ionisantes. J'ai choisi de former 3 groupes en considérant que le 1er groupe serait le groupe des sujets non exposés. Le reste de la population a été divisé en deux groupes (Tableau 16) : Score de risque moyen : 0,5 < score < 0,69 et score de risque élevé : ≥ 0,69.

Figure 10 - Répartition des scores de risque en fonction des expositions aux radiations au thorax

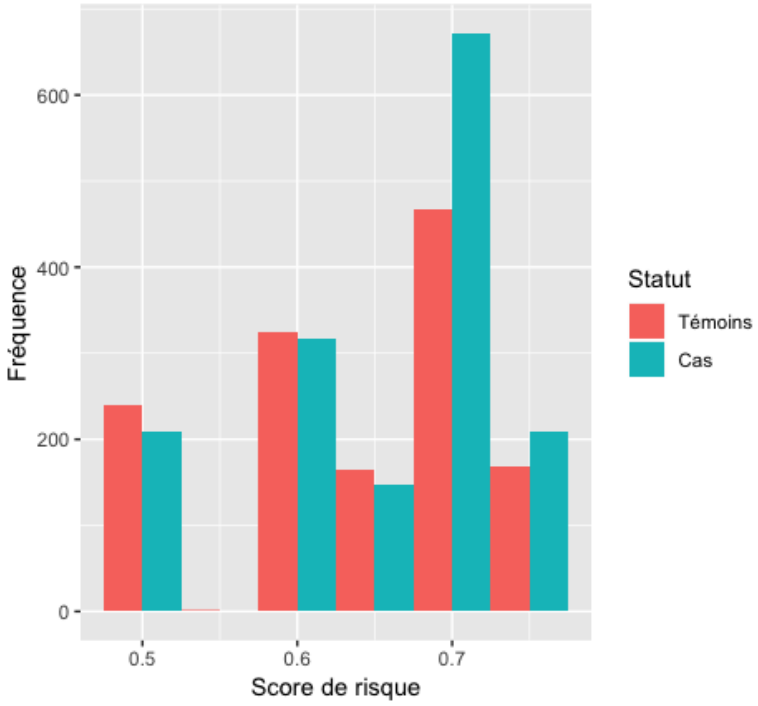


Tableau 16 - Répartition des cas et des témoins selon le score de risque établi à partir des expositions aux radiations au thorax

	Score de risque		
Statut	**Non exposé**	**Moyen**	**Élevé**
Témoin	239 (17 %)	621 (45 %)	503 (38 %)
Cas	208 (13 %)	653 (42 %)	691 (44 %)

III. Facteurs génotypiques

En parallèle de l'étude des facteurs environnementaux, j'ai étudié les SNPs. J'ai tout d'abord imputé les SNPs non génotypés.

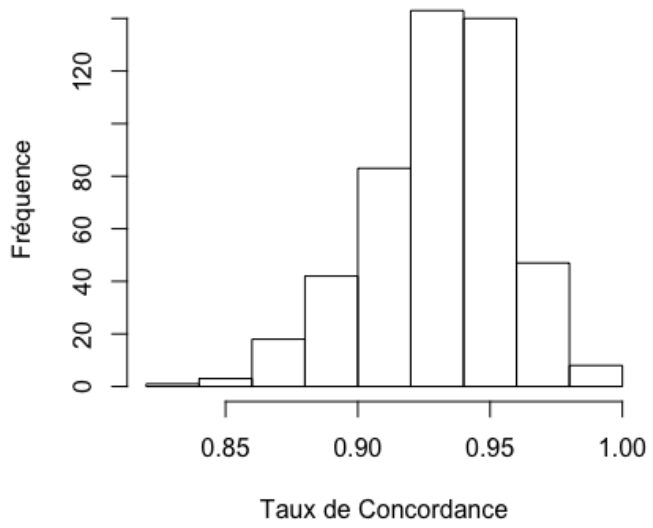
1. Imputation des SNPs non génotypés

a. Validation croisée

487 fragments de 5 Mb couvrant l'ensemble du génome ont été imputés. Afin d'évaluer la qualité d'imputation de chacun de ces fragments, j'ai utilisé la table de concordance générée par IMPUTE2. Cette table résume les résultats de la validation croisée effectuée par IMPUTE2 sur les SNPs génotypés.

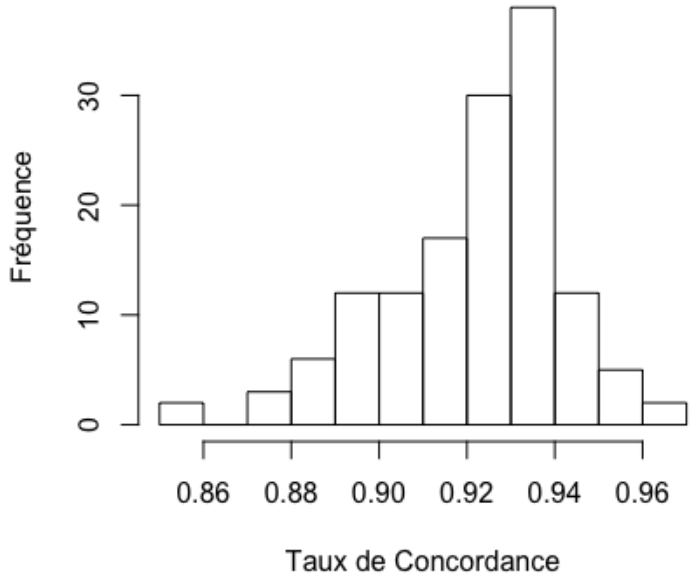
358 des fragments imputés présentent un pourcentage de concordance, entre la valeur génotypée et celle imputée, inférieur à 0,95 (Figure 11). Le taux de concordance moyen sur l'ensemble des fragments est de 0,93 et le fragment le moins bien imputé a un taux de concordance de 0,83.

Figure 11 - Distribution des taux de concordance des 487 fragments imputés



J'ai ré-imputé ces 358 régions en augmentant la taille du fragment de 5 à 10 Mb grâce à l'option *allow_large_region* proposée par IMPUTE2 avec l'hypothèse qu'en augmentant l'information utilisée pour l'imputation, sa qualité augmenterait également. Cependant, sur les 139 nouveaux fragments imputés, seuls 8 ont un taux de concordance supérieur ou égal à 0,95 (Figure 12). Ces 8 fragments étaient localisés à côté de régions imputées avec un taux de concordance supérieur ou égal à 0,95, ce qui explique l'augmentation obtenue après ré-imputation.

Figure 12 - Distribution des taux de concordance des 139 fragments ré-imputés



Cette ré-imputation sur des fragments de plus grande taille n'améliore donc pas la qualité d'imputation. Pour la suite des analyses, j'ai donc utilisé les résultats de l'imputation initiale. La faible qualité d'imputation globale de certains fragments devra donc être prise en compte lors de l'interprétation des résultats.

b. Contrôle qualité des SNPs imputés

Alors que le nombre de SNPs génotypés est d'environ 200 000, l'imputation permet d'augmenter le nombre de SNPs à 91 108 652. Le contrôle qualité réalisé à partir de la validation croisée nous permet de vérifier la qualité d'imputation globale d'un fragment mais pas celle de chaque SNP individuellement. Après vérification de l'imputation de chaque SNP,

j'ai exclu ceux dont la qualité d'imputation n'était pas suffisante en fixant comme valeur seuil $r^2 > 0,5$. 65 639 580 SNPs imputés ont un r^2 inférieur à 0,5.
J'ai également filtré les SNPs en fonction de leur fréquence. En effet, la taille modeste de notre population d'étude ne permet pas d'avoir une puissance suffisante pour mettre en évidence des SNPs peu fréquents, et les résultats positifs obtenus auraient une forte probabilité d'être des artefacts. J'ai alors choisi d'exclure tous les SNPs avec une MAF inférieure à 0,05, ce qui représente 83 283 550 SNPs.
Au final, les filtres appliqués pour garantir une bonne qualité d'imputation ($r^2 > 0,5$) et une fréquence suffisante (MAF > 0,05) a conduit à l'exclusion de 83 631 468 SNPs, soit 91,7 % des SNPs obtenus après imputation.

2. Analyse des SNPs

Dans le cadre de ma thèse, seuls les SNPs associés aux gènes intervenant dans les voies biologiques d'intérêt ont été analysés.
Parmi les SNPs qui ont passé le contrôle qualité, 133 305 SNPs se trouvent dans un gène intervenant dans la régulation des hormones sexuelles et 97 896 dans ceux intervenant dans la réparation de l'ADN. 144 gènes sont communs aux 2 voies biologiques d'intérêt, ce qui représente 22 723 SNPs. Les analyses ont donc été faites sur 208 478 SNPs (9 918 génotypés et 198 560 imputés), dans un premier temps sur la population totale de GENESIS puis par sous-groupe définis dans les Tableau 13 et Tableau 16.

a. Analyses non stratifiées

L'analyse sur la population totale de GENESIS a identifié 54 SNPs associés au risque de cancer du sein, répartis dans 4 régions du génome au seuil de Bonferroni de 10^{-4} (Tableau 17) : 1 SNP localisé dans le locus 5q14.2 sur le gène *XRCC4,* 8 SNPs en 6p11.2 sur le gène *PRIM2,* 6 SNPs en 7p11.2 sur le gène *EGFR* et 39 SNPs en 10q26.13 sur le gène *FGFR2*. L'analyse step-wise montre que ces 4 régions sont chacune représentée par un SNP indépendant : rs9293329, rs11396652, rs1253848 et rs35054928 respectivement. La région du gène *FGRF2* contient 39 SNPs significatifs au seuil $p < 1,05.10^{-4}$ (Tableau supplémentaire 1).

Tableau 17 – Top SNPs associés au risque de cancer du sein.

SNP	Chr	Locus	Gene	Localisation	Position	A1	A2	Fréquence A1	r^2	OR	P
rs9293329	5	5q14.2	XRCC4	intron	82396587	A	G	0,07	1	1,54	$4,38.10^{-5}$
rs11396652	6	6p11.2	PRIM2	intron	57466350	TA	T	0,57	0,88	0,76	$2,81.10^{-5}$
rs12538489	7	7p11.2	EGFR	intron	55099836	T	C	0,13	0,95	0,67	$2,79.10^{-5}$
rs35054928	10	10q26.13	FGFR2	intron	123340431	G	GC	0,56	0,99	0,73	$1,53.10^{-7}$

b. Analyses stratifiées sur les groupes de score de risque

Les SNPs spécifiques candidats ont été sélectionnés dans les analyses stratifiées par groupe de risque s'ils étaient associés au cancer du sein au seuil $p < 10^{-4}$ dans un des 3 groupes et si les tests d'hétérogénéité étaient significatifs au seuil $\alpha = 0,05$.

- *Expositions aux radiations*

Les analyses stratifiées sur les groupes de score de risque construits à partir des expositions aux radiations mettent en évidence 359 SNPs présents dans les gènes intervenant dans la réparation de l'ADN qui sont associés de façon significative avec le risque de cancer du sein : 326 SNPs dans le groupe des femmes non exposées aux radiations ionisantes, 4 SNPs pour celles associées à un score de risque moyen et 29 pour celles associées à un score de risque élevé (Tableaux supplémentaires 3, 5 et 7). Tous ces SNPs montrent une hétérogénéité significative au seuil $p < 0,05$ et pour les groupes à score de risque moyen et élevé, la p_{exacte} est également inférieure à 0,05 (Tableaux supplémentaires 4 et 6). Cependant, parmi les 326 SNPs significatifs dans le groupe des sujets non exposés, seuls 26 sont associés à une p_{exacte} respectant le seuil de 0,05 (Tableau supplémentaire 2).

Dans le groupe des femmes non exposées aux radiations, 26 SNPs significatifs respectent les seuils fixés. Ils sont localisés dans 4 chromosomes différents : 2 SNPs en 2q36, 9 SNPs en 3p14.1, 14 SNPs en 5q14.2 et 1 SNP en 14q23.1. Ces 4 régions sont chacune expliquée par un *top SNP* indépendant (Tableau 18). On remarque que pour les *top SNPs* localisés en 2q36 et 3p14.1 dans les gènes *COL4A4* et *MAGI1* respectivement, l'analyse sur la population entière montre une association suggérée, avec une *p-value* inférieure à 10^{-3}. Cependant, ces SNPs ne

sont pas du tout significatifs dans les groupes avec un score de risque moyen et élevé (Tableau 19). La région du gène *MAGI1* est représentée par 9 SNPs associés de façon significative, avec une p_{exacte} inférieure à 5 %. Pour les 2 autres *top SNPs,* localisés en 5q14.2 et 14q23.1, l'association n'est pas retrouvée dans l'analyse globale ni dans les autres groupes. 313 SNPs localisés en 5q14.2 dans le gène *XRCC4* (Tableau supplémentaire 2) sont associés de façon significative mais seuls 24 montrent une p_{exacte} inférieure à 5 %. Tous les SNPs trouvés significatifs dans cette région ne sont pas associés dans l'analyse globale (Tableau supplémentaire 3).

Dans le groupe de femmes associées à un score de risque moyen, 4 SNPs – dont 1 localisé en 2q31.1, 2 en 2q34 et 1 en 12p13.33 – sont significativement associés au cancer du sein (Tableau 18 et Tableau supplémentaire 4). Le SNP isolé localisé en 2q31.1 est associé de façon significative au cancer du sein seulement dans ce groupe alors que les SNPs localisés dans les deux autres régions suggèrent une association dans l'analyse sur la population entière, avec des estimations ponctuelles qui vont dans le même sens (Tableau 19). La région 2q34, et plus précisément le gène *ERBB4,* est particulière car certains SNPs de ce gène sont associés au cancer du sein dans le groupe des femmes avec un score de risque élevé (rs2076818) mais pas dans les autres groupes alors que d'autres SNP de ce même gène (par exemple, rs67127866) sont associés au cancer du sein dans le groupe des femmes à score de risque moyen et pas dans les autres groupes (Tableau 19 et Tableau supplémentaire 5).

27 autres SNPs sont trouvés associés de façon significative chez des femmes ayant un score de risque élevé (Tableau 19 et Tableaux supplémentaires 6 et 7). Ils sont associés au gène *FGFR2* en 10q26.13 qui est également trouvé associé dans l'analyse sur la population entière (Tableau 17).

Tableau 18 - Top SNPs associés au cancer du sein dans chacun des groupes de scores de risque relatifs aux expositions aux radiations au thorax

Score de risque	GENESIS															BCAC[6]		BRCA1[7]		BRCA2[8]	
	SNP	Chro	Locus	Gène	Localisation	Position[1]	Allèle A1	Allèle A2	Freq A1[2]	r^2 [3]	OR	$IC_{95\%}$	P	P_{het} [4]	P_{exacte} [5]	*OR*	*P*	*HR*	*P*	*HR*	*P*
Non exposés	rs6436654	2	2q36.3	*COL4A4*	intronique	227974113	G	A	0,62	1,00	2,05	[1,45-2,91]	$2,57.10^{-5}$	**$1,12.10^{-8}$**	**$3,60.10^{-3}$**	*1,00*	*0,61*	*1,04*	*0,05*	*0,96*	*0,17*
	rs57745762	3	3p14.1	*MAGI1*	intronique	65972545	A	AT	0,68	0,77	0,45	[0,31-0,66]	$1,99.10^{-5}$	**$8,89.10^{-17}$**	**0**	*1,00*	*0,89*	*1,01*	*0,75*	*0,97*	*0,24*
	rs201890201	5	5q14.2	*XRCC4*	intronique	82510669	AAT	A	0,63	0,98	0,54	[0,39-0,73]	$6,33.10^{-5}$	**$4,44.10^{-15}$**	0	*0,99*	*0,52*	*1,02*	*0,44*	*0,96*	*0,10*
	rs11296750	14	14q23.1	*MNAT1*	Intronique	61213232	C	CA	0,30	0,68	0,38	[0,24-0,62]	$4,12.10^{-5}$	**$6,05.10^{-13}$**	**0**	*0,97*	*0,012*	*0,99*	*0,71*	*1,00*	*0,76*
Risque moyen	rs74986298	2	2q31.1	*ITGA6*	intronique	173335568	T	C	0,06	0,96	0,40	[0,25-0,62]	$3,00.10^{-5}$	**$3,76.10^{-7}$**	**$1,00.10^{-4}$**	*1,01*	*0,72*	*0,97*	*0,48*	*0,97*	*0,62*
	rs67127866	2	2q34	*ERBB4*	intronique	212500825	C	CGT	0,91	0,56	0,43	[0,29-0,64]	$2,17.10^{-5}$	**$4,85.10^{-14}$**	**0**	*0,99*	*0,67*	*0,96*	*0,24*	*1,00*	*0,99*
	rs56303980	12	12p13.33	*RAD52*	intronique	1042625	GAA	G	0,39	0,98	1,44	[1,20-1,72]	$8,18.10^{-5}$	**$1,03.10^{-5}$**	**$4,51.10^{-2}$**	*1,01*	*0,11*	*1,03*	*0,19*	*0,99*	*0,84*
Risque élevé	rs2076818	2	2q34	*ERBB4*	intronique	212000000	T	C	0,90	0,78	2,59	[1,68-3,99]	$1,03.10^{-5}$	**$5,24.10^{-16}$**	**0**	*0,96*	*0,03*	*1,03*	*0,48*	*1,04*	*0,48*
	rs1078806	10	10q26.13	*FGFR2*	intronique	123338975	G	A	0,41	1	1,56	[1,29-1,89]	$3,95.10^{-6}$	**$8,64.10^{-11}$**	**0**	*1,27*	*$1,80.10^{-149}$*	*0,98*	*0,49*	*1,22*	*$3,92.10^{-15}$*

1. Position sur la version hg19 du génome
2. Fréquence de l'allèle A1 chez les témoins de GENESIS
3. Qualité de l'imputation
4. P-value du test d'hétérogénéité entre les 3 groupes de score de risque
5. P-value exacte d'hétérogénéité estimée
6. OR et p-value obtenus dans la population générale (consortium BCAC[210])
7. OR et p-value obtenus chez les porteurs de mutation BRCA1 (consortium CIMBA[156])
8. OR et p-value obtenus chez les porteurs de mutation BRCA2 (consortium CIMBA[156])

Tableau 19 - Top SNPs associés au cancer du sein dans chacun des groupes de scores de risque relatifs aux expositions aux radiations – résultats dans chaque groupe

Groupe où les associations sont significatives[1]	Description				Analyse population totale[2]			Non exposé[3]			Score de risque moyen[4]			Score de risque élevé[5]		
	SNP	Chr	Fréquence A1	r^2	OR	$IC_{95\%}$	P	OR	$IC_{95\%}$	P	OR	$IC_{95\%}$	P	OR	$IC_{95\%}$	P
Non exposés	rs6436654	2	0,62	1,00	1,22	[1,08-1,38]	0,001	**2,05**	**[1,45-2,91]**	**$2,57.10^{-5}$**	1,09	[0,91-1,31]	0,34	1,18	[0,97-1,44]	0,10
	rs57745762	3	0,68	0,77	0,83	[0,72-0,95]	0,008	**0,45**	**[0,31-0,66]**	**$1,99.10^{-5}$**	0,85	[0,69-1,05]	0,13	1,03	[0,82-1,30]	0,78
	rs201890201	5	0,63	0,98	0,95	[0,85-1,08]	0,45	**0,54**	**[0,39-0,73]**	**$6,33.10^{-5}$**	1,05	[0,88-1,26]	0,58	1,03	[0,84-1,26]	0,76
	rs11296750	14	0,30	0,68	0,85	[0,72-1,00]	0,06	**0,38**	**[0,24-0,62]**	**$4,12.10^{-5}$**	0,93	[0,73-1,19]	0,59	0,98	[0,75-1,29]	0,94
Risque moyen	rs74986298	2	0,06	0,96	0,81	[0,63-1,04]	0,10	0,72	[0,39-1,35]	0,32	**0,40**	**[0,25-0,62]**	**$3,00.10^{-5}$**	1,34	[0,91-1,98]	0,13
	rs67127866	2	0,91	0,56	0,71	[0,55-0,92]	0,009	1,21	[0,62-2,37]	0,57	**0,43**	**[0,29-0,64]**	**$2,17.10^{-5}$**	1,02	[0,67-1,54]	0,93
	rs56303980	12	0,39	0,98	1,15	[1,03-1,30]	0,002	1,12	[0,83-1,51]	0,45	**1,44**	**[1,20-1,72]**	**$8,18.10^{-5}$**	0,89	[0,73-1,08]	0,21
Risque élevé	rs2076818	2	0,90	0,78	1,23	[0,95-1,59]	0,12	0,68	[0,34-1,35]	0,27	0,85	[0,57-1,27]	0,42	**2,59**	**[1,68-3,99]**	**$1,03.10^{-5}$**
	rs1078806	10	0,41	1,00	1,34	[1,20-1,51]	$5,72.10^{-7}$	1,40	[1,03-1,90]	0,33	1,18	[0,99-1,41]	0,06	**1,57**	**[1,29-1,90]**	**$3,95.10^{-6}$**

1. Qualité de l'imputation
2. Résultats obtenus dans la population totale de GENESIS
3. Résultats obtenus dans le groupe non exposé aux radiations
4. Résultats obtenus dans le groupe à score de risque modéré quant aux expositions aux radiations
5. Résultats obtenus dans le groupe à score de risque élevé quant aux expositions aux radiations

Facteurs de la reproduction

Les analyses stratifiées sur les groupes construits à partir des facteurs gynéco-obstétriques ont été faites sur les SNPs trouvés associés à la fois aux gènes intervenant dans la régulation et l'action des hormones sexuelles et aux gènes impliqués dans la régulation de l'ADN.

Parmi les SNPs associés aux gènes intervenant dans la régulation et l'action des hormones sexuelles, 30 SNPs sont significatifs dans le groupe des sujets à score de risque faible et moyen. Aucun SNP n'est associé de façon significative avec le groupe à score de risque élevé. Cependant, malgré une hétérogénéité significative pour 29 des 30 SNPs, la p_{exacte} calculée avec le test de permutation est proche ou égale à 1 pour les 30 SNPs (Tableaux supplémentaires 8 à 11). Aucun SNP n'est donc conservé avec ce filtre.

26 SNPs dans les gènes intervenant dans la réparation de l'ADN sont trouvés associés au cancer du sein dans au moins un des 3 groupes et montrent une hétérogénéité significative (9 pour le groupe à score de risque faible, 5 pour le groupe à score de risque moyen et 12 pour le groupe à score de risque élevé) (Tableaux supplémentaires 12 à 17). Cependant, comme précédemment, aucun de ces SNPs n'est associé à une p_{exacte} inférieure à 0,05.

Aucun SNP candidat spécifique aux facteurs gynéco-obstétriques n'a donc été retenu comme SNP spécifique candidat selon les groupes de score de risque formés à partir des facteurs gynéco-obstétriques.

Discussion

Les analyses d'associations des facteurs de risque dans la population des femmes de GENESIS montrent, pour certains, des résultats différents de ceux observés dans la population générale. Alors qu'un âge aux premières règles précoces et l'allaitement diminuent le risque de cancer du sein dans la population générale[55,63,65], ces facteurs ne semblent pas avoir d'effet dans la population de GENESIS. Les contraceptifs oraux ont un effet encore controversé dans la population générale, mais les études qui trouvent un effet significatif, montrent une augmentation d'environ 20 % du risque de cancer chez les utilisatrices[59,60]. La diminution du risque de cancer du sein observé dans la population de GENESIS est donc surprenante. Cette observation est peut-être due au fait que les femmes appartenant à une famille à haut risque de cancer du sein reçoivent moins fréquemment une prescription de contraception orale qu'une femme de la population générale par crainte de son effet potentiellement délétère et consomment donc réellement moins de contraceptifs oraux.

L'effet de la parité observé dans la population GENESIS est également différent de celui observé dans la population générale[56,63]. Cependant, l'augmentation du risque de cancer du sein associé au premier enfant comparé à la nulliparité a aussi été observée chez les femmes à haut risque de cancer du sein porteuses d'une mutation sur *BRCA1/2*[155].

Enfin, les effets de la ménopause et des radiations ionisantes au thorax sur la population de GENESIS sont comparables à ceux observés dans la population générale.

Les divergences observées suggèrent que les cancers du sein des cas de GENESIS ont une étiologie différente de celle des cas de cancer du sein de la population générale.

L'analyse d'association des SNPs dans la population de GENESIS met en évidence 4 régions en 5q14.2, 6p11.2, 7p11.2 et 10q26.13 associées avec le risque de cancer du sein. Parmi ces régions, 39 SNPs sont localisés dans le gène *FGFR2*, gène déjà connu et mis en évidence dans les études en population générale du consortium BCAC[141], avec une estimation ponctuelle semblable ($OR_{GENESIS}$ = 0,73, $IC_{95\%}$ = [0,65-0,82] et OR_{BCAC} = 0,78, $IC_{95\%}$ = [0,76–0,80][210]) et une *p-value* égale à $6{,}4.10^{-155}$. Ce gène a également été retrouvé associé chez les porteurs d'une mutation *BRCA2* (communication personnelle). Trois nouvelles régions ont été trouvées chez les femmes de GENESIS. Ces trois régions correspondent aux gènes *PRIM2, XRCC4* et *EGFR* qui pourraient être spécifiques à la population à haut risque de GENESIS.

Les analyses stratifiées sur les groupes de scores de risque montrent des différences significatives entre les groupes construits à partir des expositions aux radiations. En effet, on remarque que dans le groupe le plus exposé (groupe à score de risque élevé) on met en évidence principalement des SNPs localisés dans le gène *FGFR2, SNPs* trouvés en population générale. Par contre, dans le groupe des sujets non exposés, il semble que des SNPs associés aux gènes *XRCC4* et *MAGI1* soient spécifiques à ce groupe. Les SNPs localisés dans le gène *FGFR2* sont trouvés associés au risque de cancer du sein dans la population de GENESIS mais également dans le groupe le plus exposé aux radiations ionisantes avec une hétérogénéité significative. Il est alors possible que l'effet mesuré dans l'analyse non stratifiée soit dû aux sujets appartenant au groupe à score de risque élevé seulement et que les SNPs associés à ce gène soient associés au cancer du sein des sujets de ce groupe uniquement.
Alors que les analyses stratifiées selon les expositions aux radiations permettent de mettre en évidence des associations différentes entre les groupes de score de risque, aucun SNP n'a été retenu lorsque la population a été stratifiée selon le risque associé aux facteurs gynéco-obstétriques.

Pour construire les groupes de risques, j'ai utilisé les résultats *a posteriori* de la régression logistique sur les facteurs de risque significatifs dans la population d'étude. Cette stratégie, certes naïve, a été utilisée pour sa simplicité de mise en pratique et d'interprétation. Cependant, sur les 2 types d'expositions environnementales étudiées, seule la stratification sur les expositions aux radiations a mis en évidence des associations hétérogènes entre les groupes de scores de risque. Le modèle de régression construit pour la stratification sur les facteurs gynéco-obstétriques n'est probablement pas optimal pour construire des groupes homogènes pour ces facteurs. Il aurait peut-être fallu intégrer tous les facteurs étudiés dans le modèle mais ces derniers sont très corrélés entre eux et les modèles de régression logistique gèrent mal ces corrélations. D'autres méthodes statistiques ont été envisagées comme l'utilisation des ACPs[211] ou de régressions LASSO[212] ou RIDGE[213]. Ces méthodes s'affranchissent de l'interdépendance entre les facteurs et permettraient donc de prendre en compte tous les facteurs d'intérêt pour construire le score de risque qui correspondrait alors mieux à l'exposition totale des sujets. Ces méthodes devront être explorées pour améliorer la construction des groupes de score de risque et, par conséquent, pour optimiser les analyses stratifiées qui en dépendent.

Les facteurs gynéco-obstétriques et les expositions aux radiations ionisantes ont probablement un impact sur la fonction des gènes intervenant dans les voies biologiques impliquées dans la réparation de l'ADN et dans la régulation et l'action des hormones sexuelles. C'est pourquoi nous avons décidé dans un premier temps de concentrer nos recherches sur les SNPs présents dans les gènes impliqués appartenant à ces voies biologiques. Par conséquent, plus de 95 % du génome n'a pas été analysé. Il serait donc intéressant d'appliquer cette stratégie d'analyse sur le génome entier, de façon agnostique et sans *a priori* sur la fonction des SNPs étudiés. Cependant, le fait d'avoir restreint notre analyse à des gènes spécifiques nous a permis d'utiliser un seuil de signification (*p-value*) pour les analyses de régression moins strict que celui utilisé dans les études *GWAS* (respectivement $p < 1{,}05.10^{-4}$ et $p < 10^{-8}$). Aucun des SNPs n'est associé au seuil $p < 10^{-8}$. Cela s'explique par la petite taille de notre population d'étude, réduite d'autant plus par la stratification de la population en trois groupes de scores de risque. Les analyses sur le génome entier devront alors être réalisées sur des populations de taille beaucoup plus importante pour espérer mettre en évidence des différences de façon significative entre les profils d'exposition.

Les SNPs analysés ont été imputés à partir des SNPs génotypés de la puce iCOGS. La validation croisée permettant de vérifier la qualité globale des imputations montre que 73 % des fragments imputés n'atteignent pas le taux de concordance attendu de 0,95, même après ré-imputation avec des fragments de plus grande taille. Il est donc possible que les SNPs de ces fragments soient mal imputés, ce qui pourrait entrainer des faux-positifs ou faux-négatifs. Sachant que la puce iCOGS est une puce à façon, sans squelette de SNPs marqueurs le long du génome, il est possible que les fragments les moins bien imputés soient ceux où la densité de SNPs génotypés est la plus faible et/ou leur répartition la plus hétérogène. De plus, on remarque que plus de 90 % des SNPs imputés sont associés à un r^2 inférieur à 0,5, la majorité de ces SNPs étant des SNPs rares (MAF < 0,05). Le nombre de SNPs exclus est très important et nous fait certainement passer à côté de SNPs candidats. L'imputation a été réalisée avec le panel de référence *1000Génomes*. Cependant, il existe d'autres panels de référence plus récents mis en place par *the Haplotype Reference Consortium*[214] (HRC) ou le programme *Trans-Omics for Precision Medicine* (TOPMed)[215], qui pourraient être plus adaptés à l'imputation de la puce iCOGS. En effet, le panel de référence *1000Génomes* contient la séquence du génome entier de 2 504 sujets[216] alors que le panel de référence de HRC en contient plus de 32 000 et celui de TOPMed plus de 62 000 sujets. Le nombre de sujets beaucoup plus important dans ces panels de référence permet de couvrir une diversité

génétique beaucoup plus large et les SNPs rares y sont, par conséquent, mieux représentés. Leur utilisation pourrait donc améliorer la qualité d'imputation des SNPs rares et réduire le nombre de SNPs exclus.

Cependant, nous avons également exclus les SNPs selon leur fréquence (MAF < 0,05). En effet, la taille modeste de la population de GENESIS rend difficile l'étude d'association de SNPs rares au risque de mettre en évidence des associations qui seraient en réalité des artefacts. Leur exclusion nous a permis d'éviter de biaiser les estimations des risques relatifs vers l'hypothèse alternative et de mettre en évidence de fausses associations avec le cancer du sein. Malgré l'homogénéité de notre population, sa taille modeste limite la puissance statistique nécessaire pour mettre en évidence des associations au seuil de Bonferroni fixé à $1{,}05.10^{-4}$. Ces analyses stratifiées selon les facteurs environnementaux devraient donc être reproduites dans des populations de taille plus importante.

Les cas de cancer du sein de la population de GENESIS ont été recrutés via les consultations d'oncogénétique. Par conséquent, plus de 90 % des cas sont des cas prévalents. Il est donc possible que les SNPs trouvés dans l'analyse globale soient des gènes de bon pronostic et non des gènes de prédisposition. Des analyses de sensibilité restreintes aux cas « pseudo-incidents » (c-à-d, cas diagnostiqués depuis moins de 5 ans) auraient pu être réalisées pour tester l'existence d'un biais de survie. Cependant, dans la mesure où seulement 39 % des cas de GENESIS sont des cas « pseudo-incidents », il n'aurait pas été possible de conclure. Par contre, aucun des gènes associés aux SNPs trouvés associés ont été décrits comme associés à la survie[217]. De plus, chaque groupe de scores de risque construits à partir des expositions aux facteurs environnementaux contient la même proportion de cas prévalents. Les associations spécifiques mises en évidence entre les groupes de scores de risque sont plus vraisemblablement dues à des différences étiologiques qu'à des différences de survie.

Les témoins de GENESIS sont des amis ou des collègues des cas index. Il est possible que ces témoins aient été motivés pour participer à l'étude à cause d'un nombre important de personnes atteintes de cancers dans leur famille. Cependant, l'analyse de l'histoire familiale des témoins réalisée par Marie-Gabrielle Dondon (Épidémiologie Génétique des Cancers, U900, Institut Curie) a montré une très faible augmentation de l'incidence des cancers dans ces familles[187]. Ce biais aurait pu induire une sous-estimation de l'effet des facteurs génétiques identifiés.

La stratégie que j'ai mise en place pour intégrer les profils d'expositions environnementales à la recherche de SNPs associés au risque de cancer du sein m'a permis de mettre en évidence des associations différentes selon l'exposition aux radiations. Les résultats que j'ai obtenus semblent valider notre hypothèse d'associations différentes selon le profil d'exposition des individus. Les résultats obtenus devront être approfondis pour notamment comprendre la fonction des SNPs identifiés. La construction des groupes de score de risque doit également être optimisée grâce à l'étude d'autres méthodes de *clustering*. Cependant, ces analyses prenant en compte le profil d'exposition doivent être poursuivies et devraient être mises en place dans des études de plus grande dimension. Des modèles de prédiction ont été développés pour prendre en compte les facteurs génétiques et non génétiques de façon simultanée[175]. Dans le cas où notre hypothèse serait confirmée, ces modèles pourraient évoluer vers une adaptation des PRS en fonction des profils d'exposition des femmes et cela tout au long de leur vie.

Deuxième Partie

Facteurs génétiques modificateurs du risque de cancer du sein chez les femmes porteuses d'une mutation dans les gènes *BRCA1* ou *BRCA2*

Introduction

Comme précédemment énoncé page 47, les gènes *BRCA1* et *BRCA2*, les deux premiers gènes trouvés associés au risque familial de cancer du sein, sont associés à des risques de cancer du sein très élevés mais ne suffisent pas pour développer la maladie. Ils sont dits à « pénétrance incomplète ». Dans l'étude récente de Kuchenbaecker *et al*.[150], le risque cumulé à 80 ans de cancer du sein a été estimé à 72 % pour les femmes porteuses d'une mutation de *BRCA1* et à 69 % pour les porteuses d'une mutation *BRCA2*[150]. De plus, ce risque varie fortement entre les femmes porteuses d'une mutation dans l'un de ces gènes[150] (voir partie « Facteurs modificateurs du risque », page 49), ce qui suggère la présence de facteurs modificateurs du risque de cancer du sein dans cette population, en particulier de facteurs génétiques modificateurs.

Plus de 180 SNPs ont été identifiés, au seuil $p < 10^{-8}$, comme étant associés au risque de cancer du sein[141,142] dans la population générale et une cinquantaine chez les femmes portant une mutation *BRCA1/2*[160,161,168,162,169,164]. Malgré le nombre important de sujets participant à l'étude CIMBA, les études faites sur cette population n'ont peut-être pas la puissance nécessaire pour détecter d'autres SNPs spécifiquement associés au risque de cancer du sein dans cette population.

Nous avons donc mis en place une nouvelle stratégie utilisant un design d'étude *case-only*, dans laquelle la fréquence des SNPs des cas porteurs d'une mutation *BRCA1/2* est comparée à celle des cas de population générale. Cette étude a pour objectif d'identifier de nouveaux SNPs qui modifient le risque de cancer du sein chez les sujets porteurs d'une mutation ainsi que d'évaluer, dans cette population spécifique, l'effet des SNPs trouvés associés dans la population générale. Cette partie de ma thèse fait l'objet d'une publication soumise au journal *Nature Communication* : Coignard J. *et al*., *A case-only study to identify genetic modifiers of breast cancer risk specifically for BRCA1 and BRCA2 mutation carriers* (page 286).

Données

I. La population d'étude

Les données des consortia internationaux BCAC et CIMBA ont constitué la population d'étude pour rechercher de nouveaux facteurs modificateurs du risque de cancer du sein chez les porteurs d'une mutation dans les gènes *BRCA1* ou *BRCA2* dans le cadre de ma thèse.

1. Consortium BCAC

BCAC[1] regroupe des études mises en place dans différents pays du monde avec pour objectif de mettre en évidence les facteurs de risque génétique du cancer du sein dans la population générale. Ce consortium a été créé en 2005[131] dans le but de générer une base de données d'une taille suffisamment importante.

a. Critères d'inclusion et données recueillies

Les études souhaitant participer au consortium BCAC doivent soit compter au moins 500 cas de cancer du sein invasif et 500 témoins provenant de la même population à risque que les cas, soit avoir au moins 1 000 cas de cancer du sein invasif. Un échantillon d'ADN de chacun de ces sujets est nécessaire pour pouvoir effectuer le génotypage. Un grand nombre de données phénotypiques (par exemple la localisation et le type de la tumeur) et épidémiologiques (l'âge, la date de naissance, l'ethnicité, l'histoire familiale, etc.) sont également enregistrées lorsque ces données sont disponibles. Les participants ont tous signé un consentement éclairé de participation à l'étude.

[1] http://bcac.ccge.medschl.cam.ac.uk, consulté le 23 octobre 2019.

b. Descriptif de la population

BCAC rassemble les données de 188 320 cas de cancer du sein et 161 669 témoins de population dite générale. Ces sujets proviennent de 108 études réalisées dans 33 pays d'Amérique du Nord, d'Amérique du Sud, d'Europe, d'Australie et d'Asie. Plus de 88 % de ces études sont des études cas-témoins. Les autres études sont des études *case-only* (11 %) ou *control-only* (0,01 %), n'ayant inclus respectivement que des cas de cancer du sein ou des témoins. La majorité des cas inclus dans le consortium BCAC ont un cancer du sein invasif (93 %) et de type canalaire (55 %) (Tableau 20 et Tableau 21).

Tableau 20 - Descriptif des types de tumeurs et des statuts des récepteurs aux œstrogènes (RE) et à la progestérone (RP) des femmes de BCAC

	Type de cancer du sein			**Statut RE**			**Statut RP**		
	Invasif	**In-situ**	**Inconnu**	**Positif**	**Négatif**	**Inconnu**	**Positif**	**Négatif**	**Inconnu**
Nombre de cas	172 571	12 480	3 269	104 455	31 640	52 225	78 163	41 974	68 183
Pourcentage	0,91	0,07	0,02	0,55	0,17	0,28	0,42	0,22	0,36

Tableau 21 - Descriptif des types histologiques des tumeurs des femmes de BCAC.

	Type histologique								
	Canalaire	**Lobulaire**	**Médullaire**	**Mixte***	**Mucineux**	**Papillaire**	**Tubulaire**	**Autre**	**Inconnu**
Nombre de cas	102 756	16 706	1 094	5 141	1 785	427	1 794	6 032	52 585
Pourcentage	0,54	0,09	0,01	0,03	0,01	0,00	0,01	0,03	0,28

*Canalaire et Lobulaire

2. Consortium CIMBA

CIMBA[m] a aussi été mis en place en 2005[156], par un groupe de chercheurs travaillant sur les facteurs génétiques modificateurs du risque de cancer du sein des porteurs d'une mutation dans les gènes *BRCA1* ou *BRCA2*. Le but de ce regroupement, comme pour le consortium BCAC, était de rassembler suffisamment de sujets pour évaluer de façon précise l'effet des facteurs génétiques modificateurs.

[m] http://cimba.ccge.medschl.cam.ac.uk, consulté le 23 octobre 2019.

a. Critères d'inclusion et données recueillies

Pour participer à CIMBA, les études doivent contenir un minimum de 50 femmes ou 20 hommes, porteurs d'une mutation dans les gènes *BRCA1* ou *BRCA2* et atteints ou non d'un cancer du sein. Les sujets inclus dans l'étude doivent être âgés de 18 ans ou plus à l'interview et avoir signé un consentement éclairé de participation à l'étude. Un minimum d'informations phénotypiques, comme l'année de naissance ou l'ethnicité, et de données familiales sont également requises.

b. Descriptif de la population

La majorité des sujets inclus dans CIMBA ont été recrutés par le biais de consultations de génétique. CIMBA rassemble les données de 80 études provenant de 38 pays différents répartis en Europe, Amérique du Nord et du Sud, Australie, Asie et Afrique soit 65 000 sujets (7 000 hommes et 56 000 femmes) porteurs d'une mutation dans le gène *BRCA1* ou *BRCA2* dont 43 000 ont été génotypés avec les puces iCOGS ou OncoArray développées par Illumina. Le type de la tumeur est connu pour 97 % des femmes atteintes d'un cancer du sein et le statut des récepteurs aux œstrogènes est renseigné pour 72 % d'entre elles (Tableau 22 et Tableau 23).

Tableau 22 - Descriptif des types de cancer du sein et du statut des récepteurs aux œstrogènes (RE) et à la progestérone (RP) des femmes de CIMBA

	Type de cancer du sein			**Statut RE**			**Statut RP**		
	Invasif	***In situ***	**Inconnu**	**Positif**	**Négatif**	**Inconnu**	**Positif**	**Négatif**	**Inconnu**
					BRCA1				
Nombre de cas	9 989	326	305	1 824	5 637	3 159	1 513	5 438	3 669
Proportion	0,94	0,03	0,03	0,17	0,53	0,3	0,14	0,51	0,35
					BRCA2				
Nombre de cas	6 354	586	189	4 014	1 150	1 965	3 090	1 610	2 429
Proportion	0,89	0,08	0,03	0,56	0,16	0,27	0,43	0,23	0,34

Tableau 23 - Descriptif des types histologiques des tumeurs des femmes de CIMBA

	Type histologique			
	Canalaire	**Lobulaire**	**Médullaire**	**Inconnu**
		BRCA1		
Nombre de cas	7 496	201	630	2 293
Proportion	0,71	0,02	0,06	0,21
		BRCA2		
Nombre de cas	5 196	466	76	1 391
Proportion	0,73	0,06	0,01	0,20

Lorsque l'on compare les données recueillies par les deux consortia, on remarque que les taux de données inconnues sont comparables. Cependant, CIMBA contient 2 fois plus de sujets ayant développé des tumeurs RE^- (38 %) que BCAC (17 %). En ce qui concerne les récepteurs RP, les taux sont équivalents entre BCAC et CIMBA (RP^+ = 41 % et 39 % respectivement).

II. Les données génotypiques

Le génotypage des sujets de BCAC et CIMBA a été effectué grâce à deux puces à ADN, les puces iCOGS[136] et OncoArray[140]. Plus de 70 % des sujets ont été génotypés avec la puce OncoArray (141 029 sujets pour BCAC ; 16 519 sujets porteurs d'une mutation de *BRCA1* et 12 775 sujets porteurs d'une mutation de *BRCA2* pour CIMBA). Dans le cadre de ce projet, seuls les sujets européens génotypés avec OncoArray ont été inclus dans les analyses. Le contrôle qualité et la préparation des données d'OncoArray ainsi que l'imputation des données manquantes ont été réalisés par Joe Dennis (*Centre for Cancer Genetic Epidemiology,* Université de Cambridge, UK).

La puce à ADN OncoArray a été créée par le consortium du même nom[140], consortium qui a pour objectif d'identifier des SNPs jouant un rôle dans les cancers les plus fréquents comme le cancer du sein, du colon, du poumon, de l'ovaire et de la prostate[140]. Cette puce est postérieure à la puce iCOGS.

1. Description de la puce OncoArray

La puce OncoArray est une puce Illumina ciblant 570 000 SNPs. Environ 260 000 d'entre eux forment un squelette (*GWAS backbone*) correspondant à un ensemble de SNPs répartis sur tout le génome. L'imputation des génotypes manquants à partir de ce squelette permettra de couvrir la totalité des SNPs connus (et présents dans les données des panels de référence) et de réaliser des analyses d'association sur le génome entier. Des associations sur de nouvelles régions du génome peuvent alors être mises en évidence.

L'autre moitié des SNPs de la puce OncoArray ont été choisis par les différents consortia impliqués. Leur sélection s'est basée sur :

- les SNPs candidats montrant une association avec une *p-value* inférieure à 10^{-5} dans les études *GWAS* précédemment réalisées par les consortia ;
- les résultats des études de *fine-mapping* effectuées sur les régions de ces SNPs ;
- les SNPs candidats provenant d'analyses de génomes entiers ou d'exomes ;
- les SNPs trouvés associés dans d'autres types de cancer ;
- les SNPs montrant une association avec un mécanisme fonctionnel pertinent vis-à-vis du cancer ;
- les SNPs associés à des traits phénotypiques corrélés au cancer comme l'IMC ou la consommation de tabac.

2. Contrôle qualité

Comme pour les données génotypiques de l'étude GENESIS obtenues grâce à la puce iCOGS, les 570 000 SNPs génotypés avec à la puce OncoArray ont passé un contrôle qualité dans le but d'exclure les SNPs mal génotypés. Ce contrôle qualité a été réalisé indépendamment pour les sujets de BCAC et de CIMBA. Il a nécessité les étapes suivantes :

a. Équilibre de Hardy-Weinberg

La première étape du contrôle qualité consiste à vérifier que les SNPs respectent l'équilibre de Hardy-Weinberg (voir pages 61). Le test a été effectué conjointement sur les cas et les témoins et stratifié par pays, avec un seuil de signification fixé à 10^{-7}. Tous les SNPs associés à une *p-value* inférieure à 10^{-7} ont été exclus.

b. « Call-rate »

Le seuil a été fixé à 95 % pour les SNPs fréquents, entrainant l'exclusion des analyses les SNPs ayant plus de 5 % de données manquantes. Ce seuil a été élevé à 98 % pour les SNPs rares, c'est-à-dire les SNPs associés à une fréquence allélique inférieure à 1 %.

c. SNPs monomorphiques

Certaines positions du génome sont polymorphiques dans des populations spécifiques et pas dans d'autres. Les SNPs présents sur OncoArray ont été mis en évidence dans plusieurs études qui ont été réalisées sur des populations différentes. Cependant, certaines de ces populations peuvent ne pas être représentées dans les populations de BCAC et de CIMBA analysées. Il est donc possible qu'aucun sujet de notre population ne soit porteur de certains SNPs présents sur OncoArray qui apparaissent monomorphiques dans notre population. Ces SNPs sont alors exclus de nos analyses.

d. Chromosome Y

La puce OncoArray contient des SNPs localisés sur le chromosome Y. Les femmes de CIMBA et de BCAC ne devraient pas avoir de valeurs pour ces derniers. Ces SNPs ne peuvent donc résulter que d'une erreur de génotypage et ont alors également été exclus.

Ces différentes étapes de contrôle qualité ont diminué le nombre de données manquantes et augmenté la fiabilité des SNPs à analyser. Les SNPs exclus lors de ce contrôle qualité seront imputés par la suite (exceptés ceux situés sur le chromosome Y).

Méthodes

I. Imputation des génotypes manquants

Comme la puce iCOGS, la puce OncoArray ne contient qu'une infime partie (0,0075 %) de l'ensemble des SNPs connus. Cependant, à la différence de la puce iCOGS, cette puce contient un squelette de SNPs tout le long du génome, ce qui permet de marquer l'ensemble des régions du génome de façon beaucoup plus précise et d'augmenter la qualité de l'imputation des génotypes manquants.

1. Paramètres de l'imputation

Les imputations des sujets de BCAC et CIMBA ont été réalisées séparément. La première étape a été de prédire les haplotypes avec le logiciel SHAPEIT[200] puis d'inférer les génotypes manquants grâce à l'imputation avec le logiciel IMPUTE2[201]. Comme pour l'imputation des génotypes manquants des femmes de l'étude GENESIS (voir page 74), le panel de référence complet ainsi que tous les SNPs présents dans la base de données de référence (quelle que soit leur fréquence chez les sujets européens du panel de référence) ont été utilisés pour les imputations. Par ailleurs, le paramètre k_hap, donnant le nombre d'haplotypes de référence utilisés pour l'imputation de chaque sujet, a été fixé à 800, à la différence de l'imputation de GENESIS où je l'ai fixé à 500.

2. Contrôle qualité

Nous avons sélectionné pour la suite des analyses les SNPs avec un score de qualité d'imputation r^2 supérieur ou égal à 0,5 uniquement, ce qui signifie que si le r^2 était inférieur à 0,5 dans au moins l'un des deux consortia, le SNP était exclu. Cependant, ce seuil a été ajusté en fonction de la différence obtenue entre la qualité d'imputation de BCAC et celle de CIMBA. En effet, comme nous l'avons dit précédemment, les sujets des deux consortia ont été imputés de façon indépendante. De ce fait, il a fallu vérifier que les SNPs avaient été

imputés avec la même qualité d'imputation, c'est-à-dire avec la même précision, entre les deux consortia. Plus les différences de qualité d'imputation entre les deux groupes d'études étaient grandes, plus nous avons été stricts sur le seuil appliqué au r^2. Pour cela nous avons calculé la différence de qualité d'imputation entre les 2 consortia pour chaque SNP ($\Delta r^2 = |r^2_{BCAC} - r^2_{CIMBA}|$) et seuls les SNPs respectant les critères détaillés dans le Tableau 24 ont été conservés. Par exemple, lorsque r^2 était supérieur à 0,9 dans BCAC et CIMBA, le SNP était conservé pour les analyses si la différence Δr^2 était inférieure à 0,05. Parmi les SNPs imputés respectant ces critères, seuls ceux associés à une MAF supérieure à 1 % dans BCAC ont été conservés. Au total, 8 873 923 SNPs, soit 45,3 % des SNPs imputés, ont été conservés pour l'analyse BRCA1 et 8 829 998, soit un taux comparable de 45,1 %, pour l'analyse BRCA2.

Tableau 24 - Critères d'inclusion des SNPs selon le Δr^2 entre BCAC et CIMBA en fonction du r^2

	Score d'imputation r^2		
	> 0,9	**]0,8-0,9]**	**[0,5-0,8]**
Δr^2	< 0,05	< 0,02	< 0,01

3. Imputation jointe des régions d'intérêt

Les analyses du génome faites sur les données de BCAC et CIMBA qui ont été imputées séparément ont permis de mettre en évidence des régions d'intérêt, c'est-à-dire des régions où une association avec un ou plusieurs SNPs a été trouvée. Ces régions ont alors été imputées une seconde fois avec les sujets de BCAC et de CIMBA réunis dans la même imputation (cas et témoins séparés) dans le but de vérifier que les résultats obtenus précédemment n'étaient pas des faux positifs dus à des différences d'imputation. Compte tenu de la taille réduite du génome à imputer à cette étape, cette ré-imputation a pu être réalisée en une seule étape avec le logiciel IMPUTE2, sans étape préliminaire de pré-haplotypage. La précision de l'imputation a ainsi pu être augmentée. Cependant, imputer des génotypes manquants sur des sujets non phasés nécessite une capacité et un temps de calcul beaucoup plus importants et nous n'avions pas les ressources informatiques nécessaires pour imputer plus de 70 000 sujets simultanément. Il a donc fallu effectuer cette ré-imputation en plusieurs fois, en répartissant

les sujets en 5 groupes appariés sur le pays d'origine selon le nombre d'individus dans chaque pays (Tableau 25) (1 : Allemagne ; 2 : 2/3 des États-Unis ; 3 : 1/3 des États-Unis, Suède et Australie ; 4 : Angleterre, Russie, Belgique, Canada et Israël et 5 : Pays-Bas, Italie, Pologne, Espagne, Danemark, Finlande, France et Grèce). Comme pour la première imputation, le paramètre k_hap, indiquant le nombre de sujets de référence utilisés, est fixé à 800.

Tableau 25 - Nombre de cas par étude et par pays.

	BCAC		**CIMBA**		
Pays	**Études**	**Nombre de cas**	**Études**	**Nombre de cas**	
				BRCA1	**BRCA2**
Australie	MCCS	1 051	NRG_ONCOLOGY	1	4
	ABCTB	951	KCONFAB	368	295
	ABCFS	1 087	BCFR-AU	25	28
	BCEES	783	VFCTG	103	70
		Total : 3 872 (6,42 %)		**Total : 497 (6,85 %)**	**Total : 397 (7,79 %)**
Belgique	LMBC	789	G-FAST	121	76
		Total : 789 (1,31 %)		**Total : 121 (1,67 %)**	**Total : 76 (1,49 %)**
Canada	CBCS	676	MCGILL	24	14
	OFBCR	1 658	BCFR-ON	88	60
	MTLGEBCS	341	OCGN	71	64
			INHERIT	37	34
		Total : 2 675 (4,43 %)		**Total : 220 (3,03 %)**	**Total : 172 (3,37 %)**
Danemark	CGPS	1411	CBCS	76	64
			OUH	191	167
		Total : 1 411 (2,34 %)		**Total : 267 (3,68 %)**	**Total : 231 (4,53 %)**
Finlande	HEBCS	281	HEBCS	53	67
	KBCP	556			
		Total : 837 (1,39 %)		**Total : 53 (0,73 %)**	**Total : 67 (1,31 %)**
France	CECILE	306	GEMO	758	563
	EPIC	433			
		Total : 739 (1,23 %)		**Total : 758 (10,45 %)**	**Total : 563 (11,04 %)**
Allemagne	ESTHER	296	GC-HBOC	1 168	646
	SKKDKFZS	1 091	DKFZ	36	14
	GESBC	351			
	GENICA	460			
	BBCC	441			
	MARIE	512			
	BSUCH	269			
	EPIC	661			
	GC-HBOC	3 634			
	HABCS	929			
		Total : 8 644 (14,33 %)		**Total : 1 204 (16,59 %)**	**Total : 660 (12,95 %)**

	BCAC		CIMBA		
Pays	**Études**	**Nombre de cas**	**Études**	**Nombre de cas**	
				BRCA1	**BRCA2**
Grèce	EPIC	182	DEMOKRITOS	132	23
	CCGP	670			
		Total : 852 (1,41 %)		**Total : 132 (1,82 %)**	**Total : 23 (0,45 %)**
Israël	BCINIS	1330	SMC	66	33
		Total : 1 330 (2,2 %)		**Total : 66 (0,91 %)**	**Total : 33 (0,65 %)**
Italie	EPIC	822	CONSIT TEAM	271	187
	MBCSG	787	IOVHBOCS	109	113
			PBCS	49	6
		Total : 1 609 (2,67 %)		**Total : 429 (5,91 %)**	**Total : 306 (6 %)**
Pays-Bas	RBCS	473	HEBON	374	199
	EPIC	709			
	ORIGO	1 055			
	ABCS	267			
		Total : 2 504 (4,15 %)		**Total : 374 (5,15 %)**	**Total : 199 (3,9 %)**
Pologne	PBCS	1 931	IHCC	77	0
	SZBCS	379			
		Total : 2 310 (3,83 %)		**Total : 77 (1,06 %)**	**Total : 0 (0 %)**
Russie	HUBCS	211	BIDMC	1	0
			NNPIO	44	2
		Total : 211 (0,35 %)		**Total : 45 (0,62 %)**	**Total : 2 (0,04 %)**
Espagne	BREOGAN	1 376	HCSC	56	76
	EPIC	337	ICO	130	185
	HCSC	426	HVH	62	63
			FPGMX	67	44
			CNIO	31	33
			iovhbocs	1	0
		Total : 2 139 (3,55 %)		**Total : 347 (4,78 %)**	**Total : 401 (7,87 %)**
Suède	SMC	1 504	SWE-BRCA	190	25
	KARBAC	497			
	MISS	697			
	PKARMA	2 991			
		Total : 5 689 (9,43 %)		**Total : 190 (2,62 %)**	**Total : 25 (0,49 %)**
Angleterre	UKBGS	1 632	OUH	1	0
	SEARCH	4 057	EMBRACE	795	768
	POSH	1 088	UKGRFOCR	13	4
	DIETCOMPLYF	711			
	BBCS	122			
	EPIC	703			
		Total : 8 313 (13,78 %)		**Total : 809 (11,15 %)**	**Total : 772 (15,14 %)**

	BCAC		CIMBA		
Pays	**Études**	**Nombre de cas**	**Études**	**Nombre de cas**	
				BRCA1	**BRCA2**
Etats-Unis	UCIBCS	490	BIDMC	43	24
	SISTER	2 016	FCCC	26	11
	MEC	672	UTMDACC	25	39
	CTS	1 156	NORTHSHORE	40	19
	NHS2	1 606	DFCI	65	46
	NBHS	677	BRICOH	52	48
	MCBCS	925	WCP	51	18
	PLCO	868	NRG_ONCOLOGY	165	141
	CPSII	3 054	OSU CCG	50	56
	BCFR-PA	132	KUMC	24	12
	MSKCC	120	UCSF	33	28
	2SISTER	1 071	GEMO	84	25
	NHS	1 590	BCFR-NC	33	22
	BCFR-UTAH	102	GEORGETOWN	5	0
	NC-BCFR	712	MAYO	122	74
	BCFR-NY	454	BCFR-PA	18	3
	TNBCC	373	NCI	42	21
	MMHS	384	COH	141	98
			BCFR-NY	37	25
			UPENN	240	168
			MSKCC	185	167
			BCFR-UT	67	54
			UCHICAGO	43	29
			UPITT	77	43
		Total : 16 402 (27,19 %)		**Total : 1 668 (22,98 %)**	**Total : 1 171 (22,97 %)**
		Total : 60 326		**Total : 7 257**	**Total : 5 098**

II. Analyse *case-only*

Nous avons proposé un design d'analyse *case-only* (Tableau 26) pour comparer la fréquence des SNPs des cas de cancer du sein de la population générale (consortium BCAC) à celle des cas porteurs d'une mutation dans les gènes *BRCA1* ou *BRCA2* (consortium CIMBA).
Les analyses des femmes porteuses d'une mutation de *BRCA1* et celles porteuses d'une mutation de *BRCA2* ont été faites séparément.

Tableau 26 – Design de l'analyse Case-only.

Mutation BRCA1/2	**Présence d'au moins un allèle mineur du SNP**	
	Oui	**Non**
Oui (cas CIMBA)	a	b
Non (cas BCAC)	c	d

Les études *case-only* s'affranchissent de la sélection de témoins, ce qui permet d'éliminer les biais potentiels dus à leur sélection. De plus, à taille d'échantillon égale, une étude *case-only* est plus puissante en moyenne d'un facteur 4 qu'une étude cas-témoin pour détecter des interactions[218].

1. Sélection des sujets

a. Définition du statut atteint ou non atteint

Dans le consortium BCAC, le statut des femmes, cas ou témoins, est défini à leur inclusion. Seules les femmes ayant déjà développé un cancer du sein à l'inclusion sont considérées comme des cas.
Les études du consortium CIMBA sont des cohortes rétrospectives et chaque femme est suivie de la naissance à l'inclusion dans l'étude ou à la survenue de l'un de ces événements : un diagnostic de cancer du sein, de l'ovaire ou d'un autre organe ou une mastectomie bilatérale prophylactique. Seules les femmes dont le premier événement au cours de leur vie est un

diagnostic de cancer du sein sont considérées comme des cas. Toutes les autres sont considérées comme des témoins.

b. Sélection des cas

- *Origine ethnique*

Seules les femmes d'origine caucasienne ont été incluses dans les analyses afin d'éviter les biais dus à une stratification de la population. La population d'étude est composée de 60,9 % de femmes d'origine européenne, 31,2 % provenant des États-Unis, 6,1 % d'Australie et 1,7 % d'Israël.

- *Type d'étude*

Dans notre étude, les femmes de BCAC représentent les femmes de la population générale. Pour respecter ce postulat, nous avons exclu les études restreintes à des types spécifiques de cancer du sein tels que les tumeurs « Her2-positif » ou les cancers du sein « triple-négatif ». Quatre études comptant au total 3 478 femmes ont donc été exclues (Annexe 7, page 280). Aucune étude de CIMBA n'a été exclue (Annexe 8, page 282). Les femmes de 65 études de BCAC et de 52 études de CIMBA provenant de 17 pays ont été incluses (Tableau 25).

- *Pays d'origine*

Bien que les analyses aient été restreintes aux femmes d'origine caucasienne, il existe également une variabilité génétique selon le pays d'origine. Les modèles de régression ont été ajustés sur le pays afin de prendre en compte cette variabilité dans les analyses d'association. Cet ajustement sur le pays permet de plus d'absorber une partie de la variabilité due à des facteurs confondants non génétiques. Il a donc été nécessaire de vérifier que chaque pays présent dans BCAC l'était également, et avec un nombre de sujets suffisant, dans CIMBA. Nous avons alors décidé d'exclure les pays où CIMBA comptait moins de 10 femmes atteintes. La Pologne et la Russie, avec respectivement 0 et 2 femmes porteuses d'une mutation de *BRCA2* dans CIMBA (Tableau 25), ont été exclues. Tous les pays ont été inclus pour l'étude BRCA1 (France, Angleterre, Canada, Israël, Pays-Bas, Italie, États-Unis,

Danemark, Finlande, Suède, Australie, Russie, Grèce, Belgique, Espagne, Pologne et Allemagne).

- *Recherche de doublons entre BCAC et CIMBA*

Les études du consortium BCAC sont des études de population générale. La prévalence des mutations *BRCA1/2* étant de 0,1 à 0,2 % dans la population générale, une faible proportion de femmes incluses dans BCAC sont porteuses d'une mutation dans les gènes *BRCA1* ou *BRCA2*. Les inclusions étant faites de façon indépendante dans les études BCAC et CIMBA, des femmes peuvent avoir été incluses dans les deux consortia. Certaines femmes de BCAC peuvent donc porter une mutation dans les gènes *BRCA1* ou *BRCA2* sans que cette information ne soit renseignée. Il a alors été nécessaire de rechercher et d'exclure de BCAC les sujets participant aux deux consortia.
Pour cela, j'ai utilisé le programme *CheckDuplicates*[n] développé par Jonathan Tyrer qui compare les génotypes des sujets deux à deux. Cette comparaison, basée sur les 35 858 SNPs non corrélés de la puce OncoArray, permet d'attribuer un « score de différence » à chaque couple de sujets BCAC/CIMBA. Plus ce score est faible, plus le lien de parenté entre les deux sujets du couple est fort. Un score de 0 signifie que les deux génotypes sont identiques et appartiennent à la même personne (ou à des jumeaux monozygotes). Ce programme m'a permis d'identifier 130 femmes porteuses d'une mutation de *BRCA1* et 83 femmes porteuses d'une mutation de *BRCA2* participant aux deux consortia. Elles ont été exclues de BCAC.

c. Sélection des témoins

Les témoins de BCAC et CIMBA ont également été analysés afin d'identifier les SNPs non indépendants des mutations *BRCA1/2* (voir page 134). Les témoins ont été sélectionnés en suivant les mêmes étapes que pour les cas. De plus, seuls les témoins n'ayant pas développé de cancer de l'ovaire ont été utilisés.

Après les différentes étapes de sélection des sujets, la population d'étude se compose de 45 881 femmes indemnes provenant de BCAC, 5 750 femmes mutées dans le gène *BRCA1* et 4 456 femmes mutées dans le gène *BRCA2*, également indemnes d'un cancer du sein et

[n] Programme développé par Jonathan Tyrer (*Centre for Cancer Genetic Epidemiology*, Université de Cambridge, UK). Non disponible en accès ouvert.

provenant de CIMBA et de 60 212 femmes atteintes d'un cancer du sein de BCAC, 7 257 femmes mutées dans le gène *BRCA1* et 5 096 femmes mutées dans le gène *BRCA2* également atteintes d'un cancer du sein (Tableau 27b).

Tableau 27 – Les différentes étapes de sélection (a) des témoins et (b) des cas

a)

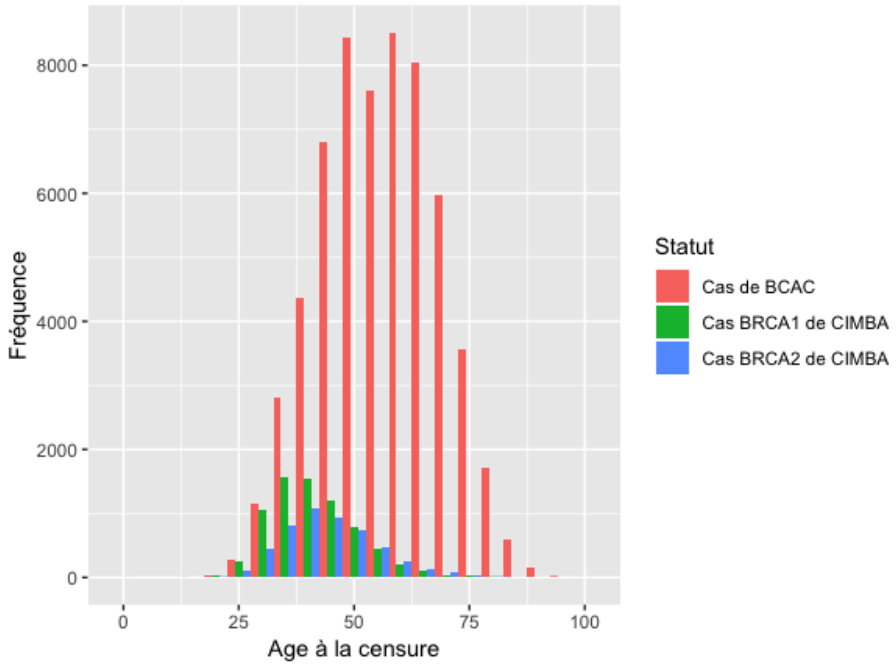
Les étapes de sélection	Analyse BRCA1		Analyse BRCA2	
	BCAC	CIMBA	BCAC	CIMBA
Témoins d'origine caucasiennes	45 888	7 190	45 888	5 046
Exclusions d'études	45 888	7 190	45 888	5 046
Exclusion des doublons	45 881	7 190	45 883	5 046
Exclusion des pays	45 881	7 190	43 549	5 046
Exclusion des cancers de l'ovaire	45 881	5 751	43 549	4 456

b)

Étapes de sélection	Analyse BRCA1		Analyse BRCA2	
	BCAC	CIMBA	BCAC	CIMBA
Cas d'origine caucasiennes	63 804	7 257	63 804	5 098
Exclusions d'études	60 326	7 257	60 326	5 098
Exclusion des doublons	60 212	7 257	60 246	5 098
Exclusion des pays	60 212	7 257	57 725	5 096
Restriction aux tumeurs RE$^-$	9 479	7 257	/	/

L'âge moyen au diagnostic de cancer du sein est de 42,2 ans (sd = 9,7) pour l'ensemble des femmes de CIMBA, de 40,9 ans (sd = 9,3) pour les porteuses d'une mutation *BRCA1* et de 44,1 ans (sd = 9,8) pour les porteuses d'une mutation de *BRCA2*. L'âge moyen au diagnostic des cas de BCAC est de 56,3 ans (sd = 12,5) (Figure 13).

Figure 13 - Distribution de l'âge au diagnostic des cas de BCAC et CIMBA



2. Sélection des SNPs : Hypothèse d'indépendance

L'analyse *case-only* permet de tester une association entre un facteur génétique G (ici, les SNPs) et un autre facteur F (ici, les mutations *BRCA1/2*) chez des femmes atteintes d'un cancer du sein. La présence d'une association entre G et F dans une population de sujets atteints suggère qu'une interaction entre G et F modifie le risque de cancer du sein. Cette analyse est sans biais sous condition que les facteurs G et F sont indépendants, c'est-à-dire P(G,F) = P(G)*P(F). Il est donc indispensable, pour obtenir des estimations non biaisées, que les mutations *BRCA1/2* et les SNPs analysés soient indépendants, c'est-à-dire qu'ils ne soient pas en déséquilibre de liaison. Pour m'en assurer, j'ai dans un premier temps exclu tous les SNPs localisés dans les gènes *BRCA1* et *BRCA2* ainsi que ceux situés à ± 500 kb de ces gènes. Cependant, le déséquilibre de liaison peut également avoir lieu entre des SNPs très éloignés, situés sur le même chromosome ou sur des chromosomes différents (on parle de déséquilibre de liaison interchromosomique[219,220]. C'est pourquoi j'ai tout d'abord effectué une analyse *control-only* comparant la fréquence des SNPs des témoins de BCAC et de CIMBA. Tous les SNPs significativement associés à la présence d'une mutation *BRCA1/2* chez les témoins, avec un seuil de signification fixé à 10^{-8}, ont été exclus de l'analyse *case-only*.

