Modélisation des arbres onco-généalogiques et application à la détermination de phénotypes cancéreux spécifiques favorisant une exploration génotypique ciblée

Fabrice Kwiatkowski

ÉCOLE DOCTORALE
DES SCIENCES FONDAMENTALES
Université Clermont Auvergne

ECOLE DOCTORALE DES SCIENCES FONDAMENTALES

Thèse

Présentée à l'Université Clermont Auvergne

pour l'obtention du grade de DOCTEUR
(Décret du 5 juillet 1984)

Spécialité
Mathématiques appliquées et applications des mathématiques

Soutenue le 2 octobre 2020

KWIATKOWSKI Fabrice

Modélisation des arbres onco-généalogiques et application à la détermination de phénotypes cancéreux spécifiques favorisant une exploration génotypique ciblée

---

Directeur de thèse : Pr Laurent SERLET
Co-directeur de thèse : Pr Yves-Jean BIGNON
Rapporteurs : Pr Jean-Christophe THALABARD (PU-PH, Université Paris-Descartes),
Pr Sergueï DACHIAN (Université de Lille)
Examinateurs : Elisabeth de TURCKHEIM (DR INRA), Mathilde GAY-BELLILE (PhD-MD, Centre Jean Perrin), Stéphanie LEGER (MCF UCA), David PEROL (MD, Centre Léon Bérard, Lyon)

Laboratoire de Mathématique Blaise Pascal

Remerciements

A mon directeur de thèse, le Professeur Laurent SERLET : mes sincères remerciements pour ton encadrement de ce projet peu conventionnel et pour l'important travail que tu as fourni et qui constitue une partie non négligeable de ce manuscrit. Je te dois une dose supplémentaire de remerciements pour avoir supporté durant ces longues années mes errements théoriques sans jamais broncher. J'ai aussi bénéficié de la chance insigne que nous partagions une forme d'humour décapante, dont la limite tend vers l'infini. Sans ta patience et ta bonne humeur, ce travail n'aurait jamais abouti.

A mon co-directeur de thèse, le Professeur Yves-Jean BIGNON : durant toutes ces années à travailler ensemble au Centre Jean Perrin, je ne saurais te remercier suffisamment de m'avoir constamment épaulé dans mes recherches, de m'avoir fait confiance jusqu'au bout… et sans doute parfois, d'avoir su freiner mon enthousiasme. Notre amitié est le résultat de cette longue aventure scientifique.

A Andrzej STOS, maître de conférence à l'UCA : merci à toi aussi de m'avoir accompagné avec Laurent dans ce long cheminement et d'avoir réalisé de nombreux développements informatiques sur des données généalogiques récalcitrantes et qui constituent une partie innovante de cette thèse. J'ai beaucoup apprécié ta gentillesse et ton aide généreuse.

Au Professeur Jean-Christophe THALABARD de l'Université Paris-Descartes et au Professeur Sergueï DACHAIN de l'Université de Lille. Je vous sais gré d'avoir accepté le difficile rôle de rapporteur pour cette thèse. Je vous adresse mes sincères remerciements pour avoir consacré une partie de votre précieux temps à examiner ce manuscrit en profondeur et, grâce à vos compétences en la matière, d'avoir suggéré des voies d'amélioration et des perspectives à cette recherche. Un grand merci aussi aux autres membres du jury, Elisabeth De TURKHEIM, directeur de recherche à l'INRA, Stéphanie LEGER, maître de conférence à l'UCA et le Dr David PEROL, directeur de la recherche clinique et de l'innovation au CLCC Léon Bérard à Lyon pour leur aimable participation à la critique de ce travail de recherche.

A mes collègues de l'équipe oncogénétique du Centre Jean Perrin, Mathilde GAY-BELLILE, Maud PRIVAT, Yannick BIDET et Mathias CAVAILLE, et ceux du service de génétique clinique du CHU, le Dr Isabelle PERTHUS et le Pr Andreï TCHIRKOV pour leur gentillesse, leurs qualités humaines, mais aussi bien évidemment leur contribution aux analyses génétiques des familles étudiées dans les modèles sans lesquelles une thèse de mathématique ne saurait avoir de validation définitive.

Dédicaces

A Dominique, mon épouse, et mes trois fils Ivan, Nicolas et Robin. Vous avez donné à ma vie cette dimension affective épanouissante sans laquelle elle n'aurait ni couleur ni saveur. Je suis chaque jour admiratif et curieux de votre devenir dans ce monde en profond changement. Puissiez-vous contribuer à le faire avancer dans une direction plus respectueuse de la vie.

A mes Directeurs du Centre Jean Perrin, les professeurs Robert PLAGNE et Jacques DAUPLAT qui, sans le dire, m'ont protégé et m'ont permis de faire évoluer ma carrière hors des sentiers battus. Ma gratitude aussi envers Jean-Pierre FERRIERE : sans toi, je n'aurais jamais embrassé le métier de statisticien. J'inclus aussi dans ces personnes qui ont contribué à mon évolution, Jacques BERLIE, oncologue au Centre René Huguenin, pour m'avoir encouragé dans le développement de SEM et pour ton amitié ; les statisticiens des CLCC Bernard ASSELAIN, Andrew KRAMAR, Xavier PAOLETTI et d'autres, toujours présents quand j'avais besoin d'aide. Cette thèse est aussi le fruit de leur générosité et de leur grande connaissance mais aussi de leur amitié sans faille. Au professeur Francis Lévi, oncologue à l'Hôpital Paul Brousse, chercheur infatigable et intransigeant : je te dois mes premiers pas dans la chronobiologie et les premières publications scientifiques qui m'ont mis le pied à l'étrier.

A Angeline, Judith, Emilie, ex-doctorantes, et à mes autres collègues pour avoir cocooné le papy pendant de longues années…

Aux défricheurs de l'inconnu, Patrice BOUCHARDON, Paul MELLERET…

Que gagne-t-on à rester dans sa zone de confort ? Un sommeil plus profond ?

Takuan Soho – moine Zen (Japon, 1573-1645)

Table des matières

Glossaire :

NB. Les mots soulignés dans le texte sont définis dans ce glossaire.

ADN :	Acide désoxyribonucléique. Macromolécule biologique en forme de double hélice constituée de bases nucléiques, ou bases azotées - adénine (A), cytosine (C), guanine (G) ou thymine (T) – et assemblés en colonne grâce à un composé « sucre – phosphate ».
Cardinal :	Effectif d'une population
CCI :	Cancer canalaire invasif (sein)
CCIS :	Cancer canalaire *in-situ* (sein), de bon pronostic
Censure :	Donnée correspondant à un état susceptible de changer dans le temps, comme l'occurrence d'une maladie (cancer ici) à un âge déterminé. Tant que la maladie n'est pas arrivée, l'information n'est que partielle est n'est valide que jusqu'au moment du point qui est fait. La censure se termine dès que la maladie arrive ou bien quand l'individu décède en en étant indemne.
Centroïde :	Centre de gravité ou barycentre
CLI :	Cancer lobulaire invasif (sein)
CLIS :	Cancer lobulaire *in-situ* (sein), de bon pronostic
Epigénétique :	L'épigénétique correspond à l'étude des changements dans l'activité des gènes, n'impliquant pas de modification de la séquence d'ADN et pouvant être transmis lors des divisions cellulaires. Contrairement aux mutations qui affectent la séquence d'ADN, les modifications épigénétiques sont réversibles [INSERM, 2015][1].
Fécondité :	Résultat de l'activité de reproduction d'une personne, d'un couple, d'un groupe ou d'une population [Haupt, 2004][2]
Fertilité :	Capacité physiologique d'une femme, d'un homme ou d'un couple à avoir un enfant vivant [Haupt, 2004]
Gène :	Unité de base d'hérédité, correspondant à une séquence déterminée d'ADN qui prédétermine une caractéristique précise d'un organisme vivant
Génome :	Ensemble du matériel génétique d'une espèce dont le code est défini dans son ADN
Génotype :	Ensemble des informations portées par le génome d'un organisme et que l'on retrouve dans chaque cellule.
Hétérozygote :	Se dit d'une mutation génétique portée par seulement un allèle du gène, laissant à ce gène une partie de sa fonction active.
Homozygote :	Se dit d'une mutation portée par les deux allèles du gène.
In-situ :	se dit d'un cancer restant limité localement à un type particulier de cellules et ne contaminant pas les tissus adjacents

LOM :	Laboratoire d'oncologie moléculaire du Centre Jean Perrin
Microbiote :	Ensemble des micro-organismes (bactéries, levures, champignons, virus) vivant dans un environnement spécifique (tube digestif, poumons, poils…) chez un hôte (animal ou végétal).
Oncogénétique :	Branche de la médecine et de la biologie qui s'occupe des cancers favorisés par des caractéristiques génétiques (ou épigénétiques). Elle inclut aujourd'hui la caractérisation génétique des tumeurs, à la recherche de potentielles cibles thérapeutiques et s'il y en a, des thérapies ciblées correspondantes.
Pénétrance :	Proportion d'individus possédant un génotype donné qui exprime le phénotype correspondant. Par exemple, 70% des femmes européennes porteuses de mutation sur les gènes BRCA ont un cancer du sein au cours de leur vie.
Phénotype :	Ensemble des caractéristiques observables d'un individu
Polymorphisme :	Existence au sein d'une population de variations individuelles d'un caractère génétique n'entraînant pas de conséquences pathologiques[3].
ROC	Receiver Operating Characteristics. Se dit d'une courbe traçant la sensibilité en ordonnées et l'anti-spécificité en abscisses d'un critère de jugement quantitatif.
Sporadique :	Se dit d'un cancer qui se développe sans que l'on ait pu déterminer des facteurs de risque, en particulier génétiques.
Taux de fécondité :	à un âge donné ou pour une tranche d'âge = nombre d'enfants nés vivants de femmes de cet âge au cours d'une période (généralement l'année), rapporté au nombre de femmes de même âge sur la même période.

Table des illustrations

1 Introduction

1.1 Contexte

L'oncogénétique est la branche de la médecine qui s'intéresse aux cancers liés à des prédispositions héréditaires, donc pour l'essentiel des prédispositions transmises via des mutations génétiques. C'est une discipline très récente puisqu'en 1988 au Centre Jean Perrin, le Pr Yves-Jean BIGNON a mis en place la première consultation oncogénétique de France. Depuis cette époque, des unités identiques se sont créées dans les grandes villes de France. On peut affirmer aujourd'hui que le service rendu de ces consultations d'oncogénétique est majeur car il permet, via une surveillance optimale des individus à risque, de diagnostiquer de manière très précoce les cancers associés aux prédispositions et ainsi d'augmenter de manière considérable leurs chances de rémission complète suite aux traitements. Par ailleurs, on peut aujourd'hui parler de prévention puisque l'on sait, suite aux diverses études épidémiologiques et aux essais cliniques réalisés, que certaines mesures de prévention (hygiène de vie, activité physique, nutrition, non consommation de substances addictives, allaitement…) permettent de diminuer sensiblement le risque de cancer des porteurs de mutation. Ces recommandations rejoignent indirectement les conclusions d'une étude de 2003 [4], mettant en évidence un risque de cancer du sein à 50 ans multiplié par 2.5 pour les femmes nées après 1940 par rapport à celles nées avant 1940 et pour laquelle les auteurs évoquaient déjà le surpoids et l'exercice physique. Accessoirement, ces mêmes mesures procurent des avantages similaires dans la population générale, même si le risque est moindre, et permettent une diminution de l'incidence tant des cancers que des maladies cardiovasculaires.

Qu'est qu'une mutation délétère ? Naturellement, un même gène peut avoir des formes variables que l'on appelle des polymorphismes, formes qui n'altèrent pas ou peu sa fonction (traduction en ARN, production de protéines ou rôle de messager). Le diagnostic d'une mutation délétère chez un patient indique que la fonction associée au gène est altérée chez lui : pour la cancérogenèse, il s'agit le plus souvent de fonctions réparatrices de l'ADN, des cassures simple ou double brin qui interviennent aléatoirement suite à des erreurs lors des mitoses cellulaires mais aussi suite aux agressions répétées de l'environnement (rayons ultraviolets du soleil, radiations ionisantes, etc.) ou plus internes comme le stress oxydatif. Dans d'autres cas, il s'agit de l'altération de la fonction apoptotique, celle qui induit la mort des cellules quand elles ne sont plus fonctionnelles. On peut enfin citer, sans que cette liste soit exhaustive, les gènes impliqués dans le maintien du télomère, sorte de permis à point des cellules qui en limite le nombre de divisions, phénomène qui participe au vieillissement physiologique.

Combien de personnes ces mutations concernent-elles ? Les mutations délétères les plus connues comme celles sur les gènes BRCA concernent moins d'un individu sur mille dans nos pays (sauf populations particulières comme les juifs ashkénazes où la prévalence de ces mutations pourrait atteindre les 10% selon Stoppa-Lyonnet [2004] [5]. Ces anomalies sont donc plutôt rares. Mais aujourd'hui, on pense que peut-être des mutations multiples sans effet individuel notable pourraient jouer un rôle grâce à une synergie entre les gènes concernés, un peu comme si des anomalies mineures agissaient de manière cumulative voire multiplicative sur un plan pathologique. Du coup, cela concernerait beaucoup plus de monde d'autant plus si des expositions à des facteurs de risque (habitudes de vie, pollution, rayons ionisants…) s'ajoutent à ce terrain génétique. Pour donner une dernière illustration quantitative de notre problématique, la base de données du service

oncogénétique du Centre Jean Perrin qui concerne essentiellement l'Auvergne (1,36 millions d'habitants), contient actuellement plus de 9 000 familles, incluant environ 200 000 personnes, bien sûr dont de nombreuses sont décédées il y a longtemps. Cela donne une idée de l'impact non négligeable de ces mutations.

La recherche des prédispositions héréditaires de cancer dans une famille consiste principalement en la construction et l'analyse de l'arbre généalogique familial, sur plusieurs générations, en précisant pour chaque membre connu le sexe et l'année de naissance et les paramètres cliniques pertinents : cancers (type et âge de diagnostic), âge de décès et cause si connue, pathologies principales, fausses couches, voire expositions à des facteurs de risque... Actuellement ces arbres sont le plus souvent analysés visuellement par l'onco-généticien. A l'issue de cette analyse, selon la prédisposition familiale diagnostiquée (syndrome familial), le médecin peut engager une recherche de mutation génétique ciblée réalisée à partir de prélèvements sanguins ou buccaux : actuellement, au vu du syndrome identifié, le généticien limite le séquençage à quelques gènes parmi une centaine de gènes de susceptibilité. Il est prévisible que cette limitation ne soit plus maintenue dans un avenir proche, compte tenu des performances toujours plus grandes des automates et des techniques de séquençage.

Etrangement aujourd'hui, cette approche ne permet de déterminer les gènes impliqués dans le syndrome familial que pour 20% environ des familles. C'est-à-dire que pour 80% des familles, aucune anomalie génétique connue ne peut être diagnostiquée. Cette méconnaissance des mutations responsables du risque de cancer pour une famille n'est pas une bonne nouvelle : en premier lieu, elle n'en diminue pas moins le risque pour ses membres. En second lieu, ne connaissant pas les mutations délétères impliquées, on ne peut pas savoir, grâce à de nouveaux prélèvements chez les descendants, lesquels sont ou ne sont pas porteurs de ces mutations et donc exposés ou pas au risque. Dans les autres familles où la mutation est connue, chaque descendant a une chance sur 2 de ne pas être porteur et donc d'échapper au risque familial, ce que l'on peut déterminer précisément pour chacun.

Il reste donc à identifier pour ces 80% de familles quels polymorphismes pourraient candidater au titre de mutation délétère, voie quelles associations de polymorphismes pourraient constituer de nouveaux groupements délétères. Les tenants et aboutissants d'un tel problème sont malheureusement complexes : notre génome contient environ 20 000 gènes correspondant à approximativement 3,2 milliards de paires de nucléotides (les fameuses lettres ATGC pour adénine, thymine, guanine et cytosine). Chercher de nouvelles mutations délétères ou pire, des associations délétères de polymorphismes, forcément moins pénétrantes que les mutations connues, c'est donc un peu chercher une aiguille dans une botte de foin… En particulier, les variables d'intérêt étant nettement plus nombreuses que les échantillons (familles, individus), le risque d'attribuer la « faute » à des faux positifs est dramatiquement élevé. Hélas ! L'expression des gènes n'est pas seulement perturbée du fait de mutations : des altérations épigénétiques, héréditaires sur seulement quelques générations, peuvent aussi temporairement bloquer ou dynamiser anormalement l'expression des gènes correspondants. Et c'est sans compter l'ADN des mitochondries, ces petites usines à énergie contenues dans nos cellules, qui affichent 37 gènes pour 16 500 paires de nucléotides. Bien évidemment, il faudrait aussi tenir compte des bactéries et des virus bactériophages de nos flores intestinale, pulmonaire, dermique… dont les représentants sont à peu près 10 fois plus nombreux que les cellules de notre corps et dont le génome réunit à peu près 100 fois plus de gènes que le génome humain[6]. Ce que l'on commence à connaître des interactions hôte-microbiotes laissent en effet penser qu'une part de l'étiologie des cancers pourrait leur être indirectement imputée[7,8]. Mais l'oncogénétique n'est pas concernée par ces derniers représentants… jusqu'à présent. Ouf !

1.2 Objectifs de la thèse

L'objectif général de cette thèse concerne donc les 80% de familles sans mutation délétère connue. Que peut-on faire pour elles ? Nous avons signalé précédemment que chercher tous azimut dans notre génome des polymorphismes potentiellement délétères, est une tâche quasi-impossible : on pense aujourd'hui qu'on ne découvrira plus de gènes d'importance majeure pour l'étiologie des cancers. Certes, l'avènement du « big data » avec l'accumulation des données issues de milliers de patients apportera probablement de nouvelles connaissances, mais les outils de datamining se heurtent à des limites conceptuelles qui, à notre avis, ne seront pas franchies avant des décennies. Une autre approche, sans doute plus modeste et moins dispendieuse, se propose de partir des phénotypes familiaux eux-mêmes – matérialisés par les arbres généalogiques – et suite à leur modélisation et à l'aide d'outils mathématiques et/ou statistiques et/ou algorithmiques, de tenter d'effectuer des regroupements de familles à risque pour lesquelles aucune mutation n'est retrouvée. Avec de tels regroupements, caractérisant des sous-groupes à risque spécifique, on pourrait effectuer chez les individus leur appartenant des analyses génétiques avec plus de chances de mettre en évidence des mutations particulières à moindre pénétrance voire des associations délétères de polymorphismes. Manipuler les arbres familiaux à l'aide d'outils informatiques dédiés pourrait vraisemblablement faciliter tout le travail préparatoire.

1.3 Synoptique de la recherche

Pour parvenir à l'objectif de rendre les arbres généalogiques informatiquement manipulables, plusieurs étapes peuvent être définies :

1. Il faut en premier lieu identifier les données pertinentes de la structure d'arbre généalogique et des mécanismes génétiques sous-jacents, en particulier les données qui caractérisent un phénotype.

2. Développer des méthodes de simulation d'arbres généalogiques proches de la réalité : cela permettrait de tester les approches sur des arbres aux caractéristiques parfaitement connues.

3. Rechercher et/ou créer des algorithmes de prédiction des génotypes au vu des phénotypes individuels ou familiaux : implémentation en langage Python, Doc.Net ou autre.

4. Tester les algorithmes sur des arbres simulés

5. Importer les arbres généalogiques de la base de données du Centre Jean Perrin sous une forme compatible (et anonyme). Discuter des familles possédant des branches ascendantes multiples et des risques combinés.

6. Calibrer les modèles efficaces sur les données simulées, sur les données réelles et comparer les algorithmes dans cet environnement.

7. Utiliser les modèles pour définir des sous-groupes à risque spécifique de cancer : utiliser ces regroupements pour séquencer certains gènes d'intérêt et rechercher des mutations récurrentes.

8. Créer une interface utilisateur pour un outil de diagnostic individuel, pouvant inclure aussi les risques non-héréditaires

1.4 Organisation de la thèse

Pour mener à bien ce travail, nous allons dans un premier temps développer notre approche pour simuler des données de test, c'est à dire des familles "réalistes" ayant des caractéristiques assez proches des familles (réelles) vues en consultation en oncogénétique. Ensuite, le concept de sous-arbre – structure d'arbre généalogique réduite au maximum – va être présenté, lui qui est censé permettre la comparaison et l'agrégation des arbres généalogiques. L'intérêt de ces sous-arbres va être évalué de deux manières :

- dans la partie 4, en utilisant l'analyse en composantes principales et le k-means clustering pour partitionner les familles. On pourra vérifier si le résultat du partitionnement est représentatif ou non par l'état mutationnel paramétré dans nos familles simulées et si ce résultat est plus efficace qu'avec les caractéristiques moyennées par familles.
- Dans la partie 5, en utilisant la classification hiérarchique ascendante et effectuer les mêmes comparaisons que ci-dessus.

Dans la partie 6, d'autres approches seront utilisées, basées sur l'utilisation d'arbres simulés moins élaborés, la première tentant de prédire les mutations à partir d'un algorithme de minimisation de distance et l'autre en utilisant des réseaux de neurones.

2 Génération d'arbres généalogiques de test

Evaluer des approches algorithmiques ou mathématiques sur des arbres généalogiques réels, mais pour lesquels on ne connaît pas toutes les caractéristiques sous-jacentes est forcément risqué : il est effectivement possible que les approches soient valides mais que les données soient au final inappropriées pour confirmer la pertinence des approches. D'un autre côté, des données particulières pourraient paraître valider les approches, mais en réalité en raison d'artefacts ou encore de non-représentativité des familles sélectionnées. L'utilisation d'arbres généalogiques parfaitement modélisés, c'est à dire avec des caractéristiques sûres bien qu'aléatoirement distribuées, est donc un préliminaire indispensable pour tester de nouvelles méthodes.

Nous avons donc réalisé un ensemble de routines qui permettent d'une part de générer ces arbres test et de l'autre, de les représenter, ceci afin d'avoir un contrôle visuel des données générées. En complément, des analyses statistiques ont été implémentées afin de s'assurer de l'adéquation entre les caractéristiques paramétrées en entrée et celles que l'on retrouve en sortie. Dans les paragraphes suivants, nous allons décrire ces trois types de fonction.

2.1 Programme POLYGENE générant les arbres généalogiques

Cette routine a été programmée en VB.NET de Microsoft, version 2010. L'intérêt de ce langage est de permettre la réalisation très aisée d'interfaces visuelles grâce à la mise à disposition de composants directement utilisables (boutons commandes, listes déroulantes, cases à cocher, glissières, boites graphiques…). Il est très adapté à la programmation événementielle et permet tous les calculs scientifiques, même s'il n'est pas optimisé pour cela. C'est un des langages les plus utilisés au monde

sans doute parce qu'il est l'héritier du BASIC des années 1960 (*Beginner's All-purpose Symbolic Instruction Code*) popularisé par Microsoft comme compagnon de son système d'exploitation DOS (*Disk Operating System*) et qui permettait aux premiers micro-ordinateurs de fonctionner.

A l'origine, c'est à dire en 2006, nous avons développé un programme informatique suite à la demande du Pr BIGNON : à l'époque, le rôle de mutations sur les gènes BRCA1 et BRCA2 dans l'oncogenèse étaient largement décrits. Toutefois, ces mutations étaient responsables de moins de 20% des prédispositions familiales. Par ailleurs, la probabilité de découvrir d'autres mutations à forte pénétrance diminuait à mesure que les recherches infructueuses sur de larges panels de gènes s'accumulaient. L'hypothèse la plus réaliste était donc que plusieurs mutations plus ou moins fréquentes en population générale, sur d'autres gènes que BRCA, pourraient agir de manière synergique. Indépendamment les unes des autres, elles n'auraient guère d'effet délétère, mais dès qu'elles se retrouveraient groupées chez certains individus/familles, elles cumuleraient les vulnérabilités associées et résulteraient en un impact carcinogénétique non négligeable.

Une approche Monte-Carlo a été choisie pour évaluer ces interactions entre mutations à faible pénétrance. Les caractéristiques des populations cibles ont été recherchées dans la littérature puis implémentées dans un programme nommé POLYGENE. Une présentation de ces développements a été faite lors du congrès de statistique et d'épidémiologie clinique à Bordeaux en 2014[9]. Quels étaient ces paramètres populationnels nécessaires au développement de POLYGENE ? Certains étaient à rechercher dans la biologie, caractéristiques de l'espèce humaine tandis que d'autres étaient contextuels, pour une part environnementaux, naturellement, mais pour une autre large part civilisationnels du fait de l'importance des coutumes permettant la vie en société chez l'homme. On peut les schématiser dans le synoptique suivant :

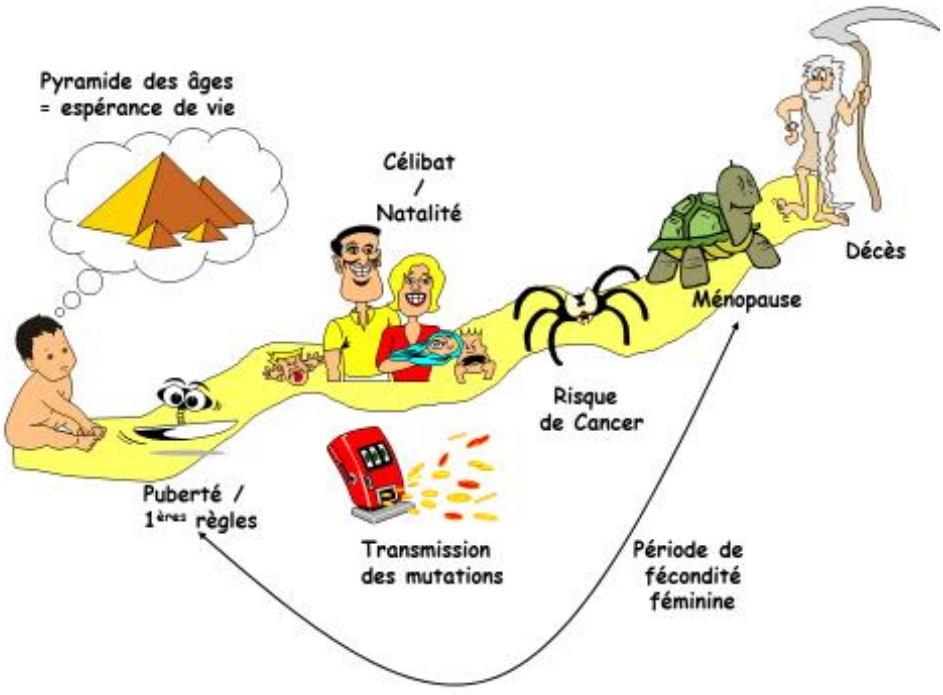


Figure 1 : synoptique des « moments » de la vie à paramétrer pour générer des arbres généalogiques

2.1.1 L'influence du type de civilisation

Quand un enfant né, selon le contexte qui préside à sa naissance, les étapes de sa vie sont quasiment prédestinées : cela se traduit en particulier par une espérance de vie caractéristique de la civilisation qui le voit naître, mais aussi par une chronologie des diverses étapes avec une probabilité que l'on peut leur attacher selon le temps. Voilà qui intéresse particulièrement le statisticien. L'espérance de vie, la fécondité, l'âge du début de la reproduction, voire des aspects comme la polygamie ou la polyandrie, varient de manière conséquente selon que l'on voit le jour dans un pays occidental comme actuellement en Europe, ou bien dans un pays dit « en voie de développement », tel certaines contrées africaines ou des groupes ethniques minoritaires comme en forment les indiens de la forêt amazonienne au Brésil. Mais on pourrait aussi réfléchir aux inégalités sociales qui, chez nous, font varier significativement l'espérance de vie. Ainsi un cadre supérieur de 60 ans bénéficie-t-il de 4 ans de vie supplémentaires en moyenne par rapport à un ouvrier[10]. Dans nos simulations, nous ne retiendrons toutefois pas ces différences fines. Trois types de contexte seront donc paramétrés et utilisés : primitif, sous-développé et développé.

Le contexte primitif (espérance de vie réduite, début de reproduction précoce, natalité importante associée à une mortalité infantile élevée) a été constitué pour évaluer si un tel contexte induisait ou non une pression de sélection sur des mutations délétères favorisant le cancer. De ces dernières, on en connaît de très anciennes qui ont perduré pendant des millénaires jusqu'à nos jours[11, 12, 13]. On pourrait penser *a priori* que les cancers liés à ces mutations n'avaient généralement pas le temps de se produire avant le délai attendu de décès. En réalité, cet argument est spécieux : on oublie souvent que ce qui pénalisait le plus fortement l'espérance de vie à ces époques, c'était la mortalité infantile. Mais pour bon nombre des personnes sortant de la prime enfance, la durée de vie dépassait fréquemment les 50 ans : par exemple et même si cela ne correspond pas à un contexte « primitif », en 1750 en France plus de la moitié des enfants de 5 ans atteignaient la cinquantaine[14] alors que l'espérance de vie à la naissance était de 29.0 ans et 27.5 ans respectivement pour les femmes et les hommes[15]. Pour la période prémoderne en occident, les conditions d'hygiène et la mortalité périnatale tant des enfants que des mères lors de l'accouchement grèvent fortement l'espérance de vie. Dans les pays en voie de développement, ce n'est que relativement récemment que cela a été corrigé suite à la généralisation de moyens médicaux modernes : les pyramides des âges de ces derniers pays se distinguent de celles des contextes primitifs du fait d'un taux nettement diminué de décès infantiles.

La recherche d'information concernant la démographie des temps anciens est facilitée par le fait que les changements économiques et sociétaux n'ont pas concerné tous les pays en même temps, loin s'en faut. Ainsi, les données relatives à nos pays européens aujourd'hui obsolètes sont-elles restées valides pour des pays dont les conditions n'ont guère changé pendant des siècles. Voici donc les études qui nous ont servi pour définir des estimateurs démographiques pour les sociétés primitives, sous-développées et développées :

- INED, l'institut national des études démographiques [Vallin, 1999] pour la France et l'Angleterre (1750 - 1997).
- GLOBOCAN par exemple pour l'Afrique saharienne en 2002 [IARC, 2002][16].
- Enquête PROVIDE pour l'Afrique du sud [Punt, 2003][17]

2.1.2 Les paramètres civilisationnels dans le détail

La **natalité** est la résultante de quantité de paramètres, certains biologiques, d'autres sociaux. Si pour les animaux, la biologie prime, chez l'homme, le paramètre qui l'influence le plus est l'environnement socio-culturel. Ainsi, un des paramètres les plus importants pour la natalité est-il l'âge du mariage (nous prendrons ce terme dans un sens générique de démarrage de la vie en couple et donc de la période de reproduction) et non pas l'âge de la puberté qui représente le possible de la fécondité, c'est-à-dire le début de la fertilité. De nos jours, l'allongement de la durée d'étude et/ou les retards imposés par la vie professionnelle des femmes retardent considérablement l'initiation de la reproduction. Le taux de fécondité par femme et par âge résume à lui seul l'ensemble de ces caractéristiques selon les populations étudiées. Pour nos calculs, nous avons utilisé les résultats de l'étude de l'INED[18] dans laquelle la fécondité des femmes nées en France est comparée à celles ayant immigré en France mais originaires d'Afrique du Nord. L'étude de Punt [2003][17] nous a permis de valider cette dernière donnée en la comparant à la fécondité des femmes de ménages (noires…) en Afrique du Sud : le pic de fécondité était retrouvé dans les deux cas à 28 ans. Des données du Bengladesh[19] fournissaient des résultats assez contradictoires : la fécondité des femmes de 1960 à 1976 montrait un pic de fécondité proche de 22 ans. C'est la précocité des mariages qui semble faire la différence entre ces résultats. Il paraît probable que ce dernier pic soit plus à même d'illustrer les sociétés primitives.

L'âge du mariage est lui aussi un paramètre sociétal. Sous l'Ancien Régime en France, la loi interdisait un mariage avant 14 ans pour les garçons et 12 ans pour les filles, ces minimums ayant été modifiés par décret en mars 1803 pour être portés à 18 ans pour les hommes et 15 ans pour les femmes. Pour autant, l'âge moyen de mariage entre 1740 et 1830 oscille pour les hommes entre 27 et 28 ans et chez les femmes entre 25 et 26 ans[22].

Selon Levy et al. [1982][23] l'âge du mariage peut être extrapolé de l'âge du 1er enfant : en France en 1980, il était de 23 ans pour les femmes avec un écart-type de 5 ans. Dans des pays développés de l'Asie, on retrouve des âges de mariage similaires toujours pour les femmes en 2000 : 24,1 ans en Thaïlande, 23,3 ans en Chine. Il est par contre nettement plus tardif au Japon (28,6 ans) ou en Corée du Sud (27,1 ans)[20].

Dans les pays sous-développés, le mariage est habituellement plus précoce. «*Dans la plus grande partie de l'Asie du Sud, le mariage est universel et précoce. En 2000, l'âge moyen au premier mariage (AMPM) des femmes était de 19 ans au Bangladesh (Bangladesh Demographic and Health Survey, 1999-2000), de 22,7 ans au Pakistan (Pakistan Reproductive Health and Family Planning Survey, 2000-2001) et de 23,6 ans au Sri Lanka (calculé à partir des données du recensement 2001)* »[21].

Dans l'hypothèse d'un environnement de pays sous-développé, nous avons fixé l'âge moyen du mariage des femmes à 18 ans avec un écart-type à 5 ans. Une borne inférieure est imposée à 13 ans et on assumera la normalité de la distribution.

La différence d'âge entre conjoints est une autre question sociétale. En occident, sous l'ancien régime, des différences importantes étaient observées mais principalement du fait des nombreux remariages suite aux décès maternels lors de l'accouchement[22]. En réalité, cette grande différence, au moins en France, n'était pas retrouvée lors du 1er mariage sur cette même période, puisqu'elle se situait autour de 2 ans. Cette différence se retrouvait à l'identique dans les années 1980[23]. Au Canada, elle a évolué

entre 1920 et 1990 de 3 ans à 2 ans[24]. Dans certaines sociétés aujourd'hui encore à l'instar de certains pays du Maghreb (Tunisie, Maroc), la différence est considérablement plus large du fait que préliminairement au mariage, les hommes doivent faire la preuve qu'ils pourront subvenir aux besoins du foyer, ce qui impose aux hommes un délai qui est imparti à l'obtention d'une assise économique. Toutefois, il semble qu'une différence minimale d'âge au mariage de 2 ans représente assez bien la généralité, quel que soit le type de civilisation et c'est ce que nous avons programmé dans POLYGENE.

Le taux de célibat, parallèlement à l'âge du mariage, est fortement influencé par l'environnement social et constitue un possible frein à la natalité observable d'une population, ce que l'on constate pendant et après les périodes de guerre chez les femmes. Ainsi, la domesticité (emploi de servantes à domicile) induisait dans la paroisse de Saint-Sulpice (près de Paris) entre 1715 et 1744 un célibat dans les métropoles pouvant atteindre 15% des femmes de 45 ans et plus[25]. De même, une religion qui impose une vie monacale durable à ses jeunes corollairement à la chasteté, grèvera nécessairement le taux net de fécondité par individu. En l'occurrence, à la fin du XVIII^{ème} siècle, en France, le célibat imposé aux femmes devenant religieuses se chiffrait à environ 12% de la population féminine adulte, ce qui est considérable [Henry, 1978][26]. Leur enquête de registre portant sur les personnes décédées après 50 ans entre 1660 et 1854 produit des chiffres édifiants : « *de 6 % à 7 % dans les générations nées vers 1675, la fréquence du célibat définitif atteint 12 % chez les femmes nées cent ans plus tard; la hausse s'accélère ensuite jusqu'à un maximum de 14 % chez les femmes nées vers 1790; une longue baisse, moins rapide que la croissance du XVIII^{e} siècle, ramène les générations nées vers 1850 au niveau des générations nées vers 1760* ».

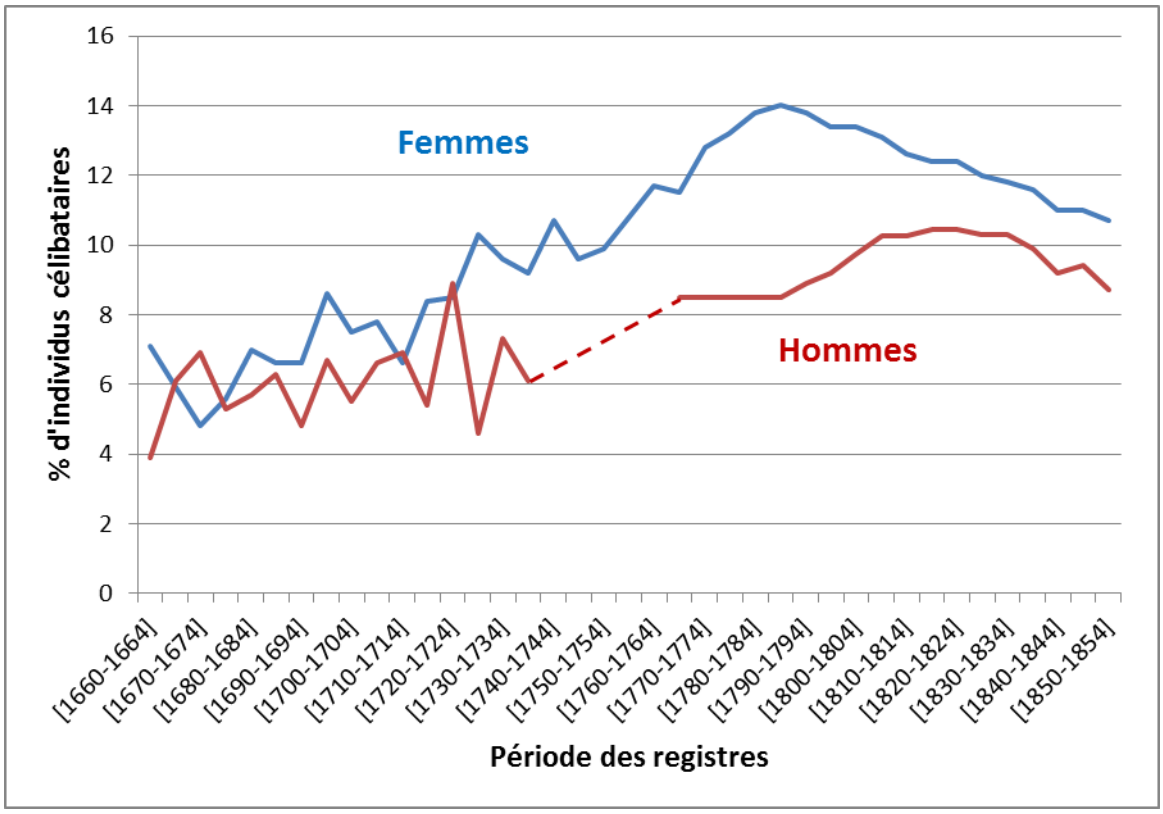


Figure 2 : proportion d'individus célibataires parmi les personnes décédées après 50 ans (période d'étude 1660-1854) – d'après Henry et al, Population, 1979

Les auteurs signalent cependant que des disparités campagnes/villes pourraient toutefois altérer ces chiffres. Aujourd'hui, la marginalisation socio-économique d'une frange de notre population pourrait conduire au célibat d'un nombre non négligeable d'individus, avec au final un impact probable sur la natalité.

Dans l'étude COSA[27], 3.5 % des femmes n'avaient jamais vécu en couple ou l'avaient vécu mais moins de 3 ans et étaient sans enfant. On a donc implémenté dans POLYGENE un taux de fixe de 4% que l'on a considéré valable pour les hommes aussi. Il est clair qu'en contexte de polygamie ou de polyandrie, cela puisse augmenter considérablement ces taux pour l'un des deux sexes. Il est par ailleurs probable que des aspects biologiques contribuent aussi au célibat, même si, exception faite des malformations, on ne connaît pas quels mécanismes précis jouent en arrière-plan. Nous parlerons plus bas d'un résultat que nous avons observé à ce sujet.

La polygamie et la polyandrie : nous avons implémenté des routines particulières gérant de la polygamie, laissant de côté la polyandrie très peu répandue actuellement (quelques rares ethnies de l'Himalaya au Tibet et au Bhoutan semblent conserver cette coutume). Il nous a semblé intéressant de vérifier si ce type d'organisation sociale pouvait favoriser la transmission des mutations délétères, en particulier dans l'hypothèse où elles induiraient une plus grande fertilité chez les hommes.

2.1.3 Paramètres liés à l'espèce

Pour la plupart, les paramètres biologiques n'ont vraisemblablement guère changé au cours du temps, comme ceux inhérent à la fertilité (âge de la puberté, des premières règles, âge de la ménopause), qui ont plus trait à l'espèce qu'à l'époque. Deux points particuliers doivent être évoqués préalablement : la mortalité en couches et le risque de malformations congénitales.

La mortalité maternelle : prise en compte dans les pyramides des âges fournies par sexe, elle ne sera donc pas paramétrée ici. Mais elle était une des principales causes de décès chez la femme au moyen âge avec un risque majoré tant pour les très jeunes primipares que pour les plus de 35 ans pouvant dépasser 1.5% par grossesse[28]. De tels chiffres sont encore retrouvés dans certains pays d'Afrique sub-saharienne avec aujourd'hui le triste record à 1.1% pour le Sierra Leone. En comparaison dans les pays développés, le taux de mortalité maternelle est inférieur à 10/100 000 [World Health Org., 2014][29].

Les malformations congénitales ont un impact non négligeable sur la démographie, puisqu'environ 3% des nouveau-nés viables vont être porteurs de malformations, dans la plupart du temps limitées à un seul système biologique (squelette, cœur, SNC…). Mais dans environ 25% de ces cas, les atteintes seront multiples ou syndromiques (ex. trisomie 21). Si aujourd'hui les échographies prénatales et certains dosages sanguins permettent de diagnostiquer les cas les plus graves et certaines atteintes syndromiques, nous avons conservé dans POLYGENE les chiffres habituels (3% de risque de malformations dont 25% multiples) et en cas de malformations multiples, nous avons implémenté un risque de mortalité de 100% décroissant sur 10 ans, les premières années correspondant au risque le plus élevé de décès comme cela se vérifiait en population générale jusque récemment. De surcroît, un lien ayant été montré dans notre étude entre les mutations favorisant le cancer et l'incidence de malformations congénitales sévères, nous avons ajouté un gradient de risque en fonction du type de mutation mis en jeu.

La fertilité des femmes a été estimée à partir de diverses sources : tout d'abord, l'étude précitée de Toulemon[18] nous a fourni des courbes de fécondité par âge, ce pour les françaises et pour les femmes originaires d'Afrique du Nord. Il peut aussi se déduire, à partir de simulations, des pyramides des âges dès lors que l'on veut assurer la stabilité d'une population théorique sur plusieurs siècles. Dans ce domaine, il faut distinguer la parité et la gravidité, la première notion se référant aux enfants vivants tandis que la gravidité correspond au nombre de fois où les femmes ont été enceintes. La différence correspond au taux de fausses couches. Dans notre cohorte COSA[27], constituée en Auvergne entre 1996 et 1999 et comprenant environ 2000 femmes dont la moitié étaient des patientes incluses en raison d'un cancer du sein ou de l'ovaire, le taux de fausse couches atteignait 20% de l'ensemble des grossesses. Nous avons retrouvé des chiffres similaires dans d'autres études[30, 31, 32]. Suite à notre étude de 2015[33], mettant en évidence des taux de natalité supérieurs chez les porteurs/porteuses de mutation BRCA en partie expliqués chez les femmes par un moindre taux de fausses couches, nous avons rendu ce taux paramétrable avec une réduction maximale de 10% en valeur absolue pour un taux standard à 20%.

La ménopause : la fertilité des femmes subit un arrêt définitif lors de la ménopause. La fin de l'ovulation, qui peut advenir assez longtemps avant le décès, est donc une étape importante à paramétrer dans nos simulations. Dans notre cohorte COSA[27], l'âge moyen de la ménopause était de 51 ± 3.8 ans, se répartissait de manière gaussienne, en accord avec les chiffres nationaux[34] et internationaux[35, 36]. Cet âge allant de 40 à 60 ans ne montre guère de variation, sinon une précocité décrite de deux ans pour les consommatrices de tabac. Nous l'avons fixé à 50 ans ± 4 ans d'écart-type avec une distribution gaussienne.

Pour **la fertilité des hommes**, nous nous sommes basés sur une petite étude réalisée à partir d'anciens registres canadiens[37]. Cette étude montrait qu'aux XVIII^{ème} et XIX^{ème} siècles que les veufs continuaient de se remarier et avaient fréquemment des enfants même après 60 ans. L'habitude était de se remarier avec des femmes beaucoup plus jeunes, de 30 ans et moins. Ces mariages d'hommes de 50 ans et plus contribuaient à une moyenne de 2.2 enfants, à comparer avec les 2.8 enfants pour les hommes mariés avant 30 ans. Sur un plan biologique, la raison de cette baisse de fertilité peut être rattachée à la qualité du sperme qui diminue avec l'âge, les spermatozoïdes devenant moins concentrés et moins motiles[38]. Toutefois, on doit relativiser l'impact de ce phénomène dans nos calculs : cette baisse due à l'âge est nettement moindre que celle qui a été mesurée sur les mêmes paramètres entre 1973 et 1992 chez l'adulte mâle français et dont les causes demeurent mal connues (habitudes de vie, alimentation, pollution… ?)[39]. Comme la baisse de fertilité est déjà transcrite dans les statistiques de natalité, nous avons ignoré ces variations liées à l'âge et nous n'avons pas fixé de limite d'âge à la fertilité des hommes. Nous avons cependant rendu plus difficile avec l'âge la recherche d'une partenaire. Enfin, similairement aux femmes porteuses de mutations BRCA qui avaient une fécondité augmentée par rapport à la population contrôle et pour lesquelles nous avons autorisé une réduction paramétrable du taux de fausses couches, nous avons implémenté une possible augmentation de natalité pour les hommes porteurs de mutation.

2.1.4 Paramétrage des conditions génétiques à tester

Mis à part les paramètres contextuels qui vont « brosser la toile de fond », de nombreux autres paramètres ont été ajoutés, ceci afin de pouvoir proposer des jeux d'essai prenant en compte des situations très variées.

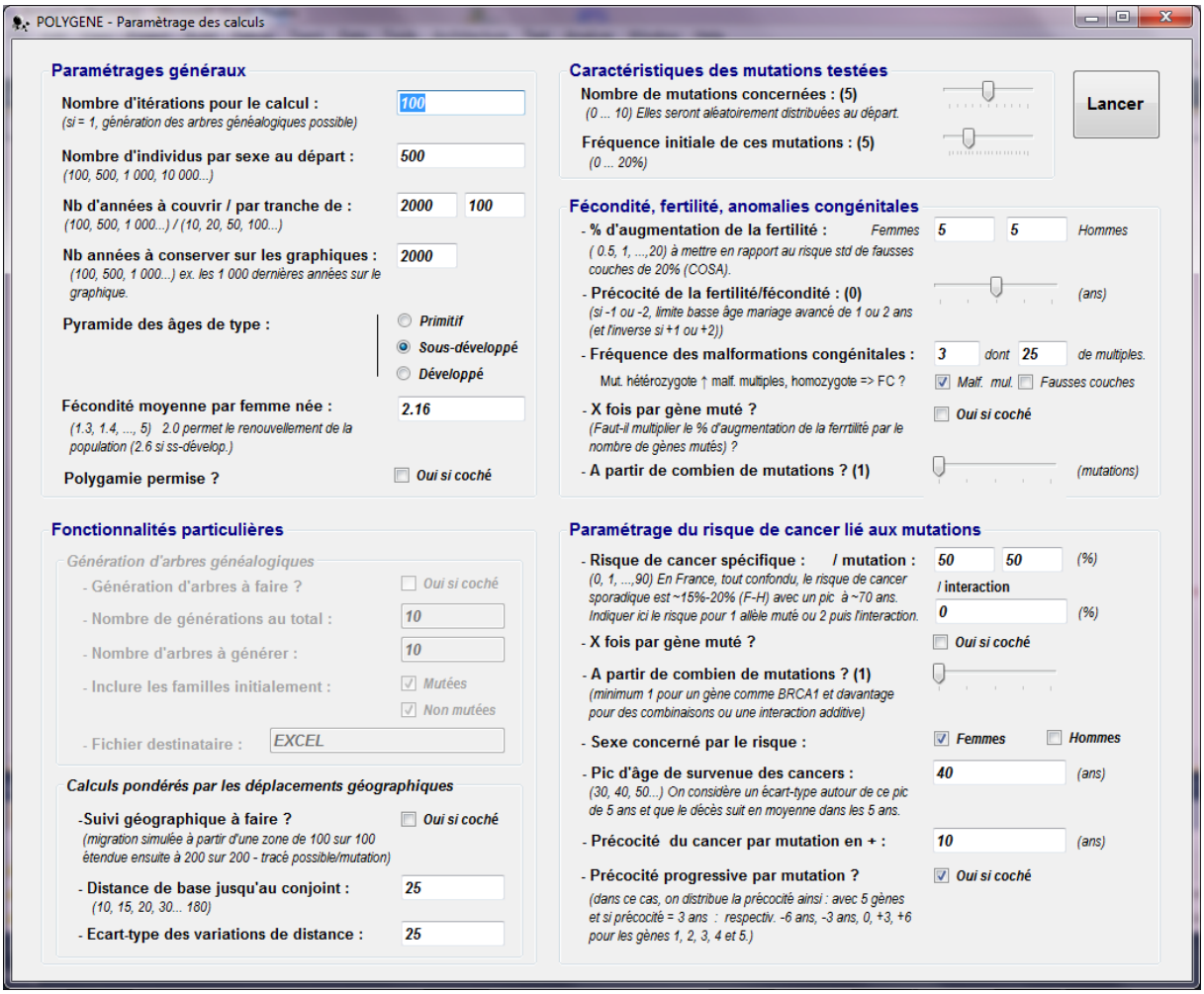


Figure 3 : interface permettant de paramétrer les arbres généalogiques

Les mutations délétères : il est possible d'en inclure entre 0 et 10. Leur fréquence initiale est comprise entre 0% et 20%. Evidemment, n'en programmer zéro ou bien leur affecter une fréquence nulle induit la génération d'une population sans risque héréditaire de cancer. Cela permet en particulier de vérifier si les autres paramètres du modèle induisent des résultats en accord avec leur définition initiale, ce dans un contexte normal. L'effet de ces mutations est à préciser à plusieurs niveaux :

- **Le pourcentage d'augmentation de fertilité** (de 0% à 20%), pour les porteurs de mutation, hommes ou femmes. Pour les femmes cela correspond à une réduction de la fréquence des fausses couches puisque ce mécanisme avait été montré dans notre étude. Quant aux hommes, comme le nombre d'enfants est programmé par femme, ce nombre est réévalué lors du mariage avec une augmentation proportionnelle au taux entré pour les porteurs de

mutation (en fait il est procédé à un tirage au sort et si le nombre généré (entre 0 et 1) est inférieur ou égal au pourcentage d'augmentation de la fertilité, on ajoute 1 enfant au nombre d'enfants attendus de l'épouse. En arrière-plan, il est envisageable que l'augmentation de fertilité de l'homme puisse provenir d'une meilleure qualité de sperme, de davantage de sécrétion de testostérone voire d'aspect physiologiques stimulant le plan motivationnel. Mais dans POLYGENE, autant pour les hommes que pour les femmes, il n'est point besoin de distinguer la cause réelle sous-jacente : pour notre programmation, seuls les effets comptent.