Pour l'analyse BRCA1, 196 SNPs localisés dans le gène *BRCA1* et 1 698 SNPs localisés à moins de 500 kb autour de *BRCA1* ont été exclus. De plus, les analyses *control-only* ont mis en évidence 2 070 SNPs associés aux mutations dans le gène *BRCA1* (Tableau supplémentaire 18). Ces SNPs (2 012 localisés dans le chromosome 17 et 58 dans d'autres chromosomes) ne sont donc pas indépendants des mutations dans le gène *BRCA1* et ont été exclus.
Pour l'analyse BRCA2, 204 SNPs localisés dans le gène *BRCA2* ainsi que 2 743 SNPs localisés à plus ou moins 500 kb de ce gène ont été exclus. Les analyses *control-only* ont mis en évidence 566 SNPs localisés dans le chromosome 13 et 60 SNPs localisés dans d'autres chromosomes non indépendants des mutations dans le gène *BRCA2* (Tableau supplémentaire 19). Ces SNPs ont donc également été exclus.

Au final, 9 068 301 SNPs ont été retenus pour l'analyse BRCA1 et 9 043 830 SNPs pour l'analyse BRCA2.

3. Les méthodes statistiques

a. Régression logistique

Nous avons utilisé un modèle de régression logistique pour tester l'interaction entre les SNPs et les mutations dans les gènes *BRCA1* ou *BRCA2*. Dans le cadre du design *case-only*, tous les sujets sont atteints de cancer du sein. La variable à expliquer, Y ici, n'est donc pas le cancer du sein mais la présence d'une mutation dans les gènes *BRCA1* ou *BRCA2*. La variable explicative sera, quant à elle, la présence ou non d'un SNP.

Programme logitRegress

Les analyses ont été faites séparément pour les femmes porteuses d'une mutation dans *BRCA1* et celles porteuses d'une mutation dans *BRCA*2 avec le programme *logitRegress*[o]. Pour cette analyse, le fichier phénotype contenait la liste des sujets, leur statut vis-à-vis de la mutation *BRCA1*/*2* et des variables d'intérêt supplémentaires sur lesquelles le modèle a été ajusté.

Ajustement du modèle

Le modèle de régression logistique a été ajusté sur plusieurs facteurs confondants.

Âge à la censure

La fréquence d'une mutation *BRCA1*/*2* varie fortement avec l'âge au diagnostic, avec une fréquence plus élevée chez les femmes ayant développé un cancer du sein à un jeune âge. Les cas de CIMBA sont donc en moyenne beaucoup plus jeunes que ceux de BCAC (42,5 ans vs 58,4 ans).

Pays d'origine

Comme expliqué précédemment, j'ai également ajusté le modèle sur le pays d'origine pour prendre en compte la variabilité génétique mais également des variables « culturelles » non mesurées qui existent entre les pays. En effet, il a été montré que la stratification génétique de

[o] Programme développé par Jonathan Tyrer (*Centre for Cancer Genetic Epidemiology*, Université de Cambridge, UK). Non disponible en accès ouvert. Ce programme est similaire au programme mlogit (https://ccge.medschl.cam.ac.uk/software/mlogit, consulté le 23 octobre 2019).

la population, qui correspond à une différence dans la fréquence d'un allèle entre les populations, est responsable de beaucoup de faux positifs ou faux négatifs dans les analyses d'association. Cela est surtout le cas pour les variants rares, qui ont tendance à être concentrés géographiquement puisque « récents »[221,222].

Nous avons réalisé une analyse en composante principale (ACP) pour visualiser la stratification de notre population d'étude. C'est une méthode mathématique d'analyse graphique de données qui cherche à résumer de façon pertinente un grand nombre N de variables initiales en un nombre réduit de k nouvelles variables, tout en perdant le moins d'information possible. Ces nouvelles variables sont des combinaisons linéaires des variables initiales et sont appelées composantes principales. La première composante principale est définie comme une combinaison des variables observées avec une variance maximale (c'est-à-dire une information maximale). La seconde composante principale représente une deuxième combinaison des variables d'origine avec une variance maximale, variance qui n'a pas été représentée par la première composante principale. Ces 2 composantes principales ne sont donc pas corrélées. Les autres composantes principales sont extraites de la même manière. Au total, jusqu'à N-1 composantes principales peuvent être calculées mais le but est de résumer les N variables initiales en le moins de composantes principales possible. En génétique des populations, l'ACP est un outil de visualisation extrêmement utilisé grâce à sa capacité à rendre compte de la structure des populations à l'aide d'un faible nombre d'axes principaux (2 axes sont généralement suffisants). Ces axes correspondent aux composantes principales et sont également appelés « axes de variation génétique »[223]. Comme l'ont montré Novembre *et al.*[224], ces axes peuvent également être interprétés en termes d'axes géographiques. Pour rendre compte de la structure de notre population, j'ai réalisé une ACP à partir des 35 858 SNPs non corrélés d'OncoArray grâce au logiciel *PCcalc*[p] développé par Jonathan Tyrer. Les 2 premières composantes principales générées ont été projetées sur un graphique afin de visualiser la stratification de notre population.

Composantes principales

Il existe également une variabilité génétique à l'intérieur même d'un pays. En effet, du fait des migrations et du mélange de populations, beaucoup d'individus ont plusieurs origines ethniques. De plus, il existe aussi une hétérogénéité génétique entre les différentes régions d'un même pays, comme l'ont montré Karakachoff *et al.*[225] dans leur étude comparant la

[p] Programme développé par Jonathan Tyrer (Centre for Cancer Genetic Epidemiology, Université de Cambridge, UK) (https://ccge.medschl.cam.ac.uk/software/pccalc/, consulté le 23 octobre 2019).

structure génétique de sujets provenant de l'ouest de la France. Ces différences génétiques intrinsèques à chaque pays ne sont donc pas prises en compte par l'ajustement sur le pays d'origine. Les modèles d'association ont alors été ajustés sur les composantes principales calculées lors de l'ACP pour prendre en compte la structure des populations dans sa totalité.
Au total, j'ai généré quinze composantes principales par femme. J'ai estimé le pourcentage de variance expliquée à partir des valeurs propres de chaque composante principale et également construit les diagrammes quantile-quantile (QQ-plot) afin d'évaluer la pertinence de l'ajustement additionnel sur ces composantes principales. Ces diagrammes permettent de comparer la distribution de la *p-value* obtenue pour chaque SNP en absence d'association (hypothèse nulle) à la distribution réellement observée, après prise en compte de l'ajustement. Le QQ-plot s'accompagne d'un facteur d'inflation génomique λ correspondant au ratio entre la médiane de la distribution observée et celle de la distribution attendue sous H_0. Ce facteur permet de quantifier le taux de résultats faux positifs et doit s'approcher le plus possible de 1,00 pour pouvoir considérer que le modèle est correctement ajusté. Ce facteur λ dépend de la taille de la population étudiée et doit être rapporté à une population composée de 1 000 cas et 1 000 témoins pour pouvoir être interprété. $\lambda_{1\,000}$ est calculé de la façon suivante :

$$\lambda_{1\,000} = (\lambda - 1)\left(\frac{1}{n} + \frac{1}{m}\right) * 500 + 1$$

où n et m correspondent au nombre de sujets de BCAC et CIMBA respectivement.

J'ai construit les QQ-plots et calculé les facteurs d'inflation $\lambda_{1\,000}$ pour les différents ajustements, en partant du modèle non ajusté pour arriver au modèle ajusté sur la totalité des 15 composantes principales calculées. Ils ont été construits à partir des résultats d'association obtenus pour les SNPs non corrélés du *GWAS backbone* d'OncoArray pour l'analyse BRCA1. Les QQ-plots et facteurs d'inflation ont été obtenus grâce aux packages « qqman » et « GenABEL » du logiciel R. Les décisions prises à partir de ces analyses valent pour l'ajustement des modèles de régression des analyses BRCA1 et BRCA2.

Correction pour tests multiples

Dans une étude *GWAS*, le nombre de SNPs testés est de l'ordre de 10 millions. Le nombre de résultats faux positifs attendus est de 0,05*10 000 000 = 500 000 si tous ces SNPs sont totalement indépendants. Le seuil de signification est donc corrigé par la méthode de Bonferroni :

$$\alpha^* = \alpha/\text{nombre de test} = 0{,}05/10\ 000\ 000 = 5.10^{-8}.$$

Ce seuil est utilisé habituellement dans les études GWAS. C'est un seuil particulièrement strict car il considère que tous les SNPs sont indépendants et ne prend pas en compte les blocs de déséquilibre de liaison qui existent tout le long du génome. Il entraine donc un fort taux de faux négatifs mais assure en contrepartie un nombre de faux positifs très faible. J'ai donc utilisé ce seuil $\alpha^* = 10^{-8}$ dans mes analyses.

b. Analyses *step-wise*

Parmi les SNPs statistiquement significatifs, un grand nombre se trouvent dans les mêmes régions génomiques et sont corrélés les uns aux autres. Une analyse de régression *step-wise* a donc été réalisée afin de définir la liste des SNPs non corrélés qui sont en interaction avec une mutation *BRCA1*/2. Le principe de cette étape est de conserver dans chaque région les SNPs les plus significatifs ayant un effet propre (on parle de SNPs marqueurs ou plus communément de *top SNPs*). Pour cela, les SNPs localisés à plus ou moins 500 kb les uns des autres ont été définis comme appartenant au même groupe et l'analyse *step-wise* a été effectuée pour chaque groupe.

Les étapes de l'analyse step-wise sont les suivantes :

- Le modèle de régression de départ (ajusté sur l'âge, le pays et les composantes principales) a été ajusté au fur et à mesure sur les SNPs trouvés significatifs, en commençant par le SNP le plus significatif.
- Après ce premier ajustement, les SNPs du groupe montrant une *p-value* supérieure au seuil $\alpha^* = 10^{-8}$ sont considérés comme étant dépendants et ayant le même effet que le SNP sur lequel le modèle a été ajusté. Ce SNP est un *top SNP*.
- Les SNPs toujours significatifs après cet ajustement additionnel ont alors un effet indépendant de ce *top SNP* et le modèle est ajusté une nouvelle fois sur le SNP le plus significatif parmi ces SNPs.
- Ces étapes sont répétées le nombre de fois nécessaire jusqu'à ce qu'il ne reste aucun SNP à ajouter dans le modèle.

Grâce à cette analyse *step-wise*, on obtient une liste de *top SNPs* ayant des effets indépendants les uns des autres pour chaque région. Cette analyse a été réalisée grâce à un script que j'ai écrit en langage Python.

c. Analyses de sensibilité par pays

Le pays d'origine a été pris en compte dans les analyses grâce à l'ajustement du modèle de régression sur ce facteur. Cependant, cet ajustement nous permet d'obtenir un effet moyen des SNPs chez les femmes européennes mais il ne permet pas de mettre en évidence des effets éventuellement spécifiques aux pays.
Un test d'hétérogénéité a été réalisé afin de tester si les associations trouvées significatives dans l'analyse globale ont les mêmes effets dans chaque pays présent dans la population d'étude. Un test de rapport de vraisemblance comparant les modèles de régression avec et sans interaction entre le *top SNP* et le pays a été réalisé. Le seuil de signification a été fixé à $\alpha = 0,05$. Compte tenu de la taille de certaines populations comme celle des États-Unis ou de l'Angleterre, cette analyse est très sensible et permet de mettre en évidence une hétérogénéité entre les pays, même lorsque les différences d'ORs sont faibles.

Pour les SNPs montrant une hétérogénéité significative entre les pays, nous avons effectué deux analyses supplémentaires pour étudier la différence d'effet du SNP d'un pays à l'autre :

- Une première analyse de régression stratifiée par pays m'a permis d'identifier si certains pays avec peu de sujets ne présentaient pas des effets différents, effets qui ne pourraient être mis en évidence dans l'analyse globale.
- Une deuxième analyse de régression, excluant un à un chaque pays, a été réalisée afin de voir si certaines associations n'étaient pas dues à un seul pays dont l'effet suffisamment important pour être mis en évidence dans l'analyse globale.

Ces deux analyses stratifiées ont été résumées dans des *forest plot*s et ont été réalisées à partir de scripts que j'ai programmés en R.

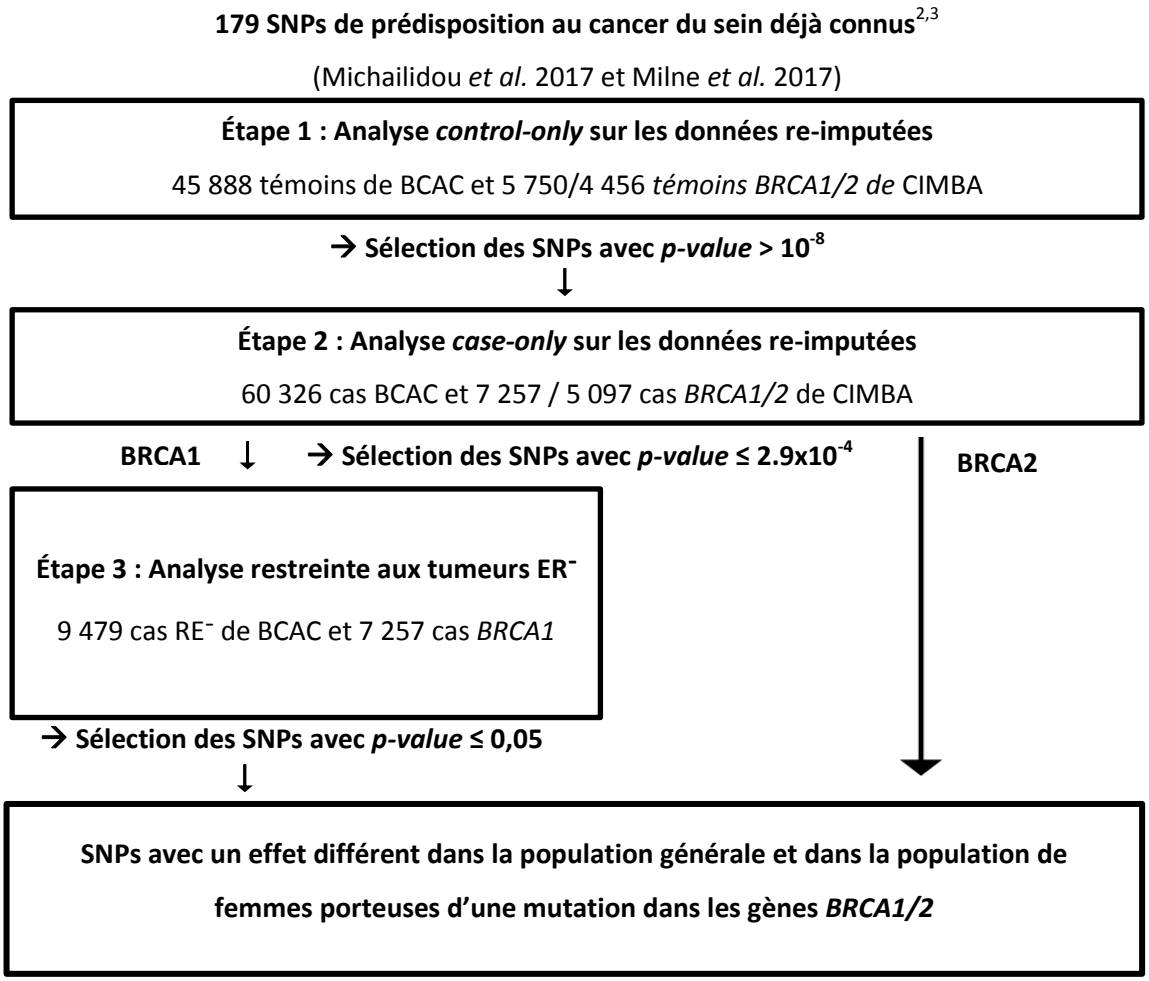
III. Stratégie d'analyse

L'analyse *case-only* a été réalisée en considérant deux catégories de SNPs selon qu'ils aient déjà été mis en évidence ou non dans les analyses faites en population générale par le consortium BCAC[141,142]. On parlera de « SNPs de prédisposition au cancer du sein déjà connus » ou de « potentiels nouveaux SNPs modificateurs ».

1. Les SNPs de prédisposition au cancer du sein déjà connus

Les régions génomiques des 179 SNPs déjà connus comme associés au risque de cancer du sein ont été ré-imputées avec les données de BCAC et CIMBA réunies. Tous les SNPs montrant une association significative au seuil $\alpha^* = 10^{-8}$ dans l'analyse *control-only* ont été exclus. Nous avons ensuite fait l'analyse *case-only* en utilisant la méthode de Bonferroni pour déterminer le seuil de signification. Celui-ci a été fixé à $\alpha^* = 0{,}05/180 = 2{,}7.10^{-4}$. Pour les femmes porteuses d'une mutation de *BRCA1*, une analyse restreinte aux tumeurs RE^- a été effectuée avec un seuil de signification fixé à $\alpha^* = 0{,}05$.

Figure 14 - Étapes d'analyse des SNPs de prédisposition au cancer du sein déjà connus dans la population générale [141,142].



2. Les potentiels nouveaux SNPs modificateurs

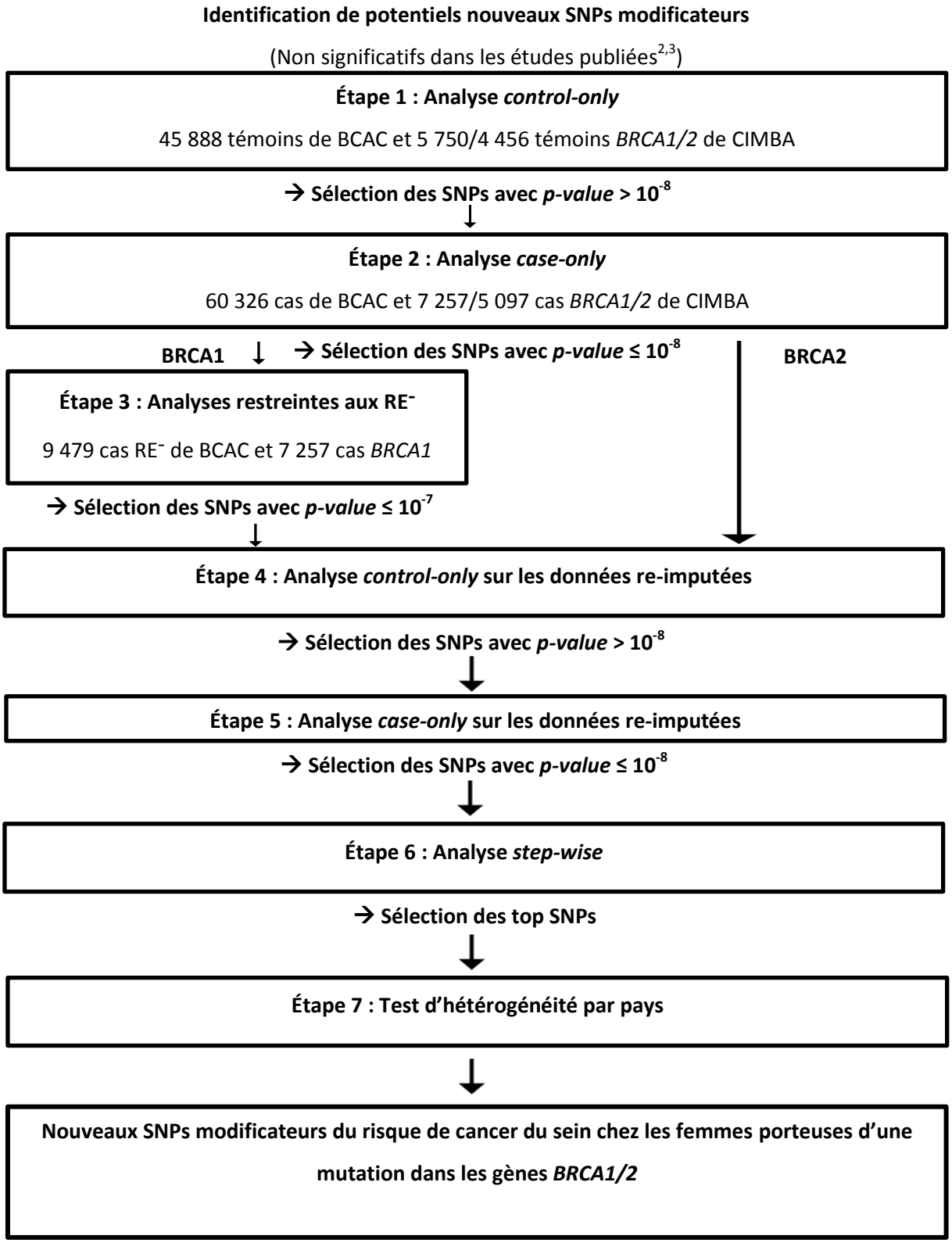
Pour les potentiels nouveaux SNPs modificateurs, les analyses ont d'abord été réalisées sur les données de CIMBA et de BCAC imputées séparément. Les étapes d'analyses sont décrites dans la Figure 15. La première étape a été d'exclure tous les SNPs trouvés significatifs dans l'analyse *control-only*. Les SNPs restants ont été analysés via l'analyse *case-only* impliquant tous les sujets de BCAC et de CIMBA sélectionnés. Le seuil de signification utilisé est le seuil *GWAS* $\alpha^* = 10^{-8}$ pour les analyses BRCA1 et BRCA2. Cependant, la proportion de tumeurs RE^- est deux fois plus importante chez les porteurs d'une mutation de *BRCA1* que chez les sujets de la population générale (Tableau 20 et Tableau 21). Nous avons donc fait une analyse supplémentaire comparant les femmes porteuses d'une mutation *BRCA1* de CIMBA

aux femmes de BCAC ayant une tumeur RE^-. Cette analyse de sensibilité a été effectuée pour :

i. s'assurer que l'association mise en évidence dans l'analyse globale n'est pas due à une différence de distribution de statut RE entre les populations de BCAC et CIMBA. La restriction aux sujets RE^- de BCAC entraine une perte de puissance pour détecter des associations. J'ai donc fixé le seuil de signification à $\alpha^* = 10^{-7}$.
ii. identifier de potentiels nouveaux SNPs modificateurs du risque de cancer du sein chez les femmes porteuses d'une mutation *BRCA1* spécifiques aux tumeurs RE^-, qui n'ont donc pas été mis en évidence dans l'analyse globale. Pour ces derniers, le seuil a été fixé à $\alpha^* = 10^{-8}$.

Pour les régions mises en évidence grâce aux étapes précédentes, une ré-imputation commune des SNPs des femmes de BCAC et de CIMBA a ensuite été réalisée et les mêmes analyses ont été répétées (exclusion des SNPs trouvés significatifs à 10^{-8} dans l'analyse *control-only*, sélection des régions significatives dans l'analyse *case-only* et restriction des analyses aux tumeurs RE^- pour les femmes porteuses d'une mutation *BRCA1*). L'analyse *step-wise* a ensuite permis de définir les meilleurs SNPs indépendants (*top SNPs*) parmi les SNPs significatifs. Pour finir, nous avons testé l'hétérogénéité par pays de chacun des *top SNPs*.

Figure 15 - Étapes d'analyse pour la recherche de potentiels nouveaux SNPs modificateurs du risque de cancer du sein chez les porteurs d'une mutation BRCA1/2[141,142].



3. Calcul des risques de cancer du sein associés aux SNPs

Avec le modèle de régression logistique utilisé dans les études d'association, on fait l'hypothèse d'un modèle d'interaction multiplicatif : l'OR estimé correspond à l'OR d'interaction entre le SNP et la mutation dans les gènes *BRCA1* ou *BRCA2* ($OR_{SNPxBRCA}$ qu'on notera OR).

Une interaction significative des « SNPs de prédisposition au cancer du sein déjà connus » avec les mutations *BRCA1/2* signifie un effet différent de ces SNPs chez les femmes porteuses d'une mutation *BRCA1/2* (CIMBA) et chez les femmes de la population générale (BCAC). Le risque relatif (RR) de cancer du sein chez les femmes *BRCA1/2* associé à ces SNPs est estimé par le produit OR x OR_{BCAC} avec :

- OR : odds-ratio obtenu dans l'analyse *case-only* ;
- OR_{BCAC} : odds-ratio obtenu dans l'étude de Michailidou *et al.*[141] pour les SNPs associés au risque de cancer du sein tous types confondus, et odds-ratio obtenu dans l'étude de Milne *et al.*[142] pour les SNPs associés aux cancers du sein RE^-.

Les potentiels nouveaux SNPs modificateurs n'ont pas été trouvés associés au risque de cancer du sein dans la population générale. Le risque relatif (RR) associé à ces SNPs dans cette population est donc fixé à 1. Par conséquent, l'effet de ces SNPs chez les femmes porteuses d'une mutation *BRCA1/2* est égal à l'OR d'interaction estimé dans l'analyse *case-only*.

4. Analyses de cartographie fine et prédiction *in silico*

L'identification de nouveaux SNPs associés au risque de développer un cancer du sein est important pour la prédiction du risque chez les porteurs d'une mutation *BRCA1/2*. Cependant, la localisation de ces SNPs ne nous permet pas d'avoir une idée sur leur effet sur le cancer. Nous avons donc réalisé une analyse fonctionnelle in silico des régions mises en évidence.
Pour cela, j'ai dans un premier temps défini les variants potentiellement causaux associés aux

régions d'intérêt grâce à une analyse conditionnelle sur la totalité des SNPs de ces régions (à la différence de l'analyse *step-wise* qui est réalisée seulement sur les SNPs significatifs). Pour chaque nouvelle région identifiée, j'ai défini un groupe de variants causaux potentiels (« *Credible Causal Variants* » CCVs) en effectuant une analyse ajustée « en cascade ». Tout d'abord, le premier groupe de CCVs défini à l'issue des analyses SNP par SNP est composé du SNP le plus significatif de la région, appelé ici SNP_1, et de tous les SNPs associés à une *p-value* égale à celle du SNP_1 à 2 ordres de grandeur près. Par exemple, si le SNP le plus significatif de la région (SNP_1) est associé à une *p-value* de 10^{-15} alors son groupe de CCVs contiendra tous les SNPs associés à une *p-value* comprise entre 10^{-15} et 10^{-13}. Ensuite, l'analyse est ajustée sur SNP_1. On définit alors un deuxième groupe de CCV si et seulement si le SNP le plus significatif (SNP_2) après ajustement sur SNP_1, est associé à une *p-value* supérieure ou égal à 10^{-6} (2 ordres de grandeur en dessous du seuil GWAS fixé à 10^{-8}). Le deuxième groupe de CCVs associés à SNP_2 est composé de tous les SNPs associés à une *p-value* égale à celle de SNP_2, à 2 ordres de grandeur près. On répète ce processus jusqu'à ce que plus aucun SNP ne soit associé à une *p-value* inférieure ou égale au seuil de 10^{-6}.

Chaque gène candidat localisé dans les régions trouvées significativement associées à la présence d'une mutation dans les gènes *BRCA1* ou *BRCA2* a été étudié en évaluant l'impact de chaque CCV sur les éléments codants ou régulateurs du génome grâce à un pipeline informatique appelé **IN**tegrated expression **QU**antitative trait and **I**n **SI**lico prediction of GWAS **T**argets (INQUISIT)[141]. Cette analyse fonctionnelle a été réalisée par Jonathan Beesley (Queensland Institute of Medical Research, Brisbane, Australie). Brièvement, les gènes sont considérés comme des cibles potentielles des CCVs s'ils ont un effet sur (1) la régulation distale du gène, (2) la régulation proximale du gène, ou (3) la séquence codante du gène. Pour cela, différentes bases de données d'expression des gènes, de régulation et d'interaction sont utilisées, comme ChIA-PET[226], PreSTIGE[227], IM-PET[228], FANTOM5[229] ou TCGA, afin d'attribuer un score à chaque catégorie de cibles potentielles (distale, proximale ou codante).

Les scores s'étendent de 0 à 8 pour la catégorie régulation distale, de 0 à 4 pour la catégorie régulation proximale et de 0 à 3 pour la catégorie séquence codante. Ces scores sont ensuite convertis en « niveau de confiance » quant à l'effet du SNP sur la fonction du gène associé :

- Niveau 1 (confiance élevée) :
    - Score de la catégorie distale > 4
    - Score de la catégorie proximale ≥ 3
    - Score de la catégorie codante > 1
- Niveau 2 (confiance modérée) :
    - Score de la catégorie distale [1-4]
    - Score de la catégorie proximale [1-3[
    - Score de la catégorie codante = 1
- Niveau 3 (confiance faible) :
    - Score de la catégorie distale = 0
    - Score de la catégorie proximale = 0
    - Score de la catégorie codante = 0

Pour les gènes associés à plusieurs scores (par exemple, les gènes associés à plusieurs SNPs ou prédits comme étant impactés dans différentes catégories), le score le plus élevé est reporté.

Résultats

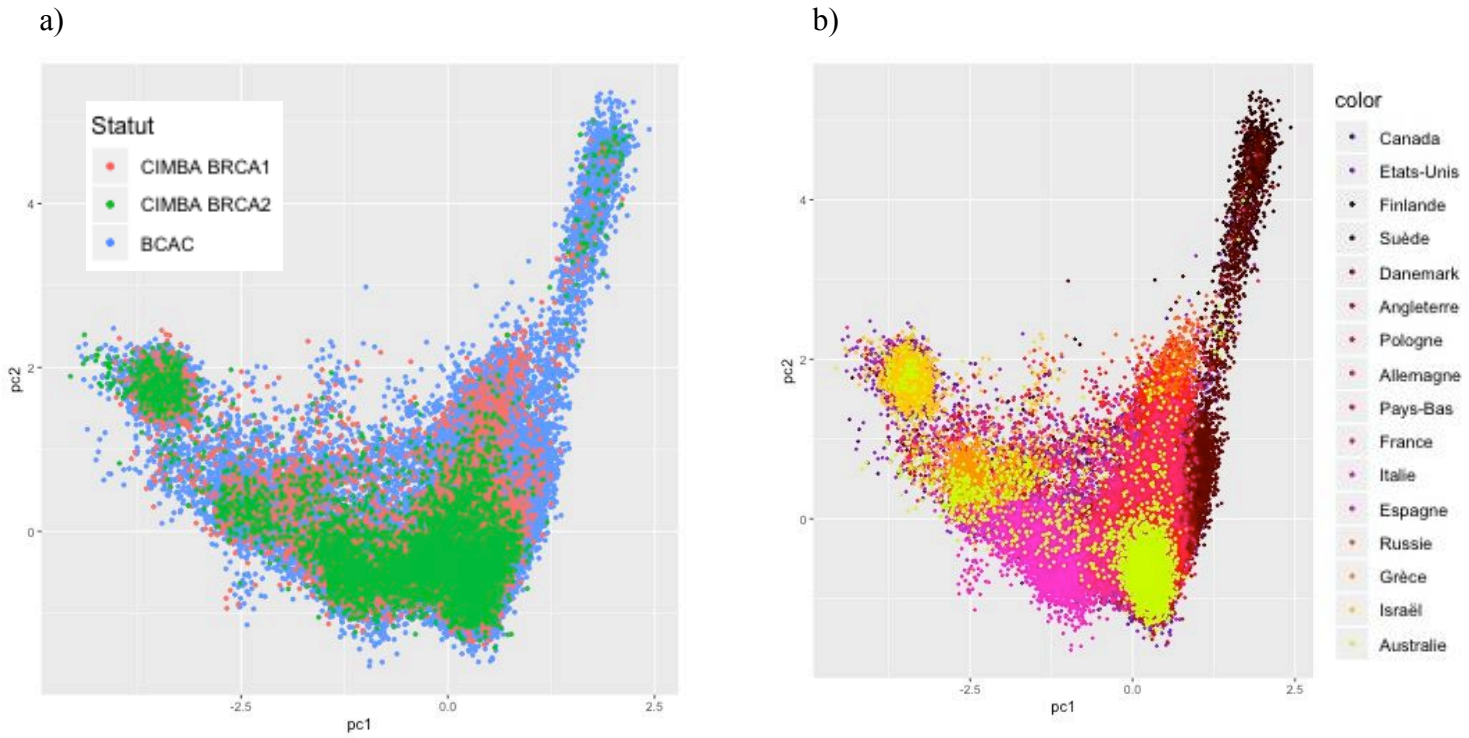
I. Variation génétique géographique

1. Variabilité génétique par pays

Les résultats de l'analyse en composante principale effectuée sur les 35 858 SNPs non corrélés de la puce OncoArray ont été utilisés pour étudier la diversité génétique entre les pays de notre échantillon d'analyse.

Seules les 15 premières composantes principales ont été calculées. Chacune d'entre elles est associée à une valeur numérique appelée valeur propre qui m'a permis d'estimer la proportion de variance expliquée par chaque composante principale. En effet, ici nous ne pouvons pas calculer le pourcentage exact de variance expliquée par chacune des composantes principales car, pour ce faire, il aurait fallu calculer la totalité des composantes principales. Le nombre de variables de départ étudiées ici étant le nombre de SNPs non corrélés, le calcul de 35 857 composantes principales aurait été beaucoup trop long et fastidieux. Cependant, à partir des valeurs propres de ces 15 premières composantes principales, nous pouvons estimer la proportion de variance de l'échantillon expliquée par chaque composante principale par rapport aux autres composantes principales en comparant leurs valeurs propres (Tableau 28). On remarque que la première composante principale explique 2,39 fois plus de variance que la deuxième et 4,98 fois plus que la troisième. Pareillement, la deuxième composante principale explique plus de 2 fois plus de variance que la troisième et la quatrième composante principale (2,08 et 2,32 respectivement). Par contre, à partir de la troisième composante principale, la proportion de variance expliquée par rapport aux composantes principales suivantes est beaucoup plus faible. Ce sont donc les deux premières composantes principales qui expliquent le plus de variabilité génétique dans notre échantillon. Nous avons alors utilisé ces deux premières composantes principales pour représenter graphiquement cette variabilité génétique, avec la première composante principale en abscisse et la deuxième en ordonnée (Figure 16).

Figure 16 - Projection des deux premières composantes principales selon le statut (a) ou le pays d'origine (b).



Le graphique a de la Figure 16 présente la projection des deux premières composantes principales selon le statut des femmes (non porteuses d'une mutation *BRCA1/2* (en bleu) de BCAC ou porteuses d'une mutation *BRCA1* (en rose) ou *BRCA2* (en vert) de CIMBA). On observe que les 3 groupes de sujets se répartissent tous de façon homogène et globalement aucun groupe ne semble avoir des caractéristiques génétiques géographiques spécifiques pour les 35 858 SNPs. Sur le graphique b de la Figure 16 les femmes sont groupées selon leur pays d'origine. Bien que les femmes de certains pays comme l'Australie (en jaune) soient réparties un peu sur tout le graphique, on remarque que celles d'autres pays, comme la Finlande ou la Suède (en marron foncé), semblent bien localisées avec des valeurs pour les deux premières composantes principales spécifiques. Lorsque l'on regarde les graphiques par pays (Figure 17), on remarque que certains pays, comme le Canada, les États-Unis ou l'Australie, ont des graphiques très proches avec des valeurs pour les deux composantes principales qui sont très dispersées. Ces 3 pays sont de très grands pays qui regroupent des sujets d'origines très diversifiées favorisant le mélange de populations. Les pays comme l'Allemagne, le Danemark, l'Angleterre ou la France montrent des nuages de points très similaires au niveau de la région du graphique où la majorité des sujets européens sont localisés. La dispersion des

nuages de points de ces pays est plus ou moins importante (par exemple, l'Allemagne vs. le Danemark). Au contraire, la plupart des pays nordiques (Finlande, Suède, Pologne) ou de l'Europe de l'est (Russie) présentent des caractéristiques génétiques bien distinctes des autres pays et leurs sujets se regroupent tous à droite du graphique plus ou moins en haut en fonction de la localisation géographique du pays. Les sujets provenant d'Israël sont aussi localisés à un endroit spécifique du graphique, et le même nuage de point est retrouvé sur les graphiques du Canada et des États-Unis, deux pays avec un grand nombre de personnes aux origines israéliennes.

Ces graphes nous confirment la présence d'une stratification génétique de la population européenne de nos données et la nécessité d'ajuster les modèles de régression logistique sur le pays d'origine.

2. Variabilité génétique restante

Afin de prendre en compte la stratification de la population dans chaque pays, stratification qui n'est pas prise en compte par l'ajustement sur le pays, nous avons également ajusté le modèle sur les composantes principales résultant de l'ACP.

Pour définir le nombre de composantes principales à intégrer dans le modèle de régression, j'ai construit 18 QQ-plots, chaque QQ-plot étant ajusté sur une variable additionnelle au QQ-plot précédent (Figure 18), et j'ai calculé les facteurs d'inflation génomique correspondants (Figure 19). On remarque que le facteur d'inflation, avec une valeur de $\lambda_{1\,000} = 1{,}184$ sans ajustement, diminue progressivement grâce aux ajustements sur l'âge et le pays d'origine ($\lambda_{1\,000} = 1{,}102$). Cependant, c'est l'ajustement sur la première composante principale qui permet de réduire le plus l'inflation en diminuant $\lambda_{1\,000}$ de 1,184 à une valeur de 1,015 jusqu'à atteindre un plateau à $\lambda_{1\,000} = 1{,}010$ à partir d'un ajustement sur les quatre premières composantes principales. Pour ne pas sur-ajuster notre modèle nous avons donc décidé de prendre en compte seulement ces quatre premières composantes principales en plus de l'âge au diagnostic et le pays.

Le modèle de régression logistique utilisé pour l'analyse *case-only* est donc le suivant :

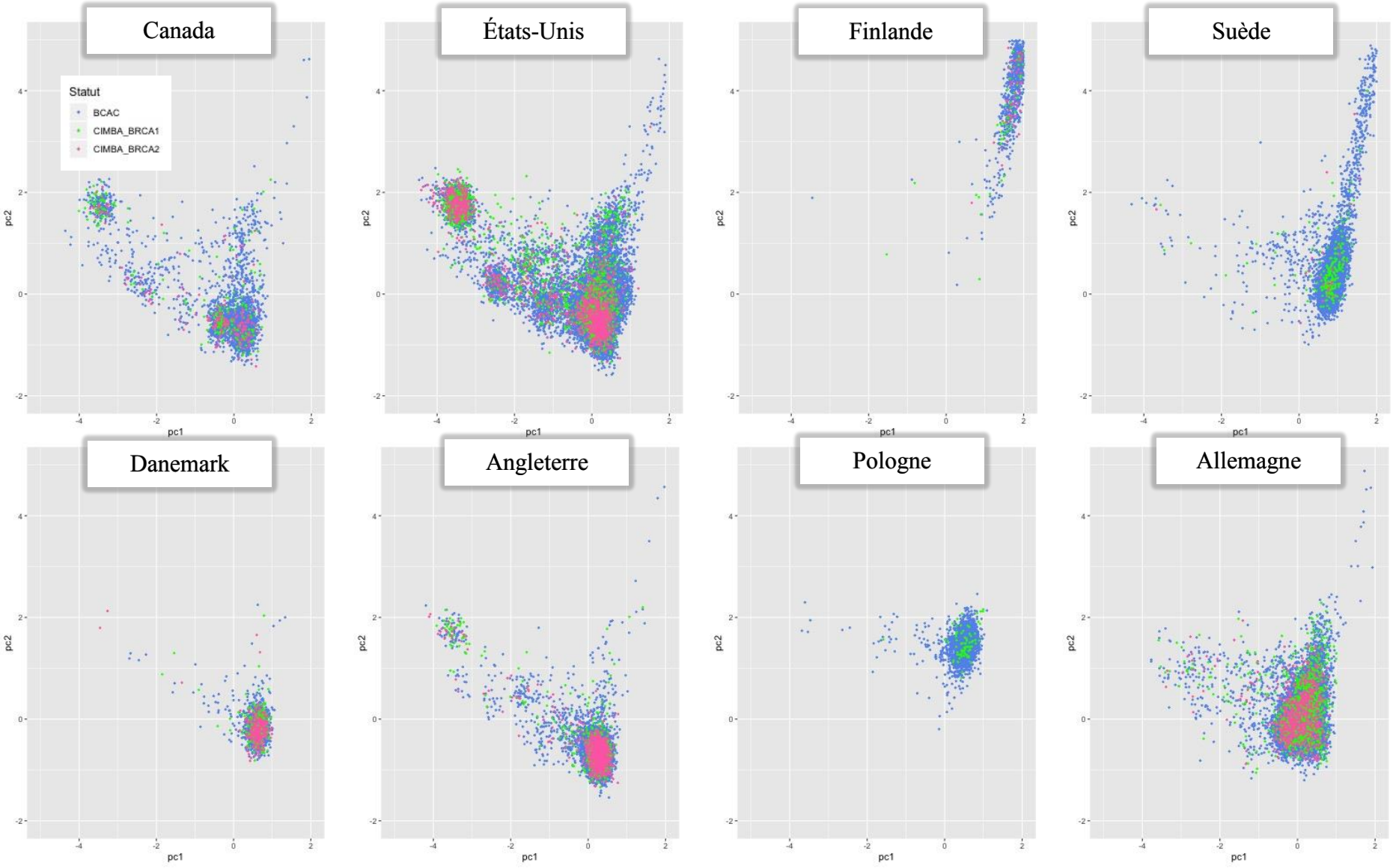
$$\boldsymbol{logit}\big(\boldsymbol{P}(\boldsymbol{BRCA1/2}|\boldsymbol{age}+\boldsymbol{pays}+\boldsymbol{PCs}+\boldsymbol{SNP})\big) = \boldsymbol{\alpha} + \boldsymbol{\beta}_{\hat{a}ge}\boldsymbol{X}_{\hat{a}ge} + \boldsymbol{\beta}_{pays}\boldsymbol{X}_{pays} + \boldsymbol{\beta}_{PC1}\boldsymbol{X}_{PC1} + \boldsymbol{\beta}_{PC2}\boldsymbol{X}_{PC2} + \boldsymbol{\beta}_{PC3}\boldsymbol{X}_{PC3} + \boldsymbol{\beta}_{PC4}\boldsymbol{X}_{PC4} + \boldsymbol{\beta}_{SNP}\boldsymbol{X}_{SNP}$$

avec $X_{\hat{a}ge}$ et $X_{PC1\text{-}4}$ des variables continues, X_{SNP} également une variable continue sous forme de dosage (voir page 84) et X_{Pays} une variable catégorielle.

Tableau 28 - Valeurs propres associées aux 15 composantes principales calculées

		Proportion de variance expliquée par rapport aux autres CP														
Composantes principales	**Valeurs propres**	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**	**PC11**	**PC12**	**PC13**	**PC14**	**PC15**
PC1	75,43	1,00	0,42	0,20	0,18	0,12	0,08	0,06	0,05	0,05	0,04	0,04	0,03	0,03	0,03	0,03
PC2	31,50	**2,39**	1,00	0,48	0,43	0,28	0,19	0,13	0,13	0,11	0,09	0,09	0,08	0,08	0,08	0,08
PC3	15,16	**4,98**	**2,08**	1,00	0,89	0,58	0,38	0,28	0,27	0,24	0,19	0,18	0,17	0,17	0,17	0,17
PC4	13,55	5,57	**2,32**	1,12	1,00	0,65	0,43	0,31	0,30	0,26	0,22	0,21	0,19	0,19	0,19	0,19
PC5	8,83	8,54	3,57	1,72	1,53	1,00	0,66	0,48	0,46	0,40	0,33	0,32	0,30	0,29	0,29	0,28
PC6	5,83	12,94	5,40	2,60	2,32	1,51	1,00	0,72	0,70	0,61	0,51	0,48	0,45	0,44	0,44	0,43
PC7	4,21	17,91	7,48	3,60	3,22	2,10	1,38	1,00	0,96	0,85	0,70	0,66	0,63	0,61	0,61	0,60
PC8	4,06	18,58	7,76	3,73	3,34	2,17	1,44	1,04	1,00	0,88	0,73	0,69	0,65	0,63	0,63	0,62
PC9	3,57	21,14	8,83	4,25	3,80	2,47	1,63	1,18	1,14	1,00	0,83	0,78	0,74	0,72	0,71	0,70
PC10	2,95	25,58	10,68	5,14	4,60	2,99	1,98	1,43	1,38	1,21	1,00	0,95	0,89	0,87	0,86	0,85
PC11	2,79	27,02	11,29	5,43	4,85	3,16	2,09	1,51	1,45	1,28	1,06	1,00	0,94	0,92	0,91	0,90
PC12	2,63	28,64	11,96	5,76	5,15	3,35	2,21	1,60	1,54	1,36	1,12	1,06	1,00	0,97	0,97	0,96
PC13	2,56	29,51	12,32	5,93	5,30	3,45	2,28	1,65	1,59	1,40	1,15	1,09	1,03	1,00	1,00	0,98
PC14	2,55	29,57	12,35	5,94	5,31	3,46	2,29	1,65	1,59	1,40	1,16	1,09	1,03	1,00	1,00	0,99
PC15	2,52	29,99	12,52	6,03	5,39	3,51	2,32	1,67	1,61	1,42	1,17	1,11	1,05	1,02	1,01	1,00

Figure 17 - Projection des deux premières composantes principales par pays.



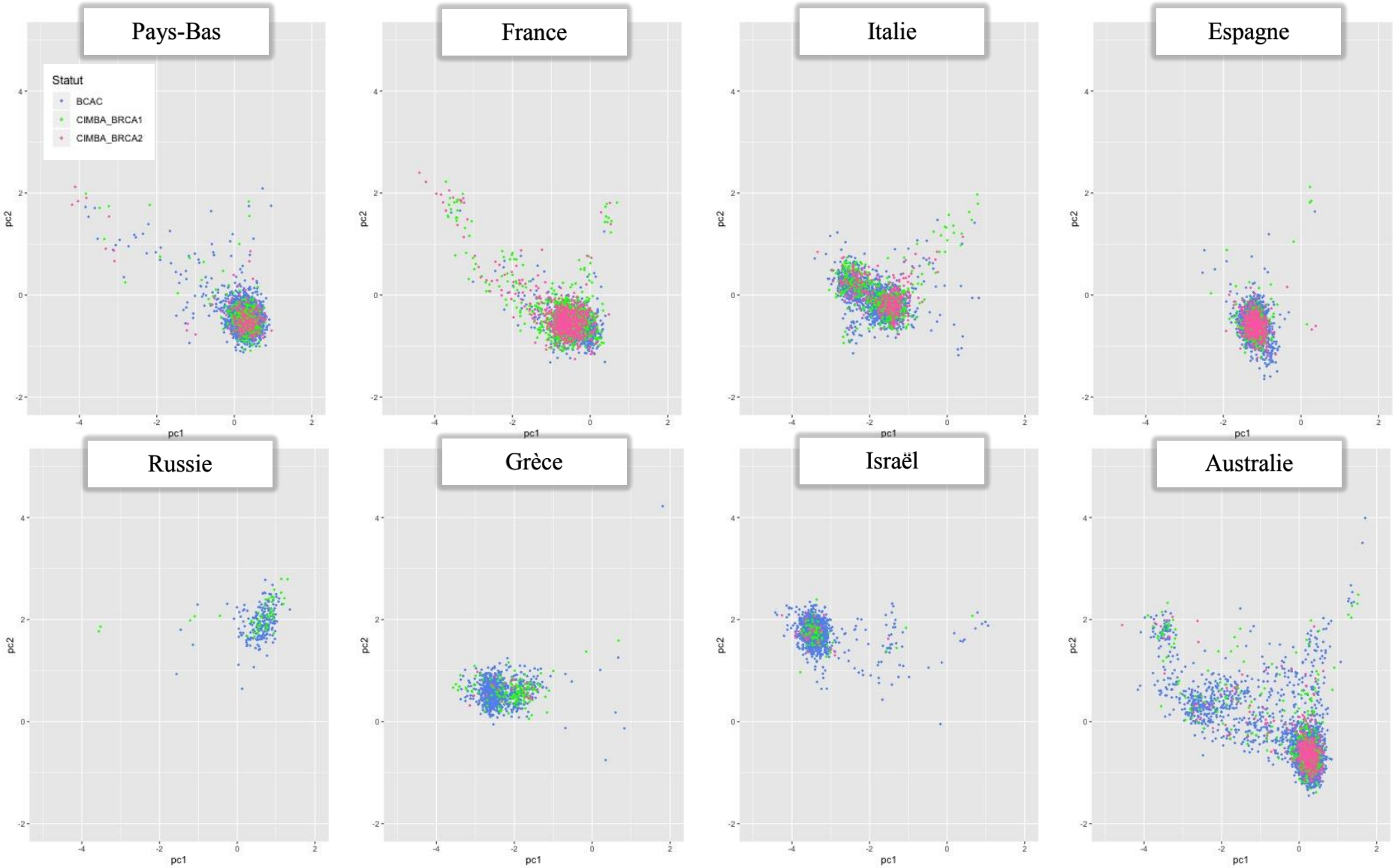
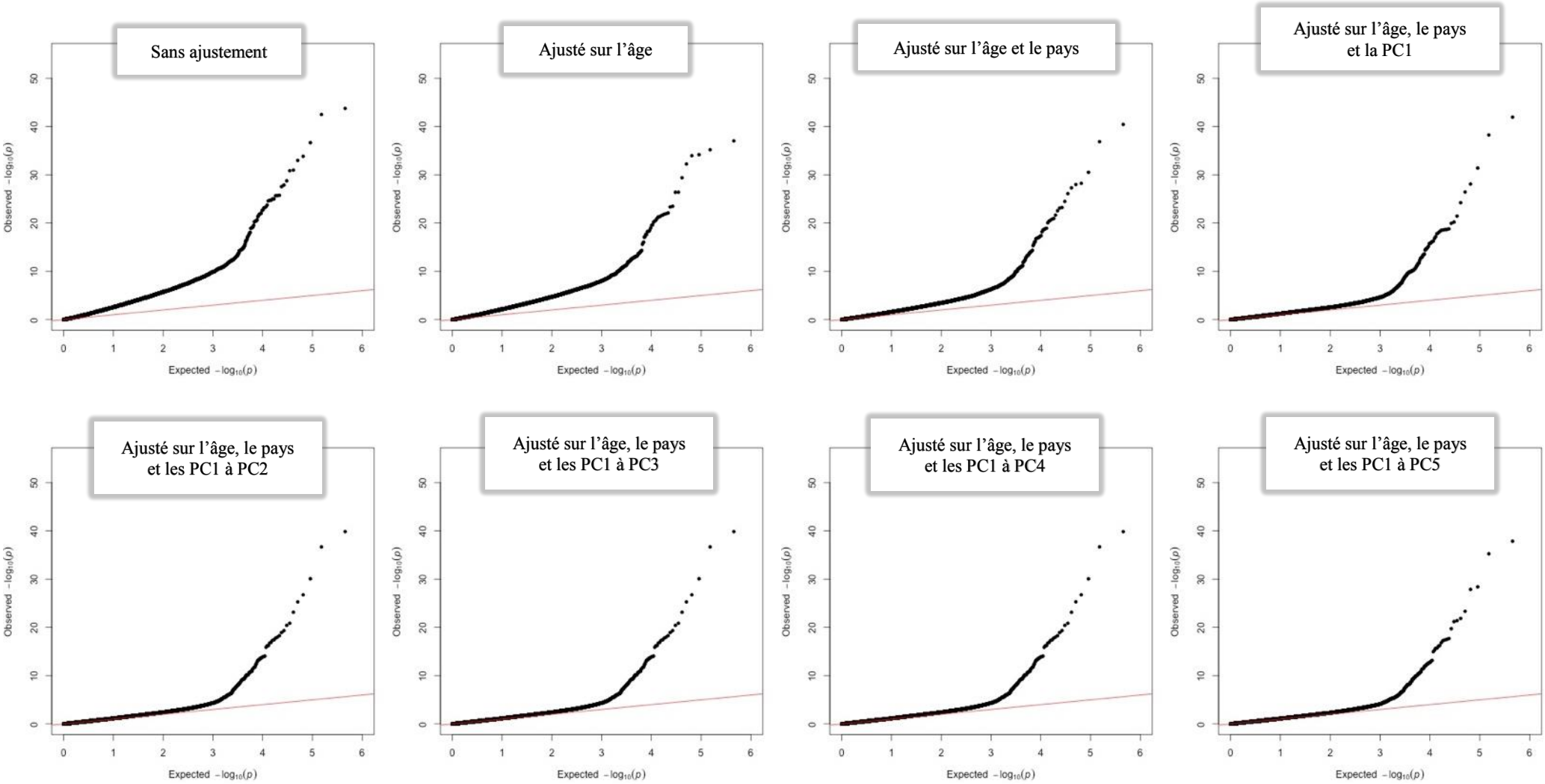

Pays-Bas
France
Italie
Espagne
Russie
Grèce
Israël
Australie
Statut
BCAC
CIMBA_BRCA1
CIMBA_BRCA2
pc1
pc2

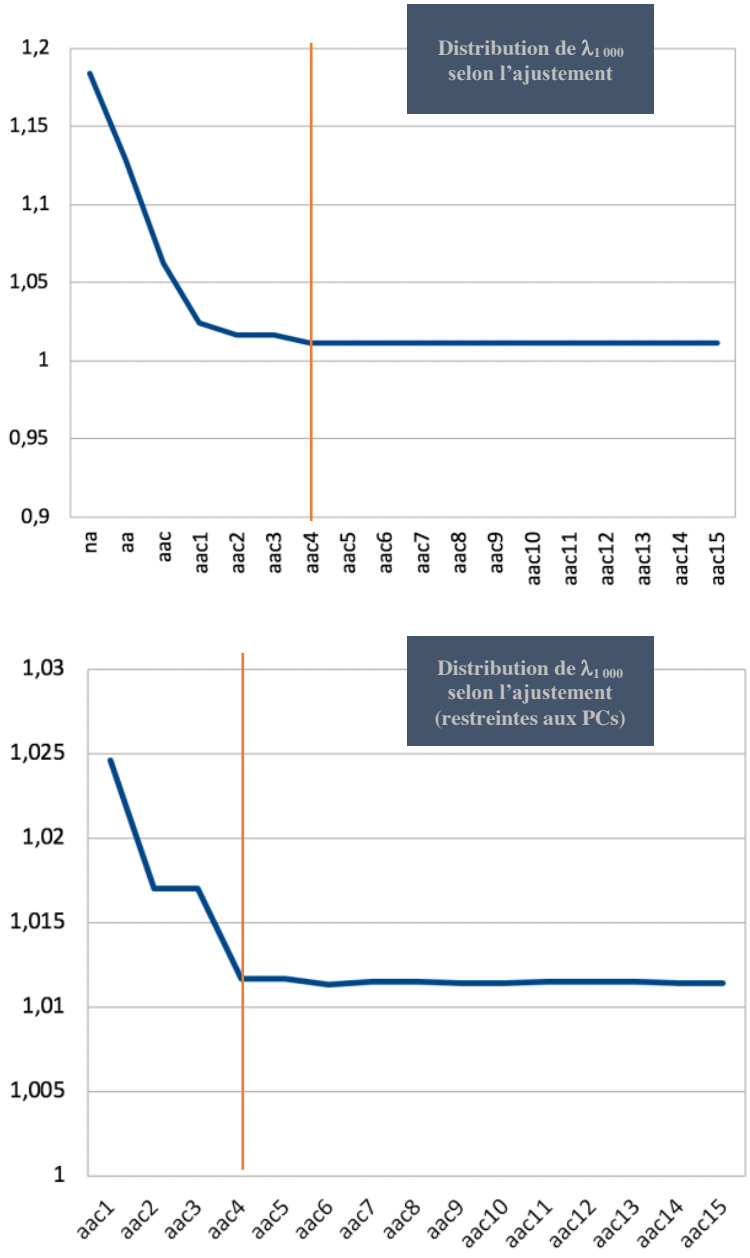
Figure 18 - QQ-plot selon les différents ajustements.



La ligne rouge montre la distribution sous l'hypothèse nulle (en absence d'association) ; en noir sont représentées les valeurs de p-value obtenues.

Figure 19 - Distribution du facteur d'inflation selon l'ajustement

Ajustement	Facteur d'inflation génomique $\lambda_{1\,000}$
Aucun	1,1849
Âge	1,1286
Âge + Pays	1,1027
Âge + Pays + PC1	1,0159
Âge + Pays + PC1-2	1,0122
Âge + Pays + PC1-3	1,0119
Âge + Pays + PC1-4	1,0100
Âge + Pays + PC1-5	1,0101
Âge + Pays + PC1-6	1,0101
Âge + Pays + PC1-7	1,0101
Âge + Pays + PC1-8	1,0101
Âge + Pays + PC1-9	1,0099
Âge + Pays + PC1-10	1,0101
Âge + Pays + PC1-11	1,0101
Âge + Pays + PC1-12	1,0101
Âge + Pays + PC1-13	1,0101
Âge + Pays + PC1-14	1,0101
Âge + Pays + PC1-15	1,0101



II. Résultats des analyses *case-only*

1. Les potentiels nouveaux SNPs modificateurs

a. Femmes porteuses d'une mutation de *BRCA1*

Une interaction significative avec les mutations dans le gène *BRCA1* a été trouvée pour 924 SNPs dans l'analyse réalisée avec tous les cas de BCAC. Les femmes porteuses d'une mutation *BRCA1* développent deux fois plus souvent une tumeur RE^- que les femmes de la population générale (38 % vs 17 %, cf. Tableau 20). J'ai ensuite restreint l'analyse aux cas de BCAC ayant une tumeur de type RE^- afin d'exclure des résultats potentiellement faux-positifs à cause de cette différence histologique. Parmi les 924 SNPs précédemment mis en évidence, 219 SNPs localisés dans 11 régions génomiques différentes restent significatifs au seuil $\alpha^* = 10^{-7}$.

Ces 11 régions ont par la suite été ré-imputées avec le logiciel IMPUTE2, conjointement avec les sujets de BCAC et CIMBA.

Dans un premier temps, il a été nécessaire de faire une nouvelle analyse *control–only* sur ces régions ré-imputées afin d'exclure celles non indépendantes des mutations *BRCA1*. Cette analyse a mis en évidence 614 SNPs (dont 613 localisés dans le chromosome 17) en déséquilibre de liaison avec les mutations dans le gène *BRCA1*. Ces SNPs ont été exclus.

L'analyse *case-only* réalisée par la suite sur les données ré-imputées a permis de confirmer l'association de 2 régions sur les 11 trouvées avant ré-imputation. Les associations trouvées dans les autres régions étaient donc probablement dues à une différence d'imputation entre les femmes de BCAC et celles de CIMBA. Dans ces 2 régions, 71 SNPs montrent une association avec le statut BRCA1 au seuil $p < 10^{-8}$ (67 SNPs localisés en 11p11.2 et 4 en 17q21.2) (Tableau supplémentaire 20). La régression step-wise réalisée sur ces 2 régions a permis d'identifier 4 top SNPs indépendants, tous imputés et associés à une MAF supérieure à 5 % (Tableau 29).

Les 3 top SNPs rs58117746, rs5820435 et rs11079012 (Tableau 29) se trouvent dans des régions introniques des gènes *KRTP4-5*, *LEPREL4* et *JUP* respectivement. Le quatrième top SNP est le SNP rs80221606 localisé en 11p11.2 dans un intron du gène *CELF1*. Ce SNP

imputé est associé à un r^2 égal à 0,76 et à une fréquence de 10 %. Ces 4 top SNPs augmentent le risque de cancer du sein dans la population des femmes porteuses d'une mutation dans le gène *BRCA1*, avec des ORs compris entre 0,86 et 1,22.

b. Femmes porteuses d'une mutation de *BRCA2*

L'analyse sur les données de BCAC et CIMBA imputées séparément montre une interaction significative avec *BRCA2* et 273 SNPs au seuil $\alpha^* = 10^{-8}$. Ces SNPs sont répartis dans 22 régions génomiques qui ont été ré-imputées avec les femmes de BCAC et CIMBA réunies.

L'analyse des données ré-imputées des témoins de BCAC et de CIMBA (analyse *control–only*) a montré que 792 SNPs, dont 787 localisés dans le chromosome 13, n'étaient pas indépendants des mutations dans le gène *BRCA2*. Ces SNPs ont été exclus. L'analyse *case-only* de ces données ré-imputées a ensuite permis de confirmer que 4 régions parmi les 22 trouvées précédemment étaient associées aux mutations *BRCA2*. Ces 4 régions comptent 102 SNPs significatifs au seuil $\alpha^* = 10^{-8}$ (Tableau supplémentaire 21).
La région associée au locus 2p14 contient 80 SNPs significatifs (Tableau supplémentaire 21) et, d'après l'analyse *step-wise*, ils sont représentés par un seul *top SNP* indépendant, rs12470785 (Tableau 30). Tous ces SNPs sont localisés dans un intron du gène *ETAA1* et augmentent modérément le risque de cancer du sein chez les porteuses d'une mutation de *BRCA2* (OR = 1,18, $IC_{95\%}$ = [1,12-1,24]). Un des SNPs significatifs de cette région, chr2_67654113_C_T, est génotypé et a un effet identique au *top SNP* (OR = 1,17, $IC_{95\%}$ = [1,12-1,23]).
Les 22 autres SNPs significatifs sont localisés dans le chromosome 13, dans 3 régions génomiques associées aux locus 13q13.1 et 13q13.2. Parmi ces 22 SNPs (Tableau supplémentaire 21), l'analyse *step-wise* a permis de mettre en évidence 3 top SNPs indépendants (Tableau 30). Deux d'entre eux sont localisés en 13q13.1 et le troisième en 13q13.2. Les ORs de ces SNPs sont compris entre 0,85 et 1,37.

Tableau 29 - Nouveaux SNPs modificateurs du risque de cancer du sein chez les femmes porteuses d'une mutation de BRCA1

Location	SNP	Chr	Position[1]	Gène	Localisation	Allèle A1	Allèle A2	r^2 [2]	Fréquence de A1[3]	OR	P	OR_{ER}[4]	P_{ER^-}[5]	HR_{CIMBA}[6]	P_{CIMBA}[7]	P_{het}[8]
11p11.2	rs80221606	11	47560211	CELF1	intron	AT	A	0,76	0,10	0,78	$1,12.10^{-10}$	0,76	$6,36.10^{-7}$	0,98	0,76	0,001
17q21.2	rs58117746	17	39305775	KRTAP4-5	codant	TGGCAGCA GCTGGGGC	T	0,60	0,39	1,18	$4,33.10^{-10}$	1,15	$7,71.10^{-5}$	1,05	0,02	$4,60.10^{-4}$
17q21.2	rs5820435	17	39961558	LEPREL4	intron	C	A	0,51	0,55	1,22	$9,55.10^{-12}$	1,17	$7,71.10^{-5}$	0,99	0,90	$1,06.10^{-8}$
17q21.2	rs11079012	17	39912880	JUP	intron	C	G	0,66	0,69	0,86	$7,06.10^{-9}$	0,85	$2,35.10^{-5}$	1,02	0,31	$1,15.10^{-7}$

1. Position sur la version hg19 du génome
2. Qualité d'imputation
3. Fréquence de l'allèle A1 chez les cas de la population générale (BCAC)
4. Odds-ratio par allèle estimé dans l'analyse case-only restreinte aux sujets de BCAC ayant une tumeur de type RE^-
5. P-value obtenue dans l'analyse case-only restreinte aux sujets de BCAC ayant un cancer du sein de type RE^-
6. Odds-ratio par allèle estimé dans l'analyse de la cohorte CIMBA (porteurs d'une mutation de BRCA1)
7. P-value obtenue dans l'analyse de la cohorte CIMBA (porteurs d'une mutation de BRCA1)
8. P-value obtenue dans l'analyse d'hétérogénéité entre pays

Tableau 30 - Nouveaux SNPs modificateurs du risque de cancer du sein chez les femmes porteuses d'une mutation de BRCA2

Location	SNP	Chr	Position[1]	Gène	Localisation	Allèle A1	Allèle A2	r^2 [2]	Fréquence A1[3]	OR	P	HR_{CIMBA}[4]	P_{CIMBA}[5]	P_{het}[6]
2p14	rs12470785	2	67634003	ETAA1	intron	A	G	0,98	0,70	1,18	$2,83.10^{-11}$	1,12	$1,69.10^{-5}$	$2,18.10^{-7}$
13q13.1	rs79183898	13	32221794	B3GALTL - RXFP2	intergénique	A	T	0,84	0,07	1,33	$2,88.10^{-10}$	1,04	0,35	$1,12.10^{-8}$
13q13.1	rs736596	13	33776506	STARD13	intron	T	G	0,66	0,09	1,37	$3,44.10^{-12}$	0,94	0,25	$4,99.10^{-11}$
13q13.2	rs4943263	13	35376357	RP11-266E6.3 - RP11-307O13.1	intergénique	C	T	0,99	0,73	0,85	$8,33.10^{-11}$	0,99	0,98	0,007

1. Position sur la version hg19 du génome
2. Qualité d'imputation
3. Fréquence de l'allèle A1 chez les cas de la population générale (BCAC)
4. Odds-ratio par allèle estimé dans l'analyse de la cohorte CIMBA (porteurs d'une mutation de BRCA2)
5. P-value obtenue dans l'analyse de la cohorte CIMBA (porteurs d'une mutation de BRCA2)
6. P-value obtenue dans l'analyse d'hétérogénéité entre pays

c. Test d'hétérogénéité par pays

Les 8 *top SNPs* mis en évidence précédemment présentent une hétérogénéité significative ($p < 0,05$) (Tableaux 13 et 14).

J'ai effectué une analyse stratifiée par pays afin de visualiser les différences d'effet de chaque SNP entre les pays. Les résultats sont résumés par des *forest plots* (Figure 20 et Figure 21). Que ce soit pour les SNPs trouvés associés au statut BRCA1 ou au statut BRCA2, les *forest plots* confirment qu'il existe une hétérogénéité entre les pays. Le SNP rs5820435 localisé sur le chromosome 17 est associé à un OR moyen de 1,22 ($IC_{95\%}$ = [1,14-1,25]) mais cette valeur fluctue entre 0,47 pour la Belgique et 2,51 pour Israël. Cependant, la stratification de la population par pays entraine une diminution importante de la taille de la population dans chaque analyse et donc une perte de précision de l'estimation. Ainsi, on remarque que l'intervalle de confiance de tous les pays est grand et que l'OR moyen obtenu dans l'analyse globale est contenu dans chacun des intervalles de confiance, excepté la Belgique, avec un OR de 0,47 et un intervalle de confiance de [0,30-0,71]. Ce SNP diminue le risque chez les belges, à l'inverse de l'association moyenne qui augmente le risque chez les femmes porteuses du SNP. Des associations spécifiques sont également retrouvées pour les autres *top SNPs* des analyses BRCA1 et BRCA2. Par exemple, le SNP rs11079012, présent sur le chromosome 17, diminue le risque de façon plus importante chez les russes (OR = 0,34, $IC_{95\%}$ = [0,17-0,67]) par rapport à l'association moyenne (OR = 0,86). Cependant, ce même SNP montre une association inverse (augmentation du risque) chez les espagnols (OR = 1,43, $IC_{95\%}$ = [1,10-1,83]). Les intervalles de confiance obtenus pour la Russie et l'Espagne ne contiennent pas l'estimation de l'OR moyen.