- Si l'augmentation de fertilité est proportionnelle au nombre de mutations et dans ce cas à partir de quel nombre (par défaut dès 1 mutation).
- **Une éventuelle précocité ou un retardement de la fertilité** pour les porteurs de mutations (compris entre -2 et +2 ans autour de l'âge de la puberté). Ce paramètre n'est pas proportionnel au nombre de mutations.
- **Le taux de malformation congénitales** et le pourcentage de ces malformations qui seront multiples (ce pourcentage est multiplié par le nombre de mutations pour les individus porteurs si la proportionnalité est demandée pour l'augmentation de la fertilité). Pour rappel, les malformations multiples sont déclarées dans POLYGENE létales dans les 10 premières années. Donc, on pourra vérifier que dans les arbres générés, tous les polymalformés décèdent dans les 10 ans qui suivent.

Les cancers mutations-dépendants : c'est un des paramètres principaux pour nos tests. Il faut savoir que pour les mutations BRCA, la pénétrance sur la vie d'une femme selon le gène peut atteindre 70% pour le cancer du sein et 40% pour les cancers de l'ovaire. Ces cancers tendent à se produire à des âges précoces, mais rarement avant 25 ans pour le cancer du sein et avant 50 ans pour le cancer des ovaires.

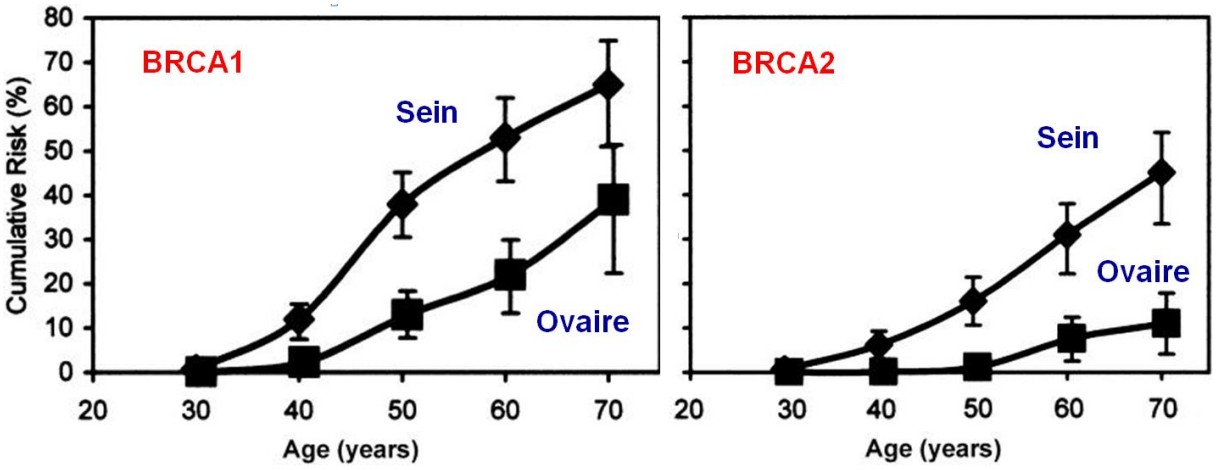


Figure 4 : Pénétrance selon l'âge des femmes des gènes BRCA, d'après Antoniu [2003][40]

Il est connu que la pénétrance de ces gènes est variable selon le contexte. Ainsi, pour les gènes BRCA, c ils induisent moins de cancer chez les femmes multipares : comme celles-ci étaient nettement plus nombreuses jadis, la pénétrance était notablement diminuée alors. Aujourd'hui, une alimentation équilibrée, une activité physique régulière associée à l'absence de tabac permettent d'envisager de réduire par 2 cette pénétrance[41].

La modélisation de la pénétrance des mutations doit tenir compte de leur aspect homozygote ou hétérozygote [42]. L'ADN est constitué de 2 allèles, chacun porteur de l'information génétique en provenance du père et de la mère. De ce fait, les anomalies génétiques - mutations ou polymorphismes - peuvent être localisées sur un seul allèle et l'on parle alors d'hétérozygotie, l'information étant hétérogène, ou bien similairement sur les deux allèles et l'on est en situation homozygote. Pour simplifier, si la mutation est délétère, une mutation hétérozygote permettra le bon fonctionnement de 50% de la fonction impartie au gène tandis que la fonction sera totalement annulée en cas d'homozygotie et pourra parfois s'avérer létale. Dans POLYGENE, deux pénétrances sont donc paramétrables, définissant l'impact des mutations selon que l'individu est porteur hétérozygote ou homozygote. Toutefois le caractère dominant ou récessif des polymorphismes n'a pas été implémenté : cet aspect semble secondaire en cancérogenèse. Enfin, sachant que certains polymorphismes n'agissent de manière délétère qu'à partir du moment où ils sont associés à d'autres polymorphismes bien définis (par exemple pour des gènes agissant en cascade), il est possible de paramétrer une pénétrance uniquement liée aux interactions. De tels cas ont déjà été mis en évidence dans les pathologies coronariennes[42]. Comme notre objectif est d'évaluer l'impact des combinaisons de plusieurs gènes sur les phénotypes, ces divers paramétrages étaient indispensables. Bien évidemment, l'étude des interactions 2 x 2 voire 3 x 3 des gènes impacte considérablement les approches mathématiques et statistiques – et les temps de calcul – puisque ces interactions élèvent au carré ou au cube le cardinal des possibilités. Nos simulations se sont limitées aux interactions 2 x 2.

Dans nos développements pour POLYGENE, nous n'avons pas tenu compte des différentes localisations de cancer possibles. Ceci pourrait être discuté, toutefois, la présence ou l'absence de cancer était suffisante pour tester l'impact de mutations délétères sur un plan démographique, ou pour tester des modélisations mathématiques et/ou statistiques. Les paramètres suivants ont été jugés suffisants pour évaluer nos hypothèses et nos approches. Ils peuvent caractériser tous les gènes ou varier selon le gène :

- La pénétrance des mutations : elle est fixée de manière globale, mais son résultat sur les populations porteuses pourra varier notablement au vu des paramètres ci-après.
- Le sexe concerné par le risque de cancer : les hommes, les femmes ou bien les deux.
- Le pic d'âge de déclaration des cancers : l'âge résultant pour chaque individu est calculé aléatoirement autour de ce pic, avec un écart-type de 5 ans.
- L'augmentation de la pénétrance selon le nombre de mutations dont est porteur chaque individu. Le risque de cancer selon le sexe est multiplié par le nombre de mutations puis ramené à 100% si le résultat dépasse. Malgré tout, même dans ce cas, l'individu peut mourir d'autre cause avant que le cancer ne se déclare.
- La précocité du cancer par mutation supplémentaire : on peut rendre l'apparition du cancer d'autant plus précoce que les mutations sont nombreuses (en années)
- Progressivité de la précocité selon la mutation : cette option permet quand on introduit plusieurs mutations, d'affecter un âge progressif à chaque mutation. Par exemple, si l'on a paramétré 5 mutations et que l'on indique un pic d'apparition de la maladie à 50 ans et une progressivité de 10 ans, un pic d'âge de déclaration est fixé à 30 pour la première mutation, 40 pour la 2ème, 50 pour la 3ème, 60 pour la 4ème et 70 pour la dernière. Cela permet d'observer si la prévalence des mutations varie dans le temps suite à la variation de l'âge de déclaration.
- L'âge du décès dû au cancer : il est fixé en moyenne à 5 ans post déclaration, avec un écart-type de 5 ans (tirage au sort pour la détermination de ce délai avec une distribution gaussienne).

Les cancers sporadiques sont censés se produire un peu au hasard du fait du risque cumulatif de mutation lié au nombre de divisions cellulaires – variable selon les tissus - à mesure que l'âge avance : la corrélation entre ce nombre de divisions et le risque de cancer par âge a été trouvé à 0.81. Ainsi, les deux tiers des cancers (0.81^2 = 66%) ne trouveraient pas d'autre explication que la « malchance », le reste relevant de causes génétiques ou environnementales[43]. Le risque de cancer sporadique par tranche d'âge et par sexe, peut donc être estimé à partir de données épidémiologique générales, telles celles fournies par le département « Cancer Research UK » anglais de 2012 à 2014[44]. Ces données sont représentatives de la population occidentale.

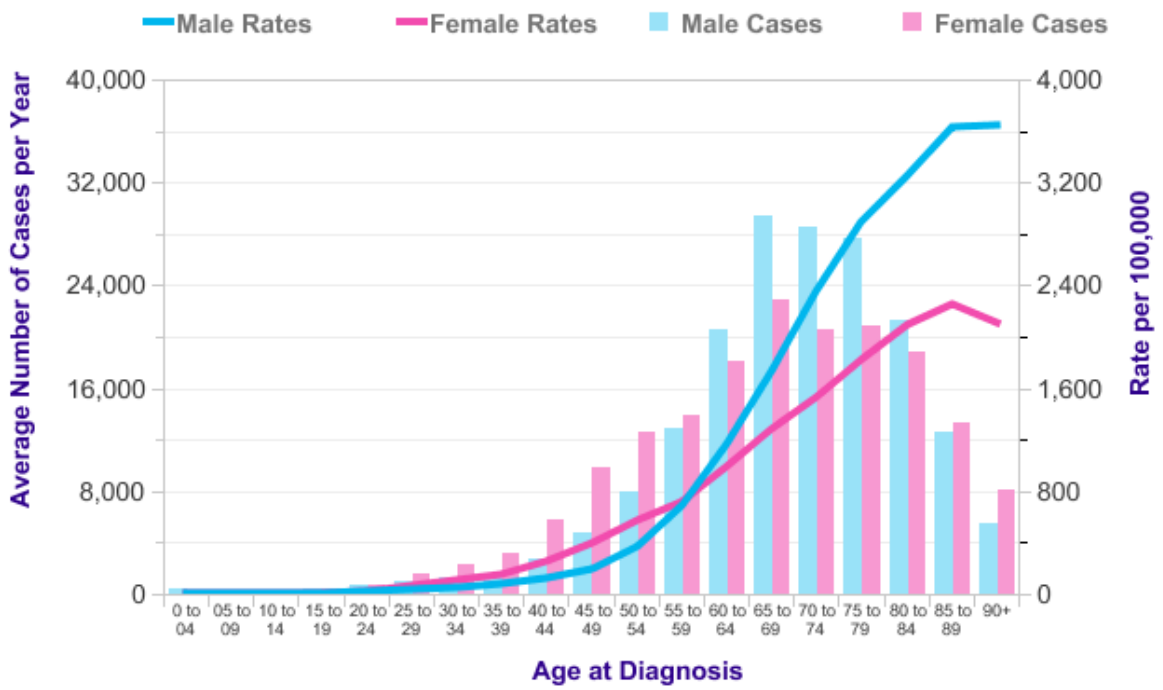


Figure 5 : nombre moyen de nouveaux cas de cancer par année et incidence annuelle par âge et par sexe au Royaume-Uni de 2012 à 2014 (ref. Cancer Research UK)

L'incidence des cancers sporadiques a été calculée à partir de l'âge du décès de chaque individu pour lequel un cancer lié à une prédisposition génétique n'avait pas été implémenté. C'est-à-dire que la détermination de la cause de décès (du fait d'un cancer sporadique ou non) a été faite au moment du décès, dès lors qu'il ne s'agit pas d'un cancer lié à une mutation introduite dans le modèle ou bien une polymalformation. Nous avons, à partir des incidences annuelles fournies par le « Cancer Research UK », estimé le risque d'avoir eu un cancer dans les 5 années précédentes. Les équations polynomiales permettant ce calcul ont été interpolées des statistiques d'incidence annuelle de la figure précédente :

Pour les femmes, ce risque était :

$$y = -2\ 10^{-10} x^5 + 4\ 10^{-8} x^4 - 2\ 10^{-6} x^3 + 6\ 10^{-5} x^2 - 0{,}0007\ x + 0{,}0024$$

Et pour les hommes :

$$y = -4\ 10^{-10} x^5 + 8\ 10^{-8} x^4 - 4\ 10^{-6} x^3 + 8\ 10^{-5} x^2 - 0{,}0004\ x + 0{,}0009$$

avec x = âge de décès – 5 (en années)

Par exemple, la probabilité qu'un homme décédant à l'âge de 66 ans le soit en raison d'un cancer sporadique est de 13% tandis qu'elle est de 58% pour un décès à 90 ans. Pour les femmes, le risque est moindre respectivement à 11% et 41% pour les âges cités.

2.1.5 Aspects méthodologiques

2.1.5.1 Méthode pour générer des variables distribuées de manière gaussienne

Il n'existe pas d'expression analytique simple pour la fonction de répartition de la fonction normale centrée réduite :

$$\frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}x^2}$$

Pour générer des valeurs réparties de manière gaussienne, nous avons donc utilisé la méthode de Box-Muller[45] réputée très rapide pour les calculs informatiques. L'intérêt annexe de cette méthode est qu'elle génère deux nombres indépendants distribués selon la loi normale centrée réduite : le premier nombre est conservé pour le calcul tandis que l'autre peut éventuellement servir de nouvelle racine pour la liste de nombres aléatoires.

2.1.5.2 Génération des formules pour le calcul des risques par âge

De manière générale, le calcul des risques variables selon l'âge, quel qu'en soit l'objet, a été effectué de manière unique, en ajustant par régression polynomiale sur les données rapportées dans la littérature ou sur les sites officiels comme l'INED. En voici un exemple utilisant les données du site de la Banque Mondiale concernant l'Afrique du sud réalisée en 1993[46] :

Tableau 1 : parité annuelle des femmes de ménage d'Afrique du Sud en 1993 par tranche de cinq ans

âge médian classe	parité cum. / 5 ans	parité simple / 5 ans	parité annuelle
17	0,19	0,19	0,038
22	0,94	0,75	0,15
27	2,23	1,29	0,258
32	3,47	1,24	0,248
37	4,4	0,93	0,186
42	5,48	1,08	0,216
47	6,29	0,81	0,162

Ces données sont alors utilisées pour bâtir une représentation graphique puis ajustées à l'aide d'une régression polynomiale comme dans la figure ci-dessous :

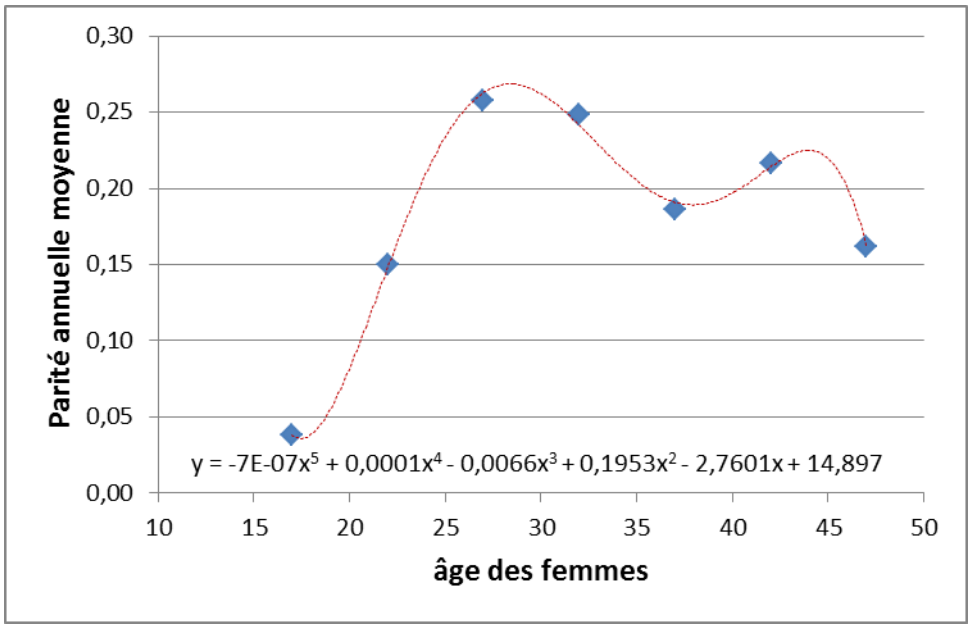


Figure 6 : graphe de répartition de la parité par tranche d'âge (données du tableau précédent) : courbe en rouge = ajustement polynomial

La formule de régression est alors implémentée dans POLYGENE et utilisée quand le cas se présente. NB. En réalité, nous avons utilisé le logiciel SEM[47] pour effectuer ces calculs, ceci pour pouvoir juger visuellement des fins de courbes (en-deçà du premier point et au-delà du dernier) qui sont absentes dans EXCEL et qui pourraient produire des artefacts.

2.1.5.3 Particularité de l'approche Monte-Carlo

Pour l'approche Monte-Carlo, 4 types d'information sont à fournir :

- Le nombre d'itérations à réaliser : le modèle arrive à saturation dès quelques centaines d'itérations.
- Le nombre d'individus par sexe au départ : un millier d'individus sont nécessaires au minimum en particulier si la fréquence des mutations délétère est faible, comme c'est le cas dans la réalité. Sinon, les familles risquent d'être « polluées » par des mutations glanées via les mariages avec d'autres familles à risque.
- Le nombre d'années de suivi démographique : 2 millénaires suffisent à objectiver les principaux impacts de mutations délétères.
- Le « saucissonnage » de la période de suivi, c'est-à-dire la périodicité des relevés des résultats, ceci afin de pouvoir générer des états donnant des indications à chaque étape. Si plusieurs millénaires sont inclus dans la période de suivi, des points séculaires sont suffisants pour la représentation graphique longitudinale.

On peut accessoirement demander un visuel des statistiques sur une période restreinte, c'est-à-dire ne couvrant pas toute la période de suivi.

2.1.6 Fonctionnalités de POLYGENE

Le programme POLYGENE offre trois principaux groupes de fonctionnalités : des statistiques sur la démographie des populations simulées, des suivis géographiques par mutation et en ce qui nous concerne ici, la génération d'arbres généalogiques. Nous allons décrire rapidement ces fonctionnalités.

2.1.6.1 Statistiques démographiques

Sans ces statistiques, impossible d'évaluer les performances des modèles. Certaines sont statiques (par exemple l'état de la population finale) tandis que d'autres sont longitudinales et retracent l'évolution des paramètres démographiques au cours des millénaires demandés. Ci-après, une statistique d'état : la pyramide des âges des populations finales, ce pour deux contextes : primitif et industrialisé. Initialement, 500 couples de 20 ans ont été définis avec des caractéristiques de natalité et de mortalité propres aux 2 contextes, puis le programme a fait évolué de manière aléatoire ces populations sur 2 000 années[i]. La pyramide est calculée sur l'ensemble des individus vivants au terme du suivi, c'est-à-dire lorsque les calculs s'arrêtent.

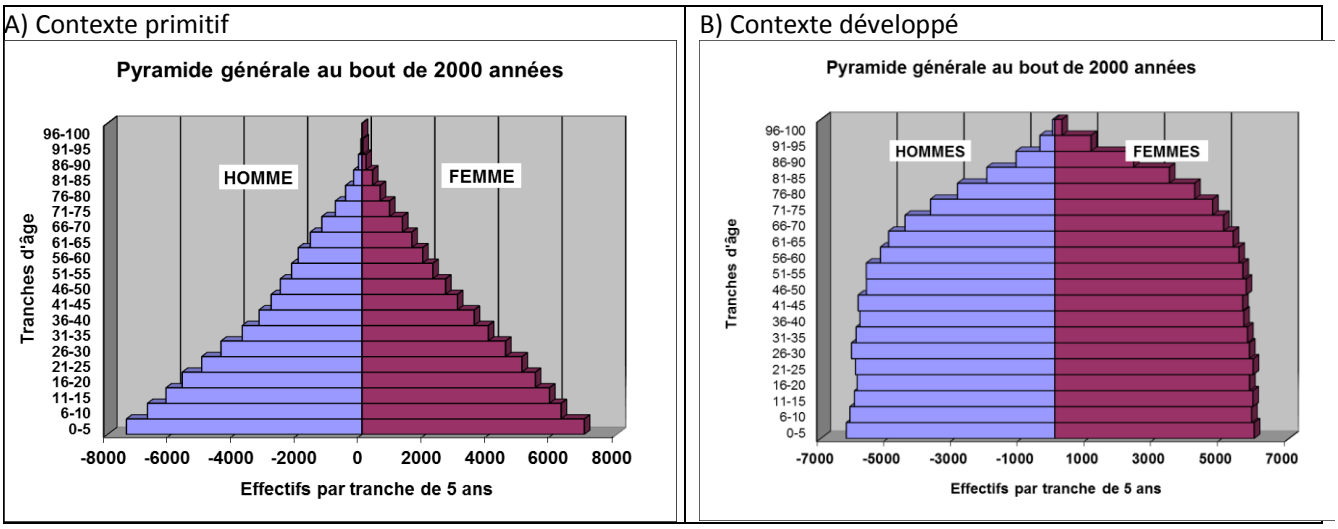


Figure 7 : exemples de pyramide des âges générée par POLYGENE pour une population de 1000 individus au départ et suivie durant 2 millénaires :
A) population primitive et B) population de pays développé

En ce qui concerne les suivis longitudinaux, à chaque étape, chaque résultat est défini par son effectif, sa moyenne et son écart-type quand il s'agit de variables quantitatives (âge, nombre d'enfants…) sinon de l'effectif et de la fréquence par classe en cas de paramètre qualitatif (cancer oui/non…). Voici un exemple de ce type d'analyse pour un contexte primitif :

[i] Sans être outrancièrement pessimiste, il est peu probable que notre civilisation industrielle ne nous permette un jour de valider le résultat d'une telle évolution sur 2000 ans…

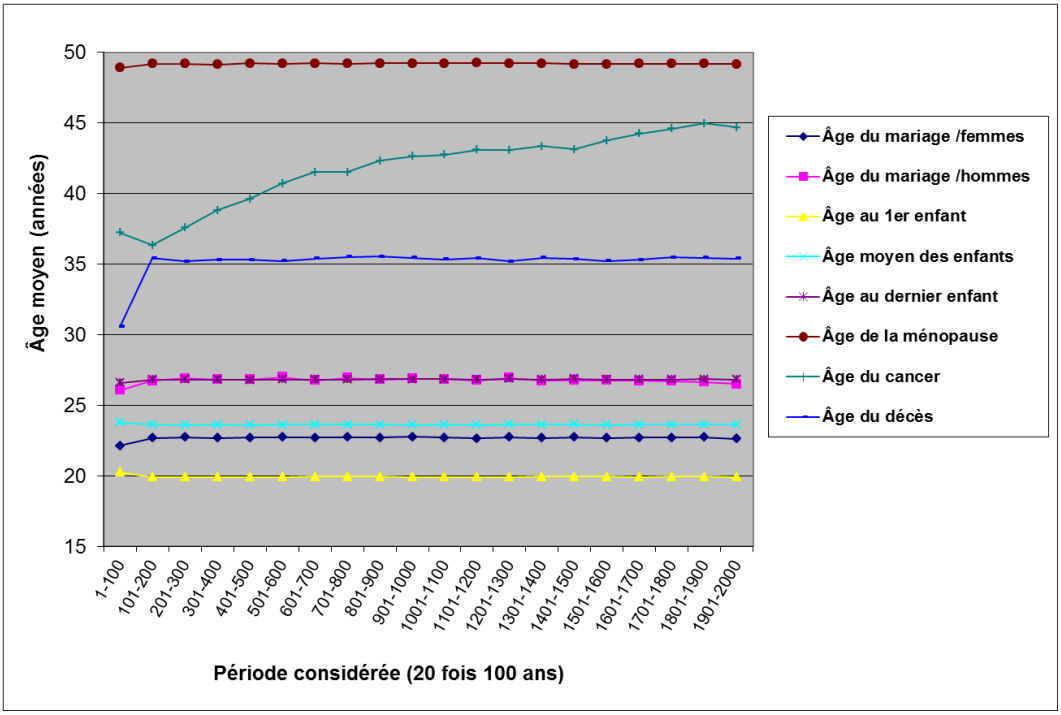


Figure 8 : exemple de suivi des âges lors des différents événements de la vie (contexte primitif – insertion de 5 mutations favorisant le cancer chez la femme de 20 à 60 ans par palier de 10 ans avec une pénétrance de 50%)

Ainsi, dans cet exemple, on observe une croissance significative de l'âge des cancers mutations-dépendants : cet âge tend à croitre jusqu'à l'âge de la ménopause, c'est-à-dire jusqu'à ce qu'il ne pénalise plus la fécondité des femmes. Ce résultat est remarquable en soi car il montre que le pic de cancer chez la femme autour de la ménopause n'est pas uniquement imputable à la durée de l'exposition hormonale (œstrogènes…), mais sans doute bien plus à la pression de sélection. Dans nos tests, ce rapprochement des courbes est dû à l'élimination au fur et à mesure des générations, des parents porteurs de mutation favorisant un cancer aux âges les plus précoces (20 ans et 30 ans), comme le montre la figure suivante :

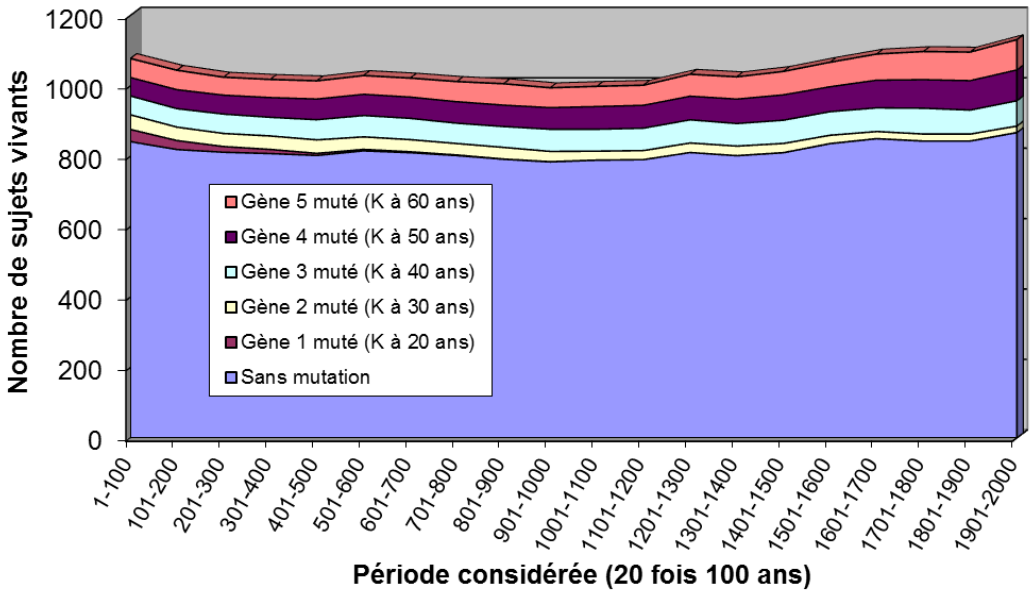


Figure 9 : Evolution démographique sur 2 millénaires selon 5 types de mutation familiale, chacune favorisant un cancer à des décennies différentes (de 20 à 60 ans). ("K" = cancer)

Dans la Figure 9 : Evolution démographique sur 2 millénaires selon 5 types de mutation familiale, chacune favorisant un cancer à des décennies différentes (de 20 à 60 ans). ("K" = cancer) la population portant une mutation favorisant un cancer à 20 ans (pénétrance = 50%) disparaît en environ 500 ans en dépit du fait qu'elle augmente la fertilité de 5%. Par contre les mutations favorisant un cancer plus tardivement (40, 50 et 60 ans) bénéficient de l'avantage en termes de fertilité et voient leur effectif croître en proportion des autres catégories, en particulier le groupe sans mutation.

Ces premières statistiques ont permis en 2015 dans PLOS-One[33], la publication d'un article princeps intitulé « *BRCA Mutations Increase Fertility in Families at Hereditary Breast/Ovarian Cancer Risk* », dans lequel nous démontrions à partir de la base de données oncogénétique du Centre Jean Perrin (en fait un sous-ensemble de 2,168 familles incluant 96,325 personnes) que les porteuses de mutation BRCA-1 ou BRCA-2 étaient bien moins souvent nullipares (9.1 % des femmes mutées versus 16.0 % pour les autres, p = 0.003), avaient significativement plus d'enfants ($1.8 \pm \sigma = 1.4$ versus 1.5 ± 1.3, p = 0.002). Mais il en était de même pour les hommes : 1.7 ± 1.3 versus 1.4 ± 1.3 (p = 0.024). Cet article est joint en annexe. Un autre résultat était important : le nombre de fausses couches. Chez les femmes appartenant à une famille mutée (sans qu'on sache ou non si elles étaient porteuses de la mutation), il était en moyenne de 0.16 ± 0.62 par femme alors qu'il passait à 0.25 ± 1.02 chez les autres femmes (p = 0.015).

Cette dernière anomalie ne nous a pas semblée sans conséquence, les fausses couches étant souvent un moyen naturel de mettre un terme à un embryon malformé. Il apparaissait donc possible qu'on trouve une incidence de malformations congénitales plus importante chez les enfants dont au moins un des deux parents était porteur de mutation favorisant le cancer. Nous avons la chance d'avoir en Auvergne un registre des malformations, le Centre d'Etude des Malformations Congénitales en Auvergne (CEMC-Auvergne) lancé en 1983 et qui colligeait en environ 10 000 enfants malformés. L'idée nous est venue de croiser cette base de données conséquente avec celle oncogénétique, contenant elle aussi de nombreux enfants nés pendant la même période et issus de familles principalement auvergnates. Comme nous n'avions pas mention des malformations congénitales dans cette dernière, seul un croisement des bases pouvait nous permettre de relier les enfants malformés à leurs parents éventuellement exposés à un risque héréditaire de cancer. Nous ne cacherons pas qu'un trésor de patience et de persévérance aura été nécessaire pour instruire les dossiers et obtenir les autorisations idoines avant la fusion de ces deux bases de données, à savoir celles de la CNIL, du CEERES (comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé) et l'avis éthique du CECIC Rhône-Alpes-Auvergne (comité d'éthique des centres d'investigations cliniques, Grenoble, IRB 5921), certains autorisations étant de surcroît devenues caduques suite aux changement de loi bioéthique (passage de la loi Huriet à la loi Jardet en 2012)… Les difficultés de la recherche ne sont pas toujours là où on les attend ! Nous sommes ensuite passés au travail sur les données à proprement parler : un appariement semi-automatique nous a permis les regroupements suivants :

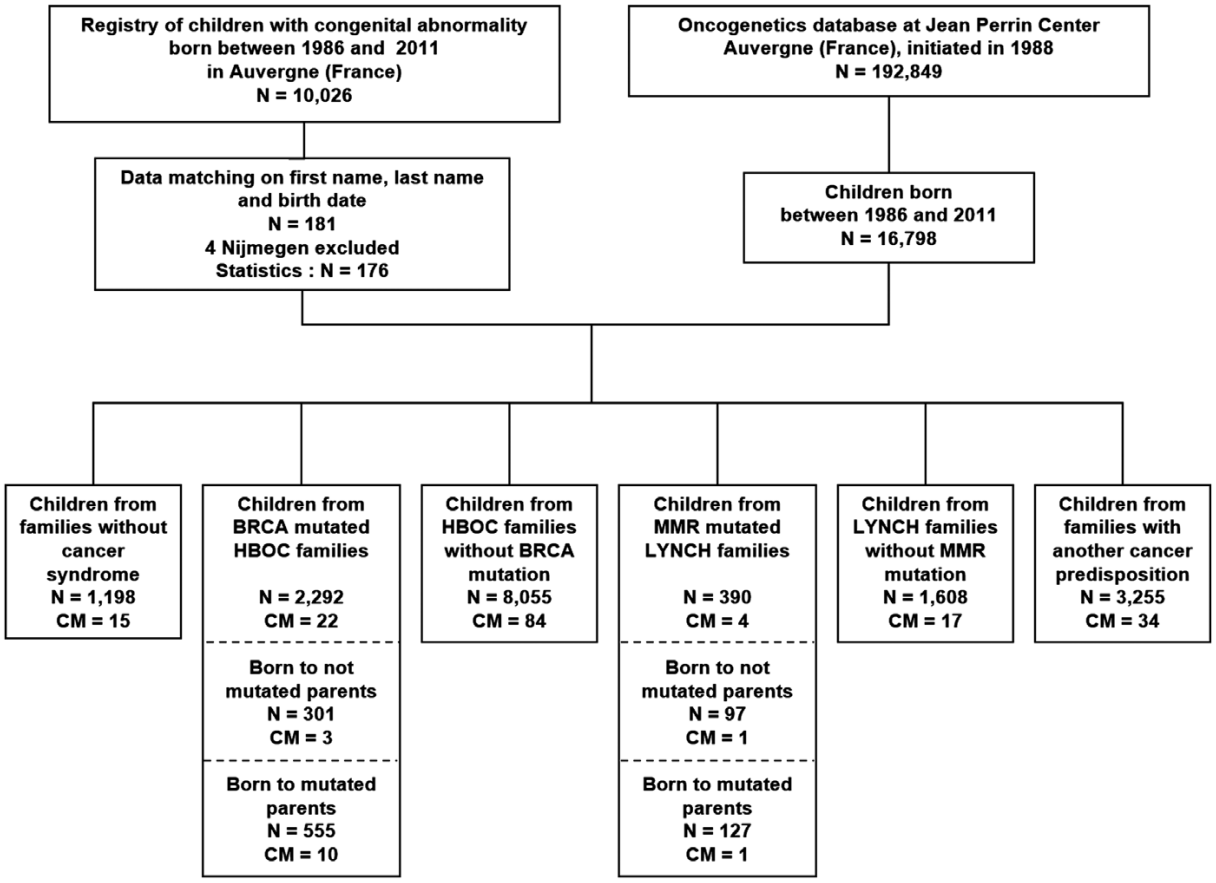


Figure 10 : diagramme d'appariement des enfants du registre régional des malformations congénitales avec ceux de la base de données oncogénétique (**CM**, malformation congénitale; **HBOC**, cancer héréditaire sein/ovaire; **Lynch**, syndrome de Lynch = risque de cancer colorectal)

La conclusion de l'article publié en 2018 dans la revue « *congenital anomalies* » était la suivante : "*compared to families without cancer syndrome, the risk of multimalformations was multiplied by 4.1 [0.8-21.7] for cancer-prone families but with no known deleterious mutation, by 6.9 [1.2-38.6] in families with a known mutation but an unknown parental mutational status and by 10.4 [2.3-46.0] when one parent carried the familial mutation... These results suggest that BRCA and MMR genes play an important role in human embryogenesis and that if their function is lowered because of heterozygote mutations, congenital malformations are either more likely (BRCA1 mutations) and/or more susceptible to concern several anatomical systems."*

Cet article est lui aussi placé en annexe. Pour la première fois au monde cette question était abordée et des arguments épidémiologiques de qualité lui étaient enfin apportés. Ce travail souffrait cependant d'une faiblesse : sur la période considérée (1986-2011), le suivi par échographie des mères était déjà largement généralisé et en cas de malformation patente de l'embryon, la consigne médicale était l'interruption de grossesse. Dans notre étude épidémiologique, nous n'avions pas intégré ces interruptions médicales de grossesse (IMG). Nous avons l'intention de compléter ce travail sur une plus longue période (jusqu'à 2017) et de prendre en compte ces IMG. En aparté, lors d'une proposition de collaboration oncogénétique avec la Biélorussie, nous avons appris à l'hôpital de Gomel, capitale régionale située en plein centre de la zone de contamination de Tchernobyl, que selon les études nationales, la pollution radioactive n'avait pas officiellement augmenté le taux de malformations congénitales. Etonnamment, l'augmentation du nombre d'IMG n'était pas prise en compte dans ces statistiques officielles.

2.1.6.2 Suivi géographique des populations selon le type de mutation

L'objet de ce module dans POLYGENE était d'évaluer comment les aspects spatiaux peuvent interférer avec la dynamique de transmission des mutations. En effet, il était important, en cas de synergie entre plusieurs gènes relativement à la cancérogenèse, d'observer si les mutations avaient tendance à s'exclure l'une l'autre, comme on pourrait le supposer. Les paramètres à fournir pour cette étude sont la distance moyenne **d** de recherche d'un conjoint ainsi que l'écart-type σ_d de cette distance à mettre en rapport à une "carte" virtuelle de 200 x 200 pixels (chaque pixel pouvant correspondre à des dizaines de kilomètres). Qu'induit ce paramétrage ? Lors de la recherche d'un conjoint, qui se fait à chaque année pour les femmes célibataires ayant l'âge de se marier, la routine recherche parmi les hommes potentiels disponibles s'ils sont dans un rayon de X pixels (= d + (σ_d multiplié par un nombre aléatoire distribué selon la loi normale). La routine développée dans POLYGENE a permis de vérifier notre hypothèse comme on peut le voir sur les graphes ci-dessous :

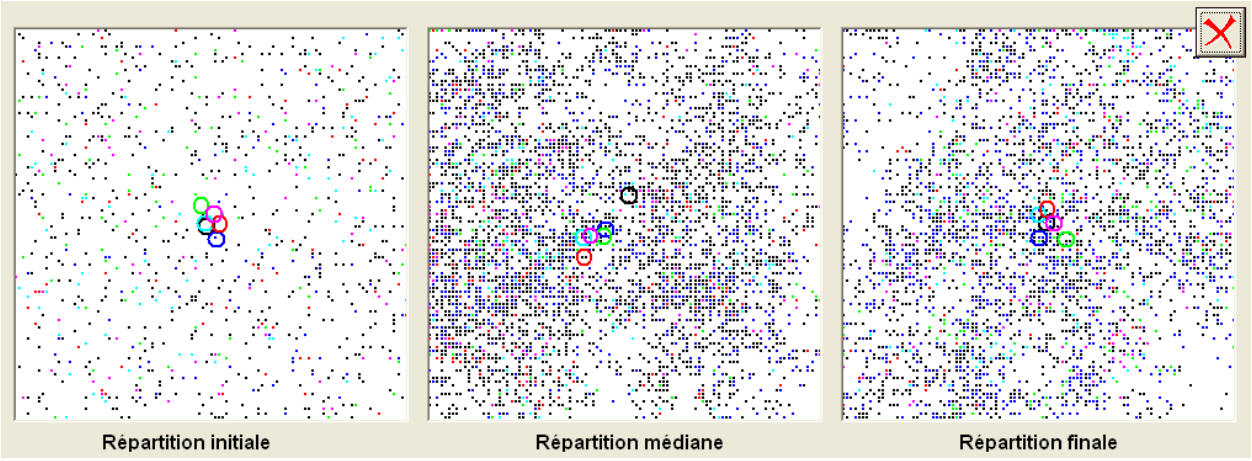


Figure 11 : évolution spatiale des clusters de population selon le type de mutation dont ils sont porteurs : pas de limitation géographique imposée. Le barycentre de chaque population représenté par un cercle de couleur demeure centré et la carte est entièrement peuplée (évolution sur 2000 ans).

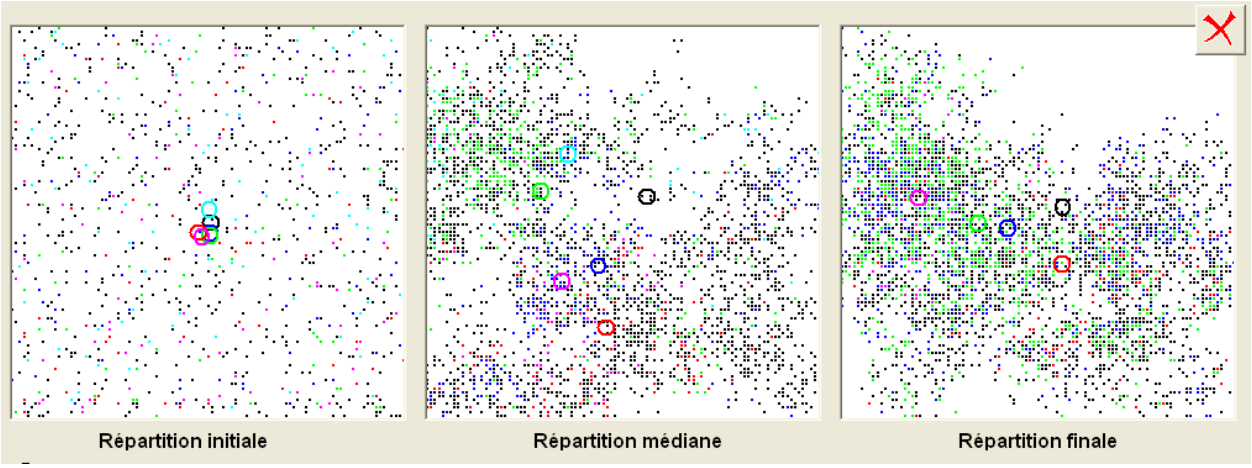


Figure 12 : évolution spatiale des clusters de population selon le type de mutation dont ils sont porteurs : déplacements limités à chaque génération. Les barycentres de chaque

population tendent à s'éloigner indiquant une ségrégation naturelle des populations (évolution sur 2000 ans). Des secteurs entiers de la carte sont inoccupés du fait de la raréfaction des conjoints induite par la limitation des déplacements.

L'ensemble des méthodes et des résultats liés à ces simulations ont été synthétisés en 2018 dans un petit journal sous le titre : « *What Selection Pressure Does to Mutations Favoring Cancer? Highlights of A Simulation Approach* »[48]. Nous l'avons joint en annexe.

2.1.6.3 Génération des arbres généalogiques

Un arbre généalogique est une manière de représenter les familles, génération après génération. En oncogénétique, contrairement aux généalogies familiales à finalité historique, on part d'un proposant (souvent un patient atteint d'un cancer) et on remonte vers ses ancêtres en ne conservant que les branches susceptibles de renseigner sur le syndrome familial. Un arbre généalogique « historique » ressemble donc à une pyramide inversée tandis qu'en oncologie, la pyramide repose sur sa base, la génération la plus nombreuse. En présence de plusieurs branches susceptibles de fournir des informations phénotypiques, il peut être alors intéressant de tracer autant d'arbres que de branches.

Ci-après un exemple de tracé d'arbre généalogique, avec l'explication des symboles :

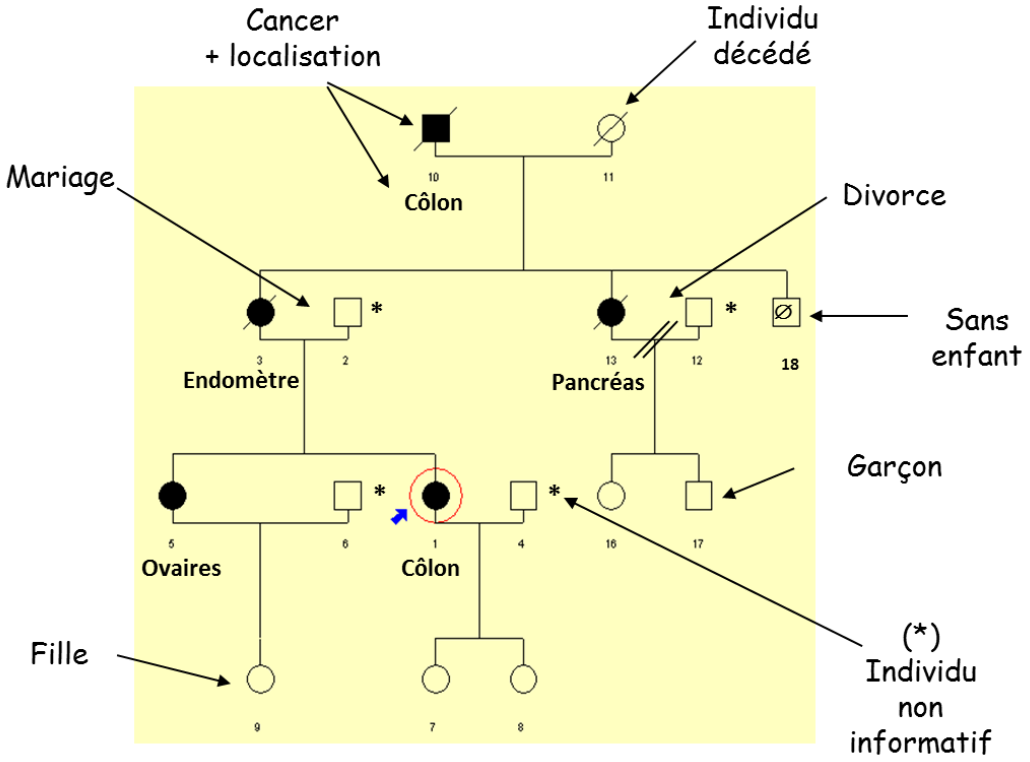


Figure 13 : principaux symboles utilisés pour le tracé des arbres généalogiques

La fonction de génération ou de simulation d'arbres onco-généalogiques a été rajoutée à POLYGENE spécialement pour la réalisation de cette thèse. Il était nécessaire de valider les diverses approches sur des arbres aux caractéristiques bien définies avant de chercher si elles étaient fonctionnelles sur les arbres de familles réelles. Dans POLYGENE, les paramètres à préciser pour leur production concerne le nombre de générations à suivre (en partant du couple initiateur) ainsi que le nombre d'arbres à collecter. Deux options sont disponibles pour la sortie des données : soit elle s'effectue dans des feuilles EXCEL (une par arbre), soit c'est une sortie fichier (.TXT ou .CSV) qui peut être choisie. Le contenu des données est identique dans les deux cas.

En cas de génération d'arbres pour un nombre restreint de familles, on sélectionne pour chaque mutation le nombre choisi de familles auxquelles s'ajoutent autant de familles non mutées. Par exemple, si l'on veut produire 5 arbres par mutation parmi 500 familles au départ sachant que chaque mutation est présente dans 2% des cas, on va choisir 5 arbres pour la mutation 1, 5 pour la mutation 2, etc. et 5 pour des familles sans mutation, soit au total 30 arbres à générer.

La programmation de cette fonction a été relativement complexe, en particulier en raison des remariages suite au décès d'un des deux conjoints. L'option choisie a été de suivre le conjoint vivant s'il appartient en ligne directe à la famille pour laquelle un arbre est constitué. Lors des mariages, si les deux conjoints appartiennent à des familles suivies, le conjoint principal est celui appartenant à une famille "mutée". Si les deux individus qui se marient proviennent de familles mutées sélectionnées pour un tracé, il y a un tirage au sort pour savoir qui rejoint la famille de l'autre. Celui qui part voit son enregistrement transféré sans sa parentèle dans la nouvelle famille, l'enregistrement initial devenant inerte, mais restant afin de garder trace de son existence et permettre le tracé de l'arbre.

Une autre gageure rencontrée a concerné la détermination d'une clef de tri qui puisse faciliter le tracé de l'arbre ensuite. Cette clef de tri se construit séquentiellement à chaque génération à partir de la clef de ses parents. Elle n'apparaît pas sur les arbres, mais elle constitue un artifice indispensable pour présenter les données individuelles lors de la constitution de l'arbre. La routine peut alors accéder alors aux membres de la famille de manière rétro-chronologique. La construction de l'arbre se fait à partir de la dernière génération (souvent la plus nombreuse), en remontant ensuite aux parents, aux parents des parents, et ainsi de suite. Quand on rencontre des individus sans enfant ni fausses couches, il suffit alors d'insérer des lignes (ou des colonnes selon le type de présentation) en prenant soin à ne pas couper les fratries des générations postérieures. Enfin, pour les algorithmes qui sont développés sur les données généalogiques, la clef de tri peut servir à identifier les liens entre les individus d'une même branche mais appartenant à des générations différentes.

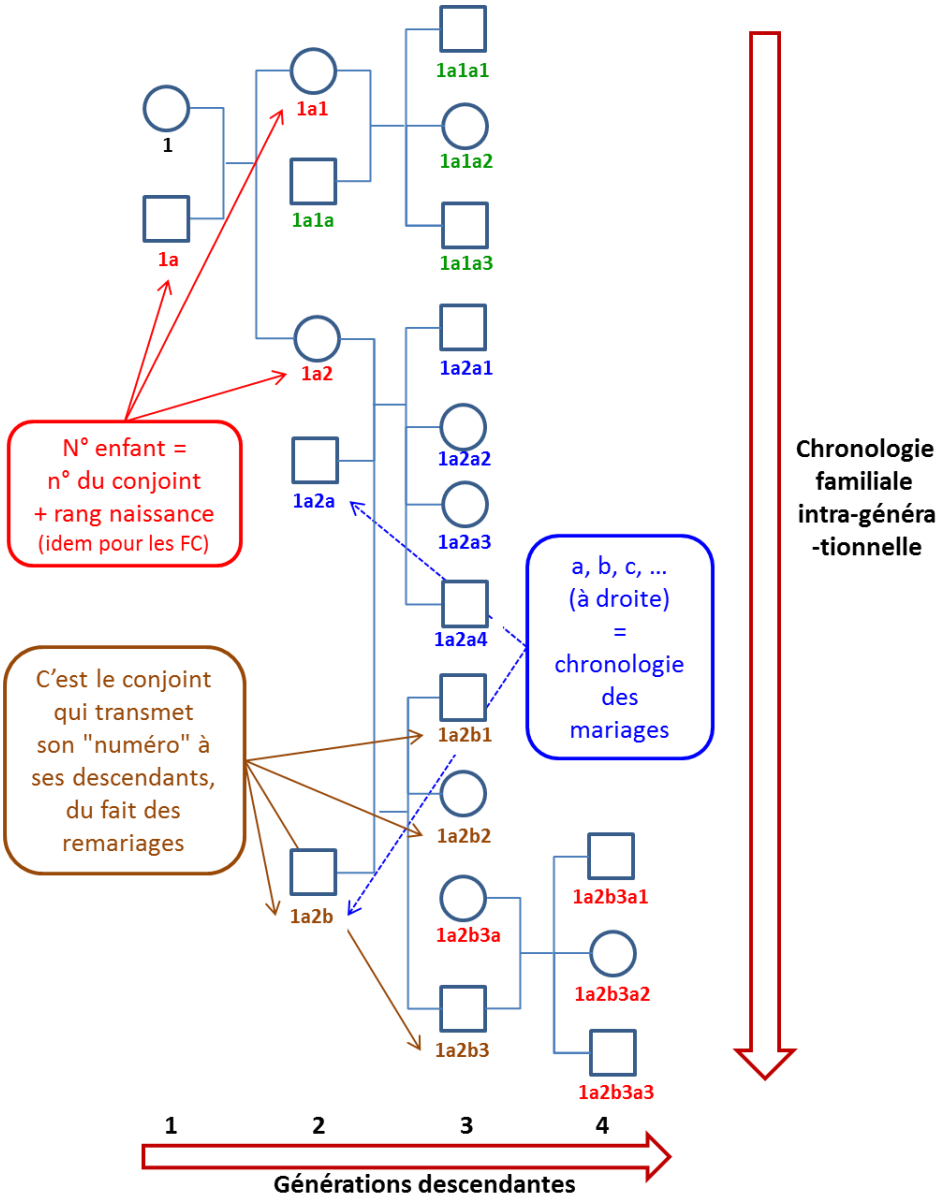


Figure 14 : constitution de la clef de tri (sous chaque cercle ou chaque rectangle) basée sur une hiérarchisation descendante intergénérationnelle. NB. « conjoint » = personne qui entre dans la famille par "mariage" ; « FC » = fausse couche

Les données générées par POLYGENE ont l'aspect suivant :

n° membre	n° famille	famille origine	Géné-ration	Né(e) en	Sexe	n° mère	Age mère	N° père	Age père	N° foyer /tri	Mutations	Malfor-mation	Cancer	Age cancer	Age 1er enf.	Age der. enf.	Age ménop.	Etat	Age décès	Nb mariages	Conjoint 1	Conjoints 2, 3, 4, 5				Age mariage 1	Age mariage 2, 3, 4, 5				Nb enfants	Nb F-C
3	2		1	-20	M					_01		0	0	0	0	0		1	76	1	4					20					3	1
4	2		1	-20	F					_01a	_10000	0	0	0	27	0	0	0	0	1	3					20					3	1
30	2		2	3	M	4	23	3	23	_01a01	_10000	0	0	0	0	0		1	69	1	86					25					0	0
86	2		2	1	F					_01a01a		0	1	28	0	0	0	0	0	1	30					28					0	0
43	2		2	6	F	4	26	3	26	_01a02		0	0	0	28	28	51	1	81	1	70					18					1	0
70	2		2	2	M					_01a02a		0	0	0	0	0		1	69	1	43					22					1	0
47	2		2	7	?	4	27	3	27	_01a03		0	0	0	0	0		1	0	0											0	0
50	2		2	8	F	4	28	3	28	_01a04	_10000	0	0	0	27	27	50	1	87	1	127					27					0	1
127	2		2	11	M					_01a04a		0	0	0	0	0		1	56	1	50					23					0	1
118	2		3	34	F	43	28	70	32	_01a02a01		0	1	87	25	31	48	2	95	1	203					23					2	1
203	2		3	31	M					_01a02a01a		0	0	0	0	0		1	85	1	118					26					2	1
128	2		3	35	?	50	27	127	23	_01a04a01		0	0	0	0	0		1	0	0											0	0
220	2		4	59	?	118	25	203	28	_01a02a01a01		0	0	0	0	0		1	0	0											0	0
229	2		4	60	F	118	26	203	29	_01a02a01a02		0	0	0	25	28	52	1	70	1	432					25					3	0
432	2		4	57	M					_01a02a01a02a		0	1	28	0	0		0	0	1	229					28					3	0
270	2		4	65	M	118	31	203	34	_01a02a01a03		0	0	0	0	0		1	80	0											0	0
433	2		5	85	M	229	25	432	28	_01a02a01a02a01		0	0	0	0	0		1	73	1	844					26					0	0
844	2		5	87	F					_01a02a01a02a01a		0	1	24	0	0	0	0	0	1	433					24					0	0
449	2		5	86	F	229	26	432	29	_01a02a01a02a02		0	0	0	29	35	50	1	85	1	924					29					4	2
924	2		5	88	M					_01a02a01a02a02a		0	1	26	0	0		0	0	1	449					26					4	2
487	2		5	88	F	229	28	432	31	_01a02a01a02a03		0	0	0	22	23	46	1	74	1	752					17					2	0
752	2		5	86	M					_01a02a01a02a03a		0	0	0	0	0		1	62	1	487					19					2	0
925	2		6	115	M	449	29	924	26	_01a02a01a02a02a01		0	0	0	0	0		1	51	0											0	0
951	2		6	116	?	449	30	924	27	_01a02a01a02a02a02		0	0	0	0	0		1	0	0											0	0
982	2		6	117	M	449	31	924	28	_01a02a01a02a02a03		0	0	0	0	0		1	69	0											0	0
998	2		6	118	M	449	32	924	29	_01a02a01a02a02a04		0	0	0	0	0		1	65	0											0	0
1020	2		6	119	?	449	33	924	30	_01a02a01a02a02a05		0	0	0	0	0		1	0	0											0	0
1069	2		6	121	M	449	35	924	32	_01a02a01a02a02a06		0	0	0	0	0		0	0	0											0	0

Figure 15 : données générées par POLYGENE par arbre généalogique

Le tracé de l'arbre généalogique résultant a nécessité le développement d'une routine particulière. Elle a été codée en VBA sous EXCEL puisque les données de POLYGENE ont été directement prévues pour figurer dans une feuille EXCEL. Ce tracé était indispensable pour s'assurer que les arbres générés avaient des configurations conformes aux attentes, c'est-à-dire aux paramètres entrés. Par ailleurs, EXCEL permet assez simplement leur réalisation sur de nombreuses générations, visibles partiellement ou entièrement en agrandissant ou en rétrécissant l'affichage, à l'aide d'un roulement de molette de la souris. Les liens interindividuels sont dessinés en utilisant les propriétés des bordures des cellules (épaisseur du trait).

Pour constituer les arbres dans EXCEL, une nouvelle symbolique a dû être définie, non conforme aux standards de présentation des arbres généalogiques (cf. Figure 13 : **principaux symboles utilisés pour le tracé des arbres généalogiques**) mais plus pratique pour notre recherche : voici la description des symboles utilisés pour cette nouvelle représentation :

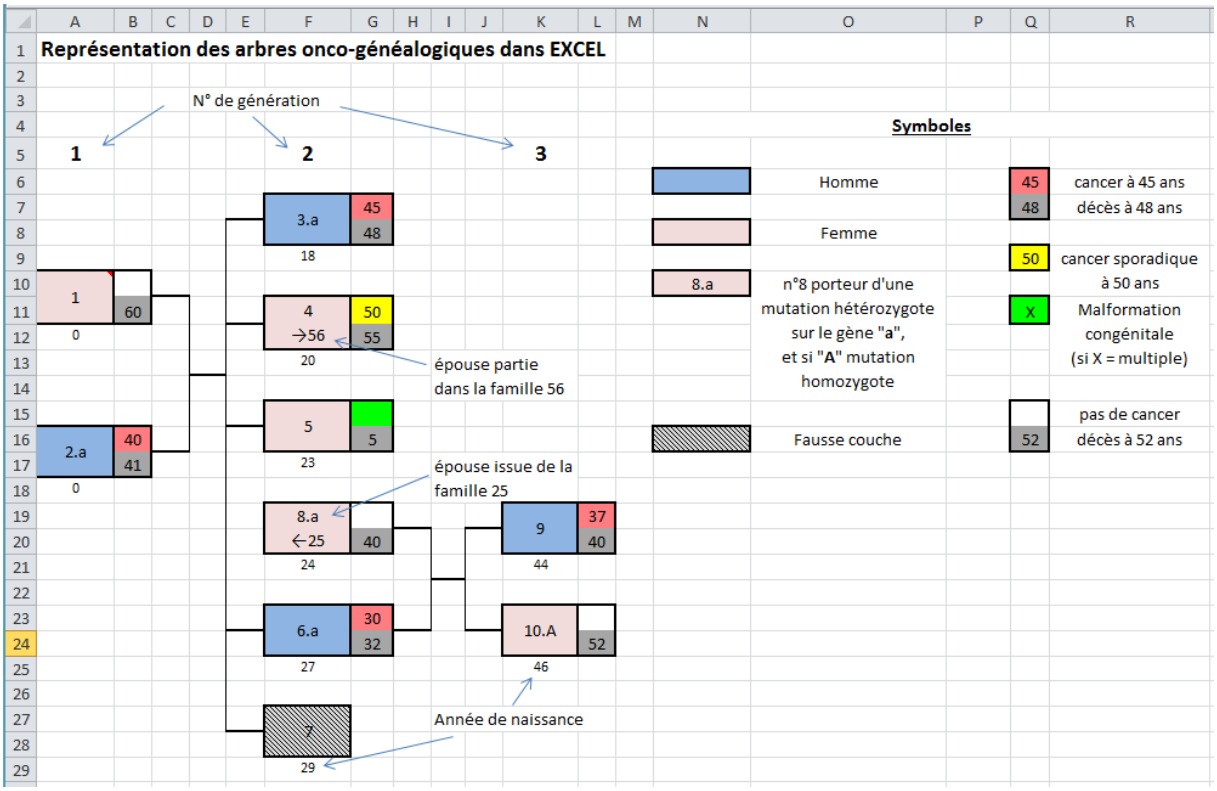


Figure 16 : description des symboles utilisés pour constituer les arbres généalogiques dans EXCEL

Une telle représentation a l'avantage d'afficher, grâce à un système de couleurs (le rose pour les filles et le bleu pour les garçons…), presque tous les paramètres intéressants d'un arbre onco-généalogique. En outre, il est possible d'insérer dans chaque cellule individuelle un commentaire avec, *in-extenso*, les caractéristiques de l'individu correspondant. Nous avons ajouté la notification des malformations congénitales ainsi que le distinguo entre cancer sporadique (en jaune) et cancer lié à la mutation(s) familiale(s) en rouge. L'homozygotie ou l'hétérozygotie d'une mutation est représentée par une lettre majuscule ou minuscule accolée au n° de sujet (ex. mère 8.a, père 6.a et fille 10.A).

Pour clore ce chapitre, voici le résultat final d'un tracé réalisé sur une des familles générées par POLYGENE :

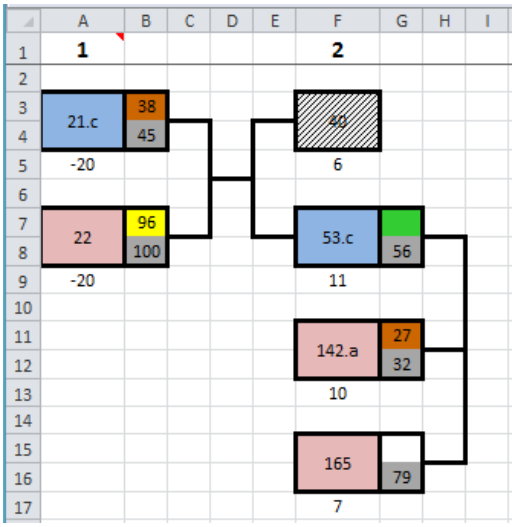


Figure 17 : arbre généalogique généré par la routine VBA dans EXCEL concernant une famille sans descendance au-delà de 2 générations. Contexte moderne et mutation familiale favorisant le cancer autour de 40 ans.

Voici comment lire le diagramme de la Figure 17 : arbre généalogique généré par la routine VBA dans EXCEL concernant une famille sans descendance au-delà de 2 générations. Contexte moderne et mutation familiale favorisant le cancer autour de 40 ans.

- membre n° 21, cellules A-B:3-4, masculin, né en l'an -20 est porteur de la mutation **c** (la 3ème sur 5 programmées). Il a été atteint d'un cancer à l'âge de 38 ans et est décédé à 45 ans.
- Membre 40 : fausse couche
- Membre n°53 : né en l'an 11 avec une malformation congénitale unique et porteur de la mutation familiale **c**. Marié à la femme n°142, issue d'une famille mutée **a** (1ère mutation sur les 5) mais décès de cette dernière à 32 ans suite à un cancer à 27 ans. Remariage avec le n°165. Sans enfant issu de ces 2 mariages => disparition de la famille
- Membre n°22 : cancer sporadique à l'âge de 96 ans et décès à 100 ans.

Voici un autre exemple partiel montrant la capacité de la routine dessinant les arbres à prendre en charge des familles de plusieurs centaines de membres :

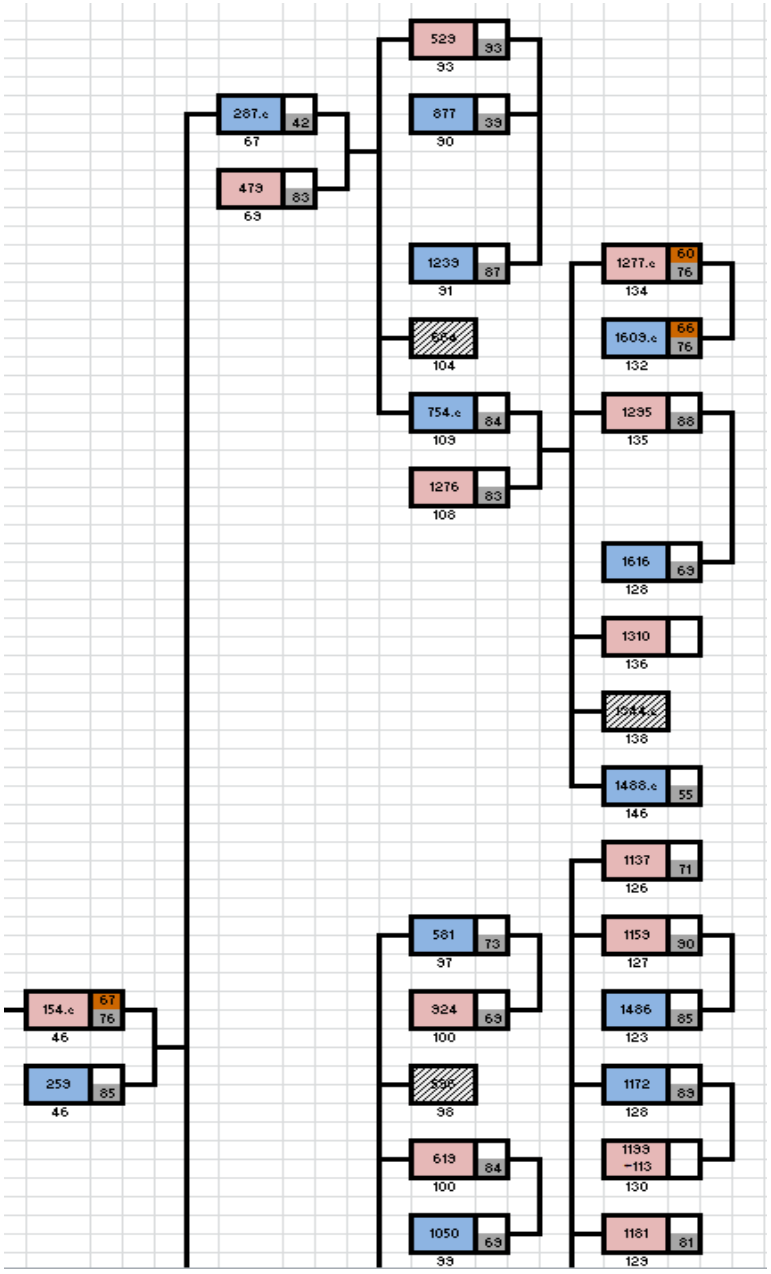


Figure 18 : vue partielle d'un arbre généalogique pour une famille de 236 membres avec mutation sur le gène n°3 (lettre c), montrant la transmission de cette mutation

3 Modélisation des arbres généalogiques : les concepts de "sous-arbre" et de "profil"

3.1 Contexte

Une grande partie de ce travail a été l'objet des stages de master I et II en « statistiques et traitement des données » de Marie ARBRE durant les années 2012 et 2013 puis de sa mission suite à son embauche durant deux ans au centre Jean Perrin grâce au financement FEDER-Conseil Régional obtenu pour ce projet. La question que nous avions à résoudre à cette époque était de trouver une méthode permettant de comparer entre eux les arbres généalogiques utilisés en oncogénétique, et donc d'évaluer et de comparer les spécificités génétiques des différentes familles à partir de leur phénotype. La difficulté est que les arbres généalogiques sont multiformes, comme on peut le voir dans la

Figure 19 avec deux exemples de familles aux effectifs très différents :

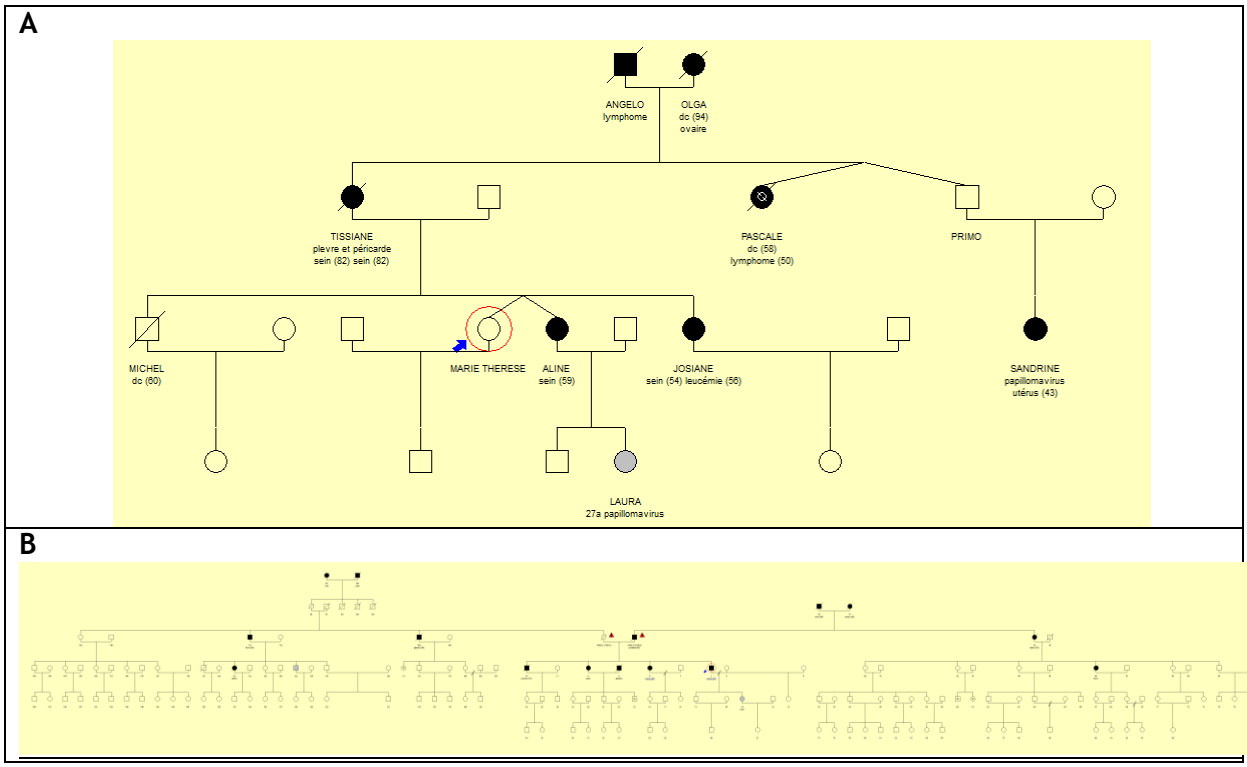


Figure 19 : exemple d'arbres généalogiques utilisés en oncogénétique

Le second arbre (B) présente une difficulté supplémentaire, celle de posséder deux branches, toutes deux porteuses d'un risque héréditaire propre. Dernière remarque, les générations les plus basses, c'est à dire les plus récentes, sont la plupart du temps peu informatives car ne contenant que des enfants ou de très jeunes adultes chez qui aucun cancer n'a eu le temps de se développer, même en présence d'une mutation très pénétrante.

3.2 La solution initiale

Avec Marie Arbre, nous avons mis au point une méthode produisant des « squelettes » d'arbres agrégeant sur deux ou trois générations l'ensemble des membres partageant un lien de filiation. Un tel squelette totalise toutes les occurrences de triplets mère-fils-fille et père-fils-fille que l'on trouve dans un arbre complet, exclusion faite des personnes non porteuses de l'information génétique "utile". Quelles sont donc ces personnes ? Ce sont les individus arrivant dans la famille suite à un mariage et qui ne sont pas porteurs (*a priori*) de la mutation délétère trouvée dans la famille d'accueil : ils ne sont donc pas prédisposées au cancer. Voici dans la Figure 20-A un exemple de sous-arbre deux générations puis un second s'étalant sur 3 générations avec la proportion des divers cancers retrouvés dans la famille par groupe de membres :

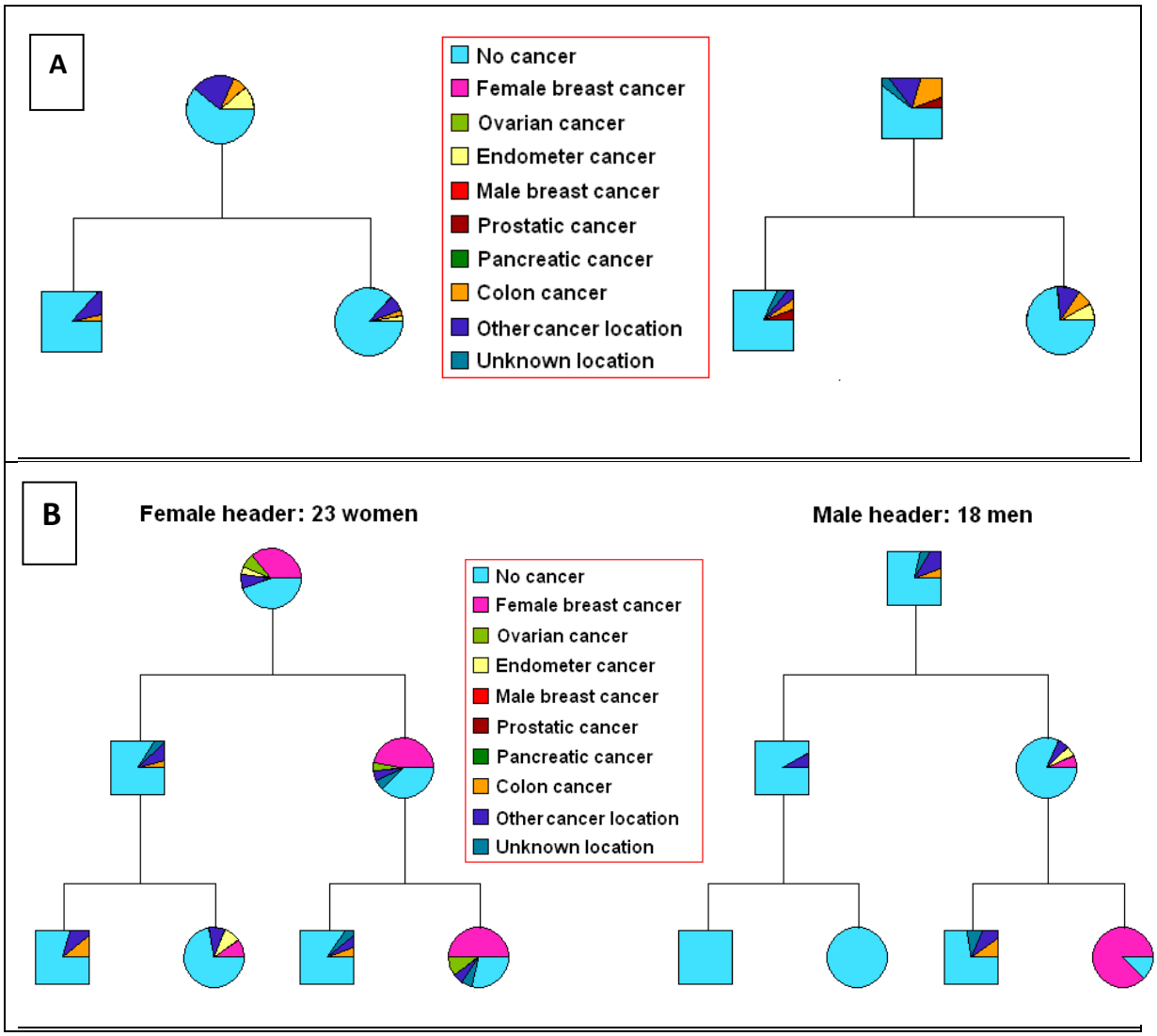


Figure 20 : modélisation des arbres généalogiques en « sous-arbres » synthétiques
A – arbres 2 générations, B – arbres 3 générations

Une fois ces structures en "sous-arbres" réalisées, elles peuvent être non seulement comparées mais aussi fusionnées puisqu'elles ont la même forme (par exemple tous les arbres généalogiques des familles porteuses d'une mutation MMR) ce qui produit ce que nous avons nommé "profils". L'intérêt de ces derniers est qu'ils associent à chaque variable des moyennes et des écarts-types nettement plus « denses »,

c'est à dire portant sur des effectifs très conséquents : il devient alors relativement facile pour chaque nouvelle famille d'établir son sous-arbre puis de déterminer de quel profil elle s'approche le plus : au final cela revient à calculer la probabilité qu'un type de mutation est présent dans la famille connaissant le profil des familles ayant cette mutation. Ces travaux ont été publiés en 2016. L'article est joint en fin de thèse.

3.3 Nouvelle approche

Depuis les années 2012-2013, la problématique a toutefois partiellement évolué : l'idée que des prédispositions au cancer ne soient pas uniquement dues à des mutations uniques à forte pénétrance, mais aussi à des mutations non ou faiblement pénétrantes sauf quand celles-ci se retrouvent associées à des polymorphismes assez répandus mais non délétères habituellement, induit des changements importants dans la constitution des sous-arbres. En effet, les personnes qui étaient écartées précédemment – arrivant par mariage dans une famille sans être porteuses de la mutation délétère familiale – doivent être désormais associées à la construction des sous-arbres car elles peuvent être porteuses de polymorphismes à « risque associatif ». Comme ces polymorphismes peuvent être très fréquents dans la population générale, il faut alors tenir compte de ces personnes, non seulement dans les sous-arbres, mais aussi dans nos modélisations et nos simulations. La différenciation des filiations paternelles et maternelles devient de ce fait moins importante et les sous-arbres peuvent être simplifiés. Ainsi en est-il de l'exemple de la Figure 21 tracé à partir de la famille simulée n° A40-3 :

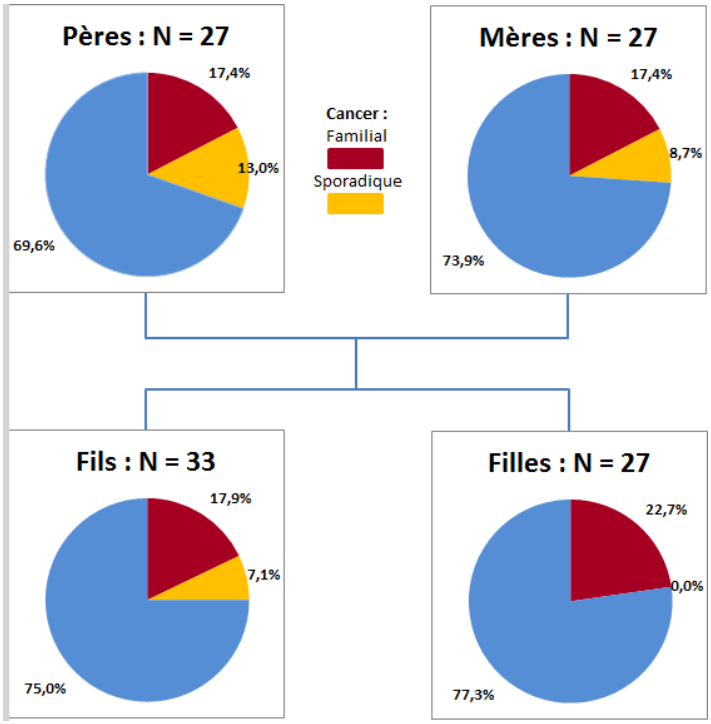


Figure 21 : sous-arbre 2-générations simplifié avec la proportion de cancers rapportée au nombre d'individus

Comme ce sous-arbre est issu de nos données simulées, seulement deux types de cancer y figurent : les cancers familiaux (secteur en rouge) et les autres sporadiques (en jaune). Ce sous-arbre est tracé grâce au développement d'une routine en VBA sous EXCEL. Cette routine fournit les informations sous forme de tableau contenant les données agrégées par position dans le sous-arbre, qui peuvent ensuite être exportées pour analyse statistique :

Tableau 2 : données servant pour la génération du sous-arbre (à partir d'une famille générée par POLYGENE porteuse d'une mutation favorisant les cancers à 50 ans)

		Pères - G1	Fils-P	Filles - P	Mères - G1	Fils-M	Filles - M
Individus	Nombre	27	33	27	27	33	27
cancers (tous types)	Nombre	7	7	5	6	7	5
	Age moyen	60,86	58,14	43,00	64,67	58,14	43,00
	Ecart-type	21,19	17,85	21,90	18,53	17,85	21,90
cancers familiaux	Nombre	4	5	5	4	5	5
	Age moyen	43,25	47,20	43,00	52,00	47,20	43,00
	Ecart-type	7,76	5,19	21,90	5,79	5,19	21,90
Cancers sporadiques	Nombre	3	2	0	2	2	0
	Age moyen	84,33	85,50	85,00	90,00	85,50	85,00
	Ecart-type	1,70	0,50	0,00	0,00	0,50	0,00
Décès	Age moyen	70,48	70,33	79,92	82,72	70,33	79,92
	Ecart-type	16,01	13,85	17,42	13,90	13,85	17,42
Enfants	Nombre moyen	2,222	1,227	2,000	2,222	1,227	2,000
	Ecart-type	1,315	1,379	1,549	1,257	1,379	1,549
Malformations chez les enfants	taux uniques	1,67%			1,67%		
	taux multiples	3,03%			3,03%		
Fausses couches	Taux			14,29%	13,04%		14,29%
Individus indemnes	Nombre	20	26	22	21	26	22

Ces données permettent déjà de calculer certains indicateurs intéressants comme le taux de transmission de la prédisposition au cancer :

Tableau 3 : taux de transmission intergénérationnel de la prédisposition au cancer. X = fils et filles ensemble et pères et mères ensemble. N = nombre de liens trouvés entre des (grands-)parents cancéreux et leurs (petits-)enfants.

	Taux de transmission (%)		
	Fils	**Filles**	**X**
Pères	16,7%	40,0%	27,3%
Mères	37,5%	66,7%	45,5%
X	28,6%	50,0%	36,4%
		IC-95 bas	16,3%
		IC-95 haut	56,5%
		N	22
	Petits-fils	**Petites-filles**	**X**
Grands-pères	50,0%	40,0%	42,9%
Grands-pères	25,0%	16,7%	21,4%
X	30,0%	27,3%	28,6%
		IC-95 bas	9,2%
		IC-95 haut	47,9%
		N	21

Ce calcul de l'héritage du risque de cancer porte uniquement sur les parents ayant eu un cancer et leurs enfants qui sont ici au nombre de 22. Parmi ces 22 enfants, 36.4% ont développé à leur tour un cancer familial. Ce taux est nécessairement inférieur à 50% puisque les enfants ont une chance sur deux d'hériter de la mutation délétère, que la pénétrance, bien que fixée à 90% dans cet exemple n'impacte pas tous les descendants mutés car ceux-ci peuvent décéder d'autre cause avant toute apparition de cancer. Le taux à 28.6% de transmission aux petits enfants est moindre qu'entre parents et enfants, ce qui est attendu puisqu'on s'attend à un taux divisé par deux.

Enfin, un dernier éclairage sur les caractéristiques du risque héréditaire de cancer peut être obtenu à partir de ce sous-arbre sous la forme de courbes de survie sans cancer. Bien sûr, le terme de survie est inadéquat ici, mais il fait référence à la méthode de calcul, celle de Kaplan-Meier[49]. Les courbes sont tracées tant pour les parents que pour les enfants. N'oublions pas que les parents appartiennent souvent aussi à la catégorie des enfants, puisqu'ils descendent eux-mêmes d'autres parents potentiellement porteurs du risque : la limite supérieure est fixée par le nombre de générations prises en compte ou simplement renseignées. Quant à la dernière génération (la plus récente), qu'elle soit prise en compte ou pas n'a guère d'importance. C'est l'intérêt de la censure dans la méthode Kaplan-Meier : l'information n'est utilisée que sur sa durée connue, mais elle l'utilise.

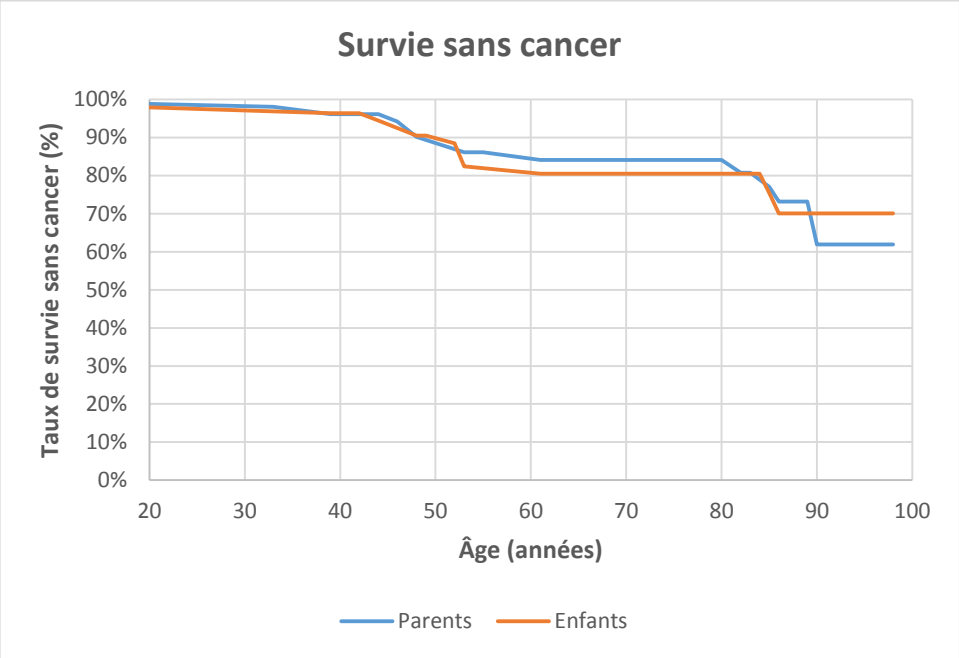


Figure 22 : courbes de survie sans cancer pour les membres de la famille A40-3

Sur ces deux courbes assez superposées, on remarque deux « chutes », une autour des 50 ans et l'autre après 80 ans. La première correspond à l'arrivée des cancers familiaux, la mutation de la famille A40-3 correspondant justement à un risque de cancer apparaissant autour de la cinquantaine. La seconde chute correspond aux cancers sporadiques paramétrés pour se développer tardivement. On notera que l'axe des temps (abscisses) débute à 20 ans.

3.4 Description du jeu de données utilisé dans les deux parties suivantes

Pour réaliser les jeux de test, un ensemble de 300 familles a été constitué à partir d'un millier générées par POLYGENE. En effet, la génération automatique des arbres produit de nombreuses familles qui s'éteignent rapidement, d'autres sans mutation qui récupèrent par alliance des mutations délétères et deviennent de ce fait non représentatives de leur catégorie, et bien sûr aussi la situation inverse arrivait souvent : des familles mutées au départ avec des enfants ne portant pas la mutation. Il était donc nécessaire d'épurer ces familles afin de ne garder à la fin que celles ayant les caractéristiques attendues, ce qui ne peut se faire que manuellement. Au final, les familles conservées avaient les caractéristiques suivantes :

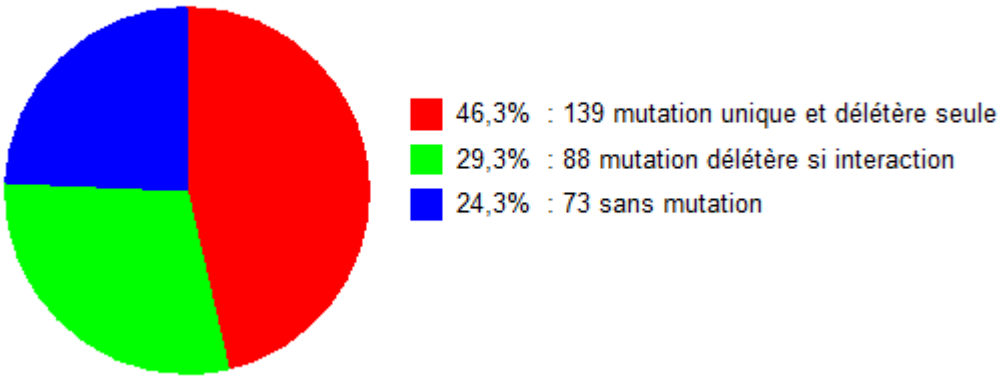


Figure 23 : répartition des 300 familles selon les 3 catégories principales de risque

Dans les familles avec mutations, 32.5% avaient une mutation favorisant un cancer programmé autour de 30 ans, 33.3% autour de 40 ans et 34.2% autour de 50 ans. Ainsi, le groupe « témoins » avait une population en nombre de familles similaire aux autres si l'on considère les groupes d'âge « cible ».

Les cancers générés par POLYGENE selon la catégorie de risque étaient répartis ainsi :

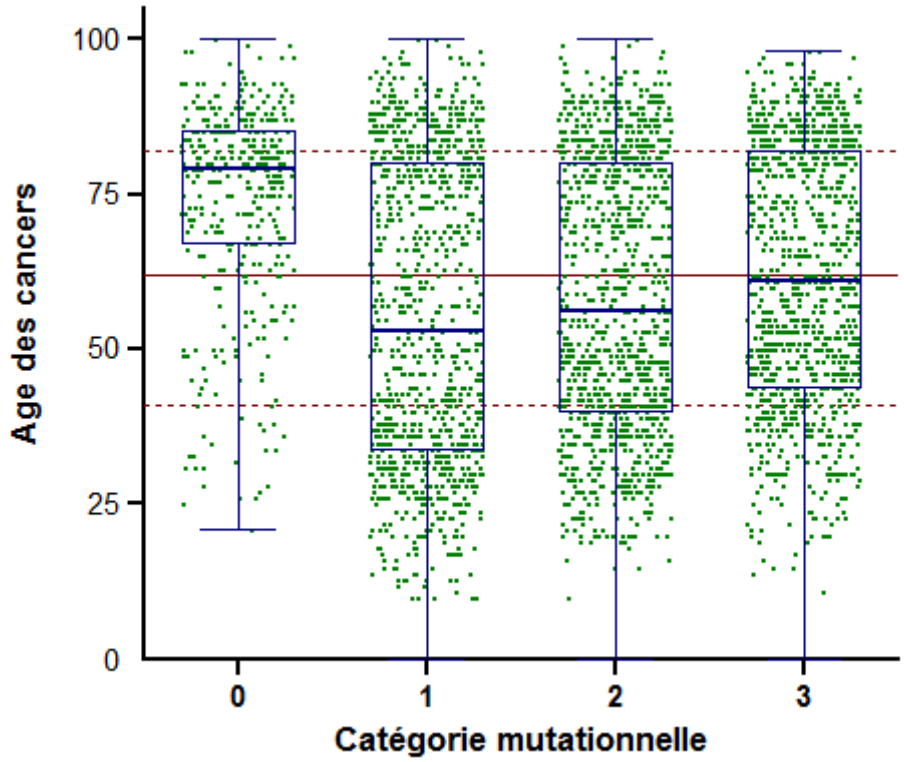


Figure 24 : répartition des âges de survenue des cancers selon la catégorie mutationnelle

Evidemment, on retrouve sur ce graphe les deux groupes de cancers générés par POLYGENE, avec en haut les cancers sporadiques qui surviennent autour de 80 ans (± 10 ans) indépendamment de la catégorie mutationnelle de la famille des individus. Quant aux cancers d'origine héréditaire, ils se différencient nettement et arrivent plus précocement dans la catégorie 1 que dans la 2 et idem entre les catégories 2 et 3. Respectivement l'âge de survenue de ces cancers chez les individus est de 35.7 ± 12.7 ans (N = 544), 40.6 ± 12.5 (N = 598) et 44.8 ± 13.0 (N = 572). On ne retrouve pas la séparation de 10 ans censée exister entre ces 3 groupes. Ceci est dû à un grand nombre de « gains » de mutation par les familles dans notre jeu de données (des mutations entrent dans la famille suite à des unions). C'est en raison de la fréquence paramétrée des mutations dans nos générations, fixée à 5% au départ alors que dans la réalité, elle est habituellement très inférieure à 1% pour les principales mutations favorisant le cancer. Cela pénalisera forcément les tests sur la performance des approches. Procéder autrement, c'est à dire avec une fréquence de 1% dans POLYGENE par exemple, aurait entraîné de multiplier par 5 le nombre initial de familles. Le temps pour en extraire manuellement le même nombre qu'ici aurait été démultiplié.

Une autre caractéristique intéressante de ce jeu d'essai est qu'elle montre l'impact différent qu'ont ces catégories de risque sur la durée de vie sans cancer :

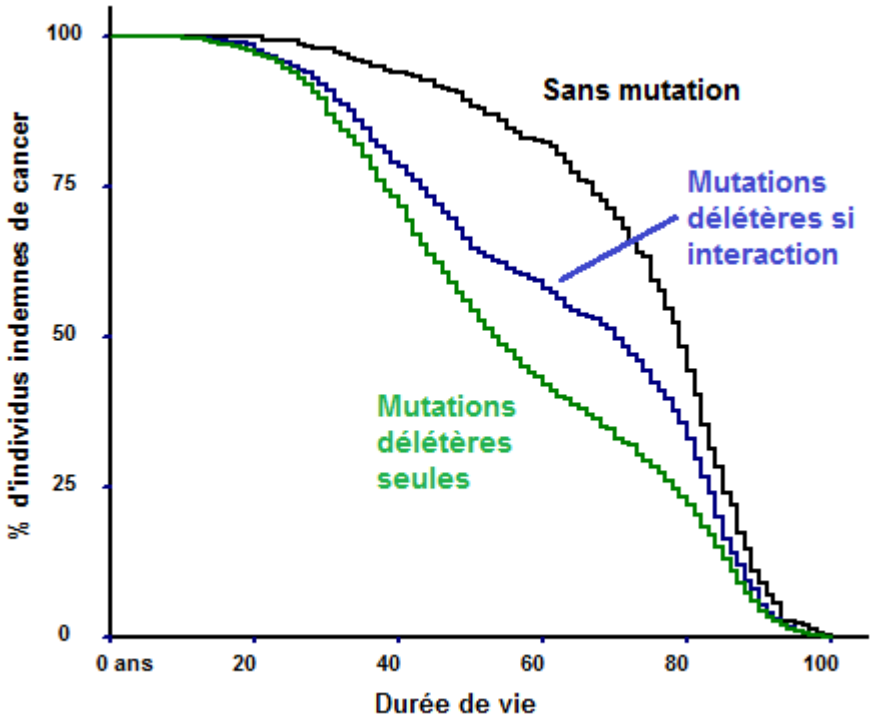


Figure 25 : courbes de "survie" sans cancer selon le type de mutation familiale

Alors que les types de mutation modulant l'âge des cancers (30, 40 ou 50 ans) sont également répartis ($p = 0.94$) dans les deux catégories de risque : mutations délétères seules versus mutations nécessitant une interaction. Les courbes de survie correspondant à ces deux groupes sont significativement différentes ($p < 10^{-7}$) en particulier en raison d'une proportion d'individus cancéreux significativement moindre quand il faut qu'un polymorphisme soit présent pour que l'aspect délétère de la mutation apparaisse : 16.9% $\pm$ 5.7% versus 20.7% $\pm$ 10.2% ($p = 0.0012$). Pour ce qui est de la courbe de survie des familles sans mutation, on note l'impact des cancers sporadiques survenant tardivement.

Un dernier regard sur le jeu de données concerne la quantité de données générées et donc la lourdeur de certains calculs ensuite :

Tableau 4 : dénombrement des familles, des membres et des données associées

	Nombre de fiches	Nombre de réponses
Rubrique : Familles	300	35 351
Rubrique : Membres	27 703	656 875

C'est donc près de 700 000 données qui sont utilisées pour les tests de validation des modèles statistiques des deux parties suivantes. Ces données sont organisées en trois tables principales qui seront traitées à tour de rôle dans nos différents tests :

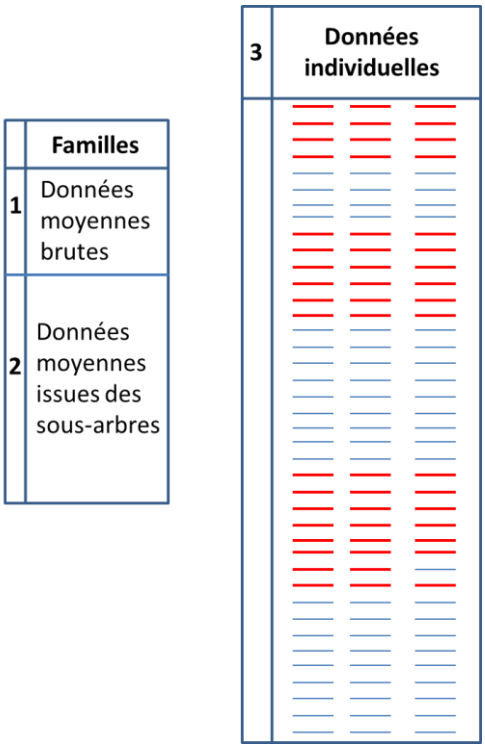


Figure 26 : les trois tables utilisées lors des différents tests dans les parties 4 et 5 de la thèse

Les données de la table 1 contiennent le dénombrement des cancers par famille, leur âge moyen de survenue et l'écart-type associé, le taux de cancer relativement au nombre de membres de chaque famille, l'âge des parents à la naissance, l'âge du mariage, les données de natalité (nombre moyen d'enfants et écart-type, taux de malformations (uniques/multiples), taux de fausses couches, âge moyen aux premier et dernier enfants), l'âge de la ménopause pour les femmes et l'âge du décès pour chacun. Les données de la table 2 contiennent ces mêmes informations, mais déclinées selon la place de le sous-arbre. On y trouve aussi les deux taux de transmission intergénérationnel des cancers, le premier entre parents et enfants et le second entre grands-parents et petits-enfants. Enfin, la table 3 contient des données identiques mais non moyennées pour chaque individu et selon le sexe de chacun.

4 Recherche de groupe de familles à risque spécifique grâce à l'analyse en composantes principales et le K-means clustering

4.1 Intérêt de l'analyse en composantes principales (ACP) pour notre problématique génétique

L'ACP est un outil extrêmement puissant de compression et de synthèse de l'information. C'est particulièrement vrai lorsque l'on est en présence d'une somme importante de données à traiter et interpréter. Nous l'avons développé dans le logiciel SEM. Que fait cette analyse ? Elle produit des variables de synthèse appelées « composantes principales », dont les principales représentent l'essentiel de la variabilité. Ce faisant, la CPA informe sur les corrélations entre variables ainsi qu'on peut le voir sur la figure suivante :

Figure 27 : diagramme des corrélations inter-variables issu de la CPA ; données analysées : données brutes moyennes des 300 familles simulées (table 1 de la Figure 26)

On constate que les variables concernant les malformations (uniques ou multiples) et les fausses couches sont regroupées ensembles tandis que l'âge moyen des cancers est à l'opposé de la variable « nombre de cancers » : en effet, plus il y a des cancers dans une famille et plus il y a de chances que ce soit d'origine familial et donc qu'ils surviennent jeune. L'âge des cancers est donc situé près du point « sans mutation » car c'est là que l'âge de survenue est le plus élevé.

Un autre intérêt de l'ACP est de proposer des regroupements des « individus » ayant des caractéristiques communes : c'est principalement cette capacité de l'ACP que l'on utilisera dans notre recherche des familles ayant des risques de cancer spécifiques mais non identifiés. Voici comment sont classées nos 300 familles en réalisant l'ACP sur les données moyennées brute :

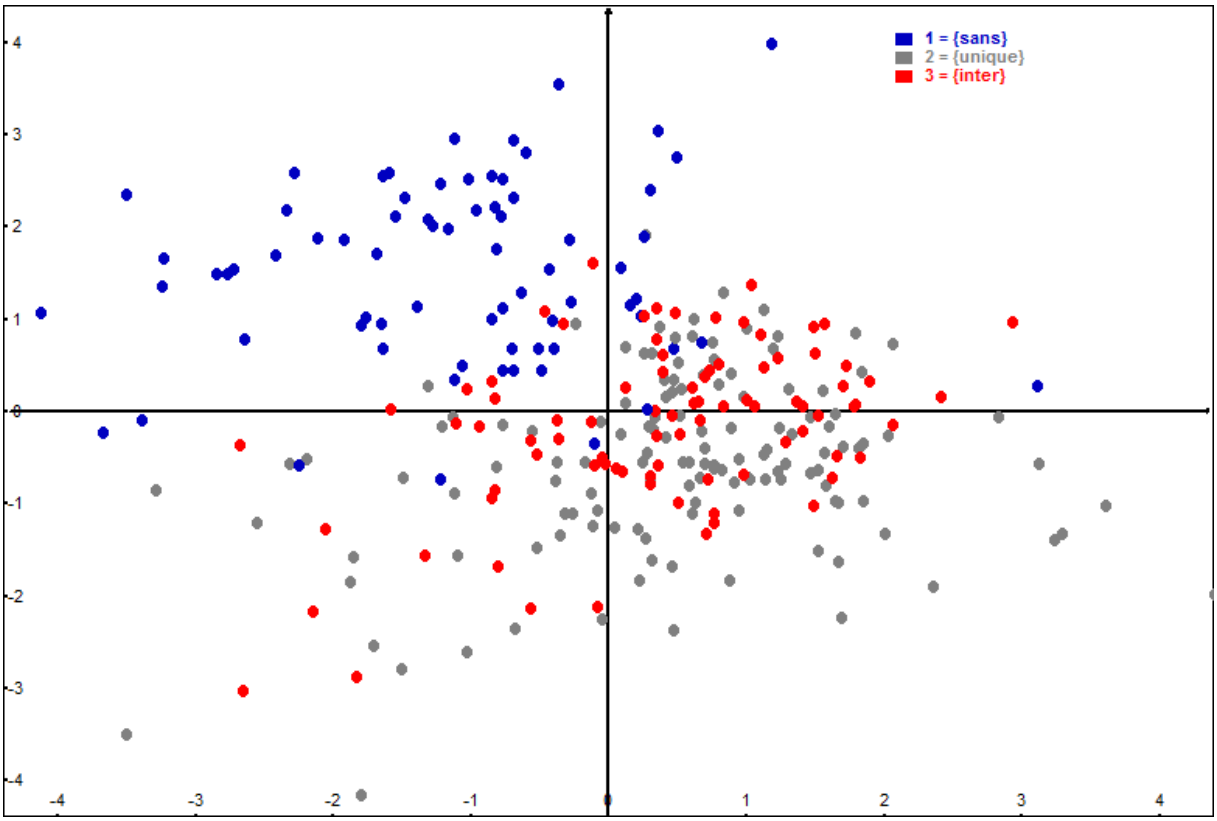
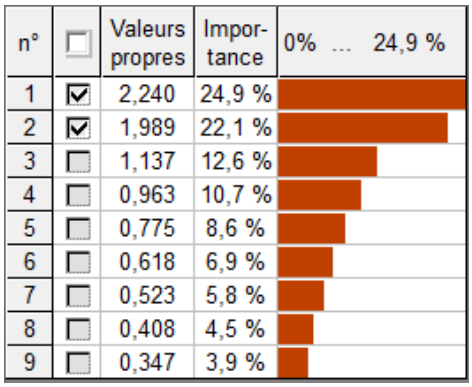


Figure 28 : graphe de répartition des familles produit par la même ACP en considérant les 2 premiers axes (les couleurs correspondent à la catégorie de risque familial)

Ce type de graphe est relativement facile à interpréter : tout d'abord, il faut noter que les deux premiers axes résument à eux seuls 47% de la variance globale ce qui est plus du double de ce que l'on pourrait attendre puisque nous avons 9 variables traitées.



n°		Valeurs propres	Impor-tance	0% ... 24,9 %
1	☑	2,240	24,9 %	
2	☑	1,989	22,1 %	
3	☐	1,137	12,6 %	
4	☐	0,963	10,7 %	
5	☐	0,775	8,6 %	
6	☐	0,618	6,9 %	
7	☐	0,523	5,8 %	
8	☐	0,408	4,5 %	
9	☐	0,347	3,9 %	

Figure 29 : "éboulis" des valeurs propres de la matrice de covariance montrant l'importance prépondérante des deux premiers axes

Ensuite, la répartition des points de couleur est elle aussi instructive : si les points bleus semblent bien se démarquer (familles sans mutation), les points orange et gris sont assez mélangés : l'ACP sur les données brutes moyennées n'arrive donc pas à distinguer les familles porteuses de mutations délétères à elles seules des familles où un polymorphisme est nécessaire pour que les mutations deviennent délétères. On peut confirmer ce point en effectuant une analyse en clusters par la méthode des k-moyennes (*k-means clustering).* Voici le tableau de contingence obtenu à partir des coordonnées des familles sur les deux principaux axes de l'ACP :

Tableau 5 : regroupement des types de risque familial suite à une analyse en cluster en prenant en considération 47% de l'information initiale

Catégories de mutation	cluster 1	cluster 2	cluster 3	total
sans	6	4	63	73
unique	112	22	5	139
interaction	67	12	9	88

Les familles non mutées se trouvent regroupées pour 86% d'entre elles dans le cluster n°3 tandis que les familles ayant une mutation unique délétère et celles porteuse d'une mutation nécessitant une interaction se distribuent de manière proportionnelle dans les deux premiers clusters (61% et 58% respectivement pour les mutations uniques versus 36% et 32% pour celles avec interaction). La prise en compte de 70% de l'information en incluant les axes 3 et 4 n'amène quasiment aucune amélioration dans la discrimination.