Ces variations peuvent être dues au fait que les analyses stratifiées donnent des estimations moins précises que l'analyse globale de par la diminution de la taille des populations étudiées. J'ai effectué de nouvelles analyses en excluant un à un chaque pays (Figure 22 et Figure 23) afin d'estimer l'impact de chaque pays dans l'estimation moyenne. On remarque alors que l'exclusion de la plupart des pays n'a pas d'impact important sur l'estimation de l'OR moyen. Cependant, pour 5 des 8 *top SNPs* (rs5820435, rs11079012, rs80221606, rs736596 et rs79183898) l'exclusion des États-Unis entraine une diminution de l'effet (l'OR tend vers 1), avec également, pour les SNPs rs79183898 et rs736596, une diminution de la force de l'association (*p-value* > 10^{-3}). Il semblerait donc que, bien qu'il ne soit pas dû seulement aux États-Unis, l'effet de ces SNPs soit plus important dans cette population que dans les autres.

Figure 20 – Analyse d'hétérogénéité par pays des nouveaux SNPs identifiés chez les porteuses d'une mutation de BRCA1.

rs5820435

Odds Ratios of association by country

Country	OR	p
australia	1.07	0.55
belgium	0.47	0.00046
canada	1.06	0.69
denmark	1.92	0.00011
finland	1.45	0.23
france	1.22	0.2
germany	1.31	5.9e-05
greece	0.92	0.72
israel	2.51	0.0013
italy	0.79	0.047
netherlands	1.07	0.6
poland	1.56	0.068
russia	2.45	0.022
spain	1.09	0.52
sweden	1.05	0.76
uk	1.2	0.018
usa	1.43	8.5e-09
All	**1.22**	**9.6e-12**

rs11079012

Odds Ratios of association by country

Country	OR	p
australia	0.81	0.043
belgium	1.55	0.032
canada	0.78	0.061
denmark	0.86	0.31
finland	0.55	0.035
france	1.05	0.76
germany	0.83	0.0024
greece	0.84	0.4
israel	0.67	0.097
italy	0.74	0.0062
netherlands	1.15	0.25
poland	1.13	0.6
russia	0.34	0.0015
spain	1.43	0.0059
sweden	0.78	0.057
uk	0.98	0.78
usa	0.73	3.8e-08
All	**0.86**	**7.1e-09**

rs58117746

Odds Ratios of association by country

Country	OR	p
australia	1.12	0.26
belgium	0.72	0.12
canada	1.02	0.9
denmark	0.88	0.38
finland	1.35	0.29
france	1.11	0.46
germany	1.14	0.043
greece	2.32	9.5e-05
israel	1.81	0.02
italy	1.2	0.1
netherlands	1.07	0.55
poland	1.1	0.67
russia	1.68	0.15
spain	0.82	0.12
sweden	1.13	0.37
uk	1.41	2.8e-06
usa	1.29	6.5e-06
All	**1.18**	**4.3e-10**

rs80221606

Odds Ratios of association by country

Country	OR	p
australia	0.68	0.011
belgium	0.58	0.099
canada	0.93	0.74
denmark	0.36	1.7e-07
finland	0.85	0.74
france	0.5	0.004
germany	0.83	0.04
greece	0.63	0.16
israel	0.6	0.22
italy	1.1	0.54
netherlands	0.59	0.0018
poland	1.15	0.67
russia	0.38	0.068
spain	1.01	0.96
sweden	0.94	0.73
uk	0.72	0.0024
usa	0.88	0.12
All	**0.78**	**1.1e-10**

Figure 21 - Analyse d'hétérogénéité par pays des nouveaux SNPs identifiés chez les porteuses d'une mutation de BRCA2.

rs736596

Odds Ratios of association by country

Country	OR	p
australia	0.92	0.6
belgium	3.55	0.0014
canada	2.05	0.00076
denmark	1.41	0.15
finland	0.53	0.16
france	1.05	0.85
germany	1.45	0.0015
greece	1.03	0.97
israel	7.69	3.8e-06
italy	0.78	0.21
netherlands	1.3	0.2
spain	1.08	0.64
sweden	2.3	0.086
uk	0.97	0.76
usa	2.01	1.7e-15
All	**1.37**	**3.4e-12**

rs4943263

Odds Ratios of association by country

Country	OR	p
australia	0.91	0.29
belgium	1.14	0.5
canada	0.72	0.01
denmark	0.67	0.0029
finland	0.92	0.69
france	0.9	0.4
germany	1.04	0.6
greece	0.53	0.061
israel	0.53	0.016
italy	0.85	0.12
netherlands	1.05	0.71
spain	0.87	0.14
sweden	1.13	0.72
uk	0.81	0.00039
usa	0.8	1.4e-05
All	**0.85**	**8.3e-11**

rs12470785

Odds Ratios of association by country

Country	OR	p
australia	1.05	0.57
belgium	0.96	0.84
canada	1.38	0.014
denmark	2.15	7.6e-09
finland	1.57	0.043
france	1.37	0.012
germany	1.24	0.00099
greece	1.43	0.34
israel	2.62	0.003
italy	0.9	0.29
netherlands	1.05	0.69
spain	0.85	0.073
sweden	0.86	0.61
uk	1.15	0.019
usa	1.24	4.1e-05
All	**1.18**	**2.8e-11**

rs79183898

Odds Ratios of association by country

Country	OR	p
australia	1.06	0.72
belgium	1.1	0.81
canada	2.05	0.0014
denmark	2.16	0.0013
finland	1.47	0.32
france	1.22	0.41
germany	0.79	0.055
greece	1.81	0.34
israel	2.94	0.019
italy	1.1	0.6
netherlands	1.15	0.53
spain	0.64	0.03
sweden	1.44	0.54
uk	1.37	0.0046
usa	1.93	8.9e-14
All	**1.33**	**2.9e-10**

Figure 22 -Analyse de sensibilité excluant chaque pays un à un pour les nouveaux SNPs identifiés chez les porteuses d'une mutation de BRCA1.

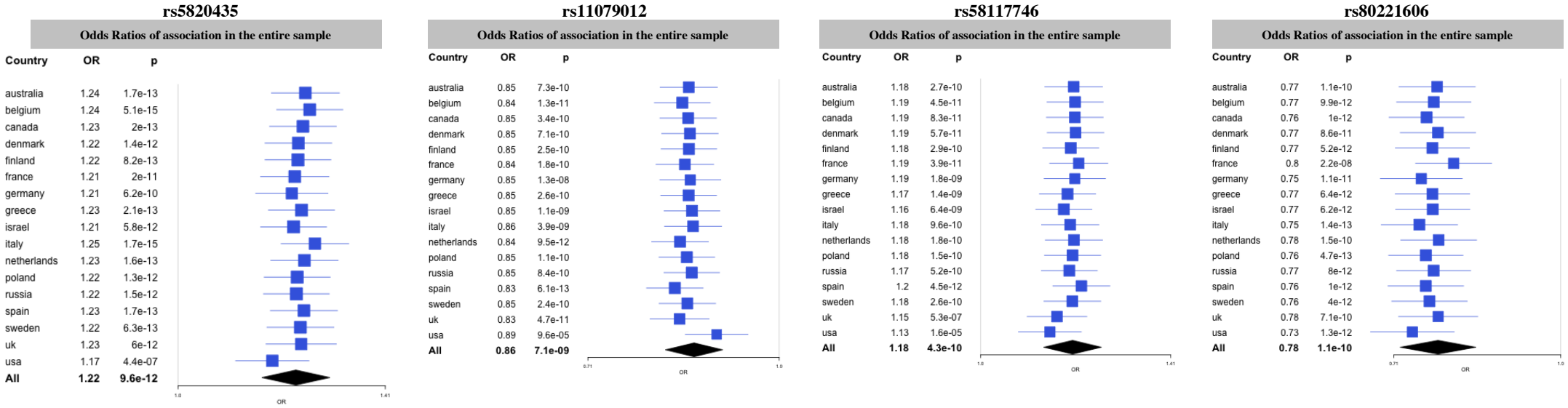
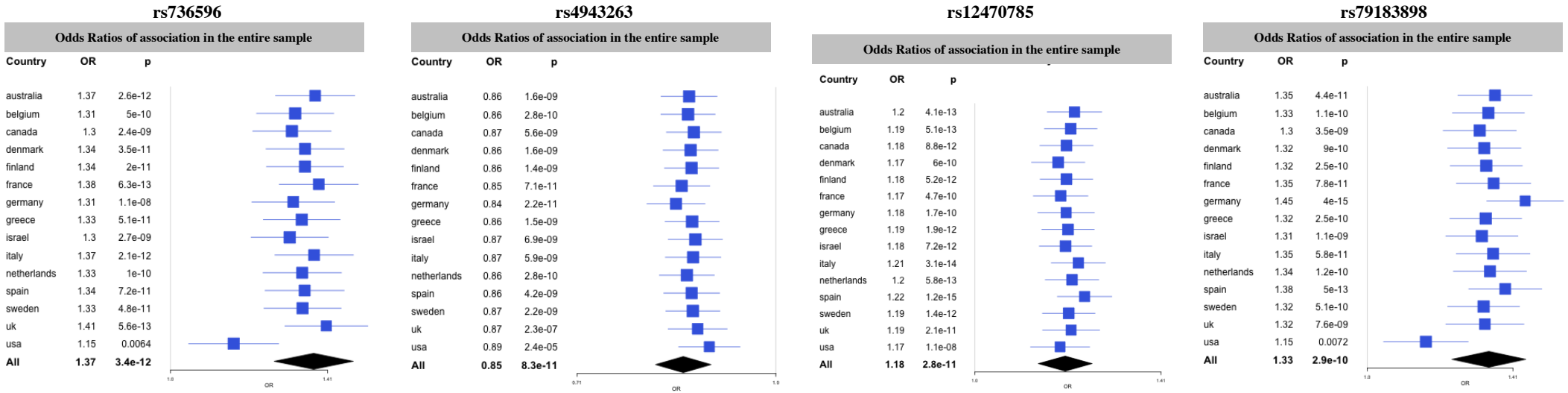


Figure 23 - Analyse de sensibilité excluant chaque pays un à un pour les nouveaux SNPs identifiés chez les porteuses d'une mutation de BRCA2.



d. Prédiction *in silico*

Les groupes de CCVs ont été recherchés pour les 2 régions précédemment trouvées associées au risque de cancer du sein chez les porteuses d'une mutation de *BRCA1*. Dans la région localisée en 11p11.2, un seul signal composé de 68 CCVs a été mis en évidence (Tableau 31). Ces 68 CCVs sont tous des SNPs imputés associés à un r^2 supérieur à 0,92 (Tableau supplémentaire 22). Dans la région localisée en 17q21.2, on trouve 9 signaux différents contenant entre 1 et 14 CCVs (Tableau 31). Deux de ces CCVs sont des SNPs génotypés et les autres ont un r^2 compris entre 0,50 et 0,98 (Tableau supplémentaire 22).

L'outil *in silico* INQUISIT[141,230] a alors été utilisé afin de définir les gènes potentiellement ciblés par ces CCVs (voir page 144). Les résultats sont résumés dans le Tableau 32. Au total, 38 gènes ont été prédits comme cibles potentielles de 6 des signaux, donc des 10 jeux de CCVs précédemment décrits. Parmi ces 38 gènes, 7 gènes localisés dans 2 régions distinctes ont été prédits avec une confiance élevée (niveau 1). Chacun de ces 7 gènes est prédit pour être régulé de façon distale par les CCVs. Parmi ces gènes cibles, les gènes *MADD*, *SPI1* et *EIF1* ont déjà été impliqués dans la biologie du cancer du sein[231–233].

De la même manière, les groupes de CCVs ont été recherchés parmi les 4 régions trouvées associées au risque de cancer du sein chez les porteuses d'une mutation *BRCA2*. Au total, 17 signaux ont été identifiés. Un premier signal composé de 78 CCVs a été mis en évidence pour la région localisée en 2p14 (Tableau 33). Un des 78 CCVs est un SNP génotypé alors que les autres sont imputés avec un r^2 entre 0,95 et 0,99 (Tableau supplémentaire 23). 12 des 17 signaux ont été mis en évidence dans la région comprenant les locus 13q13.1 (9 signaux) et 13q12.3 (3 signaux). Ces signaux sont composés de 1 à 46 CCVs. Enfin, les 4 derniers signaux ont été mis en évidence dans la région localisée en 13q13.2. Ils contiennent 3 à 40 CCVs. Parmi les CCVs mis en évidence dans le chromosome 13, 10 sont des SNPs génotypés et le r^2 des autres SNPs est supérieur à 0,58 (Tableau supplémentaire 23).

Grâce à INQUISIT, 24 gènes potentiellement ciblés par 10 des 17 signaux indépendants ont été prédits. L'un de ces gènes, *STARD13* localisé dans la région chr13:33395975-34395975, est prédit avec une confiance élevée et a déjà été reporté comme étant un suppresseur de tumeur dans les cancers du sein et du foie. Ce gène a été prédit comme une cible potentielle de 3 des 17 signaux indépendants. Tous les résultats sont présentés Tableau 34.

Tableau 31 - SNPs les plus significatifs parmi les différents signaux indépendants de l'analyse CCVs pour les femmes porteuses d'une mutation de BRCA1.

Région de *fine-mapping*[1]	Signal[2]	#CCV[3]	Locus	SNP[4]	Chr[5]	Position[6]	Gène	Localisation	Allèle A1	Allèle A2	Fréquence de A1[7]	r^2 [8]	P[9]	OR[10]	P_{ER}[11]	OR_{ER}[12]	P_{CIMBA}[13]	HR_{CIMBA}[14]
chr11:46773616-47773616	1	74	11p11.2	rs60882887	11	47475675	RAPSN, CELF1	intergénique	A	G	0,13	0,95	$2,20.10^{-10}$	0,82	$3,20.10^{-6}$	0,82	0,70	0,99
chr17:39141815-40141815	1	2	17q21.2	rs5820435	17	39961558	LEPREL4	intron	C	A	0,54	0,51	$1,10.10^{-11}$	1,22	$2,80.10^{-5}$	1,17	0,91	1,00
	2	2	17q21.2	rs7222250	17	39938469	JUP	intron	T	C	0,56	0,66	$5,50.10^{-14}$	0,81	$3,90.10^{-7}$	0,83	0,87	1,00
	3	6	17q21.2	rs9901834	17	39926811	JUP	intron	A	G	0,10	0,55	$7,20.10^{-10}$	0,72	$3,90.10^{-6}$	0,72	0,74	1,02
	4	3	17q21.2	rs58117746	17	39305775	KRTAP4-5	intron	TGGCAGCA GCTGGGGC	T	0,39	0,59	$5,50.10^{-9}$	1,17	$4,60.10^{-4}$	1,13	0,02	1,06
	5	13	17q21.2	rs2239711	17	39633317	KRT35	intron	G	A	0,70	0,92	$4,90.10^{-11}$	1,18	$2,90.10^{-4}$	1,13	0,50	1,02
	6	4	17q21.2	rs10708222	17	40137437	DNAJC7	intron	T	TA	0,17	0,59	$8,40.10^{-7}$	1,18	$6,10.10^{-4}$	1,17	0,23	0,95
	7	4	17q21.2	rs41283425	17	39925713	JUP	intron	T	C	0,06	0,53	$4,30.10^{-7}$	0,73	$1,30.10^{-5}$	0,69	0,48	0,95
	8	15	17q21.2	rs56291217	17	39858199	JUP	intron	A	AT	0,56	0,75	$6,70.10^{-8}$	1,14	$1,20.10^{-6}$	1,17	0,41	0,97
	9	1	17q21.2	rs111637825	17	40134782	DNAJC7	intron	A	G	0,06	0,89	$3,60.10^{-7}$	0,74	$3,50.10^{-4}$	0,75	0,45	0,96

1. Région significative dans l'analyse principale et utilisée pour définir les groupes de CCVs
2. Numéro du signal (le premier correspond au groupe de CCV sans aucun ajustement et les suivants à ceux trouvés par ajustement sur le SNP le plus significatif du signal précédent)
3. Nombre de CCVs dans chaque signal (SNPs avec une p-value à deux ordres de grandeur de celle du SNP le plus significatif)
4. SNP le plus significatif du signal après ajustement sur les SNPs les plus significatifs des signaux précédents (sauf pour le signal 1)
5. Chromosome
6. Position sur la version hg19 du génome
7. Fréquence de l'allèle A1 estimée dans la population générale (cas de BCAC)
8. Qualité de l'imputation
9. P-value estimée dans l'analyse *case-only* après ajustement sur le SNP le plus significatif du signal précédent (sauf pour le signal 1)
10. Odds-ratio par allèle estimé dans l'analyse *case-only* après ajustement sur le SNP le plus significatif du signal précédent (sauf pour le signal 1)
11. P-value estimée dans l'analyse *case-only* restreinte aux tumeurs RE^- de BCAC et après ajustement sur le SNP le plus significatif du signal précédent (sauf pour le signal 1)
12. Odds-ratio par allèle estimé dans l'analyse *case-only* restreinte aux tumeurs RE^- de BCAC et après ajustement sur le SNP le plus significatif du signal précédent (sauf pour le signal 1)
13. P-value estimée dans l'analyse de cohorte de CIMBA
14. Hasard-ratio par allèle estimé dans l'analyse de cohorte de CIMBA

Tableau 32 - Prédiction par INQUISIT des potentiels gènes ciblés par les CCVs trouvés dans l'analyse BRCA1

Région de *fine-mapping*[1]	Signal[2]	Gène	Biotype[3]	#SNPs signal[4]	#signal gène cible[5]	Categorie[6]	Score INQ final[7]
chr11:46773616-47773616	1	PMS2P5	pseudogene	68	1	distale	0
	1	MTCH2	protein_coding	68	1	distale	1
	1	MADD	protein_coding	68	1	distale	1
	1	PSMC3	protein_coding	68	1	distale	1
	1	RP11-750H9.5	antisense	68	1	distale	1
	1	SLC39A13	protein_coding	68	1	distale	1
	1	SPI1	protein_coding	68	1	distale	1
	8	EIF1	protein_coding	14	2	distale	1
	1	CELF1	protein_coding	68	1	distale	2
	1	FNBP4	protein_coding	68	1	distale	2
	1	KBTBD4	protein_coding	68	1	distale	2
	1	NDUFS3	protein_coding	68	1	distale	2
	1	NUP160	protein_coding	68	1	distale	2
	1	PTPMT1	protein_coding	68	1	distale	2
	1	RAPSN	protein_coding	68	1	distale	2
	1	ACP2	protein_coding	68	1	distale	2
	1	ARFGAP2	protein_coding	68	1	distale	2
	1	C1QTNF4	protein_coding	68	1	distale	2
	1	NR1H3	protein_coding	68	1	distale	2
	1	PACSIN3	protein_coding	68	1	distale	2
	1	CELF1	protein_coding	68	1	promoter	2
	1	SLC39A13	protein_coding	68	1	promoter	2
	1	MYBPC3	protein_coding	68	1	promoter	2

1. Région significative dans l'analyse principale et utilisée pour définir les groupes de CCVs
2. Numéro du signal
3. Gene/Transcript Biotypes in GENCODE & Ensembl
4. Nombre de SNPs candidats dans le signal
5. Nombre de signaux indépendants prédits pour cibler le gène correspondant
6. Catégorie d'INQUISIT : action du gène sur les régions distales, proximales (promoteur) ou codantes
7. Score INQUISIT final, avec un score de 1 qui correspond à la plus forte possibilité d'un potentiel lien entre les CCVs et le gène correspondant

Région de *fine-mapping*[1]	Signal[2]	Gène	Biotype[3]	#SNPs signal[4]	#signal gène cible[5]	Categorie[6]	Score INQ final[7]
chr17:39141815-40141815	2	JUP	protein_coding	2	3	distale	2
	3	ACLY	protein_coding	3	2	distale	2
	3	NT5C3B	protein_coding	3	2	distale	2
	3	FKBP10	protein_coding	3	2	distale	2
	3	JUP	protein_coding	3	3	distale	2
	3	NKIRAS2	protein_coding	3	1	distale	2
	5	KRT15	protein_coding	13	2	distale	2
	5	KRT36	protein_coding	13	1	distale	2
	5	KRT14	protein_coding	13	1	distale	2
	5	KRT35	protein_coding	13	1	promoter	2
	7	ACLY	protein_coding	2	2	distale	2
	7	FKBP10	protein_coding	2	2	distale	2
	7	JUP	protein_coding	2	3	distale	2
	7	NT5C3B	protein_coding	2	2	distale	2
	7	JUP	protein_coding	2	3	codante	2
	8	KRT15	protein_coding	14	2	distale	2
	8	KRT19	protein_coding	14	1	distale	2
	8	HAP1	protein_coding	14	3	distale	2
	1	RNU5E-10P	snRNA	68	1	distale	3
	1	RP11-17G12.3	antisense	68	1	distale	3
	1	MIR4487	miRNA	68	1	distale	3
	1	snoU13	snoRNA	68	1	distale	3
	1	Y_RNA	misc_RNA	68	1	distale	3
	1	AC090559.2	miRNA	68	1	distale	3
	3	EIF1	protein_coding	3	2	distale	3
	3	HAP1	protein_coding	3	3	distale	3
	7	HAP1	protein_coding	2	3	distale	3

1. Région significative dans l'analyse principale et utilisée pour définir les groupes de CCVs
2. Numéro du signal
3. Gene/Transcript Biotypes in GENCODE & Ensembl
4. Nombre de SNPs candidats dans le signal
5. Nombre de signaux indépendants prédits pour cibler le gène correspondant
6. Catégorie d'INQUISIT : action du gène sur les régions distales, proximales (promoteur) ou codantes
7. Score INQUISIT final, avec un score de 1 qui correspond à la plus forte possibilité d'un potentiel lien entre les CCVs et le gène correspondant

Tableau 33 - SNPs les plus significatifs parmi les différents signaux indépendants de l'analyse CCVs chez les femmes porteuses d'une mutation de BRCA2.

Région de fine-mapping[1]	Signal[2]	#CCV[3]	Location	SNP[4]	Chr[5]	Position[6]	Gène	Localisation	A1	A2	Fréquence de A1 [7]	r^2 [8]	p-value[9]	OR[10]	P_{CIMBA}[11]	HR_{CIMBA}[12]
chr2:67099466-68099466	1	78	2p14	rs12470785	2	67634003	ETAA1	intron	A	G	0,70	0,98	$4{,}20.10^{-11}$	1,18	$7{,}70.10^{-5}$	1,12
chr13:31015494-32515494	1	8	13q13.1	rs79183898	13	32221794	B3GALTL, RXFP2	intergénique	A	T	0,07	0,84	$1{,}10.10^{-10}$	1,33	0,36	1,04
	2	23	13q12.3	rs71434801	13	31249461	USPL1, ALOX5AP	intergénique	G	C	0,13	0,76	$3{,}40.10^{-8}$	1,22	0,84	0,99
	3	35	13q12.3	rs77197167	13	31693513	WDR95P, HSPH1	intergénique	C	T	0,09	0,76	$1{,}80.10^{-7}$	1,25	0,40	1,04
	4	7	13q12.3	rs114300732	13	31662987	WDR95P	intron	T	C	0,07	0,90	$1{,}70.10^{-8}$	0,67	0,09	1,09
	5	12	13q13.1	13:32231513:CAA:C	13	32231513	B3GALTL, RXFP2	intergénique	C	CAA	0,75	0,92	$8{,}40.10^{-7}$	1,16	0,02	0,93
	6	6	13q13.1	rs1623189	13	32232683	B3GALTL, RXFP2	intergénique	T	G	0,74	0,95	$1{,}30.10^{-31}$	0,37	0,66	0,99
chr13:33395975-34395975	1	1	13q13.1	rs736596	13	33776506	STARD13	intron	T	G	0,09	0,66	$1{,}20.10^{-12}$	1,37	0,25	0,95
	2	1	13q13.1	rs77889880	13	33776161	STARD13	intron	T	A	0,10	0,80	$3{,}00.10^{-21}$	0,51	0,02	1,12
	3	1	13q13.1	rs67776313	13	33934343	RP11-141M1.3	intron	A	AT	0,33	0,70	$7{,}70.10^{-12}$	0,81	0,46	0,98
	4	42	13q13.1	rs71196514	13	33800572	STARD13	intron	C	CT	0,38	0,67	$1{,}00.10^{-7}$	0,86	0,62	1,01
	5	52	13q13.1	rs2555605	13	33833810	STARD13	intron	T	C	0,64	1,00	$4{,}60.10^{-8}$	1,15	0,20	0,97
	6	46	13q13.1	rs74796280	13	33700860	STARD13	intron	C	A	0,06	0,96	$4{,}70.10^{-7}$	0,77	0,03	0,89
chr13:34793902-35793902	1	18	13q13.2	rs4943263	13	35376357	RP11-266E6.3, RP11-307O13.1	intergénique	C	T	0,73	0,99	$6{,}30.10^{-11}$	0,85	0,98	1,00
	2	3	13q13.2	rs2202781	13	35292372	RP11-266E6.3, RP11-307O13.1	intergénique	G	A	0,24	0,93	$3{,}10.10^{-11}$	1,20	0,60	0,98
	3	40	13q13.2	rs55675572	13	35315594	RP11-266E6.3, RP11-307O13.1	intergénique	A	T	0,40	0,77	$5{,}60.10^{-8}$	0,86	0,75	0,99
	4	21	13q13.2	rs17755120	13	35270340	RP11-266E6.3, RP11-307O13.1	intergénique	T	A	0,20	0,98	$6{,}30.10^{-7}$	0,76	0,48	0,98

1. Région significative dans l'analyse principale et utilisée pour définir les groupes de CCVs
2. Numéro du signal (le premier correspond au groupe de CCVs sans aucun ajustement et les suivants à ceux trouvés par ajustement sur le SNP le plus significatif du signal précédent)
3. Nombre de CCVs dans chaque signal
4. SNP le plus significatif du signal après ajustement sur les SNPs les plus significatifs des signaux précédents (sauf pour le signal 1)
5. Chromosome
6. Position sur la version hg19 du génome
7. Fréquence de l'allèle A1 estimée dans la population générale (cas de BCAC)
8. Qualité de l'imputation
9. P-value estimée dans l'analyse *case-only* après ajustement sur le SNP le plus significatif du signal précédent (sauf pour le signal 1)
10. Odds-ratio par allèle estimé dans l'analyse *case-only* après ajustement sur le SNP le plus significatif du signal précédent (sauf pour le signal 1)
11. P-value estimée dans l'analyse de cohorte de CIMBA
12. Hasard-ratio par allèle estimé dans l'analyse de cohorte de CIMBA

Tableau 34 - Prédiction par INQUISIT des potentiels gènes ciblés par les CCVs trouvés dans l'analyse chez les femmes porteuses d'une mutation de BRCA2

Région de *fine-mapping*[1]	Signal[2]	Gène	Biotype[3]	#SNPs signal[4]	#signal gène cible[5]	Categorie[6]	Score INQ final[7]
chr2:67099466-68099466	1	ETAA1	protein_coding	78	1	distale	2
	1	ETAA1	protein_coding	78	1	coding	2
	1	AC007392.3	lincRNA	78	1	distale	3
chr13:34793902-35793902	3	NBEA	protein_coding	39	2	distale	2
	4	NBEA	protein_coding	21	2	distale	2
chr13:33395975-34395975	6	STARD13	protein_coding	46	3	distale	1
	4	STARD13	protein_coding	3	3	distale	2
	5	STARD13	protein_coding	24	3	distale	2
	6	KL	protein_coding	46	1	distale	2
	6	PDS5B	protein_coding	46	1	distale	2
	6	RP11-141M1.1	lincRNA	46	1	distale	2
	6	RP11-37L2.1	lincRNA	46	1	distale	2
	6	STARD13-AS	processed_transcript	46	1	distale	2
	6	AL161898.1	miRNA	46	1	distale	3
	6	N4BP2L2	protein_coding	46	1	distale	3
	6	STARD13-IT1	sense_intronique	46	1	distale	3
chr13:31015494-32515494	3	B3GALTL	protein_coding	29	2	distale	0
	4	B3GALTL	protein_coding	5	2	distale	0
	2	ALOX5AP	protein_coding	7	1	distale	2
	2	HMGB1	protein_coding	7	1	distale	2
	2	USPL1	protein_coding	7	1	distale	2
	3	HSPH1	protein_coding	29	3	distale	2
	5	HSPH1	protein_coding	12	3	distale	2
	4	HSPH1	protein_coding	5	3	distale	3

1. Région significative dans l'analyse principale et utilisée pour définir les groupes de CCVs
2. Numéro du signal
3. Gene/Transcript Biotypes in GENCODE & Ensembl
4. Nombre de SNPs candidats dans le signal
5. Nombre de signaux indépendants prédits pour cibler le gène correspondant
6. Catégorie d'INQUISIT : action du gène sur les régions distales, proximales (promoteur) ou codantes
7. Score INQUISIT final, avec un score de 1 qui correspond à la plus forte possibilité d'un lien potentiel entre les CCVs et le gène correspondant

2. Les SNPs de prédisposition au cancer du sein déjà connus

Les 179 SNPs identifiés comme étant associés au cancer du sein se répartissent en 158 SNPs associés au cancer du sein tous sous-type confondus[141] et 21 SNPs associés au cancer du sein RE$^-$[142]. Parmi eux, 74 sont des SNPs imputés (soit environ 40 %). Les régions génomiques de ces derniers ont donc été ré-imputées conjointement avec les données génotypiques des femmes de CIMBA et BCAC.

a. Femmes porteuses d'une mutation dans le gène *BRCA1*

Les analyses pour les femmes porteuses d'une mutation dans le gène *BRCA1* ont également été faites en deux étapes : l'analyse *case-only* sur la population entière puis sur la population restreinte aux cas de BCAC atteintes d'un cancer du sein de type RE$^-$.

Parmi les 160 SNPs associés au risque de cancer du sein dans la population générale[141], 59 montrent une association avec le statut BRCA1 au seuil de 5 % mais seulement 16 ont une *p-value* inférieure au seuil corrigé de $2{,}7.10^{-4}$, soit 10 % (Tableau supplémentaire 24). Lorsque l'on compare les cas de CIMBA aux cas RE$^-$ de BCAC, seuls 4 SNPs, rs17426269, rs13281615, chr10_80841148_C_T et chr16_52599188_C_T, restent significatifs avec un seuil de $2{,}7.10^{-4}$, soit 2,5 % (Tableau 35). Deux autres SNPs, chr1_10566215_A_G et rs17529111, montrent une association à $2{,}7.10^{-4}$ avec le statut BRCA1 spécifiquement chez les cas de cancer du sein de type RE$^-$ (Tableau 35).
La valeur de l'association entre ces 6 SNPs et le statut BRCA1 est similaire dans les deux analyses (population entière et retreinte aux cas RE$^-$ de BCAC) et varie entre 0,85 et 1,10. Ces associations avec le statut BRCA1 suggèrent que l'effet de ces SNPs diffère de celui observé dans la population générale. Pour 3 d'entre eux, rs13281615, chr16_52599188_C_T, chr1_10566215_A_G, l'association avec le risque de cancer du sein est diminuée et tend vers 1. Pour le SNP chr10_80841148_C_T, l'association va dans la même direction que celle

de la population générale avec une amplitude plus grande et un effet plus important. Enfin, pour les 2 derniers SNPs, rs17426269 et rs17529111, l'OR estimé va dans la direction opposée à celui trouvé dans la population générale (Tableau 35).

Parmi les 22 SNPs associés au cancer du sein de type RE^- dans la population générale[142], le SNP rs66823261 localisé sur le chromosome 8 est trouvé associé dans l'analyse restreinte aux cas RE^- ($p < 2{,}9.10^{-4}$) (Tableau 35 et Tableau supplémentaire 25). L'association de ce SNP avec le cancer du sein est atténuée (Tableau 35).

b. Femmes porteuses d'une mutation dans le gène *BRCA2*

Parmi les 160 SNPs associés au risque de cancer du sein dans la population générale[141], le SNP rs11571833 est situé dans le gène *BRCA2* et a été exclu. Parmi les 159 SNPs restants, 43 sont associés au statut BRCA2 dans l'analyse *case-only* au seuil $p < 5\ \%$ (Tableau supplémentaire 26). Cependant, seulement 3 SNPs (rs62355902, rs10759243 et chr22_40876234_C_T) montrent une association significative au seuil corrigé $p < 2{,}9.10^{-4}$ avec un OR qui varie entre 0,88 et 0,89 (Tableau 36). Pour ces 3 SNPs, l'interaction avec les mutations *BRCA2* diminue l'effet associé au risque de cancer du sein (OR tend vers 1).

Tableau 35 - SNPs connus dans la population générale montrant une association dans l'analyse BRCA1

Locus	SNP	Chr	Position[1]	Gène	Allèle A1	Allèle A2	Fréquence A1[2]	r^2 [3]	OR[4]	P[5]	OR_{RE^-}[6]	P_{RE^-}[7]	OR_{BCAC}[8]	P_{BCAC}[9]	$OR_{calculé}$[10]	Variation de l'OR[11]
SNPs associés avec le cancer du sein tous sous-types confondus dans la population générale																
1p22.3	rs17426269	1	88156923	-	A	G	0,16	1	0,90	**$2,7.10^{-4}$**	0,92	0,04	1,05	$1,7.10^{-4}$	0,95	ADO
8q24.21	rs13281615	8	128355618	-	G	A	0,43	1	0,91	**$1,2.10^{-5}$**	0,94	0,04	1,11	$5,0.10^{-28}$	1,01	TV1
10q22.3	chr10_80841148_C_T	10	80841148	ZMZ1	C	T	0,6	1	1,10	**$2,2.10^{-6}$**	1,10	0,001	1,07	$1,1.10^{-14}$	1,18	AMD
16q12.1	chr16_52599188_C_T	16	52599188	TOX3	T	C	0,29	1	0,85	**$1,8.10-^{13}$**	0,91	0,003	1,23	$7,0.10^{-88}$	1,04	TV1
1p36.22	chr1_10566215_A_G	1	10566215	PEX14	G	A	0,32	1	1,07	0,001	1,12	**$1,1.10^{-4}$**	0,94	$1,8.10^{-9}$	1,05	TV1
6q14.1	rs17529111	6	82128386	-	C	T	0,23	0,96	0,92	$7,7.10^{-4}$	0,86	**$1,9.10^{-5}$**	1,02	0,04	0,88	ADO
SNPs associés avec le cancer du sein de type RE^- dans la population générale																
8p23.3	rs66823261	8	170692	RPL23AP53	C	T	0,23	0,92	-	-	0,88	**$2,3.10^{-4}$**	1,09	$5,1.10^{-9}$	0,96	TV1

1. Position sur la version hg19 du génome
2. Fréquence de l'allèle A1 chez les cas de la population générale (BCAC)
3. Qualité d'imputation
4. Odds-ratio par allèle estimé dans l'analyse case-only
5. P-value obtenue dans l'analyse case-only
6. Odds-ratio par allèle estimé dans l'analyse case-only restreinte aux sujets de BCAC ayant un cancer du sein de type RE^-
7. P-value obtenue dans l'analyse case-only restreinte aux sujets de BCAC ayant un cancer du sein de type RE^-
8. Odds-ratio par allèle estimé dans l'analyse cas-témoin de BCAC (population générale)
9. P-value obtenue dans l'analyse cas-témoin de BCAC (population générale)
10. Odds-ratio calculé ($OR \times OR_{BCAC}$)
11. Variation du risque de cancer du sein associé au SNP comparé aux estimations trouvées dans la population générale (Michailidou et al., 2017) avec TV1 = Tend Vers 1 ; AMD = Augmente dans une Même Direction ; ADO = Augmente dans une Direction Opposée

Tableau 36 - SNPs connus dans la population générale montrant une association dans l'analyse BRCA2

Locus	SNP	Chr	Position[1]	Gène	Allèle A1	Allèle A2	Fréquence A1[2]	r^2 [3]	OR[4]	P[5]	OR_{BCAC}[6]	P_{BCAC}[7]	$OR_{calculé}$[8]	Variation de l'OR[9]
5q11.2	rs62355902	5	56053723	MAP3K1	T	A	0,18	0,98	0,89	$1,1.10^{-4}$	1,18	$8,5.10^{-42}$	1,05	TV1
9q31.2	rs10759243	9	110306115	RP11-438P9.2	A	C	0,31	1	0,89	$4,6.10^{-6}$	1,06	$4,2.10^{-10}$	0,95	TV1
22q13.1	chr22_40876234_C_T	22	40876234	MKL1	C	T	0,11	1	0,88	$2,8.10^{-4}$	1,12	$5,7.10^{-16}$	0,98	TV1

1. Position sur la version hg19 du génome
2. Fréquence de l'allèle A1 chez les cas de la population générale (BCAC)
3. Qualité d'imputation
4. Odds-ratio par allèle estimé dans l'analyse case-only
5. P-value obtenue dans l'analyse case-only
6. Odds-ratio par allèle estimé dans l'analyse cas-témoin de BCAC (population générale)
7. P-value obtenue dans l'analyse cas-témoin de BCAC (population générale)
8. Odds-ratio calculé ($OR \times OR_{BCAC}$)
9. Variation du risque de cancer du sein associé au SNP comparé aux estimations trouvées dans la population générale (Michailidou et al., 2017) avec TV1 = Tend Vers 1 ; AMD = Augmente dans une Même Direction ; ADO = Augmente dans une Direction Opposée

c. Impact de la ré-imputation

Les imputations des données génotypiques des deux consortia BCAC et CIMBA ont été faites séparément. Les analyses *case-only* ont dans un premier temps été réalisées sur ces données imputées de façon indépendante. 492 SNPs sont associés significativement ($p < 10^{-8}$) au statut *BRCA1/2* répartis en 219 pour l'analyse BRCA1 et 273 pour l'analyse BRCA2. Parmi les SNPs associés au statut *BRCA1/2*, 16 SNPs sont génotypés et 476 imputés. La qualité d'imputation r^2 des SNPs imputés est supérieure à 0,5, avec un r^2 moyen égal à 0,93 (Tableau 37) et une fréquence de leur allèle mineur comprise entre 0,01 et 0,49. Seuls 59 SNPs imputés, soit 12 %, ont une fréquence inférieure à 0,05.

Tableau 37 - Qualité d'imputation des SNPs associés de façon significative avec le statut BRCA1 ou BRCA2 dans les analyses case-only réalisées sur les données imputées séparément.

	SNPs imputés – r^2						**SNPs Génotypés**
Analyse	**[0,5–0,6[**	**[0,6–0,7[**	**[0,7–0,8[**	**[0,8–0,9[**	**[0,90–0,95[**	**[0,95–1[**	
BRCA1	7 (1,47 %)	0 (0,00 %)	2 (0,42 %)	12 (2,52 %)	36 (7,56 %)	155 (32,56 %)	7
BRCA2	19 (4,00 %)	5 (1,05 %)	9 (1,89 %)	17 (3,57 %)	40 (8,41 %)	174 (36,55 %)	9

Ces 476 SNPs imputés sont localisés dans 33 régions différentes du génome. Toutes ces régions (± 500 kb du SNP le plus significatif de la région) ont été ré-imputées. 92 % de ces 476 SNPs imputés précédemment trouvés ont une *p-value* plus élevée après imputation (Figure 24a) et seuls 30 % (145 SNPs) restent significatifs au seuil $p < 10^{-8}$. Ces 145 SNPs sont situés dans 6 des 33 régions précédemment trouvées.

Tableau 38 - Distribution des SNPs après ré-imputation en fonction de la p-value et de la fréquence allélique (MAF).

MAF	***p-value* après ré-imputation**	
	$\leq 10^{-8}$	**$> 10^{-8}$**
< 0,05	9 (15,2 %)	50 (84,7 %)
≥ 0,05	136 (32,6 %)	281 (67,4 %)

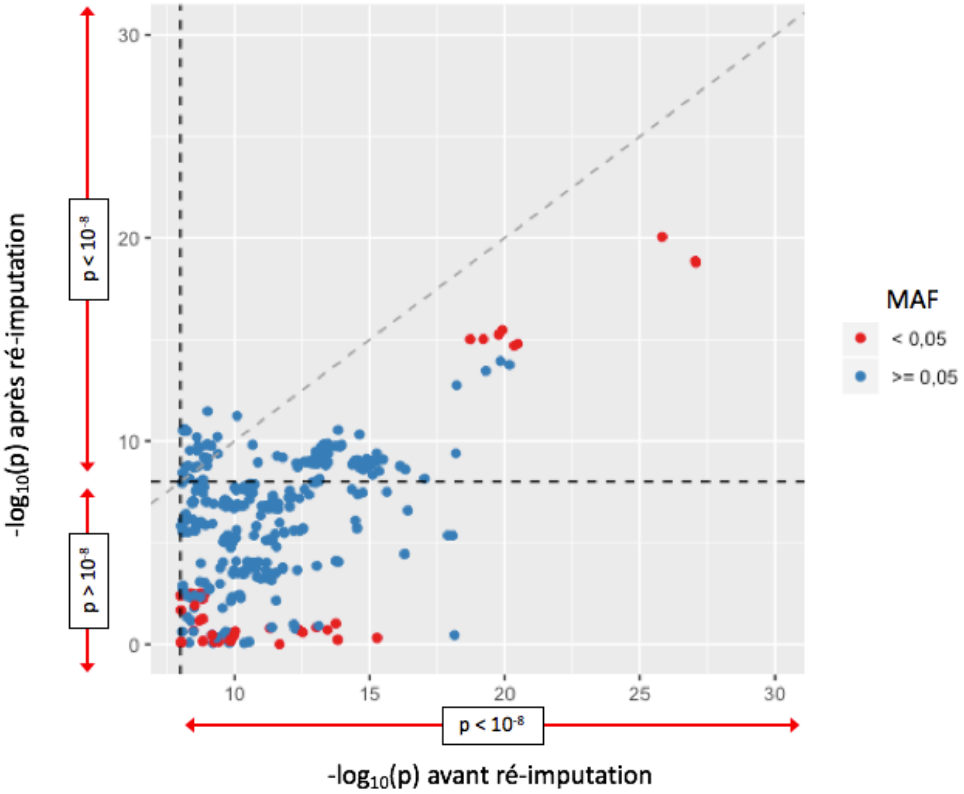
15,2 % des SNPs rares (fréquence inférieure à 5 %) et 32,6 % des SNPs fréquents restent significatifs après ré-imputation (Tableau 38 et Figure 24a). Comme prévisible, les plus grandes distorsions dans les estimations du risque associé au SNP (ORs) concernent les SNPs rares. Les SNPs plus fréquents sont associés à des ORs plus faibles dont les variations d'estimation avant et après ré-imputation sont, de fait, plus petites (Figure 24b).

La Figure 25a montre la distribution des différences entre la p-value obtenue avant et après ré–imputation avec, en abscisse, la différence d'ordre de grandeur entre les 2 p-values. Par exemple, pour 2 p-values égales à 10^{-10} et 10^{-6} avant et après ré-imputation, la différence d'ordre de grandeur sera de 4. Seuls 24 SNPs ont une p-value inférieure après ré–imputation. Parmi les 452 SNPs restants, 329 ne sont plus significatifs après ré-imputation (en rouge sur la Figure 25a). De plus, comme on peut le voir sur la Figure 25b, bien que la différence entre les ORs avant et après ré-imputation ne soit pas très importante du fait de l'effet faible associé à chacun des SNPs, 90 % des SNPs sont associés à un OR plus faible après ré-imputation. Seuls 16 SNPs ont un OR identique à 10^{-2} près, entre les deux analyses.

Outre les 145 SNPs significatifs avant et après ré-imputation, 92 nouveaux SNPs ont été mis en évidence par les analyses sur les données ré-imputées. Ces nouveaux SNPs étaient donc des faux-négatifs induits par une imputation différente des données. J'ai ré-imputé seulement 33 régions de 1 Mb, ce qui représente 0,01 % du génome. De nouveaux SNPs, et potentiellement de nouvelles régions non ré-imputées pourraient donc être associées au statut BRCA1 ou BRCA2. Il serait nécessaire de ré-imputer le génome entier des cas BCAC et de CIMBA simultanément afin d'augmenter nos chances de trouver de nouveaux SNPs.

Figure 24 - Comparaison avant et après ré-imputation a) des P-values et b) des ORs.

a)



b)

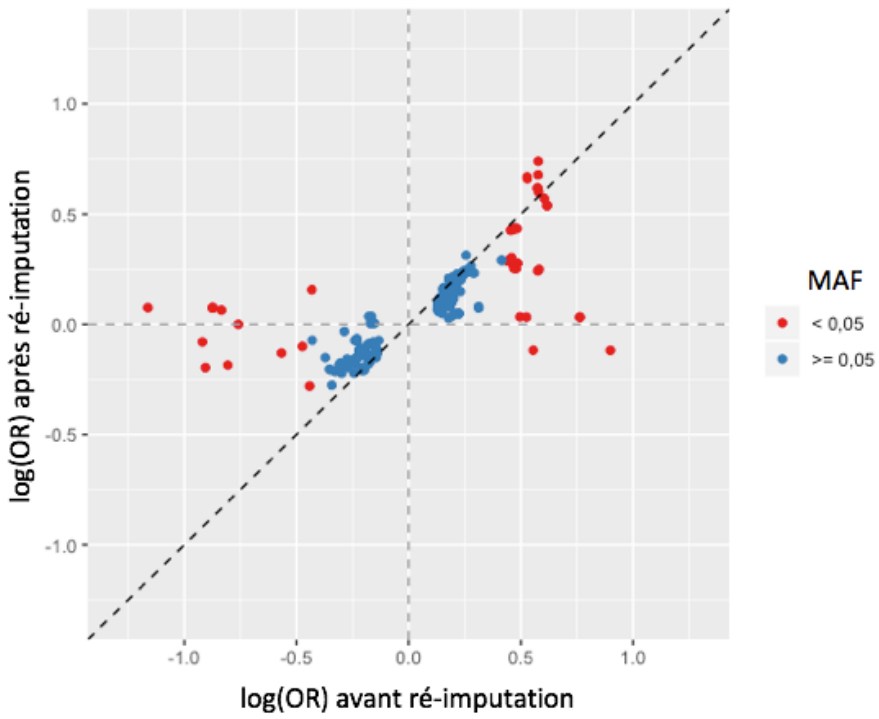
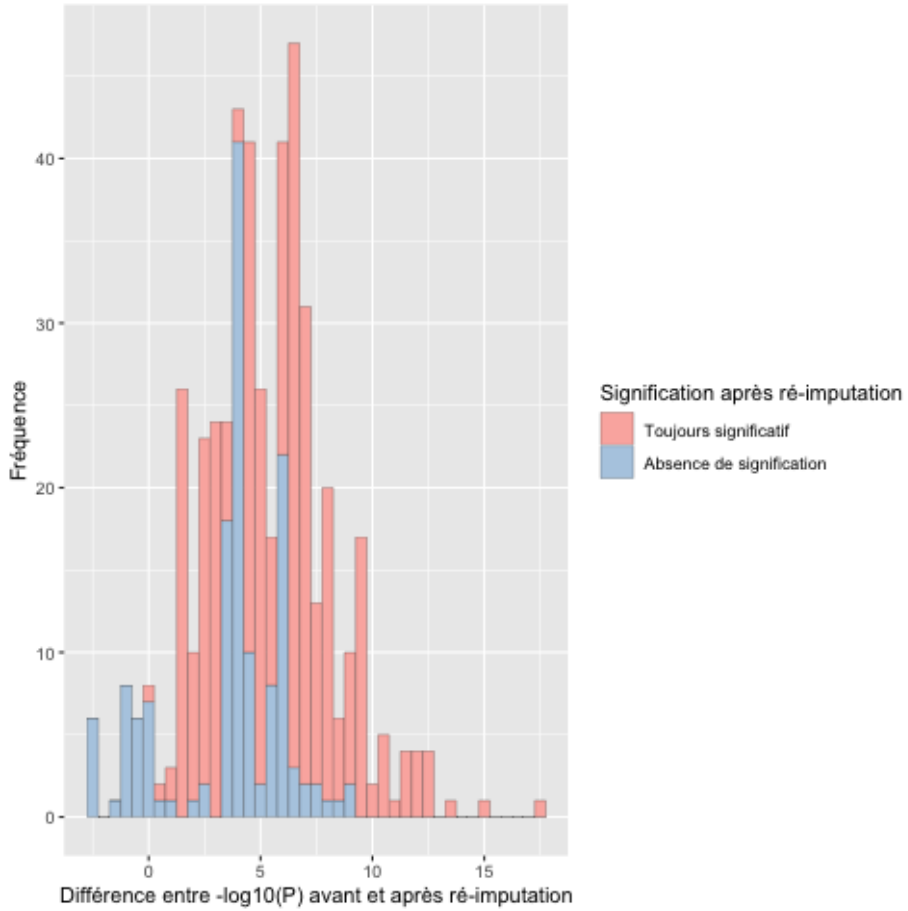
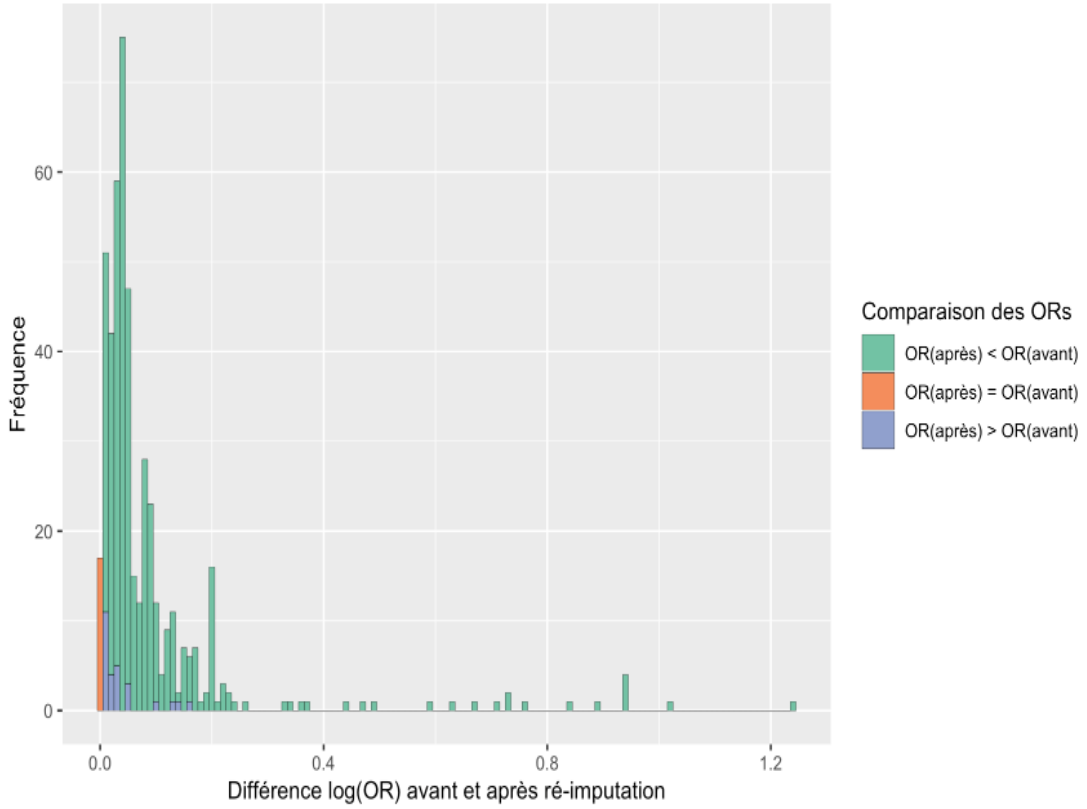


Figure 25 - Histogramme des différences observées avant et après imputation des a) p-values et b) ORs

a)



b)



Discussion

Nous avons développé une nouvelle stratégie d'analyse utilisant un design *case-only* dans le but d'identifier de nouveaux facteurs génétiques modificateurs du risque de cancer du sein chez les femmes porteuses d'une mutation dans les gènes *BRCA1* ou *BRCA2* mais également d'estimer l'effet, chez ces femmes, des SNPs déjà connus dans la population générale. Cette stratégie est basée sur les données *GWAS* des femmes de la population générale atteintes d'un cancer du sein provenant du consortium BCAC et celles porteuses d'une mutation *BRCA1/2* provenant du consortium CIMBA. Elle mène à une plus grande homogénéité de la population étudiée qu'un design cas-témoin classique. Cela permet alors d'augmenter la puissance statistique nécessaire à la détection de nouvelles associations et d'obtenir de meilleures estimations chez les porteurs d'une mutation *BRCA1/2* pour les SNPs déjà connus[234].

Nous avons montré que pour 96 % des SNPs identifiés par les études *GWAS* réalisées en population générale[132,137,138,141,158,159,161,163,165,235–256], il n'a pas été possible de mettre en évidence une interaction significative entre la présence d'une mutation *BRCA1/2* et des SNPs. Ces résultats suggèrent que ces SNPs sont également associés au risque de cancer du sein dans la population des femmes porteuses d'une mutation *BRCA1/2* avec des effets similaires à ceux trouvés dans la population générale. Cependant, il existe une interaction significative entre 10 SNPs connus comme étant associés au cancer du sein dans la population générale et les mutations *BRCA1/2,* suggérant que ces SNPs ont un effet différent chez les sujets porteurs d'une mutation *BRCA1/2* et dans la population générale (7 SNPs, rs17426269, rs13281615, chr10_80841148_C_T, chr16_52599188_C_T, chr1_10566215_C_T, rs17529111 et rs66823261, pour les porteurs d'une mutation *BRCA1* et 3, rs62355902, rs10759243 et chr22_40876234_C_T, pour les porteurs d'une mutation *BRCA2*). Les interactions observées pour ces SNPs suggèrent que 7 d'entre eux n'ont pas d'effet sur le risque de cancer du sein chez les porteurs d'une mutation *BRCA1/2*. Pour 2 SNPS, rs17426269 et rs17529111, l'interaction avec les mutations *BRCA1/2* augmente l'effet du SNP et pour le dernier SNP, chr10_80841148_C_T, l'interaction entraine un changement de direction de l'association par rapport à celle observée dans la population générale.

Cette observation – montrant que 7 des 10 SNPs identifiés comme agissant en interaction avec les mutations *BRCA1/2* sont en fait des SNPs non modificateurs du risque de cancer du

sein dans la population des porteurs d'une mutation de *BRCA1/2* – est en adéquation avec le fait que le PRS_{BCAC} prédit moins bien dans cette population[164].

Nous avons identifié 8 nouveaux SNPs indépendants associés au risque de cancer du sein chez les porteurs d'une mutation (rs80221606 en 11p11.2 et rs58117746, rs5820435 et rs11079012 en 17q21.2 pour les porteurs d'une mutation de *BRCA1* et rs12470785 en 2p14, rs79183898 et rs736596 en 13q13.1 et rs4943263 en 13q13.2 pour les porteurs d'une mutation de *BRCA2*). L'OR de ces SNPs varie entre 0,85 et 1,37 pour les porteurs d'une mutation *BRCA2* et de 0,78 à 1,22 pour les porteurs d'une mutation *BRCA1*. Ces SNPs n'ont pas été reportés dans les études précédentes[157–164,166,169,243,257–259]. Pour 5 de ces SNPs, la valeur de l'OR va dans la même direction que l'estimation ponctuelle observée dans l'analyse de cohorte de CIMBA. Deux de ces 5 SNPs, montrent une association à $p = 0{,}022$ et $p = 7{,}7.10^{-5}$ dans CIMBA (rs80221606 dans le gène *CELF1* pour les porteurs d'une mutation de *BRCA1* et rs12470785 dans *ETAA1* pour les porteurs d'une mutation de *BRCA2*) (Tableau 29 et Tableau 30). Pour les 3 autres SNPs, rs5820435 et rs11079012 localisés en 17q21.2 et rs736596 localisé en 13q13.1, l'association observée dans CIMBA n'est pas en adéquation avec les interactions observées dans l'analyse *case-only*. Ces 3 SNPs montrent également une hétérogénéité par pays, avec notamment des associations peu robustes pour certains pays comme la Belgique pour rs5820435 et la Russie pour rs11079012.

L'analyse fonctionnelle réalisée avec le pipeline INQUISIT nous a permis de prédire des gènes cibles potentiels aux SNPs localisés dans les régions trouvées associées au risque de cancer du sein. Pour les SNPs associés à *BRCA1*, 7 gènes localisés dans la région 11p11.2 du SNP rs60882887 ont été prédits avec un score de confiance élevé dont les gènes *MADD*, *SP11* et *EIF1* qui ont déjà été impliqués dans la biologie du cancer du sein[231–233]. Cependant, aucun gène cible n'a été prédit avec un score de confiance élevé pour la région 17q21.2. Cela pourrait confirmer nos suspicions d'artefacts pour les SNPs rs5820435 et rs11079012.

Seul 1 gène, *STARD13*, a été prédit comme une cible potentielle de 3 signaux localisés dans la région 13q13.1 du SNP rs736596 associé aux mutations *BRCA2*. Ce gène a été reporté comme gène suppresseur de tumeur dans le cancer du sein et ayant un rôle dans la prolifération cellulaire et le développement des métastases[260]. Bien qu'aucune association ne soit trouvée dans CIMBA pour ce SNP et qu'une hétérogénéité par pays significative ait été mise en évidence, les résultats d'INQUISIT montrent que les SNPs localisés dans cette région

pourraient être de potentiels facteurs modificateurs du risque de cancer du sein chez les femmes porteuses d'une mutation *BRCA2*.

Le gène *ETAA1* n'est associé qu'à un score INQUISIT de 2. Cependant, les estimations obtenues pour les SNPs localisés dans ce gène vont dans la même direction que ceux estimés dans la cohorte de CIMBA et les résultats montrent une association proche de la signification ($p \approx 10^{-4}$). De plus, cette région est représentée par une centaine de SNPs corrélés les uns aux autres dont un est un SNP génotypé. Malgré le mauvais score prédit par INQUISIT, nous gardons cette région comme faisant partie des régions dont l'association est la plus probable, avec celles des gènes *MADD*, *SP11*, *EIF1*, *MTCH2*, *PSMC3*, *SLC39A13*, *RP11-750H9.5* et *STARD13*.

Les prédictions *in sillico* aident à explorer un éventuel lien de causalité entre les SNPs trouvés associés au cancer du sein et leur fonction et rendent ainsi plus plausibles les associations mises en évidence. Cependant, ces analyses sont basées sur les connaissances actuelles du génome (exome, interactome, « régulatome »…) et donc incomplètes. Il semble alors déraisonnable d'exclure des SNPs montrant une forte association avec le cancer du sein sur le seul fait qu'un lien de causalité n'a pas pu être démontré. En effet, ces SNPs peuvent contribuer de façon significative à améliorer le pouvoir prédictif du PRS, que ce soit les SNPs déjà connus dans la population générale (BCAC) qui ont un effet différent dans la population des femmes porteuses d'une mutation *BRCA1/2* ou les nouveaux SNPs identifiés.
Ces deux catégories de SNPs interagissent avec les mutations *BRCA1/2*. Par quel mécanisme biologique ces gènes interagissent-ils avec les gènes *BRCA1/2* mutés ? Cette question devra être abordée par des études fonctionnelles *in vivo* ou *in sillico* pour tenter d'expliquer ces interactions statistiques.

L'analyse *case-only* repose sur l'hypothèse d'indépendance entre la présence d'une mutation *BRCA1/2* et le SNP d'intérêt[261] (pas de déséquilibre de liaison, DL). L'analyse *control-only* nous a donc permis d'exclure environ 2 000 SNPs associés à la présence d'une mutation chez les cas et chez les témoins (Tableaux supplémentaires 18 et 19). Les SNPs en DL avec les mutations *BRCA1/2* ont été mis en évidence avec un seuil *GWAS* très strict fixé à priori à $p < 10^{-8}$ dans l'analyse *control-only*. Ce seuil très bas a pu générer des faux-négatifs, c'est-à-dire des SNPs en DL avec les mutations *BRCA1/2* qui n'ont donc pas été exclus. Cependant, 33 % des témoins porteurs d'une mutation de *BRCA1/2* sont apparentés à au moins un cas

porteur d'une mutation de *BRCA1/2* du consortium CIMBA, ce qui pourrait expliquer une différence de fréquence allélique entre les témoins de CIMBA et ceux de BCAC pour certains SNPs. Ce seuil nous permet donc d'éviter d'exclure un nombre trop grand de SNPs à tort.
Les SNPs potentiellement en DL ou inter-chromosomal (DLI) avec les mutations *BRCA1/2* qui ont été exclus devront peut-être être analysés plus en profondeur afin de comprendre leur rôle dans le risque de cancer du sein. En effet, une publication récente basée sur les données de la *Framingham Heart Study* suggère qu'un DLI peut être causé par des mécanismes biogénétiques potentiellement associés à une évolution épistatique favorable ou défavorable[220].

Nos résultats ont également mis en évidence l'importance du processus d'imputation dans les études *GWAS*. Les données imputées utilisées dans un premier temps pour réaliser les analyses *case-only* sont basées sur une imputation séparée des sujets de BCAC et CIMBA. Cependant, nous avons montré que 28 des 33 régions initialement associées n'étaient plus significatives après ré-imputation jointe des sujets de BCAC et CIMBA. Cela suggère que les associations trouvées dans les régions toujours significatives sont robustes.
Avec cette stratégie, seules les régions trouvées significatives dans l'analyse faite sur les sujets de BCAC et CIMBA imputés séparément ont été ré-imputées. La ré-imputation n'a pas été réalisée sur le génome entier à cause de contraintes computationnelles. Cela a probablement conduit à un certain nombre de faux négatifs. Cependant, notre stratégie nous permet de s'assurer que le nombre de faux positifs est relativement faible parmi les SNPs avec une fréquence supérieure à 5 % et bien imputés ($r^2 > 0,5$).

L'âge moyen au diagnostic des porteuses d'une mutation *BRCA1/2* est en moyenne de 16 ans plus jeune que les cas de la population générale. Les associations observées peuvent être dues à cette différence d'âge et non à une réelle interaction avec les mutations *BRCA1/2* bien que les analyses soient ajustées sur l'âge à la censure. Cependant, le taux de mortalité dans cette période de 16 ans (entre 42,5 et 58,4 ans) est encore négligeable et ce biais est peu probable. Une autre source de biais potentiel pourrait être due au fait que 83 % des cas de CIMBA sont des cas prévalents. Les associations mises en évidence pourraient être des associations avec des gènes de bon pronostic. Cependant, aucun des SNPs identifiés n'a été associé à la survie au cancer[262].

Certains SNPs montrent une hétérogénéité par pays. Cependant, l'exclusion des pays un à un m'a permis de montrer que l'estimation était robuste à la variation géographique et que la valeur de l'OR pouvait être utilisée globalement, à l'exception de 2 SNPs, rs736596 et rs79183898, qui montrent une réduction de 12,5 % et 16 % de l'OR respectivement lorsque les États-Unis sont exclus. Néanmoins, une observation pays par pays montre que les intervalles de confiance se chevauchent mais que les estimations ponctuelles peuvent être très différentes d'un pays à l'autre. La question de la validation de ces SNPs dans chaque pays se pose donc.

Nos résultats pourraient contribuer à améliorer la prédiction du risque de cancer chez les porteurs d'une mutation dans les gènes *BRCA1* ou *BRCA2* grâce au développement d'un PRS spécifique à ces populations.

Nos résultats montrent que le PRS spécifique aux porteurs d'une mutation *BRCA1* devrait être composé des 172 SNPs trouvés associés dans la population générale et non significatifs dans notre analyse (152 SNPs trouvés associés aux cancers du sein tous sous-types confondus et à 20 SNPs trouvés associés aux tumeurs RE^-) avec comme pondération les estimations obtenues en population générale pour les tumeurs RE^-. De plus, il faudrait ajouter les 7 SNPs, rs17426269, rs13281615, chr10_80841148_C_T, chr16_52599188_C_T, chr1_10566215, rs17529111 et rs66823261 trouvés dans la population générale et associés au statut *BRCA1* mais également les 4 nouveaux SNPs mis en évidence, rs80221606, rs58117746, rs5820435 et rs11079012, en utilisant les estimations obtenues dans notre analyse.

Le PRS spécifique aux porteurs d'une mutation *BRCA2* devrait être quant à lui composé des 155 SNPs associés en population générale et non significatifs dans notre analyse, avec les estimations obtenues dans la population générale mais également les 3 SNPs trouvés dans la population générale et associés au statut BRCA2 (rs62355902, rs10759243 et chr22_40876234_C_T) et les 4 nouveaux SNPs mis en évidence, rs12470785, rs79183898, rs736596 et rs4943263, avec les betas estimés dans notre analyse. Ces PRS spécifiques devront alors être validés dans une cohorte indépendante de femmes porteuses d'une mutation dans les gènes *BRCA1/2*[3–57]. Cette approche devra être étendue au PRS composé de 313 SNPs publié au début de l'année 2019[176]. Cependant, les 2/3 des SNPs de ce nouveau PRS sont des SNPs imputés. Il sera donc nécessaire de les ré-imputer avec les sujets de BCAC et CIMBA conjointement.

Cette étude est la première sur les facteurs génétiques modificateurs du risque de cancer du sein qui a cherché à mettre en évidence des différences d'association entre la population générale et les femmes porteuses d'une mutation dans les gènes *BRCA1/2*. L'inclusion des cas de cancer du sein de la population générale a permis une augmentation de la taille de la population étudiée et donc une augmentation de la puissance statistique pour la détection de nouveaux SNPs. D'autres SNPs spécifiques aux femmes porteuses d'une mutation dans ces gènes sont encore à mettre en évidence mais nos résultats pourraient déjà contribuer à l'amélioration du PRS permettant d'estimer le risque absolu de cancer du sein pour les femmes porteuses d'une mutation *BRCA1/2*.

Conclusions Et Perspectives

Les stratégies que nous avons développées ont montré leur capacité à détecter de nouveaux facteurs génétiques impliqués dans le risque familial de cancer du sein.

L'analyse *case-only* réalisée sur les données des consortia internationaux BCAC et CIMBA nous ont permis de mettre en évidence 8 nouveaux SNPs spécifiquement associés chez les femmes porteuses d'une mutation dans les gènes *BRCA1* ou *BRCA2*. L'effort international pour regrouper les données du plus grand nombre de pays possible permet d'obtenir des jeux de données de taille importante pour mettre en évidence de nouveaux facteurs génétiques associés à des risques faibles qui pourront alors être intégrés aux PRS spécifiques aux femmes porteuses d'une mutation de *BRCA1* ou *BRCA2*. Cependant, ces données sont très hétérogènes et nos résultats montrent que les estimations ponctuelles des SNPs mis en évidence varient selon le pays d'origine. Les SNPs des PRS, que ce soit ceux de la population générale ou bien ceux spécifiques aux porteuses d'une mutation de *BRCA1* ou *BRCA2*, devraient donc être validés dans chaque pays et les estimations ponctuelles éventuellement recalculées.

Les analyses stratifiées sur les profils d'expositions environnementales des femmes de l'étude française GENESIS montrent qu'il existe une variation du risque associé à certains SNPs selon le profil d'exposition. Notamment, les analyses mettent en évidence deux régions du génome (3p14.1 et 5q14.2) associées spécifiquement aux femmes qui n'ont jamais été exposées aux radiations ionisantes et une région (10q26.13) spécifiquement associée aux femmes fortement exposées aux radiations. Cependant, la stratification sur les profils d'exposition aux facteurs gynéco-obstétriques n'a mis aucune association spécifique en évidence et cela possiblement à cause du manque de puissance de notre étude. En effet, l'étude GENESIS est une étude de dimension nationale qui contient une population environ 22 fois plus petite que celle utilisée dans l'analyse *case-only*. L'avantage de GENESIS est de porter sur une population homogène qui permet d'obtenir des estimations de risque spécifiques à la population étudiée.