4.2 Evaluation de l'intérêt des sous-arbres pour améliorer la discrimination des familles selon deux catégories de risque proches

Comme on vient de le constater ci-dessus, la difficulté n'est pas d'isoler les familles non exposées à un risque héréditaire de cancer parmi d'autres porteuses de mutations délétères, mais de faire le tri entre divers types de susceptibilité génétiques. Ici, nous avons grâce aux données simulées, un contexte composé de familles avec un risque un peu minoré du fait que la mutation délétère requiert la présence d'un polymorphisme bénin qu'il s'agit de distinguer d'autres familles porteuses de mutations ayant une pénétrance de type BRCA1-BRCA2. C'est bien dans un tel contexte qu'il faut juger

de l'intérêt ou non des sous-arbres puisque les données brutes moyennées par famille suffisent dans le contexte habituel. Si nous trouvons un moyen d'effectuer cette discrimination de manière satisfaisante, alors la même approche pourra être utilisée sur les données réelles avec un peu plus de chances de succès.

4.2.1 Critère de jugement du pouvoir discriminant des analyses en clusters

L'ACP est très utile quand on veut effectuer une analyse en clusters, c'est à dire un partitionnement des individus étudiés dans le but de regrouper ceux ayant des caractéristiques proches. En effet, non seulement l'ACP réduit l'information à ses dimensions essentielles, mais en même temps elle élimine l'effet des points aberrants (*outliers*) auquel on serait confronté si l'on effectuait le clustering directement sur les données. De ce fait, les méthodes de clustering - comme celle utilisée ici des nuées dynamiques (*k-means clustering*) - deviennent nettement plus robustes, c'est à dire moins sujettes à trouver par hasard des clusters non pertinents.

Quand on effectue des simulations, on connaît *a priori* le nombre de groupes d'individus et par conséquent le nombre de clusters à rechercher. Avec deux sous-groupes de familles bien identifiables, on dispose d'un indicateur assez simple de l'efficacité du clustering post ACP. Par exemple, on peut étudier la répartition de nos deux sous-groupes dans chaque cluster résultant : un Chi² est suffisant pour cela. Dans le tableau de contingences précédent, le Chi² était égal à 188 pour un degré de liberté de 4. Un clustering en utilisant 4 dimensions de l'ACP (soit 70% de la dispersion) amène ce Chi² à 196. L'utilisation ici de deux dimensions supplémentaire n'apporte pas grand-chose à la précision de la discrimination. Pour ce qui est de nos deux risques génétiques, on peut réitérer des tests sur des familles tirées au sort parmi les 139 familles ayant une mutation délétère par elle-même et les 88 pour lesquelles la mutation nécessite une interaction. Pour le clustering post-ACP, on se contente des premières dimensions de l'ACP synthétisant au moins 40% de l'information et on fixe le nombre de clusters à 2. En calculant à chaque fois le Chi² 2 classes x 2 clusters (le Chi² peut aussi être considéré comme une distance entre les fréquences attendues et celles observées), le résultat des tests finira par culminer, soit en pointant la similarité des approches, soit en rejetant cette hypothèse nulle.

4.2.2 Résultat des tests itératifs

Une trentaine d'itérations ont été effectuées pour évaluer si le clustering post-ACP classait mieux nos deux groupes de familles quand il utilisait les données des sous-arbres ou bien seulement les données brutes moyennées.

Tableau 6 : Evaluation de valeur discriminante des clusterings post ACP selon que les données étudiées sont celles brutes moyennées par famille ou bien celles issues des sous-arbres

	Données brutes moyennes	Données des sous-arbres
Chi² moyens	1,67	1,23
écart-types	2,01	1,29
test t apparié	p = 0,088	

Si la probabilité associée au test-t des séries appariées qui compare les deux Chi² moyens n'est pas très loin de la significativité, c'est en fait en faveur des données brutes moyennes et non pas de celles des sous-arbres. Réaliser de plus nombreuses itérations ne permettra pas d'inverser le résultat. Ce que l'on peut conclure de cette analyse, c'est que les données issues des sous-arbres ne semblent pas capables d'améliorer le distinguo entre nos deux groupes de familles. De ce fait, il est permis de tenter de faire le partitionnement des données réelles (issues des familles consultant en oncogénétique) sans passer par la réalisation des sous-arbres généalogiques.

4.3 Analyse sur données réelles relatives à des familles exposées à un risque familial de cancer sein/ovaire

4.3.1 Description de la population

Cette population, extraite de la base de données oncogénétique a servi à l'étude de la signification de la présence de cancer du sein *in-situ* pour le pronostic mutationnel des familles à risque sein/ovaire. Un article a été rédigé en 2018 à ce sujet mais n'a pas été accepté dans deux revues différentes pour publication[i]. Nous n'avons pas eu le temps de le soumettre ailleurs. Rappelons que les cancers *in-situ* sont des cancers évoluant lentement, restant localisé au sein donc ne présentant pas le caractère de gravité qu'ont les autres cancers malins. Evidemment, pour poser ce diagnostic il faut avoir analysé une pièce opératoire ou au moins du matériel extrait par biopsie. C'est donc la connaissance anatomopathologique de la tumeur qui fournit ce renseignement, en parallèle avec la typologie des autres contingents cellulaires : une tumeur du sein peut avoir une partie de ses cellules de type malin (carcinome canalaire ou lobulaire invasif) et en même temps des cellules de carcinome *in-situ*. Nous laissons de côté les autres histologies très minoritaires. Ces cellules malignes se distinguent bien sûr par leur architecture, mais aussi par tout un ensemble de caractéristiques (nombre de mitoses, nécrose, cinétique cellulaire, récepteurs hormonaux, marquages immunologiques…) qui vont orienter non seulement le pronostic mais aussi les traitements. Ces paramètres tumoraux sont donc essentiels. Leur autre particularité est qu'ils signent des processus biologiques particuliers et en conséquence des particularismes génétiques, ce qui nous concerne ici. Les tumeurs in-situ, en dépit de leur bon pronostic, ont la capacité de se transformer en tumeur invasive et c'est bien la raison pour laquelle, lorsqu'ils sont dépistés, on les opère. Mais l'existence de tumeur in-situ massifs, c'est à dire prenant le sein en masse, laisse penser que tel n'est pas forcément leur destin et il en est certainement un nombre non négligeable qui n'évolueraient jamais si elles n'étaient pas réséquées. C'est tout le problème du sur-diagnostic que certains ont estimé autour de 50% en occident[50]. Une des conclusions de notre étude cependant, était que dans les familles à risque héréditaire, le passage de l'*in-situ* à l'invasif était la règle et qu'une attitude volontariste dans leur prise en charge dès leur dépistage était la meilleure manière de préserver l'avenir.

[i] Kwiatkowski F, Gay-Bellile M, Privat M, Petit MF, Uhrhammer N, Bidet Y, Mishellany F, Bignon Y-J (2018) Does in-situ breast carcinoma characterize a different type of hereditary cancer predisposition? A retrospective study of 1,798 French HBOC families.

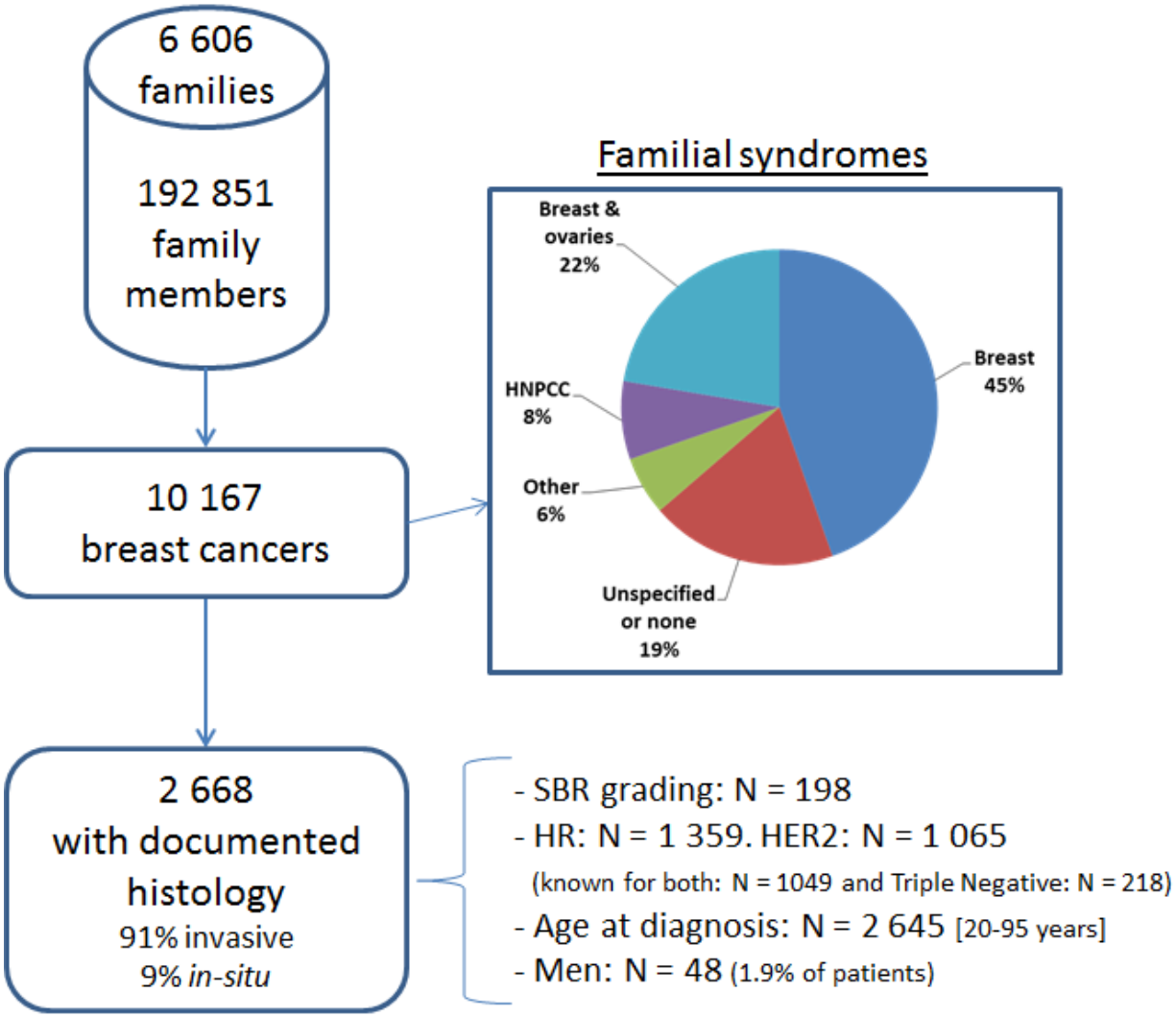


Figure 30 : extraction des patients de la base de données oncogénétique pour l'analyse en composante principale (HNPCC – syndrome familial côlon, HR – récepteurs hormonaux, HER2 – marqueur d'expression du facteur de croissance épidermique, SBR – grade de malignité de Scarff-Bloom-Richardson)

Les familles sélectionnées pour cette étude au nombre de 1 798 étaient celles pour lesquelles au moins un cancer du sein avait un diagnostic histologique. En tout, cela constituait un panel de 2 668 personnes avec un tel diagnostic sur les ≈ 93 000 individus composant ces familles. Pour de nombreuses familles on avait connaissance de l'histologie de plusieurs tumeurs différentes, pour autant de patients : l'empilement de ces diverses caractéristiques plus ou moins complémentaires nous semblait à même de faire émerger des profils génétiques familiaux spécifiques sur la base de statistiques. Voici comment se regroupaient ces types histologiques :

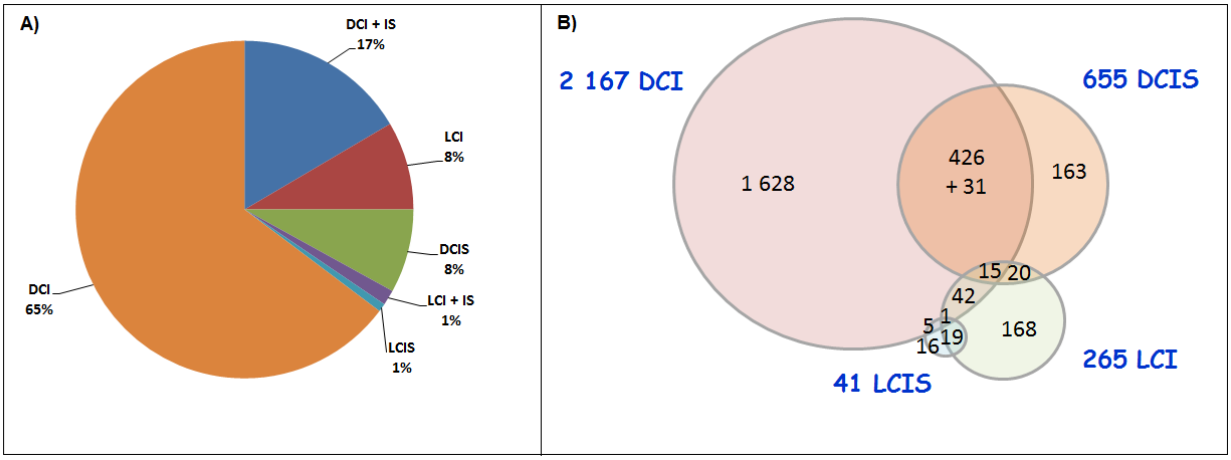


Figure 31 : A - répartition des types histologiques dans la sous-population de familles à risque sein/ovaire et B - types histologiques seuls ou coexistant dans les familles (DCI – carcinome canalaire invasif, DCIS – carcinome canalaire *in-situ*, LCI – carcinome lobulaire invasif, LCIS – carcinome lobulaire *in-situ*, + IS – avec un contingent cellulaire *in-situ*)

Si l'on se limite aux familles dont le diagnostic principal est une prédisposition sein/ovaire et chez lesquelles aucune autre mutation que BRCA1/BRCA2 n'a été diagnostiquée, ce qui inclut les familles sans mutation connue, ce nombre de familles tombe à 1352. Parfois on prendra ce sous-groupe en exemple.

4.3.2 Comment choisir le nombre optimal de clusters ?

C'est une des principales questions quand on réalise des analyses en clusters. Il ne faut tout d'abord pas perdre de vue que les méthodes de clustering « trouvent » toujours autant de clusters qu'on leur demande, que le nombre fourni soit pertinent ou non. De nombreux algorithmes ont donc été développés dans le but d'aider à répondre à cette question : il y en a plus d'une cinquantaine[51] et de nombreux ont des propriétés similaires (Milligan et al. en citait déjà une trentaine en 1985[52]). Ces méthodes peuvent se regrouper en deux grandes catégories[53] : les méthodes « locales » et les méthodes « globales ». Les méthodes locales regardent les clusters deux à deux et évaluent s'ils devraient être amalgamés ou non tandis que les méthodes globales utilisent des indicateurs calculés sur l'ensemble des données et en tant que fonction du nombre de clusters, tentent d'en optimiser la valeur. Il existe aussi des méthodes plus empiriques qui donnent de bons résultats comme le clustering à 2 étapes [Milligan, 1980][54] : on effectue en premier une classification ascendante hiérarchique (cf. partie suivante) et au vu du regroupement des données, on fixe le nombre de partitions du K-means clustering. L'inconvénient d'avoir à fixer arbitrairement la valeur de k demeure alors car il ne peut être automatisé alors qu'on aimerait bénéficier de critères de choix non partiaux.

Intuitivement, on devine que plus grand est le nombre k de clusters et plus efficace est la discrimination. En effet, si l'on a autant de clusters que d'individus inclus dans l'analyse, 100% des individus sont bien classés. Accessoirement, un tel partitionnement aboutit à un gain d'information nul. Il faut donc réfléchir à pénaliser cette tendance naturelle à accroitre le nombre de clusters dans les méthodes cherchant le partitionnement optimal.

Nous avons développé dans notre logiciel SEM plusieurs méthodes utilisant la distance euclidienne que nous détaillons ci-après. Elles sont représentatives des approches les plus fréquentes. Les premières méthodes sont principalement basées sur le calcul de scores mettant en balance les variances intra et inter-clusters. Une autre (la méthode "Silhouette") se focalise sur le bon ou le mauvais classement de chaque item et fournit un score synthétisant l'ensemble des classements. Enfin, nous avons implémenté la méthode GAP qui propose une approche totalement différente en comparant les divers clusterings à leurs résultats dans le cas d'une hypothèse nulle.

Considérons tout d'abord un ensemble de points X de coordonnées (x_1, x_2… x_p) dans un espace à P dimensions. Soient k clusters nommés C_1… C_k d'effectifs N_1… N_k et de centroïdes G_1… G_k.

4.3.2.1 La méthode Elbow

La première approche quand on veut fixer k le nombre optimal de clusters est une approche graphique représentant en abscisse la valeur k et en ordonnée la variance expliquée par les clusters qui est le rapport entre la variance inter-clusters à la variance totale. On peut aussi faire figurer en ordonnée la variance intra-classes auquel cas on dispose de la valeur pour 1 cluster. On estime qu'il faut arrêter l'incrémentation de k quand on ne gagne plus rien en variance expliquée ou encore quand la variance intra-classes commence à stagner. Ceci se manifeste par un coude (*elbow*) dans le tracé de la courbe. Malheureusement, un tel coude n'apparaît pas toujours et le jugement devient alors très difficile, d'autant plus qu'il n'y a pas de règle d'arrêt et que l'augmentation du nombre de clusters induit continuellement une diminution de la variance intra-classes, jusqu'à la nullité quand k égale le nombre de points.

4.3.2.2 La méthode de Davies-Bouldin[55]

Cette méthode utilise un score basé sur le ratio entre la variance intra-clusters et la distance inter-centroïdes. Elle cherche à mettre en perspective la « compacité » des clusters et leur éloignement. Voyons cela :

Soient $S_i = \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (X_j - G_i)^2}$ la fonction de compacité du cluster C_i, qui est la racine carrée de l'inertie du cluster sachant que tous les points ont une masse identique,

et $M_{i,j} = \sqrt{\sum_{m=1}^{p} (g_{m,i} - g_{m,j})^2}$ la distance euclidienne entre les centroïdes des clusters C_i et C_j.

Considérons maintenant pour i ≠ j le rapport $R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$. Plus il va être petit et plus les clusters C_i et C_j seront éloignés et/ou compacts. On calcule tous les $R_{i,j}$ possibles entre le cluster i et les autres. Nommons D_i sa valeur maximale : elle se produit avec les deux clusters les plus similaires et donc dans les plus mauvais cas de partitionnement. On peut alors moyenner les D_i des k clusters et l'on obtient alors l'index de Davies-Bouldin égal à $\frac{1}{k} \sum_{i=1}^{k} D_i$

Quand on cherche le nombre optimal de clusters dans un nuage de points, cette valeur DB sera la meilleure quand elle sera la plus basse. Cette méthode est donc une méthode locale en ce sens qu'elle

procède par « pénalisation » des similarités entre clusters. En aparté, elle possède des propriétés assez proches de celles du score de Dunn[56] d'application plus générale. Ces approches trouvent néanmoins leur limite en cas de chevauchement important des clusters, ce qui est hélas fréquemment le cas avec nos données cliniques (ex. âge des cancers, données de fécondité).

4.3.2.3 La méthode Maulik-Bandyopadhyay[57] (MB)

A la différence des méthodes précédentes, la méthode MB ajoute un terme de pénalisation à mesure que le nombre k de clusters augmente. Elle repose sur le calcul de la somme des distances intra-clusters de chaque point à son centroïde :

$$D_k = \sum_{i=1}^{k} \sum_{j=1}^{Ni} d(X_j, G_i)$$

Et la distance maximale inter-clusters, c'est à dire entre centroïdes :

$$E_k = max_{i,j=1}^{k} d(G_i, G_j)$$

Au final, le score pour k clusters vaut : MB(k) = $\left[\frac{1}{k} \frac{D_1}{D_k} E_k\right]^q$ (nous avons utilisé à l'instar de beaucoup d'autres la puissance q = 2)

Le premier terme de la formule sert de pénalisation quand le nombre k de clusters augmente. Le deuxième terme ressemble à la variance expliquée et le troisième terme est en partie représentatif de l'écartement des clusters. Plus il est élevé et meilleur est ce score.

L'inconvénient de cette méthode est qu'avec le terme de pénalité $\frac{1}{k}$, le score décroît rapidement au départ favorisant ainsi les petits nombres de clusters. Une version corrigée nommée AF_k a été proposée par Bayati et al que nous avons aussi implémentée[58]. Etonnamment, les auteurs éliminent le 3ème terme associé à la distance inter-clusters dans leur formule :

$$AF_k = \frac{1}{\frac{D_k}{D_1} + \propto k}$$

Après une évaluation empirique, la valeur de 0.08 a été considérée comme optimale pour α par les auteurs.

4.3.2.4 La méthode silhouette[59]

Le calcul du score Silhouette est basé sur une approche totalement différente. Il œuvre à l'échelle microscopique en s'intéressant aux items eux-mêmes et non aux clusters auxquels ils appartiennent. Son but est de vérifier si chaque item a été bien classé, c'est à dire s'il ne serait pas plus proche des items d'un autre cluster que de ceux de son propre cluster. Pour cela et pour chaque point X_i de la partition C_i, on calcule deux valeurs :

A(i) = $\frac{1}{N_i - 1} \sum_{j=1}^{N_i} d(X_i, X_j)$ qui est la moyenne des distances de X_i aux membres de son cluster

Et B(i) = $min_{k \neq k'} \frac{1}{N_{k'}} \sum_{j=1}^{N_{k'}} d(X_i, X_j)$ la distance moyenne entre le même point X_i et les points X_j du cluster le plus proche.

Le score "silouhette" du point X_i s'écrit alors $S_i = \frac{B(i)-A(i)}{\max(A(i),B(i))}$

Ce score est donc négatif si le point X_i est plus proche des points du cluster voisin que du sien propre. Dans tous les autres cas, il est positif et sa valeur quantifie sa bonne appartenance. On peut alors tracer un graphe caractéristique de cette méthode, affichant par cluster le score ce chacun de ses points, triés par ordre décroissant.

Un score moyen S peut être calculé sur l'ensemble des points : S = $\frac{1}{k} \sum_{i=1}^{k} \frac{1}{N_i} \sum_{j=1}^{N_i} S_j$

Ce score peut être utilisé pour comparer la performance de plusieurs clusterings : plus il est élevé et meilleur est la « caractérisation » des clusters. Nous en donnons un exemple ultérieurement.

4.3.2.5 La méthode GAP

Cette méthode élaborée par Tibshirani et al. en 2001[53] se base elle aussi sur l'inertie intra-clusters (si la distance euclidienne est utilisée) :

$$W_k = \sum_{i=1}^{k} \sum_{j=1}^{N_i} \|X_j - G_i\|$$

Elle propose de comparer cette valeur à celle qui serait obtenu en cas d'hypothèse nulle. Plusieurs travaux démontrent que cette dernière est obtenue en utilisant des populations ayant une répartition uniforme continue. Plus l'écart (d'où le nom "*gap*") entre l'espérance $E^*[Log(W_k^*)]$ dans ces populations et le Log(W_k) observé est important, meilleure apparaît la qualité du clustering.

Le score GAP pour k clusters et m itérations s'écrit ainsi : $GAP_k = \frac{1}{m}[Log(W_k^*)] - Log(W_k)$

Une dizaine d'itérations (m = 10) suffisent à obtenir une estimation assez stable de W_k ainsi qu'un écart-type SD_k calculé ainsi :

$$SD_k = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left[Log(W_{k,i}^* - E^*[Log(W_k^*)]\right]^2}$$

Soit $s_k = \sqrt{1 + \frac{1}{m}}\, SD_k$

Les auteurs proposent de choisir le nombre de cluster k le plus petit pour lequel :

$$GAP_k \geq GAP_{k+1} - s_{k+1}$$

Comme le signalent les auteurs, l'intérêt supplémentaire de cette approche est qu'elle fournit aussi un jugement sur l'utilité ou non du clustering puisque l'on peut s'arrêter directement à k = 1. Pour améliorer la performance du score GAP, nous avons suivi la suggestion des auteurs proposant de simuler les populations de référence à partir des axes principaux obtenus dans l'ACP (avec pour limites, les extrêmes observés par dimension). Dans de telles conditions, les auteurs signalent que score GAP s'est avéré supérieur à tous les autres scores testés. Malgré tout, en cas de clusters peu différenciés (avec de nombreuses superpositions), la méthode GAP peine à calculer le nombre optimal de cluster : la probabilité de considérer qu'un cluster est suffisant (donc pas d'avantage à partitionner) est directement proportionnelle à la proportion de points superposés.

4.3.3 Résultats obtenus avec l'ACP

Les variables disponibles dans notre base de données par famille sont fort heureusement beaucoup plus nombreuses que dans nos simulations. On y trouve le nombre de cancers du sein advenus dans la famille, ceux survenus chez les femmes et ceux chez les hommes, l'âge de déclaration par tranches de 10 ans, leurs types histologiques (carcinomes canalaires ou lobulaires, invasif, *in-situ* ou les deux), des marqueurs comme les récepteurs hormonaux, l'aspect triple-négatif. Sont aussi dénombrés les autres types de cancer (colon, poumon, prostate, pancréas, endomètre, ovaire, les autres localisations, les cancers multiples (chez un même individu). Nous n'avons cependant pas utilisées les données de natalité, mais elles étaient elles aussi disponibles.

4.3.3.1 ACP incluant les trois catégories des familles, sans mutation connue ou avec mutation BRCA1 ou BRCA2

4.3.3.1.1 Le diagramme des corrélations entre variables

L'analyse en composantes principales permet d'obtenir sur notre population de familles à risque sein/ovaire des informations intéressantes sur les proximités entre variables. Les deux principaux axes du diagramme suivant ne représentent respectivement que 12.3% et 9.3% de toute l'information :

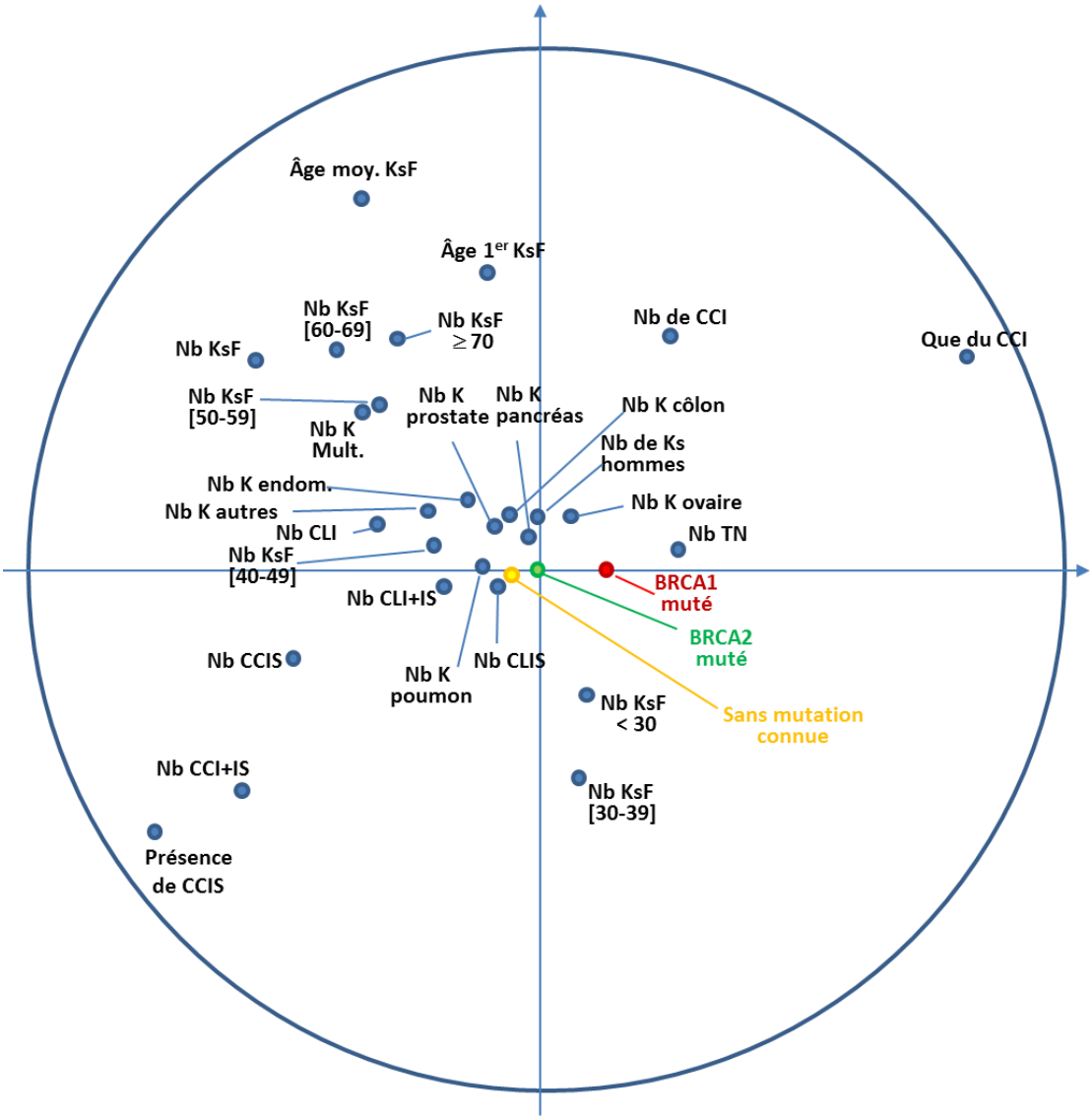
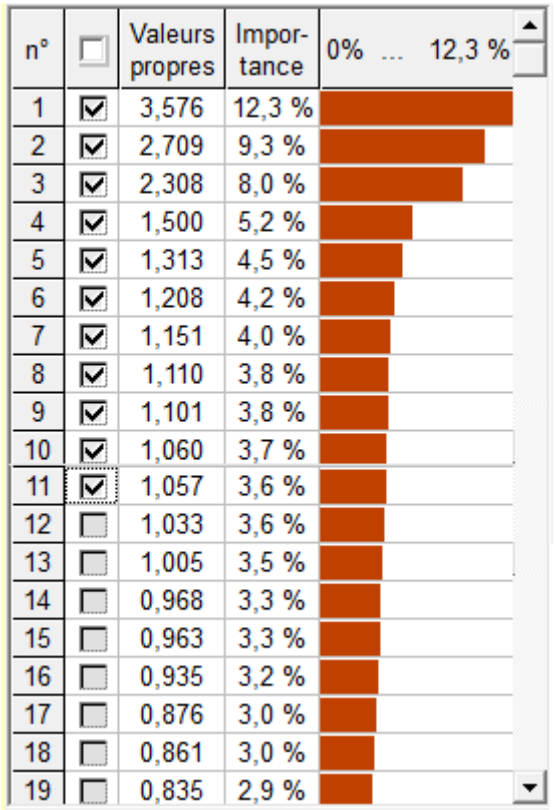


Figure 32 : Diagramme des corrélations fournis par l'ACP (K = cancer, Ks = cancer du sein, histologies des cancers du sein : TN = tumeur triple négative, CCI = carcinome canalaire invasif, CLI = carcinome lobulaire invasif, IS = in-situ, CCIS = carcinome canalaire in-situ, CLIS = cancer lobulaire in-situ)

Ce type de diagramme intéressera davantage le clinicien que le méthodologiste. On peut y remarquer que les données anatomo-pathologiques relatives aux cancers du sein sont au final peu discriminantes en regard de nos trois classes mutationnelles de familles. Néanmoins les mutations BRCA1 sont bien du côté des tumeurs du sein triples négatives (c'est à dire sans récepteurs hormonaux et ne sur-exprimant pas HER2), tumeurs de mauvais pronostic avec un taux de rechute de 25% à 3 ans. En outre, les mutations BRCA1 sont nettement du côté des familles n'ayant que des tumeurs du sein uniquement CCI, c'est à dire sans contingent *in-situ*. C'était une des conclusions de notre article, à savoir que la présence de CCIS caractérisait très probablement des processus carcinogènes différents de ceux de la voie BRCA1.

4.3.3.1.2 Clustering sur les familles à partir des dimensions principales de l'ACP

La première question qui se pose à cette étape concerne le nombre de dimensions à prendre en considération pour la réalisation du clustering. L'éboulis des 19 premières valeurs propres ici est le suivant :



n°		Valeurs propres	Impor-tance	0% ... 12,3 %
1	☑	3,576	12,3 %	
2	☑	2,709	9,3 %	
3	☑	2,308	8,0 %	
4	☑	1,500	5,2 %	
5	☑	1,313	4,5 %	
6	☑	1,208	4,2 %	
7	☑	1,151	4,0 %	
8	☑	1,110	3,8 %	
9	☑	1,101	3,8 %	
10	☑	1,060	3,7 %	
11	☑	1,057	3,6 %	
12	☐	1,033	3,6 %	
13	☐	1,005	3,5 %	
14	☐	0,968	3,3 %	
15	☐	0,963	3,3 %	
16	☐	0,935	3,2 %	
17	☐	0,876	3,0 %	
18	☐	0,861	3,0 %	
19	☐	0,835	2,9 %	

Figure 33 : éboulis des premières valeurs propres (sur les 29 disponibles) montrant une chute après la 3ème valeur

On constate la prépondérance, toute relative cependant, des 3 premières dimensions qui regroupent 30% de l'information (au lieu d'environ 9% attendu). Malgré tout, les autres dimensions sont loin d'être négligeables et il faut aller jusqu'à 18ème dimension pour tomber en-dessous des 3%. Quand on utilise les 11 premières dimensions, on a en notre possession 60% de l'information. C'est ce que nous avons utilisé pour effectuer le clustering suivant.

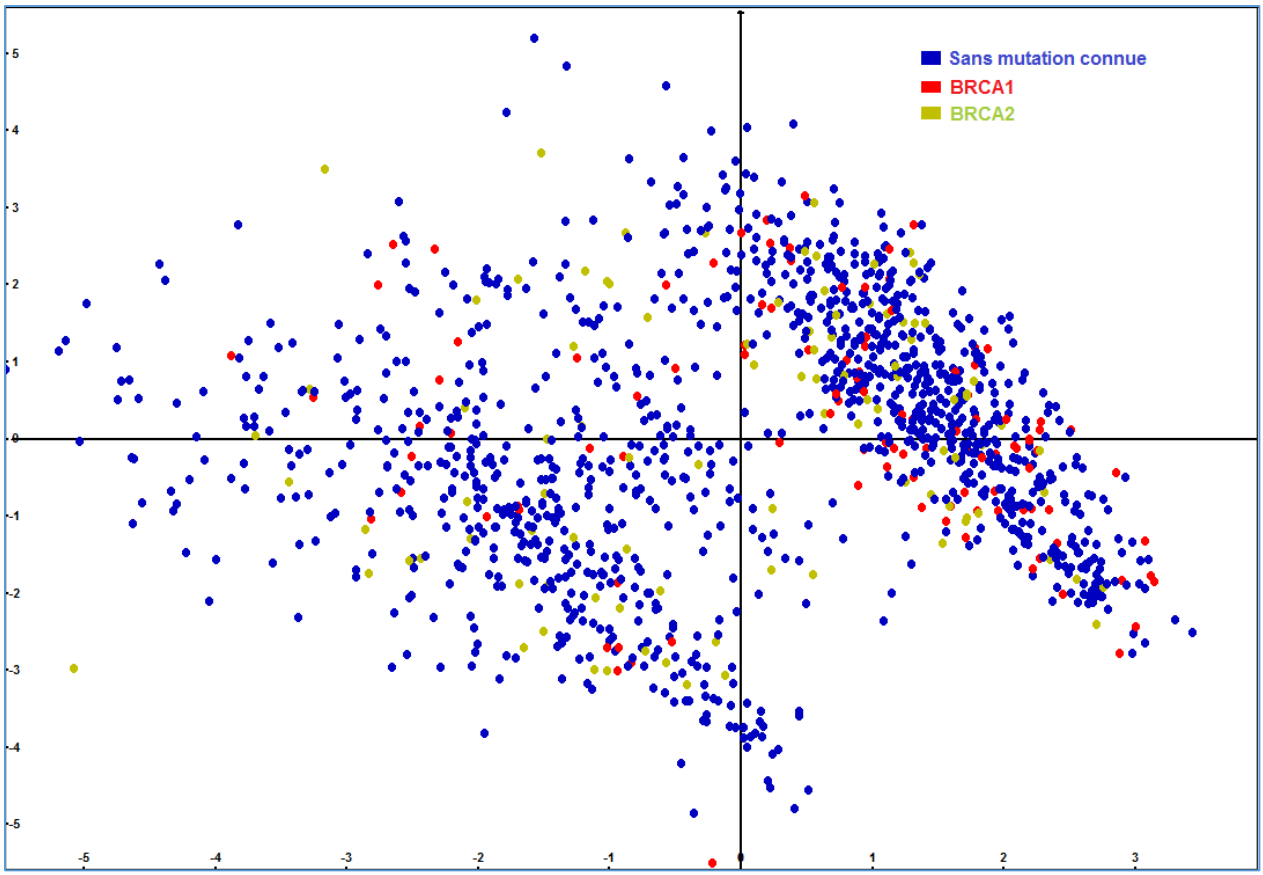


Figure 34 : répartition des 1 352 familles en prenant en compte les deux axes principaux de l'ACP

Quand on regarde les contributions des variables aux axes, l'axe n°1 (horizontal) est fonction principalement des histologies des cancers du sein : vers la droite, on trouve donc les familles où sont plus nombreuses les tumeurs triple-négatives, avec uniquement des CCI (carcinomes canalaires invasifs sans *in-situ*). Quant à l'axe n°2 (vertical), ce sont principalement les âges des cancers du sein qui le composent, avec en haut les âges de déclaration plus élevés. Deux clusters semblent se distinguer avec chacun des familles sans mutation connue, mais c'est un peu oublier que cette représentation n'utilise que 2 dimensions. Quel nombre de clusters distinguer alors ?

Nous avons calculé nos divers scores évaluant la performance des clusterings selon le nombre de clusters envisagé :

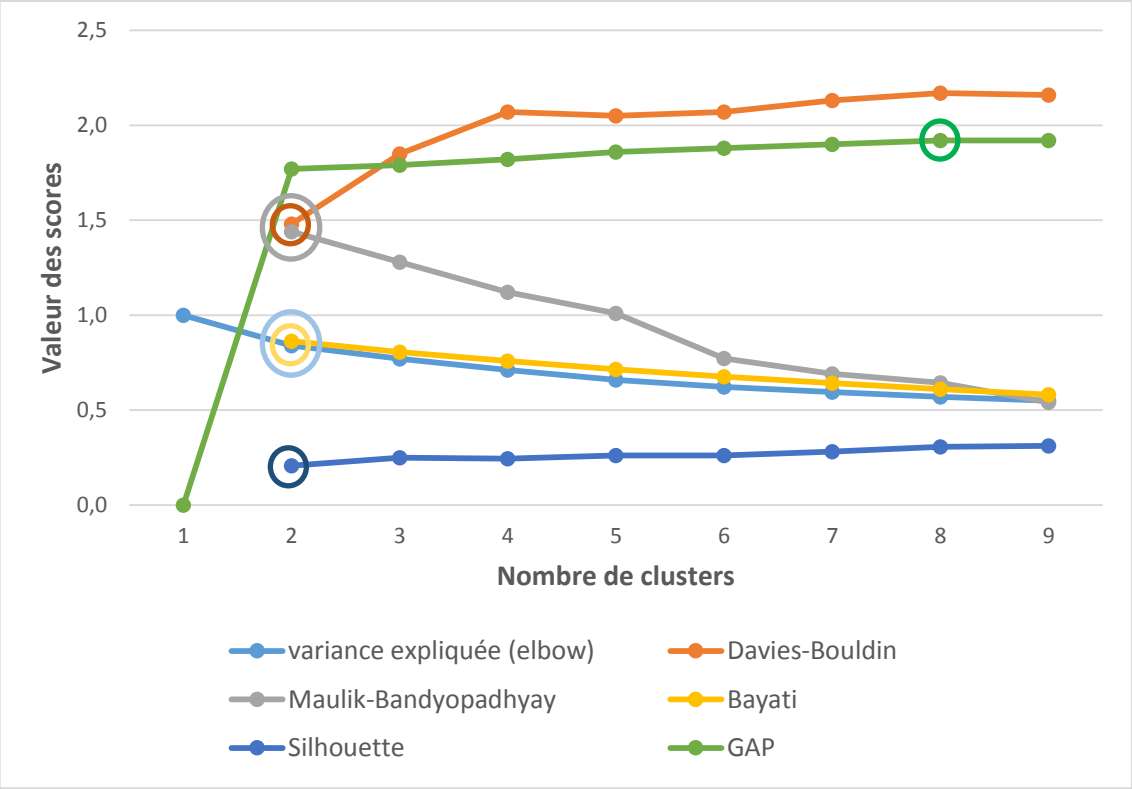


Figure 35 : comparaison des différents scores de performance des clusterings (les cercles colorés indiquent l'optimum de la méthode)

La première remarque concerne la grande concordance de ces scores, ce qui n'est pas surprenant compte tenu de leur mode de calcul, si l'on excepte le score GAP pour lequel le nombre optimal de clusters est de huit. La seconde remarque pointe la très grande ressemblance du pourcentage de variance expliquée et du score Maulik-Bandyopadhaya (MB) corrigé par Bayati et al. En effet, comme nous le faisions remarquer précédemment, ces derniers auteurs ne conservent du score MB que le ratio des variances auquel ils ajoutent une pénalisation proportionnelle au nombre de clusters.

Pour le score Elbow relatif à l'évolution de la variance expliquée, on constate sur ce graphique que le coude n'est pas franchement marqué et que ce score fait parfois la part belle à l'imagination quand il s'agit de l'interpréter.

Les points d'inflexion des courbes, quand ils existent ne sont pas inintéressants. Le score Davies-Bouldin plafonne à partir de quatre clusters tandis que celui de Maulik-Bandyopadhaya chute brusquement à partir de cinq. Le score silhouette étant presque constant, il indique que pour chacun de ces niveaux de clustering, on ne trouve guère plus de mal classés qu'avec deux clusters.

Au résultat, on doit se poser la question de scinder nos familles prédisposées aux cancers sein/ovaire en 2, 4, 5 ou 8 clusters.

4.3.3.1.3 Résultat du partitionnement en deux et quatre clusters

Le choix de deux clusters n'est pas très discriminant quant à nos classes de risque héréditaire : les deux clusters se distinguent essentiellement par leur composition de familles BRCA1 mutées (10.6% versus

4.2%, p = 0.0004) tandis que les mutations BRCA2 se répartissent en proportion égale dans les deux clusters (≈ 7% du total de chaque).

Tableau 7 : répartition des risques familiaux dans les deux clusters suite à l'ACP

Mutation	Cluster 1	Cluster 2	total
Aucune	**397**	**739**	1136
BRCA1	**19**	**96**	115
BRCA2	**32**	**69**	101
total	448	904	1352

Le partitionnement en quatre clusters sépare de meilleure manière nos risques familiaux ($p < 10^{-6}$) :

Tableau 8 : répartition des risques familiaux dans les quatre clusters

Mutation	Cluster 1	Cluster 2	Cluster 3	Cluster 4	total
Aucune	**307**	**161**	**358**	**310**	1136
BRCA1	**63**	**10**	**18**	**24**	115
BRCA2	**32**	**13**	**30**	**26**	101
total	402	184	406	360	1352

Ce clustering permet de mieux séparer les mutations BRCA1 puisque maintenant 55% d'entre elles constituent désormais 16% du cluster n°1. Ce cluster, bien que constitué majoritairement de familles sans mutation connue (à 76%) l'est beaucoup moins que les autres (à ≈87%).

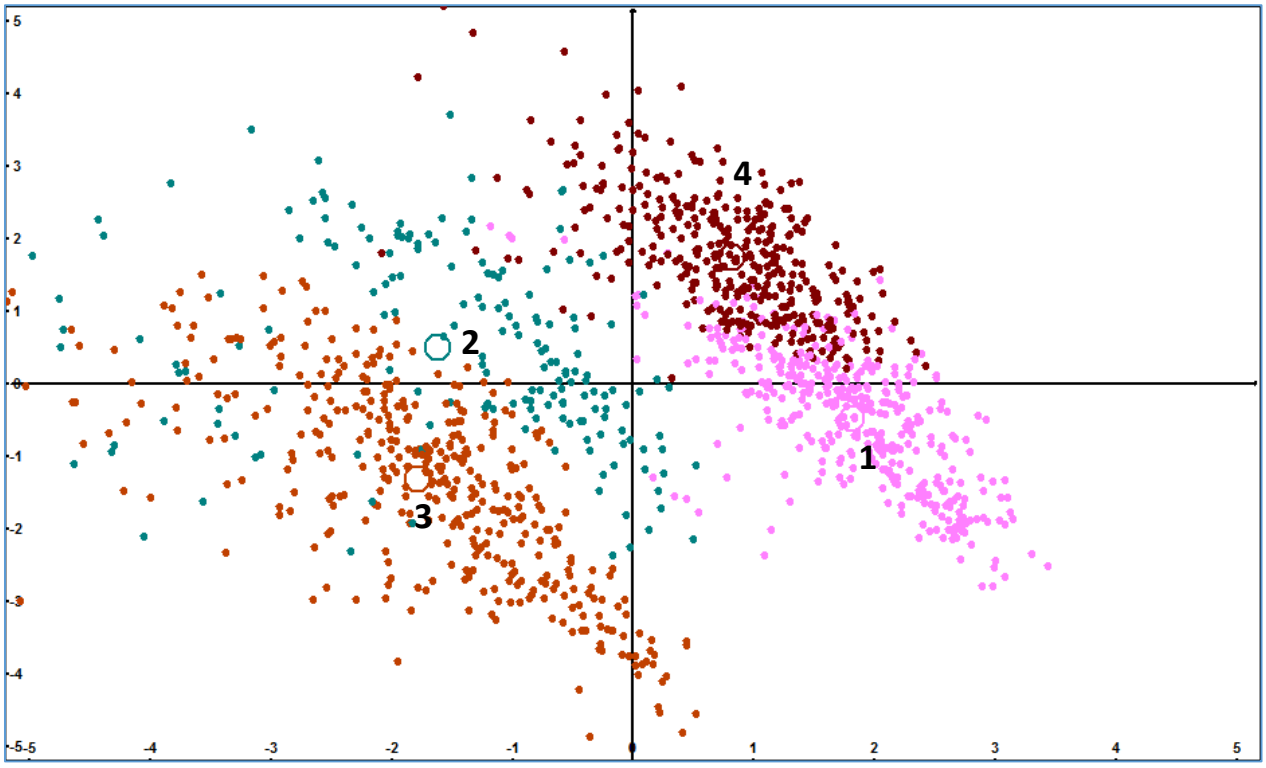


Figure 36 : partitionnement en 4 clusters des 1 352 familles de l'ACP précédente.

Ce partitionnement fait apparaître entre les deux principaux nuages un cluster médian (n°2 en cyan) tandis qu'il scinde le nuage de droite en deux en isolant le cluster n°1 en magenta en bas à droite et n°4 en rouge foncé. Le cluster n°3 en bas à gauche (orange-brun) paraît un peu mélangé avec le cluster n°2, ce qui est mis en évidence avec le graphe Silhouette suivant :

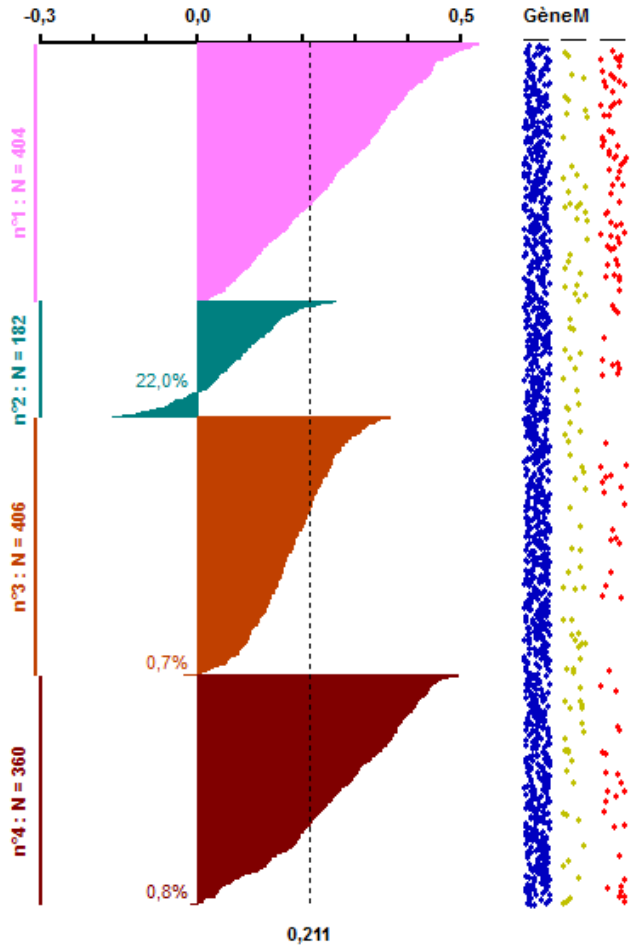


Figure 37 : graphe Silhouette : répartition des points dans chaque cluster en liaison avec le ratio de leurs distances avec les points de leur propre cluster et de ceux du cluster le plus proche (les points à droite correspondent aux familles sans mutation connue (en bleu), avec en rouge une mutation BRCA1 et en vert clair une mutation BRCA2)

Une répartition en 5 clusters, avec une probabilité associée au $Chi^2 < 10^{-7}$, ajoute dans le quart haut-gauche un cluster possédant 19% des mutations BRCA2 qu'il récupère sur le cluster n°3 : accessoirement ce cluster manque de consistance avec 32% de points qui s'éloignent trop selon le score Silhouette.

4.4 Analyse des familles à risque familial sein/ovaire sans mutation connue

L'objectif principal de notre étude est la caractérisation de familles à risque héréditaire spécifique. Nous avons vu d'une part que les données issues des sous-arbres ne procuraient pas d'avantage en

termes de prédiction et d'autre part que les analyses en clusters suite à l'ACP pouvaient permettre de partitionner nos familles en groupes assez distincts (probabilité du Chi² élevée) malgré d'importantes superpositions. Nous analysons donc dans ce paragraphe les familles ayant comme syndrome familial principal une prédisposition sein/ovaire mais sans mutation connue. Ces familles ne sont plus que 1136.

4.4.1 Recherche du nombre optimal de clusters

Les scores permettant d'établir le nombre optimal de clusters ont été réalisés sur 70% de l'information (13 axes de l'ACP), à l'aide de 50 itérations à chaque fois et après avoir limité le nombre maximal de clusters à vingt. Voici leurs évolutions :

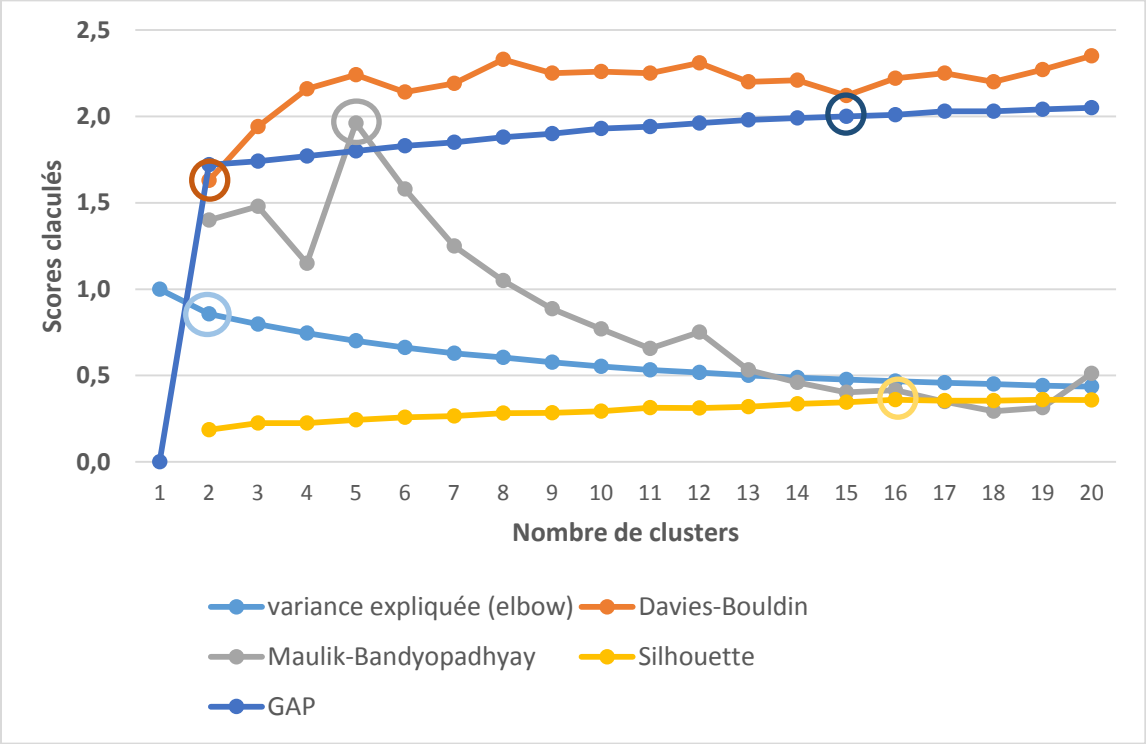


Figure 38 : évolution des différents scores évaluant le nombre optimal de clusters (les cercles colorés indiquent l'optimum par méthode)

Les méthodes Silhouette, Elbow et Davies-Bouldin donnent le même résultat à 2 clusters tandis que le score de Maulik-Bandyopadhyay est assez nettement optimal à 5 clusters et celui de GAP à 15. Si l'on regarde l'allure de la courbe Davies-Bouldin, on constate un minimum local aux nombres 6 et 15, ce qui donnerait un peu de force à aux nombres 5 ou 6 et 15. Ce diagramme confirme, s'il le fallait, qu'il n'existe pas de méthode "absolue" en particulier quand les données sont très enchevêtrées et que la responsabilité finale du choix du nombre de clusters revient à l'utilisateur. Nous avons donc opté pour une configuration en 5 clusters qui nous donne le diagramme ci-après.

4.4.2 Partitionnement en 5 clusters des 1136 familles

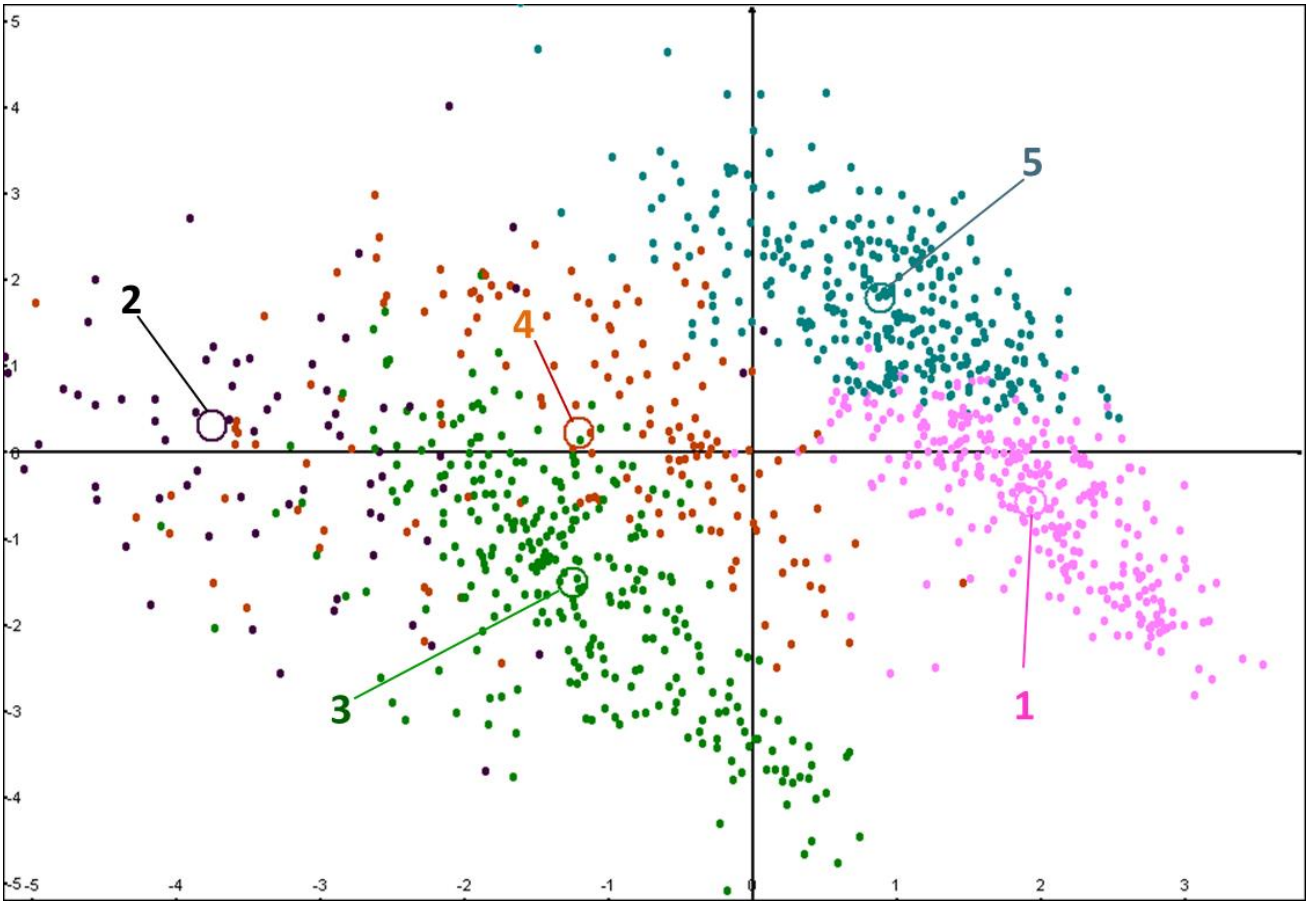


Figure 39 : analyse en 5 clusters des familles prédisposées aux cancers sein/ovaire mais sans mutation connue

Les axes 1 et 2 représentent 12.5% et 9.6% respectivement de l'information. L'axe horizontal est très corrélé la présence unique ou non de CCI, et donc l'absence de CCIS dans la partie droite du diagramme, les cancers triples-négatifs étant majoritairement au centre. A gauche par contre on trouve les histologies multiples des cancers du sein avec notamment les contingents canalaires *in-situ* et les cancers lobulaires. Les autres localisations de cancer si l'on exclut les cancers de l'ovaire viennent se situer à gauche de l'axe vertical, en particulier les cancers multiples, les cancers du poumon, de l'endomètre et du pancréas.

L'axe n° 2 est principalement associé à des âges de déclaration plus élevés (plus de 40 ans situés en bas), davantage de cancers de l'ovaire, du sein chez l'homme. Il partage avec l'axe n°1 le plus grand nombre de CCI "pur" et par conséquent en diagonale opposée sur le diagramme dessus, les CCIS dans le quart bas-gauche.

L'axe n°3 représente 7.7% de la variance totale. Les variables qui contribuent à son importance sont les cancers de l'endomètre, de la prostate, du côlon, du poumon et les cancers multiples. Cet axe va donner une indication sur les susceptibilités autres que sein/ovaire.

Leur analyse à l'aide d'un graphe "Silhouette" (ci-contre) permet d'affiner le jugement que l'on peut donner aux différents nuages.

Le cluster n°2 est le moins "pur" des clusters, avec 32% de familles sur 78 qui pourraient aussi bien se retrouver ailleurs d'après les calculs de proximité. Il se situe dans la partie des risques de cancers hétérogènes et survenant à des âges médians (centré à gauche sur la Figure 39).

Les clusters n°1, 3 et 5 avec respectivement 301, 296 et 307 familles sont assez bien discriminés et constituent vraisemblablement des groupes aux caractéristiques différentes. On sait déjà que les clusters 1 et 5 contiennent plutôt des CCI sans *in-situ* et que ce qui distingue ces deux groupes sont principalement les âges de déclaration (< 50 pour le 1er cluster et ≥ 50 pour le 5ème). Le cluster 5 a des caractéristiques BRCA1-like.

Quant au cluster 3, il rassemble les histologies mixtes des cancers du sein, mais pouvant survenir très jeune. Il apparaît assez consistant. Le cluster n°4 contient 12% de familles mal classées et son score moyen est assez faible. Situé assez proche du centre dans la Figure 39, il est dans le « ventre mou » de l'ACP.

Ce graphe nous permet en outre de connaître les familles qui sont les plus représentatives de leur cluster. Nous les avons listées par ordre décroissant de leur score (comme dans le graphe), considérant que plus leur score est élevé et plus elles sont représentatives de leur groupe. Cette liste de familles a donc été transmise à mes collègues biologistes du LOM afin de valider (ou non) la pertinence de ce partitionnement des familles grâce à l'analyse d'un panel de gènes susceptibles d'être impliqués dans la carcinogenèse en cas de mutation.

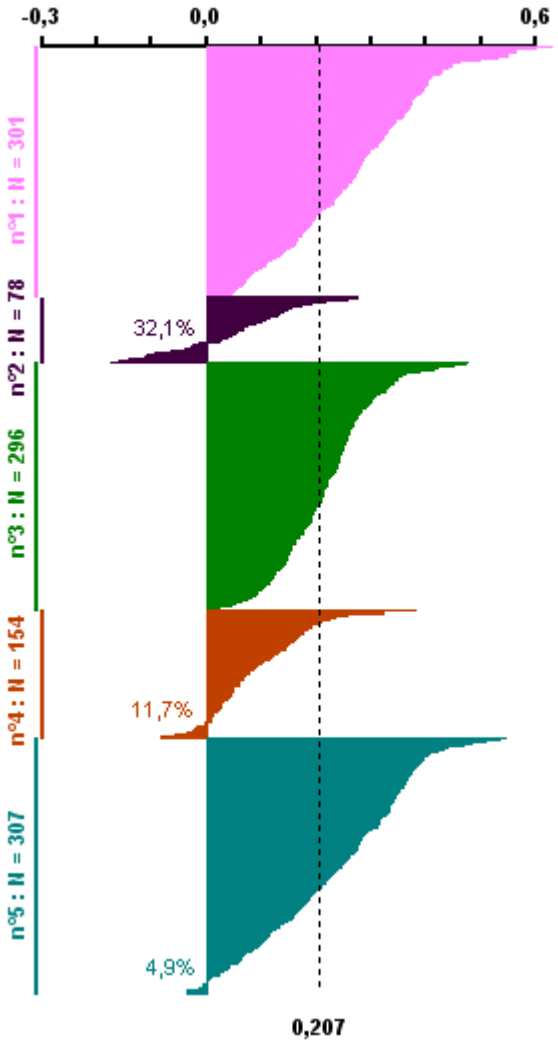
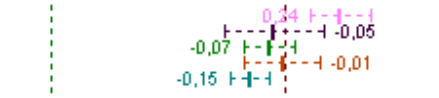
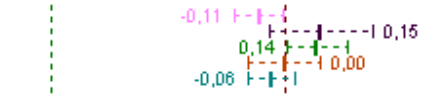
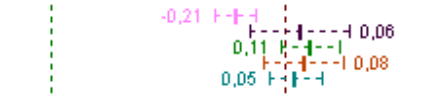
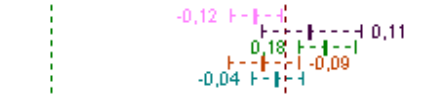
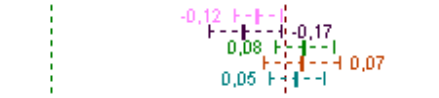
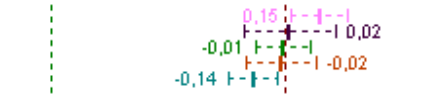
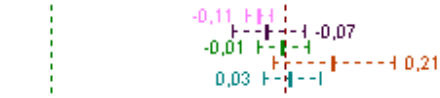
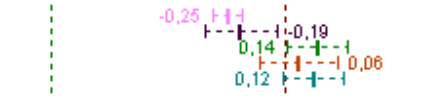
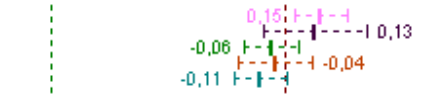
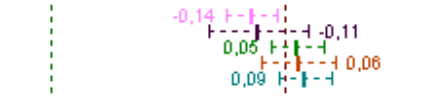


Figure 40 : graphe "Silhouette" associé au partitionnement en 5 clusters

4.4.3 Association des clusters aux variables incluses

Pour mieux caractériser les clusters, il est possible d'étudier le lien de chacun aux variables incluses initialement dans l'ACP. C'est ce que nous présentons dans le tableau suivant en ne conservant que les associations significatives :

Tableau 9 : relations entre chaque cluster et les principales variables centrées et réduites de l'ACP

Libellé des variables	Intervalles de confiance à 95% par cluster (-1 / 0 / 1)	Proba.
Nombre de cancers femmes [40-49[ans		0,0005
Nombre de cancers femmes [50-59[ans		0,000057
Nombre de cancers de la prostate		0,015
Nombre de cancers du poumon		0,0009
Nombre de cancers autres (dont non documentés)		0,0023
Nombre de types histologiques de cancer du sein différents rencontrés		0,038
Nombre de cancers du sein CCI		0,014
Nombre de cancers du sein CLI+IS		0,023
Nombre de cancers du sein invasifs triple-négatifs (CCI ou CLI) dans la famille		0,0000034
Que des CCI dans la famille ?		0,011
Existence de tumeurs dans la famille avec un contingent CCIS ?		0,03

Les couleurs des intervalles de confiance sont les mêmes que celles de la Figure 39, le premier cluster figuré en rose en début de graphe, le second en violet sombre, le troisième en vert, le quatrième en brun et le cinquième en bleu clair. Les variables différenciant peu ou pas les clusters ont été supprimées : cancers du sein avant 30 ans (p = 0.79), de 30 à 39 ans (p = 0.066), de 60 à 69 ans (p =

0.58), à 70 ans et plus (p = 0.22), âge du 1er cancer du sein dans la famille (p = 0.34), histologie du sein CCIS (p = 0.21), CCI + CCIS (p = 0.065), CLI (p = 0.46), CLIS (p = 0.32), cancers du sein chez l'homme (p = 0.61), cancers de l'ovaire (p = 0.30), de l'endomètre (p = 0.078), du pancréas (p = 0.50), du côlon (p = 0.76), localisations multiples (p = 0.81).

Le cluster n°1 est caractérisé par des cancers du sein arrivant entre 40 et 60 ans, de type CCI non triple négatifs et sans contingent *in-situ*, ceci étant valable chez la plupart des membres de la famille. De même on ne trouve guère de cancers lobulaires. Les cancers du sein dans ces familles sont assez exclusivement des CCI avec des récepteurs hormonaux positifs ou une surexpression de HER2. En outre, hormis les cancers de l'endomètre, on trouve peu d'autres cancers, et en particulier très peu de cancers du poumon ou de prostate.

Le cluster n°2 présente des caractéristiques histologiques des cancers du sein assez semblable mais arrivant plus jeune (entre 30 et 40 ans). On trouve une plus grande variété de cancers dans ces familles dont des cancers de la prostate mais peu de l'endomètre et encore moins de cancers de l'ovaire.

Le cluster n°3 rassemble les familles avec des cancers du sein survenant plutôt après 40 ans, de CCI type triple négatif mais aussi des histologies lobulaires ou *in-situ*. Les cancers de la prostate et du poumon sont plus fréquents dans ce cluster mais les autres (côlon, pancréas, endomètre…) ne sont pas rares non plus. Donc des familles avec une large palette de localisation de cancer et la présence de triples négatifs dans le sein.

Le cluster n°4 ne se distingue que par la présence en excès de cancers du sein lobulaires éventuellement associés à de l'*in-situ*. Les autres variables sont dans la moyenne.

Enfin, le cluster n°5 regroupe des familles où les cancers du sein surviennent souvent avant 40 ans, avec des histologies variées malgré un excès de triples négatifs. Les lobulaires, les *in-situ* associés ou seuls ne sont pas rares. On trouve un léger excès de cancers du pancréas et du poumon.

4.4.4 Résultats des tests génétiques réalisés sur les 5 premières familles des clusters

En dépit du coût important des tests génétiques, même quand ils sont réalisés en séquençage haut débit sur un panel de gènes comme celui développé au Centre Jean Perrin, 5 familles sans mutation connue ont été testées dans 4 des 5 clusters afin de fournir une première évaluation de la pertinence biologique de notre clustering. Ces premières analyses donnent des résultats très intéressants :

Tableau 10 : résultat des tests génétiques sur 20 familles typiques de 4 des 5 clusters (VSI = variant de signification inconnue)

Cluster 1	Cluster 3	Cluster 4	Cluster 5
mutation (classe 5) BRCA2	Négatif	Négatif	1 VSI (classe 3) dans BRCA1
2 VSI (classe 3) dans BAP1 et CHEK2	1 VSI (classe 3) dans PALB2	1 VSI (classe 3) dans RAD51D	1 VSI (classe 3) dans ATM
Négatif	2 VSI (classe 3) dans BRCA2 et MSH6	1 VSI (classe 3) dans PMS2	1 VSI (classe 3) dans CHEK2
mutation (classe 5) BRCA2	Négatif	1 VSI (classe 3) dans PALB2	1 VSI (classe 3) dans PTEN
3 VSI (classe 3) dans ATM, FANCM et MLH1	2 VSI (classe 3) dans CDH1 et MSH6	Négatif	**mutation (classe 5) dans FANCM** + 1 VSI (classe 3) dans BRCA1

Comment se lisent ces résultats ? Les variants (ou polymorphismes) de classe 3, 4 et 5 correspondent respectivement à des variants de signification inconnue (3), probablement pathogènes (4) et pathogènes (5). Quant aux variants de classe 1 et 2 (bénins ou probablement bénins), ils ne sont pas présentés dans le tableau. Les variants de classe 5 correspondent donc à des mutations reconnues comme délétères. Comme on peut le constater dans le tableau, 3 mutations délétères ont donc été retrouvées, deux sur BRCA2 et une sur FANCM. Cette dernière mutation est associée à l'anémie de Fanconi mais aussi à un risque élevé de cancers, parmi lesquels des cancers du sein triple négatifs, de l'ovaire ainsi que d'autres localisations. Nous ne sommes pas surpris de trouver les deux mutations BRCA2 dans le cluster 1, qualifié ci-dessus de groupe de familles BRCA-like. La mutation de FANCM (associée à un VSI sur BRCA1) est elle aussi idéalement située (cluster 5) puisqu'il se trouve juste dessus et se différencie essentiellement par un âge de déclaration plus élevé. Ce gène est impliqué dans la réparation de l'ADN mais on lui a récemment découvert un rôle dans le maintien du télomère[60].

De manière très intéressante, on trouve dans le cluster n°1, celui de plus mauvais pronostic, une triple association de variants de classe 3 entre à nouveau FANCM et deux gènes bien connus MLH1 qui fait partie des gènes MMR (*mismatch repair*) favorisant les cancers colorectaux et ATM (pour son rôle initialement découvert dans l'ataxie télangiectasie, une maladie principalement neurologique de mauvais pronostic). Ce dernier gène a des fonctions ubiquitaires du fait de ses multiples interactions avec les autres gènes : il intervient en particulier dans le contrôle du cycle cellulaire. Enfin, toujours dans le cluster n°1, on trouve une association de deux variants de classe 3 entre BAP1 et CHEK2. BAP1 est considéré comme un gène suppresseur de tumeur tandis que CHEK2 joue un rôle dans le contrôle du cycle cellulaire mais aussi dans la réparation des lésions de l'ADN[61].

Dans le cluster n°3 qui correspond à un risque élevé et assez jeune de cancer (mais avec des localisations et/ou des histologies variées), on note deux associations de variants de classe 3, ce deux fois avec MSH6 (lui aussi faisant partie des gènes MMR : la première association est avec BRCA2 et la seconde avec CDH1. Les mutations de CDH1 induisent un risque élevé de cancer du sein et de l'estomac, mais aussi de malformation de type fente labio-palatine. Ce gène code pour une protéine, la E-cadhérine, que l'on retrouve principalement dans les membranes des cellules épithéliales. Elle a un rôle dans les échanges inter-membranaires, l'adhésion cellulaire et enfin elle agit elle aussi comme suppresseur de tumeur.

Ces résultats préliminaires nous paraissent en faveur d'une assez bonne qualité du clustering, mais ils devront être confirmés par des résultats sur davantage de familles par cluster. Le point important, nous semble-t-il, concerne les associations de variants de classe 3. On trouve une triple association dans le groupe le plus délétère et 3 doubles associations dans des groupes eux aussi à haut risque. Peut-être ces résultats indiquent-ils que les variants de signification inconnue, lorsqu'ils s'associent, induisent un risque de cancer presque aussi élevé que les mutations connues comme sur BRCA, bien que les localisations ou les histologies malignes associées soient plus diverses. *A contrario*, sur le plan clinique, quand on trouve dans des familles des histologies et/ou de localisations cancéreuses variées, peut-être devrait-on rechercher des associations de variants non délétères *a priori.*

4.5 Conclusion de la partie 4

L'analyse en composantes principales sur les 300 familles simulées suivi d'un *k-means* clustering n'a pas montré l'intérêt des sous-arbres familiaux, au contraire même puisque les données brutes moyennées par familles se sont révélées plus discriminantes de l'état mutationnel familial que nous avions paramétré que celles tirées des sous-arbres.

Quant aux scores permettant de trouver le « meilleur » nombre de clusters suite à l'ACP, ils se sont trouvés en difficulté face à des données très superposées, imbriquées, enchevêtrées, mais c'est une difficulté connue. Le score qui nous a semblé le plus « performant » dans un tel contexte a été celui de Maulik-Bandyopadhyay. On doit toutefois nuancer ce résultat car qu'un score montre des variations « utilisables » ne signifie pas pour autant qu'il corresponde à des partitionnements optimaux dans la réalité. Le jugé de l'utilisateur, biologiste ou statisticien, apparaît donc indispensable pour proposer, à partir des résultats des divers scores, le ou les partitionnements porteurs de sens.

L'analyse qui a été faite sur nos familles prédisposées au cancer sein/ovaire, sans mutation connue et pour lesquelles nous avions au moins une histologie, a fourni des résultats non dénués d'intérêt. L'analyse des relations statistiques entre les clusters et les variables incluses dans l'ACP permet une meilleure lecture du résultat du clustering. Le partitionnement semble distinguer des familles aux caractéristiques bien séparées et nous avons transmis au laboratoire d'oncologie moléculaire du Centre Jean Perrin une liste des familles les plus représentatives de chaque groupe à fin de les tester à l'aide du panel d'environ 70 gènes possiblement impliqués dans la carcinogenèse. Au vu des résultats préliminaires rapportés ci-dessus, on peut déjà émettre l'hypothèse que des variants de signification inconnue pourraient s'avérer délétères quand ils sont associés : cela va dans le sens de l'impact probable des interactions génétiques de variants *a priori* peu délétères, problème au centre de cette thèse. Nous aurons probablement davantage de résultats de ces tests génétiques lors de la soutenance.

Enfin, sur un plan biologique, la spécificité de certains clusters aux cancers du sein de type CCI non triple négatif alors que d'autres sont associés à une grande variété d'histologies et la présence d'autres localisation de cancer (poumon, prostate…) suggère des susceptibilités génétiques sous-jacentes différentes qui semblent déjà se confirmer par les tests génétiques préliminaires.

5 La classification ascendante hiérarchique hiérarchique

5.1 Généralités sur la classification ascendante hiérarchique (CAH)

La classification ascendante hiérarchique (*hierarchical clustering*) est une méthode bien connue d'agrégation de données et elle est très utilisée en génétique. L'algorithme sur lequel elle repose est simplissime : on calcule à partir de caractéristiques communes (la plupart du temps quantitatives), la « distance » entre tous les éléments que l'on cherche à classer (objets, individus, variables biologiques…). Ensuite, on commence par rassembler les deux éléments les plus proches. Ces deux éléments sont alors supprimés du calcul et remplacés par la paire qu'ils forment. On continue ensuite en recherchant les deux autres éléments, en incluant la nouvelle paire, les plus proches, etc. De fil en aiguille, des agrégats (des paires de paires de paires…) se forment jusqu'à épuisement des éléments. Plusieurs possibilités existent pour le re-calcul des distances quand une paire est constituée, que l'on appelle "saut" (*linkage*) :

- Le saut minimum (*single linkage*) : la distance entre la paire d'éléments formée et les éléments restants est égale à la distance de l'élément considéré avec le plus proche des deux éléments assemblés.
- Le saut moyen (*average linkage*) : la distance entre la paire et les autres éléments est égale à la moyenne des distances de l'élément considéré avec les deux éléments assemblés
- Le saut maximum (*complete linkage*) : la distance entre la paire d'éléments et les éléments restants est égale à la distance de l'élément considéré avec le plus lointain des deux éléments assemblés.

La méthode du saut minimum tend à agréger un nouvel élément à un groupe existant plutôt qu'à donner naissance à un nouveau groupe d'où des dendrogrammes en escaliers. Le chainage du saut maximum est très sensible aux points aberrants qui vont altérer la qualité du résultat d'où sa faible utilisation. Le saut moyen est généralement préféré, même s'il crée des partitions souvent très déséquilibrées (de gros clusters avoisinant de très petits clusters).

Une autre méthode existe, basée non plus sur les distances, mais sur les moindres variations d'inertie quand on regroupe des éléments : la méthode de Ward[62]. Dans cette méthode, on agrège deux éléments quand cette agrégation induit le minimum de perte d'inertie interclasse (ici interfamiliale), ou, ce qui est équivalent, le minimum de gain d'inertie intra-classe. Cette méthode a l'avantage de produire des partitionnements beaucoup plus équilibrés, c'est à dire avec des dendrogrammes moins en forme d'escaliers.

Voici un exemple de résultat produit par cette méthode d'analyse dans notre logiciel SEM :

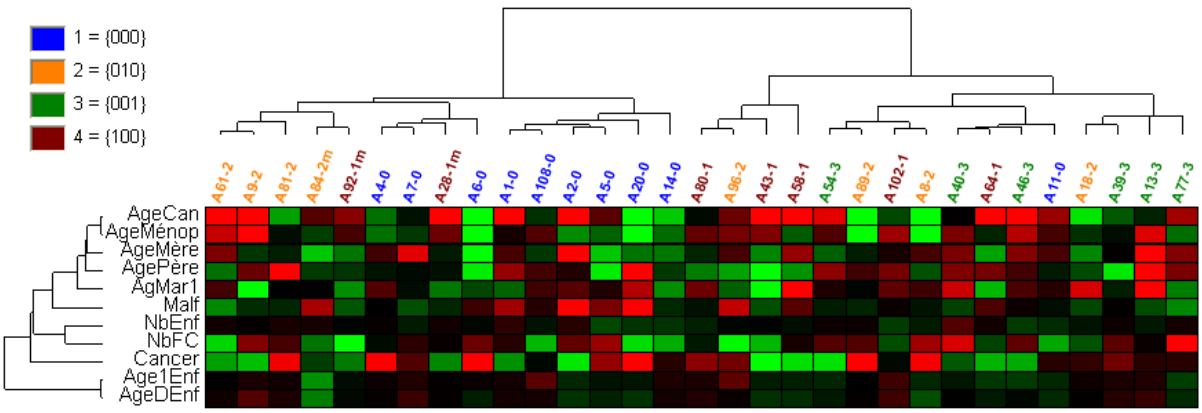


Figure 41 : exemple de CAH réalisée à partir de 31 familles générées par POLYGENE avec 3 types de mutation délétère et un groupe sans mutation (n° de famille en bleu). Les variables sont la survenue ou non d'un cancer, l'âge du cancer, un ATCD de malformation congénitale, les âges de la mère et du père lors de la naissance de l'individu considéré, de son mariage (le 1er si plusieurs), de son premier et dernier enfant, et pour une femme l'âge de la ménopause et le nombre de fausses couches.

Comment lire ce type de figure ? Trois types d'information la composent : en haut le dendrogramme des familles (arborescence), avec en bleu celles sans mutations, puis les autres colorées selon que le type de mutation favorise le cancer à 30, 40 ou 50 ans, respectivement en grenat, orange ou vert. Dans nos analyses à l'aide de familles simulées, cette connaissance du statut mutationnel facilite grandement l'interprétation : les familles apparaissent bien classées si les non mutées sont ensemble (paquet de familles en bleu) mais aussi quand les autres couleurs se trouvent dans d'autres groupes, séparément si possible. La hauteur des lignes verticales représente la distance trouvée entre les divers éléments, ici les familles ou les groupes de familles.

A gauche est représenté le dendrogramme des variables. Son calcul est indépendant de la hiérarchie familiale. De la même manière, les variables les plus « proches » sont regroupées ensemble. Enfin, au centre figure le damier, constitué d'une cellule par famille et par variable : plus elle est rouge vif et plus la variable a une valeur moyenne élevée pour la famille. Inversement une couleur verte indique une valeur au-dessous de 0 (puisque les valeurs sont centrées et réduites). Par exemple, la famille A81-2 (3ème en partant de la gauche) a de nombreux cancers : sa variable cancer (oui/non) est rouge clair (valeur élevée) tandis que l'âge du cancer est en vert clair car ses cancers arrivent jeune. Ce damier ne sert pas trop ici, même si l'on peut souvent juger de l'intérêt d'un clustering quand les couleurs du damier sont diagonalisées, les rouges dans un coin et les verts dans l'angle opposé en diagonale.

La notion de distance dans la CAH est importante et elle peut faire varier dans une assez large mesure la forme des dendrogrammes. Souvent on utilise la distance euclidienne qui est la plus facile à comprendre, mais d'autres « distances » sont possibles comme le coefficient de corrélation ou avec des variables qualitatives le calcul de Chi^2. Nous en testerons plusieurs autres pour agréger nos familles.

La CAH présente de nombreux intérêts pour le chercheur dont deux en particulier quant aux classements fournis par les dendrogrammes[63] : celui de ne pas demander d'*a priori* sur le niveau de partitionnement. On n'a pas à fixer préalablement le nombre de clusters (ce qui est requis dans le K-means clustering). Le second est d'offrir une représentation sous forme d'arborescence très propice à la réflexion quoique parfois un peu trompeuse.

5.2 Biais causés par le calcul standard de distance entre deux familles

Si la CAH constitue une approche pratique pour réaliser le regroupement d'individus à partir de leurs différentes caractéristiques, dans notre contexte, qui est de classer des familles et non des individus, il est nécessaire d'ajouter une hiérarchie supplémentaire avant tout regroupement : la hiérarchie familiale, d'où l'appellation de « clustering hiérarchique hiérarchique » ! Malheureusement, les logiciels effectuant ce type d'analyse ne sont pas capables de prendre en compte cette double hiérarchie. Nous avons donc développé des programmes particuliers permettant de travailler à partie des données individuelles sans pour autant perdre les appartenances familiales. Au final, les regroupements représentés par les dendrogrammes contiennent bien les familles (Figure 41), mais tous les calculs (distances, inertie, variance, etc.) sont effectués en revenant aux données individuelles.

Un deuxième problème se pose quand on utilise les informations oncogénétiques : celui des données manquantes. D'ordinaire, comme par exemple dans les analyses de la variance, la simulation des données manquantes permet d'obtenir des résultats optimaux tout en respectant les conditions d'utilisation des modèles. Malheureusement, avec les arbres généalogiques, la prise en compte d'individus jeunes (voire décédés jeune) interdit de simuler l'occurrence ou non de cancers et encore moins l'âge de diagnostic. Dans les analyses de survie (méthode Kaplan-Meier, modèle de Cox…), on parlerait de données censurées. Les données oncogénétiques sont donc par essence parcellaires et ce ne sont pas les données qu'il s'agit d'adapter ici mais les modèles qui les analysent.

Ce chapitre sera donc principalement consacré à la description des solutions répondant à cette double problématique : d'une part conserver dans les analyses les appartenances familiales et de l'autre garantir l'absence de biais induits par la connaissance parcellaire des données individuelles.

Prenons l'exemple de 3 familles réduites possédant respectivement 3, 4 et 5 membres. Dans le tableau suivant, se trouvent leur genre, la présence ou non d'un cancer et l'âge de son occurrence.

Tableau 11 : exemple de données familiales avec de nombreuses données manquantes (les colonnes Rcr1, Rcr2 et Rcr3 sont les résultats centrés et réduits des 3 colonnes précédentes)

Famille	Individu	Sexe	Cancer	âge cancer	Rcr1	Rcr2	Rcr3
1	***1***	1	0		1	-1	
1	***2***	0	1	50	-1	1	-0,248
1	***3***	1	0		1	-1	
2	***4***	1	1	45	1	1	-0,620
2	***5***	1	0		1	-1	
2	***6***	0	1	30	-1	1	-1,736
2	***7***	0	0		-1	-1	
3	***8***	1	0		1	-1	
3	***9***	0	1	65	-1	1	0,868
3	***10***	1	0		1	-1	
3	***11***	0	1	70	-1	1	1,240
3	***12***	0	1	60	-1	1	0,496
Moyenne		0,5	0,5	53,33	0,00	0,00	0,00
Ecart-type		0,5	0,5	13,44	1,00	1,00	1,00

Les distances euclidiennes entre familles peuvent être calculées soit en utilisant les barycentres (ou centroïdes) familiaux (Figure 42, flèches en pointillés), soit en moyennant l'ensemble des distances interindividuelles (Figure 42, flèches pleines), c'est à dire entre tous les individus d'une famille avec ceux de chaque autre.

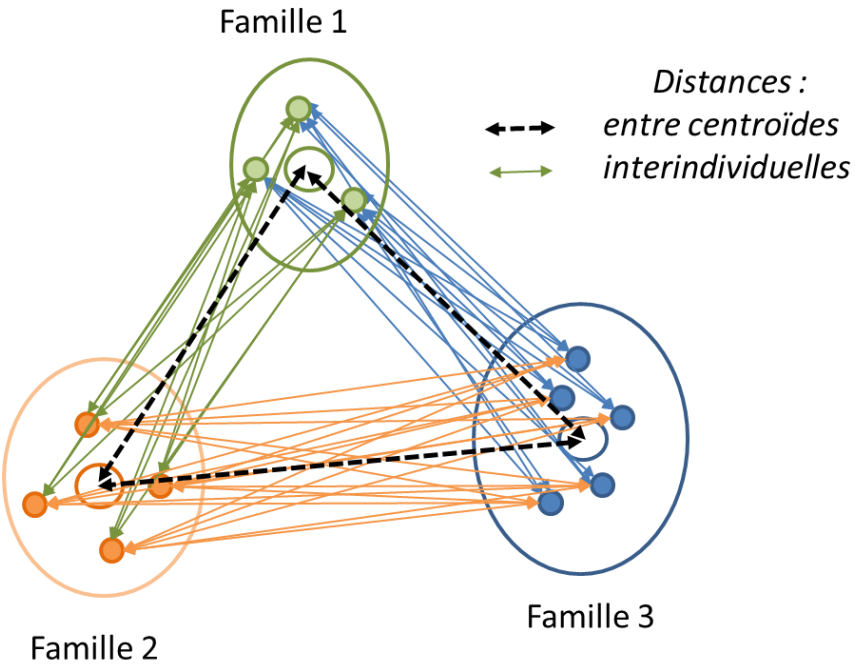


Figure 42 : deux manières de calculer les distances interfamiliales à partir de familles de même effectif que l'exemple du Tableau 11

La Figure 42 est néanmoins peu conforme à la réalité des familles en oncogénétique : celles-ci sont ici très nettement séparées. En fait, on se trouve dans la très grande majorité des cas avec des ensembles de points très superposés (Figure 43-2) et non pas avec des ensembles disjoints (Figure 43-1). Ces deux situations peuvent pourtant donner des résultats en apparence similaires si on les résume aux seules distances globales :

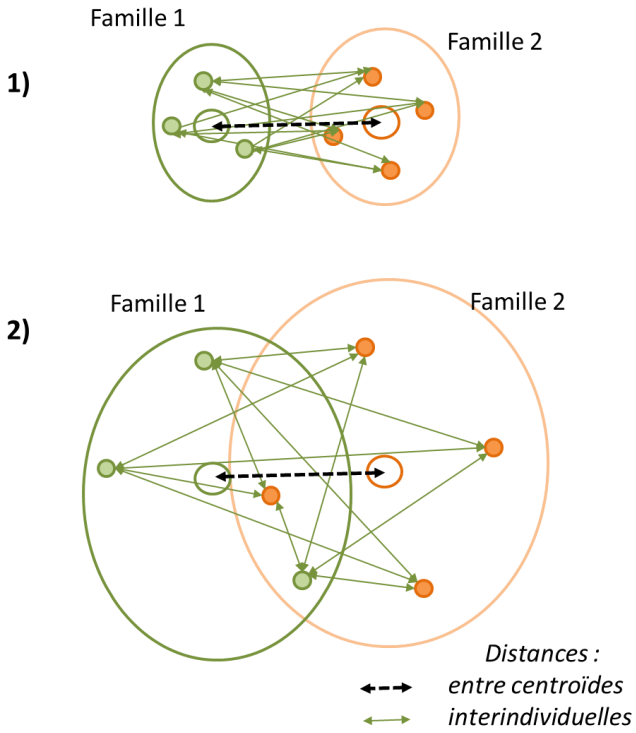


Figure 43 : importance de la dispersion des points lors de la comparaison de deux familles

Si dans la Figure 43 les deux distances inter-centroïdes est la même, la dispersion des points autour varie énormément. Quand on doit trier des familles en fonction de leur proximité, cela revient à évaluer la significativité de la distance inter-centroïdes (ou celle moyennant les distances interindividuelles) : on se retrouve dans la situation des tests statistiques standard comparant deux moyennes M_A et M_B (test t de Student par exemple).

$$t = \frac{M_A - M_B}{\left(\frac{V}{N_A} + \frac{V}{N_B}\right)^{1/2}}$$

N_A et N_B étant les effectifs des deux moyennes comparées

et V l'estimation de la variance commune = $[(N_A-1)\,V_A + (N_B-1)\,V_B] / (N_A + N_B - 2)$

Dans nos calculs de distance, il faudra donc évaluer l'intérêt donc cette correction, même si de manière générale, la dispersion des points par famille est souvent similaire.

Faisons ces calculs avec les données du Tableau 11. Comme de coutume, les données centrées et réduites servent pour les calculs ; lors des calculs des distances interindividuelles, seules les variables ayant des données en correspondance pour chaque individu sont prisent en compte. Ainsi, la distance entre un individu sain et un autre cancéreux n'inclut pas la variable « âge du cancer » alors que c'est le cas pour deux cancéreux. Le résultat moyen ne porte ensuite que sur le nombre de variables présentes.

Exemples de distance calculée à partir des données du Tableau 11 :

- Entre les individus 1 et 4 = $[((1 - 1)^2 + (-1 - 1)^2) / 2]^{½}$ = 1.414

- Entre les individus 1 et 10 = $[((1-1)^2 + (1-1)^2) / 2]^{½} = 0$
- Entre les individus 2 et 4 = $[((-1-1)^2 + (1-1)^2 + (-0.248 + 0.62)^2 / 3]^{½} = 1.175$
- Entre les individus 2 et 6 = $[((1-1)^2 + (1-1)^2 + (-0.248 + 1.736)^2 / 3]^{½} = 0.859$

Les individus 1 et 4 sont de même sexe mais le premier n'a pas de cancer, d'où le calcul sur seulement 2 paramètres. Les individus 1 et 10 sont de même sexe et n'ont pas de cancer : ils sont donc semblables d'où une distance nulle. Les individus 2 et 4 ont tous deux eu un cancer à des âges similaires mais ne sont pas de même sexe : ils demeurent malgré tout proches. Enfin Les individus 2 et 6 sont de même sexe et ont eu chacun un cancer ce qui explique leur proximité en dépit d'un écart entre les âges des cancers de 20 ans. Au vu de ces exemples, le calcul de la distance entre deux familles en moyennant toutes ces distances interindividuelles apparaît pertinent. Les distances calculées entre les 3 familles de l'exemple selon les deux modalités envisagées sont résumées dans le Tableau 12.

Tableau 12 : différences entre les distances interfamiliales selon le mode de calcul

	Distance entre familles calculée avec les :								
	A) Centroïdes			**B) Distances interindividuelles**			**Variation en % (B-A)/A**		
Familles	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
2	1,043			1,765			69,2%		
3	1,347	2,066		1,732	2,174		28,5%	5,2%	
1+2			3,217			3,413			6,1%
1+3		1,682			2,020			20,1%	
2+3	1,212			1,747			44,1%		

On constate que l'utilisation des centroïdes familiaux pour le calcul des distances interfamiliales induit des biais qui peuvent être très importants puisque l'on observe dans cet exemple des différences allant de 5% à 69% d'avec la moyenne des distances interindividuelles. L'utilisation des distances interfamiliales sur la base de leur barycentre semble donc peu appropriée et un retour aux données individuelles paraît donc indispensable si l'on veut effectuer des analyses en clusters optimales. En outre, on pourrait se contenter des barycentres pour le classement s'il existait une proportionnalité stricte entre les deux modes de calcul. Hélas, on constate que le premier mode conclut que les familles les plus proches sont les familles n°1 et 2, alors qu'avec le second, la famille n°1 est plus proche de la n°3 que de la n°2.

Si l'on applique la correction suivante de la moyenne des distances interindividuelles entre 2 familles A et B par la variance commune V :

$$d'(A,B) = \frac{d(A,B)}{\sqrt{\frac{V}{N_A} + \frac{V}{N_B}}}$$

On obtient de nouvelles distances d'. Elles apparaissent dans les 3 dernières colonnes du Tableau 13 :

Tableau 13 : distances interindividuelles moyennes et résultats après correction

	B) Distances inter-individuelles			C) Distances B corrigées		
Familles	**1**	**2**	**3**	**1**	**2**	**3**
2	1,765			3,273		
3	1,732	2,174		3,673	4,665	
1+2			3,413			8,039
1+3		2,020			4,792	
2+3	1,747			3,614		

Si les mesures se trouvent très nettement réévaluées, l'importance de cette correction réside dans le fait que l'ordre des proximités interfamiliales change : ainsi sans correction, les familles 1 et 3 sont les plus proches, tandis qu'après correction, ce sont de nouveau les familles 1 et 2.

On pourrait se poser la question de savoir au final quel calcul est le plus juste ? L'utilisation des coordonnées des centroïdes laisse supposer que l'âge moyen de la déclaration des cancers des individus malades est valide même pour les individus indemnes (qu'ils soient encore vivants ou pas d'ailleurs), ce qui est évidemment une généralisation abusive. C'est d'autant plus grave que les familles sont réduites et qu'il n'y a que de rares cancers, mais diagnostiqués jeune par exemple. Toutefois cela n'implique pas que le second mode de calcul soit exact non plus : en effet, les individus indemnes, sauf s'ils sont décédés, sont encore susceptibles de développer des cancers. En raison de ces "censures", le résultat n'est donc là encore qu'une approximation, en quelque sorte un instantané à la validité très momentanée. Malgré cette faiblesse, ce second mode de calcul est un reflet moins biaisé de la réalité et c'est sur cette base que les développements suivants seront réalisés. Quant à la correction de ce dernier mode de calcul par la variance commune, il apparaît que ce n'est pas sans conséquence et cette mesure paraît judicieuse puisqu'elle permet de tenir compte de dispersions différentes tant des individus que de leurs paramètres au sein de diverses familles. Des simulations basés sur ces différents modes de calculs seront effectuer afin d'évaluer leurs avantages et leurs inconvénients.

5.3 Quelles autres "métriques" utiliser dans le clustering ?

L'exemple précédent, bien que démonstratif, n'illustre qu'une partie de la complexification de l'analyse en clusters qu'induit la double hiérarchie. En effet en *clustering*, plusieurs "*métriques*" peuvent être utilisées pour réaliser le regroupement des familles. L'inertie utilisée avec la méthode de Ward est directement proportionnelle à la distance euclidienne entre les éléments à classer (si l'on utilise naturellement ce type de distance). Face à la difficulté liée aux données manquantes/parcellaires ou similairement du fait de caractéristiques différentes dépendantes du sexe des individus (âge de la ménopause…), on peut opter pour une approche plus précise que le calcul de distances interindividuelles : au lieu de considérer les familles comme des groupes d'individus, on peut les considérer comme des nuages de « cellules », chaque cellule étant une caractéristique d'un individu dans une famille. De cette manière, quand on effectue les calculs d'inertie, on ne prend plus en compte le nombre d'individus, mais celui de cellules, d'où une plus grande précision et un contournement de l'absence de rectangularité des données (pas le même nombre de caractéristiques renseignées par individu). Nous avons détaillé cela dans l'annexe 8.3. En résumant, si nous utilisons les données

individuelles pour les calculs, nous disposons de quatre méthodes de clustering doublement hiérarchique :

- Celle basée sur la moyenne des distances interindividuelle
- La deuxième basée sur cette même moyenne mais corrigée par une fonction de la variance commune
- La troisième part du niveau « cellulaire » et ne prend en compte que la plus faible variation d'inertie (ou de somme des carrés des écarts (SCE) à la moyenne) en utilisant la seule composante de l'inertie intra-variables.
- Enfin la quatrième qui évalue la distance interfamiliale en faisant la différence entre la variance globale une fois les paires de familles ou leurs regroupements réalisés et la variance commune calculée avec leurs variances préalables.

En outre, nous pouvons aussi effectuer des analyses en clusters standard en utilisant les données moyennées par variable et par famille – c'est à dire avec un enregistrement par famille – soit de manière primaire sans aucune considération sur la notion de génération, soit en utilisant les données de synthèse de nos sous-arbres établis sur 2 générations, telles que décrites au paragraphe 3.3. Au total, cela fait donc 6 métriques différentes à tester.

5.4 Remarques sur la réalisation des tests

Résumons tout d'abord ce qui doit être comparé pour valider nos diverses approches. L'objectif du clustering est de déterminer les familles les plus proches – en espérant que cela corresponde à une similarité phénotypique liée à la prédisposition au cancer – puis à constituer itérativement les groupes de familles.

Par ailleurs, nous devons évaluer les différences entre trois contextes génétiques principaux :

- L'absence de susceptibilité génétique (population témoin)
- La présence d'une unique mutation, délétère à elle seule, dans des familles
- L'existence d'une mutation faiblement pénétrante peu fréquente qui, en présence d'un polymorphisme non délétère assez fréquent, deviendrait très délétère.