Les deux projets ont porté sur des données génotypiques provenant de puces à ADN qui montrent certaines limites pour trouver de nouveaux SNPs associés au cancer du sein.

En effet, les SNPs inclus dans les puces à ADN ont été choisis pour leur association déjà connue avec le cancer ou d'autres phénotypes ayant un intérêt dans le cadre de la prédisposition au cancer du sein. L'imputation permet, en théorie, d'analyser le génome entier. Cependant, les régions du génome les moins denses en SNPs sur la puce utilisée sont moins bien imputées que les autres et nos résultats montrent que plus de la moitié des SNPs peuvent alors être exclus à cause de leur faible qualité d'imputation. Des régions entières de SNPs ne sont donc pas analysées. De plus, nous avons montré que l'imputation des SNPs dépendait fortement de l'étape de pré-haplotypage et que les résultats obtenus étaient très différents lorsque les haplotypes des sujets étaient prédits séparément ou simultanément, aussi bien pour les SNPs rares que fréquents.

Ces deux nouvelles stratégies pourront permettre d'expliquer une partie du risque familial encore non expliqué mais probablement pas tout. D'autres pistes doivent être envisagées comme l'analyse du génome encore non exploré avec des puces plus denses, qui permettront également d'analyser les SNPs rares. En effet, la majorité des études se sont concentrées sur les SNPs fréquents. Les SNPs génotypés sont donc majoritairement des SNPs fréquents (fréquence de 5 à 50 %). L'imputation des SNPs rares est, du fait de leur rareté, de moins bonne qualité que les SNPs fréquents. Ils sont donc pour beaucoup exclus des analyses. Les nouveaux SNPs qui pourraient expliquer le risque familial restant se trouvent peut-être parmi ces SNPs rares qui pourraient être associés à des différences fonctionnelles significativement délétères. L'effet délétère de ces SNPs expliquerait donc leur rareté.
Une solution serait d'utiliser des puces à ADN de variants rares. Cependant, comme l'explique Zuk *et al*.[263], il existe un nombre beaucoup trop important de variants rares pour les énumérer et leur fréquence ne permet pas de les analyser un à un. Il est donc nécessaire d'agréger ces variants rares (par exemple, construire un RV-PRS, pour Rare Variant Polygenic Risk Score[264]) et de comparer la distribution de la fréquence de cette agrégation entre les cas de cancer du sein et les témoins. Pour cela, il faudrait utiliser des données de séquençage à haut débit de génome ou d'exome entier. Cependant, ces études doivent être réalisées sur des milliers d'individus et bien que le prix du séquençage d'un génome ait drastiquement diminué ces dernières années, le coût est encore très important et la capacité de stockage non encore disponible.

Références Bibliographiques

1. Anbazhagan, R., Osin, P., Bartkova, J. & Nathan, B. The development of epithelial phenotypes in the human fetal and infant breast. *The Journal of Pathology* 197–206 (1998).
2. Russo, J. & Russo, I. H. Development of the human breast. *Maturitas* **49**, 2–15 (2004).
3. Beatson, G. T. On the Treatment of Inoperable Cases of Carcinoma of the Mamma: Suggestions for a New Method of Treatment, with Illustrative Cases. *Trans. Medico-Chir. Soc. Edinb.* **15**, 153–179 (1896).
4. Mulac-Jericevic, B. & Conneely, O. M. Reproductive tissue selective actions of progesterone receptors. *Reprod. Camb. Engl.* **128**, 139–146 (2004).
5. Feng, Y., Manka, D., Wagner, K.-U. & Khan, S. A. Estrogen receptor-α expression in the mammary epithelium is required for ductal and alveolar morphogenesis in mice. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14718–14723 (2007).
6. Yart, L., Lollivier, V., Marnet, P. G. & Dessauge, F. Role of ovarian secretions in mammary gland development and function in ruminants. *animal* **8**, 72–85 (2014).
7. Bachelot, A. & Binart, N. Reproductive role of prolactin. *Reproduction* **133**, 361–369 (2007).
8. Amico, J. A. & Finley, B. E. Breast Stimulation in Cycling Women, Pregnant Women and a Woman with Induced Lactation: Pattern of Release of Oxytocin, Prolactin and Luteinizing Hormone. *Clin. Endocrinol. (Oxf.)* **25**, 97–106 (1986).
9. Hovey, R. C., Trott, J. F. & Vonderhaar, B. K. Establishing a framework for the functional mammary gland: from endocrinology to morphology. *J. Mammary Gland Biol. Neoplasia* **7**, 17–38 (2002).
10. Essential Medical Physiology - 3rd Edition. https://www.elsevier.com/books/essential-medical-physiology/johnson/978-0-08-047270-6.
11. Javed, A. & Lteif, A. Development of the Human Breast. *Semin. Plast. Surg.* **27**, 5–12 (2013).
12. Brisken, C. & O'Malley, B. Hormone Action in the Mammary Gland. *Cold Spring Harb. Perspect. Biol.* **2**, (2010).
13. Dürnberger, H. & Kratochwil, K. Specificity of tissue interaction and origin of mesenchymal cells in the androgen response of the embryonic mammary gland. *Cell* **19**, 465–471 (1980).
14. Martinet, J. *Biologie de la lactation*. (Editions Quae, 1993).
15. Stanhope, R., Adams, J. & Brook, C. G. Disturbances of puberty. *Clin. Obstet. Gynaecol.* **12**, 557–577 (1985).
16. Textbook of Medical Physiology - 11th Edition. https://www.elsevier.com/books/textbook-of-medical-physiology/hall/978-0-7216-0240-0.
17. Li, X. *et al*. Single-Chain Estrogen Receptors (ERs) Reveal that the ERα/β Heterodimer Emulates Functions of the ERα Dimer in Genomic Estrogen Signaling Pathways. *Mol. Cell. Biol.* **24**, 7681–7694 (2004).
18. Macias, H. & Hinck, L. Mammary Gland Development. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 533–557 (2012).
19. Feldman, M., Ruan, W., Cunningham, B. C., Wells, J. A. & Kleinberg, D. L. Evidence that the growth hormone receptor mediates differentiation and development of the mammary gland. *Endocrinology* **133**, 1602–1608 (1993).
20. Kleinberg, D. L. & Ruan, W. IGF-I, GH, and Sex Steroid Effects in Normal Mammary Gland Development. *J. Mammary Gland Biol. Neoplasia* **13**, 353–360 (2008).

21. Hennighausen, L. & Robinson, G. W. Signaling Pathways in Mammary Gland Development. *Dev. Cell* **1**, 467–475 (2001).
22. Lange, C. Challenges to defining a role for progesterone in breast cancer. *Steroids* **73**, 914–921 (2008).
23. Longacre, T. A. & Bartow, S. A. A correlative morphologic study of human breast and endometrium in the menstrual cycle. *Am. J. Surg. Pathol.* **10**, 382–393 (1986).
24. Christopher, B. W., Victor, N. & Yvonne, M. *Remington and Klein's Infectious Diseases of the Fetus and Newborn Infant - 8th Edition*. (2015).
25. Johnson, M. C. & Cutler, M. L. Anatomy and Physiology of the Breast. *Manag. Breast Dis.* 1–39 (2016) doi:10.1007/978-3-319-46356-8_1.
26. McNeilly, A. S., Robinson, I. C., Houston, M. J. & Howie, P. W. Release of oxytocin and prolactin in response to suckling. *Br. Med. J. Clin. Res. Ed* **286**, 257–259 (1983).
27. Lupoli, B., Johansson, B., Uvnäs-Moberg, K. & Svennersten-Sjaunja, K. Effect of suckling on the release of oxytocin, prolactin, cortisol, gastrin, cholecystokinin, somatostatin and insulin in dairy cows and their calves. *J. Dairy Res.* **68**, 175–187 (2001).
28. Ricketts, D. *et al*. Estrogen and Progesterone Receptors in the Normal Female Breast. *Cancer Res.* **51**, 1817–1822 (1991).
29. Cianfrocca, M. & Goldstein, L. J. Prognostic and Predictive Factors in Early-Stage Breast Cancer. *The Oncologist* **9**, 606–616 (2004).
30. Tamoxifen for early breast cancer: an overview of the randomised trials. *The Lancet* **351**, 1451–1467 (1998).
31. Kos, Z. & Dabbs, D. J. Biomarker assessment and molecular testing for prognostication in breast cancer. *Histopathology* **68**, 70–85 (2016).
32. Bardou, V.-J., Arpino, G., Elledge, R. M., Osborne, C. K. & Clark, G. M. Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **21**, 1973–1979 (2003).
33. Dowsett, M. *et al*. Benefit from adjuvant tamoxifen therapy in primary breast cancer patients according oestrogen receptor, progesterone receptor, EGF receptor and HER2 status. *Ann. Oncol.* **17**, 818–826 (2006).
34. Leone, N. *et al*. *Projection de l'incidence et de la mortalité par cancer en France métropolitaine en 2015 - Rapport technique*. 62 (2015).
35. Ferlay, J. *et al*. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
36. Hereditary Breast and Ovarian Cancer Syndrome. *Gynecol. Oncol.* **113**, 6–11 (2009).
37. Moretta, J. *et al*. Recommandations françaises du Groupe Génétique et Cancer pour l'analyse en panel de gènes dans les prédispositions héréditaires au cancer du sein ou de l'ovaire. *Bull. Cancer (Paris)* **105**, 907–917 (2018).
38. Anders, C. K., Johnson, R., Litton, J., Phillips, M. & Bleyer, A. Breast Cancer Before Age 40 Years. *Semin. Oncol.* **36**, 237–249 (2009).
39. Howlader, N., Noone, A. & Krapcho, M. SEER Cancer Statistics Review, 1975–2009 (Vintage 2009 Populations). (2012).
40. Seitz, H. K., Pelucchi, C., Bagnardi, V. & Vecchia, C. L. Epidemiology and Pathophysiology of Alcohol and Breast Cancer: Update 2012. *Alcohol Alcohol* **47**, 204–212 (2012).
41. Mourouti, N., Kontogianni, M. D., Papavagelis, C. & Panagiotakos, D. B. Diet and breast cancer: a systematic review. *Int. J. Food Sci. Nutr.* **66**, 1–42 (2015).
42. van't Veer, P., Kok, F. J., Hermus, R. J. & Sturmans, F. Alcohol dose, frequency and age at first exposure in relation to the risk of breast cancer. *Int. J. Epidemiol.* **18**, 511–517 (1989).

43. Young, T. B. A case-control study of breast cancer and alcohol consumption habits. *Cancer* **64**, 552–558 (1989).
44. Hamajima, N., Hirose, K. & Tajima, K. Alcohol, tobacco and breast cancer – collaborative reanalysis of individual data from 53 epidemiological studies, including 58515 women with breast cancer and 95067 women without the disease. *Br. J. Cancer* **87**, 1234–1245 (2002).
45. White, A. J., D'Aloisio, A. A., Nichols, H. B., DeRoo, L. A. & Sandler, D. P. Breast cancer and exposure to tobacco smoke during potential windows of susceptibility. *Cancer Causes Control CCC* **28**, 667–675 (2017).
46. Gaudet, M. M. *et al*. Active smoking and breast cancer risk: original cohort data and meta-analysis. *J. Natl. Cancer Inst.* **105**, 515–525 (2013).
47. Gaudet, M. M. *et al*. Pooled analysis of active cigarette smoking and invasive breast cancer risk in 14 cohort studies. *Int. J. Epidemiol.* **46**, 881–893 (2017).
48. Cottet, V. *et al*. Postmenopausal Breast Cancer Risk and Dietary Patterns in the E3N-EPIC Prospective Cohort Study. *Am. J. Epidemiol.* **170**, 1257–1267 (2009).
49. Trichopoulou, A., Bamia, C., Lagiou, P. & Trichopoulos, D. Conformity to traditional Mediterranean diet and breast cancer risk in the Greek EPIC (European Prospective Investigation into Cancer and Nutrition) cohort. *Am. J. Clin. Nutr.* **92**, 620–625 (2010).
50. World Cancer Research Fund & American Institute for Cancer Reasearch. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. *Choice Rev. Online* **45**, 45-5024-45–5024 (2007).
51. Cummings, S. R. *et al*. Prevention of Breast Cancer in Postmenopausal Women: Approaches to Estimating and Reducing Risk. *JNCI J. Natl. Cancer Inst.* **101**, 384–398 (2009).
52. Hauner, H. & Hauner, D. The Impact of Nutrition on the Development and Prognosis of Breast Cancer. *Breast Care* **5**, 5–5 (2010).
53. Neil-Sztramko, S. E. *et al*. Does obesity modify the relationship between physical activity and breast cancer risk? *Breast Cancer Res. Treat.* **166**, 367–381 (2017).
54. Bergström, A., Pisani, P., Tenet, V., Wolk, A. & Adami, H.-O. Overweight as an avoidable cause of cancer in Europe. *Int. J. Cancer* **91**, 421–430 (2001).
55. Peeters, P. H. M., Verbeek, A. L. M., Krol, A., Matthyssen, M. M. M. & de Waard, F. Age at menarche and breast cancer risk in nulliparous women. *Breast Cancer Res. Treat.* **33**, 55–61 (1995).
56. Clavel-Chapelon, F. Differential effects of reproductive factors on the risk of pre- and postmenopausal breast cancer. Results from a large cohort of French women. *Br. J. Cancer* **86**, 723–727 (2002).
57. Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *Lancet Oncol.* **13**, 1141–1151 (2012).
58. Yang, X. R. *et al*. Associations of Breast Cancer Risk Factors With Tumor Subtypes: A Pooled Analysis From the Breast Cancer Association Consortium Studies. *JNCI J. Natl. Cancer Inst.* **103**, 250–263 (2011).
59. Calle, E. E. *et al*. Breast cancer and hormonal contraceptives: Collaborative reanalysis of individual data on 53297 women with breast cancer and 100239 women without breast cancer from 54 epidemiological studies. *Lancet* **347**, 1713–1727 (1996).
60. Mørch, L. S. *et al*. Contemporary Hormonal Contraception and the Risk of Breast Cancer. *N. Engl. J. Med.* **377**, 2228–2239 (2017).
61. Marchbanks, P. A. *et al*. Oral contraceptives and the risk of breast cancer. *N. Engl. J. Med.* **346**, 2025–2032 (2002).
62. Hannaford, P. C. *et al*. Cancer risk among users of oral contraceptives: cohort data

from the Royal College of General Practitioner's oral contraception study. *BMJ* **335**, 651 (2007).
63. Kelsey, J. L., Gammon, M. D. & John, E. M. Reproductive factors and breast cancer. *Epidemiol. Rev.* **15**, 36–47 (1993).
64. MacMahon, B. *et al.* Age at first birth and breast cancer risk. *Bull. World Health Organ.* **43**, 209–221 (1970).
65. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. *The Lancet* **360**, 187–195 (2002).
66. Lambe, M. *et al.* Transient Increase in the Risk of Breast Cancer after Giving Birth. *N. Engl. J. Med.* **331**, 5–9 (1994).
67. Albrektsen, G., Heuch, I., Hansen, S. & Kvåle, G. Breast cancer risk by age at birth, time since birth and time intervals between births: exploring interaction effects. *Br. J. Cancer* **92**, 167–175 (2005).
68. Nichols, H. B. *et al.* Breast Cancer Risk After Recent Childbirth: A Pooled Analysis of 15 Prospective Studies. *Ann. Intern. Med.* (2018) doi:10.7326/M18-1323.
69. Zhou, Y. *et al.* Association Between Breastfeeding and Breast Cancer Risk: Evidence from a Meta-analysis. *Breastfeed. Med.* **10**, 175–182 (2015).
70. Brind, J., Chinchilli, V. M., Severs, W. B. & Summy-Long, J. Induced abortion as an independent risk factor for breast cancer: a comprehensive review and meta-analysis. *J. Epidemiol. Community Health* **50**, 481–496 (1996).
71. Ilic, M., Vlajinac, H., Marinkovic, J. & Sipetic-Grujicic, S. Abortion and breast cancer: case-control study. *Tumori* **99**, 452–457 (2013).
72. Li, C. I. *et al.* Relationship Between Long Durations and Different Regimens of Hormone Therapy and Risk of Breast Cancer. *JAMA* **289**, 3254–3263 (2003).
73. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52 705 women with breast cancer and 108 411 women without breast cancer. *The Lancet* **350**, 1047–1059 (1997).
74. Mercado, C. L. BI-RADS Update. *Radiol. Clin. North Am.* **52**, 481–487 (2014).
75. Alunni, J. P., Marty, M. H. & Feillel, V. Classification BI-RADS : densité mammaire. (2008).
76. Boyd, N. F. *et al.* Mammographic breast density as an intermediate phenotype for breast cancer. *Lancet Oncol.* **6**, 798–808 (2005).
77. Wang, A. T., Vachon, C. M., Brandt, K. R. & Ghosh, K. Breast Density and Breast Cancer Risk: A Practical Review. *Mayo Clin. Proc.* **89**, 548–557 (2014).
78. Bertrand, K. A. *et al.* Mammographic density and risk of breast cancer by age and tumor characteristics. *Breast Cancer Res. BCR* **15**, R104 (2013).
79. Pettersson, A. *et al.* Mammographic Density Phenotypes and Risk of Breast Cancer: A Meta-analysis. *JNCI J. Natl. Cancer Inst.* **106**, (2014).
80. *Sources and effects of ionizing radiation - UNSCEAR 2000 report.* (2000).
81. Shimizu, Y., Kato, H. & Schull, W. J. Studies of the Mortality of A-Bomb Survivors: 9. Mortality, 1950-1985: Part 2. Cancer Mortality Based on the Recently Revised Doses (DS86). *Radiat. Res.* **121**, 120 (1990).
82. Tokunaga, M. *et al.* Incidence of female breast cancer among atomic bomb survivors, Hiroshima and Nagasaki, 1950-1980. *Radiat. Res.* **112**, 243–272 (1987).
83. Cook, D. C., Dent, O. & Hewitt, D. Breast cancer following multiple chest fluoroscopy: the Ontario experience. *Can. Med. Assoc. J.* **111**, 406–412 (1974).
84. Boice, J. D. & Monson, R. R. Breast cancer in women after repeated fluoroscopic

examinations of the chest. *J Natl Cancer Inst* **59**, (1977).
85. Davis, F. G., Boice, J. D., Hrubec, Z. & Monson, R. R. Cancer Mortality in a Radiation-exposed Cohort of Massachusetts Tuberculosis Patients. **49**, 6130–6136 (1989).
86. Shore, R. E. *et al*. Breast cancer among women given X-ray therapy for acute postpartum mastitis. *J. Natl. Cancer Inst*. **77**, 689–696 (1986).
87. Haddy, N. *et al*. Breast cancer following radiotherapy for a hemangioma during childhood. *Cancer Causes Control* **21**, 1807–1816 (2010).
88. Ronckers, C. M., Erdmann, C. A. & Land, C. E. Radiation and breast cancer: a review of current evidence. *Breast Cancer Res*. **7**, 21–32 (2005).
89. Jansen-van der Weide, M. C. *et al*. Exposure to low-dose radiation and the risk of breast cancer among women with a familial or genetic predisposition: a meta-analysis. *Eur. Radiol*. **20**, 2547–2556 (2010).
90. Pijpe, A. *et al*. Exposure to diagnostic radiation and risk of breast cancer among carriers of BRCA1/2 mutations: retrospective cohort study (GENE-RAD-RISK). *BMJ* **345**, e5660 (2012).
91. Pharoah, P. D. P., Day, N. E., Duffy, S., Easton, D. F. & Ponder, B. A. J. Family history and the risk of breast cancer: A systematic review and meta-analysis. *Int. J. Cancer* **71**, 800–809 (1997).
92. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. *The Lancet* **358**, 1389–1399 (2001).
93. Easton, D. F. Familial risks of breast cancer. *Breast Cancer Res*. **4**, 179 (2002).
94. Williams, W. R. & Anderson, D. E. Genetic epidemiology of breast cancer: segregation analysis of 200 Danish pedigrees. *Genet. Epidemiol*. **1**, 7–20 (1984).
95. Claus, E. B., Risch, N. & Thompson, W. D. Genetic analysis of breast cancer in the cancer and steroid hormone study. *Am. J. Hum. Genet*. **48**, 232–242 (1991).
96. Andrieu, N., Clavel, F. & Demenais, F. Familial susceptibility to breast cancer: a complex inheritance. *Int. J. Cancer* **44**, 415–418 (1989).
97. Hall, J. *et al*. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).
98. Wooster, R. *et al*. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265**, 2088–2090 (1994).
99. Apostolou, P. & Fostira, F. Hereditary Breast Cancer: The Era of New Susceptibility Genes. *BioMed Res. Int*. (2013) doi:10.1155/2013/747318.
100. Economopoulou, P., Dimitriadis, G. & Psyrri, A. Beyond BRCA: New hereditary breast cancer susceptibility genes. *Cancer Treat. Rev*. **41**, 1–8 (2015).
101. Nielsen, S. M. *et al*. Genetic Testing and Clinical Management Practices for Variants in Non-BRCA1/2 Breast (and Breast/Ovarian) Cancer Susceptibility Genes: An International Survey by the Evidence-Based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) Clinical Working Group. *JCO Precis. Oncol*. 1–42 (2018) doi:10.1200/PO.18.00091.
102. Hearle, N. *et al*. Frequency and Spectrum of Cancers in the Peutz-Jeghers Syndrome. *Clin. Cancer Res*. **12**, 3209–3215 (2006).
103. Renwick, A. *et al*. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet*. **38**, 873–875 (2006).
104. Nicolo, A. D. *et al*. A Novel Breast Cancer–Associated BRIP1 (FANCJ/BACH1) Germ-line Mutation Impairs Protein Stability and Function. *Clin. Cancer Res*. **14**, 4672–4680 (2008).
105. Akbari, M. R. *et al*. RAD51C germline mutations in breast and ovarian cancer patients. *Breast Cancer Res*. **12**, 404 (2010).

106. Damiola, F. *et al*. Rare key functional domain missense substitutions in MRE11A, RAD50, and NBN contribute to breast cancer susceptibility: results from a Breast Cancer Family Registry case-control mutation-screening study. *Breast Cancer Res*. *BCR* **16**, R58 (2014).
107. Brennan, P. Breast cancer risk in MEN1 – a cancer genetics perspective. *Clin*. *Endocrinol*. *(Oxf.)* **82**, 327–229 (2015).
108. Couch, F. J. *et al*. Associations Between Cancer Predisposition Testing Panel Genes and Breast Cancer. *JAMA Oncol*. **3**, 1190–1196 (2017).
109. Slavin, T. P. *et al*. The contribution of pathogenic variants in breast cancer susceptibility genes to familial breast cancer risk. *NPJ Breast Cancer* **3**, 22 (2017).
110. Apostolou, P. & Papasotiriou, I. Current perspectives on CHEK2 mutations in breast cancer. *Breast Cancer Targets Ther*. **9**, 331–335 (2017).
111. Howell, S. J., Hockenhull, K., Salih, Z. & Evans, D. G. Increased risk of breast cancer in neurofibromatosis type 1: current insights. *Breast Cancer Targets Ther*. **9**, 531–536 (2017).
112. Chen, X. *et al*. Associations between RAD51D germline mutations and breast cancer risk and survival in BRCA1/2-negative breast cancers. *Ann*. *Oncol*. **29**, 2046–2051 (2018).
113. Freudenheim, J. L. *et al*. Alcohol dehydrogenase 3 genotype modification of the association of alcohol consumption with breast cancer risk. *Cancer Causes Control CCC* 169–177 (1999).
114. Ewart-Toland, A. *et al*. Aurora- A/STK15 T + 91A is a general low penetrance cancer susceptibility gene: a meta-analysis of multiple cancer types. *Carcinogenesis* **26**, 1368–1373 (2005).
115. MacPherson, G. *et al*. Association of a Common Variant of the CASP8 Gene With Reduced Risk of Breast Cancer. *JNCI J*. *Natl*. *Cancer Inst*. **96**, 1866–1869 (2004).
116. Frank, B. *et al*. Re: Association of a Common Variant of the CASP8 Gene With Reduced Risk of Breast Cancer. *JNCI J*. *Natl*. *Cancer Inst*. **97**, 1012–1012 (2005).
117. Shi, Q. *et al*. Reduced DNA repair of benzo[a]pyrene diol epoxide-induced adducts and common XPD polymorphisms in breast cancer patients. *Carcinogenesis* **25**, 1695–1700 (2004).
118. Lee, S.-A. *et al*. Obesity and genetic polymorphism of ERCC2 and ERCC4 as modifiers of risk of breast cancer. *Exp*. *Mol*. *Med*. **37**, 86–90 (2005).
119. Pechlivanis, S. *et al*. Polymorphisms in the insulin like growth factor 1 and IGF binding protein 3 genes and risk of colorectal cancer. *Cancer Detect*. *Prev*. **31**, 408–416 (2007).
120. Kuschel, B. *et al*. Variants in DNA double-strand break repair genes and breast cancer susceptibility. *Hum*. *Mol*. *Genet*. **11**, 1399–1407 (2002).
121. Pooley, K. A. *et al*. Association of the Progesterone Receptor Gene with Breast Cancer Risk: A Single-Nucleotide Polymorphism Tagging Approach. *Cancer Epidemiol*. *Prev*. *Biomark*. **15**, 675–682 (2006).
122. Ambrosone, C. B. *et al*. Manganese Superoxide Dismutase (MnSOD) Genetic Polymorphisms, Dietary Antioxidants, and Risk of Breast Cancer. *Cancer Res*. **59**, 602–606 (1999).
123. Rohrbacher, M., Risch, A., Kropp, S. & Chang-Claude, J. The A-336C Insulin-Like Growth Factor Binding Protein-3 Promoter Polymorphism Is Not a Modulator of Breast Cancer Risk in Caucasian Women. *Cancer Epidemiol Biomark*. *Prev* 3 (2005).
124. Kuschel, B. *et al*. Common Polymorphisms in ERCC2 (Xeroderma pigmentosum D) are not Associated with Breast Cancer Risk. *Cancer Epidemiol*. *Prev*. *Biomark*. **14**, 1828–1831 (2005).
125. Hines, L. M. *et al*. A Prospective Study of the Effect of Alcohol Consumption and ADH3 Genotype on Plasma Steroid Hormone Levels and Breast Cancer Risk. *Cancer*

Epidemiol. *Prev*. *Biomark*. **9**, 1099–1105 (2000).
126. De Vivo, I., Hankinson, S. E., Colditz, G. A. & Hunter, D. J. The progesterone receptor Val660→Leu polymorphism and breast cancer risk. *Breast Cancer Res*. *BCR* **6**, R636–R639 (2004).
127. Egan, K. M., Thompson, P. A., Titus-Ernstoff, L., Moore, J. H. & Ambrosone, C. B. MnSOD polymorphism and breast cancer in a population-based case–control study. *Cancer Lett*. **199**, 27–33 (2003).
128. Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nat*. *Genet*. **29**, 306–309 (2001).
129. Dahlman, I. *et al*. Parameters for reliable results in genetic association studies in common disease. *Nat*. *Genet*. **30**, 149–150 (2002).
130. Colhoun, H. M., McKeigue, P. M. & Smith, G. D. Problems of reporting genetic associations with complex outcomes. *The Lancet* **361**, 865–872 (2003).
131. Commonly Studied Single-Nucleotide Polymorphisms and Breast Cancer: Results From the Breast Cancer Association Consortium. *JNCI J*. *Natl*. *Cancer Inst*. **98**, 1382–1396 (2006).
132. Easton, D. F. *et al*. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
133. Hunter, D. J. *et al*. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat*. *Genet*. **39**, 870–874 (2007).
134. Stacey, S. N. *et al*. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat*. *Genet*. **39**, 865–869 (2007).
135. Stacey, S. N. *et al*. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat*. *Genet*. **40**, 703–706 (2008).
136. Bahcall, O. COGS project and design of the iCOGS array. *Nat*. *Genet*. (2013) doi:10.1038/ngicogs.4.
137. Michailidou, K. *et al*. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat*. *Genet*. **45**, 353–361 (2013).
138. Garcia-Closas, M. *et al*. Genome-wide association studies identify four ER negative–specific breast cancer risk loci. *Nat*. *Genet*. **45**, 392-398e2 (2013).
139. Michailidou, K. *et al*. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat*. *Genet*. **47**, 373–380 (2015).
140. Amos, C. I. *et al*. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol*. *Biomark*. *Prev*. *Publ*. *Am*. *Assoc*. *Cancer Res*. *Cosponsored Am*. *Soc*. *Prev*. *Oncol*. **26**, 126–135 (2017).
141. Michailidou, K. *et al*. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
142. Milne, R. L. *et al*. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat*. *Genet*. **49**, 1767–1778 (2017).
143. Mavaddat, N. *et al*. Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. *JNCI J*. *Natl*. *Cancer Inst*. **107**, (2015).
144. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat*. *Rev*. *Cancer* **12**, 68–78 (2011).
145. Cline, M. S. *et al*. BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genet*. **14**, (2018).
146. Spurdle, A. B. *et al*. ENIGMA - Evidence-based Network for the Interpretation of Germline Mutant Alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum*. *Mutat*. **33**, 2–7

(2012).
147. Iau, P. T. C. *et al*. Are medullary breast cancers an indication for BRCA1 mutation screening? A mutation analysis of 42 cases of medullary breast cancer. *Breast Cancer Res. Treat*. **85**, 81–88 (2004).
148. Breast Cancer Linkage Consortium. Pathology of familial breast cancer: differences between breast cancers in carriers of BRCA1 or BRCA2 mutations and sporadic cases. Breast Cancer Linkage Consortium. *Lancet Lond. Engl*. **349**, 1505–1510 (1997).
149. Mavaddat, N. *et al*. Pathology of breast and ovarian cancers among BRCA1 and BRCA2 mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol*. **21**, 134–147 (2012).
150. Kuchenbaecker, K. B. *et al*. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* **317**, 2402–2416 (2017).
151. Brohet, R. M. *et al*. Oral Contraceptives and Breast Cancer Risk in the International BRCA1/2 Carrier Cohort Study: A Report From EMBRACE, GENEPSO, GEO-HEBON, and the IBCCS Collaborating Group. *J. Clin. Oncol*. **25**, 3831–3836 (2007).
152. Andrieu, N. *et al*. Pregnancies, breast-feeding, and breast cancer risk in the International BRCA1/2 Carrier Cohort Study (IBCCS). *J. Natl. Cancer Inst*. **98**, 535–544 (2006).
153. Friebel, T. M., Domchek, S. M. & Rebbeck, T. R. Modifiers of Cancer Risk in BRCA1 and BRCA2 Mutation Carriers: Systematic Review and Meta-Analysis. *JNCI J. Natl. Cancer Inst*. **106**, (2014).
154. Schrijver, L. H. *et al*. Oral Contraceptive Use and Breast Cancer Risk: Retrospective and Prospective Analyses From a BRCA1 and BRCA2 Mutation Carrier Cohort Study. *JNCI Cancer Spectr*. **2**, (2018).
155. Terry, M. B. *et al*. The Influence of Number and Timing of Pregnancies on Breast Cancer Risk for Women With BRCA1 or BRCA2 Mutations. *JNCI Cancer Spectr*. **2**, pky078 (2018).
156. Chenevix-Trench, G. *et al*. An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). *Breast Cancer Res*. **9**, 104 (2007).
157. Couch, F. J. *et al*. Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk. *PLoS Genet*. **9**, (2013).
158. Antoniou, A. C. *et al*. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor–negative breast cancer in the general population. *Nat. Genet*. **42**, 885–892 (2010).
159. Gaudet, M. M. *et al*. Identification of a BRCA2-Specific Modifier Locus at 6p24 Related to Breast Cancer Risk. *PLoS Genet*. **9**, (2013).
160. Antoniou, A. C. *et al*. Common Breast Cancer-Predisposition Alleles Are Associated with Breast Cancer Risk in BRCA1 and BRCA2 Mutation Carriers. *Am. J. Hum. Genet*. **82**, 937–948 (2008).
161. Antoniou, A. C. *et al*. Common variants in LSP1, 2q35 and 8q24 and breast cancer risk for BRCA1 and BRCA2 mutation carriers. *Hum. Mol. Genet*. **18**, 4442–4456 (2009).
162. Antoniou, A. C. *et al*. Common alleles at 6q25.1 and 1p11.2 are associated with breast cancer risk for BRCA1 and BRCA2 mutation carriers. *Hum. Mol. Genet*. **20**, 3304–3321 (2011).
163. Bojesen, S. E. *et al*. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet*. **45**, 371-384e2 (2013).

164. Kuchenbaecker, K. B. *et al*. Associations of common breast cancer susceptibility alleles with risk of breast cancer subtypes in BRCA1 and BRCA2 mutation carriers. *Breast Cancer Res*. *BCR* **16**, (2014).
165. Couch, F. J. *et al*. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. *Nat*. *Commun*. **7**, (2016).
166. Lawrenson, K. *et al*. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast–ovarian cancer susceptibility locus. *Nat*. *Commun*. **7**, (2016).
167. Hamdi, Y. *et al*. Association of breast cancer risk in BRCA1 and BRCA2 mutation carriers with genetic variants showing differential allelic expression: identification of a modifier of breast cancer risk at locus 11q22.3. *Breast Cancer Res*. *Treat*. **161**, 117–134 (2017).
168. Antoniou, A. C. *et al*. Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carriers: implications for risk prediction. *Cancer Res*. **70**, 9742–9754 (2010).
169. Antoniou, A. C. *et al*. Common variants at 12p11, 12q24, 9p21, 9q31.2 and in ZNF365 are associated with breast cancer risk for BRCA1 and/or BRCA2 mutation carriers. *Breast Cancer Res*. *BCR* **14**, R33 (2012).
170. Kuchenbaecker, K. B. *et al*. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *JNCI J*. *Natl*. *Cancer Inst*. **109**, (2017).
171. Antoniou, A. C. & Easton, D. F. Models of genetic susceptibility to breast cancer. *Oncogene* **25**, 5898–5905 (2006).
172. Park, D. J. *et al*. Rare mutations in XRCC2 increase the risk of breast cancer. *Am*. *J*. *Hum*. *Genet*. **90**, 734–739 (2012).
173. Park, D. J. *et al*. Rare mutations in RINT1 predispose carriers to breast and Lynch syndrome-spectrum cancers. *Cancer Discov*. **4**, 804–815 (2014).
174. Melchor, L. & Benítez, J. The complex genetic landscape of familial breast cancer. *Hum*. *Genet*. **132**, 845–863 (2013).
175. Lee, A. *et al*. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet*. *Med*. *Off*. *J*. *Am*. *Coll*. *Med*. *Genet*. (2019) doi:10.1038/s41436-018-0406-9.
176. Mavaddat, N. *et al*. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am*. *J*. *Hum*. *Genet*. **104**, 21–34 (2019).
177. Simonds, N. I. *et al*. Review of the Gene-Environment Interaction Literature in Cancer: What do we know? *Genet*. *Epidemiol*. **40**, 356–365 (2016).
178. Milne, R. L. *et al*. Assessing interactions between the associations of common genetic susceptibility variants, reproductive history and body mass index with breast cancer risk in the breast cancer association consortium: a combined case-control study. *Breast Cancer Res*. *BCR* **12**, R110 (2010).
179. Campa, D. *et al*. Interactions Between Genetic Variants and Breast Cancer Risk Factors in the Breast and Prostate Cancer Cohort Consortium. *JNCI J*. *Natl*. *Cancer Inst*. **103**, 1252–1263 (2011).
180. Butt, S. *et al*. Genetic predisposition, parity, age at first childbirth and risk for breast cancer. *BMC Res*. *Notes* **5**, 414 (2012).
181. Andersen, S. W. *et al*. The associations between a polygenic score, reproductive and menstrual risk factors and breast cancer risk. *Breast Cancer Res*. *Treat*. **140**, 427–434 (2013).
182. Warren Andersen, S. *et al*. Reproductive windows, genetic loci and breast cancer risk. *Ann*. *Epidemiol*. **24**, 376–382 (2014).
183. Nickels, S. *et al*. Evidence of Gene–Environment Interactions between Common Breast Cancer Susceptibility Loci and Established Environmental Risk Factors. *PLOS Genet*.

9, e1003284 (2013).
184. Rudolph, A. *et al*. Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the Breast Cancer Association Consortium. *Int. J. Epidemiol*. **47**, 526–536 (2018).
185. Schoeps, A. *et al*. Identification of new genetic susceptibility loci for breast cancer through consideration of gene-environment interactions. *Genet. Epidemiol*. **38**, 84–93 (2014).
186. Dierssen-Sotos, T. *et al*. Reproductive risk factors in breast cancer and genetic hormonal pathways: a gene-environment interaction in the MCC-Spain project. *BMC Cancer* **18**, (2018).
187. Sinilnikova, O. M. *et al*. GENESIS: a French national resource to study the missing heritability of breast cancer. *BMC Cancer* **16**, (2016).
188. Purcell, S. *et al*. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet*. **81**, 559–575 (2007).
189. Edwards, A. W. F. G. H. Hardy (1908) and Hardy–Weinberg Equilibrium. *Genetics* **179**, 1143–1150 (2008).
190. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. **28**, 27–30 (2000).
191. Bennett, D. A. How can I deal with missing data in my study? *Aust. N. Z. J. Public Health* **25**, 464–469 (2001).
192. *Stata Statistical Software: Release 14*. (2017).
193. Rubin, D. B. Multiple imputations in sample surveys - a phenomenological bayesian approach to nonresponse. *Proc. Surv. Res. Methods Sect. Am. Stat. Assoc*. 20–28 (1978).
194. Schafer, J. L. & Graham, J. W. Missing data: Our view of the state of the art. *Psychol. Methods* **7**, 147–177 (2002).
195. van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res*. **16**, 219–242 (2007).
196. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data | Wiley Series in Probability and Statistics*. (2002).
197. Graham, J. W., Olchowski, A. E. & Gilreath, T. D. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci. Off. J. Soc. Prev. Res*. **8**, 206–213 (2007).
198. Kitts, A. & Sherry, S. *The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation*. (National Center for Biotechnology Information (US), 2011).
199. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
200. O'Connell, J. *et al*. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLOS Genet*. **10**, e1004234 (2014).
201. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet*. **5**, (2009).
202. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype Estimation Using Sequencing Reads. *Am. J. Hum. Genet*. **93**, 687–696 (2013).
203. Frazer, K. A. *et al*. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
204. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet*. **44**, 955–959 (2012).
205. *R: a language and environment for statistical computing*. https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing (2018).

206. Robert-Bobée, I. *2,1 enfants par femme pour les générations nées entre 1947 et 1963*. https://www.insee.fr/fr/statistiques/1379743.
207. Volant, S. *Un premier enfant à 28,5 ans en 2015 : 4,5 ans plus tard qu'en 1974*. https://www.insee.fr/fr/statistiques/2668280.
208. Marcus, R. P. *et al*. The evolution of radiation dose over time: Measurement of a patient cohort undergoing whole-body examinations on three computer tomography generations. *Eur. J. Radiol*. **86**, 63–69 (2017).
209. Barcellos-Hoff, M. H. New biological insights on the link between radiation exposure and breast cancer risk. *J. Mammary Gland Biol. Neoplasia* **18**, 3–13 (2013).
210. BCAC - The Breast Cancer Association Consortium —. http://bcac.ccge.medschl.cam.ac.uk/.
211. Jolliffe, I. T. *Principal Component Analysis*. (Springer-Verlag, 2002).
212. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol*. **58**, 267–288 (1996).
213. Jain, R. K. Ridge regression and its application to medical data. *Comput. Biomed. Res. Int. J*. **18**, 363–368 (1985).
214. McCarthy, S. *et al*. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet*. **48**, 1279–1283 (2016).
215. Trans-Omics for Precision Medicine (TOPMed) Program | National Heart, Lung, and Blood Institute (NHLBI). https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program.
216. Sudmant, P. H. *et al*. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
217. Escala-Garcia, M. *et al*. Genome-wide association study of germline variants and breast cancer-specific mortality. *Br. J. Cancer* **120**, 647–657 (2019).
218. Andrieu, N. & Goldstein, A. M. Epidemiologic and Genetic Approaches in the Study of Gene-Environment Interaction: an Overview of Available Methods. *Epidemiol. Rev*. **20**, 137–147 (1998).
219. Hohenlohe, P. A., Bassham, S., Currey, M. & Cresko, W. A. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philos. Trans. R. Soc. B Biol. Sci*. **367**, 395–408 (2012).
220. Kulminski, A. M. Complex phenotypes and phenomenon of genome-wide inter-chromosomal linkage disequilibrium in the human genome. *Exp. Gerontol*. **46**, 979–986 (2011).
221. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet*. **44**, 243–246 (2012).
222. Jiang, Y., Epstein, M. P. & Conneely, K. N. Assessing the Impact of Population Stratification on Association Studies of Rare Variation. *Hum. Hered*. **76**, 28–35 (2013).
223. Price, A. L. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*. **38**, 904 (2006).
224. Novembre, J. *et al*. Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
225. Karakachoff, M. *et al*. Fine-scale human genetic structure in Western France. *Eur. J. Hum. Genet*. **23**, 831–836 (2015).
226. Fullwood, M. J. *et al*. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
227. Corradin, O. *et al*. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res*. **24**, 1–13 (2014).
228. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer–promoter interactome

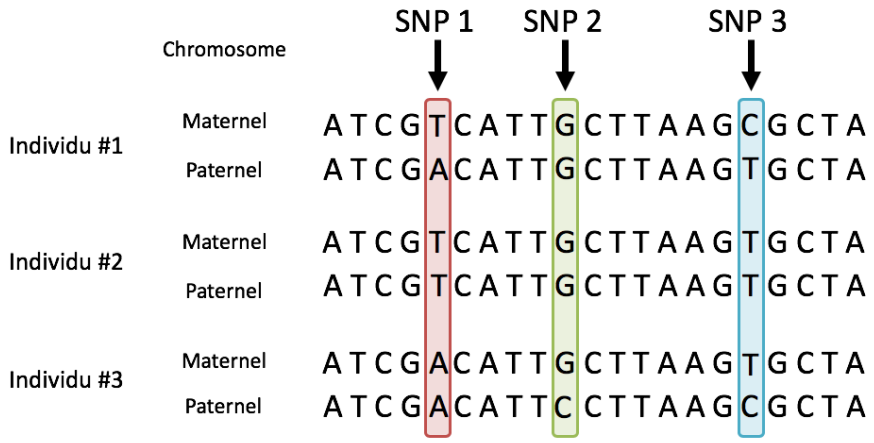
in human cells. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2191–E2199 (2014).
229. Andersson, R. *et al*. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
230. Fachal, L. *et al*. Fine-mapping of 150 breast cancer risk regions identifies 178 high confidence target genes. *bioRxiv* 521054 (2019) doi:10.1101/521054.
231. Turner, A. *et al*. MADD knock-down enhances doxorubicin and TRAIL induced apoptosis in breast cancer cells. *PloS One* **8**, e56817 (2013).
232. Zheng, T., Wang, A., Hu, D. & Wang, Y. Molecular mechanisms of breast cancer metastasis by gene expression profile analysis. *Mol. Med. Rep.* **16**, 4671–4677 (2017).
233. Sharma, D. K., Bressler, K., Patel, H., Balasingam, N. & Thakor, N. Role of Eukaryotic Initiation Factors during Cellular Stress and Cancer Progression. *J. Nucleic Acids* **2016**, (2016).
234. Whittemore, A. S. Assessing environmental modifiers of disease risk associated with rare mutations. *Hum. Hered.* **63**, 134–143 (2007).
235. Thomas, G. *et al*. A multi-stage genome-wide association in breast cancer identifies two novel risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.* **41**, 579–584 (2009).
236. Finucane, H. K. *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
237. Lin, W.-Y. *et al*. Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. *Hum. Mol. Genet.* **24**, 285–298 (2015).
238. Milne, R. L. *et al*. Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium. *Hum. Mol. Genet.* **23**, 6096–6111 (2014).
239. Haiman, C. A. *et al*. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor–negative breast cancer. *Nat. Genet.* **43**, 1210–1214 (2011).
240. Ghoussaini, M. *et al*. Evidence that the 5p12 Variant rs10941679 Confers Susceptibility to Estrogen-Receptor-Positive Breast Cancer through FGF10 and MRPS30 Regulation. *Am. J. Hum. Genet.* **99**, 903–911 (2016).
241. Glubb, D. M. *et al*. Fine-Scale Mapping of the 5q11.2 Breast Cancer Locus Reveals at Least Three Independent Risk Variants Regulating MAP3K1. *Am. J. Hum. Genet.* **96**, 5–20 (2015).
242. Siddiq, A. *et al*. A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum. Mol. Genet.* **21**, 5373–5384 (2012).
243. Dunning, A. M. *et al*. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat. Genet.* **48**, 374–386 (2016).
244. Sawyer, E. *et al*. Genetic Predisposition to In Situ and Invasive Lobular Carcinoma of the Breast. *PLOS Genet.* **10**, e1004285 (2014).
245. Turnbull, C. *et al*. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–507 (2010).
246. Orr, N. *et al*. Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. *Hum. Mol. Genet.* **24**, 2966–2984 (2015).
247. Darabi, H. *et al*. Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression. *Am. J. Hum. Genet.* **97**, 22–34 (2015).
248. Meyer, K. B. *et al*. Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1. *Am. J. Hum. Genet.* **93**, 1046–1060 (2013).
249. French, J. D. *et al*. Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. *Am. J. Hum. Genet.* **92**, 489–

503 (2013).
250. Zeng, C. *et al*. Identification of independent association signals and putative functional variants for breast cancer risk through fine-scale mapping of the 12p11 locus. *Breast Cancer Res*. *BCR* **18**, (2016).
251. Ghoussaini, M. *et al*. Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat*. *Genet*. **44**, 312–318 (2012).
252. Udler, M. S. *et al*. Fine scale mapping of the breast cancer 16q12 locus. *Hum*. *Mol*. *Genet*. **19**, 2507–2515 (2010).
253. Darabi, H. *et al*. Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). *Sci*. *Rep*. **6**, 32512 (2016).
254. Long, J. *et al*. Genome-Wide Association Study in East Asians Identifies Novel Susceptibility Loci for Breast Cancer. *PLOS Genet*. **8**, e1002532 (2012).
255. Cai, Q. *et al*. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat*. *Genet*. **46**, 886–890 (2014).
256. Long, J. *et al*. A Common Deletion in the APOBEC3 Genes and Breast Cancer Risk. *JNCI J*. *Natl*. *Cancer Inst*. **105**, 573–579 (2013).
257. Antoniou, A. C. *et al*. RAD51 135G→C Modifies Breast Cancer Risk among BRCA2 Mutation Carriers: Results from a Combined Analysis of 19 Studies. *Am*. *J*. *Hum*. *Genet*. **81**, 1186–1200 (2007).
258. Garcia-Closas, M. *et al*. Heterogeneity of Breast Cancer Associations with Five Susceptibility Loci by Clinical and Pathological Characteristics. *PLoS Genet*. **4**, e1000054 (2008).
259. Silva, L. D. & Lakhani, S. R. Pathology of hereditary breast cancer. *Mod*. *Pathol*. **23**, S46–S51 (2010).
260. HANNA, S. *et al*. StarD13 is a tumor suppressor in breast cancer that regulates cell motility and invasion. *Int*. *J*. *Oncol*. **44**, 1499–1511 (2014).
261. Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat*. *Med*. **13**, 153–162 (1994).
262. Schmidt, M. K. *et al*. Breast Cancer Survival of BRCA1/BRCA2 Mutation Carriers in a Hospital-Based Cohort of Young Women. *JNCI J*. *Natl*. *Cancer Inst*. **109**, (2017).
263. Zuk, O. *et al*. Searching for missing heritability: Designing rare variant association studies. *Proc*. *Natl*. *Acad*. *Sci*. **111**, E455–E464 (2014).
264. Girard, E. *et al*. Familial breast cancer and DNA repair genes: Insights into known and novel susceptibility genes from the GENESIS study, and implications for multigene panel testing. *Int*. *J*. *Cancer* **144**, 1962–1974 (2019).

Annexes

Single Nucleotide Polymorphism (SNP)

Un SNP est la variation d'ADN la plus fréquente entre les individus. Elle correspond à une substitution d'un nucléotide qui est retrouvée chez au moins 1 % de la population. Un SNP est présent tous les 300 nucléotides environ, ce qui revient à plusieurs millions de SNPs différents par génome. Ils représentent 90 % de l'ensemble des variations génétiques humaines et sont à la base de la diversité entre les individus.
Notre ADN est double brin, avec une copie maternelle et une copie paternelle. Pour chaque position du génome nous avons donc le même nucléotide répété deux fois, sauf à l'emplacement des SNPs où les deux copies peuvent être différentes. On parle alors de deux allèles et la combinaison de ces deux allèles forment le génotype.



Individu	SNP 1			SNP 2			SNP 3		
	Allèle 1	Allèle 2	Génotype	Allèle 1	Allèle 2	Génotype	Allèle 1	Allèle 2	Génotype
1	T	A	TA	G	G	GG	C	T	CT
2	T	T	TT	G	G	GG	T	T	TT
3	A	A	AA	G	C	GC	T	C	TC

Exemple de trois SNPs sur trois individus

En distinguant l'ADN maternel et paternel, il existe 4 combinaisons de génotypes possibles. Un SNP avec une variation de A vers T, peut avoir comme génotype AA, AT, TA ou TT. En pratique, on ne fait pas de différence entre les deux génotypes hétérozygotes AT et TA et on considère donc 3 génotypes différents. Un des deux allèles est plus fréquemment retrouvé dans la population. C'est l'allèle majeur, noté « A ». L'autre est appelé allèle mineur et est noté « a ». Lorsque l'on parle de la fréquence d'un SNP, on se réfère à la fréquence de l'allèle mineur : « Minor Allele Frequency » (MAF). La majorité des SNPs sont bi-alléliques, mais il existe également des SNPs tri- et quadri-alléliques.
En plus des substitutions il existe des délétions et des insertions, regroupées sous le terme INDELS. Ces derniers ont les mêmes caractéristiques que les SNPs, et par abus de langage nous parlerons de SNPs pour les 3 types de variations (substitutions, délétions, insertions).

Annexe 2 - Questionnaire épidémiologique de l'étude GENESIS

Etude GENESIS

Questionnaire épidémiologique

Ce questionnaire est constitué de 43 questions. Si vous ne savez pas répondre ou si vous rencontrez des difficultés, n'hésitez pas à contacter l'assistante de recherche :

Séverine Eon-Marchais
Inserm U794 & Service de Biostatistiques

Téléphone : 01 55 43 14 73
Télécopie : 01 55 43 14 69

E-mail : severine.eonmarchais@curie.net

Institut Curie
26 rue d'Ulm
75248 Paris Cedex 05
France

Si, pour certaines questions, vous ne vous souvenez vraiment plus de la réponse, merci de noter « je ne sais plus » ou un point d'interrogation afin que nous sachions que vous n'avez pas oublié d'y répondre.

Le temps moyen nécessaire pour compléter ce questionnaire est de 35 minutes.

Étude GENESIS

Réservé *Réservé*

NIPCC |__|__|__|__|__|

NC : ..... **NF :** ..... **NI :** .....

TQ |1|

1) Votre identification

Votre nom : .....
(Nom marital) *(Nom de jeune fille)*

Vos prénoms : .....

Votre date de naissance : |__|__| |__|__| |__|__|__|__| ...../...../.....

Jour *Mois* *Année*

Votre lieu de naissance : . Ville : ..... |__|__|__|__|__|

. Département : ..... |__|__|

. Pays : ..... |__|__|

Votre n° de téléphone : |__|__| |__|__| |__|__| |__|__| |__|__|

Date à laquelle vous répondez au questionnaire : |__|__| |__|__| |__|__|__|__| ...../...../.....

Jour *Mois* *Année*

2) Quel est votre statut marital actuel ?

- Vous n'avez jamais vécu en couple ☐
- Vous vivez en couple ☐ depuis combien d'années ? |__|__|
- Vous ne vivez plus en couple ☐ depuis combien d'années ? |__|__| |__| |__|__|

3) Quel est votre niveau d'études ?

- Pas d'études ☐
- Certificat d'études ☐
- BEPC-CAP ☐☐☐ ☐☐☐ ☐☐☐ ☐☐
- Bac à Bac+2 ☐
- Bac+3 à Bac+4 ☐
- Bac+5 et plus ☐ |__|

4) Êtes-vous droitière ou gauchère ?

- Droitière ☐

- Gauchère ☐
- Gauchère contrariée ☐
- Ambidextre ☐

Réservé

5) À quel groupe de population* appartenez-vous ?

- Caucasienne / Blanche ☐
- Ashkénaze ☐
- Noire-africaine ☐
- Afro-caribéenne ☐
- Arabe ☐
- Berbère ☐
- Asiatique ☐
- Indienne / Pakistanaise ☐

** Si vous êtes métisse (née d'une mère et d'un père d'origine différente), veuillez cocher les groupes dont vous êtes issue.*

6) Quelles ont été vos différentes professions (exercées pendant un an au moins) **?**

Profession ? (en clair)	*A partir de quel âge ?*	*Pendant combien d'années ?*
……	ans	
……	ans	
……	ans	
……	ans	
……	ans	
……	ans	

7) Quelle est la dernière profession de votre conjoint(e) actuel(le) ?

(avant sa retraite si il (ou elle) est retraité(e)) : .....

8) Quel est votre poids actuel ?

kg

Si vous avez eu un cancer du sein, indiquez votre poids avant le diagnostic de ce cancer :

kg

9) Quelle est votre taille ? m cm

Réservé

10) Quelle a été votre consommation d'alcool au cours de votre vie ?

Pour chaque période de votre vie, cochez la case correspondant au nombre de verres d'alcool consommés par semaine. (*On se base sur un verre de table classique, type duralex. Pour information une canette équivaut à 2 verres).*

- Votre consommation de cidre, bière, vin, vin cuit, etc.

	Nombre de verres par semaine					
Période	0	Moins de 1	1 à 5	6 à 10	Plus de 10	
Avant l'âge de 18 ans :	☐	☐	☐	☐	☐	☐
De 18 à 30 ans :	☐	☐	☐	☐	☐	☐
De 31 à 40 ans :	☐	☐	☐	☐	☐	☐
De 41 à 50 ans :	☐	☐	☐	☐	☐	☐
De 51 à 60 ans :	☐	☐	☐	☐	☐	☐
De 61 à 70 ans :	☐	☐	☐	☐	☐	☐
Après 70 ans :	☐	☐	☐	☐	☐	☐

- Votre consommation d'alcools forts type whisky, vodka, pastis, digestif, etc.

	Nombre de verres par semaine					
Période	0	Moins de 1	1 à 5	6 à 10	Plus de 10	
Avant l'âge de 18 ans :	☐	☐	☐	☐	☐	☐
De 18 à 30 ans :	☐	☐	☐	☐	☐	☐
De 31 à 40 ans :	☐	☐	☐	☐	☐	☐
De 41 à 50 ans :	☐	☐	☐	☐	☐	☐
De 51 à 60 ans :	☐	☐	☐	☐	☐	☐
De 61 à 70 ans :	☐	☐	☐	☐	☐	☐
Après 70 ans :	☐	☐	☐	☐	☐	☐

Réservé

11) Fumez-vous actuellement ?

Oui ☐ Non ☐ |__|

Si oui, - Combien de cigarettes par jour ? |__|__| |__|__|

- Depuis quel âge ? |__|__| ans |__|__|

- Pendant combien d'années ?
(déduire les années d'interruption) |__|__| |__|__|

Si non, - Avez-vous déjà fumé ? Oui ☐ Non ☐ |__|

Si oui, . Combien de cigarettes par jour ? |__|__| |__|__|

. Pendant combien d'années ?
(déduire les années d'interruption) |__|__| |__|__|

. A quel âge avez-vous commencé à fumer ? |__|__| ans |__|__|

. A quel âge avez-vous cessé de fumer ? |__|__| ans |__|__|

Réservé

12) Êtes-vous ou avez-vous été exposée au tabagisme passif (exposition dans un espace clos à la fumée de tabac) **?**

Oui ☐ Non ☐

Si oui, **précisez à quelles occasions vous avez été exposée :**

- **Votre mère fumait pendant qu'elle était enceinte de vous** Oui ☐ Non ☐

Si oui, pouvez vous préciser :

. Nombre de cigarettes par jour ?

- **Votre mère et/ou votre père fumaient pendant votre enfance** Oui ☐ Non ☐

Si oui, précisez votre exposition :

. Nombre d'heures par semaine ?

. A partir de quel âge ? ans

. Nombre d'années d'exposition ?

- **Votre conjoint et/ou vos enfants fument ou ont fumé** Oui ☐ Non ☐

Si oui, précisez votre exposition :

. Nombre d'heures par semaine ?

. A partir de quel âge ? ans

. Nombre d'années d'exposition ?

- **Vous êtes ou avez été professionnellement exposée** Oui ☐ Non ☐

(ex : travail dans un café/restaurant, collègues qui fument ou ont fumé dans votre bureau, etc.)

Si oui, précisez votre exposition :

. Nombre d'heures par semaine ?

. A partir de quel âge ? ans

. Nombre d'années d'exposition ?

- **Vous êtes ou avez été exposée à d'autres occasions** Oui ☐ Non ☐

(ex : sorties dans un café/restaurant, une discothèque, etc.)

Si oui, précisez votre exposition :

. Nombre d'heures par semaine ?

. A partir de quel âge ? ans

. Nombre d'années d'exposition ?

Réservé

13) Quelle est votre activité physique ?

Complétez les tableaux **a et b** en cochant **pour chaque intensité**, la fréquence qui correspond le mieux à votre activité. Indiquez en dessous de la case cochée, le nombre d'heures par semaine (h/sem) en moyenne pendant lesquelles vous exercez une activité physique.

a - Au cours de votre activité professionnelle et/ou de votre vie quotidienne ?

	Fréquence *(cochez une seule case par ligne)*				
	jamais	de manière occasionnelle *(moins d'1 fois par semaine)*	de manière régulière *(1 à 3 fois par semaine)*	de manière quotidienne *(4 à 7 fois par semaine)*	
Intensité légère : *Ex : marcher sans se presser, monter les escaliers, effectuer des travaux légers de ménage, etc.*	☐	☐ ___ h/sem	☐ ___ h/sem	☐ ___ h/sem	___ ___
Intensité moyenne : *Ex : marcher d'un bon pas, aller au travail en vélo, travaux de ménage qui nécessitent des efforts, etc.*	☐	☐ ___ h/sem	☐ ___ h/sem	☐ ___ h/sem	___ ___
Intensité élevée : *Ex : travail physique, monter les escaliers en courant, etc.*	☐	☐ ___ h/sem	☐ ___ h/sem	☐ ___ h/sem	___ ___

b – Au cours de vos loisirs (de type sportif et autres) ?	Fréquence *(cochez une seule case par ligne)*				*Réservé*
	jamais	de manière occasionnelle *(moins d'1 fois par semaine)*	de manière régulière *(1 à 3 fois par semaine)*	de manière quotidienne *(4 à 7 fois par semaine)*	
Intensité légère : *Ex : faire des exercices d'étirement, effectuer des travaux légers de jardinage, promener son chien, etc.*	☐	☐ h/sem	☐ h/sem	☐ h/sem	
Intensité moyenne : *Ex : nager ou danser sans forcer, faire de la gymnastique à son domicile, travaux de jardinage qui nécessitent des efforts, etc.*	☐	☐ h/sem	☐ h/sem	☐ h/sem	
Intensité élevée : *Ex : faire de la gymnastique dans une salle de sport, faire du jogging, nager ou danser à un rythme soutenu, faire du vélo d'appartement, faire un sport (tennis, handball…), etc.*	☐	☐ h/sem	☐ h/sem	☐ h/sem	

Réservé

ANTECEDENTS PERSONNELS DE MALADIE

14) Avez-vous des antécédents personnels de maladies cardio-vasculaires ?

Oui ☐ Non ☐ |__|

Si oui, précisez :

			A quel âge ?	
Hypertension artérielle	Oui ☐	Non ☐	|__|__| ans	|__| |__|__|
Hypotension artérielle	Oui ☐	Non ☐	|__|__| ans	|__| |__|__|
Phlébite	Oui ☐	Non ☐	|__|__| ans	|__| |__|__|
Embolie pulmonaire	Oui ☐	Non ☐	|__|__| ans	|__| |__|__|
Infarctus du myocarde	Oui ☐	Non ☐	|__|__| ans	|__| |__|__|
Angine de poitrine	Oui ☐	Non ☐	|__|__| ans	|__| |__|__|
Accident vasculaire cérébral	Oui ☐	Non ☐	|__|__| ans	|__| |__|__|
Autre (en clair) :	Oui ☐	Non ☐	|__|__| ans	|__| |__|__|

15) Au cours de votre vie, avez-vous eu une hyper-cholestérolémie (taux élevé de cholestérol dans le sang) ?

Oui ☐ Non ☐ Ne sait pas ☐ |__|

Si oui,

- A partir de quel âge ? |__|__| ans |__|__|
- Pendant combien d'années au total ? |__|__| |__|__|
  (périodes de traitement incluses)

16) Avez-vous des antécédents personnels de diabète (hors diabète gestationnel) **?**

Oui ☐ Non ☐ |__|

Si oui, quel type de diabète ?

- Insulinodépendant ☐ *Date de diagnostic* |__|__| *Mois* |__|__|__|__| *Année* |__| ……/……
- Non insulinodépendant ☐ *Date de diagnostic* |__|__| *Mois* |__|__|__|__| *Année* |__| ……/……

Réservé

17) Avez-vous eu une maladie bénigne du sein ?

Oui ☐ Non ☐ |__|

Si oui, précisez :

			A quel sein ? *(droit=D, gauche=G, les deux=DG)*	***A quel âge ?***	
Kyste	Oui ☐	Non ☐	|__|__|	|__|__| ans	|__|__|__|__| |__|__|__|__| |__|
Maladie fibrokystique	Oui ☐	Non ☐	|__|__|	|__|__| ans	|__|__|__|__| |__|__|__|__| |__|
Abcès au sein	Oui ☐	Non ☐	|__|__|	|__|__| ans	|__|__|__|__| |__|__|__|__| |__|
Mastose	Oui ☐	Non ☐	|__|__|	|__|__| ans	|__|__|__|__| |__|__|__|__| |__|
Micro calcifications	Oui ☐	Non ☐	|__|__|	|__|__| ans	|__|__|__|__| |__|__|__|__| |__|
Autres	Oui ☐	Non ☐	|__|__|	|__|__| ans	|__|__|__|__| |__|__|__|__| |__|

Si autres, précisez : .....

- Où avez-vous été examinée ?
  - . Ville, département : .....
  - . Etablissement : .....
  - . Nom du médecin : .....
  - . En quelle année : |__|__|__|__|
  - . Quel était votre nom usuel à cette époque ? :.....

18) Avez-vous eu une maladie bénigne de l'utérus ?					Réservé
				Oui ☐ Non ☐	
Si oui, précisez :			***A quel âge ?***	***Quels traitements ?***	
Fibrome utérin	Oui ☐	Non ☐	ans	……	
Endométriose	Oui ☐	Non ☐	ans	……	
Polypes ou hyperplasie	Oui ☐	Non ☐	ans	……	
Malformations	Oui ☐	Non ☐	ans	……	
Autres	Oui ☐	Non ☐	ans	……	
Si autres, précisez : ……					

Réservé

19) Au cours de votre vie, avez-vous eu une maladie bénigne de la thyroïde ou une intervention chirurgicale à la thyroïde ?

Oui ☐ Non ☐ ☐

Si oui, précisez (plusieurs réponses possibles) :

			A quel âge ?	
Hyperthyroïdie	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Hypothyroïdie	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Adénome	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Nodule	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Kyste	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Goitre	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Maladie de Basedow	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Maladie de Hashimoto	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Ablation partielle de la thyroïde	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Ablation totale de la thyroïde	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Autres (précisez)	Oui ☐	Non ☐	☐☐ ans	☐ ☐☐☐
Ne sait pas	☐		☐☐ ans	☐ ☐☐☐

- Où avez-vous été examinée ?
  - . Ville, département : .....
  - . Etablissement : .....
  - . Nom du médecin : .....
  - . En quelle année : ☐☐☐☐
  - . Quel était votre nom usuel à cette époque ? :.....

Réservé

20) Avez-vous eu un (ou des) cancer(s) ?

Oui ☐ Non ☐

Si oui, précisez

(Si plusieurs cancers, remplissez une rubrique par localisation) :

1ère localisation (ex : sein, colon, poumon, etc.) : ……

S'il s'agit d'un organe double (ex : seins, reins) précisez le côté de l'organe atteint (droit=D, gauche=G, les deux=DG)

- En quelle année ?

- Où avez-vous été soignée ?
  - . Ville, département : .....
  - . Etablissement : .....
  - . Nom du médecin : .....
  - . Quel était votre nom usuel à cette époque ? :.....

2ème localisation (ex : sein, colon, poumon, etc.) : ……

S'il s'agit d'un organe double (ex : seins, reins) précisez le côté de l'organe atteint (droit=D, gauche=G, les deux=DG)

- En quelle année ?

- Où avez-vous été soignée ?
  - . Ville, département : .....
  - . Etablissement : .....
  - . Nom du médecin : .....
  - . Quel était votre nom usuel à cette époque ? :.....

Réservé

21) Avez-vous pris des médicaments pendant longtemps (un an ou plus) pour le traitement d'une maladie ?

Oui ☐ Non ☐ |__|

Si oui, précisez :

Nom du médicament ? *(en clair)*	***Pour quelle maladie ?***	***Age au début du traitement?***	***Durée du traitement ?*** *(en mois)*	
.....	.....	|__|__| ans	|__|__|__|	|__|__|__||__|__|__||__|__|__|__|
.....	.....	|__|__| ans	|__|__|__|	|__|__|__||__|__|__||__|__|__|__|
.....	.....	|__|__| ans	|__|__|__|	|__|__|__||__|__|__||__|__|__|__|
.....	.....	|__|__| ans	|__|__|__|	|__|__|__||__|__|__||__|__|__|__|
.....	.....	|__|__| ans	|__|__|__|	|__|__|__||__|__|__||__|__|__|__|

ANTECEDENTS GYNECO-OBSTETRIQUES

22) A quel âge avez-vous eu vos premières règles ? |__|__| |__|__|__|

- Si vous ne savez pas précisément :
  - . Plutôt jeune (avant 12 ans) ☐
  - . Plutôt moyen (entre 12 et 14 ans) ☐
  - . Plutôt tard (après 15 ans) ☐
  - . Vous n'avez jamais eu de règles ☐
  - . Vous ne savez plus ☐ |__|

23) Quelle est (ou était) la périodicité de vos règles (en dehors des contraceptifs oraux) **?**

- . Régulière avec des cycles de 25 à 31 jours ☐
- . Régulière avec des cycles de 24 jours et moins ☐
- . Régulière avec des cycles de 32 jours et plus ☐
- . Irrégulière ☐
- . Vous ne savez plus ☐ |__|

24) Avez-vous déjà été enceinte ?	*Réservé*
Oui ☐ Non ☐	☐
Si oui, - Combien de fois avez-vous été enceinte ?	☐☐
- Combien d'enfants nés vivants avez-vous eu ?	☐☐

Complétez les tableaux ci-dessous pour chacune de vos grossesses en commençant par la première.

a - Grossesses ayant abouti à la naissance d'un enfant (vivant ou mort-né) ou grossesse en cours :

Grossesse	*Age au début de la grossesse*	*Durée de la grossesse* ***en mois***	*Date de naissance mois/an*	*Enfant né vivant (oui/non) ou grossesse en cours*	*Grossesse multiple notez le nombre d'enfants* *jumeaux = 2 triplés = 3 …*	*Allaitement*	
						oui/non	*durée en mois*
exemple	24	9 mois	09/2004	oui	1	oui	2,5
exemple	26	-	-	en cours	2	-	-
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							

b - Grossesses interrompues suite à une fausse couche, une IVG (Interruption Volontaire de Grossesse), **une ITG** (Interruption Thérapeutique de Grossesse) ou **une GEU** (Grossesse Extra-Utérine)

Grossesse	*Age au début de la grossesse*	*Durée de la grossesse* ***en semaines***	*Date d'interruption mois/an*	*Interruption par une fausse couche oui/non*	*Interruption par une IVG, ITG ou GEU oui/non*
exemple	22	4 semaines	12/2002	non	oui
1b					
2b					
3b					
4b					

Réservé

25) Avez-vous déjà utilisé des contraceptifs hormonaux (pilule, patch, implant sous-cutané, stérilet Mirena®, etc.) **?**

(Si vous avez eu un cancer du sein, répondez pour la période avant votre cancer).