Le jeu d'essai doit donc incorporer ces trois catégories de familles, en nombre suffisant pour que des tests répétés puissent être réalisés (c'est à dire sur des sous-populations tirées au sort à chaque fois). Enfin, pour répondre à la problématique de cette thèse, nous devons disposer des données familiales sous trois formats :

- Les données brutes moyennées par famille, telles qu'elles sont généralement utilisées dans les régressions logistiques évaluant les facteurs prédictifs de mutations (ex. score d'Eisinger[64]).
- Les données de synthèse des sous-arbres généalogiques qui sont des données analogues, mais ventilées sur les membres de la famille de base (père, mère, fils et fille)
- Les données de base individuelles incluant les appartenances familiales

Les tests itératifs, effectués chacun suite à un tirage au sort des familles au sein du jeu de données des 300 familles (décrit au §3.4), devront être réalisé pour chaque format des données, à savoir en utilisant nos 4 méthodes de clustering sur les données individuelles, et des clusterings standards pour les deux premiers formats (moyennes par familles et moyennes par sous-arbre).

Pour finir, il nous faut choisir un moyen de comparer entre eux les différents classements issus des clusterings, ce que nous développons dans le paragraphe suivant.

5.5 Méthodes de comparaison des dendrogrammes résultant du clustering

La question ici est : comment juger que tel ou tel classement issu d'une analyse en clusters est meilleur qu'un autre ? Par exemple, dans la Figure 44 ci-dessous, qu'est ce qui nous permet d'affirmer, hormis l'expérience, que le classement du clustering n°4 est meilleur que les trois autres ? Si le clustering de chaque type d'approche produisait un nombre constant de clusters, on pourrait utiliser plusieurs tests pour comparer chaque partitionnement aux catégories mutationnelles paramétrées comme un simple test du Chi² (ici pour le clustering de la Figure 41, la probabilité associée au Chi², si l'on coupe le dendrogramme en 3 classes, est égale à 0.0012), ou encore le Kappa de Cohen qui mesure l'accord entre observateurs lors d'un codage qualitatif en catégories).

Malheureusement, toutes ces solutions sont inapplicables ici car le clustering fournit rarement des groupements quantitativement « homogènes » (par exemple 3 ou 4 clusters bien identifiés) : c'est d'autant plus vrai si l'on utilise des méthodes d'agrégation comme le saut minimum, moyen ou maximum où les dendrogrammes ressemblent à des escaliers montant plus ou moins lentement. Par ailleurs, des groupements, même justifiés, peuvent apparaître à divers endroits du dendrogramme, ce qui introduit un second niveau de complication : comment détermine-t-on que tel cluster correspond ou non à tel groupe de départ ?

Considérons l'exemple suivant cherchant à classer 35 familles, les bleues étant porteuses d'une mutation délétère de style BRCA1-BRCA2 et celles en orange de mutations qui demandent la présence d'un polymorphisme pour devenir délétère. Les pénétrances sont toujours de 80% et la fréquence du polymorphisme dans la population est de 30%. Ces familles peuvent donc avoir des phénotypes assez divergents mais si l'on considère les âges moyens des cancers, ils seront identiques : une gageure donc pour l'algorithme de classement.

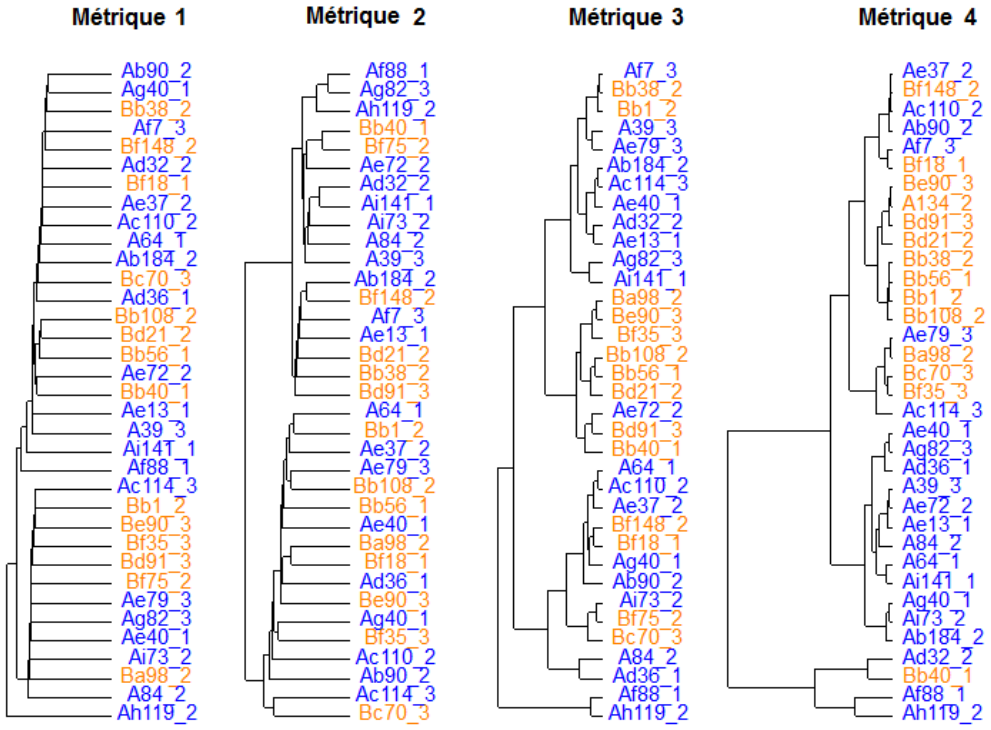


Figure 44 : 4 dendrogrammes issus de la même méthode de clustering (Ward) mais avec des métriques différentes

Subjectivement, on va conclure que la 4ème métrique classe un peu mieux que la 3ème et beaucoup mieux que les 2 premières.

Que proposer alors ? Une première méthode assez rudimentaire mais qui a l'avantage d'être robuste est envisageable : celle de compter les voisins "corrects" de chaque élément classé (c'est à dire issus de la même classe au départ). En cas de classement optimal ou non, il est invariant quel que soit l'ordre des clusters. On peut en outre faire le même calcul pour les triplets corrects, etc. Et si l'on continue ainsi, on peut calculer Mrgr, le niveau moyen de regroupement (ou effectif moyen des groupes) :

$$\text{Mrgr} = \frac{1}{N} \sum_{i=1}^{k} n_i \, E_i$$ où E_i correspond à la valeur de la classe (son effectif donc)

n_i le nombre de groupes d'effectif E_i et k le nombre de classes

Cette moyenne sera d'autant plus élevée que les groupements seront fournis. Par ailleurs, on peut assez facilement juger de l'efficacité de nos 3 indicateurs : il suffit pour cela de développer une méthode analogue à la statistique GAP servant à déterminer le nombre optimal de partitions en k-means clustering, ce que l'on a vu dans notre chapitre consacré à l'analyse en composantes principales (cf. § 4.3.2.5). En effet, rien n'interdit de calculer sur des données aléatoires quels seraient les nombres moyens de bons voisinages (pour 1 ou 2 voisins) ou encore le niveau de regroupement moyen. En répétant ce calcul une trentaine de fois (ou plus), on obtient des paramètres de répartition de ces valeurs (moyenne et écart-type) assez stables qui se distribuent selon la loi normale. Il suffit alors d'évaluer la compatibilité ou non des valeurs trouvées suite aux clusterings avec celles issues du hasard. Ceci a été fait pour les classements de nos 35 familles de la Figure 44 et une cinquantaine d'itérations pour le classement aléatoire. Le nombre moyen de bons voisinages unaires est en moyenne de 17.1 $\pm$ un écart-type de 2.7, celui de deux bons voisins à 8.2 $\pm$ 3.2 et enfin le regroupement moyen s'établit à 3.0 $\pm$ 0.7. Bien sûr ces chiffres sont fonction du nombre de catégories et des effectifs par catégorie.

Tableau 14 : analyse de voisinage des 35 familles avec les dendrogrammes de la Figure 44

	Nombre de voisins de même couleur			Nombre de paires de voisins de même couleur			Taille moyenne des groupements		
Métrique	observé	attendu	p	observé	attendu	p	observé	attendu	p
1	18	$17,1 \pm 2,7$	0,58	11	$8,2 \pm 3,2$	0,35	3,26	$3,0 \pm 0,7$	0,59
2	17		0,90	8		0,99	3,00		0,96
3	24		0,01	15		0,04	5,09		0,0032
4	26		0,0008	21		0,000056	8,80		$2\ 10^{-15}$

Ces résultats confirment la conclusion intuitive tirée ci-dessus. Les métriques 3 et 4 aboutissent à des classements significativement différents de ce que l'on pourrait observer du fait du hasard : par exemple des niveaux de regroupement à 5.09 et 8.80 respectivement alors qu'on s'attendrait à 3.0 ± 0.7. Par contre les deux premières métriques ne produisent pas de classement s'écartant du hasard.

Au vu de ces résultats, le regroupement moyen et le simple pourcentage d'individus avec un bon voisinage semblent des méthodes de bonne qualité pour estimer la qualité du classement suite au clustering. Sans doute serait-t-il intéressant de persévérer dans cette recherche d'estimateurs, en particulier sachant que des voisins de même catégorie initiale mais appartenant à des clusters franchement distants (mais pourtant contigus), ne seront pas distingués des proximités intra-cluster avec nos algorithmes. Peut-être l'utilisation des distances entre chaque élément sur lesquelles se base le clustering pourrait permettre de moduler la notion de voisinage. Enfin d'autres approches n'ont pas donné de résultat intéressant pour la comparaison des classements : une première rapportant le nombre de bons voisins à celui de mauvais voisins et une seconde basée sur un calcul de vraisemblance du bon voisinage.

5.6 Validation des méthodes sur les 300 familles fictives

Cette validation s'est déroulée en deux temps. Le premier a consisté à évaluer les 6 modèles en utilisant nos trois groupes de familles : celles sans mutation, les deuxièmes avec une mutation délétère à elle seule et les troisièmes porteuses de mutations délétères seulement quand elles interagissent avec un polymorphisme assez fréquent sans impact lui-même. Dans ces trois groupes, les familles sans mutation sont les plus éloignées : ce contexte est intéressant pour évaluer l'aptitude des 6 modèles à bien discriminer les 3 catégories et non la première des deux autres seulement. Le deuxième temps consiste à mettre le focus sur les deux groupes de familles mutées : en réalisant les tests sur ces seules familles, on évalue l'aptitude des 6 modèles à distinguer des différences moins prononcées de phénotype.

5.6.1 Aptitude des diverses méthodes à distinguer les 3 catégories de risque : sans mutation, mutation unique et mutation nécessitant une interaction

Les tests ont été réalisés itérativement sur des sous-populations d'au moins 100 familles tirées au sort au sein des 300 familles de la base. Une vingtaine de jeux d'essais étaient suffisants pour permettre une évaluation correcte des méthodes :

Tableau 15 : comparaison des résultats des 6 approches sur le classement des familles dans les 3 catégories de risque. Première partie, le % d'amélioration de classement de chaque méthode et seconde partie les probabilités médianes associées.

		1 voisin	2 voisins	Niveau de regroupement
gain moyen	distances interindividuelles	28%	74%	42%
	distances corrigées	10%	31%	18%
	delta inertie intra-variables	40%	106%	84%
	delta variances	4%	21%	9%
	données moyennes / fam.	**62%**	**172%**	**147,7%**
	données des sous-arbres	51%	139%	111%
probabilité médiane	distances interindividuelles	0,010	0,002	0,002
	distances corrigées	0,201	0,096	0,127
	delta inertie intra-variables	$1{,}8\ 10^{-4}$	$9{,}4\ 10^{-6}$	10^{-8}
	delta variances	0,524	0,427	0,404
	données moyennes / fam.	**6,7 10-8**	**6,1 10-11**	**10-8**
	données des sous-arbres	$7{,}7\ 10^{-6}$	$6{,}5\ 10^{-8}$	10^{-8}

Le gain moyen de chaque méthode est mesuré grâce au pourcentage calculé entre la valeur observée et celle attendue suite à la cinquantaine d'itérations sur données aléatoires. Les probabilités de la seconde partie du tableau sont les médianes des probabilités obtenues lors de 20 jeux d'essais.

L'inefficacité des méthodes basées sur les distances interindividuelles, qu'elles soient corrigées ou non, est confirmée au vu du Tableau 15. Le calcul basé sur la différence entre variances globale et commune ne s'avère pas très discriminant non plus. Quant au clustering basé sur la minimisation de l'inertie intra-variables, il apparaît presque aussi efficace pour le classement des familles que les clusterings directs sur les données familiales moyennées ou encore celles issues des sous-arbres. Parmi ces deux dernières méthodes, la plus simple, basée sur l'utilisation des données moyennées par famille, apparaît la plus efficace, avec des gains de voisinage systématiquement supérieurs aux chiffres attendus mais aussi aux résultats des autres méthodes.

Un autre résultat transparaît quand on compare le niveau moyen de regroupement, mais cette fois-ci par catégorie. Dans les clusterings effectués avec les moyennes brutes par variable (ligne 5), les familles se retrouvaient par "paquets moyens" de 8.0 (IC-95% [6.2 - 9.7]), 4.7 [3.7 - 5.7] et 2.2 [1.9 - 2.4] respectivement pour (1) les familles sans mutation, (2) avec mutation unique et enfin (3) avec mutation nécessitant une interaction. Ces différences entre niveau de regroupement étaient assorties d'une probabilité p = 0.002 quand on comparait les groupes (1) et (2), p = 0.0004 pour les groupes (2) et (3) et enfin p = 0.00004 pour les groupes (1) et (3). Ces chiffres indiquent que cette méthode de clustering arrive mieux à identifier les familles sans mutation que celles des deux autres catégories, ce qui n'est pas optimal pour notre problématique générale.

Les approches plus complexes utilisant les données individuelles ne semble donc pas apporter de bénéfice majeur quand on se trouve en présence de critères assez facilement discriminants, ici globalement avec ou sans risque mutationnel.

5.6.2 Aptitude des diverses méthodes à distinguer les mutations délétères seules des mutations nécessitant une interaction

Les mêmes calculs ont été effectués mais cette fois-ci dans le but de distinguer les mutations uniques ne nécessitant pas la présence d'autre polymorphisme pour être délétère, de celles nécessitant cette interaction. Dans ce jeu d'essai, le polymorphisme avait une fréquence de 30% et la pénétrance pour les deux types de risque de cancer était de 80%.

Tableau 16 : comparaison des résultats des 6 approches sur le classement des familles dans leur efficacité à distinguer les mutations délétères selon qu'elles nécessitent ou non une interaction avec un polymorphisme

	1 voisin			2 voisins			niveau de regroupement		
	Gain moyen	IC-95%	p	Gain moyen	IC-95%	p	Gain moyen	IC-95%	p
dist. Interindividuelles	8%	[3,3; 11,9]	0,271	18%	[9,0; 26,3]	0,352	15%	[5,3; 23,7]	0,404
distances corrigées	4%	[0,2; 8,1]	0,358	14%	[4,9; 22,1]	0,268	10%	[2,0; 19,2]	0,400
delta inertie intra-var.	**15%**	**[11,6; 18,4]**	**0,046**	**39%**	**[32,0; 45,4]**	**0,017**	**67%**	**[50,3; 84,1]**	**0,000027**
delta variances	3%	[-0,2; 5,5]	0,474	7%	[-0,5; 14,8]	0,341	6%	[-1,8; 13,7]	0,361
données moyennes / fam.	9%	[4,9; 12,8]	0,167	30%	[21,0; 38,1]	0,021	45%	[31,5; 57,9]	0,0055
données des sous-arbres	8%	[5,0; 11,3]	0,352	21%	[14,6; 26,9]	0,176	30%	[14,6; 45,2]	0,124

Quand il s'agit de distinguer plus finement des situations, c'est à dire quand les différences sont plus nuancées, le clustering à partir des données individuelles est nettement plus performant si l'on utilise pour regrouper les classes la minimisation de l'inertie intra-variables (ligne 3). Les gains, quelle que soit la colonne, sont toujours supérieurs et les probabilités associées toujours significatives alors que pour les données familiales moyennes (ligne 5) et celles issues des sous-arbres (ligne 6) apparaissent ici moins discriminantes. Si l'on compare les métriques 3 et 5 selon les 3 critères à l'aide d'un test-t de Student (pour séries appariées), les probabilités sont respectivement de 0.02, 0,10 et 0,06, donc en limite de significativité. Le doublement du nombre de tests devrait permettre de renforcer la significativité.

5.7 Comparaison des méthodes de clustering quant à la distinction des mutations BRCA1-BRCA2

Le dernier test que nous avons effectué porte sur population réelle de 408 familles dont au moins un des membres est porteur d'une mutation BRCA (231 familles mutées BRCA1 et 177 mutées BRCA2). Ces familles sont issues de notre base de données oncogénétique et chaque famille devait, pour être sélectionnée ici, avoir au moins une dizaine de membres enregistrés. Comme on le sait, le risque de cancer du sein est à peu près similaire dans ces deux catégories de mutation BRCA, mais la prédisposition à d'autres cancers peut faire une différence : par exemple pour BRCA1 on observe davantage de cancer du côlon et pour BRCA2, plus de cancer du sein chez l'homme et de cancer du

pancréas[65]. Nous avons comparé le clustering standard basé sur les données moyennées par famille (CAH) au clustering doublement hiérarchique (CAH²) calculé à partir des données individuelles et en considérant la métrique optimale « plus petite variation d'inertie intra-variable ». Les résultats que nous trouvons sont moins favorables à nos hypothèses qu'attendu. Ils sont présentés dans le tableau suivant :

Tableau 17 : comparaison de l'aptitude des deux méthodes de clustering à distinguer les familles mutées BRCA1 ou BRCA2 (probabilités associées aux divers tests et pour les Chi² après retrait des clusters avec un effectif < 5)

	CAH²	CAH
1 bon voisin	0,25	0,0024
2 bons voisins	0,19	0,00053
Groupement moyen	0,0071	< 10-7
Chi² 2 clusters	0.55	0,0019
Chi² 4 clusters	0,0046	0,00037
Chi² 8 clusters	0,0049	0,0022
Chi² 10 clusters	0,0011	0,0000048
Chi² 20 clusters	0,008	0,000012

Dans les trois premières lignes de ce tableau, on retrouve les tests qui nous ont servis pour comparer nos dendrogrammes et dans les 5 dernières, nous avons calculés les Chi² associant les clusters aux deux mutations, c'est à dire par exemple si dans les deux premiers clusters (les plus globaux), on note une proportion différente ou non de mutations BRCA1 ou BRCA2. Au vu du tableau ci-dessus, la CAH standard sur données moyennées par famille s'avère toujours plus efficace et discriminante que le clustering doublement hiérarchique. Si cette analyse ne permet pas de mettre en avant notre méthode de clustering avec retour aux données individuelles, elle confirme cependant l'intérêt de nos indicateurs de bon classement des dendrogrammes, en particulier le regroupement moyen du fait de leur cohérence avec les analyses de Chi².

En aparté, on notera dans les dendrogrammes par variable le regroupement des cancers du pancréas, de la prostate et de l'endomètre. Par ailleurs, on peut observer dans les deux dendrogrammes un lien entre le nombre de fausses couches et le cancer de l'ovaire :

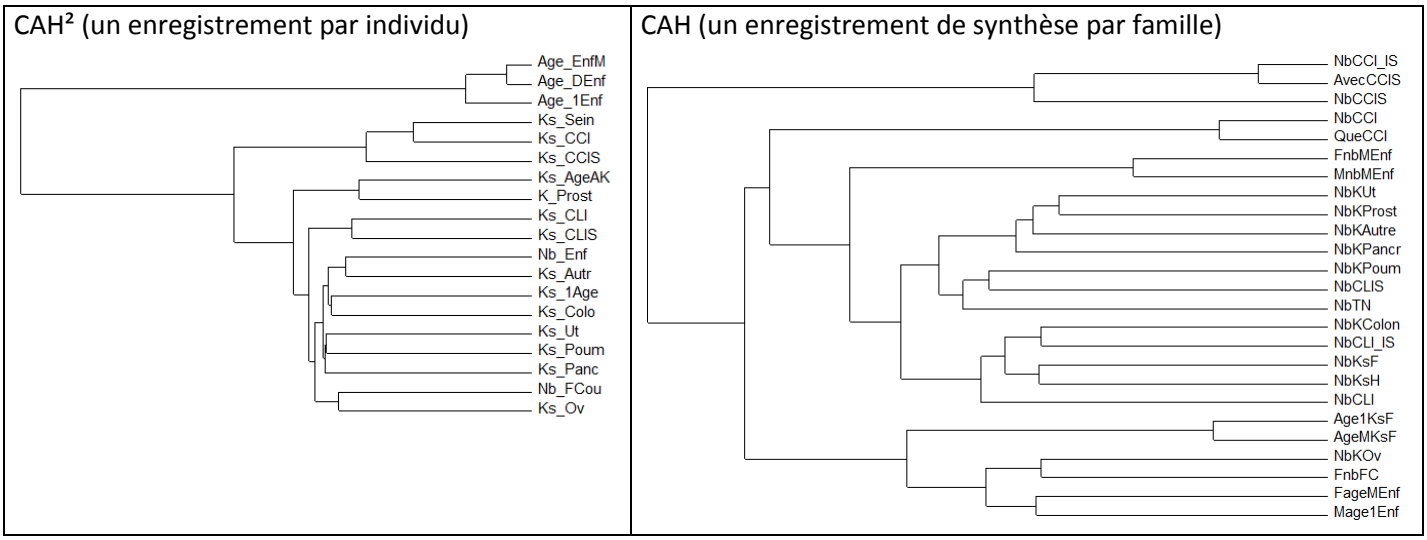


Figure 45 : dendrogrammes des variables associés aux deux méthodes de clustering (codification des variables : Age_EnfM, FAgeMEnf et MAgeMEnf = âge moyen à la naissance des enfants ; idem mais avec 1 ou D à la place de M pour le premier et le dernier enfant ; Nb_Fcou et FnbFC = fausses couches ; Ks_1Age et Ks_AgeAK = âge de déclaration du cancer ; les autres codes correspondent aux localisations)

Ce que l'on doit ajouter, c'est qu'il est quand même très difficile de comparer l'efficacité des deux méthodes sur des bases « égales » en raison du nombre différent de variables synthétiques que l'on peut générer par familles ainsi que l'on peut le constater par la hauteur plus importante du dendrogramme de droite. Malgré tout, la significativité des tests effectués laisse peu de place au doute quant à la moindre efficacité de la CAH² dans cette population.

5.8 Conclusion de la partie 5

Comme dans le clustering suite à l'ACP (partie 4), l'utilisation des données générées par les sous-arbres n'apporte aucun gain dans la discrimination entre les divers types de mutation. Par ailleurs, si la prise en compte dans la CAH des données individuelles, introduisant ainsi un nouveau niveau hiérarchique, a semblé améliorer les performances en termes de discrimination fine avec les données simulées, nous ne confirmons pas son avantage sur les données réelles. D'autres tests seraient sans doute nécessaires pour conclure définitivement sur ce point. Les analyses sur les données moyennées, étant capables de discriminer les mutations BRCA1 et BRCA2, elles devraient permettre d'isoler des sous-groupes de familles à risque spécifique malgré des phénotypes assez ressemblants. L'inclusion des données anatomopathologiques des tumeurs du sein et la connaissance des autres cancers dans la famille produisent des partitionnements plus efficace des familles sans mutation connue, possiblement intéressants quant à la recherche d'association entre des mutations faiblement délétères et divers polymorphismes. La gageure suivante est de mettre en évidence les associations opérant dans chaque sous-groupe. Des méthodes ont été développées récemment pour prendre en compte de manière améliorée la problématique de la multiplication des associations possibles (≈28 000² pour 2 gènes, 28 000³ pour trois gènes, etc.) en procédant à des approches par étape[66].

Un article est prévu pour publier ce travail méthodologique et présenter nos résultats.

6 Algorithmes de prédiction des génotypes par les phénotypes

Dans les parties précédentes, nous nous sommes attachés à simuler des données ayant le plus possible des caractéristiques communes avec les données issues des familles exposées à un risque élevé de cancer, en particulier en utilisant les informations telles que la natalité selon les populations étudiées. Ensuite nous avons poussé nos investigations en utilisant des données réelles. Mais il est aussi possible d'obtenir un grand nombre d'informations en réalisant des simulations à partir de données modélisées de manière plus restrictive, d'une part sans tenir compte des paramètres populationnels et de l'autre à l'aide d'arbres généalogiques stéréotypés. Toutefois, les phénotypes et les mutations introduits dans ces familles correspondent aux susceptibilités génétiques standard des familles porteuses de mutations BRCA1-BRCA2. Dans ce cadre, les travaux réalisés sont de deux ordres : les premiers portent sur l'évaluation de la qualité informationnelle des arbres généalogiques selon la configuration qui leur est donnée. Les seconds comparent différents schémas mutationnels (pas de mutation délétère, des mutations délétères à elles seules et enfin des mutations à effet croisé) à l'aide de deux outils mathématiques usuels : la minimisation de distance et les réseaux de neurones.

6.1 Design et performance du pronostic de mutation délétère basés sur les antécédents familiaux

Contrairement aux parties précédentes, il s'agit ici de questionner la structure des arbres généalogiques et la qualité de l'information qu'ils fournissent grâce à un cadre mathématique rigoureux doublé d'une approche contextuelle simplifiée. Ainsi, pour la constitution des modèles, nous allons partir comme au paragraphe 3.4, d'une maladie de type unique nommée K qui correspondra ici à un cancer de localisation quelconque. Evidemment il pourrait aussi s'agir de n'importe quelle autre maladie (cardiovasculaire, neurodégénérative, etc.). Conformément à l'objet de notre étude, l'occurrence et la précocité de cette maladie seront favorisées par la présence de mutations délétères constitutionnelles[i], donc transmissibles.

Que les antécédents soient recueillis par l'onco-généticien spécialement dans le but d'établir un diagnostic familial ou qu'ils soient préexistants dans une base de données, il est nécessaire de se questionner sur le périmètre souhaitable de cette exploration familiale, c'est à dire la forme et la taille de l'arbre généalogique à prendre en compte pour que le pronostic mutationnel soit le meilleur, c'est à dire *in fine* que les tests génétiques aient de plus grandes chances de détecter des mutations délétères connues. Pour cela,

- un plus grand nombre de parents est-il un gage de meilleure prédiction ?
- est-il souhaitable de remonter autant de générations que possible ?
- faut-il inclure les fratries des lignées ancestrales de même que les cousins plus ou moins éloignés ?
- le nombre de cas de maladie fait-il varier sensiblement la précision du pronostic ?

[i] On oppose les mutations constitutionnelles, obtenues d'un de ses parents, aux mutations somatiques, présentes dans les cellules cancéreuses, se produisant de manière plus ou moins spontanées (des facteurs de risque comme le tabac, l'alcool, interviennent ici) et plus fréquemment dans les cellules à fort taux de renouvellement (muqueuses, peau, voies digestives, etc.).

- quelles sensibilité et spécificité peut-on typiquement atteindre par ce genre de méthode et quelle est leur variabilité en fonction des paramètres du modèle ?

6.1.1 Description du modèle et des données

Le fondement du modèle aléatoire est que les individus d'une même famille ont des génotypes stochastiquement liés entre eux par les lois de Mendel (une chance sur deux d'hériter une mutation d'un de ses parents[i]). Mais ces génotypes sont cachés et toutes les données statistiquement utilisables (les phénotypes) sont des expressions aléatoires de ces génotypes. Mathématiquement, il s'agit d'un processus de Markov caché, indexé par un arbre.

Deux difficultés se présentent alors :

- L'explosion combinatoire du nombre de génotypes possible rend inenvisageable – sauf pour de toutes petites familles – toute approche reposant sur l'examen exhaustif de tous les génotypes éventuels, c'est à dire une méthode par force brute[ii].
- Ensuite, les mutations en question restant relativement rares en fréquence dans la population générale, l'effet à analyser reste fin et il faut le distinguer du bruit aléatoire qui prend ici la forme de survenue sporadique de la maladie, ce à un âge généralement plus avancé.

Dans les calculs suivants, la probabilité de survenue de la maladie selon qu'une mutation est impliquée ou non aura la même répartition que ce qui est constaté pour les cancers sein/ovaire avec ou sans mutation BRCA :

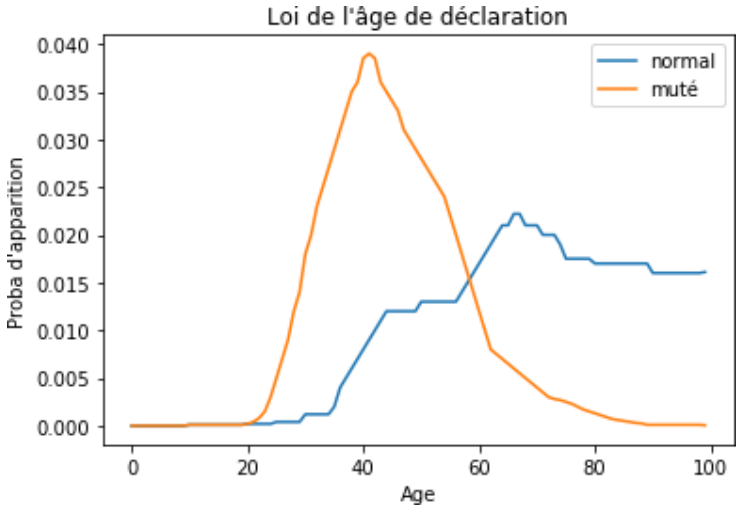


Figure 46 : lois de probabilité utilisées pour les modélisations dans les arbres généalogiques de l'apparition de cancers

[i] Cette règle pourrait être nuancée si une mutation biaise le risque qu'une fausse couche se produise, comme nous l'avons montré au 2.1.6.1. Si les gènes sont bien répartis de manière aléatoire lors de la méiose (production des gamètes), donc avec une probabilité de 1/2 de transmettre une mutation à chaque spermatozoïde ou ovule, le résultat final en termes de reproduction (les embryons viables) peut être déséquilibré en faveur d'une dissémination des mutations délétères avec une fréquence de nouveau-nés porteurs supérieure à 1/2.

[ii] Par force brute signifie qu'on calcule tous les cas possibles. Par exemple, si l'on désire « craquer » un mot de passe ainsi, on essaye toutes les combinaisons possibles de caractères, d'où de besoins en temps machine potentiellement infinis si le mot de passe est long.

En outre, nous ne considérons que le cas des mutations hétérozygotes, sachant que les mutations homozygotes, comme nous l'avons indiqué précédemment, sont généralement létales dès l'embryogenèse. Nous considérons donc que le génome de chaque individu est une quantité aléatoire ayant deux états possibles {normal, muté} avec les probabilités associées suivantes :

Tableau 18 : probabilité d'hériter d'une mutation de ses parents selon qu'ils sont ou non porteurs d'une mutation délétère

↓ mère / père →	Normal	muté
Normale	0	0.50
mutée	0.50	0.75

Nous supposons connu tous les paramètres de ce modèle :

- La fréquence de la mutation dans la population générale, f_{mut}
- La pénétrance de K pour les sujets non mutés, p_0
- La pénétrance de K pour les sujets mutés, p_1

Toutefois, on pourra faire varier ces paramètres afin d'envisager des pénétrances liées à des mutations moins délétères ou encore à des interactions entre mutations et polymorphismes très peu pénétrants seuls. Les méthodes pour évaluer ces différents contextes seront discutées au chapitre 0.

Ici, nous supposons que les données disponibles par individu sont son année de naissance, la survenue ou non de la maladie et si c'est le cas, l'âge de déclaration. Une telle approche si elle paraît sommaire présente l'avantage d'être facilement généralisable. Pour quantifier les performances des algorithmes utilisés, nous allons procéder à des simulations de données selon le modèle suivant : pour chaque arbre généalogique, nos générerons d'abord les génotypes (cachés) : pour tout individu, son génotype est obtenu aléatoirement mais en fonction de celui de ses parents s'il en a ; s'il n'en a pas dans l'arbre généalogique, on utilisera la fréquence f_{mut}. Nous affectons ensuite aléatoirement des années de naissance aux individus de façon à avoir des âges en concordance avec les générations. Enfin les phénotypes sont générés en tenant compte des génotypes et de la valeur des paramètres p_0 et p_1.

6.1.2 Principe de l'algorithme

Le calcul du risque familial de mutation au vu des cas de maladie dans la famille par le calcul de probabilité conditionnelle, tel que décrit ci-après est connu depuis longtemps[67]. Toutefois, sa mise en œuvre pour chaque cas précis est complexe. Pour cette raison les praticiens utilisent plutôt des méthodes de scoring qui nécessite de faire quelques additions ou soustractions mais sans pouvoir prétendre à la même performance[68].

Notons F l'ensemble des phénotypes de tous les membres de la famille – les données familiales connues – et désignons par G la compilation des génotypes (inconnus) des individus, en particulier G(i) désignera le génotype du $i^{ème}$ individu. Pour le modèle à mutation simple dans lequel nous travaillons jusqu'à mention du contraire, nous noterons G(i) = 1 quand l'individu i est muté et G(i) = 0 sinon.

Un calcul de probabilité élémentaire à partir du théorème de Bayes montre que :

(1) $$P\,(G(i) = 1 \mid F) = \frac{1}{1+\frac{numérateur}{dénominateur}}$$

Où

$$numérateur = \sum_{g:g(i)=0} P(F|G = g)\, P(G = g)$$

Et

$$dénominateur = \sum_{g:g(i)=1} P(F|G = g)\, P(G = g)$$

Dans chacune de ces deux expressions, la somme est étendue à tous les génotypes possibles satisfaisant la contrainte mentionnée sur le génotype g(i) de l'individu i. Quant à $P(F|G = g)$, c'est la probabilité (conditionnelle) du phénotype F sachant que le génotype est g. Nous noterons $L(F|g)$ cette vraisemblance (conditionnelle) qui se calcule par une formule explicite.

Notons que le calcul du numérateur et du dénominateur requièrent soit de scanner tous les génotypes possibles, soit de procéder par une méthode de Monte-Carlo. La première méthode, dite « par force brute », est envisageable pour les familles peu nombreuses et nécessite le calcul de la vraisemblance de chaque génotype g, i.e. $P(G = g)$. Notons d'ailleurs qu'une fraction importante de génotypes sont impossibles, donc de probabilité nulle ; la sommation est de facto réduite aux génotypes admissibles. La méthode de Monte-Carlo consiste à simuler une suite de génotypes, soit par procédé direct, soit par un algorithme de Metropolis[69]. Avec cette méthode, la formule (1) se réécrit :

(2) $$\frac{numérateur}{dénominateur} = \lim_{N\ grand} \frac{\sum_{j=1}^{N} 1_{\{G_j(i)=0\}}\, L(F\,|G_j)}{\sum_{j=1}^{N} 1_{\{G_j(i)=1\}}\, L(F\,|G_j)}$$

où G_1, G_2,… forme une suite de tirages aléatoires parmi l'ensemble des génomes possibles, suivant les lois de Mendel et les hypothèses contextuelles mentionnées précédemment. La fonction 1_A fait référence à la fonction indicatrice binaire qui vaut 1 si A est réalisée et 0 sinon. Numériquement, il faut s'assurer d'un nombre N de simulations suffisant pour qu'à la fois le numérateur et surtout le dénominateur soient estimés de façon correcte.

La probabilité (conditionnelle) de mutation obtenue par la formule (1) est une quantification du risque de mutation d'un individu mais aussi indirectement du risque familial. Il peut lui être préféré un prédicteur binaire "muté/non muté" que l'on obtiendra par seuillage : pour un niveau de seuil s, l'individu i est pronostiqué muté si $P(G(i) = muté|F) > s$. D'un certain point de vue, on peut regretter cette conversion en pronostic binaire d'un risque quantitatif *a priori* plus informatif. Toutefois cette conversion permet ensuite d'analyser le prédicteur avec les indicateurs de performance usuels, c'est à dire en termes de sensibilité (pourcentage de bien prédits parmi les mutés) et de spécificité (pourcentage de bien prédits parmi les non mutés). On présentera la traditionnelle courbe ROC (*Receiver Operating Characteristics*) et dans la suite de ce chapitre, les différents prédicteurs seront comparés via la position de leurs courbes ROC. On peut aussi envisager d'évaluer la performance sur l'ensemble de l'arbre généalogique, ce qui revient à moyenner les prédicteurs sur tous les individus de l'arbre. Les résultats sont d'ailleurs en général meilleurs mais nous les laisserons de côté pour nous concentrer sur le cas individuel.

Sur un plan plus pratique, les différentes courbes ROC présentées ci-après sont le résultat d'entre 10 000 et 100 000 itérations des calculs avec une limite basse pour le nombre de cas contribuant au calcul du dénominateur d'au moins 2 000.

6.1.3 Standardisation des arbres généalogiques

La question traitée est l'influence de la taille et de la forme de l'arbre généalogique des individus pris en compte. Dans une situation réelle, l'arbre généalogique est contraint dans sa forme par le nombre d'enfants nés des différents couples et par la capacité d'obtenir pour chacun des individus des informations fiables sur son phénotype. Ce dernier point s'applique particulièrement aux générations anciennes, parce qu'elles ont eu le temps de présenter la maladie. En conséquence, la taille et surtout la forme peuvent présenter une grande diversité.

Pour faciliter nos calculs, nous allons utiliser une forme générique d'arbre régulier dont on fait varier les deux paramètres que sont la taille (nombre d'enfants) et la hauteur (nombre de générations), et pour répondre à des questions particulières ultérieurement, nous définirons d'autres formes typiques.

Nous appelons arbre régulier à $n_g \geq 2$ générations et $n_e \geq 1$ enfants par couple et on note Reg(n_g , n_e) tout arbre dont le squelette est la lignée ancestrale d'un individu numéroté 0 qui remonte à n_g générations en incluant les ascendants des deux sexes et à ce squelette, on rajoute n_e enfants pour chacun des couples. Voici schématisé un arbre de type Reg(4, 3) :

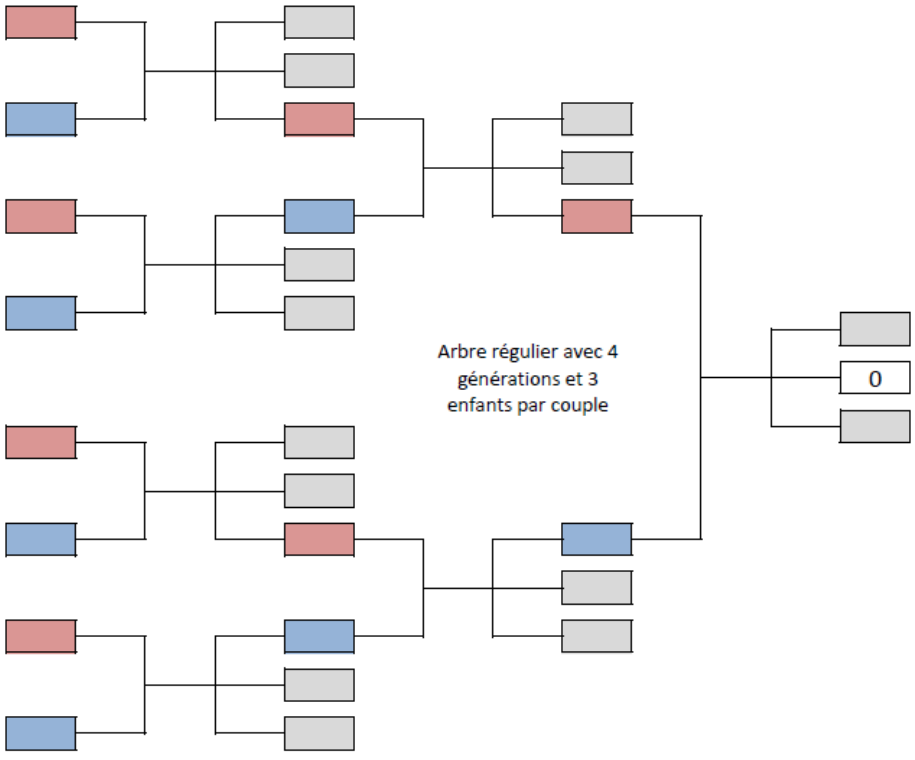


Figure 47 : schématisation d'un arbre régulier de 4 générations et de 3 enfants par couple.

Pour les comparaisons, on peut fixer n_g et faire varier $n_e \geq 1$, ou bien en fixant $n_e \geq 1$ et en faisant varier $n_g \geq 2$. Une idée intuitive, voire naïve et que plus il y a d'individu pris en compte (richesse du

phénotype), et plus la performance du prédicteur augmente, vraisemblablement jusqu'à une certaine limite. Une question connexe mais plus subtile est d'évaluer l'importance de la forme à taille fixée : vaut-il mieux plus de générations mais moins d'individus par générations que le contraire ? En particulier nous comparerons les performances des prédicteurs pour deux arbres comportant tous deux 15 individus. Le premier est Reg(4, 1) que nous appellerons par la suite « Grd grd par. » car il fait intervenir jusqu'aux arrière-grands-parents. Le voici :

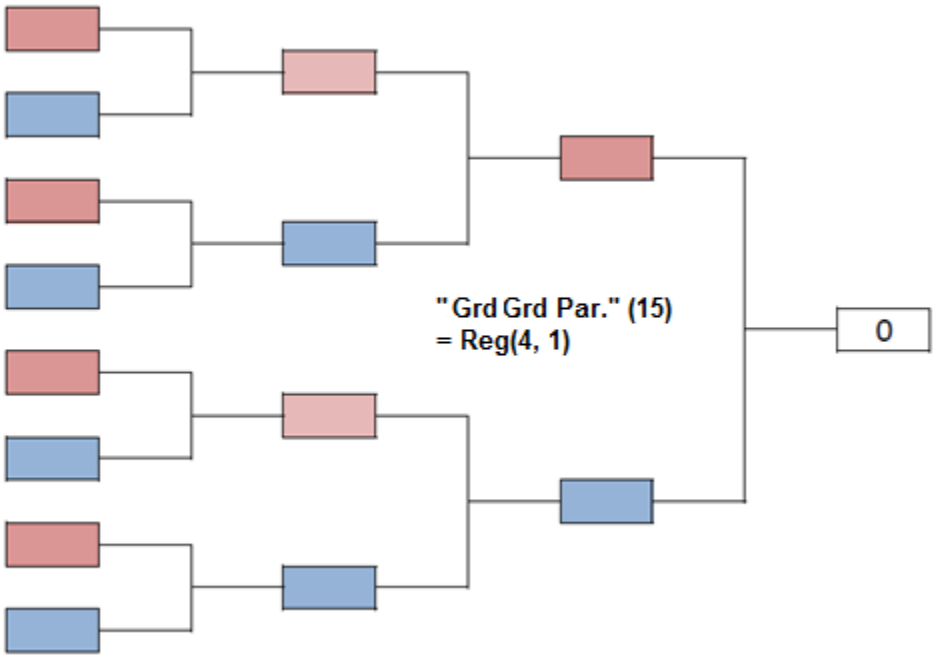


Figure 48 : arbre régulier avec 4 générations et 1 enfant par couple

Le second que nous appelons « Wide 3 gen. » comporte lui aussi 15 membres mais que 3 générations grâce à l'ajout des oncles/tantes et des cousins.

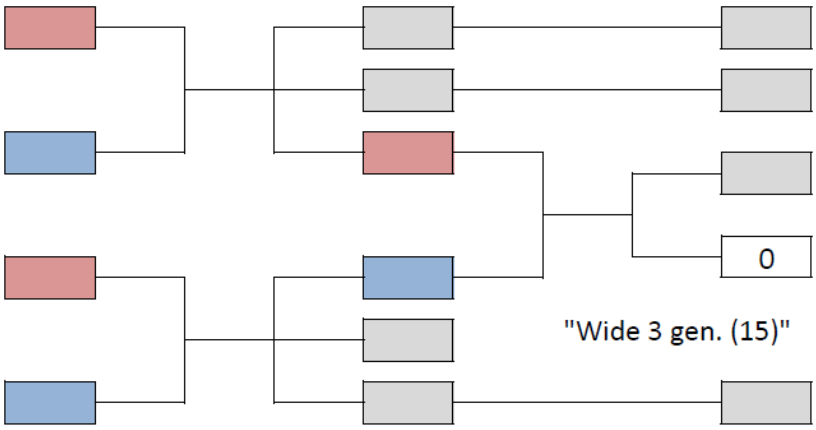


Figure 49 : design de l'arbre « wide 3 gen. » à 15 membres.

Dans le cas de l'arbre régulier Reg(n_g , n_e) avec $n_g \geq 3$ et $n_e \geq 2$, l'individu 0 a des oncles (ou tantes) mais ses cousins ne figurent pas. De façon générale, il pourrait être intéressant d'isoler l'influence d'individus un peu plus éloignés de la ligne ascendante directe de l'individu 0. Nous ferons une

comparaison de la famille « uncles » à 11 individus par rapport à une famille « cousins » à 19 individus, qui contient strictement la précédente en ajoutant 8 cousins à l'individu 0 et qui sont dessinés en vert dans le schéma ci-dessous :

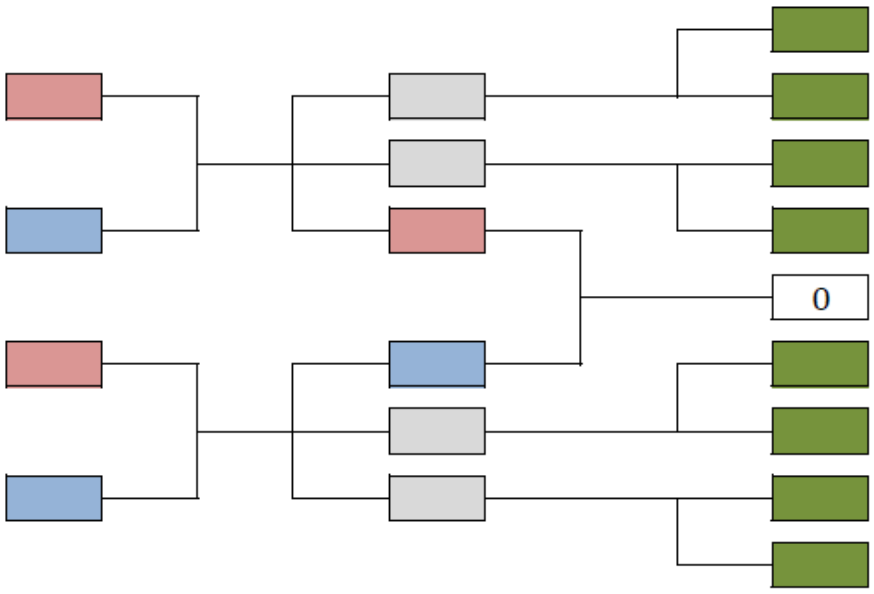


Figure 50 : Représentation de l'arbre « uncles » (11) après adjonction de 8 cousins (en vert)

Dans l'étude de l'influence de la forme, nous étudions aussi le cas d'un arbre généalogique asymétrique au sens où seuls un des parents de l'individu 0 est renseigné. Par exemple nous comparons les prédicteurs obtenus pour les deux familles suivantes de tailles presque identiques :

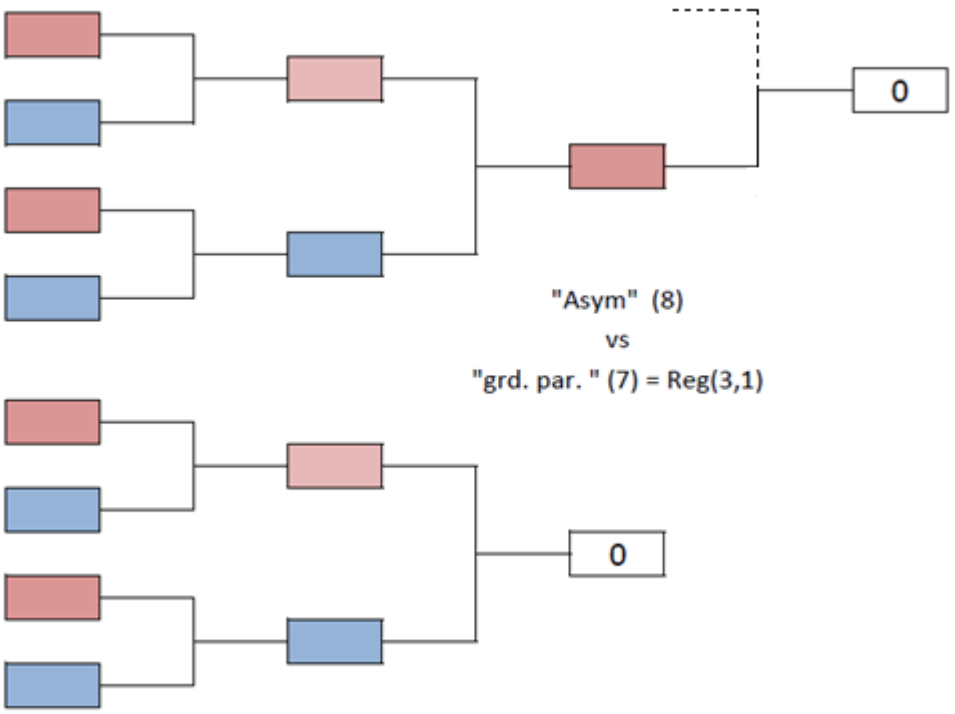


Figure 51 : Asymétrie créée entre deux arbres Reg(3, 1) par le déplacement de l'individu 0 à la génération suivante et l'absence de mère pour lui.

Dans les diagrammes précédents, l'individu 0 sur lequel porte le calcul du risque mutationnel est positionné à droite et son risque est calculé en fonction d'individus de génération égale ou antérieure représentés plus à gauche. On peut se demander à quel point les descendants peuvent eux aussi fournir une information valable sur le génotype d'un individu. Bien sûr cela ne peut fournir de renseignements que si un individu a une date de naissance assez ancienne – bien que nous avons montré (Annexe 8.2.1) que le nombre d'enfants et de fausses couches, pouvait être indicateur du risque mutationnel – ce qui peut en limiter l'intérêt pratique mais ce problème reste une question intéressante. En particulier, à taille d'arbre fixée, les ascendants donnent-ils plus d'information que les descendants ?

Rappelons que pour obtenir nos résultats numériques nous simulons des familles avec inévitablement un choix sur la loi d'année de naissance selon les générations. En l'occurrence, les années de naissances sont choisies de façon à ce que l'âge (potentiel) des individus soit uniforme dans [30, 50], [50, 70] et [70, 90] respectivement pour les 3 premières générations précédant l'individu 0. Nous utilisons le mot « potentiel » car l'individu peut être en réalité décédé plus tôt, soit de la maladie K étudiée, soit d'une autre cause. La prise en compte de l'âge est indispensable car la probabilité d'avoir contracté la maladie K avant une date donnée dépend de la durée pendant laquelle l'individu a déjà vécu. Par conséquent les individus « jeunes », par exemple de la première génération apportent *a priori* peu d'information ce qui laisse à penser que la prédiction avec les ascendants est plus performante que celle avec les descendants. Pour mener à bien cette comparaison cependant, on comparera les prédicteurs associés aux arbres généalogiques « Grd Par. » = Reg(3, 1) à 7 individus et Arr. Grd Par. » = Reg(4, 1) à 15 individus aux prédicteurs associés aux arbres suivants :

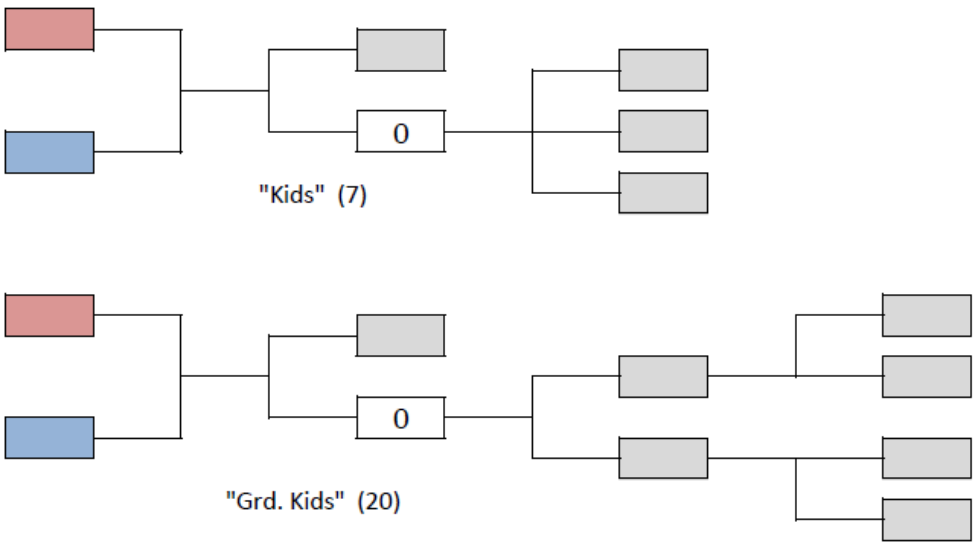


Figure 52 : deux types d'arbre avec une ou deux générations descendantes

6.1.4 Résultats

Notons que les résultats présentés dans cette section sont établis avec diverses valeurs typiques des paramètres : de l'ordre de quelques pourcents pour f_{mut}, la fréquence des mutations dans la population, et p_0 le risque de cancer durant sa vie pour les individus sans mutation et de l'ordre de 50% pour p_1 celui avec mutation.

On étudiera aussi comment la performance du prédicteur varie selon les paramètres avec – disons-le de suite – des résultats assez prévisibles. Par contre nous montrerons les effets plutôt surprenant du conditionnement du phénotype. En effet, il paraît normal de générer les arbres de façon à ce qu'au moins l'un des membres voire plusieurs présentent la maladie K. C'est ce nombre qui va orienter vers le diagnostic de « famille à risque » et donc à qui les tests génétiques seront proposés. Nous verrons qu'un tel conditionnement peut modifier la hiérarchie entre prédicteurs. Egalement, nous verrons si la qualité de la prédiction est meilleure quand la famille présente un cas ou des cas de la maladie.

6.1.4.1 Effet du nombre de générations

Nous considérons une population où existe une mutation délétère et des arbres réguliers Reg(n_g, 2) pour un nombre de générations n_g variant de 2 à 6 et un nombre d'enfants par couple égal à 2. Les courbes ROC résultant des sensibilité et spécificité du prédicteur à affecter les individus 0 à leur bon groupe sont les suivantes :

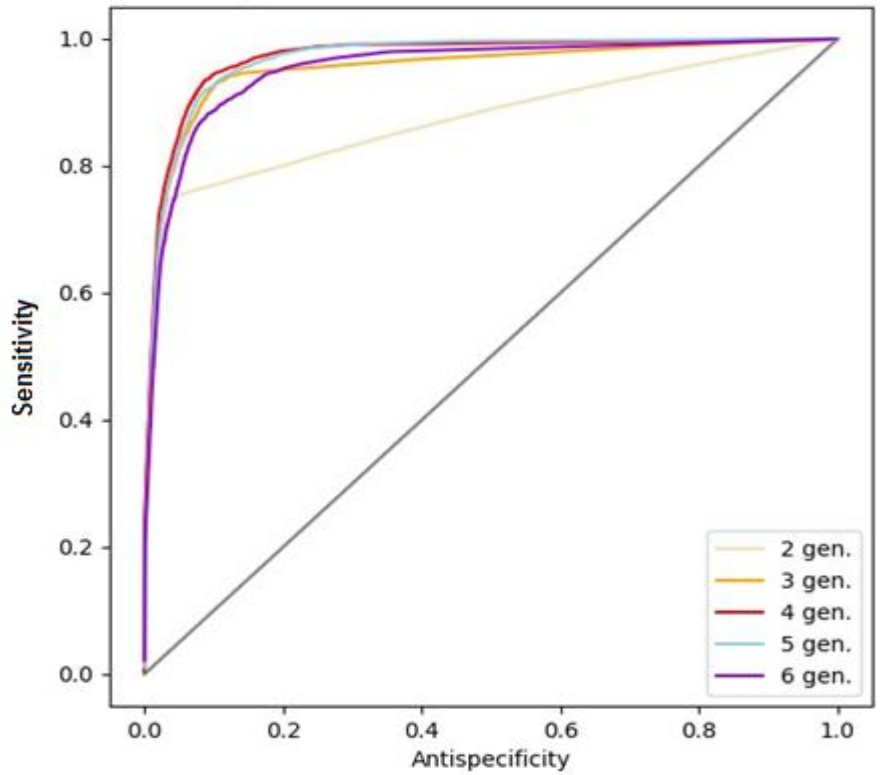


Figure 53 : influence du nombre de générations sur la performance du modèle (contexte : une mutation et deux enfants par couple).

Visiblement les données sur 2 générations seulement, c'est à dire l'individu et ses parents ne suffisent pas pour obtenir un pronostic performant. Le passage à 3 générations permet un gain important tandis que l'ajout de la 4ème génération ne procure qu'une amélioration marginale. Au delà, tout ajout

d'information sur les générations antérieures est inutile voire nuisible : les performances régressent légèrement. On constate donc que les performances du prédicteur n'augmentent plus avec la taille au-delà de 4 générations, mais se dégradent. La même chose se produit quand on fixe le nombre d'enfants par couple à 1.

6.1.4.2 Nombre d'enfants

La question de la performance selon le nombre d'enfants dans l'arbre est plus théorique que la précédente question. En effet, dans le précédent cas, le médecin/statisticien voulant faire un pronostic peut choisir le nombre de générations qu'il renseigne. Pour les enfants, leur existence, réelle, n'est pas l'objet d'un choix, même si le praticien peut choisir de n'en inclure qu'un : celui qui est dans la lignée ascendante de l'individu étudié (ex. le cas index[i]). Les courbes ROC obtenues pour les familles Reg(4, n_e) ou n_e varie entre 1 et 5 sont les suivantes :

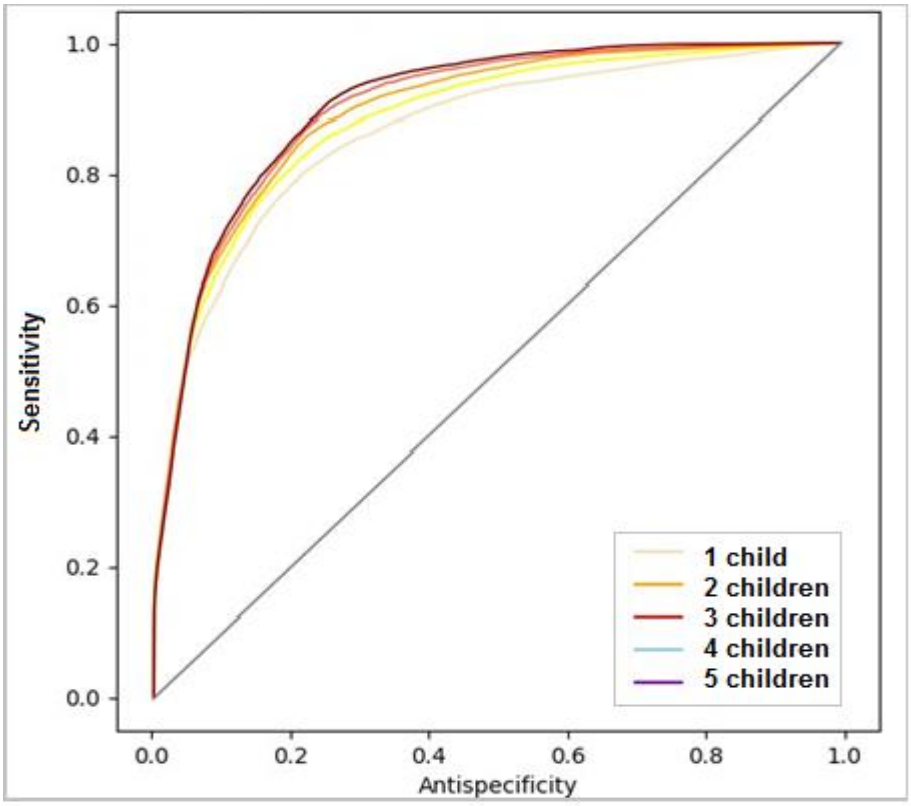


Figure 54 : influence du nombre d'enfants par couple sur la performance du modèle (contexte : une mutation et 4 générations).

On observe que plus l'arbre est large, meilleure est la prédiction avec un gain non insignifiant par enfant ajouté. Cela signifie que l'ajout d'information venant d'individus directement raccordés à la lignée ascendante est bénéfique.

6.1.4.3 Hauteur contre largeur

[i] Un cas dans une famille est dit index s'il s'agit du 1er membre de la famille qui vient en consultation. Le plus souvent, il a déclaré un cancer correspondant au syndrome familial, même si ce n'est pas obligatoire.

On vient de voir que des gains de performance des prédicteurs étaient obtenus grâce à l'augmentation du nombre de générations jusqu'à 4 (hauteur) tandis qu'il n'y avait pas de limite à ces gains pour le nombre d'enfants par fratrie (largeur). D'où la question : à taille égale (même nombre d'individus), mieux vaut-il disposer d'un arbre large ou bien d'un arbre haut ? Si l'on se réfère aux courbes ROC suivantes relatives aux arbres "Grd. Grd Par." et "Wide 3 gen." de 15 individus chacun, il semblerait qu'un arbre large fasse un peu mieux qu'un arbre haut.

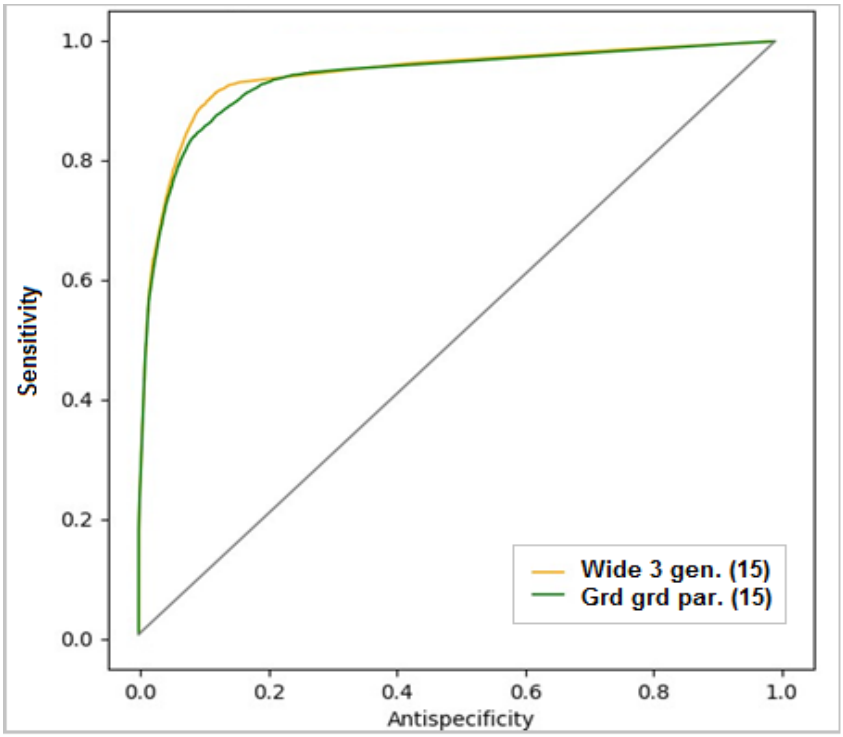


Figure 55 : comparaison de la performance parmi les arbres de 15 membres selon qu'ils sont hauts ou larges

Toutefois, le conditionnement de la présence d'au moins un cas de maladie par arbre suffit à produire un résultat nettement inverse :

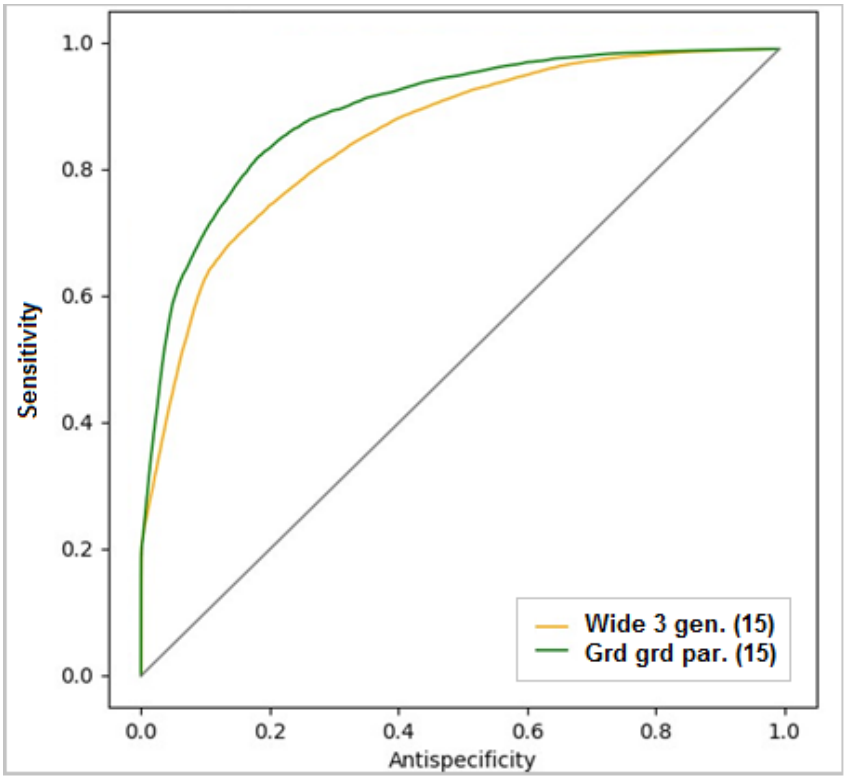


Figure 56 : comparaison de la performance parmi les arbres de 15 membres selon qu'ils sont hauts ou larges mais en "forçant" la présence d'au moins un cas de maladie par arbre

Compte tenu de la présence de cas de cancer soit sporadiques, soit familiaux dans les arbres généalogiques que l'on rencontre habituellement en oncogénétique, on peut estimer que le second cas de figure est plus pertinent et donc qu'à nombre de membres identique, la hauteur d'un arbre est plus informative que sa largeur, même si comme on l'a vu un peu plus haut qu'au-delà de 5 générations, il n'y aurait plus de bénéfice à glaner de l'information.

6.1.4.4 Utilité des cousins

Nous avons montré précédemment l'intérêt de l'information correspondant aux individus directement raccordés à la lignée ancestrale de l'individu étudié. Il est logique d'étudier maintenant le cas de parents un peu plus éloignés en commençant par les cousins. On compare donc les prédicteurs associés à la famille « uncles » et « cousins » entre eux mais aussi à l'arbre Reg(3, 1) sans oncle ni cousin :

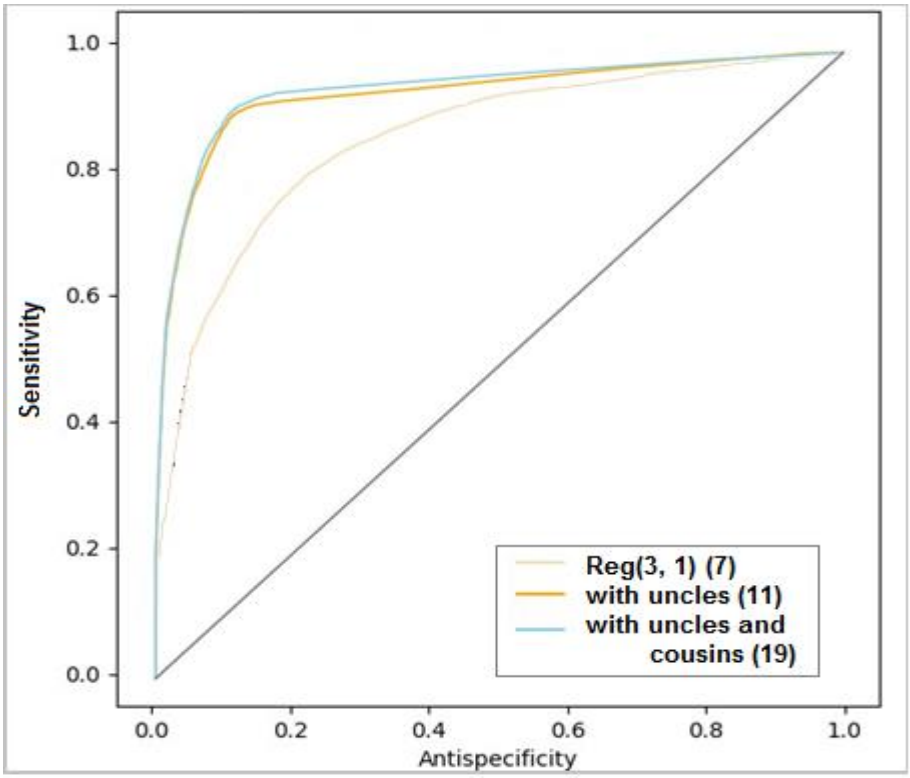


Figure 57 : influence sur le prédicteur de la présence des oncles et des cousins dans l'arbre (contexte : 1 mutation, 3 générations, 1 enfant par couple, sauf lors de l'ajout des oncles et des cousins)

Cette figure est intéressante car elle montre que les liens indirects sont eux aussi très informatifs – ceci en écho à la conclusion quant au nombre d'enfants par couple – mais que l'ajout des cousins quand on a les oncles n'apporte pas d'éléments nettement contributifs. En conditionnant par la présence d'au moins un malade par famille, le verdict est quasiment identique avec un bénéfice accru toutefois avec l'ajout des cousins, mais une légère dégradation globale des prédicteurs. On reviendra ultérieurement sur ce phénomène peu intuitif.

6.1.4.5 Asymétrie de l'arbre

On se place dans le cas extrême où il manque toute information phénotypique sur le père ou la mère ainsi que toute leur ascendance. Est-ce alors illusoire d'espérer un pronostic de mutation ? Nous comparons les prédicteurs associés des trois arbres suivants :

- La famille « asym » formée du père et des grands et arrière-grands-parents paternels
- La famille « Grd grd par. » avec les grands et arrière-grands-parents des deux côtés.
- La famille « Grd par. » avec grands-parents mais sans arrière-grands-parents des deux côtés.

Les courbes ROC suivantes sont obtenues :

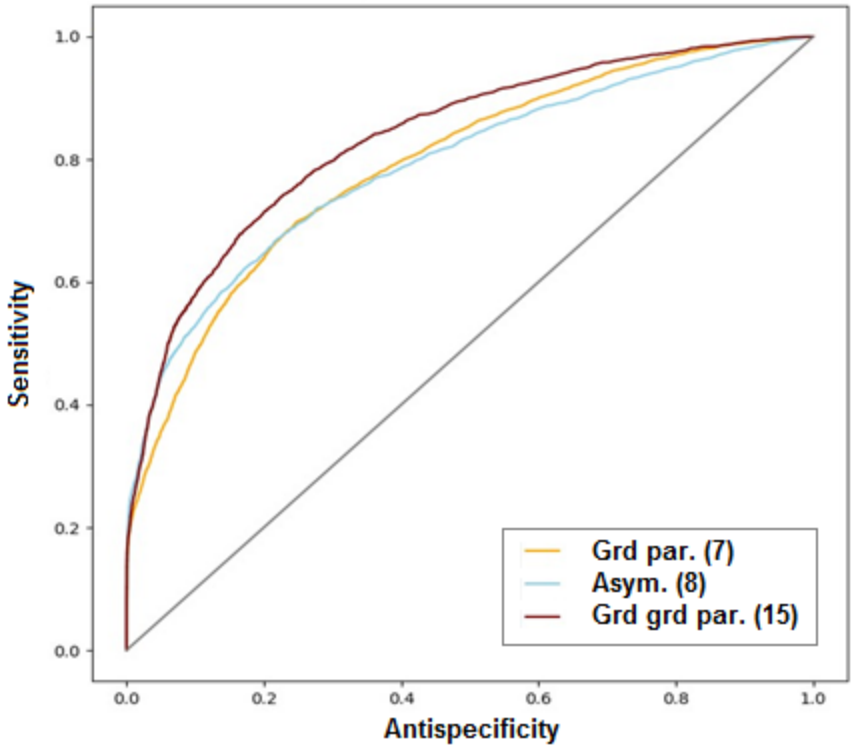


Figure 58 : Effet de l'asymétrie sur la performance du prédicteur (contexte : 1 mutation et un enfant par couple)

La perte de symétrie induit donc bien une perte d'information importante par rapport à un arbre avec ses deux branches, ce à quoi on pouvait s'attendre, l'arbre asymétrique produisant une courbe Roc analogue à un arbre symétrique avec une génération de moins.

6.1.4.6 Ascendants contre descendants

Nous avons affirmé précédemment que les prédictions utilisant les descendants devaient être moins bonnes que celles incluant les ascendants puisque les générations récentes sont moins longtemps exposées au risque de maladie. Nous avons testé cette affirmation et voici le résultat dans le contexte standard (A) et dans celui du conditionnement d'au moins un cas de maladie par arbre (B) :

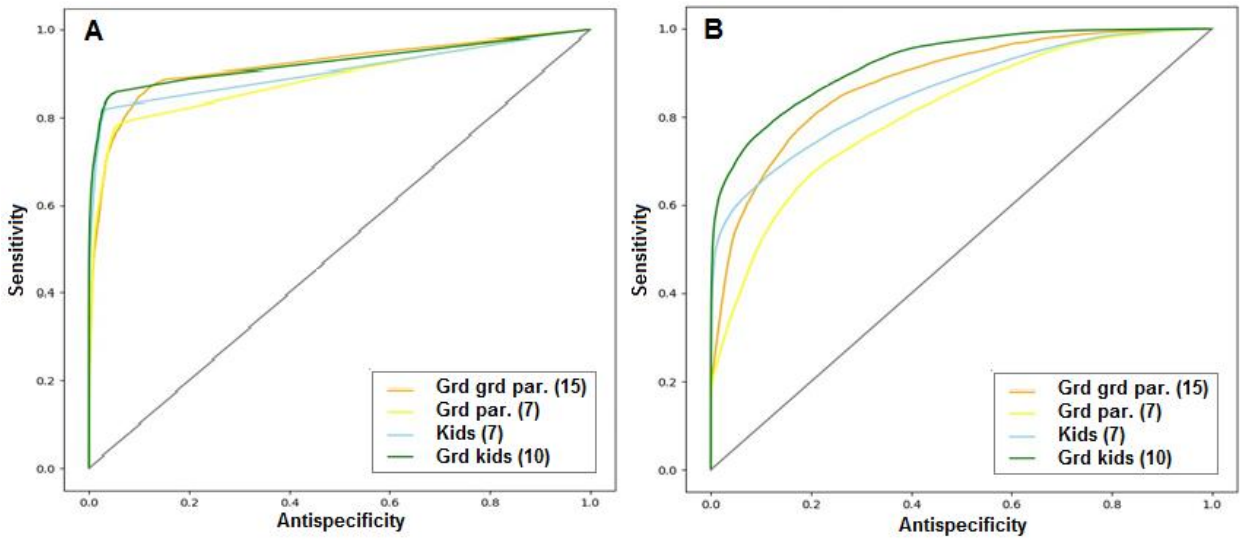


Figure 59 : comparaison des prédicteurs selon que l'arbre contient plutôt des générations ascendantes ou descendantes et selon que l'on "force" (B) ou non (A) l'occurrence d'au moins une maladie par arbre.

Ces résultats vont à l'encontre de notre *a priori*. La prédiction utilisant les enfants ou petits-enfants est meilleure à nombre de membres par arbre identique. Peut-être cela se produit-il parce qu'un arbre contenant des enfants et/ou des petits impose une condition d'âge élevé chez l'individu 0.

6.1.4.7 Influence des paramètres f_{mut}, p_0 et p_1

On constate que le prédicteur est d'autant meilleur, aussi bien en sensibilité qu'en spécificité, que la prévalence de la mutation est faible. Pour des arbres de type Reg(4, 2) contenant 22 individus, les cinq courbes ROC ci-dessous correspondent à des fréquences de mutation croissantes mais respectivement des aires sous la courbe (AUC) de plus en plus faibles :

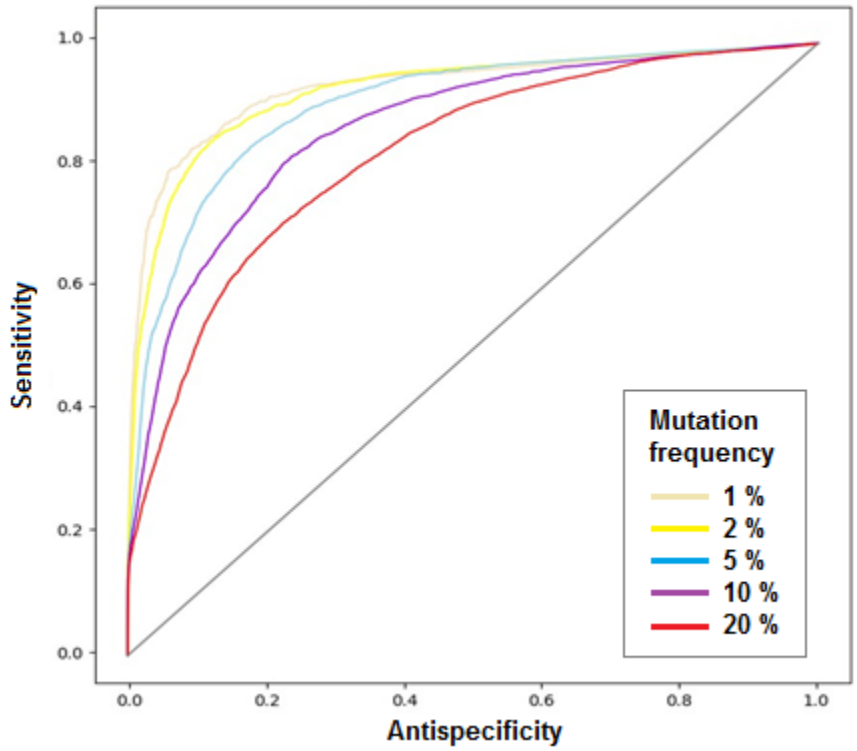


Figure 60 : performance des prédicteurs selon la prévalence de la mutation (contexte : arbres à 4 générations et deux enfants par couples, donc 22 individus)

La pénétrance de la maladie K chez les porteurs de mutation est un autre paramètre du modèle. Nous avons testé des pénétrances p_1 de 20%, 40% et 70% (similaire à celle des mutations BRCA). Mais de son côté, le risque p_0 de survenue de la maladie chez les individus non mutés (wild type) au cours de sa vie peut varier grandement et nous avons testé des valeurs allant de 0.05% à 10% :

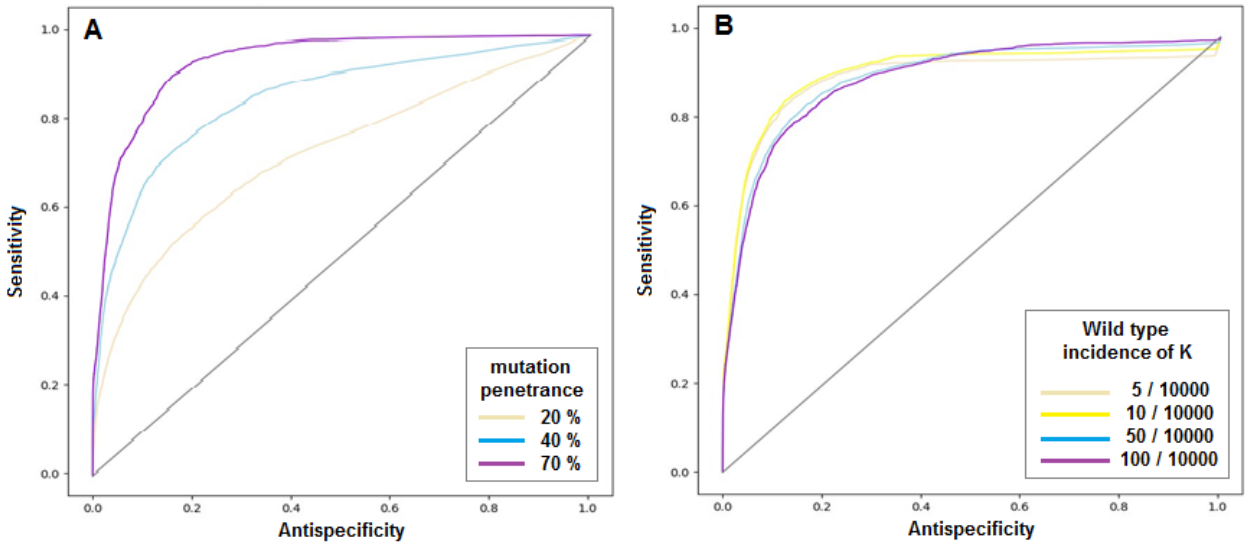


Figure 61 : influence sur la prédiction de la pénétrance de la mutation délétère (A) ou de son incidence au cours de la vie des individus non mutés (B).

Il ressort qu'une pénétrance élevée en cas de mutation améliore la qualité du prédicteur tandis qu'à l'inverse, une incidence plus élevée de la maladie dans la population générale minore cette qualité. En clair, plus grand est l'écart entre la pénétrance et l'incidence de la maladie quel que soit l'âge et meilleur devient le prédicteur.

Au total ces effets sont bien conformes à ce que l'on attend : plus la maladie K se cantonne aux individus mutés et plus la prédiction est facile. Toutefois ces paramètres n'ont pas tous la même influence. L'incidence de la maladie chez les individus sains n'est guère influente alors que la pénétrance chez les porteurs de mutation a une forte répercussion sur la performance du pronostic. La fréquence de mutation est aussi très influente et son sens de variation paraît moins intuitif : plus la mutation est fréquente et plus sa détection est difficile.

6.1.4.8 Influence du conditionnement

Comme le pronostic individuel de mutation est prioritairement proposé aux familles à risque, c'est à dire celles qui présentent déjà un certain nombre de cas de la maladie, il paraît intéressant de voir comment évoluent les performances du prédicteur quand on conditionne par le nombre total de malades dans la famille. Concrètement pour une famille Reg(4, 2) à 22 individus, dont on sait qu'elle permet de très bonnes performances, nous "forçons" le nombre minimal de malades successivement à 0, 1, 2 ou 3 et nous obtenons les courbes ROC suivantes :

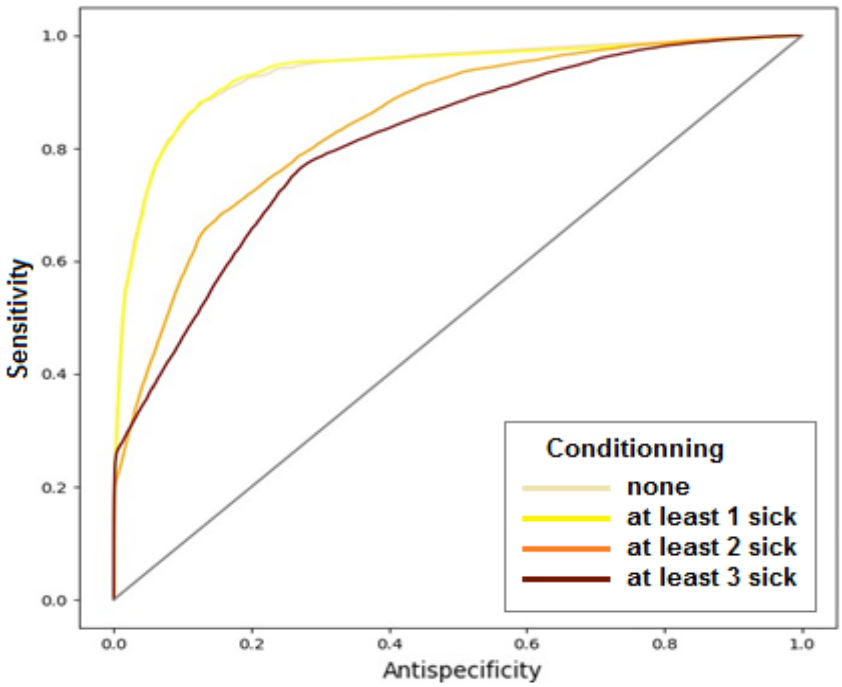


Figure 62 : Performance du prédicteur en fonction du conditionnement du nombre minimal de cas de maladie par famille (contexte : arbre Reg(4, 2), une mutation délétère)

Les deux premières courbes sont superposées et ne se distinguent guère. On remarque qu'au-delà de 1 cas de maladie, le conditionnement diminue l'efficacité du prédicteur, ce qui en fin de compte est assez logique car il devient de plus en plus difficile pour lui de distinguer entre les familles mutées des autres dès lors que ces dernières ont systématiquement des cas de maladie.

6.1.4.9 Conclusion sur l'arbitrage coût/efficacité

Les expérimentations menées précédemment sur des arbres généalogiques de taille croissante montrent qu'il existe une taille optimale pour l'optimisation de la prédiction. Cette notion d'optimum est renforcée par l'accroissement du coût de calcul avec la taille des arbres, qui peut donc décaler l'optimum pratique vers une taille encore un peu moins importante. Une différence doit aussi être faite relativement au coût de collecte des phénotypes, selon que les données soient pré-existantes dans une base de données ou recueillies spécialement pour le diagnostic de mutation. Typiquement, on cherchera à renseigner parents, grands-parents et arrière-grands-parents. Cette lignée ancestrale peut être complétée, quoique avec un gain de performance relatif. Il faut sans doute éviter d'aller trop loin sous peine de faire plutôt moins bien, comme illustré par les trois courbes ROC ci-dessous :

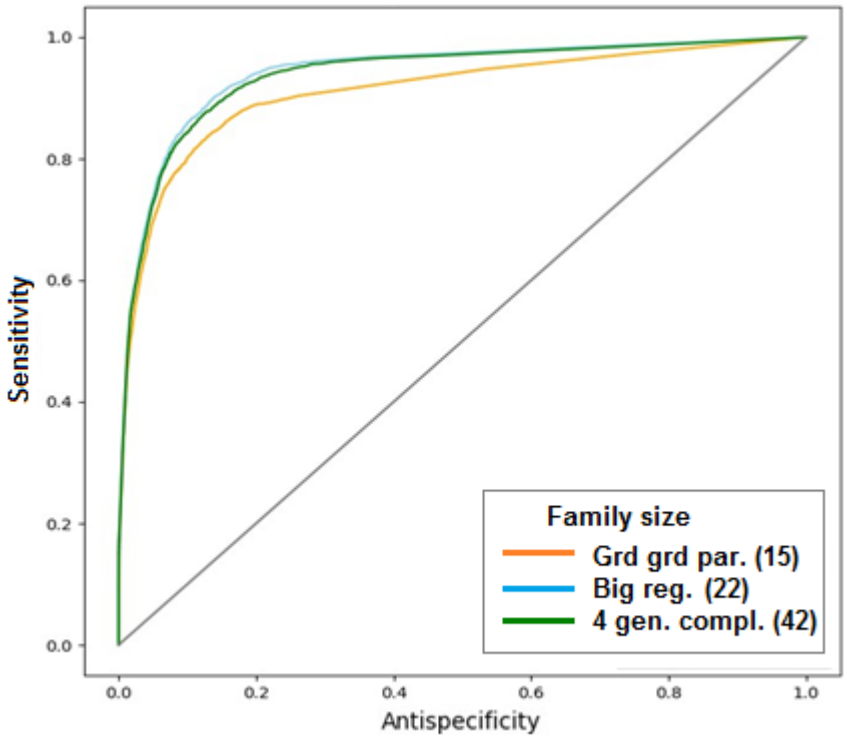


Figure 63 : performance du prédicteur selon que la taille des arbres simulés passe de 15 membres à 22 membres puis à 42 membres

On constate sur cette figure qu'on n'obtient plus de gain en ajoutant les 20 derniers membres, ce qui suggère qu'une recherche très exhaustive des apparentés collatéraux peut être plus un coût qu'une source d'information utile.

Ces arbres sont représentés de manière emboitée dans le schéma suivant :

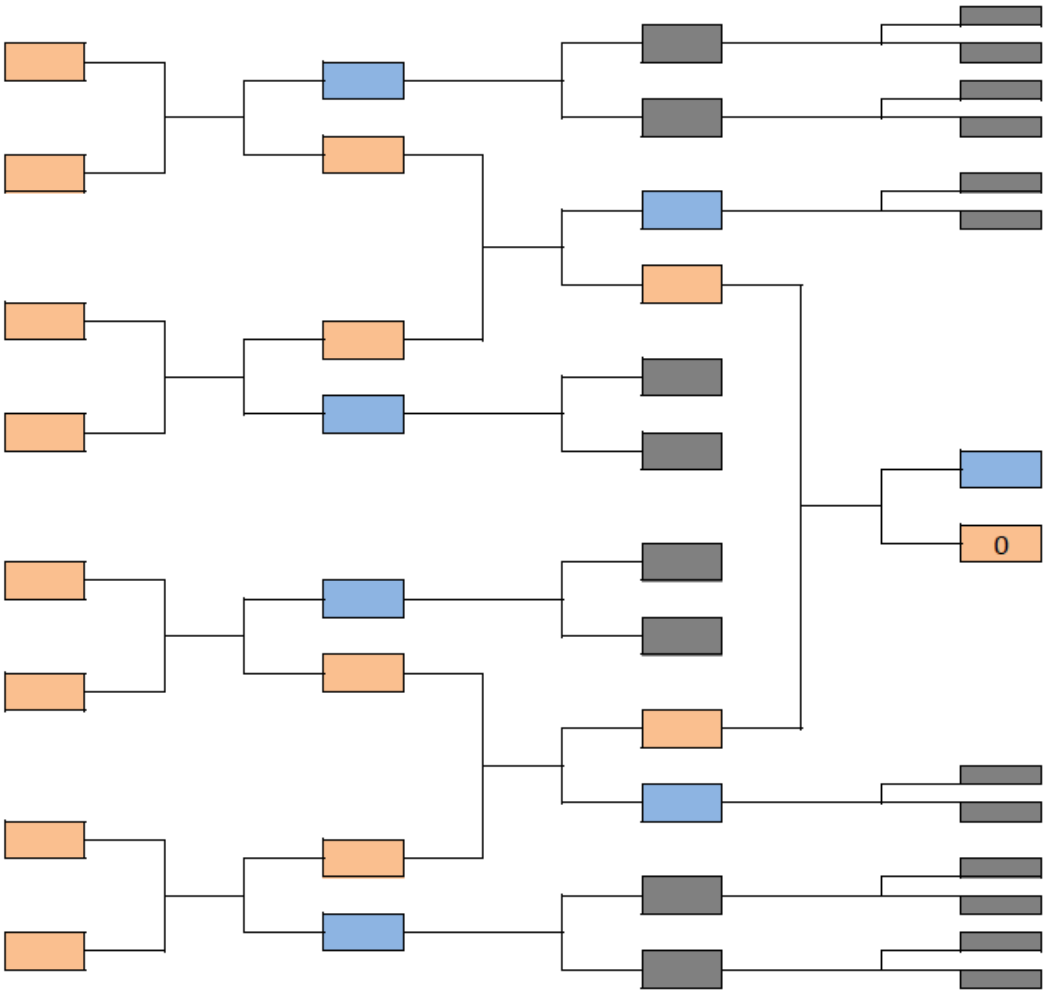


Figure 64 : Les trois arbres emboités correspondant aux courbes ROC ci-dessus : 15 membres en rose pour l'arbre de base, + 7 membres en bleu pour le deuxième (N = 22) et enfin + 20 membres en gris pour le dernier (N = 42)

6.1.5 Généralisation et prolongements

6.1.5.1 Modèle à deux mutations

Dans ce modèle, deux mutations à effet croisé sont implémentées. On pourrait sans doute les appeler des polymorphismes car séparément, elles n'ont qu'une pénétrance minime et aucun impact sur l'âge de déclaration, mais quand elles sont présentes ensemble dans le génome, le risque de cancer augmente considérablement ainsi que sa précocité, avec une pénétrance similaire aux précédents modèles à mutation simple. Le phénotype est donc régi de façon similaire, mais ce n'est pas la présence d'une mutation qui génère le sur-risque mas la présence simultanée des deux.

La différence entre les modèles simple et double réside dans la transmission des mutations qui certes, obéit aux lois de Mendel, mais aboutit à des caractéristiques différentes. Ainsi, dans le modèle double nous considérons que le génome d'un individu prend ses valeurs dans un ensemble de 4 éléments :

{ normal, muté 1, muté 2, muté 1 et 2 } = { (0, 0), (1, 0), (0, 1), (1, 1) }

Pour chaque mutation 1 ou 2, la probabilité de transmission des parents à leurs enfants est toujours régie par le Tableau 18. Dans le modèle à mutation simple, on avait vu qu'en enfant de deux parents porteurs de la même mutation hétérozygote a la probabilité 3/4 d'être porteur. Dans le modèle à double mutation, un enfant dont les deux parents sont doublement muté (toujours de manière hétérozygote), n'a maintenant plus qu'une probabilité $(3/4)^2 \approx 0.56$ d'être à son tour doublement muté. Notons que nous continuons à ne considérer que les mutations hétérozygotes en ignorant les cas homozygotes, censés auparavant être létaux. Mais dans le cas de polymorphismes non franchement délétères, les possibilités d'homozygotie pourraient légitimement être aussi considérées. Cela compliquerait d'autant les modèles [i] sans vraisemblablement apporter d'information supplémentaire aussi les avons-nous écartées.

Les paramètres de ce modèle à double mutation sont de ce fait un peu plus nombreux, à savoir :

- La fréquence de la mutation 1 dans la population générale, f_{mut1}
- La fréquence de la mutation 2 dans la population générale, f_{mut2}
- La pénétrance de K pour les sujets non mutés, p_0
- La pénétrance de K pour les sujets mutés 1 uniquement, p_1
- La pénétrance de K pour les sujets mutés 2 uniquement, p_2
- La pénétrance de K pour les sujets doublement mutés, $p_{1,2}$

Dans les cas étudiés ci-après, les valeurs de p_0, p_1 et p_2 sont proches est faibles tandis que la valeur de $p_{1,2}$ est élevée. Quant à la loi de l'âge de déclaration, elle est supposée identique, pour les porteurs de double mutation au cas des porteurs de mutation simple des chapitres précédents. Le problème revient donc à la prédiction de l'état doublement muté qui est le seul à constituer un risque significatif de maladie. La formule de probabilité de double mutation est une adaptation évidente du modèle à mutation simple (1).

(3) $$P\,(G(i) = (1,1) \mid F) = \frac{1}{1+\frac{numérateur}{dénominateur}}$$

Où

$$numérateur = \sum_{g:g(i)\neq(1,1)} P(F|G=g)\,P(G=g)$$

Et

$$dénominateur = \sum_{g:g(i)=(1,1)} P(F|G=g)\,P(G=g)$$

Dans ce modèle à deux mutations, le nombre de génotypes à scanner dans la méthode par force brute est 4^N ou N désigne la taille de l'arbre. Donc, hormis pour des arbres de taille extrêmement réduite, la méthode par simulation doit lui être préférée et s'écrit comme variante évidente de (2).

Les performances obtenues dans ce modèle à deux mutations sont globalement un peu moins bonnes que dans le modèle à une mutation. Mais les phénomènes décrits précédemment persistent. En particulier la performance atteint un optimum pratique autour d'une quinzaine d'individus comme le montre le peu de gain d'efficacité lors du passage à 22 membres puis une perte en passant à 42 dans le graphique suivant :

[i] En particulier il faudrait statuer sur le caractère létal ou non d'une double homozygotie, sur une différence de pénétrance selon qu'une des deux mutations est homozygote ou non, etc.

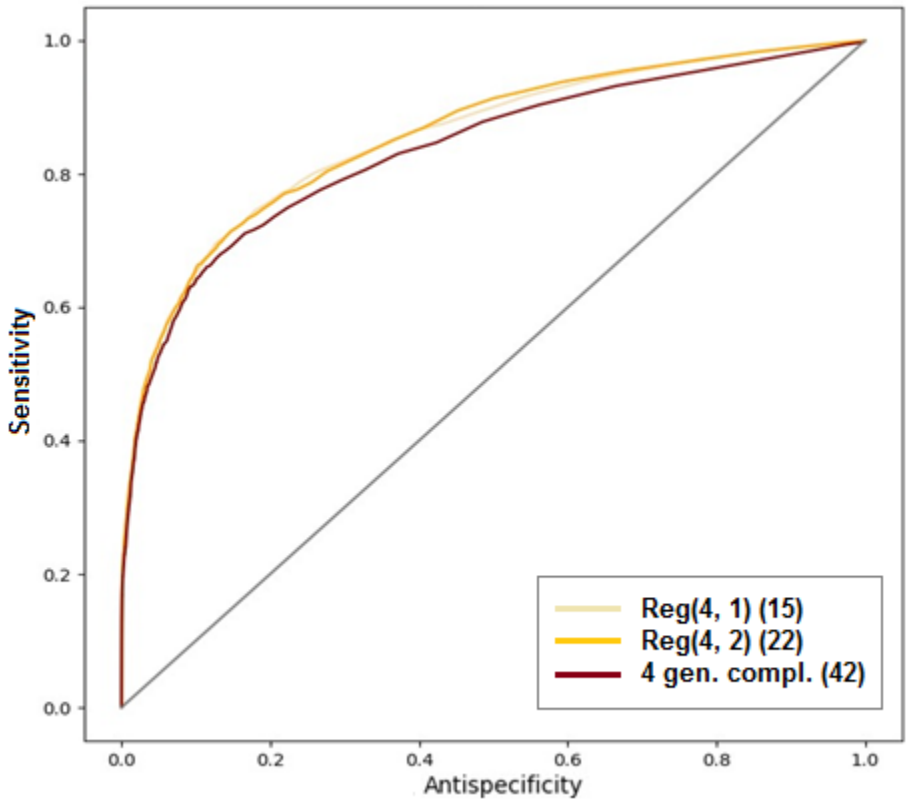


Figure 65 : comparaison de la performance du prédicteur selon le nombre de membres inclus par arbre (contexte : deux mutations interagissant et conditionnement d'au moins un cas de maladie par arbre)

6.1.5.2 Autres prédicteurs

Le calcul de la probabilité du génotype conditionnellement au phénotype est l'estimation la plus fiable du risque de mutation pour un individu. Toutefois d'autres prédicteurs se présentent naturellement à l'esprit. Compte tenu de la popularité des méthodes de maximisation de vraisemblance, un prédicteur pourrait être la valeur du génome qui maximise la vraisemblance d'un phénotype. Pour de petites familles, ce calcul de maximisation peut être fait par force brute, mais pour des tailles plus réalistes, il faut s'orienter vers un algorithme de *recuit simulé*[i]. Plaçons-nous dans le cas du modèle à mutation unique et notons F [resp. G] la variable qui compile les phénotypes [resp. génotypes] des individus de la famille. Nous notons toujours n°0 l'individu pour lequel le pronostic doit être fait. Désignons par $\hat{G}_0$ et $\hat{G}_1$ respectivement les génotypes qui maximisent la vraisemblance du phénotype avec l'individu 0 non-muté ou muté respectivement, i.e. :

$$L(F|G = \hat{G}_i) = \max\{ L(F|G = g); g(0) = i\} \quad pour\ i \in \{0,1\}$$

[i] Le recuit simulé est une méthode de programmation empirique inspirée d'un processus utilisé en métallurgie où l'on alterne des cycles de refroidissement lent et de réchauffage (recuit) qui ont pour effet de minimiser l'énergie du matériau. Cette méthode est transposée en optimisation pour trouver les extrema d'une fonction.

Un prédicteur naturel de mutation est :

$$p_1 = 1_{\{L(F|\hat{G}_1) \geq L(F|\hat{G}_0)\}}$$

C'est à dire que l'état mutationnel de l'individu 0 serait celui constaté dans le génome qui rend le phénotype le plus probable. Malheureusement, les performances de ce prédicteur sont catastrophiques. On peut le modifier pour obtenir une variante utilisable en introduisant

$$p_s = 1_{\{L(F|\hat{G}_1) \geq s\, L(F|\hat{G}_0)\}}$$

où s est un seuil à ajuster. On obtient ainsi par variation de s une famille de prédicteurs basée sur le rapport de vraisemblance. Dans le cas de petits arbres généalogiques, ces prédicteurs obtiennent des résultats très proches de la prédiction par seuillage de la probabilité conditionnelle que nous avons utilisée jusqu'à présent. En effet, sur le graphe ci-dessous, les courbes ROC obtenues par méthode de force brute pour chacun des deux types de prédicteur se confondent. Il en est de même pour la courbe relative au prédicteur par seuillage de la probabilité conditionnelle, mais cette fois calculée par simulation.