Oui ☐ Non ☐ |__|

***Si oui*,**

- A quel âge avez-vous pris des contraceptifs hormonaux pour la 1ère fois ? |__|__| ans |__|__|

- Si vous n'avez jamais été enceinte :

- Pendant combien de temps avez-vous pris des contraceptifs hormonaux (durée totale en mois) ? |__|__|__| |__|__|__|

- Si vous avez été enceinte, avez-vous pris des contraceptifs hormonaux :

• Avant votre 1ère grossesse ? (qu'elle ait abouti à la naissance d'un enfant ou qu'elle ait été interrompue)

Oui ☐ Non ☐ |__|

Si oui, pendant combien de mois ? |__|__|__| |__|__|__|

• Après votre 1ère grossesse et avant votre dernière grossesse ?

Oui ☐ Non ☐ |__|

Si oui, pendant combien de mois ? |__|__|__| |__|__|__|

• Après votre dernière grossesse ? Oui ☐ Non ☐ |__|

Si oui, pendant combien de mois ? |__|__|__| |__|__|__|

Réservé

26) Avez-vous pris au cours de votre vie, ou prenez-vous actuellement un traitement hormonal (hormones sexuelles) **pour d'autres raisons que la contraception et la ménopause?**

Par exemple, les traitements pour favoriser la survenue d'une grossesse (et ce, que la grossesse ait eu lieu ou non) : fécondation *in vitro*, prise d'inducteurs d'ovulation (type Clomid®), etc. ; pour régulariser les cycles ; pour traiter des maladies bénignes (utérus, sein, ovaire, etc.)

Oui ☐ Non ☐

Si oui, précisez :

Age au début du traitement ?	***Nom du (ou des) médicament(s) ?***	***Raison du traitement ?*** *(infertilité, etc.)*	***Nombre de jours par mois ?***	***Durée du traitement ?*** *(en mois)*
ans	.....	.....		
ans	.....	.....		
ans	.....	.....		
ans	.....	.....		
ans	.....	.....		

27) Avez-vous eu une hystérectomie (ablation de l'utérus) **?**

Oui ☐ Non ☐

Si oui :

- A quel âge ? ans
- Quelle en était la cause ?

(ex : endométriose, fibrome, GEU, etc.) ……

28) Avez-vous eu une ovariectomie (ablation d'un ou des deux ovaires) **?**

Oui ☐ Non ☐

Si oui, était-ce une ablation de :			***A quel âge ?***	***Quelle en était la cause ?*** *(ex : kyste, tumeur, etc.)*
- L'ovaire gauche	Oui ☐	Non ☐	ans	……..
- L'ovaire droit	Oui ☐	Non ☐	ans	……..

Réservé

29) Etes-vous ménopausée ? (sans règles depuis 1 an ou plus)

Oui ☐ Non ☐ |__|

Si oui, depuis quel âge ? ☐|__|__| ans |__|__|

Est-ce ? (plusieurs réponses possibles)

- Ménopause naturelle ☐
- Ménopause artificielle par hystérectomie *(ablation de l'utérus)* ☐
- Ménopause artificielle par ovariectomie *(ablation des ovaires)* ☐
- Ménopause artificielle par chimiothérapie ☐
- Ménopause artificielle par radiothérapie ☐
- Ménopause artificielle par radiothérapie et chimiothérapie ☐
- Ménopause artificielle autre (précisez) ……. ☐ |__|

……..……..

30) Avez-vous pris ou prenez-vous un traitement hormonal pour la ménopause ?

Oui ☐ Non ☐ |__|

Si oui, précisez :

Age au début du traitement ?	*Nom du ou des médicament(s) ?*	*Durée (en mois) ?*	
|__|__| ans	.....	|__|__|__|	|__|__| |__|__||__|__|__|
|__|__| ans	.....	|__|__|__|	|__|__| |__|__||__|__|__|
|__|__| ans	.....	|__|__|__|	|__|__| |__|__||__|__|__|
|__|__| ans	.....	|__|__|__|	|__|__| |__|__||__|__|__|
|__|__| ans	.....	|__|__|__|	|__|__| |__|__||__|__|__|
|__|__| ans	.....	|__|__|__|	|__|__| |__|__||__|__|__|

31) Avez-vous pris ou prenez-vous des phytoestrogènes (oestrogènes végétaux) **pour la ménopause ?** (Phytosoya®, Ymea®, Sojyam®, Evestrel®, etc.)

Oui ☐ Non ☐ |__|

Si oui, précisez

Age au début du traitement ?	*Nom du ou des médicament(s) ?*	*Durée (en mois)?*	
|__|__| ans	.....	|__|__|__|	|__|__| |__|__||__|__|__|
|__|__| ans	.....	|__|__|__|	|__|__| |__|__||__|__|__|
|__|__| ans	.....	|__|__|__|	|__|__| |__|__||__|__|__|

Réservé

EXPOSITIONS AUX IRRADIATIONS

32) Avez-vous exercé ou exercez-vous actuellement une profession où des rayons X ou des éléments radioactifs étaient ou sont utilisés ?

Oui ☐ Non ☐ ☐

Si oui :

. Pendant combien d'années ? ☐☐ ☐☐

. A partir de quel âge ? ☐☐ ans ☐☐

. Portez-vous ou avez-vous porté un badge (dosimètre individuel) ? Oui ☐ Non ☐ ☐

. Dans quel type d'établissement / entreprise ?

- Un hôpital, une clinique, une PMI *(Protection Maternelle et Infantile)* ? Oui ☐ Non ☐ ☐
- Chez un dentiste ? Oui ☐ Non ☐ ☐
- Un institut de recherche ? Oui ☐ Non ☐ ☐
- Une industrie utilisant des substances radioactives *(centre nucléaire, etc.)* ? Oui ☐ Non ☐ ☐
  Précisez : ……
- Une autre structure ? Oui ☐ Non ☐ ☐
  Précisez : ……

MEDECINE PREVENTIVE OU DU TRAVAIL

33) Avez-vous eu des examens pulmonaires de type radiographique (radiographie ou scopie) **dans le cadre de la <u>médecine préventive</u>** (à l'école, examens prénuptiaux) **ou dans le cadre de la <u>médecine du travail</u> ?**

(Si vous avez eu un cancer du sein, répondez pour la période avant votre cancer).

Oui ☐ Non ☐ ☐

Si oui :

. A quel âge avez-vous eu le 1er examen pulmonaire ? ☐☐ ans ☐☐

. Combien en avez-vous eu au total ? ☐☐☐ ☐☐☐
(avant votre cancer du sein, si vous en avez eu un)

. A combien de temps remonte le dernier examen *(en mois) ?* ☐☐☐ ☐☐☐
(avant votre cancer du sein, si vous en avez eu un)

Réservé

EXAMENS RADIOGRAPHIQUES POUR PATHOLOGIES

(Ne pas tenir compte des échographies ni des IRM)

34) Avez-vous eu des examens de type radiographique (radiographie, scopie, tomographie, bronchographie, scintigraphie, angiographie, scanner ou autre) **pour des pathologies pulmonaires** (tuberculose, primo-infection, corps étrangers dans l'appareil respiratoire, dilatation des bronches, bronchite chronique, emphysème, fibrose pulmonaire ou autre) **ou pour un bilan pré-opératoire en vue d'une anesthésie générale ?**

Oui ☐ Non ☐ ⌊__⌋

Si oui, précisez :

A quel âge ?	*Quel type d'examen ?* (en clair)	*Pour quelle raison ?* (en clair)	
Ex : ⌊2⊥5⌋ ans	Radiographie	Bilan préopératoire	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋

35) Avez-vous eu des examens de type radiographique (scintigraphie osseuse, myélographie, radiographie, tomographie, scanner ou autre) **aux épaules, aux côtes ou à la colonne vertébrale** (fractures, scolioses, traumatismes, arthrose, etc.) **?**

Oui ☐ Non ☐ ⌊__⌋

Si oui, précisez :

A quel âge ?	*Quel type d'examen ?* (en clair)	*Pour quelle raison ?* (en clair)	
Ex □□: ⌊3⊥5⌋ ans	Radiographie	Fracture	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋
⌊__⊥__⌋ ans	.....	.....	⌊__⊥__⌋ ⌊__⊥__⌋

Réservé

36) Avez-vous eu des examens panoramiques aux dents ? (clichés de l'ensemble de la dentition)

Oui ☐ Non ☐

Si oui, précisez :

A quel âge ?	***Pour quelle raison ?*** *(en clair)*
ans	.....
ans	.....
ans	.....
ans	.....
ans	.....

37) Avez-vous eu des examens de type radiographique (scintigraphie, scanner) **pour une maladie de la thyroïde** (hypo ou hyperthyroïdie) ?

Oui ☐ Non ☐

Si oui, précisez :

A quel âge ?	***Quel type d'examen ?*** *(en clair)*	***Pour quelle raison ?*** *(en clair)*
ans	.....	.....
ans	.....	.....
ans	.....	.....

38) Avez-vous eu des examens de type radiographique (radiographie, scopie, scintigraphie, coronarographie, artériographie, phlébographie ou autre) **au cœur et aux vaisseaux thoraciques** (infarctus, embolie pulmonaire, malformation cardiaque, souffle au cœur, etc.) **?**

Oui ☐ Non ☐

Si oui, précisez :

A quel âge ?	***Quel type d'examen ?*** *(en clair)*	***Pour quelle raison ?*** *(en clair)*
ans	.....	.....
ans	.....	.....
ans	.....	.....
ans	.....	.....
ans	.....	.....
ans	.....	.....
ans	.....	.....

Réservé

EXAMENS RADIOGRAPHIQUES AUX SEINS

39) Avez-vous eu des mammographies (radio des seins) ?

(Si vous avez eu un cancer du sein, répondez pour la période avant votre cancer).

Oui ☐ Non ☐ |__|

Si oui, précisez :

	A quel âge ?	***A quel sein ?*** *(droit=D, gauche=G, les deux=DG)*	***Pour quelle raison ?***	***Nombre total de clichés ?*** *(sein D + sein G)*	
1ère mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
2e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
3e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
4e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
5e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
6e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
7e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
8e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
9e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
10e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
11e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
12e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
13e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
14e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|
15e mammo	|__|__| ans	|__|__|	.....	|__|__|	|__|__||__||__||__||__|__|

Réservé

TRAITEMENTS PAR RAYONNEMENTS IONISANTS

(radiothérapie, curiethérapie, iode radioactif, etc.)

40) Avez-vous déjà été soignée par des rayons pour une maladie bénigne (angiomes, etc.)?

(Si vous avez eu un cancer du sein, répondez pour la période avant votre cancer).

Oui ☐ Non ☐

Si oui, précisez :

	OUI	NON	*A quel âge ?*	*Pour quelle raison ? (en clair)*
Tête (cerveau, maxillaires, végétations, etc.)	☐	☐	ans	.....
Cou (thyroïde, pharynx, etc.)	☐	☐	ans	.....
Poumon droit	☐	☐	ans	.....
Poumon gauche	☐	☐	ans	.....
Sein droit	☐	☐	ans	.....
Sein gauche	☐	☐	ans	.....
Ovaire droit	☐	☐	ans	.....
Ovaire gauche	☐	☐	ans	.....
Utérus	☐	☐	ans	.....
Peau	☐	☐	ans	.....
Autres (en clair)	☐	☐	ans	.....

.....

41) Avez-vous déjà été soignée par des rayons pour une maladie maligne ?

(Si vous avez eu un cancer du sein, répondez pour la période avant votre cancer).

Oui ☐ Non ☐

Si oui, précisez :

	OUI	NON	*A quel âge ?*	*Pour quelle raison ? (en clair)*
Tête (cerveau, maxillaires, végétations, etc.)	☐	☐	ans	.....
Cou (thyroïde, pharynx, etc.)	☐	☐	ans	.....
Poumon droit	☐	☐	ans	.....
Poumon gauche	☐	☐	ans	.....
Ovaire droit	☐	☐	ans	.....
Ovaire gauche	☐	☐	ans	.....
Utérus	☐	☐	ans	.....
Peau	☐	☐	ans	.....
Autres (en clair)	☐	☐	ans	.....

.....

COMPOSITION DE VOTRE FAMILLE ET ANTECEDENTS FAMILIAUX DE CANCER

42) Complétez les tableaux ci-dessous concernant la composition de votre famille (vos frères et sœurs, parents et grands-parents, oncles et tantes paternels et maternels) et les antécédents de cancer dans votre famille.

Vos Frères et Sœurs - *ne pas mentionner vos demi-frères et demi-sœurs* -

Prénom	**Sexe** F=Féminin M=Masculin	**Année de naissance**	**Cette personne a-t-elle eu un cancer ?**	**Si oui :** Age au diagnostic et localisation du cancer *Age* *Localisation(s)*	**Cette personne est-elle décédée ?**	**Si oui :** Age au décès ?	**Si non :** Age aux dernières nouvelles ?
1.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
2.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
3.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
4.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
5.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐

Si vous n'avez pas assez de place pour tous vos frères et sœurs, continuez en page 28

Votre famille paternelle

Prénom	**Sexe** F=Féminin M=Masculin	**Année de naissance**	**Cette personne a-t-elle eu un cancer ?**	**Si oui :** Age au diagnostic et localisation du cancer *Age* *Localisation(s)*	**Cette personne est-elle décédée ?**	**Si oui :** Age au décès ?	**Si non :** Age aux dernières nouvelles ?
Votre Père							
----------	[M]	[_][_][_][_]	Oui ☐ Non ☐ Ne sait pas ☐	1. [_][_] ---------- 2. [_][_] ----------	Oui ☐ Non ☐ Ne sait pas ☐	[_][_]	[_][_]
Votre Grand-père paternel							
----------	[M]	[_][_][_][_]	Oui ☐ Non ☐ Ne sait pas ☐	1. [_][_] ---------- 2. [_][_] ----------	Oui ☐ Non ☐ Ne sait pas ☐	[_][_]	[_][_]
Votre Grand-mère paternelle							
----------	[F]	[_][_][_][_]	Oui ☐ Non ☐ Ne sait pas ☐	1. [_][_] ---------- 2. [_][_] ----------	Oui ☐ Non ☐ Ne sait pas ☐	[_][_]	[_][_]

Votre famille paternelle (suite)

Prénom	**Sexe** F=Féminin M=Masculin	**Année de naissance**	**Cette personne a-t-elle eu un cancer ?**	**Si oui :** Age au diagnostic et localisation du cancer *Age* *Localisation(s)*	**Cette personne est-elle décédée ?**	**Si oui :** Age au décès ?	**Si non :** Age aux dernières nouvelles ?
Vos Oncles et Tantes paternels (frères et sœurs de votre père) - *Ne pas mentionner les demi-frères et demi-sœurs de votre père* -							
1.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
2.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
3.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
4.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
5.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
6.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐

Si vous n'avez pas assez de place pour tous vos oncles et tantes paternels, continuez en page 28

Votre famille maternelle

Prénom	**Sexe** F=Féminin M=Masculin	**Année de naissance**	**Cette personne a-t-elle eu un cancer ?**	**Si oui :** Age au diagnostic et localisation du cancer *Age* *Localisation(s)*	**Cette personne est-elle décédée ?**	**Si oui :** Age au décès ?	**Si non :** Age aux dernières nouvelles ?
Votre Mère							
----------	F		Oui ☐ Non ☐ Ne sait pas ☐	1. ___ ---------- 2. ___ ----------	Oui ☐ Non ☐ Ne sait pas ☐	___	___
Votre Grand-père maternel							
----------	M		Oui ☐ Non ☐ Ne sait pas ☐	1. ___ ---------- 2. ___ ----------	Oui ☐ Non ☐ Ne sait pas ☐	___	___
Votre Grand-mère maternelle							
----------	F		Oui ☐ Non ☐ Ne sait pas ☐	1. ___ ---------- 2. ___ ----------	Oui ☐ Non ☐ Ne sait pas ☐	___	___

Votre famille maternelle (suite)

Prénom	**Sexe** F=Féminin M=Masculin	**Année de naissance**	**Cette personne a-t-elle eu un cancer ?**	**Si oui :** Age au diagnostic et localisation du cancer *Age* *Localisation(s)*	**Cette personne est-elle décédée ?**	**Si oui :** Age au décès ?	**Si non :** Age aux dernières nouvelles ?
Vos Oncles et Tantes maternels (frères et soeurs de votre mère) : - *Ne pas mentionner les demi-frères et demi-sœurs de votre mère* -							
1.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
2.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
3.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
4.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
5.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
6.----------	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐

Si vous n'avez pas assez de place pour tous vos oncles et tantes maternels, continuez en page 28

Si vous n'avez pas assez de place pour tous les membres de votre famille, utilisez le tableau ci-dessous.

Prénom	**Lien de parenté** 1 : frère-sœur, 2 : oncle-tante paternel 3 : oncle-tante maternel	**Sexe** **F=Féminin** **M=Masculin**	**Année de naissance**	**Cette personne a-t-elle eu un cancer ?**	**Si oui :** Age au diagnostic et localisation du cancer *Age* *Localisation(s)*	**Cette personne est-elle décédée ?**	**Si oui :** Age au décès ?	**Si non :** Age aux dernières nouvelles ?
1.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
2.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
3.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
4.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
5.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
6.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐

Prénom	**Lien de parenté** 1 : frère-sœur, 2 : oncle-tante paternel 3 : oncle-tante maternel	**Sexe** **F=Féminin** **M=Masculin**	**Année de naissance**	**Cette personne a-t-elle eu un cancer ?**	**Si oui :** Age au diagnostic et localisation du cancer *Age* *Localisation(s)*	**Cette personne est-elle décédée ?**	**Si oui :** Age au décès ?	**Si non :** Age aux dernières nouvelles ?
7.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
8.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
9.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
10.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
11.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐
12.----------	☐	☐	☐☐☐☐	Oui ☐ Non ☐ Ne sait pas ☐	1. ☐☐ ---------- 2. ☐☐ ----------	Oui ☐ Non ☐ Ne sait pas ☐	☐☐	☐☐

43) Si des événements de votre vie n'ont pas été évoqués dans ce questionnaire, ou si vous avez des commentaires à nous communiquer, notez les dans cette rubrique :

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

Nous vous remercions vivement de participer à cette étude et d'avoir pris le temps de compléter ce questionnaire.

Merci de vérifier que vous avez répondu à toutes les questions et de nous renvoyer ce questionnaire en utilisant la grande enveloppe T jointe.

Le groupe de l'Etude GENESIS

Dans le cadre de la recherche biomédicale à laquelle la Fédération Nationale des Centres de Lutte Contre le Cancer (FNCLCC) vous propose de participer, un traitement de vos données personnelles va être mis en œuvre par l'équipe de l'unité Inserm U794 localisée dans le Service de Biostatistiques de l'Institut Curie pour permettre d'analyser les résultats de la recherche au regard de l'objectif de cette dernière qui vous a été présenté. À cette fin, les données médicales vous concernant et les données relatives à vos habitudes de vie, ainsi que, dans la mesure où ces données sont nécessaires à la recherche, vos origines ethniques, seront transmises au Promoteur de la recherche ou aux personnes ou sociétés agissant pour son compte, en France ou à l'étranger. Ces données seront identifiées par un numéro de code. Ces données pourront également, dans les conditions assurant leur confidentialité, être transmises aux autorités de santé françaises ou étrangères. Conformément aux dispositions de loi relative à l'informatique aux fichiers et aux libertés, vous disposez d'un droit d'accès et de rectification. Vous disposez également d'un droit d'opposition à la transmission des données couvertes par le secret professionnel susceptibles d'être utilisées dans le cadre de cette recherche et d'être traitées. Vous pouvez également accéder directement ou par l'intermédiaire d'un médecin de votre choix à l'ensemble de vos données médicales en application des dispositions de l'article L. 1111-7 du Code de la Santé Publique. Ces droits s'exercent auprès du médecin qui vous suit dans le cadre de la recherche et qui connaît votre identité.

Annexe 3 - Gènes impliqués dans les voies de signalisation des hormones

Gène	Voies Biologiques KEGG
ABCC8	Insulin secretion
ACACA	Insulin signaling pathway
ACACB	Insulin resistance, Insulin signaling pathway
ACTB	Thyroid hormone signaling pathway Oxytocin signaling pathway
ACTG1	Thyroid hormone signaling pathway Oxytocin signaling pathway
ADCY1	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
ADCY2	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
ADCY3	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
ADCY4	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
ADCY5	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
ADCY6	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion

	Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
ADCY7	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
ADCY8	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
ADCY9	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
ADCYAP1	Insulin secretion
ADCYAP1R1	Insulin secretion
AGTR1	Cortisol synthesis and secretion
AKR1C1	Steroid hormone biosynthesis
AKR1C2	Steroid hormone biosynthesis
AKR1C3	Steroid hormone biosynthesis Ovarian steroidogenesis
AKR1C4	Steroid hormone biosynthesis
AKR1D1	Steroid hormone biosynthesis
AKT1	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
AKT2	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
AKT3	Estrogen signaling pathway

	Prolactin signaling pathway Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
ALB	Thyroid hormone synthesis
ALOX5	Ovarian steroidogenesis
ANAPC1	Progesterone-mediated oocyte maturation
ANAPC10	Progesterone-mediated oocyte maturation
ANAPC11	Progesterone-mediated oocyte maturation
ANAPC13	Progesterone-mediated oocyte maturation
ANAPC2	Progesterone-mediated oocyte maturation
ANAPC4	Progesterone-mediated oocyte maturation
ANAPC5	Progesterone-mediated oocyte maturation
ANAPC7	Progesterone-mediated oocyte maturation
ARAF	Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
ASGR1	Thyroid hormone synthesis
ASGR2	Thyroid hormone synthesis
ATF2	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion
ATF4	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway
ATF6B	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion
ATP1A1	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion
ATP1A2	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion
ATP1A3	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion
ATP1A4	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion
ATP1B1	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion
ATP1B2	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion

ATP1B3	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion
ATP1B4	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion
AURKA	Progesterone-mediated oocyte maturation
AXL	EGFR tyrosine kinase inhibitor resistance
BAD	Thyroid hormone signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
BAX	EGFR tyrosine kinase inhibitor resistance
BCL2	Estrogen signaling pathway EGFR tyrosine kinase inhibitor resistance
BCL2L1	EGFR tyrosine kinase inhibitor resistance
BCL2L11	EGFR tyrosine kinase inhibitor resistance
BMP15	Ovarian steroidogenesis
BMP4	Thyroid hormone signaling pathway
BMP6	Ovarian steroidogenesis
BRAF	Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
BUB1	Progesterone-mediated oocyte maturation
CACNA1C	Cortisol synthesis and secretion Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
CACNA1D	Cortisol synthesis and secretion Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
CACNA1F	Cortisol synthesis and secretion Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
CACNA1G	Cortisol synthesis and secretion
CACNA1H	Cortisol synthesis and secretion
CACNA1I	Cortisol synthesis and secretion
CACNA1S	Cortisol synthesis and secretion Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
CACNA2D1	Oxytocin signaling pathway
CACNA2D2	Oxytocin signaling pathway
CACNA2D3	Oxytocin signaling pathway
CACNA2D4	Oxytocin signaling pathway
CACNB1	Oxytocin signaling pathway
CACNB2	Oxytocin signaling pathway
CACNB3	Oxytocin signaling pathway
CACNB4	Oxytocin signaling pathway

CACNG1	Oxytocin signaling pathway
CACNG2	Oxytocin signaling pathway
CACNG3	Oxytocin signaling pathway
CACNG4	Oxytocin signaling pathway
CACNG5	Oxytocin signaling pathway
CACNG6	Oxytocin signaling pathway
CACNG7	Oxytocin signaling pathway
CACNG8	Oxytocin signaling pathway
CALM1	Estrogen signaling pathway Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway
CALM2	Estrogen signaling pathway Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway
CALM3	Estrogen signaling pathway Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway
CALML3	Estrogen signaling pathway Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway
CALML4	Estrogen signaling pathway Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway
CALML5	Estrogen signaling pathway Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway
CALML6	Estrogen signaling pathway Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway
CAMK1	Oxytocin signaling pathway
CAMK1D	Oxytocin signaling pathway
CAMK1G	Oxytocin signaling pathway
CAMK2A	Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
CAMK2B	Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
CAMK2D	Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
CAMK2G	Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
CAMK4	Oxytocin signaling pathway

CAMKK2	Oxytocin signaling pathway
CANX	Thyroid hormone synthesis
CASP9	Thyroid hormone signaling pathway
CBL	Insulin signaling pathway
CBLB	Insulin signaling pathway
CCKAR	Insulin secretion
CCNA1	Progesterone-mediated oocyte maturation
CCNA2	Progesterone-mediated oocyte maturation
CCNB1	Progesterone-mediated oocyte maturation
CCNB2	Progesterone-mediated oocyte maturation
CCNB3	Progesterone-mediated oocyte maturation
CCND1	Prolactin signaling pathway Thyroid hormone signaling pathway Oxytocin signaling pathway
CCND2	Prolactin signaling pathway
CD36	Insulin resistance
CD38	Oxytocin signaling pathway
CDC16	Progesterone-mediated oocyte maturation
CDC23	Progesterone-mediated oocyte maturation
CDC25A	Progesterone-mediated oocyte maturation
CDC25B	Progesterone-mediated oocyte maturation
CDC25C	Progesterone-mediated oocyte maturation
CDC26	Progesterone-mediated oocyte maturation
CDC27	Progesterone-mediated oocyte maturation
CDC42	GnRH signaling pathway
CDK1	Progesterone-mediated oocyte maturation
CDK2	Progesterone-mediated oocyte maturation
CDKN1A	Oxytocin signaling pathway
CEL	Steroid biosynthesis
CGA	Ovarian steroidogenesis Prolactin signaling pathway Thyroid hormone synthesis GnRH signaling pathway
CHRM3	Insulin secretion
CISH	Prolactin signaling pathway
COMT	Steroid hormone biosynthesis
CPEB1	Progesterone-mediated oocyte maturation
CPEB2	Progesterone-mediated oocyte maturation
CPEB3	Progesterone-mediated oocyte maturation
CPEB4	Progesterone-mediated oocyte maturation
CPT1A	Insulin resistance
CPT1B	Insulin resistance
CREB1	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin resistance Insulin secretion

CREB3	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin resistance Insulin secretion
CREB3L1	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin resistance Insulin secretion
CREB3L2	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin resistance Insulin secretion
CREB3L3	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin resistance Insulin secretion
CREB3L4	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin resistance Insulin secretion
CREB5	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin resistance Insulin secretion
CREBBP	Thyroid hormone signaling pathway
CRK	Insulin signaling pathway
CRKL	Insulin signaling pathway
CRTC2	Insulin resistance
CSN2	Prolactin signaling pathway
CTNNB1	Thyroid hormone signaling pathway
CTSD	Estrogen signaling pathway
CYP11A1	Steroid hormone biosynthesis Ovarian steroidogenesis Cortisol synthesis and secretion
CYP11B1	Steroid hormone biosynthesis Cortisol synthesis and secretion
CYP11B2	Steroid hormone biosynthesis
CYP17A1	Steroid hormone biosynthesis Ovarian steroidogenesis Prolactin signaling pathway Cortisol synthesis and secretion
CYP19A1	Steroid hormone biosynthesis Ovarian steroidogenesis
CYP1A1	Steroid hormone biosynthesis Ovarian steroidogenesis
CYP1A2	Steroid hormone biosynthesis

CYP1B1	Steroid hormone biosynthesis Ovarian steroidogenesis
CYP21A2	Steroid hormone biosynthesis Cortisol synthesis and secretion
CYP24A1	Steroid biosynthesis
CYP27B1	Steroid biosynthesis
CYP2E1	Steroid hormone biosynthesis
CYP2J2	Ovarian steroidogenesis
CYP2R1	Steroid biosynthesis
CYP3A4	Steroid hormone biosynthesis
CYP3A5	Steroid hormone biosynthesis
CYP3A7	Steroid hormone biosynthesis
CYP51A1	Steroid biosynthesis
CYP7A1	Steroid hormone biosynthesis
CYP7B1	Steroid hormone biosynthesis
DHCR24	Steroid biosynthesis
DHCR7	Steroid biosynthesis
DHRS11	Steroid hormone biosynthesis
DIO1	Thyroid hormone signaling pathway
DIO2	Thyroid hormone signaling pathway
DIO3	Thyroid hormone signaling pathway
DUOXA2	Thyroid hormone synthesis
EBAG9	Estrogen signaling pathway
EBP	Steroid biosynthesis
EEF2	Oxytocin signaling pathway
EEF2K	Oxytocin signaling pathway
EGF	EGFR tyrosine kinase inhibitor resistance
EGFR	Estrogen signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway
EGR1	GnRH signaling pathway
EIF4E	Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
EIF4E1B	Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
EIF4E2	Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
EIF4EBP1	Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
ELF5	Prolactin signaling pathway
ELK1	Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway
EP300	Thyroid hormone signaling pathway
ERBB2	EGFR tyrosine kinase inhibitor resistance
ERBB3	EGFR tyrosine kinase inhibitor resistance
ESR1	Estrogen signaling pathway

	Prolactin signaling pathway Thyroid hormone signaling pathway
ESR2	Estrogen signaling pathway Prolactin signaling pathway
EXOC7	Insulin signaling pathway
FASN	Insulin signaling pathway
FBP1	Insulin signaling pathway
FBP2	Insulin signaling pathway
FDFT1	Steroid biosynthesis
FFAR1	Insulin secretion
FGF2	EGFR tyrosine kinase inhibitor resistance
FGFR2	EGFR tyrosine kinase inhibitor resistance
FGFR3	EGFR tyrosine kinase inhibitor resistance
FKBP4	Estrogen signaling pathway
FKBP5	Estrogen signaling pathway
FLOT1	Insulin signaling pathway
FLOT2	Insulin signaling pathway
FOS	Estrogen signaling pathway Prolactin signaling pathway Oxytocin signaling pathway
FOXO1	Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway
FOXO3	Prolactin signaling pathway EGFR tyrosine kinase inhibitor resistance
FSHB	Ovarian steroidogenesis GnRH signaling pathway
FSHR	Ovarian steroidogenesis
FXYD2	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion
FZR1	Progesterone-mediated oocyte maturation
G6PC	Insulin resistance Insulin signaling pathway
G6PC2	Insulin resistance Insulin signaling pathway
G6PC3	Insulin resistance Insulin signaling pathway
GAB1	EGFR tyrosine kinase inhibitor resistance
GABBR1	Estrogen signaling pathway
GABBR2	Estrogen signaling pathway
GALT	Prolactin signaling pathway
GAS6	EGFR tyrosine kinase inhibitor resistance
GATA4	Thyroid hormone signaling pathway
GCG	Insulin secretion
GCK	Prolactin signaling pathway Insulin secretion Insulin signaling pathway
GFPT1	Insulin resistance

GFPT2	Insulin resistance
GIP	Insulin secretion
GLP1R	Insulin secretion
GNA11	Cortisol synthesis and secretion Insulin secretion GnRH signaling pathway
GNAI1	Estrogen signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
GNAI2	Estrogen signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
GNAI3	Estrogen signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
GNAO1	Estrogen signaling pathway Oxytocin signaling pathway
GNAQ	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
GNAS	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
GNRH1	GnRH signaling pathway
GNRH2	GnRH signaling pathway
GNRHR	GnRH signaling pathway
GPER1	Estrogen signaling pathway
GPR119	Insulin secretion
GPX1	Thyroid hormone synthesis
GPX2	Thyroid hormone synthesis
GPX3	Thyroid hormone synthesis
GPX5	Thyroid hormone synthesis
GPX6	Thyroid hormone synthesis
GPX7	Thyroid hormone synthesis
GPX8	Thyroid hormone synthesis
GRB2	Estrogen signaling pathway Prolactin signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway
GRM1	Estrogen signaling pathway
GSK3B	Prolactin signaling pathway Thyroid hormone signaling pathway Insulin resistance

	Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
GSR	Thyroid hormone synthesis
GUCY1A1	Oxytocin signaling pathway
GUCY1A2	Oxytocin signaling pathway
GUCY1B1	Oxytocin signaling pathway
GYS1	Insulin resistance Insulin signaling pathway
GYS2	Insulin resistance Insulin signaling pathway
HBEGF	Estrogen signaling pathway GnRH signaling pathway
HDAC1	Thyroid hormone signaling pathway
HDAC2	Thyroid hormone signaling pathway
HDAC3	Thyroid hormone signaling pathway
HGF	EGFR tyrosine kinase inhibitor resistance
HIF1A	Thyroid hormone signaling pathway
HK1	Insulin signaling pathway
HK2	Insulin signaling pathway
HK3	Insulin signaling pathway
HKDC1	Insulin signaling pathway
HRAS	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway
HSD11B1	Steroid hormone biosynthesis
HSD11B2	Steroid hormone biosynthesis
HSD17B1	Steroid hormone biosynthesis Ovarian steroidogenesis
HSD17B12	Steroid hormone biosynthesis
HSD17B2	Steroid hormone biosynthesis Ovarian steroidogenesis
HSD17B3	Steroid hormone biosynthesis
HSD17B6	Steroid hormone biosynthesis
HSD17B7	Steroid hormone biosynthesis Ovarian steroidogenesis Steroid biosynthesis
HSD17B8	Steroid hormone biosynthesis
HSD3B1	Steroid hormone biosynthesis Ovarian steroidogenesis Cortisol synthesis and secretion
HSD3B2	Steroid hormone biosynthesis Ovarian steroidogenesis Cortisol synthesis and secretion
HSP90AA1	Estrogen signaling pathway Progesterone-mediated oocyte maturation
HSP90AB1	Estrogen signaling pathway

	Progesterone-mediated oocyte maturation
HSP90B1	Estrogen signaling pathway Thyroid hormone synthesis
HSPA1A	Estrogen signaling pathway
HSPA1B	Estrogen signaling pathway
HSPA1L	Estrogen signaling pathway
HSPA2	Estrogen signaling pathway
HSPA5	Thyroid hormone synthesis
HSPA6	Estrogen signaling pathway
HSPA8	Estrogen signaling pathway
IGF1	Ovarian steroidogenesis EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
IGF1R	Ovarian steroidogenesis EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
IKBKB	Insulin resistance Insulin signaling pathway
IL6	Insulin resistance EGFR tyrosine kinase inhibitor resistance
IL6R	EGFR tyrosine kinase inhibitor resistance
INPPL1	Insulin signaling pathway
INS	Ovarian steroidogenesis Prolactin signaling pathway Insulin resistance Insulin secretion Insulin signaling pathway Progesterone-mediated oocyte maturation
INSR	Ovarian steroidogenesis Insulin resistance Insulin signaling pathway
IRF1	Prolactin signaling pathway
IRS1	Insulin resistance Insulin signaling pathway
IRS2	Insulin resistance Insulin signaling pathway
IRS4	Insulin signaling pathway
ITGAV	Thyroid hormone signaling pathway
ITGB3	Thyroid hormone signaling pathway
ITPR1	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis GnRH signaling pathway Oxytocin signaling pathway
ITPR2	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis GnRH signaling pathway Oxytocin signaling pathway
ITPR3	Estrogen signaling pathway Cortisol synthesis and secretion

	Thyroid hormone synthesis Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
IYD	Thyroid hormone synthesis
JAK1	EGFR tyrosine kinase inhibitor resistance
JAK2	Prolactin signaling pathway EGFR tyrosine kinase inhibitor resistance
JUN	Estrogen signaling pathway GnRH signaling pathway Oxytocin signaling pathway
KAT2A	Thyroid hormone signaling pathway
KAT2B	Thyroid hormone signaling pathway
KCNA4	Cortisol synthesis and secretion
KCNJ11	Insulin secretion
KCNJ12	Oxytocin signaling pathway
KCNJ14	Oxytocin signaling pathway
KCNJ2	Oxytocin signaling pathway
KCNJ3	Estrogen signaling pathway Oxytocin signaling pathway
KCNJ4	Oxytocin signaling pathway
KCNJ5	Estrogen signaling pathway Oxytocin signaling pathway
KCNJ6	Estrogen signaling pathway Oxytocin signaling pathway
KCNJ9	Estrogen signaling pathway Oxytocin signaling pathway
KCNK2	Cortisol synthesis and secretion
KCNK3	Cortisol synthesis and secretion
KCNMA1	Insulin secretion
KCNMB1	Insulin secretion
KCNMB2	Insulin secretion
KCNMB3	Insulin secretion
KCNMB4	Insulin secretion
KCNN1	Insulin secretion
KCNN2	Insulin secretion
KCNN3	Insulin secretion
KCNN4	Insulin secretion
KCNU1	Insulin secretion
KDR	EGFR tyrosine kinase inhibitor resistance
KIF22	Progesterone-mediated oocyte maturation
KRAS	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation

KRT10	Estrogen signaling pathway
KRT12	Estrogen signaling pathway
KRT13	Estrogen signaling pathway
KRT14	Estrogen signaling pathway
KRT15	Estrogen signaling pathway
KRT16	Estrogen signaling pathway
KRT17	Estrogen signaling pathway
KRT18	Estrogen signaling pathway
KRT19	Estrogen signaling pathway
KRT20	Estrogen signaling pathway
KRT23	Estrogen signaling pathway
KRT24	Estrogen signaling pathway
KRT25	Estrogen signaling pathway
KRT26	Estrogen signaling pathway
KRT27	Estrogen signaling pathway
KRT28	Estrogen signaling pathway
KRT31	Estrogen signaling pathway
KRT32	Estrogen signaling pathway
KRT33A	Estrogen signaling pathway
KRT33B	Estrogen signaling pathway
KRT34	Estrogen signaling pathway
KRT35	Estrogen signaling pathway
KRT36	Estrogen signaling pathway
KRT37	Estrogen signaling pathway
KRT38	Estrogen signaling pathway
KRT39	Estrogen signaling pathway
KRT40	Estrogen signaling pathway
KRT9	Estrogen signaling pathway
LDLR	Ovarian steroidogenesis Cortisol synthesis and secretion
LHB	Ovarian steroidogenesis Prolactin signaling pathway GnRH signaling pathway
LHCGR	Ovarian steroidogenesis Prolactin signaling pathway
LIPA	Steroid biosynthesis
LIPE	Insulin signaling pathway
LRP2	Thyroid hormone synthesis
LRTOMT	Steroid hormone biosynthesis
LSS	Steroid biosynthesis
MAD1L1	Progesterone-mediated oocyte maturation
MAD2L1	Progesterone-mediated oocyte maturation
MAD2L2	Progesterone-mediated oocyte maturation
MAP2K1	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin signaling pathway

	EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
MAP2K2	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway
MAP2K3	GnRH signaling pathway
MAP2K4	GnRH signaling pathway
MAP2K5	Oxytocin signaling pathway
MAP2K6	GnRH signaling pathway
MAP2K7	GnRH signaling pathway
MAP3K1	GnRH signaling pathway
MAP3K2	GnRH signaling pathway
MAP3K3	GnRH signaling pathway
MAP3K4	GnRH signaling pathway
MAPK1	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
MAPK10	Prolactin signaling pathway Insulin resistance Insulin signaling pathway GnRH signaling pathway Progesterone-mediated oocyte maturation
MAPK11	Prolactin signaling pathway GnRH signaling pathway Progesterone-mediated oocyte maturation
MAPK12	Prolactin signaling pathway GnRH signaling pathway Progesterone-mediated oocyte maturation
MAPK13	Prolactin signaling pathway GnRH signaling pathway Progesterone-mediated oocyte maturation
MAPK14	Prolactin signaling pathway GnRH signaling pathway Progesterone-mediated oocyte maturation
MAPK3	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway

	Progesterone-mediated oocyte maturation
MAPK7	GnRH signaling pathway Oxytocin signaling pathway
MAPK8	Prolactin signaling pathway Insulin resistance Insulin signaling pathway GnRH signaling pathway Progesterone-mediated oocyte maturation
MAPK9	Prolactin signaling pathway Insulin resistance Insulin signaling pathway GnRH signaling pathway Progesterone-mediated oocyte maturation
MC2R	Cortisol synthesis and secretion
MDM2	Thyroid hormone signaling pathway
MED1	Thyroid hormone signaling pathway
MED12	Thyroid hormone signaling pathway
MED12L	Thyroid hormone signaling pathway
MED13	Thyroid hormone signaling pathway
MED13L	Thyroid hormone signaling pathway
MED14	Thyroid hormone signaling pathway
MED16	Thyroid hormone signaling pathway
MED17	Thyroid hormone signaling pathway
MED24	Thyroid hormone signaling pathway
MED27	Thyroid hormone signaling pathway
MED30	Thyroid hormone signaling pathway
MED4	Thyroid hormone signaling pathway
MEF2C	Oxytocin signaling pathway
MET	EGFR tyrosine kinase inhibitor resistance
MKNK1	Insulin signaling pathway
MKNK2	Insulin signaling pathway
MLX	Insulin resistance
MLXIP	Insulin resistance
MLXIPL	Insulin resistance
MMP14	GnRH signaling pathway
MMP2	Estrogen signaling pathway GnRH signaling pathway
MMP9	Estrogen signaling pathway
MOS	Progesterone-mediated oocyte maturation
MRAP	Cortisol synthesis and secretion
MSMO1	Steroid biosynthesis
MTOR	Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
MYC	Thyroid hormone signaling pathway
MYL6	Oxytocin signaling pathway
MYL6B	Oxytocin signaling pathway

MYL9	Oxytocin signaling pathway
MYLK	Oxytocin signaling pathway
MYLK2	Oxytocin signaling pathway
MYLK3	Oxytocin signaling pathway
MYLK4	Oxytocin signaling pathway
NCEH1	Cortisol synthesis and secretion
NCOA1	Estrogen signaling pathway Thyroid hormone signaling pathway
NCOA2	Estrogen signaling pathway Thyroid hormone signaling pathway
NCOA3	Estrogen signaling pathway Thyroid hormone signaling pathway
NCOR1	Thyroid hormone signaling pathway
NF1	EGFR tyrosine kinase inhibitor resistance
NFATC1	Oxytocin signaling pathway
NFATC2	Oxytocin signaling pathway
NFATC3	Oxytocin signaling pathway
NFATC4	Oxytocin signaling pathway
NFKB1	Prolactin signaling pathway Insulin resistance
NFKBIA	Insulin resistance
NOS3	Estrogen signaling pathway Insulin resistance Oxytocin signaling pathway
NOTCH1	Thyroid hormone signaling pathway
NOTCH2	Thyroid hormone signaling pathway
NOTCH3	Thyroid hormone signaling pathway
NOTCH4	Thyroid hormone signaling pathway
NPR1	Oxytocin signaling pathway
NPR2	Oxytocin signaling pathway
NR0B1	Cortisol synthesis and secretion
NR1H2	Insulin resistance
NR1H3	Insulin resistance
NR4A1	Cortisol synthesis and secretion
NR5A1	Cortisol synthesis and secretion
NRAS	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway
NRG1	EGFR tyrosine kinase inhibitor resistance
NRG2	EGFR tyrosine kinase inhibitor resistance
NSDHL	Steroid biosynthesis
OGA	Insulin resistance
OGT	Insulin resistance
OPRM1	Estrogen signaling pathway

ORAI1	Cortisol synthesis and secretion
OXT	Oxytocin signaling pathway
OXTR	Oxytocin signaling pathway
PAX8	Thyroid hormone synthesis
PBX1	Cortisol synthesis and secretion
PCK1	Insulin resistance Insulin signaling pathway
PCK2	Insulin resistance Insulin signaling pathway
PCLO	Insulin secretion
PDE3B	Insulin signaling pathway Progesterone-mediated oocyte maturation
PDE8A	Cortisol synthesis and secretion
PDE8B	Cortisol synthesis and secretion
PDGFA	EGFR tyrosine kinase inhibitor resistance
PDGFB	EGFR tyrosine kinase inhibitor resistance
PDGFC	EGFR tyrosine kinase inhibitor resistance
PDGFD	EGFR tyrosine kinase inhibitor resistance
PDGFRA	EGFR tyrosine kinase inhibitor resistance
PDGFRB	EGFR tyrosine kinase inhibitor resistance
PDIA4	Thyroid hormone synthesis
PDPK1	Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway
PDX1	Insulin secretion
PFKFB2	Thyroid hormone signaling pathway
PGR	Estrogen signaling pathway Progesterone-mediated oocyte maturation
PHKA1	Insulin signaling pathway
PHKA2	Insulin signaling pathway
PHKB	Insulin signaling pathway
PHKG1	Insulin signaling pathway
PHKG2	Insulin signaling pathway
PIK3CA	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
PIK3CB	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
PIK3CD	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway

	Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
PIK3CG	Oxytocin signaling pathway
PIK3R1	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
PIK3R2	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
PIK3R3	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance Progesterone-mediated oocyte maturation
PIK3R5	Oxytocin signaling pathway
PIK3R6	Oxytocin signaling pathway
PKLR	Insulin signaling pathway
PKMYT1	Progesterone-mediated oocyte maturation
PLA2G4A	Ovarian steroidogenesis GnRH signaling pathway Oxytocin signaling pathway
PLA2G4B	Ovarian steroidogenesis GnRH signaling pathway Oxytocin signaling pathway
PLA2G4C	Ovarian steroidogenesis GnRH signaling pathway Oxytocin signaling pathway
PLA2G4D	Ovarian steroidogenesis GnRH signaling pathway Oxytocin signaling pathway
PLA2G4E	Ovarian steroidogenesis GnRH signaling pathway Oxytocin signaling pathway
PLA2G4F	Ovarian steroidogenesis GnRH signaling pathway Oxytocin signaling pathway
PLCB1	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion

	GnRH signaling pathway Oxytocin signaling pathway
PLCB2	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
PLCB3	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
PLCB4	Estrogen signaling pathway Cortisol synthesis and secretion Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion GnRH signaling pathway Oxytocin signaling pathway
PLCD1	Thyroid hormone signaling pathway
PLCD3	Thyroid hormone signaling pathway
PLCD4	Thyroid hormone signaling pathway
PLCE1	Thyroid hormone signaling pathway
PLCG1	Thyroid hormone signaling pathway EGFR tyrosine kinase inhibitor resistance
PLCG2	Thyroid hormone signaling pathway EGFR tyrosine kinase inhibitor resistance
PLCZ1	Thyroid hormone signaling pathway
PLD1	GnRH signaling pathway
PLD2	GnRH signaling pathway
PLK1	Progesterone-mediated oocyte maturation
PLN	Thyroid hormone signaling pathway
POMC	Cortisol synthesis and secretion
PPARA	Insulin resistance
PPARGC1A	Insulin resistance Insulin signaling pathway
PPARGC1B	Insulin resistance
PPP1CA	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PPP1CB	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PPP1CC	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PPP1R12A	Oxytocin signaling pathway

PPP1R12B	Oxytocin signaling pathway
PPP1R12C	Oxytocin signaling pathway
PPP1R3A	Insulin resistance Insulin signaling pathway
PPP1R3B	Insulin resistance Insulin signaling pathway
PPP1R3C	Insulin resistance Insulin signaling pathway
PPP1R3D	Insulin resistance Insulin signaling pathway
PPP1R3E	Insulin resistance Insulin signaling pathway
PPP1R3F	Insulin signaling pathway
PPP3CA	Oxytocin signaling pathway
PPP3CB	Oxytocin signaling pathway
PPP3CC	Oxytocin signaling pathway
PPP3R1	Oxytocin signaling pathway
PPP3R2	Oxytocin signaling pathway
PRKAA1	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PRKAA2	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PRKAB1	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PRKAB2	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PRKACA	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
PRKACB	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
PRKACG	Estrogen signaling pathway Ovarian steroidogenesis Cortisol synthesis and secretion

	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion Insulin signaling pathway GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
PRKAG1	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PRKAG2	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PRKAG3	Insulin resistance Insulin signaling pathway Oxytocin signaling pathway
PRKAR1A	Insulin signaling pathway
PRKAR1B	Insulin signaling pathway
PRKAR2A	Insulin signaling pathway
PRKAR2B	Insulin signaling pathway
PRKCA	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway
PRKCB	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin resistance Insulin secretion EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway
PRKCD	Estrogen signaling pathway Insulin resistance GnRH signaling pathway
PRKCE	Insulin resistance
PRKCG	Thyroid hormone synthesis Thyroid hormone signaling pathway Insulin secretion EGFR tyrosine kinase inhibitor resistance Oxytocin signaling pathway
PRKCI	Insulin signaling pathway
PRKCQ	Insulin resistance
PRKCZ	Insulin resistance Insulin signaling pathway
PRL	Prolactin signaling pathway
PRLR	Prolactin signaling pathway
PTEN	Insulin resistance EGFR tyrosine kinase inhibitor resistance
PTGS2	Ovarian steroidogenesis Oxytocin signaling pathway

PTK2B	GnRH signaling pathway
PTPA	Insulin resistance
PTPN1	Insulin resistance Insulin signaling pathway
PTPN11	Insulin resistance
PTPRF	Insulin resistance Insulin signaling pathway
PYGB	Insulin resistance Insulin signaling pathway
PYGL	Insulin resistance Insulin signaling pathway
PYGM	Insulin resistance Insulin signaling pathway
RAB3A	Insulin secretion
RAF1	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway Oxytocin signaling pathway Progesterone-mediated oocyte maturation
RAPGEF1	Insulin signaling pathway
RAPGEF4	Insulin secretion
RARA	Estrogen signaling pathway
RCAN1	Thyroid hormone signaling pathway Oxytocin signaling pathway
RCAN2	Thyroid hormone signaling pathway
RELA	Prolactin signaling pathway Insulin resistance
RGS2	Oxytocin signaling pathway
RHEB	Thyroid hormone signaling pathway Insulin signaling pathway
RHOA	Oxytocin signaling pathway
RHOQ	Insulin signaling pathway
RIMS2	Insulin secretion
ROCK1	Oxytocin signaling pathway
ROCK2	Oxytocin signaling pathway
RPS6	Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
RPS6KA1	Insulin resistance Progesterone-mediated oocyte maturation
RPS6KA2	Insulin resistance Progesterone-mediated oocyte maturation
RPS6KA3	Insulin resistance Progesterone-mediated oocyte maturation
RPS6KA6	Insulin resistance Progesterone-mediated oocyte maturation
RPS6KB1	Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance

RPS6KB2	Insulin resistance Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
RPTOR	Insulin signaling pathway
RXRA	Thyroid hormone signaling pathway
RXRB	Thyroid hormone signaling pathway
RXRG	Thyroid hormone signaling pathway
RYR1	Oxytocin signaling pathway
RYR2	Insulin secretion Oxytocin signaling pathway
RYR3	Oxytocin signaling pathway
SC5D	Steroid biosynthesis
SCARB1	Ovarian steroidogenesis Cortisol synthesis and secretion
SERPINA7	Thyroid hormone synthesis
SH2B2	Insulin signaling pathway
SHC1	Estrogen signaling pathway Prolactin signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
SHC2	Estrogen signaling pathway Prolactin signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
SHC3	Estrogen signaling pathway Prolactin signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
SHC4	Estrogen signaling pathway Prolactin signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance
SIN3A	Thyroid hormone signaling pathway
SLC16A10	Thyroid hormone signaling pathway
SLC16A2	Thyroid hormone signaling pathway
SLC26A4	Thyroid hormone synthesis
SLC27A1	Insulin resistance
SLC27A2	Insulin resistance
SLC27A3	Insulin resistance
SLC27A4	Insulin resistance
SLC27A5	Insulin resistance
SLC27A6	Insulin resistance
SLC2A1	Thyroid hormone signaling pathway Insulin resistance Insulin secretion
SLC2A2	Prolactin signaling pathway Insulin resistance Insulin secretion
SLC2A4	Insulin resistance Insulin signaling pathway

SLC5A5	Thyroid hormone synthesis
SLC9A1	Thyroid hormone signaling pathway
SLCO1C1	Thyroid hormone signaling pathway
SNAP25	Insulin secretion
SOAT1	Steroid biosynthesis
SOAT2	Steroid biosynthesis
SOCS1	Prolactin signaling pathway Insulin signaling pathway
SOCS2	Prolactin signaling pathway Insulin signaling pathway
SOCS3	Prolactin signaling pathway Insulin resistance Insulin signaling pathway
SOCS4	Prolactin signaling pathway Insulin signaling pathway
SOCS5	Prolactin signaling pathway
SOCS6	Prolactin signaling pathway
SOCS7	Prolactin signaling pathway
SORBS1	Insulin signaling pathway
SOS1	Estrogen signaling pathway Prolactin signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway
SOS2	Estrogen signaling pathway Prolactin signaling pathway Insulin signaling pathway EGFR tyrosine kinase inhibitor resistance GnRH signaling pathway
SP1	Estrogen signaling pathway Cortisol synthesis and secretion
SPDYA	Progesterone-mediated oocyte maturation
SPDYC	Progesterone-mediated oocyte maturation
SPDYE1	Progesterone-mediated oocyte maturation
SPDYE11	Progesterone-mediated oocyte maturation
SPDYE16	Progesterone-mediated oocyte maturation
SPDYE17	Progesterone-mediated oocyte maturation
SPDYE18	Progesterone-mediated oocyte maturation
SPDYE2	Progesterone-mediated oocyte maturation
SPDYE2B	Progesterone-mediated oocyte maturation
SPDYE3	Progesterone-mediated oocyte maturation
SPDYE4	Progesterone-mediated oocyte maturation
SPDYE5	Progesterone-mediated oocyte maturation
SPDYE6	Progesterone-mediated oocyte maturation
SQLE	Steroid biosynthesis
SRC	Estrogen signaling pathway Prolactin signaling pathway Thyroid hormone signaling pathway EGFR tyrosine kinase inhibitor resistance

	GnRH signaling pathway Oxytocin signaling pathway
SRD5A1	Steroid hormone biosynthesis
SRD5A2	Steroid hormone biosynthesis
SRD5A3	Steroid hormone biosynthesis
SREBF1	Insulin resistance Insulin signaling pathway
STAR	Ovarian steroidogenesis Cortisol synthesis and secretion
STAT1	Prolactin signaling pathway Thyroid hormone signaling pathway
STAT3	Prolactin signaling pathway Insulin resistance EGFR tyrosine kinase inhibitor resistance
STAT5A	Prolactin signaling pathway
STAT5B	Prolactin signaling pathway
STK10	Progesterone-mediated oocyte maturation
STS	Steroid hormone biosynthesis
STX1A	Insulin secretion
SULT1E1	Steroid hormone biosynthesis
SULT2B1	Steroid hormone biosynthesis
TBC1D4	Thyroid hormone signaling pathway Insulin resistance
TFF1	Estrogen signaling pathway
TG	Thyroid hormone synthesis
TGFA	Estrogen signaling pathway EGFR tyrosine kinase inhibitor resistance
TH	Prolactin signaling pathway
THRA	Thyroid hormone signaling pathway
THRB	Thyroid hormone signaling pathway
TM7SF2	Steroid biosynthesis
TNF	Insulin resistance
TNFRSF11A	Prolactin signaling pathway
TNFRSF1A	Insulin resistance
TNFSF11	Prolactin signaling pathway
TP53	Thyroid hormone signaling pathway
TPO	Thyroid hormone synthesis
TRIB3	Insulin resistance
TRIP10	Insulin signaling pathway
TRPM2	Oxytocin signaling pathway
TRPM4	Insulin secretion
TSC1	Insulin signaling pathway
TSC2	Thyroid hormone signaling pathway Insulin signaling pathway
TSHB	Thyroid hormone synthesis
TSHR	Thyroid hormone synthesis
TTF1	Thyroid hormone synthesis
TTF2	Thyroid hormone synthesis

TTR	Thyroid hormone synthesis
UGT1A1	Steroid hormone biosynthesis
UGT1A10	Steroid hormone biosynthesis
UGT1A3	Steroid hormone biosynthesis
UGT1A4	Steroid hormone biosynthesis
UGT1A5	Steroid hormone biosynthesis
UGT1A6	Steroid hormone biosynthesis
UGT1A7	Steroid hormone biosynthesis
UGT1A8	Steroid hormone biosynthesis
UGT1A9	Steroid hormone biosynthesis
UGT2A1	Steroid hormone biosynthesis
UGT2A2	Steroid hormone biosynthesis
UGT2A3	Steroid hormone biosynthesis
UGT2B10	Steroid hormone biosynthesis
UGT2B11	Steroid hormone biosynthesis
UGT2B15	Steroid hormone biosynthesis
UGT2B17	Steroid hormone biosynthesis
UGT2B28	Steroid hormone biosynthesis
UGT2B4	Steroid hormone biosynthesis
UGT2B7	Steroid hormone biosynthesis
VAMP2	Insulin secretion
VEGFA	EGFR tyrosine kinase inhibitor resistance
WNT4	Thyroid hormone signaling pathway

Annexe 4 - Gènes intervenant dans la réparation de l'ADN et le cycle cellulaire

Gène	Voies Biologiques KEGG
ABL1	Cell cycle
ABRAXAS1	Homologous recombination
AKT1	PI3K-Akt signaling pathway
AKT2	PI3K-Akt signaling pathway
AKT3	PI3K-Akt signaling pathway
ANAPC1	Cell cycle
ANAPC10	Cell cycle
ANAPC11	Cell cycle
ANAPC13	Cell cycle
ANAPC2	Cell cycle
ANAPC4	Cell cycle
ANAPC5	Cell cycle
ANAPC7	Cell cycle
ANGPT1	PI3K-Akt signaling pathway
ANGPT2	PI3K-Akt signaling pathway
ANGPT4	PI3K-Akt signaling pathway
APEX1	Base excision repair
APEX2	Base excision repair
AREG	PI3K-Akt signaling pathway
ATF2	PI3K-Akt signaling pathway
ATF4	PI3K-Akt signaling pathway
ATF6B	PI3K-Akt signaling pathway
ATM	Homologous recombination Cell cycle
ATR	Fanconi anemia pathway Cell cycle
ATRIP	Fanconi anemia pathway
BABAM1	Homologous recombination
BABAM2	Homologous recombination
BAD	PI3K-Akt signaling pathway
BARD1	Homologous recombination
BCL2	PI3K-Akt signaling pathway
BCL2L1	PI3K-Akt signaling pathway
BCL2L11	PI3K-Akt signaling pathway
BDNF	PI3K-Akt signaling pathway
BLM	Fanconi anemia pathway Homologous recombination
BRCA1	Fanconi anemia pathway PI3K-Akt signaling pathway Homologous recombination
BRCA2	Fanconi anemia pathway Homologous recombination
BRCC3	Homologous recombination

BRIP1	Fanconi anemia pathway Homologous recombination
BUB1	Cell cycle
BUB1B	Cell cycle
BUB3	Cell cycle
CASP9	PI3K-Akt signaling pathway
CCNA1	Cell cycle
CCNA2	Cell cycle
CCNB1	Cell cycle
CCNB2	Cell cycle
CCNB3	Cell cycle
CCND1	PI3K-Akt signaling pathway Cell cycle
CCND2	PI3K-Akt signaling pathway Cell cycle
CCND3	PI3K-Akt signaling pathway Cell cycle
CCNE1	PI3K-Akt signaling pathway Cell cycle
CCNE2	PI3K-Akt signaling pathway Cell cycle
CCNH	Cell cycle Nucleotide excision repair
CD19	PI3K-Akt signaling pathway
CDC14A	Cell cycle
CDC14B	Cell cycle
CDC16	Cell cycle
CDC20	Cell cycle
CDC23	Cell cycle
CDC25A	Cell cycle
CDC25B	Cell cycle
CDC25C	Cell cycle
CDC26	Cell cycle
CDC27	Cell cycle
CDC37	PI3K-Akt signaling pathway
CDC45	Cell cycle
CDC6	Cell cycle
CDC7	Cell cycle
CDK1	Cell cycle
CDK2	PI3K-Akt signaling pathway Cell cycle
CDK4	PI3K-Akt signaling pathway Cell cycle
CDK6	PI3K-Akt signaling pathway Cell cycle
CDK7	Cell cycle Nucleotide excision repair
CDKN1A	PI3K-Akt signaling pathway Cell cycle

CDKN1B	PI3K-Akt signaling pathway Cell cycle
CDKN1C	Cell cycle
CDKN2A	Cell cycle
CDKN2B	Cell cycle
CDKN2C	Cell cycle
CDKN2D	Cell cycle
CENPS	Fanconi anemia pathway
CENPX	Fanconi anemia pathway
CETN2	Nucleotide excision repair
CHAD	PI3K-Akt signaling pathway
CHEK1	Cell cycle
CHEK2	Cell cycle
CHRM1	PI3K-Akt signaling pathway
CHRM2	PI3K-Akt signaling pathway
CHUK	PI3K-Akt signaling pathway
COL1A1	PI3K-Akt signaling pathway
COL1A2	PI3K-Akt signaling pathway
COL2A1	PI3K-Akt signaling pathway
COL4A1	PI3K-Akt signaling pathway
COL4A2	PI3K-Akt signaling pathway
COL4A3	PI3K-Akt signaling pathway
COL4A4	PI3K-Akt signaling pathway
COL4A5	PI3K-Akt signaling pathway
COL4A6	PI3K-Akt signaling pathway
COL6A1	PI3K-Akt signaling pathway
COL6A2	PI3K-Akt signaling pathway
COL6A3	PI3K-Akt signaling pathway
COL6A5	PI3K-Akt signaling pathway
COL6A6	PI3K-Akt signaling pathway
COL9A1	PI3K-Akt signaling pathway
COL9A2	PI3K-Akt signaling pathway
COL9A3	PI3K-Akt signaling pathway
COMP	PI3K-Akt signaling pathway
CREB1	PI3K-Akt signaling pathway
CREB3	PI3K-Akt signaling pathway
CREB3L1	PI3K-Akt signaling pathway
CREB3L2	PI3K-Akt signaling pathway
CREB3L3	PI3K-Akt signaling pathway
CREB3L4	PI3K-Akt signaling pathway
CREB5	PI3K-Akt signaling pathway
CREBBP	Cell cycle
CRTC2	PI3K-Akt signaling pathway
CSF1	PI3K-Akt signaling pathway
CSF1R	PI3K-Akt signaling pathway

CSF3	PI3K-Akt signaling pathway
CSF3R	PI3K-Akt signaling pathway
CSH1	PI3K-Akt signaling pathway
CSH2	PI3K-Akt signaling pathway
CUL1	Cell cycle
CUL4A	Nucleotide excision repair
CUL4B	Nucleotide excision repair
DBF4	Cell cycle
DCLRE1C	Non-Homologous End-Joining
DDB1	Nucleotide excision repair
DDB2	Nucleotide excision repair
DDIT4	PI3K-Akt signaling pathway
DNA2	DNA replication
DNTT	Non-Homologous End-Joining
E2F1	Cell cycle
E2F2	Cell cycle
E2F3	Cell cycle
E2F4	Cell cycle
E2F5	Cell cycle
EFNA1	PI3K-Akt signaling pathway
EFNA2	PI3K-Akt signaling pathway
EFNA3	PI3K-Akt signaling pathway
EFNA4	PI3K-Akt signaling pathway
EFNA5	PI3K-Akt signaling pathway
EGF	PI3K-Akt signaling pathway
EGFR	PI3K-Akt signaling pathway
EIF4B	PI3K-Akt signaling pathway
EIF4E	PI3K-Akt signaling pathway
EIF4E1B	PI3K-Akt signaling pathway
EIF4E2	PI3K-Akt signaling pathway
EIF4EBP1	PI3K-Akt signaling pathway
EME1	Fanconi anemia pathway Homologous recombination
EME2	Fanconi anemia pathway
EP300	Cell cycle
EPHA2	PI3K-Akt signaling pathway
EPO	PI3K-Akt signaling pathway
EPOR	PI3K-Akt signaling pathway
ERBB2	PI3K-Akt signaling pathway
ERBB3	PI3K-Akt signaling pathway
ERBB4	PI3K-Akt signaling pathway
ERCC1	Fanconi anemia pathway Nucleotide excision repair
ERCC2	Nucleotide excision repair
ERCC3	Nucleotide excision repair

ERCC4	Fanconi anemia pathway Nucleotide excision repair
ERCC5	Nucleotide excision repair
ERCC6	Nucleotide excision repair
ERCC8	Nucleotide excision repair
EREG	PI3K-Akt signaling pathway
ESPL1	Cell cycle
EXO1	Mismatch repair
F2R	PI3K-Akt signaling pathway
FAAP100	Fanconi anemia pathway
FAAP24	Fanconi anemia pathway
FAN1	Fanconi anemia pathway
FANCA	Fanconi anemia pathway
FANCB	Fanconi anemia pathway
FANCC	Fanconi anemia pathway
FANCD2	Fanconi anemia pathway
FANCE	Fanconi anemia pathway
FANCF	Fanconi anemia pathway
FANCG	Fanconi anemia pathway
FANCI	Fanconi anemia pathway
FANCL	Fanconi anemia pathway
FANCM	Fanconi anemia pathway
FASLG	PI3K-Akt signaling pathway
FEN1	Non-Homologous End-Joining Base excision repair DNA replication
FGF1	PI3K-Akt signaling pathway
FGF10	PI3K-Akt signaling pathway
FGF16	PI3K-Akt signaling pathway
FGF17	PI3K-Akt signaling pathway
FGF18	PI3K-Akt signaling pathway
FGF19	PI3K-Akt signaling pathway
FGF2	PI3K-Akt signaling pathway
FGF20	PI3K-Akt signaling pathway
FGF21	PI3K-Akt signaling pathway
FGF22	PI3K-Akt signaling pathway
FGF23	PI3K-Akt signaling pathway
FGF3	PI3K-Akt signaling pathway
FGF4	PI3K-Akt signaling pathway
FGF5	PI3K-Akt signaling pathway
FGF6	PI3K-Akt signaling pathway
FGF7	PI3K-Akt signaling pathway
FGF8	PI3K-Akt signaling pathway
FGF9	PI3K-Akt signaling pathway
FGFR1	PI3K-Akt signaling pathway
FGFR2	PI3K-Akt signaling pathway

FGFR3	PI3K-Akt signaling pathway
FGFR4	PI3K-Akt signaling pathway
FLT1	PI3K-Akt signaling pathway
FLT3	PI3K-Akt signaling pathway
FLT3LG	PI3K-Akt signaling pathway
FLT4	PI3K-Akt signaling pathway
FN1	PI3K-Akt signaling pathway
FOXO3	PI3K-Akt signaling pathway
FZR1	Cell cycle
G6PC	PI3K-Akt signaling pathway
G6PC2	PI3K-Akt signaling pathway
G6PC3	PI3K-Akt signaling pathway
GADD45A	Cell cycle
GADD45B	Cell cycle
GADD45G	Cell cycle
GH1	PI3K-Akt signaling pathway
GH2	PI3K-Akt signaling pathway
GHR	PI3K-Akt signaling pathway
GNB1	PI3K-Akt signaling pathway
GNB2	PI3K-Akt signaling pathway
GNB3	PI3K-Akt signaling pathway
GNB4	PI3K-Akt signaling pathway
GNB5	PI3K-Akt signaling pathway
GNG10	PI3K-Akt signaling pathway
GNG11	PI3K-Akt signaling pathway
GNG12	PI3K-Akt signaling pathway
GNG13	PI3K-Akt signaling pathway
GNG2	PI3K-Akt signaling pathway
GNG3	PI3K-Akt signaling pathway
GNG4	PI3K-Akt signaling pathway
GNG5	PI3K-Akt signaling pathway
GNG7	PI3K-Akt signaling pathway
GNG8	PI3K-Akt signaling pathway
GNGT1	PI3K-Akt signaling pathway
GNGT2	PI3K-Akt signaling pathway
GRB2	PI3K-Akt signaling pathway
GSK3B	PI3K-Akt signaling pathway Cell cycle
GTF2H1	Nucleotide excision repair
GTF2H2	Nucleotide excision repair
GTF2H2C	Nucleotide excision repair
GTF2H3	Nucleotide excision repair
GTF2H4	Nucleotide excision repair
GTF2H5	Nucleotide excision repair
GYS1	PI3K-Akt signaling pathway

GYS2	PI3K-Akt signaling pathway
HDAC1	Cell cycle
HDAC2	Cell cycle
HES1	Fanconi anemia pathway
HGF	PI3K-Akt signaling pathway
HMGB1	Base excision repair
HRAS	PI3K-Akt signaling pathway
HSP90AA1	PI3K-Akt signaling pathway
HSP90AB1	PI3K-Akt signaling pathway
HSP90B1	PI3K-Akt signaling pathway
IBSP	PI3K-Akt signaling pathway
IFNA1	PI3K-Akt signaling pathway
IFNA10	PI3K-Akt signaling pathway
IFNA13	PI3K-Akt signaling pathway
IFNA14	PI3K-Akt signaling pathway
IFNA16	PI3K-Akt signaling pathway
IFNA17	PI3K-Akt signaling pathway
IFNA2	PI3K-Akt signaling pathway
IFNA21	PI3K-Akt signaling pathway
IFNA4	PI3K-Akt signaling pathway
IFNA5	PI3K-Akt signaling pathway
IFNA6	PI3K-Akt signaling pathway
IFNA7	PI3K-Akt signaling pathway
IFNA8	PI3K-Akt signaling pathway
IFNAR1	PI3K-Akt signaling pathway
IFNAR2	PI3K-Akt signaling pathway
IFNB1	PI3K-Akt signaling pathway
IGF1	PI3K-Akt signaling pathway
IGF1R	PI3K-Akt signaling pathway
IGF2	PI3K-Akt signaling pathway
IGH	PI3K-Akt signaling pathway
IKBKB	PI3K-Akt signaling pathway
IKBKG	PI3K-Akt signaling pathway
IL2	PI3K-Akt signaling pathway
IL2RA	PI3K-Akt signaling pathway
IL2RB	PI3K-Akt signaling pathway
IL2RG	PI3K-Akt signaling pathway
IL3	PI3K-Akt signaling pathway
IL3RA	PI3K-Akt signaling pathway
IL4	PI3K-Akt signaling pathway
IL4R	PI3K-Akt signaling pathway
IL6	PI3K-Akt signaling pathway
IL6R	PI3K-Akt signaling pathway
IL7	PI3K-Akt signaling pathway
IL7R	PI3K-Akt signaling pathway

INS	PI3K-Akt signaling pathway
INSR	PI3K-Akt signaling pathway
IRS1	PI3K-Akt signaling pathway
ITGA1	PI3K-Akt signaling pathway
ITGA10	PI3K-Akt signaling pathway
ITGA11	PI3K-Akt signaling pathway
ITGA2	PI3K-Akt signaling pathway
ITGA2B	PI3K-Akt signaling pathway
ITGA3	PI3K-Akt signaling pathway
ITGA4	PI3K-Akt signaling pathway
ITGA5	PI3K-Akt signaling pathway
ITGA6	PI3K-Akt signaling pathway
ITGA7	PI3K-Akt signaling pathway
ITGA8	PI3K-Akt signaling pathway
ITGA9	PI3K-Akt signaling pathway
ITGAV	PI3K-Akt signaling pathway
ITGB1	PI3K-Akt signaling pathway
ITGB3	PI3K-Akt signaling pathway
ITGB4	PI3K-Akt signaling pathway
ITGB5	PI3K-Akt signaling pathway
ITGB6	PI3K-Akt signaling pathway
ITGB7	PI3K-Akt signaling pathway
ITGB8	PI3K-Akt signaling pathway
JAK1	PI3K-Akt signaling pathway
JAK2	PI3K-Akt signaling pathway
JAK3	PI3K-Akt signaling pathway
KDR	PI3K-Akt signaling pathway
KIT	PI3K-Akt signaling pathway
KITLG	PI3K-Akt signaling pathway
KRAS	PI3K-Akt signaling pathway
LAMA1	PI3K-Akt signaling pathway
LAMA2	PI3K-Akt signaling pathway
LAMA3	PI3K-Akt signaling pathway
LAMA4	PI3K-Akt signaling pathway
LAMA5	PI3K-Akt signaling pathway
LAMB1	PI3K-Akt signaling pathway
LAMB2	PI3K-Akt signaling pathway
LAMB3	PI3K-Akt signaling pathway
LAMB4	PI3K-Akt signaling pathway
LAMC1	PI3K-Akt signaling pathway
LAMC2	PI3K-Akt signaling pathway
LAMC3	PI3K-Akt signaling pathway
LIG1	Base excision repair Mismatch repair DNA replication

	Nucleotide excision repair
LIG3	Base excision repair
LIG4	Non-Homologous End-Joining
LPAR1	PI3K-Akt signaling pathway
LPAR2	PI3K-Akt signaling pathway
LPAR3	PI3K-Akt signaling pathway
LPAR4	PI3K-Akt signaling pathway
LPAR5	PI3K-Akt signaling pathway
LPAR6	PI3K-Akt signaling pathway
MAD1L1	Cell cycle
MAD2L1	Cell cycle
MAD2L2	Cell cycle
MAGI1	PI3K-Akt signaling pathway
MAGI2	PI3K-Akt signaling pathway
MAP2K1	PI3K-Akt signaling pathway
MAP2K2	PI3K-Akt signaling pathway
MAPK1	PI3K-Akt signaling pathway
MAPK3	PI3K-Akt signaling pathway
MBD4	Base excision repair
MCL1	PI3K-Akt signaling pathway
MCM2	Cell cycle DNA replication
MCM3	Cell cycle DNA replication
MCM4	Cell cycle DNA replication
MCM5	Cell cycle DNA replication
MCM6	Cell cycle DNA replication
MCM7	Cell cycle DNA replication
MDM2	PI3K-Akt signaling pathway Cell cycle
MET	PI3K-Akt signaling pathway
MLH1	Fanconi anemia pathway Mismatch repair
MLH3	Mismatch repair
MLST8	PI3K-Akt signaling pathway
MNAT1	Nucleotide excision repair
MPG	Base excision repair
MRE11	Non-Homologous End-Joining Homologous recombination
MSH2	Mismatch repair
MSH3	Mismatch repair
MSH6	Mismatch repair
MTCP1	PI3K-Akt signaling pathway
MTOR	PI3K-Akt signaling pathway

MUS81	Fanconi anemia pathway Homologous recombination
MUTYH	Base excision repair
MYB	PI3K-Akt signaling pathway
MYC	PI3K-Akt signaling pathway Cell cycle
NBN	Homologous recombination
NEIL1	Base excision repair
NEIL2	Base excision repair
NEIL3	Base excision repair
NFKB1	PI3K-Akt signaling pathway
NGF	PI3K-Akt signaling pathway
NGFR	PI3K-Akt signaling pathway
NHEJ1	Non-Homologous End-Joining
NOS3	PI3K-Akt signaling pathway
NR4A1	PI3K-Akt signaling pathway
NRAS	PI3K-Akt signaling pathway
NTF3	PI3K-Akt signaling pathway
NTF4	PI3K-Akt signaling pathway
NTHL1	Base excision repair
NTRK1	PI3K-Akt signaling pathway
NTRK2	PI3K-Akt signaling pathway
OGG1	Base excision repair
ORC1	Cell cycle
ORC2	Cell cycle
ORC3	Cell cycle
ORC4	Cell cycle
ORC5	Cell cycle
ORC6	Cell cycle
OSM	PI3K-Akt signaling pathway
OSMR	PI3K-Akt signaling pathway
PALB2	Fanconi anemia pathway Homologous recombination
PARP1	Base excision repair
PARP2	Base excision repair
PARP3	Base excision repair
PARP4	Base excision repair
PCK1	PI3K-Akt signaling pathway
PCK2	PI3K-Akt signaling pathway
PCNA	Base excision repair Cell cycle Mismatch repair DNA replication Nucleotide excision repair
PDGFA	PI3K-Akt signaling pathway
PDGFB	PI3K-Akt signaling pathway
PDGFC	PI3K-Akt signaling pathway

PDGFD	PI3K-Akt signaling pathway
PDGFRA	PI3K-Akt signaling pathway
PDGFRB	PI3K-Akt signaling pathway
PDPK1	PI3K-Akt signaling pathway
PGF	PI3K-Akt signaling pathway
PHLPP1	PI3K-Akt signaling pathway
PHLPP2	PI3K-Akt signaling pathway
PIK3AP1	PI3K-Akt signaling pathway
PIK3CA	PI3K-Akt signaling pathway
PIK3CB	PI3K-Akt signaling pathway
PIK3CD	PI3K-Akt signaling pathway
PIK3CG	PI3K-Akt signaling pathway
PIK3R1	PI3K-Akt signaling pathway
PIK3R2	PI3K-Akt signaling pathway
PIK3R3	PI3K-Akt signaling pathway
PIK3R5	PI3K-Akt signaling pathway
PIK3R6	PI3K-Akt signaling pathway
PKMYT1	Cell cycle
PKN1	PI3K-Akt signaling pathway
PKN2	PI3K-Akt signaling pathway
PKN3	PI3K-Akt signaling pathway
PLK1	Cell cycle
PMS2	Fanconi anemia pathway Mismatch repair
POLA1	DNA replication
POLA2	DNA replication
POLB	Base excision repair
POLD1	Homologous recombination Base excision repair Mismatch repair DNA replication Nucleotide excision repair
POLD2	Homologous recombination Base excision repair Mismatch repair DNA replication Nucleotide excision repair
POLD3	Homologous recombination Base excision repair Mismatch repair DNA replication Nucleotide excision repair
POLD4	Homologous recombination Base excision repair Mismatch repair DNA replication Nucleotide excision repair
POLE	Base excision repair DNA replication

	Nucleotide excision repair
POLE2	Base excision repair DNA replication Nucleotide excision repair
POLE3	Base excision repair DNA replication Nucleotide excision repair
POLE4	Base excision repair DNA replication Nucleotide excision repair
POLH	Fanconi anemia pathway
POLI	Fanconi anemia pathway
POLK	Fanconi anemia pathway
POLL	Non-Homologous End-Joining Base excision repair
POLM	Non-Homologous End-Joining
POLN	Fanconi anemia pathway
PPP2CA	PI3K-Akt signaling pathway
PPP2CB	PI3K-Akt signaling pathway
PPP2R1A	PI3K-Akt signaling pathway
PPP2R1B	PI3K-Akt signaling pathway
PPP2R2A	PI3K-Akt signaling pathway
PPP2R2B	PI3K-Akt signaling pathway
PPP2R2C	PI3K-Akt signaling pathway
PPP2R2D	PI3K-Akt signaling pathway
PPP2R3A	PI3K-Akt signaling pathway
PPP2R3B	PI3K-Akt signaling pathway
PPP2R3C	PI3K-Akt signaling pathway
PPP2R5A	PI3K-Akt signaling pathway
PPP2R5B	PI3K-Akt signaling pathway
PPP2R5C	PI3K-Akt signaling pathway
PPP2R5D	PI3K-Akt signaling pathway
PPP2R5E	PI3K-Akt signaling pathway
PRIM1	DNA replication
PRIM2	DNA replication
PRKAA1	PI3K-Akt signaling pathway
PRKAA2	PI3K-Akt signaling pathway
PRKCA	PI3K-Akt signaling pathway
PRKDC	Non-Homologous End-Joining Cell cycle
PRL	PI3K-Akt signaling pathway
PRLR	PI3K-Akt signaling pathway
PTEN	PI3K-Akt signaling pathway
PTK2	PI3K-Akt signaling pathway
PTTG1	Cell cycle
PTTG2	Cell cycle
RAC1	PI3K-Akt signaling pathway

RAD21	Cell cycle
RAD23A	Nucleotide excision repair
RAD23B	Nucleotide excision repair
RAD50	Non-Homologous End-Joining Homologous recombination
RAD51	Fanconi anemia pathway Homologous recombination
RAD51B	Homologous recombination
RAD51C	Fanconi anemia pathway Homologous recombination
RAD51D	Homologous recombination
RAD52	Homologous recombination
RAD54B	Homologous recombination
RAD54L	Homologous recombination
RAF1	PI3K-Akt signaling pathway
RB1	Cell cycle
RBBP8	Homologous recombination
RBL1	Cell cycle
RBL2	PI3K-Akt signaling pathway Cell cycle
RBX1	Cell cycle Nucleotide excision repair
RELA	PI3K-Akt signaling pathway
RELN	PI3K-Akt signaling pathway
REV1	Fanconi anemia pathway
REV3L	Fanconi anemia pathway
RFC1	Mismatch repair DNA replication Nucleotide excision repair
RFC2	Mismatch repair DNA replication Nucleotide excision repair
RFC3	Mismatch repair DNA replication Nucleotide excision repair
RFC4	Mismatch repair DNA replication Nucleotide excision repair
RFC5	Mismatch repair DNA replication Nucleotide excision repair
RHEB	PI3K-Akt signaling pathway
RMI1	Fanconi anemia pathway
RMI2	Fanconi anemia pathway
RNASEH1	DNA replication
RNASEH2A	DNA replication
RNASEH2B	DNA replication
RNASEH2C	DNA replication
RPA1	Fanconi anemia pathway

	Homologous recombination Mismatch repair DNA replication Nucleotide excision repair
RPA2	Fanconi anemia pathway Homologous recombination Mismatch repair DNA replication Nucleotide excision repair
RPA3	Fanconi anemia pathway Homologous recombination Mismatch repair DNA replication Nucleotide excision repair
RPA4	Fanconi anemia pathway Homologous recombination Mismatch repair DNA replication Nucleotide excision repair
RPS6	PI3K-Akt signaling pathway
RPS6KB1	PI3K-Akt signaling pathway
RPS6KB2	PI3K-Akt signaling pathway
RPTOR	PI3K-Akt signaling pathway
RXRA	PI3K-Akt signaling pathway
SEM1	Homologous recombination
SFN	Cell cycle
SGK1	PI3K-Akt signaling pathway
SGK2	PI3K-Akt signaling pathway
SGK3	PI3K-Akt signaling pathway
SKP1	Cell cycle
SKP2	Cell cycle
SLX1A	Fanconi anemia pathway
SLX1B	Fanconi anemia pathway
SLX4	Fanconi anemia pathway
SMAD2	Cell cycle
SMAD3	Cell cycle
SMAD4	Cell cycle
SMC1A	Cell cycle
SMC1B	Cell cycle
SMC3	Cell cycle
SMUG1	Base excision repair
SOS1	PI3K-Akt signaling pathway
SOS2	PI3K-Akt signaling pathway
SPP1	PI3K-Akt signaling pathway
SSBP1	Homologous recombination Mismatch repair DNA replication
STAG1	Cell cycle
STAG2	Cell cycle

STK11	PI3K-Akt signaling pathway
SYCP3	Homologous recombination
SYK	PI3K-Akt signaling pathway
TCL1A	PI3K-Akt signaling pathway
TCL1B	PI3K-Akt signaling pathway
TDG	Base excision repair
TEK	PI3K-Akt signaling pathway
TELO2	Fanconi anemia pathway
TFDP1	Cell cycle
TFDP2	Cell cycle
TGFA	PI3K-Akt signaling pathway
TGFB1	Cell cycle
TGFB2	Cell cycle
TGFB3	Cell cycle
THBS1	PI3K-Akt signaling pathway
THBS2	PI3K-Akt signaling pathway
THBS3	PI3K-Akt signaling pathway
THBS4	PI3K-Akt signaling pathway
THEM4	PI3K-Akt signaling pathway
TLR2	PI3K-Akt signaling pathway
TLR4	PI3K-Akt signaling pathway
TNC	PI3K-Akt signaling pathway
TNN	PI3K-Akt signaling pathway
TNR	PI3K-Akt signaling pathway
TNXB	PI3K-Akt signaling pathway
TOP3A	Fanconi anemia pathway Homologous recombination
TOP3B	Fanconi anemia pathway Homologous recombination
TOPBP1	Homologous recombination
TP53	PI3K-Akt signaling pathway Cell cycle
TSC1	PI3K-Akt signaling pathway
TSC2	PI3K-Akt signaling pathway
TTK	Cell cycle
UBE2T	Fanconi anemia pathway
UIMC1	Homologous recombination
UNG	Base excision repair
USP1	Fanconi anemia pathway
VEGFA	PI3K-Akt signaling pathway
VEGFB	PI3K-Akt signaling pathway
VEGFC	PI3K-Akt signaling pathway
VEGFD	PI3K-Akt signaling pathway
VTN	PI3K-Akt signaling pathway
VWF	PI3K-Akt signaling pathway
WDR48	Fanconi anemia pathway

WEE1	Cell cycle
WEE2	Cell cycle
XPA	Nucleotide excision repair
XPC	Nucleotide excision repair
XRCC1	Base excision repair
XRCC2	Homologous recombination
XRCC3	Homologous recombination
XRCC4	Non-Homologous End-Joining
XRCC5	Non-Homologous End-Joining
XRCC6	Non-Homologous End-Joining
YWHAB	PI3K-Akt signaling pathway Cell cycle
YWHAE	PI3K-Akt signaling pathway Cell cycle
YWHAG	PI3K-Akt signaling pathway Cell cycle
YWHAH	PI3K-Akt signaling pathway Cell cycle
YWHAQ	PI3K-Akt signaling pathway Cell cycle
YWHAZ	PI3K-Akt signaling pathway Cell cycle
ZBTB17	Cell cycle

Déséquilibre de liaison et haplotypes

Un génotype correspond à la combinaison de l'allèle provenant du chromosome paternel et de l'allèle provenant du chromosome maternel, sans précision du chromosome hérité sur lequel il se trouve. Un haplotype correspond à une combinaison ordonnée d'allèles situés sur un même chromosome et donc transmis ensemble. En effet, tous les SNPs ne sont pas indépendants les uns des autres. Certaines régions du génome, pouvant s'étendre sur plusieurs kilobases (kb), sont toujours transmises ensemble lors de la méiose et ne subissent pas de recombinaison. Ainsi, certaines combinaisons d'allèles sont plus fréquentes que ne le voudrait le hasard si l'on avait une association aléatoire. On parle de déséquilibre de liaison (DL).

Dans le cas des SNPs bi-alléliques, nous avons deux allèles à chaque locus. Prenons deux SNPs A et B ; le SNP A possède deux allèles, A1 et A2, associés à des fréquences respectives pA1 et pA2. De la même manière, pour le SNP B, nous avons les allèles B1 et B2 avec les fréquences pB1 et pB2. Il existe alors quatre combinaisons gamétiques possibles : A1-B1, A1-B2, A2-B1 et A2-B2. Dans le cas d'une transmission aléatoire, et donc en absence de DL, on aura pour chacune des combinaisons $PA_iB_i = PA_i * PB_i$. Dans le cas d'un DL entre ces deux loci, $P_{A_iB_i} \neq P_{A_i} * P_{B_i}$. Le coefficient de déséquilibre de liaison D qui mesure la force du DL correspond à l'écart par rapport à la ségrégation aléatoire, soit $D = P_{A_iB_i} - P_{A_i} * P_{B_i}$. Lorsqu'il n'y a pas de déséquilibre de liaison D = 0.

Ce concept de DL est très utilisé pour la recherche d'associations entre une maladie et un marqueur génétique. Si une association est trouvée, cela suggère qu'il existe un DL entre le gène de prédisposition et le marqueur. Ce dernier a de grandes chances de se situer à « proximité » du locus de la maladie, ce qui facilite alors l'identification du gène. Le DL est aussi utilisé pour l'inférence des haplotypes. En effet, si deux SNPs sont en DL alors il est très probable que les allèles correspondants se trouvent sur le même haplotype. Cette probabilité dépend de D.

a)

		SNP_1		
		AA	**AG**	**GG**
SNP_2	**TT**	9	10	11
	TA	11	15	14
	AA	13	11	6

b)

		SNP_1		
		AA	AG	GG
SNP_2	**TT**	0	0	25
	TA	0	53	0
	AA	22	0	0

Exemple de fréquences de deux SNP qui a) ne sont pas et b) sont en DL.
(SNP_1 : A_1 = A et A_2 = G ; SNP_2 : A_1 = T et A_2 = A)

Par exemple, considérons un échantillon de 100 individus et deux SNPs (allèles A et G pour SNP_1 et A et T pour SNP_2). Il faut calculer la fréquence de la combinaison de ces deux SNPs dans notre population pour savoir s'ils sont en DL. Si ce n'est pas le cas, les fréquences des combinaisons des différents génotypes possibles sont égales (voir tableau a). Au contraire, le tableau b illustre un DL complet entre ces deux SNPs. Il faut s'intéresser aux individus homozygotes pour les deux SNPs pour définir les haplotypes. En effet, les individus hétérozygotes pour les deux SNPs ne permettent pas de dire si c'est l'allèle A ou l'allèle G du SNP_1 qui est transmis avec l'allèle A du SNP_2. Par contre, grâce aux individus homozygotes, on voit que l'allèle G du SNP_1 est toujours transmis avec l'allèle T du SNP_2. On peut alors dire que ces deux allèles des SNP_1 et SNP_2 appartiennent au même haplotype.

Annexe 6 - Risque de cancer du sein associé aux facteurs gynéco-obstétriques – Analyses stratifiées sur l'année de naissance.

	≤ 1945					[1946-1959]					≥ 1960				
	Témoins	Cas	OR*	$IC_{95\%}$	P	Témoins	Cas	OR*	$IC_{95\%}$	P	Témoins	Cas	OR*	$IC_{95\%}$	P
Âge aux premières règles															
< 12 ans	141	236	1			353	385	1			182	146	1		
[12-15] ans	115	175	1,02	[0,67-1,54]	0,92	278	322	0,92	[0,69-1,22]	0,58	137	100	0,93	[0,65-1,34]	0,72
> 15 ans	41	80	1,69	[0,95-2,98]	0,07	79	110	0,98	[0,64-1,51]	0,94	46	23	0,66	[0,37-1,9]	0,17
Inconnu	4	9	1,97	[0,45-8,57]	0,36	5	3	0,66	[0,09-4,79]	0,68	0	2	NA	NA	NA
Période[1]															
[25-31] jours	205	309	1			452	488	1			214	166	1		
< 25 jours	20	34	1,27	[0,59-2,68]	0,53	67	59	0,83	[0,51-1,35]	0,45	27	23	0,97	[0,51-1,82]	0,92
> 31 jours	10	14	0,61	[0,22-1,83]	0,38	43	45	1,33	[0,77-2,30]	0,31	13	10	1,05	[0,43-2,52]	0,92
Irrégulières	45	104	1,44	[0,86-2,41]	0,17	124	183	1,45	[1,03-2,05]	0,03	82	48	0,65	[0,42-1,02]	0,06
Inconnu	21	39	2,19	[1,03-4,65]	0,04	29	45	0,84	[0,45-1,57]	0,58	29	24	0,77	[0,41-1,45]	0,42
Utilisation de contraceptifs hormonaux[2]															
Non	124	232	1			121	133	1			18	23	1		
Oui	142	232	0,42	[0,27-0,63]	< 0,001	535	613	0,54	[0,37-0,79]	0,001	321	232	0,65	[0,30-1,42]	0,28
Inconnu	35	36	0,41	[0,20-0,84]	0,02	59	74	0,62	[0,35-1,10]	0,10	26	16	0,53	[0,19-1,47]	0,22
Âge à la première utilisation[2]															
≤ 20 ans	4	6	1			210	219	1			250	177	1		
> 20 ans	135	222	0,92	[0,21-4,09]	0,91	323	388	2,99	[2,13-4,20]	< 0,001	69	54	1,41	[0,91-2,18]	0,12
Jamais	124	232	2,24	[0,50-9,98]	0,29	121	133	3,80	[2,42-5,96]	< 0,001	18	23	1,68	[0,76-3,67]	0,19
Inconnu	38	40	0,96	[0,19-4,70]	0,96	61	80	2,14	[1,26-3,62]	0,004	28	17	0,86	[0,43-1,74]	0,68
Durée d'utilisation[2]															
≤ 5 ans	62	92	1			152	206	1			62	56	1		
> 5 ans	80	140	1,46	[0,82-2,57]	0,19	383	407	0,70	[0,50-0,98]	0,03	259	176	0,73	[0,47-1,14]	0,17
Jamais	124	232	3,02	[1,74-5,26]	< 0,001	121	133	1,43	[0,92-2,23]	0,10	18	23	1,20	[0,52-2,79]	0,67
Inconnu	35	36	1,24	[0,56-2,73]	0,59	59	74	0,88	[0,51-1,52]	0,66	26	16	0,63	[0,28-1,40]	0,26
Grossesses menées à terme[3]															
Non	39	49	1			95	95	1			54	44	1		
Oui	261	451	1,37	[0,76-2,47]	0,28	620	724	1,02	[0,68-1,53]	0,90	310	227	1,03	[0,63-1,67]	0,91
Inconnu	1	0	NA	NA	NA	0	1	NA	NA	NA	1	0	NA	NA	NA

* Analyses ajustées sur l'âge à la censure en continu et le niveau d'éducation (élevé/intermédiaire, primaire et pas d'études).

1. * + la prise de contraceptifs hormonaux (oui/non).
2. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans), le statut ménopausique (oui/non) et le nombre de grossesses menées à terme (0, 1, 2, ≥ 3).
3. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans) et le statut ménopausique (oui/non).

	≤ 1945					[1946-1959]					≥ 1960				
	Témoins	Cas	OR*	$IC_{95\%}$	P	Témoins	Cas	OR*	$IC_{95\%}$	P	Témoins	Cas	OR*	$IC_{95\%}$	P
Nombre de grossesses menées à terme[4]															
0	39	49	1			95	95	1			54	44	1		
1	32	73	1,01	[0,41-2,73]	0,89	90	146	1,49	[0,78-2,79]	0,22	50	54	1,95	[0,77-4,94]	0,16
2	118	183	1,24	[0,58-2,70]	0,58	314	375	1,15	[0,67-1,98]	0,60	167	111	1,42	[0,61-3,27]	0,41
≥ 3	111	195	1,89	[0,89-4,00]	0,09	216	203	0,85	[0,50-1,47]	0,55	93	62	1,50	[0,65-3,45]	0,34
Inconnu	1	0	NA	NA	NA	0	1	NA	NA	NA	1	0	NA	NA	NA
Âge à la première grossesse menée à terme[3]															
< 20 ans	45	96	1			117	160	1			21	27	1		
[20-25[ans	133	215	1,05	[0,62-1,77]	0,86	303	325	1,03	[0,69-1,52]	0,88	126	94	0,73	[0,36-1,45]	0,36
[25-30[ans	66	98	0,89	[0,47-1,69]	0,72	148	173	1,15	[0,73-1,82]	0,53	120	73	0,70	[0,34-1,44]	0,34
≥ 30 ans	17	41	1,52	[0,62-3,75]	0,36	52	65	1,61	[0,89-2,98]	0,11	43	33	0,93	[0,42-2,08]	0,86
Jamais	39	49	0,75	[0,37-1,53]	0,43	95	95	1,08	[0,65-1,79]	0,76	54	44	0,74	[0,34-1,62]	0,45
Inconnu	1	1	NA	NA	NA	0	2	NA	NA	NA	1	0	NA	NA	NA
Allaitement[5]															
Non	135	223	1			326	370	1			148	140			
Oui	164	272	1,05	[0,68-1,63]	0,81	385	442	0,98	[0,72-1,33]	0,90	215	131	0,73	[0,49-1,09]	0,12
Inconnu	2	5	69,3	[5,47-879]	0,001	4	8	1,87	[0,39-8,85]	0,43	2	0	NA	NA	NA
Durée d'allaitement[5]															
Jamais	135	223	1			327	370	1			148	140	1		
≤ 10 mois	135	224	1,04	[0,67-1,63]	0,96	346	387	0,94	[0,69-1,29]	0,74	174	108	0,74	[0,49-1,12]	0,15
> 10 mois	26	43	1,07	[0,49-2,32]	0,68	37	50	1,35	[0,73-2,47]	0,33	40	22	0,64	[0,33-1,29]	0,21
Inconnu	3	5	5,17	[1,05-25,4]	0,09	5	13	1,72	[0,44-6,60]	0,43	3	1	NA	NA	NA
Grossesses interrompues[6]															
Non	203	335	1			436	491	1			235	177	1		
Oui	97	164	0,89	[0,59-1,34]	0,58	279	328	1,11	[0,84-1,45]	0,45	129	93	1,09	[0,75-1,58]	0,64
Inconnu	1	1	NA	NA	NA	0	1	NA	NA	NA	1	1	NA	NA	NA
Nombre de grossesses interrompues[6]															
0	203	335	1			436	491	1			235	177	1		
1	65	111	0,90	[0,55-1,47]	0,67	185	225	1,32	[0,96-1,81]	0,08	80	56	1,03	[0,66-1,59]	0,90
2	26	40	1,07	[0,52-2,19]	0,85	75	72	0,75	[0,47-1,21]	0,24	32	27	1,42	[0,77-2,61]	0,25
≥ 3	6	13	1,85	[0,29-3,85]	0,91	19	31	0,91	[0,41-2,03]	0,82	17	10	0,83	[0,35-1,97]	0,67
Inconnu	1	1	NA	NA	NA	0	1	NA	NA	NA	1	1	NA	NA	NA

3. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans) et le statut ménopausique (oui/non).
4. * + 3 + l'âge à la première grossesse menée à terme (< 20 ans, [20-25[, [25-30[, ≥ 30 ans, nullipare) et la prise de contraceptifs hormonaux (oui/non).
5. * + 4 + le nombre de grossesses menées à terme (0, 1, 2, ≥ 3).
6. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans), le statut ménopausique (oui/non), le nombre de grossesses menées à terme (0, 1, 2, ≥ 3) et l'âge à la première grossesse (< 20 ans, [20-25[, [25-30[, ≥ 30 ans, nullipare).

	≤ 1945					[1946-1959]					≥ 1960				
	Témoins	Cas	OR*	$IC_{95\%}$	P	Témoins	Cas	OR*	$IC_{95\%}$	P	Témoins	Cas	OR*	$IC_{95\%}$	P
Statut ménopausique[7]															
Non	0	150				145	552	1			343	267	1		
Oui	301	335				553	224	1,08	[0,74-1,57]	0,67	13	3	0,54	[0,14-2,04]	0,37
Inconnu	0	15				17	44	1,69	[0,88-3,24]	0,11	9	1	NA	NA	NA
Origine de la ménopause[7]															
Naturelle	283	309	1			528	213	1			12	3	#		
Ovariectomie	18	26	1,29	[0,62-2,65]	0,49	24	11	0,66	[0,28-1,54]	0,34	1	0	NA	NA	NA
Non ménopausée	0	150	NA	NA	NA	145	552	0,89	[0,62-1,30]	0,56	343	267	1,62	[0,43-6,09]	0,47
Inconnu	0	15	NA	NA	NA	18	44	1,46	[0,77-2,77]	0,24	9	1	NA	NA	NA
Âge à la ménopause[7]															
[40-55] ans	227	269	1			461	194	1			12	3	1		
< 40 ans	4	4	0,54	[0,11-2,63]	0,44	10	6	0,26	[0,03-1,23]	0,09	1	0	NA	NA	NA
> 55 ans	21	27	1,51	[0,75-3,01]	0,24	24	7	2,36	[0,95-5,87]	0,06	0	0	NA	NA	NA
Non ménopausée	0	150	NA	NA	NA	145	552	0,89	[0,62-1,29]	0,54	343	267	1,7	[0,45-6,41]	0,43
Inconnu	49	50	1,00	[0,57-1,76]	0,99	75	61	1,55	[0,97-2,46]	0,06	9	1	NA	NA	NA
Traitements hormonaux substitutifs[8]															
Non	106	310	1			464	714	1			359	270	1		
Oui	166	141	0,55	[0,36-0,84]	0,006	219	88	1,43	[0,99-2,07]	0,05	6	1	NA	NA	NA
Inconnu	29	49	1,06	[0,55-2,05]	0,86	32	18	1,95	[0,99-4,12]	0,07	0	0	NA	NA	NA
IMC[9]															
[18,5-25]	181	298	1			461	38	1			256	191	1		
< 18,5	4	14	1,62	[0,34-7,70]	0,54	18	38	1,82	[0,86-3,83]	0,11	10	16	1,87	[0,77-4,50]	0,16
> 25	116	187	1,27	[0,86-1-86]	0,23	236	210	0,97	[0,72-1,31]	0,85	99	64	0,84	[0,57-1,26]	0,41
Inconnu	0	1	NA	NA	NA	0	2	NA	NA	NA	0	0	NA	NA	NA

7. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans), et le nombre de grossesses menées à terme (0, 1, 2, ≥ 3).
8. * + l'âge aux premières règles ([12-15], < 12 ans, > 15 ans) et le statut ménopausique (oui/non).
9. * + le statut ménopausique (oui/non).

Annexe 7 - Études participant au consortium BCAC

			Inclus dans l'analyse :	
Acronyme de l'étude	**Nom de l'étude**	**Pays**	***Case-only***	***Control-only***
ABCFS	Australian Breast Cancer Family Study	Australie	oui	oui
ABCTB	Australian Breast Cancer Tissue Bank	Australie	oui	oui
BCEES	Breast Cancer Employment and Environment Study	Australie	oui	oui
MCCS	Melbourne Collaborative Cohort Study	Australie	oui	oui
LMBC	Leuven Multidisciplinary Breast Centre	Belgique	oui	oui
CBCS	Canadian Breast Cancer Study	Canada	oui	oui
MTLGEBCS	Montreal Gene-Environment Breast Cancer Study	Canada	oui	oui
OFBCR	Ontario Familial Breast Cancer Registry	Canada	oui	oui
CGPS	Copenhagen General Population Study	Danemark	oui	oui
HEBCS	Helsinki Breast Cancer Study	Finlande	oui	oui
KBCP	Kuopio Breast Cancer Project	Finlande	oui	oui
CECILE	CECILE Breast Cancer Study	France	oui	oui
BBCC	Bavarian Breast Cancer Cases and Controls	Allemagne	oui	oui
BSUCH	Breast Cancer Study of the University of Heidelberg	Allemagne	oui	oui
ESTHER	ESTHER Breast Cancer Study	Allemagne	oui	oui
GC-HBOC	German Consortium for Hereditary Breast & Ovarian Cancer	Allemagne	oui	oui
GENICA	Gene Environment Interaction and Breast Cancer in Germany	Allemagne	oui	oui
GEPARSIXTO	A randomized phase II trial investigating the addition of carboplatin to neoadjuvant therapy for triple-negative and HER2-positive early breast cancer	Allemagne	non	non
GESBC	Genetic Epidemiology Study of Breast Cancer by Age 50	Allemagne	oui	oui
HABCS	Hannover Breast Cancer Study	Allemagne	oui	oui
MARIE	Mammary Carcinoma Risk Factor Investigation	Allemagne	oui	oui
PREFACE	Evaluation of Predictive Factors regarding the Effectivity of Aromatase Inhibitor Therapy	Allemagne	non	non
SKKDKFZS	Städtisches Klinikum Karlsruhe Deutsches Krebsforschungszentrum Study	Allemagne	oui	non
SUCCESSB	Simultaneous Study of Gemcitabine-Docetaxel Combination adjuvant treatment	Allemagne	non	non
SUCCESSC	Simultaneous Study of Docetaxel Based Anthracycline Free Adjuvant Treatment Evaluation	Allemagne	non	non
CCGP	Crete Cancer Genetics Program	Grèce	oui	oui
BCINIS	Breast Cancer In Northern Israel Study	Israël	oui	oui
MBCSG	Milan Breast Cancer Study Group	Italie	oui	oui
ABCS	Amsterdam Breast Cancer Study	Pays-Bas	oui	oui
ORIGO	Leiden University Medical Centre Breast Cancer Study	Pays-Bas	oui	oui
RBCS	Rotterdam Breast Cancer Study	Pays-Bas	oui	oui
PBCS	NCI Polish Breast Cancer Study	Pologne	oui	oui
SZBCS	IHCC-Szczecin Breast Cancer Study	Pologne	oui	oui
HUBCS	Hannover-Ufa Breast Cancer Study	Russie	oui	oui
BREOGAN	Breast Oncology Galicia Network	Espagne	oui	oui
HCSC	Hospital Clinico San Carlos	Espagne	oui	non
KARBAC	Karolinska Breast Cancer Study	Suède	oui	non
MISS	Melanoma Inquiry of Southern Sweden	Suède	oui	oui

			Inclus dans l'analyse :	
Acronyme de l'étude	**Nom de l'étude**	**Pays**	***Case-only***	***Control-only***
pKARMA	Karolinska Mammography Project for Risk Prediction of Breast Cancer – Case-Control Study	Suède	oui	oui
SMC	Swedish Mammography Cohort	Suède	oui	oui
BBCS	British Breast Cancer Study	Angleterre	oui	oui
DIETCOMPLYF	DietCompLyf Breast Cancer Survival Study	Angleterre	oui	non
POSH	Prospective Study of Outcomes in Sporadic Versus Hereditary Breast Cancer	Angleterre	oui	non
SEARCH	Study of Epidemiology and Risk factors in Cancer Heredity	Angleterre	oui	oui
UKBGS	UK Breakthrough Generations Study	Angleterre	oui	oui
UKOPS	UK Ovarian Cancer Population Study	Angleterre	non	oui
2SISTER	The Two Sister Study	USA	oui	non
BCFR-NY	New York Breast Cancer Family Registry	USA	oui	oui
BCFR-PA	Philadelphia Breast Cancer Family Registry	USA	oui	non
BCFR-UTAH	Utah Breast Cancer Family Registry	USA	oui	non
CPSII	Cancer Prevention Study-II Nutrition Cohort	USA	oui	oui
CTS	California Teachers Study	USA	oui	oui
MCBCS	Mayo Clinic Breast Cancer Study	USA	oui	oui
MEC	Multiethnic Cohort	USA	oui	oui
MMHS	Mayo Mammography Health Study	USA	oui	oui
MSKCC	Memorial Sloan-Kettering Cancer Center Study	USA	oui	non
NBHS	Nashville Breast Health Study	USA	oui	oui
NC-BCFR	Northern California Breast Cancer Family Registry	USA	oui	oui
NHS	Nurses Health Study	USA	oui	oui
NHS2	Nurses Health Study 2	USA	oui	oui
PLCO	The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial	USA	oui	oui
SISTER	The Sister Study	USA	oui	oui
UCIBCS	UCI Breast Cancer Study	USA	oui	oui
EPIC	European Prospective Investigation Into Cancer and Nutrition (BPC3)	Varié	oui	oui
TNBCC	Triple Negative Breast Cancer Consortium Study	Varié	oui	non

Annexe 8 - Études participant au consortium CIMBA

Acronyme de l'étude	Nom de l'étude	Pays
BCFR-AU	Australian site of the Breast Cancer Family Registry	Australie
KCONFAB	Kathleen Cuningham Consortium for Research into Familial Breast Cancer	Australie
VFCTG	Victorian Familial Cancer Trials Group	Australie
G-FAST	Ghent University Hospital	Belgique
BCFR-ON/OCGN	Ontario site of the Breast Cancer Family Registry/Ontario Cancer Genetics Network	Canada
INHERIT	INterdisciplinary HEalth Research Internal Team BReast CAncer susceptibility	Canada (Quebec)
MCGILL	McGill University	Canada (Quebec)
CBCS	Copenhagen Breast Cancer Study	Danemark
OUH	Odense University Hospital	Danemark
HEBCS	Helsinki Breast Cancer Study	Finlande
GEMO	Genetic Modifiers of cancer risk in BRCA1/2 mutation carriers	France/USA
GC-HBOC	German Familial Breast Group	Allemagne
DKFZ	German Cancer Research Center	Allemagne/Pakistan/Colombie
DEMOKRITOS	National Centre for Scientific Research Demokritos	Grèce
SMC	Sheba Medical Centre	Israël
CONSIT TEAM	CONsorzio Studi ITaliani sui Tumori Ereditari Alla Mammella	Italie
IOVHBOCS	Istituto Oncologico Veneto	Italie
PBCS	Università di Pisa	Italie
HEBON	Hereditary Breast and Ovarian cancer study the Netherlands	Pays-Bas
IHCC	International Hereditary Cancer Centre	Pologne
NNPIO	N.N. Petrov Institute of Oncology	Russie
CNIO	Spanish National Cancer Centre	Espagne
FPGMX	Fundación Pública Galega de Medicina Xenómica	Espagne
HCSC	Hospital Clinico San Carlos	Espagne
HVH	University Hospital Vall d'Hebron	Espagne
ICO	Institut Català d'Oncologia	Espagne
SWE-BRCA	Swedish Breast Cancer Study	Sweden
EMBRACE	Epidemiological Study of Familial Breast Cancer	Angleterre
UKGRFOCR	UK and Gilda Radner Familial Ovarian Cancer Registries	Angleterre/USA
BCFR-NC	Northern California site of the Breast Cancer Family Registry	USA
BCFR-NY	New York site of the Breast Cancer Family Registry	USA
BCFR-PA	Philadelphia site of the Breast Cancer Family Registry	USA
BCFR-UT	Utah site of the Breast Cancer Family Registry	USA
BIDMC	Beth Israel Deaconess Medical Center	USA
BRICOH	Beckman Research Institute of the City of Hope	USA
COH	City of Hope Cancer Center	USA
DFCI	Dana Farber Cancer Institute	USA
FCCC	Fox Chase Cancer Center	USA
GEORGETOWN	Georgetown University	USA
KUMC	University of Kansas Medical Center	USA
MAYO	Mayo Clinic	USA

Acronyme de l'étude	Nom de l'étude	Pays
MSKCC	Memorial Sloane Kettering Cancer Center	USA
NCI	National Cancer Institute	USA
NORTHSHORE	NorthShore University HealthSystem	USA
OSU CCG	The Ohio State University Comprehensive Cancer Center	USA
UCHICAGO	University of Chicago	USA
UCSF	University of California San Francisco	USA
UPENN	University of Pennsylvania	USA
UPITT	Cancer Family Registry University of Pittsburg	USA
UTMDACC	University of Texas MD Anderson Cancer Center	USA
WCP	Women's Cancer Program at Cedars-Sinai Medical Center	USA
NRG_ONCOLOGY	NRG Oncology	USA/Australie

VALORISATIONS SCIENTIFIQUES

Article soumis :

1- **Juliette Coignard***, Michael Lush, Joe Dennis, [>200 coauteurs], Jacques Simard, Georgia Chenevix-Trench, Douglas Easton, Nadine Andrieu, Antonis Antoniou on behalf of CIMBA and BCAC consortia. **A case-only study for identifying genetic modifiers of breast cancer risk for BRCA1/2 mutation carriers.** *(Soumis au journal Nature communication).*

Communications orales :

1- **Juliette Coignard***, Christine Lonjou, Marie-Gabrielle Dondon, Séverine Eon-Marchais, Francesca Damiola, Laure Barjhoux, Morgane Marcou, Carole Verny-Pierre, Valérie Sornin, Lucie Toulemonde, Juana Beauvallet, Dorothée Le Gal, Noura Mebirouk, Muriel Belotti, Olivier Caron, Marion Gauthier-Villars, Isabelle Coupier, Bruno Buecher, Alain Lortholary, Catherine Dugast, Paul Gesta, Jean-Pierre Fricker, Catherine Nogues, Laurence Faivre, Elisabeth Luporsi, Pascaline Berthet, Capucine Delnatte, Valérie Bonadona, Christine M Maugard, Pascal Pujol, Christine Lasset, Michel Longy, Yves-Jean Bignon, Claude Adenis, Laurence Venat-Bouvet, Liliane Demange†, Hélène Dreyfus, Marc Frenay, Laurence Gladieff, Isabelle Mortemousque, Séverine Audebert-Bellanger, Florent Soubrier, Sophie Giraud, Sophie Lejeune-Dumoulin, Annie Chevrier, Jean-Marc Limacher, Jean Chiesa, Anne Fajac, Anne Floquet, François Eisinger, Julie Tinat, Chrystelle Colas, Sandra Fert-Ferrer, Clotilde Penet, Thierry Frebourg, Marie-Agnès Collonge-Rame, Emmanuelle Barouk-Simonet, Valérie Layet, Dominique Leroux, Odile Cohen-Haguenauer, Fabienne Prieur, Emmanuelle Mouret- Fourme, François Cornelis, Philippe Jonveaux, Odile Bera, Eve Cavaciuti, Sylvie Mazoyer, Olga M. Sinilnikova†, Fabienne Lesueur, Dominique Stoppa-Lyonnet, Nadine Andrieu. **Impact des facteurs de la reproduction sur l'association entre les gènes de la régulation des œstrogènes et le risque de cancer du sein : une stratégie pour l'étude GENESIS**, 8èmes Assises de Génétique humaine et médicale, Lyon, France, Février 2016.

2- **Juliette Coignard***, Joe Dennis, Michael Lush, Jacques Simard, Georgia Chenevix-Trench, Douglas Easton, Nadine Andrieu, Antonis, Antoniou on behalf of CIMBA and BCAC consortia, **Combined CIMBA- BCAC case only analysis**, **18th CIMBA meeting**, Limassol, Chypres, 13 Janvier 2017.

3- **Juliette Coignard***, Joe Dennis, Michael Lush, Jacques Simard, Georgia Chenevix-Trench, Douglas Easton, Nadine Andrieu, Antonis Antoniou on behalf of CIMBA and BCAC consortia, **Combined CIMBA- BCAC case only analysis**, **20th BCAC meeting**, Limassol, Chypres, 12 Janvier 2017.

4- **Juliette Coignard*,** Joe Dennis, Michael Lush, Jacques Simard, Georgia Chenevix-Trench, Douglas Easton, Nadine Andrieu, Antonis Antoniou on behalf of CIMBA and BCAC consortia, **Combined CIMBA-BCAC case only analysis**, **20th BCAC meeting**, Saint-Jacques de Compostelle, Espagne, 17 Septembre 2017.

5- **Juliette Coignard*,** Joe Dennis, Michael Lush, Jacques Simard, Georgia Chenevix-Trench, Douglas Easton, Nadine Andrieu, Antonis Antoniou on behalf of CIMBA and BCAC consortia, **Assessment of the joint effects of SNPs and BRCA1/2 mutations on breast cancer risk: A combined CIMBA-BCAC case only analysis**, **21th BCAC meeting**, Edimbourg, Royaume-Uni, 15 Juin 2018.

6- **Juliette Coignard*,** Joe Dennis, Michael Lush, Jacques Simard, Georgia Chenevix-Trench, Douglas Easton, Nadine Andrieu, Antonis Antoniou on behalf of CIMBA and BCAC consortia, **Assessment of the joint effects of SNPs and BRCA1/2 mutations on breast cancer risk: A combined CIMBA-BCAC case only analysis**, **22th BCAC meeting**, Springdale, Utah, États-Unis, 9 Avril 2019.

Communications affichées :

1- **Juliette Coignard*,** Joe Dennis, Michael Lush, Jacques Simard, Georgia Chenevix-Trench, Douglas Easton, Nadine Andrieu, Antonis Antoniou on behalf of CIMBA and BCAC consortia, **Assessment of the joint effects of common genetic variants and BRCA1/2 mutations on breast cancer risk**, **the « young researchers in life science » conference,** Institut Imagine, Paris, 15-17 Mai 2017.

2- **Juliette Coignard***, Joe Dennis, Michael Lush, Jacques Simard, Georgia Chenevix-Trench, Douglas Easton, Nadine Andrieu, Antonis Antoniou on behalf of CIMBA and BCAC consortia. **Assessment of the joint effects of common genetic variants and BRCA1/2 mutations on breast cancer risk**, **8èmes Assises de Génétique humaine et médicale**, Nantes, 24-26 Janvier 2018.

Articles de collaborations :

1- Guerra M, **Coignard J** *et al*., **Diagnostic Chest X-Rays and Breast Cancer Risk in High Risk Population: findings in the GENESIS Study**. (En préparation).

2- Girard E, Eon-Marchais S, Olaso R, Renault AL, Damiola F, Dondon MG, Barjhoux L, Goidin D, Meyer V, Le GD, Beauvallet J, Mebirouk N, Lonjou C, **Coignard J**, Marcou M, Cavaciuti E, Baulard C, Bihoreau MT, Cohen-Haguenauer O, Leroux D, Penet C, Fert-Ferrer S, Colas C, Frebourg T, Eisinger F, Adenis C, Fajac A, Gladieff L, Tinat J, Floquet A, Chiesa J, Giraud S, Mortemousque I, Soubrier F, Audebert-Bellanger S, Limacher JM, Lasset C, Lejeune-Dumoulin S, Dreyfus H, Bignon YJ, Longy M, Pujol P, Venat-Bouvet L, Bonadona V, Berthet P, Luporsi E, Maugard CM, Nogues C, Delnatte C, Fricker JP, Gesta P, Faivre L, Lortholary A, Buecher B, Caron O, Gauthier-Villars M, Coupier I, Servant N, Boland A, Mazoyer S, Deleuze JF, Stoppa-Lyonnet D, Andrieu N, Lesueur F (2019) **Familial breast cancer and DNA repair genes: Insights into known and novel susceptibility genes from the GENESIS study, and implications for multigene panel testing.** *Int J Cancer* 144 (8): 1962-1974, doi:10.1002/ijc.31921

ARTICLE SOUMIS

A case-only study to identify genetic modifiers of breast cancer risk specifically for *BRCA1* and *BRCA2* mutation carriers.

Juliette Coignard[1-6], Michael Lush[4], Jonathan Beesley[7], Tracy A. O'Mara[7], Joe Dennis[4], Jonathan P. Tyrer[8], Daniel R. Barnes[4], Lesley McGuffog[4], Goska Leslie[4], Manjeet K. Bolla[4], Muriel A. Adank[9], Simona Agata[10], Thomas Ahearn[11], Kristiina Aittomäki[12], Irene L. Andrulis[13,14], Hoda Anton-Culver[15], Volker Arndt[16], Norbert Arnold[17,18], Kristan J. Aronson[19], Banu K. Arun[20], Annelie Augustinsson[21], Jacopo Azzollini[22], Daniel Barrowdale[4], Caroline Baynes[8], Heko Becher[23], Marina Bermisheva[24], Leslie Bernstein[25], Katarzyna Białkowska[26], Carl Blomqvist[27,28], Stig E. Bojesen[29-31], Bernardo Bonanni[32], Ake Borg[33], Hiltrud Brauch[34-36], Hermann Brenner[16,36,37], Barbara Burwinkel[38,39], Saundra S. Buys[40], Trinidad Caldés[41], Maria A. Caligo[42], Daniele Campa[43,44], Brian D. Carter[45], Jose E. Castelao[46], Jenny Chang-Claude[44,47], Stephen J. Chanock[11], Wendy K. Chung[48], Kathleen B.M. Claes[49], Christine L. Clarke[50], GEMO Study Collaborators[1-3], EMBRACE Collaborators[4], J. Margriet Collée[51], Don M. Conroy[8], Kamila Czene[52], Mary B. Daly[53], Peter Devilee[54,55], Orland Diez[56,57], Yuan Chun Ding[25], Susan M. Domchek[58], Thilo Dörk[59], Isabel dos-Santos-Silva[60], Alison M. Dunning[8], Miriam Dwek[61], Diana M. Eccles[62], A. Heather Eliassen [63,64], Christoph Engel[65], Mikael Eriksson[52], D. Gareth Evans[66,67], Peter A. Fasching[68,69], Henrik Flyger[70], Florentia Fostira[71], Eitan Friedman[72,73], Lin Fritschi[74], Debra Frost[4], Manuela Gago-Dominguez[75,76], Susan M. Gapstur[45], Judy Garber[77], Vanesa Garcia-Barberan[78], Montserrat García-Closas[11], José A. García-Sáenz[78], Mia M. Gaudet[45], Simon A. Gayther[79], Andrea Gehrig[80], Vassilios Georgoulias[81], Graham G. Giles[82-84], Andrew K. Godwin[85], Mark S. Goldberg[86,87], David E. Goldgar[88], Anna González-Neira[89], Mark H. Greene[90], Pascal Guénel[91], Lothar Haeberle[92], Eric Hahnen[93,94], Christopher A. Haiman[95], Niclas Håkansson[96], Per Hall[52,97], Ute Hamann[98], Patricia A. Harrington[8], Steven N. Hart[99], Wei He[52], Frans B.L. Hogervorst[9], Antoinette Hollestelle[100], John L. Hopper[83], Darling J. Horcasitas[101], Peter J. Hulick[102,103], David J. Hunter[64,104,105], Evgeny N. Imyanitov[106], KConFab Investigators[107,108], HEBON Investigators[109], ABCTB Investigators[110], Agnes Jager[100], Anna Jakubowska[26,111], Paul A. James[108,112], Uffe Birk Jensen[113], Esther M. John[114], Michael E. Jones[115], Rudolf Kaaks[44], Pooja Middha Kapoor[44,116], Beth Y. Karlan[117,118], Renske Keeman[119], Elza Khusnutdinova[24,120], Johanna I. Kiiski[121], Yon-Dschun Ko[122], Veli-Matti Kosma[123-125], Peter Kraft[64,104], Allison W. Kurian[114,126], Yael Laitman[72], Diether Lambrechts[127,128], Loic Le Marchand[129], Jenny Lester[117,118], Fabienne Lesueur[1-3], Tricia Lindstrom[99], Adria Lopez-Fernández[130], Jennifer T. Loud[90], Craig Luccarini[8], Arto Mannermaa[123-125], Siranoush Manoukian[22], Sara Margolin[97,131], John W.M. Martens[100], Noura Mebirouk[1-3], Alfons Meindl[132], Austin Miller[133], Roger L. Milne[82-84], Marco Montagna[10], Katherine L. Nathanson[58], Susan L. Neuhausen[25], Heli Nevanlinna[121], Finn C. Nielsen[134], Katie M. O'Brien[135], Olufunmilayo I. Olopade[136], Janet E. Olson[99], Håkan Olsson[21], Ana Osorio[89,137], Laura Ottini[138], Tjoung-Won Park-Simon[59], Michael T. Parsons[7], Inge Sokilde Pedersen[139-141], Beth Peshkin[142], Paolo Peterlongo[143], Julian Peto[60], Paul D.P. Pharoah[4,8], Kelly-Anne Phillips[7,83,108], Eric C. Polley[99], Bruce Poppe[49], Nadege Presneau[61], Miquel Angel Pujana[144], Kevin Punie[145], Paolo Radice[146], Johanna Rantala[147], Muhammad U. Rashid[98,148], Gad Rennert[149], Hedy S. Rennert[149], Mark Robson[150], Atocha Romero[151], Maria Rossing[134],

Emmanouil Saloustros[152], Dale P. Sandler[135], Regina Santella[153], Maren T. Scheuner[154], Marjanka K. Schmidt[119,155], Gunnar Schmidt[156], Christopher Scott[99], Priyanka Sharma[157], Penny Soucy[158], Melissa C. Southey[84,159,160], John J. Spinelli[161,162], Zoe Steinsnyder [163], Jennifer Stone[83,164], Dominique Stoppa-Lyonnet[165-167], Anthony Swerdlow[115,168], Rulla M. Tamimi[63,64,104], William J. Tapper[62], Jack A. Taylor[135,169], Mary Beth Terry[153], Alex Teulé[170], Darcy L. Thull[171], Marc Tischkowitz[172,173], Amanda E. Toland[174], Diana Torres[98,175], Alison H. Trainer[112,176], Thérèse Truong[91], Nadine Tung[177], Celine M. Vachon[178], Ana Vega[179], Joseph Vijai[150,163], Qin Wang[4], Barbara Wappenschmidt[93,94], Clarice R. Weinberg[180], Jeffrey N. Weitzel[181], Camilla Wendt[131], Alicja Wolk[96,182], Siddhartha Yadav[183], Xiaohong R. Yang[11], Drakoulis Yannoukakos[71], Wei Zheng[184], Argyrios Ziogas[15], Kristin K. Zorn[185], Sue K. Park[186-188], Mads Thomassen[189], Kenneth Offit[150,163], Rita K. Schmutzler[93,94], Fergus J. Couch[190], Jacques Simard[158], Georgia Chenevix-Trench[7], Douglas F. Easton[4,8], Nadine Andrieu[1-3,5], Antonis C. Antoniou[4]

1. Genetic Epidemiology of Cancer team Inserm U900 Paris: France, 75005.
2. Institut Curie Paris: France, 75005.
3. Mines ParisTech Fontainebleau: France, 77305.
4. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care University of Cambridge, Cambridge, UK, CB1 8RN.
5. PSL University Paris, France, 75006.
6. Paris Sud University, Orsay, France.
7. Department of Genetics and Computational Biology QIMR Berghofer Medical Research Institute, vol. Locked Bag 2000, Herston, QLD 4029 Brisbane, Queensland, Australia.
8. Centre for Cancer Genetic Epidemiology, Department of Oncology University of Cambridge, Cambridge CB1 8RN, UK.
9. Family Cancer Clinic, The Netherlands Cancer Institute, Antoni van Leeuwenhoek hospital, vol. P.O. Box 90203, 1006 BE Amsterdam, The Netherlands.
10. Immunology and Molecular Oncology, Unit Veneto Institute of Oncology IOV – IRCCS, 35128 Padua, Italy.
11. Division of Cancer Epidemiology and Genetics National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20850, USA.
12. Department of Clinical Genetics, Helsinki University Hospital University of Helsinki, vol. P.O. BOX160(Meilahdentie 2), 00029 HUS Helsinki, Finland.
13. Fred A. Litwin Center for Cancer Genetics Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada.
14. Department of Molecular Genetics University of Toronto Toronto, ON: Canada, M5S 1A8.
15. Department of Epidemiology, Genetic Epidemiology Research Institute University of California Irvine, Irvine, CA 92617, USA.
16. Division of Clinical Epidemiology and Aging Research German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.
17. Department of Gynaecology and Obstetrics University Hospital of Schleswig-Holstein, Campus Kiel, Christian-Albrechts University Kiel, Kiel 24118, Germany.
18. Institute of Clinical Molecular Biology University Hospital of Schleswig-Holstein, Campus Kiel, Christian-Albrechts University Kiel, Kiel 24118, Germany.
19. Department of Public Health Sciences, and Cancer Research Institute Queen's University, Kingston, ON K7L 3N6, Canada.
20. Department of Breast Medical Oncology University of Texas MD Anderson Cancer Center Houston, TX: USA, 77030.

21. Department of Cancer Epidemiology, Clinical Sciences Lund University, Lund 222 42, Sweden.
22. Unit of Medical Genetics, Department of Medical Oncology and Hematology Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan 20133, Italy.
23. Institute for Medical Biometrics and Epidemiology University Medical Center Hamburg-Eppendorf, Hamburg 20246, Germany.
24. Institute of Biochemistry and Genetics Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa 450054, Russia.
25. Department of Population Sciences Beckman Research Institute of City of Hope Duarte, CA 91010, USA.
26. Department of Genetics and Pathology Pomeranian Medical University Szczecin, Poland, 71-252.
27. Department of Oncology, Helsinki University Hospital University of Helsinki, vol. P.O. BOX 180 (Haartmaninkatu 4), 00029 HUS, Helsinki 00290, Finland.
28. Department of Oncology Örebro University Hospital, Örebro 70185, Sweden.
29. Copenhagen General Population Study, Herlev and Gentofte Hospital Copenhagen University Hospital, Herlev 2730, Denmark.
30. Department of Clinical Biochemistry, Herlev and Gentofte Hospital Copenhagen University Hospital, Herlev 2730, Denmark.
31. Faculty of Health and Medical Sciences University of Copenhagen, Copenhagen 2200, Denmark.
32. Division of Cancer Prevention and Genetics IEO, European Institute of Oncology IRCCS, Milan 20141, Italy.
33. Department of Oncology Lund University and Skåne University Hospital, Lund 222 41, Sweden.
34. Dr Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart 70376, Germany.
35. iFIT-Cluster of Excellence University of Tübingen, Tübingen 72074, German.
36. German Cancer Consortium (DKTK) German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.
37. Division of Preventive Oncology German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg 69120, Germany.
38. Molecular Epidemiology Group, C080 German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.
39. Molecular Biology of Breast Cancer, University Womens Clinic Heidelberg University of Heidelberg, Heidelberg 69120, Germany.
40. Department of Medicine Huntsman Cancer Institute, Salt Lake City, UT 84112, USA.
41. Molecular Oncology Laboratory CIBERONC, Hospital Clinico San Carlos, IdISSC (Instituto de Investigación Sanitaria del Hospital Clínico San Carlos), Madrid 28040, Spain.
42. SOD Genetica Molecolare University Hospital, Pisa, Italy.
43. Department of Biology University of Pisa, Pisa 56126, Italy.
44. Division of Cancer Epidemiology German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.
45. Behavioral and Epidemiology Research Group American Cancer Society Atlanta, GA 30303, USA.
46. Oncology and Genetics Unit Instituto de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo 36312, Spain.

47. Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH) University Medical Center Hamburg-Eppendorf, Hamburg 20246, Germany.
48. Departments of Pediatrics and Medicine Columbia University New York, NY: USA, 10032.
49. Centre for Medical Genetics Ghent University, Gent 9000, Belgium.
50. Westmead Institute for Medical Research University of Sydney, Sydney, New South Wales 2145, Australia.
51. Department of Clinical Genetics Erasmus University Medical Center, vol. P.O. Box 2040, 3000 CA, Rotterdam 3015 CN, The Netherlands.
52. Department of Medical Epidemiology and Biostatistics Karolinska Institutet, Stockholm 3015 CN, Sweden.
53. Department of Clinical Genetics Fox Chase Cancer Center Philadelphia, PA 19111, USA.
54. Department of Pathology Leiden University Medical Center, Leiden 2333 ZA, The Netherlands.
55. Department of Human Genetics Leiden University Medical Center, vol. P.O. Box 9600, 2300 RC, Leiden 2333 ZA, The Netherlands.
56. Oncogenetics Group Vall dHebron Institute of Oncology (VHIO), Barcelona 8035, Spain.
57. Clinical and Molecular Genetics Area University Hospital Vall dHebron, Barcelona 8035, Spain.
58. Basser Center for BRCA, Abramson Cancer Center University of Pennsylvania, Philadelphia, PA 19066, USA.
59. Gynaecology Research Unit Hannover Medical School, Hannover 30625, Germany.
60. Department of Non-Communicable Disease Epidemiology London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK.
61. School of Life Sciences University of Westminster, London W1B 2HW, UK.
62. Faculty of Medicine University of Southampton, Southampton SO17 1BJ, UK.
63. Channing Division of Network Medicine, Department of Medicine Brigham and Womenís Hospital and Harvard Medical School Boston, MA 02115, USA.
64. Department of Epidemiology Harvard TH Chan School of Public Health, Boston, MA 02115, USA.
65. Institute for Medical Informatics, Statistics and Epidemiology University of Leipzig, Leipzig 04107, Germany.
66. Genomic Medicine, Division of Evolution and Genomic Sciences The University of Manchester, Manchester Academic Health Science Centre, Manchester Universities Foundation Trust, St Mary's Hospital, Manchester M13 9WL, UK.
67. Genomic Medicine, North West Genomics hub Manchester Academic Health Science Centre, Manchester Universities Foundation Trust, St Mary's Hospital, Manchester M13 9WL, UK.
68. David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology University of California at Los Angeles, Los Angeles, CA 90095, USA.
69. Department of Gynecology and Obstetrics, Comprehensive Cancer Center ER-EMN University Hospital Erlangen, Friedrich-Alexander-University, Erlangen-Nuremberg, Erlangen 91054, Germany.
70. Department of Breast Surgery, Herlev and Gentofte Hospital Copenhagen University Hospital, Herlev, 2730, Denmark.

71. Molecular Diagnostics Laboratory, INRASTES National Centre for Scientific Research íDemokritosí, Athens 15310, Greece.
72. The Susanne Levy Gertner Oncogenetics Unit Chaim Sheba Medical Center, Ramat Gan 52621, Israel.
73. Sackler Faculty of Medicine Tel Aviv University, Ramat Aviv 69978, Israel.
74. School of Public Health Curtin University, Perth, Western Australia 6102, Australia.
75. Genomic Medicine Group, Galician Foundation of Genomic Medicine Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela 15706, Spain.
76. Moores Cancer Center University of California, San Diego La Jolla, CA 92037, USA.
77. Cancer Risk and Prevention Clinic Dana-Farber Cancer Institute, Boston, MA 02215, USA.
78. Medical Oncology Department, Hospital Clínico San Carlos Instituto de Investigación Sanitaria San Carlos (IdISSC), Centro Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid 28040, Spain.
79. Center for Bioinformatics and Functional Genomics and the Cedars Sinai Genomics Core Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA.
80. Department of Human Genetics University Würzburg, Würzburg 97074, Germany.
81. Department of Medical Oncology University Hospital of Heraklion, Heraklion 711 10, Greece.
82. Cancer Epidemiology Division Cancer Council Victoria, Melbourne, Victoria 3004, Australia.
83. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria 3010, Australia.
84. Precision Medicine, School of Clinical Sciences at Monash Health Monash University, Clayton, Victoria 3168, Australia.
85. Department of Pathology and Laboratory Medicine Kansas University Medical Center Kansas City, KS: USA, 66160.
86. Department of Medicine McGill University Montréal, QC: Canada, H4A 3J1.
87. Division of Clinical Epidemiology, Royal Victoria Hospital McGill University Montréal, QC: Canada, H4A 3J1.
88. Department of Dermatology Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT 84112, USA.
89. Human Cancer Genetics Programme Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain.
90. Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics National Cancer Institute, Bethesda, MD 20850-9772, USA.
91. Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP) INSERM, University Paris-Sud, University Paris-Saclay, Villejuif 94805, France.
92. Department of Gynaecology and Obstetrics, University Hospital Erlangen Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen 91054, Germany.
93. Center for Hereditary Breast and Ovarian Cancer Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne 50937, Germany.
94. Center for Integrated Oncology (CIO) Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne 50937, Germany.

95. Department of Preventive Medicine, Keck School of Medicine University of Southern California, Los Angeles, CA 90033, USA.
96. Institute of Environmental Medicine Karolinska Institutet, vol. Box 210, Stockholm 171 77, Sweden.
97. Department of Oncology Södersjukhuset, Stockholm 118 83, Sweden.
98. Molecular Genetics of Breast Cancer German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.
99. Department of Health Sciences Research Mayo Clinic Rochester, MN 55905, USA.
100. Department of Medical Oncology, Family Cancer Clinic Erasmus MC Cancer Institute, vol. P.O. Box 2040, 3000 CA Rotterdam: The Netherlands.
101. New Mexico Oncology Hematology Consultants University of New Mexico Albuquerque, NM, USA.
102. Center for Medical Genetics NorthShore University HealthSystem, Evanston, IL 60201, USA.
103. The University of Chicago Pritzker School of Medicine Chicago, IL 60637, USA.
104. Program in Genetic Epidemiology and Statistical Genetics Harvard TH Chan School of Public Health Boston, MA 02115, USA.
105. Nuffield Department of Population Health University of Oxford, Oxford OX3 7LF, UK.
106. NN Petrov Institute of Oncology, St. Petersburg 197758, Russia,.
107. Peter MacCallum Cancer Center Melbourne, Victoria 3000, Australia.
108. Sir Peter MacCallum Department of Oncology The University of Melbourne, Melbourne, Victoria 3000, Australia.
109. The Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON) Coordinating center: The Netherlands Cancer Institute, vol. P.O. Box 90203, 1006 BE Amsterdam 1066 CX, The Netherlands.
110. Australian Breast Cancer Tissue Bank, Westmead Institute for Medical Research University of Sydney, Sydney, New South Wales 2145, Australia.
111. Independent Laboratory of Molecular Biology and Genetic Diagnostics Pomeranian Medical University, Szczecin 71-252, Poland.
112. Parkville Familial Cancer Centre Peter MacCallum Cancer Center, Melbourne, Victoria 3000, Australia.
113. Department of Clinical Genetics Aarhus University Hospital Aarhus N: Denmark, 8200.
114. Department of Medicine, Division of Oncology Stanford Cancer Institute, Stanford University School of Medicine Stanford, CA 94304, USA.
115. Division of Genetics and Epidemiology The Institute of Cancer Research, London SM2 5NG, UK.
116. Faculty of Medicine University of Heidelberg, Heidelberg 69120, Germany.
117. David Geffen School of Medicine, Department of Obstetrics and Gynecology University of California at Los Angeles, Los Angeles, CA 90095, USA.
118. Womenís Cancer Program at the Samuel Oschin Comprehensive Cancer Institute Cedars-Sinai Medical Center, Los Angeles, CA 90048,USA.
119. Division of Molecular Pathology The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam 1066 CX, The Netherlands.
120. Department of Genetics and Fundamental Medicine Bashkir State Medical University, Ufa 450000, Russia.

121. Department of Obstetrics and Gynecology, Helsinki University Hospital University of Helsinki, vol. P.O. BOX 700 (Haartmaninkatu 8), 00029 HUS, Helsinki 450000, Finland.
122. Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH Johanniter Krankenhaus, Bonn 53177, Germany.
123. Translational Cancer Research Area University of Eastern Finland, Kuopio 70210, Finland.
124. Institute of Clinical Medicine, Pathology and Forensic Medicine University of Eastern Finland, Kuopio 70210, Finland.
125. Imaging Center, Department of Clinical Pathology Kuopio University Hospital, Kuopio 70210, Finland.
126. Department of Health Research and Policy - Epidemiology Stanford University School of Medicine, Stanford, CA 94305, USA.
127. VIB Center for Cancer Biology, Leuven 3001, Belgium.
128. Laboratory for Translational Genetics, Department of Human Genetics University of Leuven, Leuven 3000, Belgium.
129. Epidemiology Program University of Hawaii Cancer Center, Honolulu, HI 96813, USA.
130. High Risk and Cancer Prevention Group Vall díHebron Institute of Oncology, Barcelona 08035, Spain.
131. Department of Clinical Science and Education, Södersjukhuset Karolinska Institutet, Stockholm 118 83, Sweden.
132. Department of Gynecology and Obstetrics University of Munich, Campus Grosshadern, Munich 81377, Germany.
133. NRG Oncology, Statistics and Data Management Center Roswell Park Cancer Institute, Buffalo, NY 14263, USA.
134. Center for Genomic Medicine Rigshospitalet, Copenhagen University Hospital, Copenhagen DK-2100, Denmark.
135. Epidemiology Branch National Institute of Environmental Health Sciences, NIH Research Triangle Park, NC 27709, USA.
136. Center for Clinical Cancer Genetics The University of Chicago, Chicago, IL 60637, USA.
137. Centro de Investigación en Red de Enfermedades Raras (CIBERER), Madrid 28029, Spain.
138. Department of Molecular Medicine University La Sapienza, Rome 00161, Italy.
139. Molecular Diagnostics Aalborg University Hospital, Aalborg 9000, Denmark.
140. Clinical Cancer Research Center Aalborg University Hospital, Aalborg 9000, Denmark.
141. Department of Clinical Medicine Aalborg University, Aalborg 9000, Denmark.
142. Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20007, USA.
143. Genome Diagnostics Program IFOM - the FIRC (Italian Foundation for Cancer Research) Institute of Molecular Oncology, Milan 20139, Italy.
144. Translational Research Laboratory IDIBELL (Bellvitge Biomedical Research Institute), Catalan Institute of Oncology, CIBERONC, Barcelona 08908, Spain.
145. Leuven Multidisciplinary Breast Center, Department of Oncology Leuven Cancer Institute, University Hospitals Leuven, Leuven 3000, Belgium.
146. Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan 20133, Italy.
147. Clinical Genetics Karolinska Institutet, Stockholm 171 76, Sweden.

148. Department of Basic Sciences Shaukat Khanum Memorial Cancer Hospital and Research Centre (SKMCH & RC), Lahore 54000, Pakistan.
149. Clalit National Cancer Control Center Carmel Medical Center and Technion Faculty of Medicine, Haifa 35254, Israel.
150. Clinical Genetics Service, Department of Medicine Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA.
151. Medical Oncology Department Hospital Universitario Puerta de Hierro, Madrid 28222, Spain.
152. Department of Oncology University Hospital of Larissa, Larissa 411 10, Greece.
153. Department of Epidemiology, Mailman School of Public Health Columbia University, New York, NY 10032, USA.
154. Cancer Genetics and Prevention Program University of California San Francisco, San Francisco, CA 94143-1714, USA.
155. Division of Psychosocial Research and Epidemiology The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam 1066 CX, The Netherlands,
156. Institute of Human Genetics Hannover Medical School Hannover: Germany, 30625.
157. Department of Internal Medicine, Division of Medical Oncology University of Kansas Medical Center, Westwood, KS 66205, USA.
158. Genomics Center Centre Hospitalier Universitaire de Québec - Université Laval Research Center, Québec City, QC G1V 4G2, Canada.
159. Department of Clinical Pathology The University of Melbourne, Melbourne, Victoria 3010, Australia.
160. Cancer Epidemiology Division Cancer Council Victoria, Melbourne, Victoria 3004, Australia.
161. Population Oncology BC Cancer, Vancouver, BC V5Z 1G1, Canada.
162. School of Population and Public Health University of British Columbia, Vancouver, BC V6T 1Z4, Canada.
163. Clinical Genetics Research Lab, Department of Cancer Biology and Genetics Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA.
164. The Curtin UWA Centre for Genetic Origins of Health and Disease Curtin University and University of Western Australia, Perth, Western Australia 6000, Australia.
165. Service de Génétique Institut Curie, Paris 75005, France.
166. Department of Tumour Biology INSERM U830, Paris 75005, France.
167. Université Paris Descartes, Paris 75006, France.
168. Division of Breast Cancer Research Institute of Cancer Research, London, UK.
169. Epigenetic and Stem Cell Biology Laboratory National Institute of Environmental Health Sciences, NIH Research Triangle Park, NC 27709, USA.
170. Hereditary Cancer Program ONCOBELL-IDIBELL-IDIBGI-IGTP, Catalan Institute of Oncology, CIBERONC, Barcelona, Spain.
171. Department of Medicine Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA.
172. Program in Cancer Genetics, Departments of Human Genetics and Oncology McGill University, Montréal, QC H4A 3J1, Canada.
173. Department of Medical Genetics University of Cambridge, vol. Box 134, Level 6 Addenbrookeís Treatment Centre, Addenbrookeís Hosptital, Cambridge CB2 0QQ, UK.
174. Department of Cancer Biology and Genetics The Ohio State University, Columbus, OH 43210, USA.
175. Institute of Human Genetics Pontificia Universidad Javeriana, Bogota, Colombia.

176. Department of medicine University Of Melbourne, Melbourne, Victoria 3002, Australia.
177. Department of Medical Oncology Beth Israel Deaconess Medical Center, Boston, MA 02215, USA.
178. Department of Health Science Research, Division of Epidemiology Mayo Clinic, Rochester, MN 55905, USA.
179. Fundación Pœblica galega Medicina Xenómica-SERGAS, Grupo de Medicina Xenómica-USC CIBERER, IDIS, Santiago de Compostela, Spain.
180. Biostatistics and Computational Biology Branch National Institute of Environmental Health Sciences, NIH Research Triangle Park, NC 27709, USA.
181. Clinical Cancer Genomics City of Hope, Duarte, CA 91010, USA.
182. Department of Surgical Sciences Uppsala University, Uppsala 751 05, Sweden.
183. Department of Oncology Mayo Clinic, Rochester, MN 55905: USA.
184. Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center Vanderbilt University School of Medicine, Nashville, TN 37232, USA.
185. Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA.
186. Department of Preventive Medicine Seoul National University College of Medicine, Seoul 03080, Korea.
187. Department of Biomedical Sciences Seoul National University Graduate School, Seoul 03080, Korea.
188. Cancer Research Institute Seoul National University, Seoul 03080, Korea.
189. Department of Clinical Genetics Odense University Hospital, Odence C 5000, Denmark.
190. Department of Laboratory Medicine and Pathology Mayo Clinic, Rochester, MN 55905, USA.

[$]Joint senior authors

Correspondence to:

Antonis C. Antoniou, PhD, Strangeways Research Laboratory, Department of Public Health and Primary Care, University of Cambridge, Worts Causeway, Cambridge CB1 8RN, U.K. (email: aca20@medschl.cam.ac.uk; telephone: +44 1223 748630)

or

Nadine Andrieu, PhD, Cancer Genetic Epidemiology Team, INSERM Unit 900, Institut Curie, 26 rue d'Ulm, 75005 Paris, France (e-mail: nadine.andrieu@curie.fr; telephone: 0033 (0) 172 38 93 83)

Abstract (words counting: 149, max=150):

Breast cancer (BC) risk for *BRCA1* and *BRCA2* mutation carriers varies by genetic and familial factors. About 50 common variants have been shown to modify BC risk for mutation carriers. All but three, were identified in general population studies. Other mutation carrier-specific susceptibility variants may exist, but studies of mutation carriers have so far been underpowered.

We conducted a novel case-only genome-wide association study comparing genotype frequencies between 60,212 general population BC cases and 13,007 cases with *BRCA1* or *BRCA2* mutations.

We identified four novel variants associated with BC for *BRCA1* and four for *BRCA2* mutation carriers, $P<10^{-8}$, at 8 loci, which are not associated with risk in the general population. The SNPs include rs60882887 in 11p11.2 where *MADD*, *SP11* and *EIF1*, genes previously implicated in BC biology, are predicted as potential targets. These findings will contribute towards customising BC polygenic risk scores for *BRCA1* and *BRCA2* mutation carriers.

Introduction (words counting: 4,060, max in nature communication=5,000 (material & methods paragraphs excluded))

Breast cancer (BC) is the most common cancer in women worldwide[1] and BC family history is one of the most important risk factors for the disease. Women with a history of BC in a first-degree relative are about two times more likely to develop BC than women without a family history[2]. Around 15-20% of the familial risk of BC can be explained by rare mutations in the *BRCA1* or *BRCA2* genes[3]. A recent prospective cohort study estimated the cumulative risk of BC by 80 years to be 72% for *BRCA1* mutation carriers and 69% for *BRCA2* mutation carriers[4]. This study also demonstrated that BC risk for mutation carriers varies by family history of BC in first and second degree relatives, suggesting the existence of other genetic factors that modify BC risks[4].

A total of 179 common BC susceptibility single nucleotide polymorphisms (SNPs) or small insertions or deletions (INDELs) have been identified through genome-wide association studies (GWAS) in the general population[1,5–35]. Although risk alleles at individual SNPs (hereafter used as generic term to refer to common variants, which also includes the small INDELs) are associated with modest increases in BC risk, it has been shown that they combine multiplicatively on risk, resulting in substantial levels of BC risk stratification in the population[36–38]. Similarly, more than 50 of the common genetic BC susceptibility variants have also been shown to be associated with BC for *BRCA1* and *BRCA2* mutation carriers[5,6,15,18,20,39–48] and their joint effects, summarised as polygenic risk scores (PRS), result in large differences in the absolute risks of developing BC for mutation carriers at the extremes of the PRS distribution[49]. BC GWAS for *BRCA1* and *BRCA2* mutation carriers have been carried out through the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA)[50]. However, despite the large number of *BRCA1 and BRCA2* mutation carriers included, the power to detect genetic modifiers of risk remains limited in comparison to that available in the general population[7]. To date, no variants specifically associated with BC risk for *BRCA1* and *BRCA2* carriers have been identified.

Here, we applied a novel strategy using a case-only GWAS design[51,52], in which SNP genotype frequencies in 7,257 *BRCA1* and 5,097 *BRCA2* mutation carrier BC cases were compared to those in 60,212 BC cases from the Breast Cancer Association Consortium (BCAC), unselected for mutation status. We aimed (1) to identify novel SNPs that modify BC risk for *BRCA1* or *BRCA2* mutation carriers but are not associated with risk in the general population and (2) for the known 179 BC susceptibility SNPs, assess whether there is

evidence of an interaction between the SNPs and *BRCA1* or *BRCA2* mutations and therefore evaluate whether the SNP effect size estimates applicable to mutation carriers are different.

Results

Sample characteristics

A total of 60,212 BCAC cases and 7,257 *BRCA1* mutation carrier cases were available for the BRCA1 case-only analyses and 57,725 BCAC cases and 5,097 *BRCA2* mutation carrier cases were available for the BRCA2 case-only analyses (see Figure 1). A total of 45,881 BCAC controls and 5,750 unaffected *BRCA1* mutation carriers were available for the BRCA1 control-only analyses and 43,549 BCAC controls and 4,456 unaffected *BRCA2* mutation carriers for the BRCA2 control-only analyses (see Figure 2). Only women of European ancestry were included with 60.9% samples from European countries, 31.1% from the USA, 6.1% from Australia and 1.7% from Israel (Supplemental Tables 1-4). The mean age at BC diagnosis for mutation carrier cases in CIMBA was 42.5 years (40.9 for *BRCA1* mutation carriers; 44.1 for *BRCA2* mutation carriers) and 58.4 years for cases in BCAC.

The analytical process for assessing interactions with known BC susceptibility SNP is summarised in Figure 3 and for the detection of novel modifiers in Figure 4.

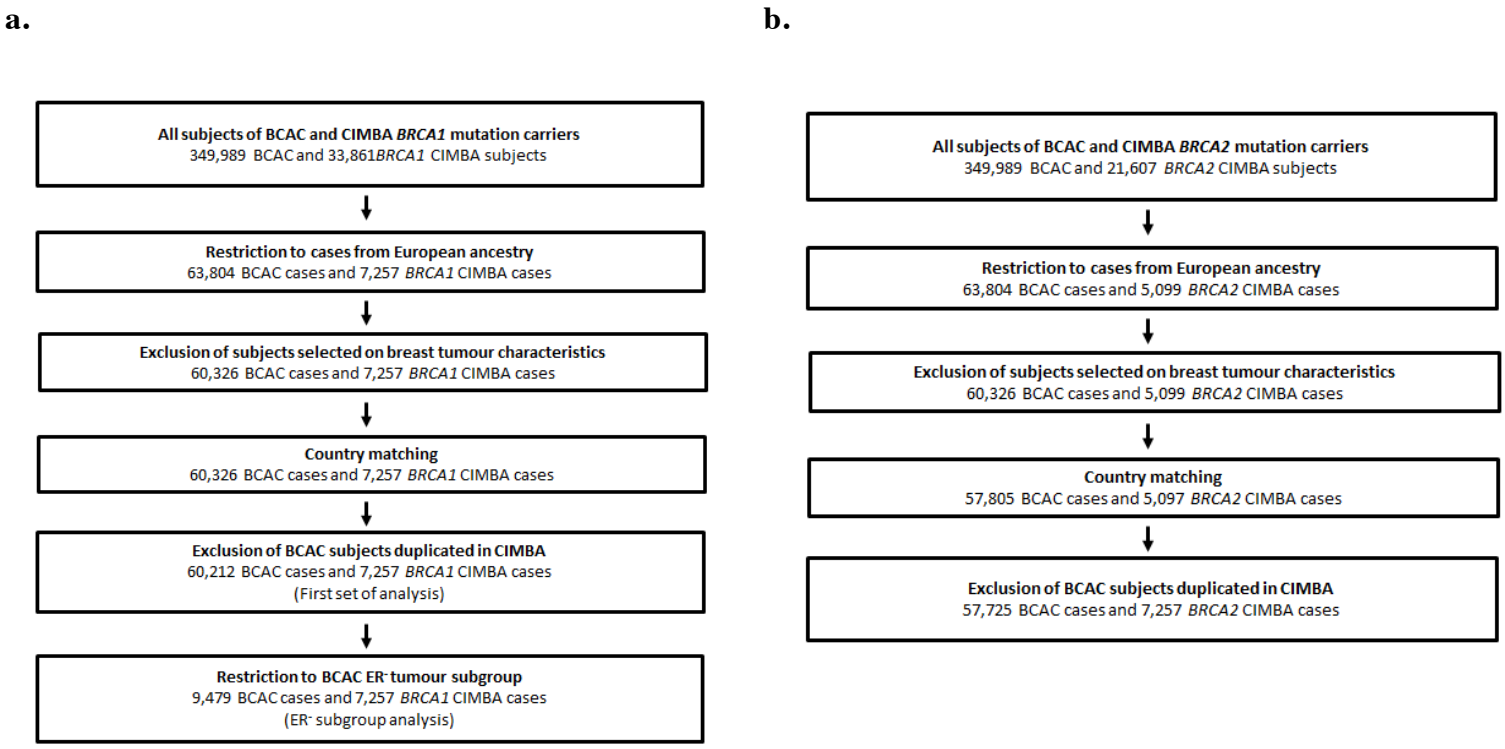


Figure 26 - Flowchart of the sample selection for a) *BRCA1* and b) *BRCA2* case-only analysis. * 4 studies were excluded because they were included in clinical trials based on breast tumour characteristics as HER-2 receptor status (see Supplementary Table 2)

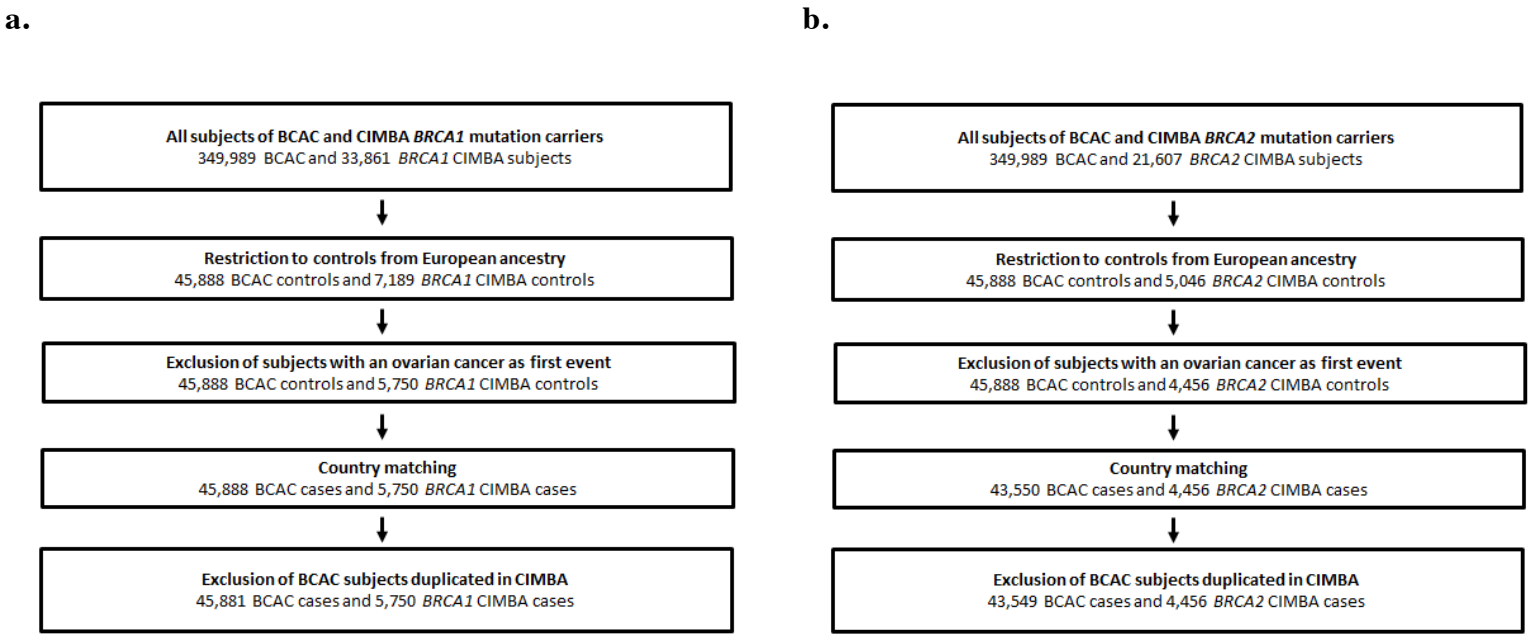


Figure 2 - Flowchart of the sample selection for a) BRCA1 and b) BRCA2 control-only analysis

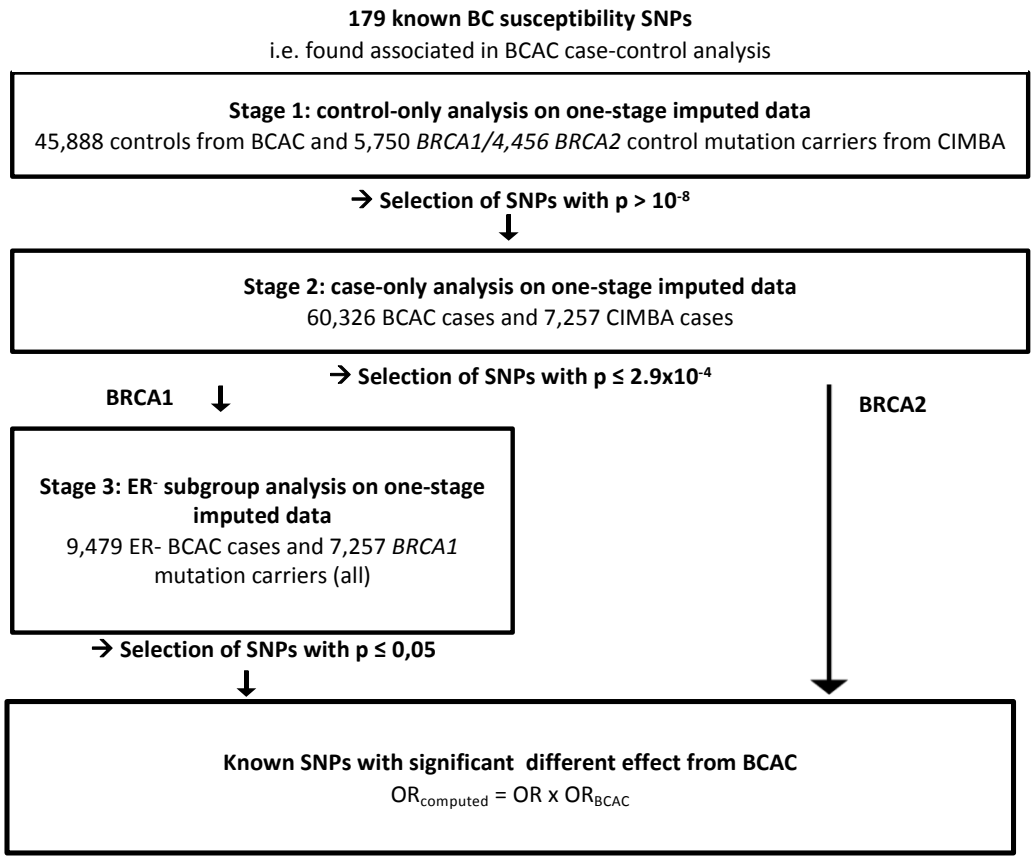


Figure 3 – Flowchart of the strategy followed for analysing the associations for the 179 known BC susceptibility SNPs

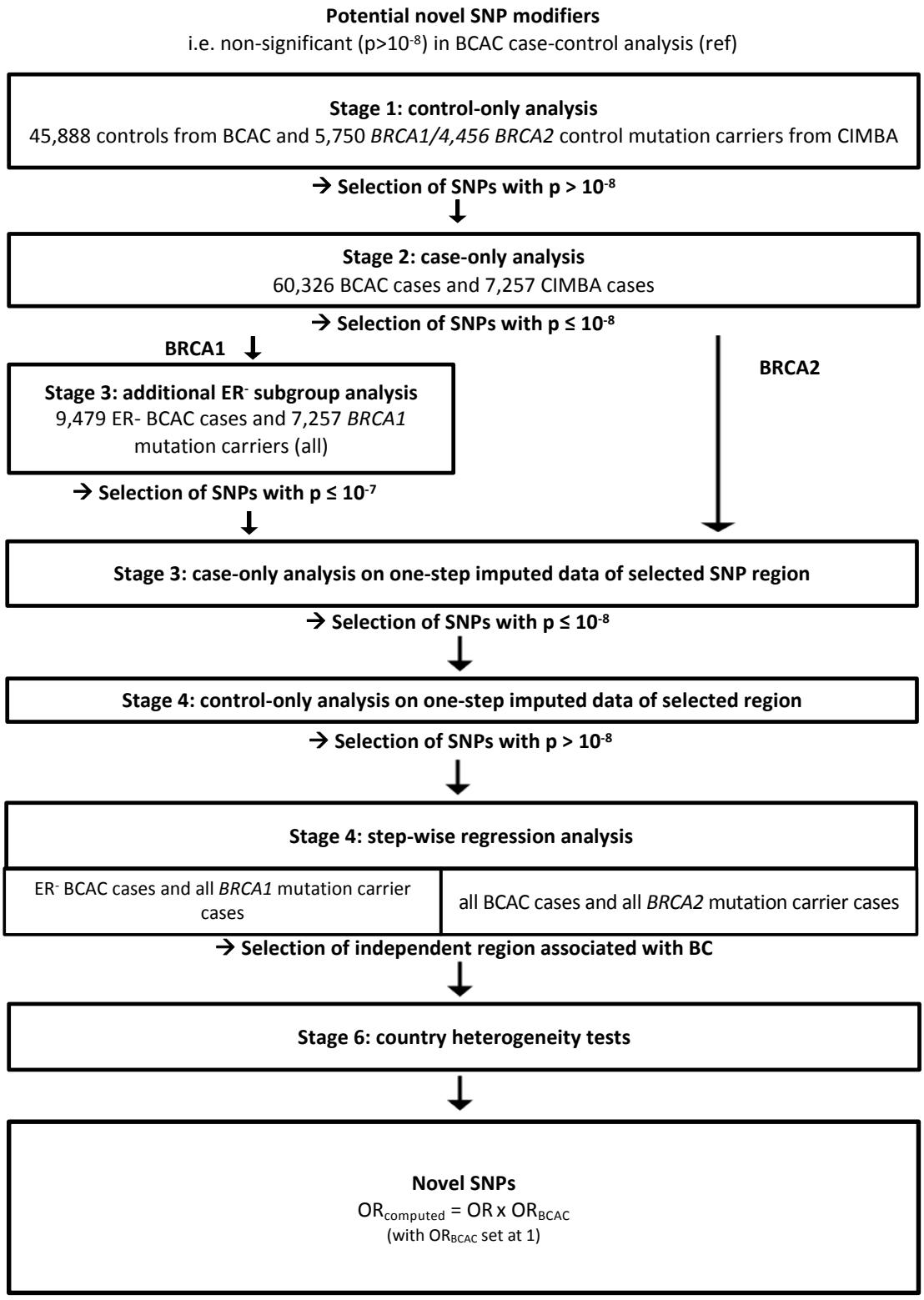


Figure 4 – Flowchart of the strategy followed for identifying potentially novel SNP modifiers.

Assessing independence of SNP frequency with mutation carrier status

Under a case-only study design, it is important to establish independence between the SNPs and *BRCA1* or *BRCA2* mutation carrier status[53]. This was assessed at genome-wide level using a control-only analysis which included controls from BCAC and unaffected mutation carriers from CIMBA with SNP data imputed based on the 1000 genomes project. Genotypes had been imputed separately by each consortium[7,50]. In the analysis of *BRCA1* mutation carriers, 2,164 SNPs were excluded because they were located in or within 500kb of *BRCA1*. 2,070 SNPs were excluded from further analyses because they showed associations at $p<10^{-8}$ with *BRCA1* mutation carrier status in the control-only analysis (2,012 SNPs located on chromosome 17 and 58 on other chromosomes). In the analysis of *BRCA2* mutation carriers, 2,947 SNPs were excluded because they were located in or within 500kb of *BRCA2*. A further 626 SNPs were excluded from further analyses because they were found to be associated with *BRCA2* mutation carrier status in the control-only analysis (566 SNPs on chromosome 13, and 60 on other chromosomes). A total of 9,068,301 SNPs remained for the *BRCA1* case-only association analysis and 9,043,830 SNPs for the *BRCA2* case-only analysis.

Assessing interactions between known BC susceptibility SNPs and *BRCA1* or *BRCA2* mutation carrier status

Based on published data, 179 SNPs were considered as established BC susceptibility SNPs (Figure 3); 158 SNPs were associated with overall BC risk[35] and 21 additional SNPs were found to be associated through studies in ER-negative breast cancer[48] (see Supplemental Table 11 in Milne *et al*. 2017). One of the 158 SNPs, rs11571833 located within *BRCA2* was excluded from the BRCA2 analysis. The detailed results are shown in Supplemental Tables 5, 6 and 7.

BRCA1 mutation carriers

Previous studies have demonstrated heterogeneity in the associations of the SNPs with ER-positive and ER-negative breast cancer[35]. Since *BRCA1* mutation carriers develop primarily ER-negative BC, to comprehensively assess the evidence of interaction with *BRCA1* mutation status, we followed a two-step process; we first assessed the associations using all BC cases from BCAC and then we restricted the comparison to BCAC ER-negative BC cases. Of the 158 SNPs[35], 59 were associated with *BRCA1* mutation carrier status when compared to all BC cases ($P<0.05$, Supplemental Table 5). However, after adjusting for multiple testing, only four of these SNPs were associated ($P<2.7\times10^{-4}$) and also showed evidence of association ($P<0.05$)

when compared with ER-negative BC cases (Table 1). Two additional SNPs on chromosome 1 and 6 (chr1_10566215_A_G and rs17529111) were associated at $P<2.7x10^{-4}$ with *BRCA1* mutation status only when compared with ER-negative BCAC cases. The OR estimates for association with *BRCA1* mutation status for these six SNPs were similar under both case-only analyses (all BC and ER-negative BC cases analyses) and varied from 0.85 to 1.07, suggesting that the magnitude of their associations with BC risk for *BRCA1* mutation carriers differs from that observed in the general population. For the other 152 SNPs, there was no evidence of association with *BRCA1* mutation status when compared against the ER-negative BC cases from BCAC (Supplemental Table 5), suggesting that the OR estimated using case-control data from BCAC are also applicable to *BRCA1* mutation carriers.

Among the 21 ER-negative SNPs reported in Milne *et al.*[48], only one (rs66823261) demonstrated significant evidence of association in the ER-negative case only analysis (OR=0.88, $p<2.7\ 10^{-4}$) (Table 1 and Supplemental Table 6). For the 20 other showing no association, the ORs estimated in Milne *et al.*[48] would be applicable to *BRCA1* mutation carriers.

To estimate the association of the seven significant SNPs with BC for *BRCA1* mutation carriers ($OR_{computed}$), the OR estimated using case-control data from BCAC (OR_{BCAC}) was multiplied by the OR estimated using the case-only analysis (OR). For three SNPs, rs17426269, chr10_80841148_C_T and rs17529111, the magnitude of the association with BC for *BRCA1* carriers was greater than that in the general population (OR_{BCAC}) and for two of them, the $OR_{computed}$ is in the opposite direction than the OR_{BCAC} (Table 1). For the four other SNPs (rs13281615, chr16_52599188_C_T, chr1_10566215_A_G and rs66823261), the estimated interaction OR resulted in the OR for associations with *BRCA1* BC risk being closer to 1 (Table 1).

Among the remaining 172 SNPs (152+20) that showed no associations with *BRCA1* mutation status, the estimated $OR_{computed}$ was smaller (i.e. closer to 1) than those estimated in the general population (OR_{BCAC}) for 146 (85%) (Supplemental Table 5 and 6). Based on the analysis of ER^- tumors, the proportion of SNPs for which $OR_{computed}$ was closer to 1 than OR_{BCAC} was 59% (Supplemental Table 5 and 6).

BRCA2 mutation carriers

Among the 157 SNPs known to be associated with BC risk in the general population, 43 were associated with *BRCA2* mutation carrier status at P<0.05 in the case-only analysis that included all cases of BC of BCAC (Supplemental Table 7). However, only three SNPs

(rs62355902, rs10759243 and chr22_40876234_C_T) showed associations after adjusting for multiple testing ($P<2.7x10^{-4}$) with OR estimates in the range of 0.88 to 0.89 (Table 2).

For these three SNPs, the observed interaction resulted in the magnitude of association with BC risk for *BRCA2* mutation carriers ($OR_{computed}$) to be closer to 1 (Table 2).

For the 154 SNPs that showed no significant associations with *BRCA2* mutation status, 79% had ORs of BC for *BRCA2* mutation carriers ($OR_{computed}$) that were closer to 1 when compared to the ORs estimated using data in the general population (OR_{BCAC}) (Supplemental Table 7).

Location	SNP name	Chr[1]	Position[2]	Nearest gene	Estimated effect allele	Referent Allele	Frequency[3]	r2	OR[4]	P[5]	OR_{ER-}[6]	P_{ER-}[7]	OR_{BCAC}[8]	P_{BCAC}[9]	$OR_{computed}$[10]	Variation in risk[11]
SNPs associated with all BC subtypes in the general population																
1p22.3	rs17426269	1	88156923	-	A	G	0.16	1	0.90	$2.70e^{-04}$	0.92	$4.22e^{-02}$	1.05	$1.70e^{-04}$	0.95	IOD
8q24.21	rs13281615	8	128355618	-	G	A	0.43	1	0.91	$1.20e^{-05}$	0.94	$4.14e^{-02}$	1.11	$5.00e^{-28}$	1.01	TT1
10q22.3	chr10_80841148_C_T	10	80841148	*ZMZ1*	T	C	0.40	1	0.91	$2.20e^{-06}$	0.91	$1.01e^{-03}$	0.93	$1.10e^{-14}$	0.84	ISD
16q12.1	chr16_52599188_C_T	16	52599188	*TOX3*	T	C	0.29	1	0.85	1.80E-13	0.91	$2.80e^{-03}$	1.23	$7.00e^{-88}$	1.04	TT1
SNPs associated with ER-negative BC in the general population																
1p36.22	chr1_10566215_A_G	1	10566215	*PEX14*	G	A	0.32	1	1.07	$1.30e^{-03}$	1.12	$1.10e^{-04}$	0.94	$1.80e^{-09}$	1.05	TT1
6q14.1	rs17529111	6	82128386	-	C	T	0.23	0.96	0.92	$7.70e^{-04}$	0.86	$1.96e^{-05}$	1.02	$4.20e^{-02}$	0.88	IOD
8p23.3	rs66823261	8	170692	*RPL23AP53*	C	T	0.23	0.92	-	-	0.88	$2.37e^{-04}$	1,09*	$5,09e^{-09}$*	0.96	TT1

Table 1 – SNPs demonstrating associations in the BRCA1 case-only analysis: considering SNPs with known BC (Michailidou *et al.* 2017) and ER-negative -specific BC (Milne *et al.* 2017) associations in the general population.

1 Chromosome
2 Build 37 position
3 Frequency of the allele for which effect is estimated in BCAC cases (OncoArray dataset)
4 Per allele odds ratio estimated in the case-only analysis
5 p-value in the case-only analysis (after allowing for multiple testing, p=2.7x10-4)
6 Per allele odds-ratio estimated in the case-only ER-negative subgroup analysis
7 p-value in the case-only ER-negative subgroup analysis.
8 Per allele odds-ratio estimated in BCAC (Michailidou *et al.* 2017), except for * (Milne *et al.* 2017)
9 p-value in BCAC (Michailidou *et al.* 2017), except for * (Milne *et al.* 2017). For SNPs with PBCAC>10-8, significance was attained in merging data of Oncoarray, iCOGS and 11 different breast cancer GWAS in Michailidou *et al.* 2017 or Milne *et al.* 2017
10 Per allele computed odds-ratio (OR x ORBCAC)
11 compared with Michailidou et al's OR estimates: TT1= Tends To 1, ISD= Increase in Same Direction, IOD= Increase in Opposite Direction

Location	SNP name	Chr[1]	Position[2]	Nearest gene	Estimated effect allele	Referent Allele	Frequency[3]	r2	OR[4]	P[5]	OR_{BCAC}[6]	P_{BCAC}[7]	$OR_{computed}$[8]	Variation in risk[9]
5q11.2	rs62355902	5	56053723	*MAP3K1*	T	A	0.18	0.98	0.89	$1.10e^{-04}$	1.18	$8.50e^{-42}$	1.05	TT1
9q31.2	rs10759243	9	110306115	*RP11-438P9.2*	A	C	0.31	1	0.89	$4.60e^{-06}$	1.06	$4.20e^{-10}$	0.95	TT1
22q13.1	chr22_40876234_C_T	22	40876234	*MKL1*	C	T	0.11	1	0.88	$2.8e^{-04}$	1.12	$5.70e^{-16}$	0.98	TT1

Table 2 – SNPs demonstrating associations in the BRCA2 case-only analysis: considering SNPs with known BC (Michailidou *et al.* 2017) associations in the general population.

1 Chromosome
2 Build 37 position
3 Frequency of the allele for which effect is estimated in BCAC cases (OncoArray dataset)
4 Per allele odds ratio estimated in the case-only analysis
5 p-value in the case-only analysis (after allowing for multiple testing, p*=2.7x10-4)
6 Per allele odds-ratio estimated in BCAC (Michailidou *et al.* 2017)
7 p-value in BCAC (Michailidou *et al.* 2017). For SNPs with PBCAC>10-8, significance was attained in merging data of Oncoarray, iCOGS and 11 different breast cancer GWAS in Michailidou *et al.* 2017
8 Per allele computed odds-ratio (OR x ORBCAC)
9 compared with Michailidou et al's OR estimates: TT1= Tends To 1, ISD= Increase in Same Direction, IOD= Increase in Opposite Direction

Novel SNP modifiers

To identify novel SNPs that modify BC risks for *BRCA1* and *BRCA2* mutation carriers, we investigated the associations in the case-only design for SNPs that were not established as BC susceptibility variants for the general population (Figure 4).

BRCA1 mutation carriers

A total of 924 SNPs showed associations at $P<10^{-8}$ in all BC case-only analysis. To ensure that none of these associations are driven by differences in the distribution of ER-positive and ER-negative tumours in BCAC cases, an intermediate step was applied, in which we re-analysed the associations after restricting the BCAC data to only ER-negative cases. 220 of these SNPs remained significant at $P<10^{-7}$ located in 11 distinct genomic regions. SNPs were considered to belong in the same region if they were located within 500kb of each other.

To ensure that none of these associations was driven by differences in the genotype imputation in the BCAC and CIMBA data (which had been carried out separately), all the SNPs in these 11 distinct genomic regions were re-imputed in the BCAC and CIMBA samples jointly and the associations for all SNPs in the regions were re-assessed in the control-only and case-only analyses. After the exclusion of 614 SNPs (613 on chromosome 17) that showed associations in the control-only analysis, 71 SNPs in two regions remained significant at $P<10^{-8}$ (Supplementary Table 8) in the case-only analyses including all BCAC cases. None of these SNPs had been previously reported in GWAS in the general population (p-values of association ranged from 0.51 to 5.9×10^{-5} with effect sizes in the range 0.96 - 1.04 in BCAC case-control analyses)[35,48]. A forward step-wise regression analysis within each of these two regions (restricted to the SNPs exhibiting associations at $p<10^{-8}$) starting with the most significant SNP and adding sequentially the other SNPs, identified a set of four conditionally independent SNPs ("top SNPs") (Table 3): all SNPs were imputed, with $r^2>0.5$, and had minor allele frequency (MAF) > 10%. Three of the "top SNPs" are located in 17q21.2. rs58117746 is an insertion of 16 bp within an exon of *KRTAP4-5* leading to a frameshift of the amino acid sequence. rs5820435 and rs11079012 are both intronic and located in *LEPREL4* (also named *P3H4*) and *JUP*, respectively, while rs80221606 is intronic and located in 11p11.2, within *CELF1*. The OR estimates of these four "top SNPs" ranged from 0.78 to 1.22. All showed evidence of heterogeneity in the OR by country ($P<0.05$) (Table 3); however, in a leave-one-out analysis, in which each country was left out in turn, the overall associations remained similar (Supplemental Figures 1 and 2) suggesting that no individual country had a big impact on the observed associations.

BRCA2 mutation carriers

The *BRCA2* case-only analysis identified 273 SNPs, located across 22 regions, with evidence of association at $P<10^{-8}$. After the joint re-imputation of the SNPs in these 22 regions, only 102 SNPs located in four regions (2p14, 13q13.1 and 13q13.2) remained associated at $P<10^{-8}$ (Supplemental Table 9). The step-wise regression analysis suggested that associations in each of the four regions were driven by a single variant ("top SNPs") (Table 4). All four variants were imputed ($r^2>0.5$) and had MAF higher than 5%. At 2p14, rs12470785 ($r^2=0.98$) is within an intron of *ETAA1*. At 13q13.1, rs79183898 ($r^2=0.84$) is located between *B3GALTL* and *RXFP2* and rs736596 ($r^2=0.66$) is within an intron of *STARD13*. At 13q13.2, rs4943263 ($r^2=0.99$) is located between *RP11-266E6.3* and *RP11-307O13.1*. None of these SNPs had been previously reported to be associated with BC risk in BCAC studies in the general population (*p-values* from 0.01 to 0.90 in BCAC case-control analyses)[35,48]. The OR estimates of these four SNPs ranged from 0.85 to 1.37. All showed evidence of heterogeneity in the OR by country at p=0.05 (Table 4). In the leave-one-country-out sensitivity analysis the two intergenic SNPs, rs79183898 and rs736596 were no longer significant at $P<10^{-4}$ when studies from the USA were excluded from the analysis and the OR estimates were substantially attenuated (Supplemental Figures 3 and 4).

Location	SNP name[1]	Chr[2]	Position[3]	Nearest gene	Localisation	Estimated effect allele	Referent Allele	r^2 [4]	Frequency[5]	OR[6]	P[7]	OR_{ER}[8]	P_{ER-}[9]	HR_{CIMBA}[10]	P_{CIMBA}[11]	OR_{BCAC}[12]	P_{BCAC}[13]	P_{het}[14]	Target gene[15]
11p11.2	rs80221606	11	47560211	*CELF1*	intronic	AT	A	0.76	0.10	0.78	$1.12e^{-10}$	0.76	$6.36e^{-07}$	0.98	$7.60e^{-01}$	1.04	0.01	$1.39e^{-03}$	Level 2
17q21.2	rs58117746	17	39305775	*KRTAP4-5*	pepshift	TGGCAGCAG CTGGGGC	T	0.60	0.39	1.18	$4.33e^{-10}$	1.15	$7.71e^{-05}$	1.05	$2.20e^{-02}$	1.02	0.26	$4.60e^{-04}$	-
17q21.2	rs5820435	17	39961558	*LEPREL4*	intronic	A	C	0.51	0.45	0.82	$9.55e^{-12}$	0.85	$7.71e^{-05}$	1.01	$9.00e^{-01}$	1.02	0.07	$1.06e^{-08}$	-
17q21.2	rs11079012	17	39912880	*JUP*	intronic	G	C	0.66	0.31	1.17	$7.06e^{-09}$	1.18	$2.35e^{-05}$	0.98	$3.10e^{-01}$	1.01	0.51	$1.15e^{-07}$	Level 2

Table 3 - List of " potential novel SNP modifiers" associated in the case-only analysis for *BRCA1* mutation carriers.

1 The most significant SNP of each region
2 Chromosome
3 Build 37 position
4 Imputation accuracy
5 Frequency of the allele for which effect is estimated in BCAC cases (OncoArray dataset)
6 Per allele odds ratio estimated in the case-only analysis
7 p-value in the case-only analysis (after allowing for multiple testing, p=2.7x10-4)
8 Per allele odds-ratio estimated in the case-only ER-negative subgroup analysis
9 p-value in the case-only ER-negative subgroup analysis
10 Per allele hazard ratio estimated in CIMBA cohort analysis
11 Pvalue found in CIMBA cohort analysis
12 Per allele odds-ratio estimated in BCAC (Michailidou *et al.* 2017)
13 p-value in BCAC (Michailidou *et al.* 2017). For SNPs with PBCAC>10-8, significance was attained in merging data of Oncoarray, iCOGS and 11 different breast cancer GWAS in Michailidou *et al.* 2017
14 Pvalue of the heterogeneity test by country
15 INQUISIT score level: 1 = most functional evidence supporting potential link between CCVs and target gene

Location	SNP name[1]	Chr[2]	Position[3]	Nearest gene	Localisation	Estimated effect allele	Referent Allele	r^2 [4]	Frequency [5]	OR[6]	P[7]	HR_{CIMBA} [8]	P_{CIMBA} [9]	OR_{BCAC} [10]	P_{BCAC} [11]	P_{HET} [12]	Target gene[13]
2p14	rs12470785	2	67634003	*ETAA1*	intron	G	A	0.98	0.30	0.84	$2.83e^{-11}$	0.89	$1.69e^{-05}$	0.98	0.03	$2.18e^{-07}$	Level 2
13q13.1	rs79183898	13	32221794	*B3GALTL - RXFP2*	intergenic	A	T	0.84	0.07	1.33	$2.88e^{-10}$	1.04	$3.55e^{-01}$	1.01	0.54	$1.12e^{-08}$	-
13q13.1	rs736596	13	33776506	*STARD13*	intron	T	G	0.66	0.09	1.37	$3.44e^{-12}$	0.94	$2.54e^{-01}$	0.98	0.45	$4.99e^{-11}$	Level 1
13q13.2	rs4943263	13	35376357	*RP11-266E6.3 - RP11-307O13.1*	intergenic	T	C	0.99	0.27	1.17	$8.33e^{-11}$	1.01	$9.83e^{-01}$	1.00	0.47	$6.94e^{-03}$	Level 2

Table 4 - List of " potential novel SNP modifiers" associated in the case-only analysis for *BRCA2* mutation carriers.

1 The most significant SNP of each region
2 Chromosome
3 Build 37 position
4 Imputation accuracy
5 Frequency of the allele for which effect is estimated in BCAC cases (OncoArray dataset)
6 Per allele odds ratio estimated in the case-only analysis
7 p-value in the case-only analysis
8 Per allele hazard ratio estimated in CIMBA cohort analysis
9 Pvalue found in CIMBA cohort analysis
10 Per allele odds-ratio estimated in BCAC (Michailidou *et al.* 2017)
11 p-value in BCAC (Michailidou *et al.* 2017). For SNPs with PBCAC>10-8, significance was attained in merging data of Oncoarray, iCOGS and 11 different breast cancer GWAS in Michailidou *et al.* 2017
12 Pvalue of the heterogeneity test by country
13 INQUISIT score level: 1 = most functional evidence supporting potential link between CCVs and target gene

***In silico* analyses on credible causal variants (CCV)**

In order to determine the likely target genes of each region of the eight novel mutation carriers' BC risk-associated SNPs, we first defined credible set of SNPs candidates to be causal (credible causal variants [CCVs]) (see methods).

Sets of CCVs were sought for the two regions found in the previous step-wise analyses to be associated with risk in *BRCA1* mutation carriers. In the region located at 11p11.2, only one signal composed of 74 CCVs was found (Table 5). All of these 74 CCVs were imputed with a r^2 higher than 0.92 (Supplemental Table 10). In the region located in 17q21.2, we found nine signals which contained from one to 13 CCVs (Table 5). Two of these CCVs were genotyped and the others had an r^2 between 0.50 and 0.98 (Supplemental Table 10).
We used INQUISIT[35,54] to prioritize target genes by intersecting each CCV with publicly available annotation data from breast cells and tissues (see Methods). The results for *BRCA1* mutation carriers are summarized in Supplemental Table 11. For *BRCA1* mutation carriers, we predicted 38 unique target genes for six of the 10 independent signals. Seven target genes in two regions (*MTCH2, MADD*, *PSMC3, RP11-750H9.5, SLC39A13, SPI1* and *EIF1*) were predicted with high confidence (designated "Level 1", scoring range between Level 1 [highest confidence] to Level 3 [lowest confidence]). All seven Level 1 genes were predicted to be distally regulated by CCVs.

Similarly, sets of CCVs were sought from the four regions found in the previous step-wise analyses to be associated with risk in *BRCA2* mutation carriers. A total of 17 signals were found. One signal composed of 78 CCVs was found in the region located at 2p14 (Table 6). One CCV was genotyped and the others were imputed with r^2 between 0.95 and 0.99 (Supplemental Table 12). Twelve signals were found from the two regions previously found in 13q13.1 which contained from one to 46 CCVs. The analysis in the region of rs79183898 in 13q13.1 found three signals out of the 12, which are located in 13q12.3 (with top SNPs: rs71434801, rs77197167, rs114300732). Finally, four signals in the previously identified region located in 13q13.2 containing from three to 40 CCVs were found. Among all CCVs, 11 are genotyped and the imputed ones have a r^2 higher than 0.58 (Table 6 and Supplemental Table 12).
For *BRCA2* mutation carriers, we predicted 24 unique target genes for 10 of the 17 independent signals, including one high confidence target gene, *STARD13* at chr13:33395975-34395975. *STARD13* was also predicted to be targeted by three independent signals. All results are presented in Supplemental Table 13.

Fine mapping region[1]	Signal[2]	#CCV[3]	Location	SNP name[4]	Chr[5]	Position[6]	Nearest gene	Localisation	Estimated effect allele	Referen t Allele	Frequency [7]	r^2 [8]	P[9]	OR[10]	P_{ER}[11]	OR_{ER}[12]	P_{CIMBA}[13]	HR_{CIMBA}[14]
chr11:46773616-47773616	1	74	11p11.2	rs60882887	11	47475675	*RAPSN, CELF1*	intergenic	A	G	0.14	0.95	2.20E-10	0.82	3.20E-06	0.82	7.00E-01	0.99
chr17:39141815-40141815	1	2	17q21.2	rs5820435	17	39961558	*LEPREL4*	intronic	A	C	0.45	0.51	1.10E-11	0.82	2.80E-05	0.85	9.10E-01	1.00
	2	2	17q21.2	rs7222250	17	39938469	*JUP*	intronic	C	T	0.44	0.66	5.50E-14	1.23	3.90E-07	1.20	8.70E-01	1.00
	3	6	17q21.2	rs9901834	17	39926811	*JUP*	intronic	A	G	0.10	0.55	7.20E-10	0.72	3.90E-06	0.72	7.40E-01	1.02
	4	3	17q21.2	rs58117746	17	39305775	*KRTAP4-5*	intronic	TGGCAGCAGCTGGGGC	T	0.39	0.60	5.50E-09	1.17	4.60E-04	1.13	2.20E-02	1.06
	5	13	17q21.2	rs2239711	17	39633317	*KRT35*	intronic	A	G	0.29	0.93	4.90E-11	0.85	2.90E-04	0.88	5.00E-01	0.98
	6	4	17q21.2	rs10708222	17	40137437	*DNAJC7*	intronic	T	TA	0.17	0.60	8.40E-07	1.18	6.10E-04	1.17	2.28E-01	0.95
	7	4	17q21.2	rs41283425	17	39925713	*JUP*	intronic	T	C	0.06	0.54	4.30E-07	0.73	1.30E-05	0.69	4.82E-01	0.95
	8	15	17q21.2	rs56291217	17	39858199	*JUP*	intronic	AT	A	0.44	0.76	6.70E-08	0.88	1.20E-06	0.85	4.06E-01	1.03
	9	1	17q21.2	rs111637825	17	40134782	*DNAJC7*	intronic	A	G	0.06	0.89	3.60E-07	0.74	3.50E-04	0.75	4.47E-01	0.96

Table 5 - List of most significant SNPs in the CCV analysis for *BRCA1* mutation carriers.

1 Significant region in the main analysis used to look for credible causal variants (CCV)
2 Signal number (the first one corresponds to the CCV set without any adjustment and the following are those with adjustment on each most significant SNP of the previous signals)
3 Number of credible causal variants in each signal (SNP with p-value at 2 order of magnitude of the most significant one)
4 The most significant SNP after adjustment on the most significant SNPs of the previous signals (except for these of the signal 1)
5 Chromosome
6 Build 37 position
7 Frequency of the allele for which effect is estimated in BCAC cases (OncoArray dataset)
8 Imputation accuracy
9 P-value in the case-only analysis after adjustment on the most significant SNPs of the previous signals (except for these of the signal 1)
10 Per allele odds ratio estimated in the case-only analysis after adjustment on the most significant SNPs of the previous signals (except for these of the signal 1)
11 P-value in the case-only analysis restricted to ER-negative BCAC cases and after adjustment with the most significant SNP of the previous signals (except for these of the signal 1)
12 Per allele odds ratio estimated in the case-only analysis restricted to ER-negative BCAC cases and after adjustment with the most significant SNP of the previous signals (except for these of the signal 1)
13 P-value found in CIMBA cohort analysis
14 Per allele hazard ratio estimated in CIMBA cohort analysis

Fine mapping region[1]	Signal[2]	#CCV[3]	Location	SNP name[4]	Chr[5]	Position[6]	Nearest gene	Localisation	Estimated effect allele	Referent Allele	Frequency [7]	r^2 [8]	P[9]	OR[10]	P_{CIMBA} [11]	HR_{CIMBA} [12]
chr2:67099466-68099466	1	78	2p14	rs12470785	2	67634003	*ETAA1*	intronic	G	A	0.30	0.98	$4.20e^{-11}$	0.85	$7.70e^{-05}$	0.89
chr13:31015494-32515494	1	8	13q13.1	rs79183898	13	32221794	*B3GALTL, RXFP2*	intergenic	A	T	0.07	0.84	$1.10e^{-10}$	1.33	$3.60e^{-01}$	1.04
	2	23	13q12.3	rs71434801	13	31249461	*USPL1, ALOX5AP*	intergenic	G	C	0.13	0.76	$3.40e^{-08}$	1.22	$8.40e^{-01}$	0.99
	3	35	13q12.3	rs77197167	13	31693513	*WDR95P, HSPH1*	intergenic	C	T	0.09	0.76	$1.80e^{-07}$	1.25	$4.00e^{-01}$	1.04
	4	7	13q12.3	rs114300732	13	31662987	*WDR95P*	intronic	T	C	0.07	0.90	$1.70e^{-08}$	0.67	$8.80e^{-02}$	1.09
	5	12	13q13.1	13:32231513:CAA:C	13	32231513	*B3GALTL, RXFP2*	intergenic	CAA	C	0.25	0.92	$8.40e^{-07}$	0.86	$1.70e^{-02}$	1.08
	6	6	13q13.1	rs1623189	13	32232683	*B3GALTL, RXFP2*	intergenic	G	T	0.26	0.95	$1.30e^{-31}$	2.70	$6.60e^{-01}$	1.01
chr13:33395975-34395975	1	1	13q13.1	rs736596	13	33776506	*STARD13*	intronic	T	G	0.09	0.66	$1.20e^{-12}$	1.37	$2.50e^{-01}$	0.95
	2	1	13q13.1	rs77889880	13	33776161	*STARD13*	intronic	T	A	0.10	0.80	$3.00e^{-21}$	0.51	$1.90e^{-02}$	1.12
	3	1	13q13.1	rs67776313	13	33934343	*RP11-141M1.3*	intronic	A	AT	0.33	0.70	$7.70e^{-12}$	0.81	$4.60e^{-01}$	0.98
	4	42	13q13.1	rs71196514	13	33800572	*STARD13*	intronic	C	CT	0.38	0.67	$1.00e^{-07}$	0.86	$6.20e^{-01}$	1.01
	5	52	13q13.1	rs2555605	13	33833810	*STARD13*	intronic	C	T	0.36	1.00	$4.60e^{-08}$	0.87	$2.00e^{-01}$	1.03
	6	46	13q13.1	rs74796280	13	33700860	*STARD13*	intronic	C	A	0.06	0.96	$4.70e^{-07}$	0.77	$3.10e^{-02}$	0.89
chr13:34793902-35793902	1	18	13q13.2	rs4943263	13	35376357	*RP11-266E6.3, RP11-307O13.1*	intergenic	T	C	0.27	0.99	$6.30e^{-11}$	1.18	$9.80e^{-01}$	1.00
	2	3	13q13.2	rs2202781	13	35292372	*RP11-266E6.3, RP11-307O13.1*	intergenic	G	A	0.24	0.93	$3.10e^{-11}$	1.20	$6.00e^{-01}$	0.98
	3	40	13q13.2	rs55675572	13	35315594	*RP11-266E6.3, RP11-307O13.1*	intergenic	A	T	0.40	0.77	$5.60e^{-08}$	0.86	$7.50e^{-01}$	0.99
	4	21	13q13.2	rs17755120	13	35270340	*RP11-266E6.3, RP11-307O13.1*	intergenic	T	A	0.20	0.98	$6.30e^{-07}$	0.76	$4.80e^{-01}$	0.98

Table 6 - List of most significant SNPs in the CCV analysis for *BRCA2* mutation carriers.

1 Significant region in the main analysis used to look for credible causal variants (CCV)
2 Signal number (the first one correspond to the CCV set without any adjustement and the following are those with adjstement on each most significant SNP of the previous signals)
3 Number of credible causal variants in each signals (SNP with p-vaue at 2 order of magnitude of the most significant one)
4 The most significant SNP after adjustement on the most significant SNPs of the previous signals (except for these of the signal 1)
5 Chromosome
6 Build 37 position
7 Frequency of the allele for which effect is estimated in BCAC cases (OncoArray dataset)
8 Imputation accuracy
9 p-value in the case-only analysis after adjustement on the most significant SNPs of the previous signals (except for these of the signal 1)
10 Per allele odds ratio estimated in the case-only analysis after adjustement on the most significant SNPs of the previous signals (except for these of the signal 1)
11 Pvalue found in CIMBA cohort analysis
12 Per allele hazard ratio estimated in CIMBA cohort analysis

Discussion

To identify novel genetic modifiers of BC risk for *BRCA1* and *BRCA2* mutation carriers and to further clarify the effects of known BC susceptibility SNPs on BC risk for carriers, a novel case-only analysis strategy was used based on GWAS data from unselected BC cases in BCAC and mutation carriers with BC from CIMBA. This strategy provides increased statistical power for detecting new associations and for clarifying the risk associations of known BC susceptibility SNPs in mutation carriers[55].

Of the 179 known BC susceptibility SNPs identified through GWAS in the general population[5–35], only 10 showed evidence of interaction with *BRCA1* or *BRCA2* mutation carrier status after taking the tumour ER-status into account. However, 82% of all 179 showed an OR point estimate closer to 1 than the one estimated in the general population. This suggests that, while most SNPs associated with risk in the general population are also associated with risk in carriers, the average effects size is smaller. However, the effect sizes in the general population may be exaggerated as this dataset contributed to the discovery of most of the loci[49]. For 10 SNPs, an interaction was observed with *BRCA1* or *BRCA2* mutation carrier status, suggesting that these SNPs may have different effect sizes in *BRCA1* or *BRCA2* mutation carriers compared to the general population (seven for *BRCA1* mutation carriers and three for *BRCA2* mutation carriers). The observed interactions specifically suggest that seven SNPs may have no effect on BC risk for mutation carriers, two SNPs lead to an increase in the magnitude of association for mutation carriers compared to the observed BC OR estimates in the general population and one leads to an association which is in the opposite direction to that observed in the general population.

We also identified eight novel conditionally independent common SNPs associated with breast cancer risk (four for *BRCA1* mutation carriers, four for *BRCA2* mutation carriers). These have not been reported in previous association studies[5,6,15,18,20,39–47]. The case-only OR estimates for these SNPs varied from 0.85 to 1.37 for *BRCA2* mutation carriers and from 0.78 to 1.22 for *BRCA1* mutation carriers. For five of these SNPs the estimated ORs from the case-only analysis results were in the same direction as the estimated HRs from previously reported GWAS using cohort analyses restricted in *BRCA1* and *BRCA2* mutation carriers in CIMBA[56]. Two of these five SNPs also demonstrated some evidence of association in mutation carriers ($p=2.2x10^{-2}$ for rs58117746 for *BRCA1* mutation carriers; and $p=7.7x10^{-5}$ for

rs12470785 in *ETAA1* for *BRCA2* mutation carriers; Tables 3 and 4). For the remaining three variants, rs5820435 and rs11079012 at 17q21.2 and rs736596 at 13q13.1, the associations in *BRCA1* or *BRCA2* mutation carriers in the CIMBA data were not consistent with the observed interactions and might be artefactual. One possibility is that the associations with SNPs on 17q and 13q in *BRCA1* and *BRCA2* carriers respectively, reflect confounding due to linkage disequilibrium with specific mutations. Although we excluded variants with evidence of association in the control only analyses, it is possible that residual confounding due to specific mutations was still present.

Seven genes at a locus at 11p11.2 marked by rs60882887, were predicted with high confidence as targets, including *MADD*, *SP11* and *EIF1* which have previously been reported to be associated with BC biology[57–59]. However, no likely target genes were predicted at the 17q21.2 region. The lack of target gene predictions may be due to reliance on breast cell line data which does not represent the *in vivo* tissue of interest or due to the fact that the target transcripts are not annotated.

Only one gene, *STARD13*, was predicted as a potential target of SNPs at 13q13.1. This tumor suppressor gene has been previously implicated in metastasis, proliferation and development[60]. However, rs736596, localized at 13q13.1, showed no association in CIMBA analyses and the association observed in our case-only analysis showed heterogeneity by country.

At the 2p14 locus, INQUISIT-predicted target genes included *ETAA1* with lower confidence. The OR estimates obtained in the case-only analysis for the SNPs located in this gene were consistent with the HR estimated in previously reported CIMBA analyses[56]. Moreover, around one hundred correlated SNPs, were associated with *BRCA2* mutation carrier status at $p<10^{-8}$, including the genotyped SNP chr2_67654113_C_T.

The validity of the case-only analysis as evidence of interaction relies on the assumption of independence between the mutation status and the SNPs under investigation[61]. Therefore, based on the control-only analyses, we excluded approximately 2,000 SNPs because of their association with *BRCA1* or *BRCA2* mutation carrier status, which also showed an association with risk in the case-only analyses (Supplementary Figure 5). Nevertheless, these SNPs, which are potentially in LD or in interchromosomal linkage disequilibrium with *BRCA1* or *BRCA2* mutations, may warrant further investigation to understand their potential role in BC risk. A recent publication using data from the Framingham Heart Study suggested that

interchromosomal linkage disequilibrium can be caused by bio-genetic mechanisms possibly associated with favourable or unfavourable epistatic evolution[62]. SNPs, for which no association with mutation carrier status was found at the significance level of 10^{-8} were assumed to be independent of the mutation status. However, this does not necessarily rule out residual LD between the novel SNPs on chromosomes 13 and 17 and *BRCA1 or BRCA2* mutations. Therefore, the OR estimates for these SNPs might be biased and may further explain the lack of evidence of association in the CIMBA only analyses.

Our findings highlight the importance of imputation in GWAS. The imputed genome-wide genotype data used in the main case-only association analyses were based on carrying out the imputation separately for the BC cases from BCAC and CIMBA. We found that 28 out of the 33 regions associated with *BRCA1* or *BRCA2* mutation carrier status were no longer associated with risk after re-imputing all samples together. By re-imputing all the data together we ensured that the associations observed for the remaining regions are robust to potential differences in the imputation accuracy between the BCAC and CIMBA samples.

Under our analytical strategy, only the regions for which evidence of associated with BC risk was observed were re-imputed using all BCAC and CIMBA samples combined. This re-imputation was not done at genome-wide level due to computational constraints and this may have led to false-negative associations being excluded for further evaluation as potential novel modifiers. Future analyses should aim to analyse the genome-wide associations after the genome-wide re-imputation across the combined BCAC and CIMBA dataset. Our approach suggests that the number of false-positives is likely to be small for common and well imputed SNPs.

Due to the recruitment of participants in CIMBA studies primarily through genetic counselling, the mean age at diagnosis of mutation carriers was 16 years younger than the BC cases participating in BCAC. Although all analyses were adjusted for age, the observed associations might be related to the ageing process instead of interactions with mutation carrier status. Another source of bias could be related to the fact that there are 1.5 times more prevalent cases among CIMBA (68.1%) than BCAC (42.3%) with a delay between diagnosis and study recruitment of 6.83 years and 2.07 years respectively. An observed association might be due to a differential survival between CIMBA and BCAC cases. However, none of the identified SNPs has been found to be associated with BC survival[63].

The majority (92.5%) of cases and controls in BCAC were not tested for *BRCA1/2* mutations at the time of enrolment, potentially leading to some attenuation in the interaction OR (as some BCAC cases will be carriers). However, most BCAC studies were population-based case-control studies and the proportion of cases and controls that carry pathogenic *BRCA1/2* mutations will be small (<5%), hence any attenuation is likely to be negligible.

Despite heterogeneity in the interaction ORs by country for some SNPs, results were generally robust to the exclusion of each country sequentially except, for two SNPs (rs79183898 and rs736596) found associated with *BRCA2* mutation carrier status; for these, the association seemed to be driven by data from the USA. For the other SNPs, the observed heterogeneity may be due to random error, given the relatively small sample sizes of each country. However, if these differences are real, future PRS for *BRCA1* and *BRCA2* carriers should consider the country specific differences.

This is the first analysis of genetic modifiers of BC risk that investigated the differences in the association of common genetic variants with BC risk in the general population and in women with *BRCA1* or *BRCA2* mutations. The inclusion of unselected BC cases resulted in an increased sample size and hence a gain in statistical power for identifying novel SNPs. More detailed fine mapping and functional analysis will be required to elucidate the role of the novel variants identified in BC development for *BRCA1* and *BRCA2* mutation carriers. Our findings should contribute to the improved performance of BC PRS for absolute risk prediction for *BRCA1* and *BRCA2* mutation carriers, which will help inform decisions on the best timing for risk reducing surgery, risk reduction medication, or start of surveillance.

Methods (2,657 words (max is 3,000 for methods in nature communication))

Study sample

We used data from two international consortia, BCAC[64] and CIMBA[56]. BCAC included data from 108 studies of BC from 33 countries in North America, Europe and Australia, the majority (88%) of which were case-control studies. The majority of BCAC cases/controls were not tested for *BRCA1/2* mutations at the time of enrolment. However, most studies were population-based, hence the proportion of cases and controls that carry pathogenic *BRCA1/2* mutations will be small. CIMBA participants were women with pathogenic mutations in *BRCA1* or *BRCA2*. All participants were at least 18 years old. The majority of mutation carriers were recruited through cancer genetics clinics and enrolled into national or regional studies. Data were available on 30,500 *BRCA1* mutation carriers and 20,500 *BRCA2* mutation carriers from 77 studies in 32 countries. A total of 188,320 BC cases and 161,669 controls were available from both consortia. All studies provided information on disease status, age at diagnosis or at interview. Oestrogen receptor status was available for 72% of BCAC cases. All subjects provided written informed consent and participated in studies with protocols approved by ethics committees at each participating institution.

Sample selection

Details of the BCAC study designs have been published elsewhere[7]. Cases were women diagnosed with BC. To define disease status in CIMBA participants, women were censored at the first of the following events: age at BC diagnosis, age at ovarian cancer diagnosis, other cancer, bilateral prophylactic mastectomy or age at study recruitment. Subjects censored at a BC diagnosis were considered as cases.

A control-only analysis was carried out to test the independence between the SNPs and the *BRCA1* and *BRCA2* mutation carrier status. In BCAC, controls were defined as individuals unaffected by BC at study recruitment[35]. In CIMBA, participants were considered as controls if they were unaffected at recruitment.

Only women of European ancestry were included. To minimise the chance of observing spurious associations due to differences in the distribution of BC cases in the population by tumour characteristics (defined as unselected BC cases), 3,478 BCAC cases from 4 studies were excluded because they were included in clinical trials based on breast tumour

characteristics as HER-2 receptor status (see Supplementary Table 2). Because all the analyses were adjusted for country, to ensure that the number of subjects in each country stratum was large enough, we excluded the CIMBA data from any country for which there were less than ten BC cases with *BRCA1* or *BRCA2* mutation. Consequently, data from Poland and Russia were excluded from the BRCA2 analyses (Supplemental Table 3). Finally, duplicate subjects between BCAC and CIMBA were excluded from the BCAC data (114 and 80 subjects from the BRCA1 and BRCA2 case-only analyses, respectively; eight subjects from control-only analyses).

A total of 60,212 BCAC cases and 7,257 *BRCA1* mutation carrier cases were available for the BRCA1 case-only analyses and 57,725 BCAC cases and 5,097 *BRCA2* mutation carrier cases were available for the BRCA2 case-only analyses (Figure 1). A total of 45,881 BCAC controls and 5,750 *BRCA1* mutation carrier controls were available for the BRCA1 control-only analyses and 43,549 BCAC controls and 4,456 *BRCA2* mutation carrier controls for the BRCA2 control-only analyses (Figure 2).

Genotype data

All the study samples were genotyped using the OncoArray Illumina beadchip[67]. The array includes a backbone of approximately 260,000 SNPs that provide genome-wide coverage of most common variants, together with markers of interest for breast and other cancers identified through GWAS, fine-mapping of known susceptibility regions, and other approaches. Further details on the design of the array have been published elsewhere[65].

Quality control

A standard genotype quality control process was followed for both the BCAC and CIMBA samples which has been described in detail elsewhere[35,48]. Briefly, this involved excluding SNPs located on chromosome Y; SNPs with call rates < 95%; SNPs with minor allele frequency (MAF) <0.05 and call rate <98%; monomorphic SNPs; and SNPs for which evidence of departure from Hardy-Weinberg equilibrium was observed ($P<10^{-7}$ based on a country-stratified test).

Imputation

Genotypes for ~21 Million SNPs were imputed for all subjects using the 1000 Genomes Phase III data (released October 2014) as reference panel, as described previously[66]. Briefly,

the number of reference haplotypes used as templates when imputing missing genotypes was fixed to 800 (-k_hap = 800). A two-stage imputation approach was used: phasing with SHAPEIT[67,68] and imputation with IMPUTE2[69] using 5Mb non-overlapping intervals. Genotypes were imputed for all SNPs that were found polymorphic (MAF>0.1%) in either European or Asian populations.

The genome-wide imputation process described above was carried out separately for the BCAC and CIMBA samples. However, this may potentially lead to spurious associations if there are differences in the quality of the imputation (measured using the imputation accuracy r^2 metric[70]) for a given SNP between the two datasets. To address this, a stringent approach was employed which involved including only SNPs for which the difference in r^2 between the BCAC and CIMBA SNP imputations (Δr^2) was minimal relative to their r^2 values. SNPs with $r^2>0.9$ in both BCAC and CIMBA were kept in the analyses only if $\Delta r^2<0.05$; SNPs with $0.8< r^2\leq 0.9$ in both BCAC and CIMBA were kept if $\Delta r^2<0.02$ and, SNPs with $0.5<r^2\leq 0.8$ in both BCAC and CIMBA were kept if $\Delta r^2<0.01$. All SNPs with $r^2<0.5$ in either CIMBA or BCAC were excluded. Only SNPs with a MAF greater than 0.01 in BCAC cases were included.

Consequently, 9,072,535 SNPs were included in the BRCA1 analyses (402,336 genotyped and 8,670,199 imputed SNPs) and 9,047,403 SNPs in the BRCA2 analyses (402,397 genotyped and 8,645,006 imputed SNPs).

Statistical analysis

Case-only and control-only analyses

The comparison of SNP frequency between CIMBA cases and BCAC cases (case-only analyses), or between unaffected CIMBA subjects and BCAC controls (control-only analyses), was performed using logistic regression adjusted for age at BC diagnosis in the case-only analyses and for age at interview for BCAC controls or at censure for CIMBA unaffected subjects in the control-only analyses, as well as for country and principal components (PCs) to account for population structure. Separate analyses were carried out for *BRCA1* and *BRCA2* mutation carriers. To define the number of principal components (PC) for inclusion in the models, principal component analysis was carried out using 35,858 uncorrelated genotyped SNPs on the OncoArray and purpose-written software (http://ccge.medschl.cam.ac.uk/software/pccalc/). The inflation statistic was calculated and

converted to an equivalent statistic for a study of 1,000 subjects for each outcome ($\lambda_{1,000}$) by adjusting for effective study size:

$$\lambda_{1,000} = (\lambda - 1)\left(\frac{1}{n} + \frac{1}{m}\right) * 500 + 1$$

where n and m are the numbers of BCAC and CIMBA subjects respectively. The models were adjusted with the first four PCs ($\lambda_{1,000}$ with and without PCs in the model = 1.03 and 1.21, respectively) since additional PCs did not result in further reduction in the inflation of the test statistics.

Strategy for determining significant associations

The analytical process is summarised in Figures 3 and 4. A fundamental assumption when using a case-only design in this context is that the SNPs and mutation carrier status are independent[61]. To confirm independence, SNPs likely to be in linkage disequilibrium (LD) with *BRCA1* or *BRCA2* mutations, i.e. those located in or within 500 kb of either gene, were excluded. However, LD also exist between variants at long-distance on the same chromosome or even on a different chromosome (interchromosomal linkage disequilibrium)[62,71]. Therefore, control-only analyses were performed to further exclude SNPs associated with mutation carrier status in unaffected women[72], using a stringent statistical significance level of 10^{-8}).

After excluding SNPs in LD or in interchromosomal linkage disequilibrium with *BRCA1* or *BRCA2* mutations, case-only analyses were performed to assess the association between SNPs and *BRCA1* or *BRCA2* mutation carrier status. We considered two categories of SNPs depending on whether they had been previously found to be associated with BC in published BCAC studies[35,48]. For known BC susceptibility SNPs (Figure 3) we used a significance threshold of 2.7×10^{-4} (applying Bonferroni correction to 179 tests) and for "*potential novel SNP modifier*" (Figure 4) a stringent significance threshold of 10^{-8} was used.

Because *BRCA1* mutation-associated tumours are more often ER-negative than those in the general population[73], a subsequent case-only analysis was performed restricting the BCAC cases to those with ER-negative disease. We used this strategy for two reasons. First, to exclude associations driven by differences in the tumour ER-status distributions between *BRCA1* carriers and BCAC cases. Therefore, in the *BRCA1* analysis, SNPs were considered to be associated with mutation carrier status only if they were also associated in the ER-negative case-only analysis at a prior defined significance threshold of 10^{-7} for novel SNP modifiers and of 0.05 for the established BC susceptibility SNPs. The second reason we applied this

strategy was to identify novel SNP modifiers specific to *BRCA1*/ER-negative tumours that had not been detected in the overall analysis; for this we applied a significance threshold of 10^{-8}.

To confirm that potentially novel associations in the case-only analysis were not driven by differences in the imputation accuracy between the CIMBA and BCAC data, each of the regions defined as +/-500 kb around the associated SNP, were re-imputed for the combined CIMBA and BCAC samples. The more accurate one-stage imputation was carried out, using IMPUTE2 without pre-phasing. Associations with all the SNPs in the re-imputed regions were then re-evaluated using the control-only and case-only analytical approaches described above. Finally, we used a step-wise regression analysis using a significance threshold of 10^{-8} in order to determine whether associations with SNPs in the same region are independent and to define the conditionally independent SNPs ("top SNPs").

Among the 179 established BC susceptibility SNPs, 107 were genotyped and 71 were imputed. As previously, although none of these 71 SNPs were excluded based on their Δr^2, to exclude potentially spurious associations, regions around these 71 SNPs were re-imputed using the one-stage imputation applied to BCAC and CIMBA data combined, and before performing the control-only and case-only analyses.

Determining the magnitude of association.

For the potentially novel SNP modifiers the risk ratio of BC applicable to mutation carriers was assumed to be equal to the OR estimate from the case-only analysis (with the hypothesis that their relative risk equals 1 in the general population, given that none of them were found to be associated with BC in BCAC)[55].

For the known BC susceptibility SNPs, a significant association in the case-only analysis implies that the magnitude of association is different for *BRCA1* or *BRCA2* mutation carriers than for the general population. Therefore, the risk ratio of BC for mutation carriers was computed as the product of $OR \times OR_{BCAC}$ where OR was obtained from the case-only analysis, and OR_{BCAC} was the odds ratio of association obtained from either Michailidou *et al*.[35] for the SNPs associated with overall BC risk and from Milne *et al*.[48] for the SNPs associated with ER-negative BC.

For all associated SNPs in case-only analyses, heterogeneity by country was assessed using likelihood ratio tests that compared models with and without a SNP by country interaction term. When the heterogeneity test was significant at $P<0.05$, a leave-one-out analysis was

performed, by excluding each country in turn to assess the influence of a data from a specific country on the overall association.

Credible causal variants

For each novel region, we defined sets of credible causal variants (CCVs) to use in the prediction of the likely target genes. For this purpose, we defined a first set of CCVs including the "top SNP" of the region of interest and the SNPs with p-values of association within two orders of magnitude of the "top SNP" association. Then, we sequentially performed logistic regression analyses using all other SNPs in the region, adjusted for the "top SNP". We defined a second set of CCVs which included the "most significant SNP" after adjusting for the "top SNP" and the SNPs with p-values within two orders of magnitude of the "most significant SNP" association. This was repeated (conditioning on the previously found most significant SNPs) to define additional sets of CCVs as long as at least one p-value remained $<10^{-6}$.

***In silico* analyses**

eQTL Analysis

Data from breast cancer tumors and adjacent normal breast tissue were accessed from The Cancer Genome Atlas[74] (TCGA). Germline SNP genotypes (Affymetrix 6.0 arrays) from individuals of European ancestry were processed and imputed to the 1000 Genomes reference panel (October 2014)[35]. Tumor tissue copy number was estimated from the Affymetrix 6.0 and called using the GISTIC2 algorithm[75]. Complete genotype, RNA-seq and copy number data were available for 679 genetically European patients (78 with adjacent normal tissue). Further, RNA-seq for normal breast tissue and imputed germline genotype data were available from 80 females from the GTEx Consortium[76]. Genes with a median expression level of 0 RPKM across samples were removed, and RPKM values of each gene were log2 transformed. Expression values of samples were quantile normalized. Genetic variants were evaluated for association with the expression of genes located within ±2Mb of the lead variant at each risk region using linear regression models, adjusting for ESR1 expression. Tumor tissue was also adjusted for copy number variation, as previously described[77]. eQTL analyses were performed using the MatrixEQTL program in R[78].

INQUISIT analyses

Each candidate target genes were evaluated by assessing each CCV's potential impact on regulatory or coding features using a computational pipeline, INtegrated expression QUantitative trait and In SIlico prediction of GWAS Targets (INQUISIT)[35,54]. Briefly, genes were considered as potential targets of candidate causal variants through effects on: (1) distal gene regulation, (2) proximal regulation, or (3) a gene's coding sequence. We intersected CCV positions with multiple sources of genomic information chromatin interaction analysis by paired-end tag sequencing (ChIA-PET[79]) in MCF7 cells and genome-wide chromosome conformation capture (Hi-C) in HMECs[82]. We used breast cell line computational enhancer–promoter correlations (PreSTIGE[80], IM-PET[81], FANTOM5[82]) breast cell super-enhancer[83], breast tissue-specific expression variants (eQTL) from multiple independent studies (TCGA (normal breast and breast tumor) and GTEx breast – see eQTL methods), transcription factor and histone modification chromatin immunoprecipitation followed by sequencing (ChIP-seq) from the ENCODE and Roadmap Epigenomics Projects together with the genomic features found to be significantly enriched for all known breast cancer CCVs[54], gene expression RNA-seq from several breast cancer lines and normal samples (ENCODE) and topologically associated domain (TAD) boundaries from T47D cells (ENCODE[84]). To assess the impact of intragenic variants, we evaluated their potential to alter primary protein coding sequence and splicing using Ensembl Variant Effect Predictor[85] using MaxEntScan and dbscSNV modules for splicing alterations based on "ada" and "rf" scores. Nonsense and missense changes were assessed with the REVEL ensemble algorithm, with CCVs displaying REVEL scores > 0.5 deemed deleterious.

Each target gene prediction category (distal, promoter or coding) was scored according to different criteria. Genes predicted to be distally-regulated targets of CCVs were awarded points based on physical links (for example ChIA-PET), computational prediction methods, or eQTL associations. All CCVs were considered as potentially involved in distal regulation. Intersection of a putative distal enhancer with genomic features found to be significantly enriched[54] were further upweighted. Multiple independent interactions were awarded an additional point. CCVs in gene proximal regulatory regions were intersected with histone ChIP-Seq peaks characteristic of promoters and assigned to the overlapping transcription start sites (defined as -1.0 kb - +0.1 kb). Further points were awarded to such genes if there was evidence for eQTL association, while a lack of expression resulted in down-weighting as potential targets. Potential coding changes including missense, nonsense and predicted

splicing alterations resulted in addition of one point to the encoded gene for each type of change, while lack of expression reduced the score. We added an additional point for predicted target genes that were also breast cancer drivers (278 genes[35,54]). For each category, scores potentially ranged from 0-8 (distal); 0-4 (promoter) or 0-3 (coding). We converted these scores into 'confidence levels': Level 1 (highest confidence) when distal score > 4, promoter score >= 3 or coding score > 1; Level 2 when distal score <= 4 and >=1, promoter score = 1 or = 2, coding score = 1; and Level 3 when distal score < 1 and > 0, promoter score < 1 and > 0, and coding < 1 and > 0. For genes with multiple scores (for example, predicted as targets from multiple independent risk signals or predicted to be impacted in several categories), we recorded the highest score.

Authors' contributions

ACA, DFE and NA conceived the study design. JC, NA and ACA drafted the initial manuscript, while the complete writing group consisted of JC, NA, ACA, GCT and DFE. JC performed the statistical analyses, and JB the INQUISIT predictions.

All authors read and approved the final manuscript. The funders had no role in the design of the study, the collection, analysis, or interpretation of the data, the writing of the manuscript, or the decision to submit the manuscript for publication.

References

1. Ferlay, J. *et al*. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
2. Pharoah, P. D. P., Day, N. E., Duffy, S., Easton, D. F. & Ponder, B. A. J. Family history and the risk of breast cancer: A systematic review and meta-analysis. *Int. J. Cancer* **71**, 800–809 (1997).
3. Nelson, H. D. *et al*. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: a systematic review to update the U.S. Preventive Services Task Force recommendation. *Ann. Intern. Med.* **160**, 255–266 (2014).
4. Kuchenbaecker, K. B. *et al*. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* **317**, 2402–2416 (2017).
5. Antoniou, A. C. *et al*. Common variants in LSP1, 2q35 and 8q24 and breast cancer risk for BRCA1 and BRCA2 mutation carriers. *Hum. Mol. Genet.* **18**, 4442–4456 (2009).
6. Antoniou, A. C. *et al*. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor–negative breast cancer in the general population. *Nat. Genet.* **42**, 885–892 (2010).
7. Michailidou, K. *et al*. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–361 (2013).
8. Thomas, G. *et al*. A multi-stage genome-wide association in breast cancer identifies two novel risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.* **41**, 579–584 (2009).
9. Finucane, H. K. *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
10. Garcia-Closas, M. *et al*. Genome-wide association studies identify four ER negative–specific breast cancer risk loci. *Nat. Genet.* **45**, 392-398e2 (2013).
11. Couch, F. J. *et al*. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. *Nat. Commun.* **7**, (2016).
12. Lin, W.-Y. *et al*. Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. *Hum. Mol. Genet.* **24**, 285–298 (2015).
13. Milne, R. L. *et al*. Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium. *Hum. Mol. Genet.* **23**, 6096–6111 (2014).
14. Haiman, C. A. *et al*. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor–negative breast cancer. *Nat. Genet.* **43**, 1210–1214 (2011).
15. Bojesen, S. E. *et al*. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* **45**, 371-384e2 (2013).
16. Ghoussaini, M. *et al*. Evidence that the 5p12 Variant rs10941679 Confers Susceptibility to Estrogen-Receptor-Positive Breast Cancer through FGF10 and MRPS30 Regulation. *Am. J. Hum. Genet.* **99**, 903–911 (2016).
17. Glubb, D. M. *et al*. Fine-Scale Mapping of the 5q11.2 Breast Cancer Locus Reveals at Least Three Independent Risk Variants Regulating MAP3K1. *Am. J. Hum. Genet.* **96**, 5–20 (2015).
18. Gaudet, M. M. *et al*. Identification of a BRCA2-Specific Modifier Locus at 6p24 Related to Breast Cancer Risk. *PLoS Genet.* **9**, (2013).
19. Siddiq, A. *et al*. A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum. Mol. Genet.* **21**, 5373–5384 (2012).
20. Dunning, A. M. *et al*. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat. Genet.* **48**, 374–386 (2016).
21. Sawyer, E. *et al*. Genetic Predisposition to In Situ and Invasive Lobular Carcinoma of the Breast. *PLOS Genet.* **10**, e1004285 (2014).
22. Easton, D. F. *et al*. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
23. Turnbull, C. *et al*. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–507 (2010).
24. Orr, N. *et al*. Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. *Hum. Mol. Genet.* **24**, 2966–2984 (2015).

25. Darabi, H. *et al*. Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression. *Am. J. Hum. Genet*. **97**, 22–34 (2015).
26. Meyer, K. B. *et al*. Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1. *Am. J. Hum. Genet*. **93**, 1046–1060 (2013).
27. French, J. D. *et al*. Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. *Am. J. Hum. Genet*. **92**, 489–503 (2013).
28. Zeng, C. *et al*. Identification of independent association signals and putative functional variants for breast cancer risk through fine-scale mapping of the 12p11 locus. *Breast Cancer Res. BCR* **18**, (2016).
29. Ghoussaini, M. *et al*. Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat. Genet*. **44**, 312–318 (2012).
30. Udler, M. S. *et al*. Fine scale mapping of the breast cancer 16q12 locus. *Hum. Mol. Genet*. **19**, 2507–2515 (2010).
31. Darabi, H. *et al*. Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). *Sci. Rep*. **6**, 32512 (2016).
32. Long, J. *et al*. Genome-Wide Association Study in East Asians Identifies Novel Susceptibility Loci for Breast Cancer. *PLOS Genet*. **8**, e1002532 (2012).
33. Cai, Q. *et al*. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat. Genet*. **46**, 886–890 (2014).
34. Long, J. *et al*. A Common Deletion in the APOBEC3 Genes and Breast Cancer Risk. *JNCI J. Natl. Cancer Inst*. **105**, 573–579 (2013).
35. Michailidou, K. *et al*. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
36. Mavaddat, N. *et al*. Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. *JNCI J. Natl. Cancer Inst*. **107**, (2015).
37. Pashayan, N., Morris, S., Gilbert, F. J. & Pharoah, P. D. P. Cost-effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer: A Life-Table Model. *JAMA Oncol*. (2018). doi:10.1001/jamaoncol.2018.1901
38. Mavaddat, N. *et al*. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet*. **104**, 21–34 (2019).
39. Antoniou, A. C. *et al*. RAD51 135G→C Modifies Breast Cancer Risk among BRCA2 Mutation Carriers: Results from a Combined Analysis of 19 Studies. *Am. J. Hum. Genet*. **81**, 1186–1200 (2007).
40. Garcia-Closas, M. *et al*. Heterogeneity of Breast Cancer Associations with Five Susceptibility Loci by Clinical and Pathological Characteristics. *PLoS Genet*. **4**, e1000054 (2008).
41. Antoniou, A. C. *et al*. Common Breast Cancer-Predisposition Alleles Are Associated with Breast Cancer Risk in BRCA1 and BRCA2 Mutation Carriers. *Am. J. Hum. Genet*. **82**, 937–948 (2008).
42. Silva, L. D. & Lakhani, S. R. Pathology of hereditary breast cancer. *Mod. Pathol*. **23**, S46–S51 (2010).
43. Antoniou, A. C. *et al*. Common alleles at 6q25.1 and 1p11.2 are associated with breast cancer risk for BRCA1 and BRCA2 mutation carriers. *Hum. Mol. Genet*. **20**, 3304–3321 (2011).
44. Antoniou, A. C. *et al*. Common variants at 12p11, 12q24, 9p21, 9q31.2 and in ZNF365 are associated with breast cancer risk for BRCA1 and/or BRCA2 mutation carriers. *Breast Cancer Res. BCR* **14**, R33 (2012).
45. Couch, F. J. *et al*. Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk. *PLoS Genet*. **9**, (2013).
46. Kuchenbaecker, K. B. *et al*. Associations of common breast cancer susceptibility alleles with risk of breast cancer subtypes in BRCA1 and BRCA2 mutation carriers. *Breast Cancer Res. BCR* **16**, (2014).
47. Lawrenson, K. *et al*. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast–ovarian cancer susceptibility locus. *Nat. Commun*. **7**, (2016).
48. Milne, R. L. *et al*. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet*. **49**, 1767–1778 (2017).

49. Kuchenbaecker, K. B. *et al*. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *JNCI J. Natl. Cancer Inst.* **109**, (2017).
50. Chenevix-Trench, G. *et al*. An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). *Breast Cancer Res.* **9**, 104 (2007).
51. Pierce, B. L. & Ahsan, H. Case-only Genome-wide Interaction Study of Disease Risk, Prognosis and Treatment. *Genet. Epidemiol.* **34**, 7–15 (2010).
52. Ottman, R. Gene–Environment Interaction: Definitions and Study Designs. *Prev. Med.* **25**, 764–770 (1996).
53. Andrieu, N. & Goldstein, A. M. Epidemiologic and Genetic Approaches in the Study of Gene-Environment Interaction: an Overview of Available Methods. *Epidemiol. Rev.* **20**, 137–147 (1998).
54. Fachal, L. *et al*. Fine-mapping of 150 breast cancer risk regions identifies 178 high confidence target genes. *bioRxiv* 521054 (2019). doi:10.1101/521054
55. Whittemore, A. S. Assessing environmental modifiers of disease risk associated with rare mutations. *Hum. Hered.* **63**, 134–143 (2007).
56. CIMBA - Consortium of Investigators of Modifiers of BRCA1/2 —. Available at: http://cimba.ccge.medschl.cam.ac.uk/. (Accessed: 13th March 2017)
57. Turner, A. *et al*. MADD knock-down enhances doxorubicin and TRAIL induced apoptosis in breast cancer cells. *PloS One* **8**, e56817 (2013).
58. Zheng, T., Wang, A., Hu, D. & Wang, Y. Molecular mechanisms of breast cancer metastasis by gene expression profile analysis. *Mol. Med. Rep.* **16**, 4671–4677 (2017).
59. Sharma, D. K., Bressler, K., Patel, H., Balasingam, N. & Thakor, N. Role of Eukaryotic Initiation Factors during Cellular Stress and Cancer Progression. *J. Nucleic Acids* **2016**, (2016).
60. HANNA, S. *et al*. StarD13 is a tumor suppressor in breast cancer that regulates cell motility and invasion. *Int. J. Oncol.* **44**, 1499–1511 (2014).
61. Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat. Med.* **13**, 153–162 (1994).
62. Kulminski, A. M. Complex phenotypes and phenomenon of genome-wide inter-chromosomal linkage disequilibrium in the human genome. *Exp. Gerontol.* **46**, 979–986 (2011).
63. Escala-Garcia, M. *et al*. Genome-wide association study of germline variants and breast cancer-specific mortality. *Br. J. Cancer* **120**, 647–657 (2019).
64. BCAC - The Breast Cancer Association Consortium —. Available at: http://bcac.ccge.medschl.cam.ac.uk/. (Accessed: 13th March 2017)
65. Amos, C. I. *et al*. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **26**, 126–135 (2017).
66. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
67. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
68. O'Connell, J. *et al*. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLOS Genet.* **10**, e1004234 (2014).
69. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* **5**, (2009).
70. Browning, B. L. & Browning, S. R. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
71. Hohenlohe, P. A., Bassham, S., Currey, M. & Cresko, W. A. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 395–408 (2012).
72. Umbach, D. M. & Weinberg, C. R. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat. Med.* **16**, 1731–1743 (1997).
73. Spurdle, A. B. *et al*. Refined histopathological predictors of BRCA1 and BRCA2mutation

status: a large-scale analysis of breast cancer characteristics from the BCAC, CIMBA, and ENIGMA consortia. *Breast Cancer Res*. *BCR* **16**, (2014).
74. Comprehensive molecular portraits of human breast tumors. *Nature* **490**, 61–70 (2012).
75. Mermel, C. H. *et al*. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. **12**, R41 (2011).
76. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat*. *Genet*. **45**, 580–585 (2013).
77. Li, Q. *et al*. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633–641 (2013).
78. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinforma*. *Oxf*. *Engl*. **28**, 1353–1358 (2012).
79. Fullwood, M. J. *et al*. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
80. Corradin, O. *et al*. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res*. **24**, 1–13 (2014).
81. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer–promoter interactome in human cells. *Proc*. *Natl*. *Acad*. *Sci*. *U*. *S*. *A*. **111**, E2191–E2199 (2014).
82. Andersson, R. *et al*. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
83. Hnisz, D. *et al*. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
84. Dixon, J. R. *et al*. Integrative detection and analysis of structural variation in cancer genomes. *Nat*. *Genet*. **50**, 1388–1398 (2018).
85. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biol*. **17**, 122 (2016).

Fundings

Juliette Coignard is supported by a fellowship of INCa Institut National du Cancer N°2015-181, la Ligue Nationale contre le Cancer IP/SC-15229 and Olga Sinilnikova's fellowship (2016).

BCAC Funding

BCAC is funded by Cancer Research UK [C1287/A16563, C1287/A10118], the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and by the European Community´s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report.

Genotyping of the OncoArray was funded by the NIH Grant U19 CA148065, and Cancer UK Grant C1287/A16563 and the PERSPECTIVE project supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344) and, the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec and the PSRSIIRI-701 grant, and the Quebec Breast Cancer Foundation.

The Australian Breast Cancer Family Study (ABCFS) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. J.L.H. is a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellow. M.C.S. is a NHMRC Senior Research Fellow. The ABCS study was supported by the Dutch Cancer Society [grants NKI 2007-3839; 2009 4363]. The Australian Breast Cancer Tissue Bank (ABCTB) was supported by the National Health and Medical Research Council of Australia, The Cancer Institute NSW and the National Breast Cancer Foundation. The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen. The BBCS is funded by Cancer Research UK and Breast Cancer Now and acknowledges NHS funding to the NIHR Biomedical Research Centre, and the National Cancer Research Network (NCRN). The BCEES was funded by the National Health and Medical Research Council, Australia and the Cancer Council Western Australia and acknowledges funding from the National Breast Cancer Foundation (JS). For the BCFR-NY, BCFR-PA, BCFR-UT this work was supported by grant UM1 CA164920 from the National Cancer Institute. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the BCFR. The BREast Oncology GAlician Network (BREOGAN) is funded by Acción Estratégica de Salud del Instituto de Salud Carlos III FIS PI12/02125/Cofinanciado FEDER; Acción Estratégica de Salud del Instituto de Salud Carlos III FIS Intrasalud (PI13/01136); Programa Grupos Emergentes, Cancer Genetics Unit, Instituto de Investigacion Biomedica Galicia Sur. Xerencia de Xestion Integrada de Vigo-SERGAS, Instituto de Salud Carlos III, Spain; Grant 10CSA012E, Consellería de Industria Programa Sectorial de Investigación Aplicada, PEME I + D e I + D Suma del Plan Gallego de Investigación, Desarrollo e Innovación Tecnológica de la Consellería de Industria de la Xunta de Galicia, Spain; Grant EC11-192. Fomento de la Investigación Clínica Independiente,

Ministerio de Sanidad, Servicios Sociales e Igualdad, Spain; and Grant FEDER-Innterconecta. Ministerio de Economia y Competitividad, Xunta de Galicia, Spain. The BSUCH study was supported by the Dietmar-Hopp Foundation, the Helmholtz Society and the German Cancer Research Center (DKFZ). CBCS is funded by the Canadian Cancer Society (grant # 313404) and the Canadian Institutes of Health Research. CCGP is supported by funding from the University of Crete. The CECILE study was supported by Fondation de France, Institut National du Cancer (INCa), Ligue Nationale contre le Cancer, Agence Nationale de Sécurité Sanitaire, de l'Alimentation, de l'Environnement et du Travail (ANSES), Agence Nationale de la Recherche (ANR). The CGPS was supported by the Chief Physician Johan Boserup and Lise Boserup Fund, the Danish Medical Research Council, and Herlev and Gentofte Hospital. The CNIO-BCS was supported by the Instituto de Salud Carlos III, the Red Temática de Investigación Cooperativa en Cáncer and grants from the Asociación Española Contra el Cáncer and the Fondo de Investigación Sanitario (PI11/00923 and PI12/00070). The CTS was initially supported by the California Breast Cancer Act of 1993 and the California Breast Cancer Research Fund (contract 97-10500) and is currently funded through the National Institutes of Health (R01 CA77398, UM1 CA164917, and U01 CA199277). Collection of cancer incidence data was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885. The University of Westminster curates the DietCompLyf database funded by Against Breast Cancer Registered Charity No. 1121258 and the NCRN. The coordination of EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by: Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), Federal Ministry of Education and Research (BMBF) (Germany); the Hellenic Health Foundation, the Stavros Niarchos Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Health Research Fund (FIS), PI13/00061 to Granada, PI13/01162 to EPIC-Murcia, Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, ISCIII RETIC (RD06/0020) (Spain); Cancer Research UK (14136 to EPIC-Norfolk; C570/A16491 and C8221/A19170 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk, MR/M012190/1 to EPIC-Oxford) (United Kingdom). The ESTHER study was supported by a grant from the Baden Württemberg Ministry of Science, Research and Arts. Additional cases were recruited in the context of the VERDI study, which was supported by a grant from the German Cancer Aid (Deutsche Krebshilfe). The GC-HBOC (German Consortium of Hereditary Breast and Ovarian Cancer) is supported by the German Cancer Aid (grant no 110837, coordinator: Rita K. Schmutzler, Cologne). This work was also funded by the European Regional Development Fund and Free State of Saxony, Germany (LIFE - Leipzig Research Centre for Civilization Diseases, project numbers 713-241202, 713-241202, 14505/2470, 14575/2470). The GENICA was funded by the Federal Ministry of Education and Research (BMBF) Germany grants 01KW9975/5, 01KW9976/8, 01KW9977/0 and 01KW0114, the Robert Bosch Foundation, Stuttgart, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, the Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, as well as the Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany. The GESBC was supported by the Deutsche Krebshilfe e. V. [70492] and the German Cancer Research

Center (DKFZ). The HABCS study was supported by the Claudia von Schilling Foundation for Breast Cancer Research, by the Lower Saxonian Cancer Society, and by the Rudolf Bartling Foundation. The HEBCS was financially supported by the Helsinki UniversityHospital Research Fund, the Finnish Cancer Society, and the Sigrid Juselius Foundation. The HUBCS was supported by a grant from the German Federal Ministry of Research and Education (RUS08/017), and by the Russian Foundation for Basic Research and the Federal Agency for Scientific Organizations for support the Bioresource collections and RFBR grants 14-04-97088, 17-29-06014 and 17-44-020498. Financial support for KARBAC was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and Bert von Kantzows foundation. The KARMA study was supported by Märit and Hans Rausings Initiative Against Breast Cancer. The KBCP was financially supported by the special Government Funding (EVO) of Kuopio University Hospital grants, Cancer Fund of North Savo, the Finnish Cancer Organizations, and by the strategic funding of the University of Eastern Finland. kConFab is supported by a grant from the National Breast Cancer Foundation, and previously by the National Health and Medical Research Council (NHMRC), the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia. Financial support for the AOCS was provided by the United States Army Medical Research and Materiel Command [DAMD17-01-1-0729], Cancer Council Victoria, Queensland Cancer Fund, Cancer Council New South Wales, Cancer Council South Australia, The Cancer Foundation of Western Australia, Cancer Council Tasmania and the National Health and Medical Research Council of Australia (NHMRC; 400413, 400281, 199600). G.C.T. and P.W. are supported by the NHMRC. RB was a Cancer Institute NSW Clinical Research Fellow. LMBC is supported by the 'Stichting tegen Kanker'. The MARIE study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR I, 106332, 108253, 108419, 110826, 110828], the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. MBCSG is supported by grants from the Italian Association for Cancer Research (AIRC; IG2014 no.15547) to P. Radice. The MCBCS was supported by the NIH grants CA192393, CA116167, CA176785 an NIH Specialized Program of Research Excellence (SPORE) in Breast Cancer [CA116201], and the Breast Cancer Research Foundation and a generous gift from the David F. and Margaret T. Grohne Family Foundation. The Melbourne Collaborative Cohort Study (MCCS) cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database. The MEC was support by NIH grants CA63464, CA54281, CA098758, CA132839 and CA164973. The MISS study is supported by funding from ERC-2011-294576 Advanced grant, Swedish Cancer Society, Swedish Research Council, Local hospital funds, Berta Kamprad Foundation, Gunnar Nilsson. The MMHS study was supported by NIH grants CA97396, CA128931, CA116201, CA140286 and CA177150. MSKCC is supported by grants from the Breast Cancer Research Foundation and Robert and Kate Niehaus Clinical Cancer Genetics Initiative. The work of MTLGEBCS was supported by the Quebec Breast Cancer Foundation, the Canadian Institutes of Health Research for the “CIHR Team in Familial Risks of Breast Cancer” program – grant # CRN-87521 and the Ministry of Economic Development, Innovation and Export Trade – grant # PSR-SIIRI-701. The NBHS was supported by NIH grant R01CA100374. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported

by P30 CA68485. The Northern California Breast Cancer Family Registry (NC-BCFR) and Ontario Familial Breast Cancer Registry (OFBCR) were supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The Carolina Breast Cancer Study was funded by Komen Foundation, the National Cancer Institute (P50 CA058223, U54 CA156733, U01 CA179715), and the North Carolina University Cancer Research Fund. The NHS was supported by NIH grants P01 CA87969, UM1 CA186107, and U19 CA148065. The NHS2 was supported by NIH grants UM1 CA176726 and U19 CA148065. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. Genotyping for PLCO was supported by the Intramural Research Program of the National Institutes of Health, NCI, Division of Cancer Epidemiology and Genetics. The PLCO is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, National Institutes of Health. The POSH study is funded by Cancer Research UK (grants C1275/A11699, C1275/C22524, C1275/A19187, C1275/A15956 and Breast Cancer Campaign 2010PR62, 2013PR044. The RBCS was funded by the Dutch Cancer Society (DDHK 2004-3124, DDHK 2009-4318. SEARCH is funded by Cancer Research UK [C490/A10124, C490/A16561] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. The University of Cambridge has received salary support for PDPP from the NHS in the East of England through the Clinical Academic Reserve. The Sister Study (SISTER) is supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES044005 and Z01-ES049033). The Two Sister Study (2SISTER) was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES044005 and Z01-ES102245), and, also by a grant from Susan G. Komen for the Cure, grant FAS0703856. SKKDKFZS is supported by the DKFZ. The SMC is funded by the Swedish Cancer Foundation and the Swedish Research Council (VR 2017-00644) grant for the Swedish Infrastructure for Medical Population-based Life-course Environmental Research (SIMPLER). The SZBCS and IHCC were supported by Grant PBZ_KBN_122/P05/2004 and the program of the Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in 2019-2022 project number 002/RID/2018/19 amount of financing 12 000 000 PLN. The TNBCC was supported by: a Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), a grant from the Breast Cancer Research Foundation, a generous gift from the David F. and Margaret T. Grohne Family Foundation. The UCIBCS component of this research was supported by the NIH [CA58860, CA92044] and the Lon V Smith Foundation [LVS39420]. The UKBGS is funded by Breast Cancer Now and the Institute of Cancer Research (ICR), London. The UKOPS study was funded by The Eve Appeal (The Oak Foundation) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

CIMBA Funding

CIMBA: The CIMBA data management and data analysis were supported by Cancer Research – UK grants C12292/A20861, C12292/A11174. ACA is a Cancer Research -UK Senior Cancer Research Fellow. GCT and ABS are NHMRC Research Fellows. iCOGS: the

European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 - the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer (CRN-87521), and the Ministry of Economic Development, Innovation and Export Trade (PSR-SIIRI-701), Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The PERSPECTIVE project was supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the Ministry of Economy, Science and Innovation through Genome Québec, and The Quebec Breast Cancer Foundation.
BCFR: UM1 CA164920 from the National Cancer Institute. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the BCFR. BIDMC: Breast Cancer Research Foundation. CNIO: Spanish Ministry of Health PI16/00440 supported by FEDER funds, the Spanish Ministry of Economy and Competitiveness (MINECO) SAF2014-57680-R and the Spanish Research Network on Rare diseases (CIBERER). COH-CCGCRN: Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under grant number R25CA112486, and RC4CA153828 (PI: J. Weitzel) from the National Cancer Institute and the Office of the Director, National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. CONSIT TEAM: Funds from Italian citizens who allocated the 5x1000 share of their tax payment in support of the Fondazione IRCCS Istituto Nazionale Tumori, according to Italian laws (INT-Institutional strategic projects '5x1000') to S. Manoukian. Associazione Italiana Ricerca sul Cancro (AIRC; IG2015 no.16732) to P. Peterlongo. DEMOKRITOS: European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program of the General Secretariat for Research & Technology: SYN11_10_19 NBCA. Investing in knowledge society through the European Social Fund. DKFZ: German Cancer Research Center. EMBRACE: Cancer Research UK Grants C1287/A10118 and C1287/A11990. D. Gareth Evans and Fiona Lalloo are supported by an NIHR grant to the Biomedical Research Centre, Manchester. The Investigators at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust are supported by an NIHR grant to the Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. Ros Eeles and Elizabeth Bancroft are supported by Cancer Research UK Grant C5047/A8385. Ros Eeles is also supported by NIHR support to the Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. FCCC: The University of Kansas Cancer Center (P30 CA168524) and the Kansas Bioscience Authority Eminent Scholar Program. A.K.G. was funded by R0 1CA140323, R01 CA214545, and by the Chancellors Distinguished Chair in Biomedical Sciences Professorship. A.Vega is supported by the Spanish Health Research Foundation, Instituto de Salud Carlos III (ISCIII), partially supported by FEDER funds through Research Activity Intensification Program (contract grant numbers: INT15/00070, INT16/00154, INT17/00133), and through Centro de Investigación Biomédica en Red de Enferemdades Raras CIBERER (ACCI 2016: ER17P1AC7112/2018); Autonomous Government of Galicia (Consolidation and structuring program: IN607B), and

by the Fundación Mutua Madrileña (call 2018). GC-HBOC: German Cancer Aid (grant no 110837, Rita K. Schmutzler) and the European Regional Development Fund and Free State of Saxony, Germany (LIFE - Leipzig Research Centre for Civilization Diseases, project numbers 713-241202, 713-241202, 14505/2470, 14575/2470). GEMO: Ligue Nationale Contre le Cancer; the Association “Le cancer du sein, parlons-en!” Award, the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program and the French National Institute of Cancer (INCa grants 2013-1-BCB-01-ICH-1 and SHS-E-SP 18-015). GEORGETOWN: the Non-Therapeutic Subject Registry Shared Resource at Georgetown University (NIH/NCI grant P30-CA051008), the Fisher Center for Hereditary Cancer and Clinical Genomics Research, and Swing Fore the Cure. G-FAST: Bruce Poppe is a senior clinical investigator of FWO. Mattias Van Heetvelde obtained funding from IWT. HCSC: Spanish Ministry of Health PI15/00059, PI16/01292, and CB-161200301 CIBERONC from ISCIII (Spain), partially supported by European Regional Development FEDER funds. HEBCS: Helsinki University Hospital Research Fund, the Finnish Cancer Society and the Sigrid Juselius Foundation. HEBON: the Dutch Cancer Society grants NKI1998-1854, NKI2004-3088, NKI2007-3756, the Netherlands Organization of Scientific Research grant NWO 91109024, the Pink Ribbon grants 110005 and 2014-187.WO76, the BBMRI grant NWO 184.021.007/CP46 and the Transcan grant JTC 2012 Cancer 12-054. HEBON thanks the registration teams of Dutch Cancer Registry (IKNL; S. Siesling, J. Verloop) and the Dutch Pathology database (PALGA; L. Overbeek) for part of the data collection. ICO: The authors would like to particularly acknowledge the support of the Asociación Española Contra el Cáncer (AECC), the Instituto de Salud Carlos III (organismo adscrito al Ministerio de Economía y Competitividad) and “Fondo Europeo de Desarrollo Regional (FEDER), una manera de hacer Europa” (PI10/01422, PI13/00285, PIE13/00022, PI15/00854, PI16/00563 and CIBERONC) and the Institut Català de la Salut and Autonomous Government of Catalonia (2009SGR290, 2014SGR338 and PERIS Project MedPerCan). INHERIT: Canadian Institutes of Health Research for the “CIHR Team in Familial Risks of Breast Cancer” program – grant # CRN-87521 and the Ministry of Economic Development, Innovation and Export Trade – grant # PSR-SIIRI-701. IOVHBOCS: Ministero della Salute and “5x1000” Istituto Oncologico Veneto grant. kConFab: The National Breast Cancer Foundation, and previously by the National Health and Medical Research Council (NHMRC), the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia. MAYO: NIH grants CA116167, CA192393 and CA176785, an NCI Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201),and a grant from the Breast Cancer Research Foundation. MCGILL: Jewish General Hospital Weekend to End Breast Cancer, Quebec Ministry of Economic Development, Innovation and Export Trade. Marc Tischkowitz is supported by the funded by the European Union Seventh Framework Program (2007Y2013)/European Research Council (Grant No. 310018). MSKCC: the Breast Cancer Research Foundation, the Robert and Kate Niehaus Clinical Cancer Genetics Initiative, the Andrew Sabin Research Fund and a Cancer Center Support Grant/Core Grant (P30 CA008748). NCI: the Intramural Research Program of the US National Cancer Institute, NIH, and by support services contracts NO2-CP-11019-50, N02-CP-21013-63 and N02-CP-65504 with Westat, Inc, Rockville, MD. NNPIO: the Russian Foundation for Basic Research (grants 17-00-00171 and 18-515-45012). NRG Oncology: U10 CA180868, NRG SDMC grant U10 CA180822, NRG Administrative Office and the NRG Tissue Bank (CA 27469), the NRG Statistical and Data Center (CA 37517) and the Intramural Research Program, NCI. OSUCCG: Ohio State University Comprehensive Cancer Center. PBCS: Italian Association of Cancer Research (AIRC) [IG 2013 N.14477] and Tuscany Institute for Tumors (ITT) grant 2014-2015-2016. SMC: the Israeli Cancer Association. SWE-BRCA: the Swedish Cancer Society. UCHICAGO: NCI

Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA125183), R01 CA142996, 1U01CA161032 and by the Ralph and Marion Falk Medical Research Trust, the Entertainment Industry Fund National Women's Cancer Research Alliance and the Breast Cancer research Foundation. UCSF: UCSF Cancer Risk Program and Helen Diller Family Comprehensive Cancer Center. UPENN: Breast Cancer Research Foundation; Susan G. Komen Foundation for the cure, Basser Research Center for BRCA. UPITT/MWH: Hackers for Hope Pittsburgh. VFCTG: Victorian Cancer Agency, Cancer Australia, National Breast Cancer Foundation. WCP: Dr Karlan is funded by the American Cancer Society Early Detection Professorship (SIOP-06-258-01-COUN) and the National Center for Advancing Translational Sciences (NCATS), Grant UL1TR000124. HVH: Supported by the Carlos III National Health Institute funded by FEDER funds – a way to build Europe – PI16/11363.
MT Parsons is supported by a grant from Newcastle University. Kelly-Anne Phillips is an Australian National Breast Cancer Foundation Fellow.

Acknowledgements

BCAC acknowledgements
We thank all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out. ABCFS thank Maggie Angelakos, Judi Maskiell, Gillian Dite. ABCS thanks the Blood bank Sanquin, The Netherlands. ABCTB Investigators: Christine Clarke, Rosemary Balleine, Robert Baxter, Stephen Braye, Jane Carpenter, Jane Dahlstrom, John Forbes, Soon Lee, Debbie Marsh, Adrienne Morey, Nirmala Pathmanathan, Rodney Scott, Allan Spigelman, Nicholas Wilcken, Desmond Yip. Samples are made available to researchers on a non-exclusive basis. BBCS thanks Eileen Williams, Elaine Ryder-Mills, Kara Sargus. BCEES thanks Allyson Thomson, Christobel Saunders, Terry Slevin, BreastScreen Western Australia, Elizabeth Wylie, Rachel Lloyd. The BCINIS study would not have been possible without the contributions of Dr. K. Landsman, Dr. N. Gronich, Dr. A. Flugelman, Dr. W. Saliba, Dr. E. Liani, Dr. I. Cohen, Dr. S. Kalet, Dr. V. Friedman, Dr. O. Barnet of the NICCC in Haifa, and all the contributing family medicine, surgery, pathology and oncology teams in all medical institutes in Northern Israel. The BREOGAN study would not have been possible without the contributions of the following: Manuela Gago-Dominguez, Jose Esteban Castelao, Angel Carracedo, Victor Muñoz Garzón, Alejandro Novo Domínguez, Maria Elena Martinez, Sara Miranda Ponte, Carmen Redondo Marey, Maite Peña Fernández, Manuel Enguix Castelo, Maria Torres, Manuel Calaza (BREOGAN), José Antúnez, Máximo Fraga and the staff of the Department of Pathology and Biobank of the University Hospital Complex of Santiago-CHUS, Instituto de Investigación Sanitaria de Santiago, IDIS, Xerencia de Xestion Integrada de Santiago-SERGAS; Joaquín González-Carreró and the staff of the Department of Pathology and Biobank of University Hospital Complex of Vigo, Instituto de Investigacion Biomedica Galicia Sur, SERGAS, Vigo, Spain. BSUCH thanks Peter Bugert, Medical Faculty Mannheim. CBCS thanks study participants, co-investigators, collaborators and staff of the Canadian Breast Cancer Study, and project coordinators Agnes Lai and Celine Morissette. CCGP thanks Styliani Apostolaki, Anna Margiolaki, Georgios Nintos, Maria Perraki, Georgia Saloustrou, Georgia Sevastaki, Konstantinos Pompodakis. CGPS thanks staff and participants of the Copenhagen General Population Study. For the excellent technical assistance: Dorthe Uldall Andersen, Maria Birna Arnadottir, Anne Bank, Dorthe Kjeldgård Hansen. The Danish Cancer Biobank is acknowledged for providing infrastructure for the collection of blood samples for the cases. CNIO-BCS thanks Guillermo Pita, Charo Alonso, Nuria Álvarez, Pilar Zamora, Primitiva Menendez, the Human Genotyping-CEGEN Unit (CNIO). The CTS Steering Committee includes Leslie Bernstein, Susan Neuhausen, James Lacey, Sophia Wang, Huiyan Ma, and Jessica Clague DeHart at the Beckman Research Institute of City of Hope, Dennis Deapen, Rich Pinder, and Eunjung Lee at the University of Southern California, Pam Horn-Ross, Peggy Reynolds, Christina Clarke Dur and David Nelson at the Cancer Prevention Institute of California, Hoda Anton-Culver, Argyrios Ziogas, and Hannah Park at the University of California Irvine, and Fred Schumacher at Case Western University. DIETCOMPLYF thanks the patients, nurses and clinical staff involved in the study. The DietCompLyf study was funded by the charity Against Breast Cancer (Registered Charity Number 1121258) and the NCRN. We thank the participants and the investigators of EPIC (European Prospective Investigation into Cancer and Nutrition). ESTHER thanks Hartwig Ziegler, Sonja Wolf, Volker Hermann, Christa Stegmaier, Katja Butterbach. GC-HBOC thanks Stefanie Engert, Heide Hellebrand, Sandra Kröber and LIFE - Leipzig Research Centre for Civilization Diseases (Markus Loeffler, Joachim Thiery, Matthias Nüchter, Ronny Baber). The GENICA Network: Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, and University of Tübingen, Germany [HB, Wing-Yee Lo], German Cancer

Consortium (DKTK) and German Cancer Research Center (DKFZ) [HB], Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2180 - 390900677 [HB], Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany [YDK, Christian Baisch], Institute of Pathology, University of Bonn, Germany [Hans-Peter Fischer], Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany [Ute Hamann], Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany [Thomas Brüning, Beate Pesch, Sylvia Rabstein, Anne Lotz]; and Institute of Occupational Medicine and Maritime Medicine, University Medical Center Hamburg-Eppendorf, Germany [Volker Harth]. HABCS thanks Michael Bremer. HEBCS thanks Kirsimari Aaltonen, Irja Erkkilä. HUBCS thanks Shamil Gantsev. KARMA and SASBAC thank the Swedish Medical Research Counsel. KBCP thanks Eija Myöhänen, Helena Kemiläinen. kConFab/AOCS wish to thank Heather Thorne, Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study (which has received funding from the NHMRC, the National Breast Cancer Foundation, Cancer Australia, and the National Institute of Health (USA)) for their contributions to this resource, and the many families who contribute to kConFab. LMBC thanks Gilian Peuteman, Thomas Van Brussel, EvyVanderheyden and Kathleen Corthouts. MARIE thanks Petra Seibold, Dieter Flesch-Janys, Judith Heinz, Nadia Obi, Alina Vrieling, Sabine Behrens, Ursula Eilber, Muhabbet Celik, Til Olchers and Stefan Nickels. MBCSG (Milan Breast Cancer Study Group): Mariarosaria Calvello, Davide Bondavalli, Aliana Guerrieri Gonzaga, Monica Marabelli, Irene Feroce, and the personnel of the Cogentech Cancer Genetic Test Laboratory. The MCCS was made possible by the contribution of many people, including the original investigators, the teams that recruited the participants and continue working on follow-up, and the many thousands of Melbourne residents who continue to participate in the study. We thank the coordinators, the research staff and especially the MMHS participants for their continued collaboration on research studies in breast cancer. MSKCC thanks Marina Corines, Lauren Jacobs. MTLGEBCS would like to thank Martine Tranchant (CHU de Québec – Université Laval Research Center), Marie-France Valois, Annie Turgeon and Lea Heguy (McGill University Health Center, Royal Victoria Hospital; McGill University) for DNA extraction, sample management and skilful technical assistance. J.S. is Chair holder of the Canada Research Chair in Oncogenetics. NBHS and SBCGS thank study participants and research staff for their contributions and commitment to the studies. For NHS and NHS2 the study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required. We would like to thank the participants and staff of the NHS and NHS2 for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data. OFBCR thanks Teresa Selander, Nayana Weerasooriya. ORIGO thanks E. Krol-Warmerdam, and J. Blom for patient accrual, administering questionnaires, and managing clinical information. PBCS thanks Louise Brinton, Mark Sherman, Neonila Szeszenia-Dabrowska, Beata Peplonska, Witold Zatonski, Pei Chao, Michael Stagner. The ethical approval for the POSH study is MREC /00/6/69, UKCRN ID: 1137. We thank staff in the Experimental Cancer Medicine Centre (ECMC) supported Faculty of Medicine Tissue Bank and the Faculty of Medicine DNA Banking resource. RBCS thanks Jannet Blom, Saskia Pelders, Annette Heemskerk and the Erasmus MC Family Cancer Clinic. We thank the SEARCH and EPIC teams. SKKDKFZS thanks all study participants, clinicians, family doctors, researchers and

technicians for their contributions and commitment to this study. SZBCS thanks Ewa Putresza. UCIBCS thanks Irene Masunaka. UKBGS thanks Breast Cancer Now and the Institute of Cancer Research for support and funding of the Breakthrough Generations Study, and the study participants, study staff, and the doctors, nurses and other health care providers and health information sources who have contributed to the study. We acknowledge NHS funding to the Royal Marsden/ICR NIHR Biomedical Research Centre. We acknowledge funding to the Manchester NIHR Biomedical Research Centre (IS-BRC-1215-20007). The authors thank the WHI investigators and staff for their dedication and the study participants for making the program possible.

CIMBA acknowledgments

All the families and clinicians who contribute to the studies; Catherine M. Phelan for her contribution to CIMBA until she passed away on 22 September 2017; Sue Healey, in particular taking on the task of mutation classification with the late Olga Sinilnikova; Maggie Angelakos, Judi Maskiell, Gillian Dite, Helen Tsimiklis; members and participants in the New York site of the Breast Cancer Family Registry; members and participants in the Ontario Familial Breast Cancer Registry; Vilius Rudaitis and Laimonas Griškevičius; Drs Janis Eglitis, Anna Krilova and Aivars Stengrevics; Yuan Chun Ding and Linda Steele for their work in participant enrollment and biospecimen and data management; Bent Ejlertsen and Anne-Marie Gerdes for the recruitment and genetic counseling of participants; Alicia Barroso, Rosario Alonso and Guillermo Pita; all the individuals and the researchers who took part in CONSIT TEAM (Consorzio Italiano Tumori Ereditari Alla Mammella), in particular: Bernard Peissel, Dario Zimbalatti, Daniela Zaffaroni, Alessandra Viel, Giuseppe Giannini Liliana Varesco, Viviana Gismondi, Maria Grazia Tibiletti, Daniela Furlan, Antonella Savarese, Aline Martayan, Stefania Tommasi, Brunella Pilato and the personnel of the Cogentech Cancer Genetic Test Laboratory, Milan, Italy. Ms. JoEllen Weaver and Dr. Betsy Bove; FPGMX: members of the Cancer Genetics group (IDIS): Ana Blanco, Miguel Aguado, Uxía Esperón and Belinda Rodríguez; IFE - Leipzig Research Centre for Civilization Diseases (Markus Loeffler, Joachim Thiery, Matthias Nüchter, Ronny Baber); We thank all participants, clinicians, family doctors, researchers, and technicians for their contributions and commitment to the DKFZ study and the collaborating groups in Lahore, Pakistan (Muhammad U. Rashid, Noor Muhammad, Sidra Gull, Seerat Bajwa, Faiz Ali Khan, Humaira Naeemi, Saima Faisal, Asif Loya, Mohammed Aasim Yusuf) and Bogota, Colombia (Diana Torres, Ignacio Briceno, Fabian Gil). Genetic Modifiers of Cancer Risk in BRCA1/2 Mutation Carriers (GEMO) study is a study from the National Cancer Genetics Network UNICANCER Genetic Group, France. We wish to pay a tribute to Olga M. Sinilnikova, who with Dominique Stoppa-Lyonnet initiated and coordinated GEMO until she sadly passed away on the 30th June 2014. The team in Lyon (Olga Sinilnikova, Mélanie Léoné, Laure Barjhoux, Carole Verny-Pierre, Sylvie Mazoyer, Francesca Damiola, Valérie Sornin) managed the GEMO samples until the biological resource centre was transferred to Paris in December 2015 (Noura Mebirouk, Fabienne Lesueur, Dominique Stoppa-Lyonnet). We want to thank all the GEMO collaborating groups for their contribution to this study: Coordinating Centre, Service de Génétique, Institut Curie, Paris, France: Muriel Belotti, Ophélie Bertrand, Anne-Marie Birot, Bruno Buecher, Sandrine Caputo, Anaïs Dupré, Emmanuelle Fourme, Marion Gauthier-Villars, Lisa Golmard, Claude Houdayer, Marine Le Mentec, Virginie Moncoutier, Antoine de Pauw, Claire Saule, Dominique Stoppa-Lyonnet, and Inserm U900, Institut Curie, Paris, France: Fabienne Lesueur, Noura Mebirouk.Contributing Centres : Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon - Centre Léon Bérard, Lyon, France: Nadia Boutry-Kryza, Alain Calender, Sophie Giraud, Mélanie Léone. Institut Gustave Roussy, Villejuif, France: Brigitte Bressac-de-Paillerets,

Olivier Caron, Marine Guillaud-Bataille. Centre Jean Perrin, Clermont–Ferrand, France: Yves-Jean Bignon, Nancy Uhrhammer. Centre Léon Bérard, Lyon, France: Valérie Bonadona, Christine Lasset. Centre François Baclesse, Caen, France: Pascaline Berthet, Laurent Castera, Dominique Vaur. Institut Paoli Calmettes, Marseille, France: Violaine Bourdon, Catherine Noguès, Tetsuro Noguchi, Cornel Popovici, Audrey Remenieras, Hagay Sobol. CHU Arnaud-de-Villeneuve, Montpellier, France: Isabelle Coupier, Pascal Pujol. Centre Oscar Lambret, Lille, France: Claude Adenis, Aurélie Dumont, Françoise Révillion. Centre Paul Strauss, Strasbourg, France: Danièle Muller. Institut Bergonié, Bordeaux, France: Emmanuelle Barouk-Simonet, Françoise Bonnet, Virginie Bubien, Michel Longy, Nicolas Sevenet, Institut Claudius Regaud, Toulouse, France: Laurence Gladieff, Rosine Guimbaud, Viviane Feillel, Christine Toulas. CHU Grenoble, France: Hélène Dreyfus, Christine Dominique Leroux, Magalie Peysselon, Rebischung. CHU Dijon, France: Amandine Baurand, Geoffrey Bertolone, Fanny Coron, Laurence Faivre, Caroline Jacquot, Sarab Lizard. CHU St-Etienne, France: Caroline Kientz, Marine Lebrun, Fabienne Prieur. Hôtel Dieu Centre Hospitalier, Chambéry, France: Sandra Fert Ferrer. Centre Antoine Lacassagne, Nice, France: Véronique Mari. CHU Limoges, France: Laurence Vénat-Bouvet. CHU Nantes, France: Stéphane Bézieau, Capucine Delnatte. CHU Bretonneau, Tours and Centre Hospitalier de Bourges France: Isabelle Mortemousque. Groupe Hospitalier Pitié-Salpétrière, Paris, France: Chrystelle Colas, Florence Coulet, Florent Soubrier, Mathilde Warcoin. CHU Vandoeuvre-les-Nancy, France: Myriam Bronner, Johanna Sokolowska. CHU Besançon, France: Marie-Agnès Collonge-Rame, Alexandre Damette. CHU Poitiers, Centre Hospitalier d'Angoulême and Centre Hospitalier de Niort, France: Paul Gesta. Centre Hospitalier de La Rochelle : Hakima Lallaoui. CHU Nîmes Carémeau, France : Jean Chiesa. CHI Poissy, France: Denise Molina-Gomes. CHU Angers, France : Olivier Ingster; Ilse Coene en Brecht Crombez; Ilse Coene and Brecht Crombez; Alicia Tosar and Paula Diaque; Drs .Sofia Khan, Taru A. Muranen, Carl Blomqvist, Irja Erkkilä and Virpi Palola; The Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON) consists of the following Collaborating Centers: Coordinating center: Netherlands Cancer Institute, Amsterdam, NL: M.A. Rookus, F.B.L. Hogervorst, F.E. van Leeuwen, S. Verhoef, M.K. Schmidt, N.S. Russell, D.J. Jenner; Erasmus Medical Center, Rotterdam, NL: J.M. Collée, A.M.W. van den Ouweland, M.J. Hooning, C. Seynaeve, C.H.M. van Deurzen, I.M. Obdeijn; Leiden University Medical Center, NL: C.J. van Asperen, J.T. Wijnen, R.A.E.M. Tollenaar, P. Devilee, T.C.T.E.F. van Cronenburg; Radboud University Nijmegen Medical Center, NL: C.M. Kets, A.R. Mensenkamp; University Medical Center Utrecht, NL: M.G.E.M. Ausems, R.B. van der Luijt, C.C. van der Pol; Amsterdam Medical Center, NL: C.M. Aalfs, T.A.M. van Os; VU University Medical Center, Amsterdam, NL: J.J.P. Gille, Q. Waisfisz, H.E.J. Meijers-Heijboer; University Hospital Maastricht, NL: E.B. Gómez-Garcia, M.J. Blok; University Medical Center Groningen, NL: J.C. Oosterwijk, A.H. van der Hout, M.J. Mourits, G.H. de Bock; The Netherlands Foundation for the detection of hereditary tumours, Leiden, NL: H.F. Vasen; The Netherlands Comprehensive Cancer Organization (IKNL): S. Siesling, J.Verloop; the ICO Hereditary Cancer Program team led by Dr. Gabriel Capella; the ICO Hereditary Cancer Program team led by Dr. Gabriel Capella; Dr Martine Dumont for sample management and skillful assistance; Ana Peixoto, Catarina Santos and Pedro Pinto; members of the Center of Molecular Diagnosis, Oncogenetics Department and Molecular Oncology Research Center of Barretos Cancer Hospital; Heather Thorne, Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study (which has received funding from the NHMRC, the National Breast Cancer Foundation, Cancer Australia, and the National Institute of Health (USA)) for their contributions to this resource, and the many families who contribute to kConFab; the investigators of the Australia New Zealand NRG Oncology group; members and participants

in the Ontario Cancer Genetics Network; Leigha Senter, Kevin Sweet, Caroline Craven, Julia Cooper, Amber Aielts, and Michelle O'Conor; HVH : acknowledgments to the Cellex Foundation for providing research facilities and equipment.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique

Title: Novel strategies for the study of genetic factors associated with familial breast cancer.

Keywords : Breast cancer, Risk factors, SNPs, Interactions, Prediction

Abstract: One of the most important risk factors for breast cancer (BC) is having a family history of BC. Around 20% of the familial BC risk is explained by rare mutations in the genes *BRCA1* and *BRCA2* (*BRCA1/2*). An additional 30% of the risk is accounted for mutations in other known genes, like *ATM* or *TP53*, and by common genetic variants, called single nucleotide polymorphism (SNPs), identified in population-based GWAS. Therefore, the majority of the familial forms of BC remains unexplained. Furthermore, there are large variations in the estimation of the BC lifetime risk for *BRCA1/2* mutation carriers. It has been shown that some SNPs identified in the general population by GWAS (Genome Wide Association Studies) modified BC risk for *BRCA1/2* mutation carriers. Therefore, little is known on how these SNPs interact with *BRCA1/2* mutations since association studies have been performed within the population of *BRCA1/2* mutation carriers so far.

In the first part of this PhD project, I developed a novel strategy to analyze genetic factors by integrating simultaneously environmental and lifestyle factors. This strategy was used to analyze the data of GENESIS study composed of pairs of sisters affected by BC without *BRCA1/2* mutation and controls from the general population. 5,000 BC cases and controls were genotyped for the 200,000 SNPs targeted by the iCOGS array. Groups of subjects were created according to their exposition profile reflecting expositions to radiation or reproductive factors. Analyses stratified on groups built according to their reproduction factors exposures did not highlighted specific variants. However, analyses stratified on groups reflecting the chest X-ray exposures showed potential specific SNPs for women who had never been exposed to chest X-ray, in genes *XRCC4* and *MAGI1*, and for women highly exposed to X-ray exposures, in gene *FGFR2*, already known in the general population.

The second aim was to identify and characterize genetic modifiers of BC risk for *BRCA1/2* mutation carriers using data from the international consortia CIMBA (Consortium of Investigators of Modifiers of *BRCA1/2*) and BCAC (Breast Cancer Association Consortium). I developed a *case-only GWAS* analysis where we compare genotype frequencies between 60,212 unselected BC cases from the BCAC and 13,007 BC cases from CIMBA. We identified 4 novel variants associated with BC for *BRCA1* mutation carriers and 4 for *BRCA2* mutation carriers at $P<10^{-8}$. *MADD*, *SPI1* and *EIF1* genes, already associated with BC biology, was predicted by the tool INQUISIT, to be target genes of the potential causal variants located in the locus 11p11.2 associated with BRCA1 status.

These new SNPs could be used to improve polygenic risk scores (PRS). Studies considering the exposure profile should be implemented in larger population. The models could then evolve towards an adaptation of the PRS according to women's exposure profiles and that throughout their life.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique

Titre : Nouvelles stratégies pour l'étude des facteurs génétiques impliqués dans le cancer du sein familial.

Mots clés : Cancer du sein, Facteurs de risque, SNPs, Interactions, Prédiction

Résumé : Avoir une apparentée atteinte d'un cancer du sein (CS) multiplie par 2 le risque de développer un CS. Environ 20 % du risque familial de CS est attribué à une mutation fortement pénétrante dans les gènes *BRCA1* ou *BRCA2*. D'autres gènes connus, comme *ATM* ou *TP53*, ainsi que des variants génétiques fréquents (SNP) identifiés dans des études pangénomiques (GWAS) réalisées en population générale, expliqueraient 30 % supplémentaires des cas familiaux. La majorité des formes familiales de CS restent donc inexpliquées. Par ailleurs, il existe de fortes variations du risque parmi les porteurs d'une mutation *BRCA1/2*. Il a été montré que certains SNP identifiés dans les *GWAS* modifient leur risque. Cependant, l'effet propre de ces SNP n'a pu être estimé puisque ces études ont été réalisées chez les porteurs de mutation *BRCA1/2*.

Dans une première partie de ma thèse, j'ai développé une stratégie intégrant aux analyses sur les facteurs génétiques des facteurs « environnementaux ». Cette stratégie a été utilisées pour analyser les données de l'étude GENESIS qui inclut des paires de sœurs atteintes de CS et sans mutation *BRCA1/2* et des témoins de population générale. Elle regroupe 5 000 cas de CS et témoins génotypés pour les 200 000 SNP de la puce iCOGS. Les femmes de GENESIS ont été réparties dans des groupes selon leur profil d'expositions aux radiations ou aux facteurs gynéco-obstétriques. Alors que l'analyse stratifiée sur les groupes construits à partir des expositions aux facteurs gynéco-obstétriques ne nous permet pas de mettre en évidence de potentiels SNPs spécifiques, l'analyse stratifiée sur les groupes construits à partir des expositions aux radiations a permis de mettre en évidence des SNPs spécifiques potentiels aux femmes non exposées, dans les gènes *XRCC4* et *MAGI1*, et à celle fortement exposées aux radiations, dans le gène *FGFR2*, déjà trouvés en population générale.

Le deuxième objectif de ma thèse visait à optimiser la caractérisation des gènes *BRCA1/2* en étudiant leurs interactions avec des SNPs modificateurs à partir des données des consortia internationaux CIMBA (Consortium of Investigators of Modifiers of BRCA1/2) et BCAC (Breast Cancer Association Consortium). J'ai développé une stratégie d'analyse *GWAS case-only*, comparant la fréquence de 60 212 cas de cancer du sein de la population générale (BCAC) et 13,007 cas porteurs d'une mutation *BRCA1/2* (CIMBA). J'ai identifié 4 nouvelles régions associés au CS chez les femmes porteuses d'une mutation *BRCA1* et 4 autres chez les femmes porteuses d'une mutation *BRCA2*. Les gènes *MADD*, *SPI1* et *EIF1*, déjà associées à la biologie du cancer du sein dans d'autres études, ont été prédits par l'outil INQUISIT comme étant des gènes cibles des potentiels variants causaux se trouvant dans la région 11p11.2 associée au statut BRCA1.

Ces nouveaux SNPs mis en évidence pourraient être utilisés pour améliorer les prédictions de risque des PRS (Polygénic Risk Score). Les analyses prenant en compte des profils d'exposition devraient être poursuivies sur des études de grande dimension. Les modèles pourraient alors évoluer vers une adaptation des PRS en fonction des profils d'expositions des femmes et cela tout au long de leur vie.