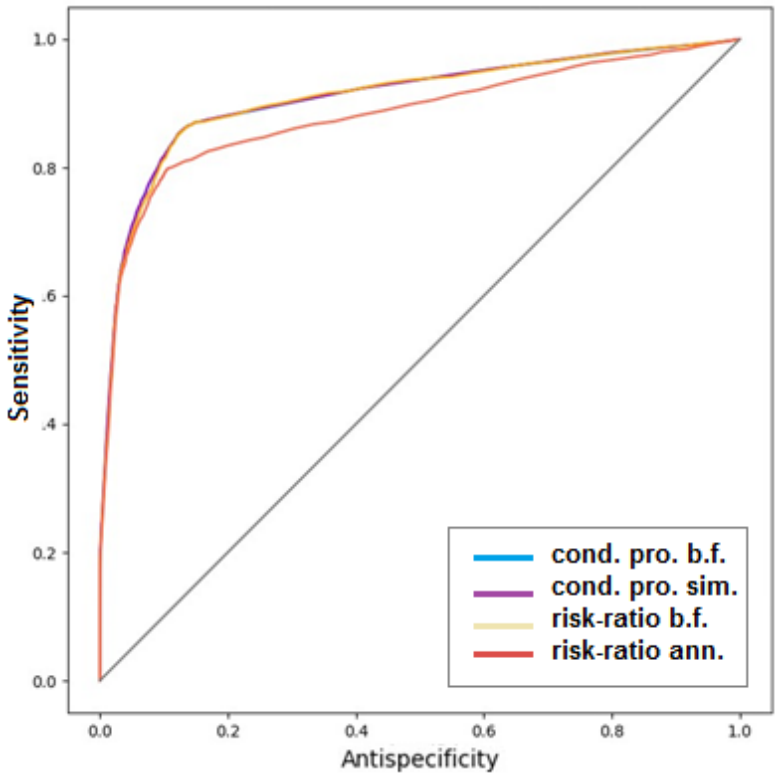


Figure 66 : comparaison de la performance des deux prédicteurs (seuillage de la probabilité conditionnelle ou rapport de vraisemblance) obtenus soit par force brute, soit par recuit simulé (*ann*. pour *simulated annealing*) (contexte : familles Reg(3, 1) de 7 individus donc et une mutation délétère)

Dans le cas de la maximisation de la vraisemblance du génome par recuit simulé, le résultat est moins bon et il faut certainement conclure que ce recuit simulé doit être amélioré soit dans le choix du noyau d'exploration de l'espace des génomes, soit dans le réglage du schéma de température (nombre de longueur des paliers de température).

Dans le cas du conditionnement à au moins un cas de maladie on obtient comme on l'a déjà vu des performances inférieures mais selon une hiérarchie identique :

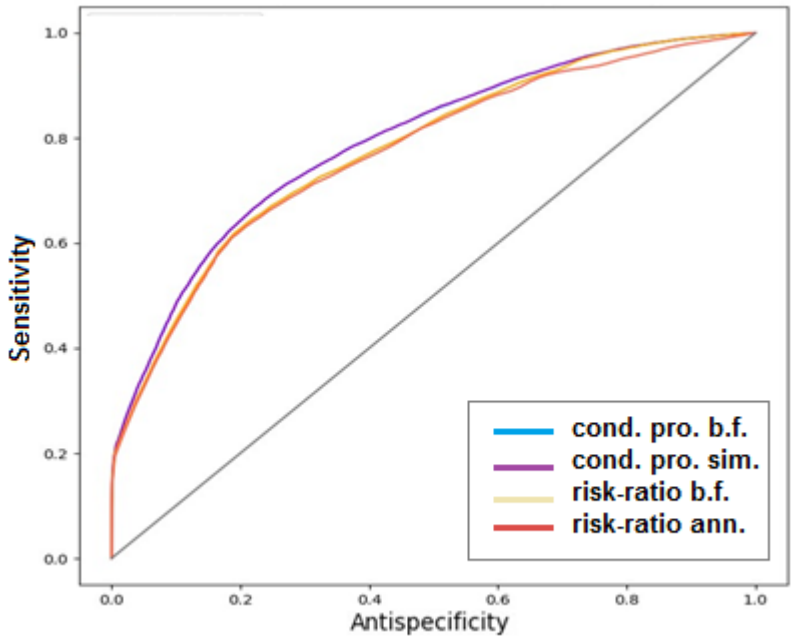


Figure 67 : comparaison de la performance des deux prédicteurs (seuillage de la probabilité conditionnelle ou rapport de vraisemblance) obtenus soit par force brute, soit par recuit simulé (contexte : familles Reg(3, 1) de 7 individus donc et une mutation délétère mais conditionnement d'au moins un cas de maladie par arbre)

Pour une famille plus nombreuse comme Reg(4,2) ayant 22 individus, il faut renoncer à la méthode par force brute et la comparaison se fait entre simulation pour le prédicteur associé à la probabilité conditionnelle et recuit simulé pour le prédicteur associé au rapport de vraisemblances des génotypes optimaux :

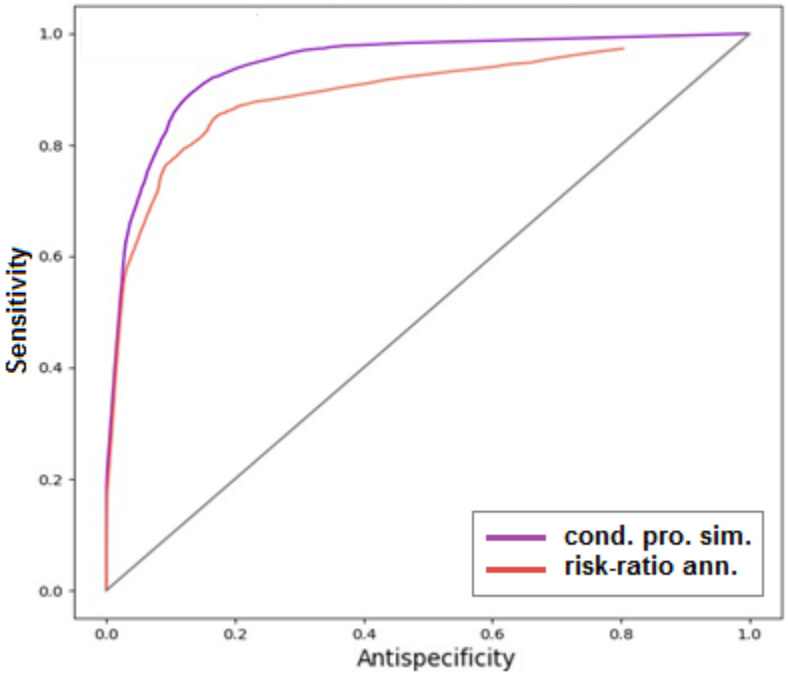


Figure 68 : comparaison de la performance des deux prédicteurs (seuillage de la probabilité conditionnelle ou rapport de vraisemblance) obtenus par recuit simulé (contexte : familles Reg(4, 2) de 22 individus donc et une mutation délétère

Ce résultat signifie soit que le prédicteur par rapport de vraisemblance est moins bon quand la taille de l'arbre augmente, soit que son calcul par recuit simulé est à perfectionner. Cette question reste ouverte pour le moment.

6.1.6 Compléments envisagés

Nous avons limité l'énoncé de nos résultats aux cas qui nous paraissaient les plus intéressants. Le temps de calcul est aussi un facteur limitant car pour obtenir les courbes ROC régulières comme ci-dessus et – espérons-le – fiables, le temps de calcul se compte en heures voire en jours. Cela dit, divers compléments sont envisageables :

- Reprendre dans le modèle à double mutation la totalité des expérimentations faites en cas de mutation simple
- Evaluer un modèle à double mutation avec prise en compte des mutations homozygotes
- Améliorer la méthode de recuit relative au prédicteur basé sur le rapport de vraisemblances pour qu'elle égale la méthode par force brute dans le cas où cette dernière est envisageable.
- Examiner la sensibilité des prédicteurs aux erreurs d'estimation des paramètres ou même au changement de modèle, c'est à dire simuler des familles selon le modèle à une mutation et les analyser avec le modèle pour deux mutations, et inversement.

6.2 Algorithmes statistiques pour l'analyse des mutations génétiques délétères

Nous avons porté l'attention dans le chapitre précédent sur les caractéristiques morphologiques des arbres généalogiques et de leurs impacts sur la performance de la prédiction du génotype connaissant le phénotype. Dans ce chapitre, nous allons revenir au cœur de la problématique en laissant de côté ces aspects morphologiques, et utiliser certains algorithmes susceptibles de permettre une prédiction mutationnelle optimisée.

A partir de familles à géométrie tout d'abord invariable, ces algorithmes seront testés dans trois cas de figure différents (modalités) déjà envisagés plus haut :

- absence de mutation délétère en dépit de la présence de cancers sporadiques (0)
- phénotype cancéreux causé par une unique mutation très pénétrante et favorisant la déclaration de la maladie à un âge précoce (1)
- phénotype cancéreux causé par l'interaction entre un gène muté peu fréquent et délétère s'il se trouve en présence d'un polymorphisme bénin beaucoup plus fréquent, lui (2).

Ils viseront à répondre à deux questions en particulier :

- est-il possible de sélectionner, entre les modalités 0, 1 ou 2, celui correspondant au génotype qui aura été paramétré, ce avec une bonne fiabilité et dans le cadre d'un échantillon de taille réaliste ?
- Peut-on obtenir des estimateurs corrects des paramètres ?

6.2.1 Méthode de validation des algorithmes

Comme précédemment, pour tester les algorithmes, il faut générer aléatoirement plusieurs centaines de jeux de données formés d'un grand nombre de familles décrites par leur arbre généalogique. Ces arbres seront de forme régulière afin de pouvoir évaluer la dépendance de nos algorithmes vis à vis des caractéristiques géométriques de ces arbres. Typiquement et à l'instar de ce qui a été réalisé dans le chapitre précédent, on fera varier deux paramètres : le nombre de générations entre 2 et 5 et le nombre d'enfants par couples entre 2 et 4. Pour chacun de ces jeux de données, les paramètres phénotypiques sont tirés au sort en fonction du génotype choisi et avec les mêmes règles quant à la pénétrance et l'âge de déclaration que ci-dessus. On effectue ensuite les itérations nécessaires pour étudier comment les algorithmes décident si les données relèvent de la modalité 0, 1 ou 2. Enfin les erreurs relatives d'estimation sont mesurées.

Deux approches mathématiques, détaillées plus bas) sont utilisées :

- Une première que nous appellerons « minimisation de distance » qui consiste à choisir le modèle et les paramètres qui s'adaptent le mieux aux données relativement à un résumé statistiques de ces données (décrit plus bas)
- Une seconde basée sur un réseau de neurones pour classifier nos jeux de données comme relevant des modèles 0, 1 ou 2.

Il faut noter que ces méthodes ont leurs avantages respectifs que nous détaillerons ultérieurement. Par ailleurs, la mise en œuvre pratique de chacune d'elles impose de nombreux choix techniques comme celui du résumé statistique, de la dimension de l'espace des paramètres, du réglage de nombreux paramètres de fonctionnement des algorithmes (caractéristiques du réseau neuronal,

algorithme de recherche du minimum, nombre de simulations pour les répliques, etc.). Les résultats présentés ci-dessous peuvent donc relever de différentes variantes car il n'est pas possible de tout recalculer à chaque nouvelle version. En particulier, certaines simulations dans la phase de validation d'algorithmes ont nécessité des temps de calcul de l'ordre plusieurs semaines sur des machines de niveau standard. Par contre, les algorithmes implémentés pour la phase décisionnelle où ils sont censés servir, ont des durées d'exécution très acceptables.

6.2.2 Présentation détaillée des modèles et des données

6.2.2.1 Généralités sur les données

Le terme « données » est un terme ambigu car il peut recouvrir à la fois des données réelles (issues de bases de données) auxquelles l'algorithme devra finalement s'appliquer et les données « simulées » de manière aussi réaliste que possible servant à paramétrer, évaluer et valider l'algorithme. De même les données réelles peuvent recouvrer diverses formes selon leur origine, et leur variété peut introduire des complications pour l'algorithme qui s'avèrent parfois très pénalisantes. En particulier, au niveau des arbres généalogiques, on peut redouter d'avoir à prendre en compte des formes et des tailles très variables.

Il s'ensuit que dans un premier temps, nous devons discuter le fonctionnement et la performance des approches sur des données « génériques » que l'on peut assez facilement simuler avec des caractéristiques totalement explicites. Toutefois, les algorithmes doivent être conçus pour s'adapter à tout type de données et, en particulier, être robustes face à une grande variété d'arbres généalogiques. Ce sont ces données génériques que l'on appelle ici données « réalistes » par opposition à d'autres jeux de données simplifiés permettant de calibrer nos modèles. Un bon exemple de données réalistes est fourni par notre générateur d'arbres POLYGENE (cf. chapitre 2).

Dans ce chapitre, les arbres simulés ont des génotypes obtenus de la façon suivante : les individus n'ayant pas de parents identifiés ont des génotypes tirés au hasard avec la fréquence de mutation qui est un des paramètres du modèle. Pour les individus qui ont des parents, le génotype est obtenu par les lois de Mendel (simplifiées) telles qu'explicitées précédemment. Une fois le génotype obtenu, on génère le phénotype avec les lois décrites au § 6.1.1. Enfin, selon le type de conditionnement (présence obligatoire ou non d'au moins x cas malades), on applique une méthode n'acceptant que les tirages qui présentent le nombre minimal requis. Ce conditionnement, rappelons-le, est nécessaire pour compenser le biais de sélection qui préside à l'établissement de toute base de données clinique dans un hôpital ou un centre de lutte contre le cancer. Les familles étant constituées à partir le plus souvent de patients venant se faire traiter pour leur maladie, on dispose donc de familles avec au minimum 1 cas de maladie, en l'occurrence le cas index. On peut même fixer cette limite basse au-delà de 1 car si les patients consultent, c'est parce qu'il y a d'autres cas de maladie dans leur famille. Dans la pratique des expérimentations numérique, nous conditionnons souvent par un nombre minimal de 1 pour les petites familles, 2 voire 3 pour les arbres de plus grande taille.

6.2.2.2 Géométrie des arbres généalogiques

Comme nous l'avons montré précédemment, un arbre symétrique (c'est à dire avec les branches maternelle et paternelle) est plus adapté à la validation méthodologique car plus informatif. Nous choisissons donc des arbres simulés de forme régulière appartenant à un catalogue fixé de sorte qu'il soit possible aussi d'évaluer certaines variations de forme (nombre de générations, nombre d'enfants par couple) sur la performance des algorithmes. Si l'on emploie la terminologie définie au paragraphe 6.1.3, nous utiliserons pour ce travail les arbres réguliers possédant entre 2 et 5 générations et entre 1 et 4 enfants par couple, donc des arbres allant de Reg(2, 1) à Reg(5, 4).

6.2.2.3 Phénotype et génotype

L'objectif est toujours de déterminer le génotype à partir du phénotype, ce dernier étant censé être le seul connu. Ce phénotype comprend le fait d'avoir développé ou non la maladie K et son âge de survenue. Cet âge est important car dans les familles porteuses de mutation délétère, il correspond à la durée d'exposition de l'individu au risque familial de maladie. Pour les cancers sporadiques, il y a peu ou prou la même correspondance, avec une loi de l'âge beaucoup plus reculée et une probabilité d'occurrence de beaucoup inférieure. Voici ci-dessous une représentation de la loi de l'âge utilisée dans ces deux cas de cancer :

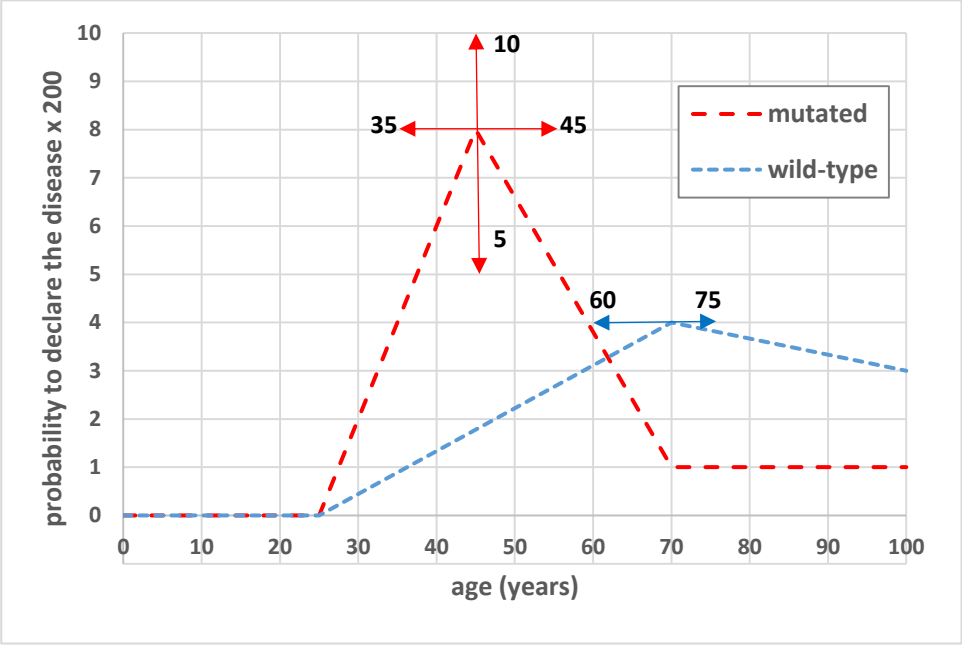


Figure 69 : loi de l'âge de déclaration de la maladie pour les individus mutés (en rouge) et pour les non mutés (en bleu) avec l'étendue de variation possible des points charnières.

Cette loi de l'âge garde une forme cumulative conforme aux distributions que l'on constate dans le cadre habituel des cancers, ainsi que le montrent les courbes suivantes inspirées des statistiques réelles tirées de la base de données oncogénétique du Centre Jean Perrin portant sur plusieurs milliers

de familles. Elles concernent la prédisposition sein/ovaire avec d'un côté les femmes porteuses d'une mutation sur les gènes BRCA et de l'autre celles qui sont exemptes de ces mutations malgré le fait qu'elles ont été retrouvées dans leur famille.

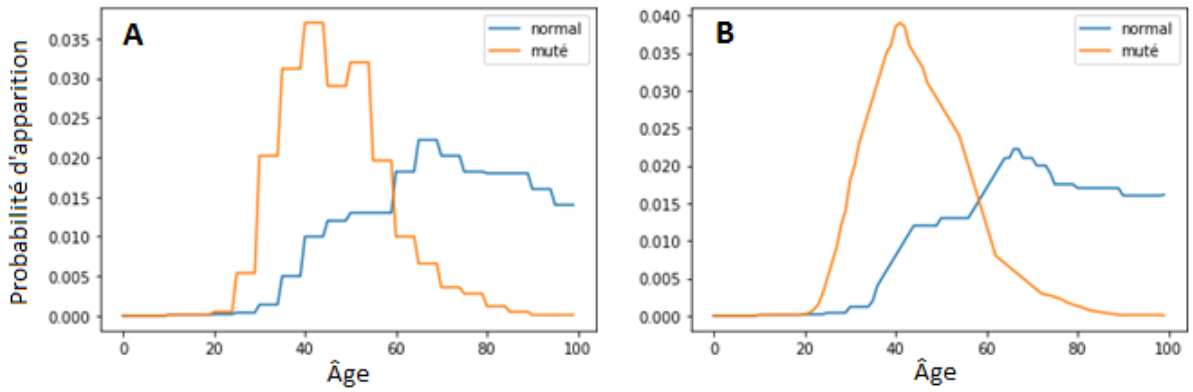


Figure 70 : loi de l'âge de déclaration du cancer calculées à partir de la BdD oncogénétique selon que les femmes sont porteuses d'une mutation BRCA (courbes en jaune) ou non (courbes en bleu).
A : courbes brutes et B : courbes débruitées et lissées

Les modèles paramétrés sont une approximation affine de deux courbes brutes, où certains points de cette ligne polygonale sont rendus mobiles. Les variations sont telles que la maladie, malgré cette correction, reste nettement plus probable à un âge précoce dans le cas muté que dans le cas non muté.

Toutefois nous avons autorisé une variabilité des points charnières. Nous avons veillé à ce que les variations de ces points charnières dans nos algorithmes permettent de maintenir entre les sujets mutés et les autres une différenciation nette de déclaration de la maladie. Cela donne ainsi deux familles de lois de probabilité cumulatives qui nous pouvons représenter comme ci-dessous :

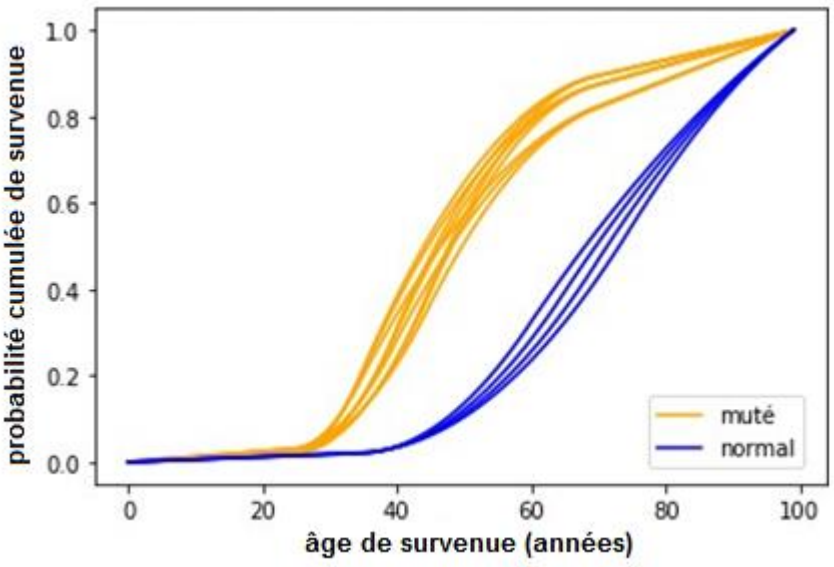


Figure 71 : familles de lois de probabilité cumulatives de l'âge de survenue du cancer selon que les individus sont mutés ou non

Les paramètres utilisés dans ce chapitre dans le cas des situations à une ou deux mutations sont les mêmes que ceux décrits aux § 6.1.1 et 6.1.5. On notera que pour simplifier l'espace des paramètres dans le modèle à 2 mutations en interaction, on peut figer les rapports de pénétrance p_1/p_0 et p_2/p_0 autour de la valeur 1, si l'on considère qu'il faut nécessairement être porteur des deux mutations pour avoir un risque plus élevé de maladie.

D'une manière générale, les connaissances épidémiologiques nous permettent au moins de décider d'intervalles de variation raisonnables pour ces paramètres. Une approche plus bayésienne serait de formuler une loi *a priori* sur ces paramètres mais nous n'avons pas suivi cette démarche compte tenu du contexte dans lequel ce travail pourrait servir. En effet, les mutations concernées qu'elles soient délétères seules ou via des interactions, n'ont pas besoin d'être connues car les données n'incluent aucun génotype d'individu. Il serait donc illogique de supposer une connaissance statistique de ces mutations. Notons aussi que les fréquences des mutations f_{mut}, f_1 ou f_2 dans la population générale ne doivent pas être mal interprétées. En effet les bases de données d'un hôpital ou d'un centre de recherche oncogénétique sont exposées au biais de sélection de contenir essentiellement des familles victimes de la maladie K. Il en résulte que pour un individu pris au hasard dans la base de données, la proportion de porteurs de mutation est sensiblement supérieure à celle que l'on trouverait dans la population générale. Cet effet sera illustré par des valeurs numériques dans la sous-section (6.2.9). Dans nos algorithmes, les familles que nous simulons utilisent donc les paramètres f_{mut}, f_1 ou f_2 valables pour la population générale mais nous imposons à ces familles de contenir un ou plusieurs cas de maladie K (selon la taille des arbres généalogiques générés).

Enfin, la maladie K est supposée, encore une fois, de type unique : on ignore en particulier les variations histologiques des cancers qui sont autant d'informations importantes dans la qualification habituelle du risque familial en consultation oncogénétique. Toutefois deux localisations différentes seront testées au final, d'abord la susceptibilité sein/ovaire et le syndrome de Lynch (côlon) ceci afin d'évaluer les approches sur des risques concernant principalement un sexe ou bien les deux.

Quant à lui, le génotype est caractérisé par des mutations hétérozygotes, comme précédemment, et la transmission de ces mutations entre parents et enfants suit la loi de Mendel avec les probabilités décrites au Tableau 18.

6.2.3 Principe de fonctionnement de nos algorithmes

Deux approches mathématiques sont utilisées, une première par minimisation de la distance entre les données réalistes et d'autres répliquées par nos algorithmes, puis une seconde basée sur un réseau de neurones.

6.2.3.1 Approche par minimisation de distance

Pour tout jeu de données, l'algorithme va apporter trois informations :

(1) Une estimation des paramètres les plus pertinents dans les modèles 0, 1 ou 2 quant à leur aptitude à représenter les données
(2) Le calcul de la distance du jeu de données réalistes à ces jeux générés par algorithme selon un résumé statistique
(3) La sélection du modèle représentant le mieux le jeu initial de données comme étant celui de distance minimale avec estimation des paramètres correspondants.

Le point fondamental est que l'algorithme fonde sa décision exclusivement sur l'examen de plusieurs statistiques mesurées sur le jeu de données. Nous appelons ces statistiques « le résumé statistique des données » ; elles ont une interprétation relativement simple qui sera détaillée plus loin. Elles peuvent être calculées facilement sur tout jeu de données réalistes voire réelles, ce directement par le système de gestion de base de données. Dans le cas très particulier où tous les arbres généalogiques seraient identiques, l'importation des données pourrait ne consister en tout et pour tout qu'en l'entrée des valeurs de ces statistiques dans l'algorithme décisionnel qui nécessite, lui, d'être codé en langage suffisamment évolué (Python en ce qui nous concerne). Dans le cas de données réelles où la variabilité des arbres généalogiques est importante, il faut a priori importer les données correspondantes, codées sous une forme adéquate, puis calculer les résumés statistiques.

Notons $data_N$, les données correspondant à N familles et $S(data_N)$ le résumé statistique de ces données qui prend ses valeurs dans un espace de dimension d, quel que soit le modèle. Indexons par $m \in \{0, 1, 2\}$ les trois modalités que nous souhaitons considérer, par τ_m le domaine de variation de leurs paramètres inclus dans R^2 pour m = 0, R^3 pour m = 1 et R^6 pour m = 2 et par P_θ^m la loi de probabilité induite par la modalité m pour une valeur de $\theta \in \tau_m$ de ses paramètres.

En s'inspirant de la classique méthode des moments, pour chaque modalité m, on pourrait souhaiter estimer les paramètres par l'estimateur $\hat{\theta}$ tel que :

$$S(data_N) = E_{\hat{\theta}_0}^{m_0}[S]$$

Toutefois, si $S(data_N)$ tend vers $E_{\hat{\theta}}^{m}[S]$ pour les vraies valeurs m_0 et θ_0 quand N devient grand, rien n'assure qu'à un rang fini N, cette équation admette une solution ni qu'elle soit unique. Nous introduisons donc pour $m \in \{0, 1, 2\}$, le meilleur jeu de paramètres représentant les données dans la modalité m comme étant la valeur $\hat{\theta}_m$ de $\theta \in \tau_m$ qui minimise :

$$dist(S(data_N), E_\theta^m[S])$$

où dist(., .) est une distance à préciser sur R^d dans lequel le résumé statistique prend ses valeurs. Dans cette version, l'existence du minimum est garantie dès que $E_\theta^m[S]$ est continu par rapport au paramètre θ variant dans un domaine fermé et borné τ_m mais l'unicité n'est pas garantie *a priori*.

Malheureusement, il n'y a pas d'espoir de disposer de formules explicites pour $E_\theta^m[S]$ qui ne pourra, en réalité, n'être approché que par simulation en calculant $S(gendata_{N'})$ où $gendata_{N'}$ sont les données issues de N' arbres généalogiques de la loi P_θ^m. Nous allons donc considérer pour tout m la valeur $\hat{\theta}_m$ qui minimise

$$dist\left(S(data_N),\ S(gendata_{N'})\right)$$

où $gendata_{N'}$ provient d'un jeu de N' arbres générées selon la loi P_θ^m. Cette valeur n'est donc pas univoque car elle dépend des tirages aléatoires $S(gendata_{N'})$ que l'on obtient sous les différentes lois $P_\theta^m, \theta \in \tau_m$. Toutefois, N' étant notre choix, nous contrôlons la proximité entre $S(gendata_N)$ et $E_\theta^m[S]$.

Ainsi la recherche de minimum s'effectue selon un certain algorithme numérique que nous préciserons et pour lequel la fonction objectif est bruitée aléatoirement.

La dernière étape est la sélection de la meilleure modalité qui consiste simplement à choisir la modalité présentant la distance minimale aux données dans le sens ci-dessus. L'algorithme comporte donc deux choix principaux :

- Celui du résumé statistique
- Celui de l'algorithme de minimisation

Le résumé statistique doit prendre ses valeurs dans un espace de dimension *d* suffisamment grande pour espérer que l'application $\theta \to E_\theta^m[S]$ soit injective. Typiquement, nous prenons *d* strictement supérieur au nombre de paramètres maximal.

L'algorithme de minimisation que nous devons mettre en place a une fonction objectif aléatoirement bruitée ce qui exclut les algorithmes classiques des cas différentiables. Nous procédons en deux temps :

- La recherche systématique sur un réseau pour déterminer les zones de faible valeur
- Une recherche plus précise par marche aléatoire sur les zones trouvées précédemment

Dans le premier temps de cette méthode, on peut se contenter de calculer $S(gendata_{N'})$ pour une petite valeur de *N'* car peu de précision est requise. Cela est heureux car dans cette étape le nombre de points augmente rapidement en fonction du nombre de mailles *k* sur chaque paramètre, par exemple par k^6 pour la modalité 2 qui peut avoir jusqu'à 6 paramètres (voire 3 paramètres en plus sur les lois). La seconde étape affine le résultat en partant des meilleurs jeux de paramètres trouvés dans la première étape par une marche au hasard autour de cette position. Compte tenu du bruit affectant la fonction objectif notamment dans la première phase, l'algorithme se doit d'explorer plusieurs des meilleures valeurs obtenues.

Il y a donc un certain nombre de paramètres à calibrer dans cet algorithme de recherche de minimum :

- Le nombre de simulations pour la première phase qui ne doit pas être trop grand
- Le nombre de points calculés dans l'espace des paramètres, dans la première phase
- Le nombre de points sélectionnés à l'issue de cette première étape
- Le nombre de simulation *N'* dans la seconde phase
- Le nombre de pas de la marche aléatoire que l'on fait ensuite autour de chaque point sélectionné
- La taille de ces pas qui évolue progressivement à la baisse
- Le résultat final adopté : réel minimum ou moyenne des *k* meilleurs avec *k* à choisir

On peut identifier différentes sources d'erreur qui se présentent dans la démarche décrite précédemment. D'abord $S(data_N)$ est une approximation de $E_{\theta_0}^{m_0}[S]$ où m_0 *et* θ_0 sont les vraies valeurs, à l'ordre $1 / \sqrt{N}$ pour une valeur de N qui est dictée par les données disponibles. Le tableau qui suit confirme ce fait : il donne pour $m_0 \in \{0, 1, 2\}$ les valeurs moyennes de $dist(S_N, S'_N)$ où S_N et S'_N sont deux jeux de données indépendants selon la loi $P_{\theta_0}^{m_0}$ avec θ_0 tiré uniformément dans τ_{m_0} et $N \in \{100, 1\,000, 10\,000\}$. Les arbres généalogiques comportent 4 générations et 3 enfants par couple :

Tableau 19 : influence du nombre d'arbres sur la distance moyenne entre données simulées et données issues des algorithmes.

N	100	1000	10000
$m_0 = 0$	0.43	0.12	0.035
$m_0 = 1$	0.30	0.090	0.027
$m_0 = 2$	0.32	0.095	0.028

Notons qu'une fois $S(data_N)$ calculée, la valeur de *N* n'intervient plus dans la vitesse d'exécution de l'algorithme qui est ainsi meilleur mais pas plus lent si le nombre de données augmente.

La recherche de minimum se base dans sa seconde étape sur une simulation à *N'* pas. Compte tenu de la remarque précédente, il est raisonnable que *N'* reste au plus dans l'ordre de *N*. L'idée de moyenner au niveau du résultat final est de s'affranchir du bruit de simulation.

La valeur des paramètres réalisant l'infimum de distance parmi tous les points où le calcul de distance a été fait sera notre estimateur des paramètres. Le problème d'existence du minimum de distance que nous évoquions précédemment est purement théorique puisque la machine arrivera à coup sûr à trouver le minimum parmi toutes les valeurs des paramètres qu'elle a essayé. Par contre, le problème de l'unicité et même de reproductibilité est réel. L'injectivité de la fonction $\theta \rightarrow E_\theta^m[S]$ est inconnue et, compte tenu du caractère aléatoire de la fonction $dist\left(S(data_N),\ S(gendata_{N'})\right)$ minimisée, le résultat obtenu est intrinsèquement aléatoire au-delà même l'incertitude numérique.

6.2.3.2 Méthode par réseau de neurones

La fonction $E_\theta^m[S] \rightarrow \theta$ n'est pas plus calculable explicitement que la fonction $\theta \rightarrow E_\theta^m[S]$ mais il est par contre tout à fait envisageable de la faire apprendre à un réseau de neurones. En fait, il s'agira d'entraîner le réseau de neurones sur la fonction $S(gendata_{N'}) \rightarrow \theta$. Dans cette phase d'apprentissage, la valeur de *N'* n'est limitée que par le temps de calcul et la fonction $S(gendata_{N'})$ peut approximer $E_\theta^m[S]$ d'aussi près qu'on le souhaite. C'est à dire que, sur la structure généalogique dont on dispose, qu'elle soit simplifiée dans la phase théorique ou réelle dans la phase opérationnelle, on réplique les simulations un grand nombre de fois pour éliminer au maximum le bruit de simulation sur les statistiques $E_\theta^m[S]$.

De plus, comme on dispose d'un modèle pour la génération des données « réalistes », les données d'apprentissage sont disponibles en quantité illimitée. Il en résultera probablement une performance en termes d'estimation des paramètres une performance qui devrait dépasser la méthode par minimisation de distance.

Détaillons un peu les réseaux que nous utilisons. L'architecture est simple puisqu'elle consiste en quelques couches séquentielles :

- Pour la modalité à 0 mutations, on utilise 2 couches cachées à 128 et 32 neurones

- Pour la modalité à 1 mutation, on utilise 4 couches à respectivement 64, 32, 16 et 8 neurones
- Pour la modalité à 2 mutations, on utilise 4 couches à 128, 64, 64 et 32 neurones.

Dans tous les cas, on utilise la fonction d'activation relu et l'algorithme d'optimisation Adam. Les régresseurs cherchent à minimiser l'erreur relative (*mean absolute percentage error*), et le classificateur utilise l'entropie croisée (*categorical cross-entropy*) pour la fonction objectif.

Etant donné le temps et les ressources de calcul, nous n'avons pas pu effectuer des tests exhaustifs quant aux différentes architectures. Cependant, dans notre expérimentation, on a observé que pour le modèle à 0 mutation, un réseau simple à 2 couches cachées fournit de meilleurs résultats qu'un réseau plus profond. Ce phénomène n'a pas été observé pour les réseaux à 1 ou 2 mutations.

Les réseaux ont été entraînés sur un jeu de données de 25 000 observations simulées selon chacun des 3 modalités. Chaque simulation a été répliquée 500 fois afin d'obtenir des statistiques plus fiables.

6.2.4 Performances en sélection de modèle par minimisation de distance

Nous pouvons construire la matrice de confusion présentant la répartition des classements obtenus pour 100 jeux de données générés parmi les 3 modalités. Ces jeux comportaient chacun 1000 familles, donc 1000 arbres généalogiques de même forme comprenant des individus répartis sur 3 générations avec un nombre fixé de 2 enfants par couple, soit au total 10 000 individus par jeu.

Tableau 20 : adéquation entre la modalité des jeux et le résultat donné par l'algorithme

↓ Modalité des jeux / classement →	**0**	**1**	**2**
0 - pas de mutation	100	0	0
1 - une mutation délétère	0	92	8
2 - deux mutations interagissant	0	4	96

Ce taux de bon classement montre l'efficacité de l'algorithme pour un nombre de données relativement faible dans un contexte de numérisation systématique des données médicales. En particulier, la modalité 0 (sans mutation) est très facilement discriminée. Si l'on fait jouer les paramètres des arbres, la reconnaissance de la modalité est optimale quand on a quatre générations et quatre enfants par couple :

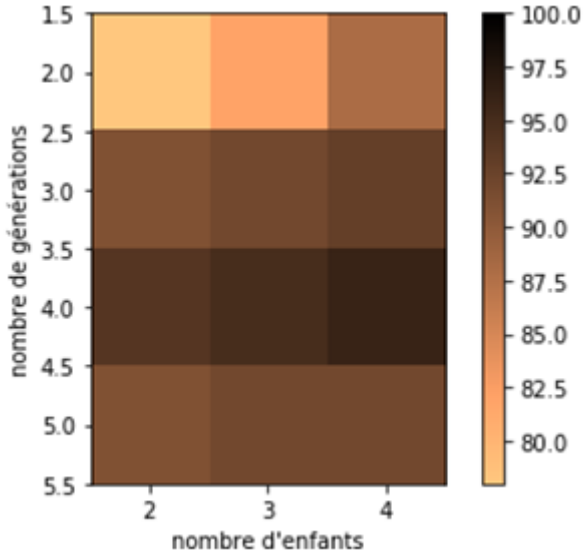


Figure 72 : taux de bon classement des arbres selon le nombre de générations et le nombre d'enfants (contexte : une mutation)

En dehors de tailles d'arbre réduites, la capacité de classement est excellente avec des valeurs typiques dépassant 85%. A l'instar de ce qui avait été remarqué dans le chapitre précédent, on note qu'une très grande taille de famille non seulement n'améliore les performances, mais tend à les dégrader. Une justification possible est que chaque individu supplémentaire apporte non seulement de l'information mais aussi de l'aléa et qu'une fois le modèle saturé, c'est cet aléa qui pénalise.

Il en est de même pour distinguer les jeux de données simulés selon la 2ème modalité que l'algorithme reconnaît très fréquemment :

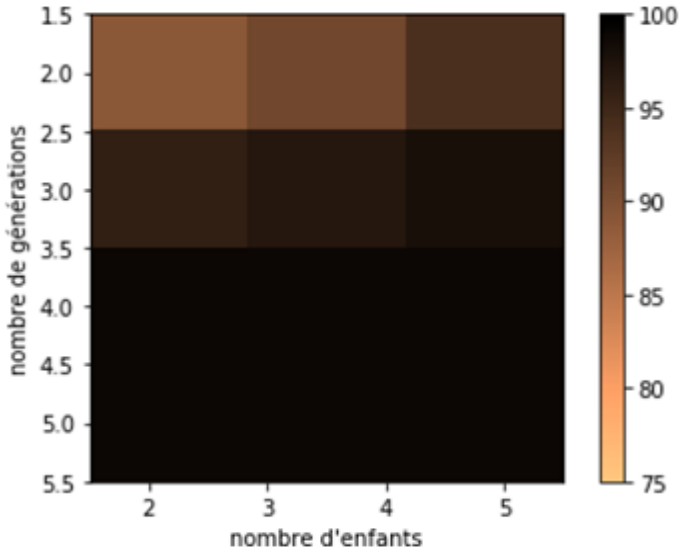


Figure 73 : taux de bon classement des arbres selon le nombre de générations et le nombre d'enfants (contexte : 2 mutations interagissant)

La saturation du modèle s'obtient dès la quatrième génération, quel que soit le nombre d'enfants (à partir de 2) : le taux de bien classé flirte alors avec le 100%

6.2.5 Performance de l'estimation paramétrique par minimisation de distance

Simultanément à la détermination de la modalité qui se révèle comme on l'a vu ci-dessus très efficace, l'algorithme procède à l'estimation des paramètres. Nous rappelons que les tests ont été conduits en tirant au hasard ces paramètres et uniformément dans des intervalles assez larges. Cela permet de calculer l'erreur relative moyenne sur tous les jeux de données où l'algorithme a choisi la bonne modalité.

Dans le cas de la modalité à une mutation, il y a 3 paramètres à estimer et l'erreur relative est moyennée sur ces 3 paramètres :

Tableau 21 : Erreur moyenne sur les 3 paramètres dans le cas de la distinction des modalités 0 et 1

Nombre d'enfants → ↓ nombre de générations	**2 enfants**	**3 enfants**	**4 enfants**
2 générations	40 %	28 %	20 %
3 générations	18 %	17 %	15 %
4 générations	11 %	11 %	11 %

Dans le cas de la modalité à 2 mutations interagissant, il y a 6 paramètres (3 par mutation) et les erreurs moyennes sont données par le tableau suivant :

Tableau 22 : Erreur moyenne sur les 6 paramètres en cas de modalité à 2 mutations interagissant

Nombre d'enfants → ↓ nombre de générations	**2 enfants**	**3 enfants**	**4 enfants**
2 générations	18 %	19 %	19 %
3 générations	15 %	14 %	14 %
4 générations	13 %	13 %	13 %

On voit que si on exclut les arbres à 2 générations seulement, on obtient des estimations des paramètres qui semblent tout à fait appréciables dans le contexte oncogénétique applicatif.

6.2.6 Estimation par réseau de neurone

L'apprentissage se fait à l'aide des mêmes types d'arbre, de format fixe comportant 3 générations et 2 enfants par couple et en conditionnant à au moins un malade par famille. Le tableau ci-dessous résume les résultats obtenus sur un échantillon test de valeurs du résumé statistique. Celui-ci est calculé comme précédemment pour la méthode par minimisation.

Tableau 23 : Erreur d'estimation moyenne des paramètres par le réseau de neurones

modalité →	**0**	**1**	**2**
Erreur d'estimation (%)	19.3 %	5.6 %	11.6 %

Pour la modalité sans mutation (0) qui est la moins bien estimée, d'autres techniques d'estimation seraient sûrement préférables, comme par exemple un estimateur à noyau pour la loi de l'âge de

déclaration. Le réseau de neurones s'applique aussi en classification (trouver la modalité). Nous allons détailler cela ci-dessous dans le cas moins académique d'une généalogie réelle.

6.2.7 Résultats sur une base de données oncogénétique

Nous avons fait fonctionner les algorithmes décrits précédemment sur un ensemble de 395 arbres généalogiques (anonymisés) provenant du service oncogénétique du Centre Jean Perrin. En tout, cela représente 11 700 individus appartenant à des familles de syndrome sein/ovaire : il s'agit là d'un exemple typique de situation à laquelle il était intéressant d'appliquer notre méthodologie. Cette étape nous permet aussi un étalonnage de la confiance à accorder à tout nouveau résultat.

L'étape suivante est de tester le calibrage des modèles avec une nouvelle population où seuls les phénotypes sont connus dans un but désormais de classification.

La version que nous testons dans ce qui suit est la version « spéciale syndrome sein/ovaire » que nous avons paramétrée pour s'adapter à ce cancer touchant presque exclusivement les femmes. Pour la structure héréditaire, les deux sexes sont à prendre en compte mais seuls les individus féminins (ou presque) apportent de l'information par leur phénotype. Comme l'information utilisable ne provient que de la moitié des membres des familles, la performance est notablement altérée par rapport à une maladie qui concernerait les deux sexes.

6.2.7.1 Classification par minimisation de distance

Nous avons lancé les algorithmes de classification sur 3 130 jeux de données, mais avec une généalogie réelle, celle issue de chacune des familles. Après un temps de calcul conséquent, nous obtenons la matrice de confusion suivante :

Tableau 24 : taux de concordance de la classification par minimisation sur un jeu de données correspondant à une généalogie réaliste

Modalité calculée → ↓ Modalité réelle	**0**	**1**	**2**
0	99 %	1 %	0 %
1	0 %	87 %	13 %
2	0 %	33 %	67 %

Le taux de concordance est moins bon qu'avec la généalogie simplifiée, alors que l'algorithme a bénéficié de nettes améliorations. Mais le caractère sexué du risque de la maladie est pénalisant. Peut-être serait-il possible de faire mieux en optimisant le réglage des paramètres de fonctionnement. Nous n'avons pas encore disposé du temps machine et du temps humain pour le faire.

6.2.7.2 Classification avec le réseau de neurones

Nous donnerons les détails des réseaux construits dans un paragraphe ultérieur. Nous avons voulu voir les performances sur le jeu de données précédent. Voici la matrice de confusion :

Tableau 25 : taux de concordance de la classification par réseau de neurones

Modalité calculée → **↓ Modalité réelle**	**0**	**1**	**2**
0	96.9 %	3.12 %	0.0 %
1	0.02 %	92.16 %	7.8 %
2	0 %	25 %	75 %

Si l'on utilise des données 10 fois plus volumineuses, on améliore la précision :

Tableau 26 : taux de concordance de la classification par réseau de neurones avec une population 10 fois plus importante

Modalité calculée → **↓ Modalité réelle**	**0**	**1**	**2**
0	99.98 %	0.02 %	0.0 %
1	0.0 %	97.76 %	2.24 %
2	0.0 %	8.54 %	91.46 %

Enfin si on utilise une population 500 fois plus importante que précédemment, on observe une convergence de la méthode :

Tableau 27 : taux de concordance de la classification par réseau de neurones avec une population 50 fois plus importante

Modalité calculée → **↓ Modalité réelle**	**0**	**1**	**2**
0	100 %	0.0 %	0.0 %
1	0.0 %	100 %	0 %
2	0.0 %	0.14 %	99.86 %

6.2.8 Discussion sur les choix techniques

6.2.8.1 Le résumé statistique

Pour les résultats que nous avons donnés ci-dessus, le résumé statistique S d'un arbre généalogique est formé de huit quantités dont voici la définition :

- S_1 : la fréquence de K sur l'ensemble des individus dont les parents ne sont pas renseignés (non inclus dans l'arbre)
- S_2 : la fréquence de K pour les individus dont les parents sont renseignés (donc inclus dans l'arbre)
- S_3 : l'âge moyen de déclaration de K pour les individus non dans l'arbre
- S_4 : l'âge moyen de déclaration de K pour les individus de l'arbre
- S_5 : la proportion de malades de K dont les deux parents sont renseignés mais indemnes de K
- S_6 : le nombre moyen de parents malades de K chez les individus malades
- S_7 : le score moyen de proximité phénotypique entre un individu et ses ascendants (parents ou grands-parents)

- S_8 : le score moyen de proximité phénotypique entre un individu et ses éventuels frères et sœurs

Dans le cas d'un jeu de données où les familles ont toutes un seul enfant par couple, la statistique S_8 n'est pas définie et peut donc être omise. Les statistiques S_7 et S_8 sont les plus évoluées : elles consistent pour chaque individu à calculer un score de ressemblance entre son phénotype et ceux de ses parents, grands-parents, frères et sœurs en tenant compte d'une pondération selon le degré de parenté. Pour ce calcul, on se sert d'un barème algébrique (positif ou négatif) quantifiant la proximité de deux phénotypes. Pour donner un exemple concret, voici ce que nous avons utilisé dans la dernière version :

- Pas de K chez aucun des deux individus comparés $\rightarrow$ +1
- K chez l'un mais pas chez l'autre $\rightarrow$ -3
- K précoce chez les deux $\rightarrow$ +10
- K non doublement précoce mais à des âges voisins $\rightarrow$ +5
- Double K ne rentrant pas dans les deux cas précédents $\rightarrow$ +3

Les scores sont calculés pour tous les individus, puis moyennés par famille. Une dernière moyenne est calculée sur l'ensemble des familles.

Ces méthodes ont un coût qui est indépendant de la dimension d du résumé statistique. On peut donc sans problème enrichir ce résumé notamment en ajoutant la loi complète de l'âge à la déclaration de K séquencée par tranche d'âge. Ce choix est en particulier fait quand on souhaite estimer en plus des paramètres standards ceux régissant la loi d'âge de déclaration de K.

6.2.8.2 Choix de la distance

La distance "dist" apparaissant dans nos formules est une mesure de ressemblance sur R^d, l'espace dans lequel notre résumé statistique prend ses valeurs. On ne demande pas que cela soit une distance au sens topologique mathématique. Elle doit être choisie de façon à s'affranchir des unités et intervalles de variation des statistiques S_j. Par exemple, certaines statistiques sont des fréquences assez petites tandis que d'autres comme les âges se comptent en année. Pour le calcul numérique effectif nous avons utilisé :

$$dist\left[\left(x_j, 1 \leq j \leq d\right), \left(y_j, 1 \leq j \leq d\right)\right] = \sqrt{\sum_{j=1}^{d} \left[1 - \max\left(\frac{x_j}{y_j}, \frac{y_j}{x_j}\right)\right]^2}$$

Dans cette formule nous avons préféré les carrés aux valeurs absolues compte tenu de la tradition statistique pour des données non-éparses. Des coefficients peuvent être rajoutés pour pondérer l'influence de diverses statistiques mais leur ajustement rajoute un niveau supplémentaire de réglage dont nous avons fait l'économie.

6.2.9 Etude de la variabilité

6.2.9.1 Influence de la géométrie de l'arbre et du conditionnement

Il est crucial de savoir si le résumé statistique choisi est invariant ou peu variable selon la taille et la forme de l'arbre généalogique ou si *a contrario*, il faudra tenir compte de la répartition des formes d'arbre lors de l'utilisation de données réelles. Les variables du résumé statistique ont été choisies

parce que globales par arbre. Elles peuvent en outre être calculées sur des parties d'arbre de manière à exprimer des particularismes locaux, voire ce qui se passe au niveau d'un individu et de ses parents. Comme on l'a vu dans la partie 3, de tels sous-arbres sont les mêmes quels que soient les arbres complets dont ils sont issus et cela laisse espérer des statistiques assez invariantes par rapport à la géométrie des arbres. Néanmoins, la réalité numérique est différente comme le montrent les expérimentations qui vont suivre. Pour celles-ci, nous travaillons sur la classe d'arbres réguliers Reg(x, y) telle que définie précédemment, avec x représentant le nombre de générations et y le nombre d'enfants par couple. Comme on le voit dans le tableau suivant, le nombre de membres dans une famille est strictement croissant avec le nombre de générations, puis à l'intérieur des générations par nombre d'enfants fixé par couple. Ce nombre d'individus servira comme axe des abscisses dans les graphes ci-après.

Tableau 28 : nombre d'individus par famille selon le paramétrage du nombre de générations et d'enfants par couple (les valeurs inférieures à 2 ont été ignorées)

Générations → ↓ Enfants/couple	**2**	**3**	**4**	**5**
2	4	10	22	46
3	5	13	29	61
4	6	16	36	76
5	7	19	43	91

L'autre élément à prendre en compte dans la comparaison de ces statistiques est le conditionnement des arbres à contenir un nombre minimal de cas malades. Nous avons vu que ce conditionnement est nécessaire pour rendre compte du biais de sélection à l'œuvre dans la constitution des bases de données réelles, lesquelles, en oncogénétique, comprennent systématiquement des familles ayant des cancers. Les graphiques suivants représentent l'évolution des huit statistiques qui forment le résumé statistique en fonction de la taille de l'arbre et du conditionnement. Les trois courbes qui figurent par graphe se différencient par le conditionnement du nombre minimal de malades dans chaque arbre : pour la courbe jaune-orange, il n'y a pas de conditionnement, tandis que pour les courbes bleu clair et violette on a conditionné à 1 et 2 malades minimum respectivement.

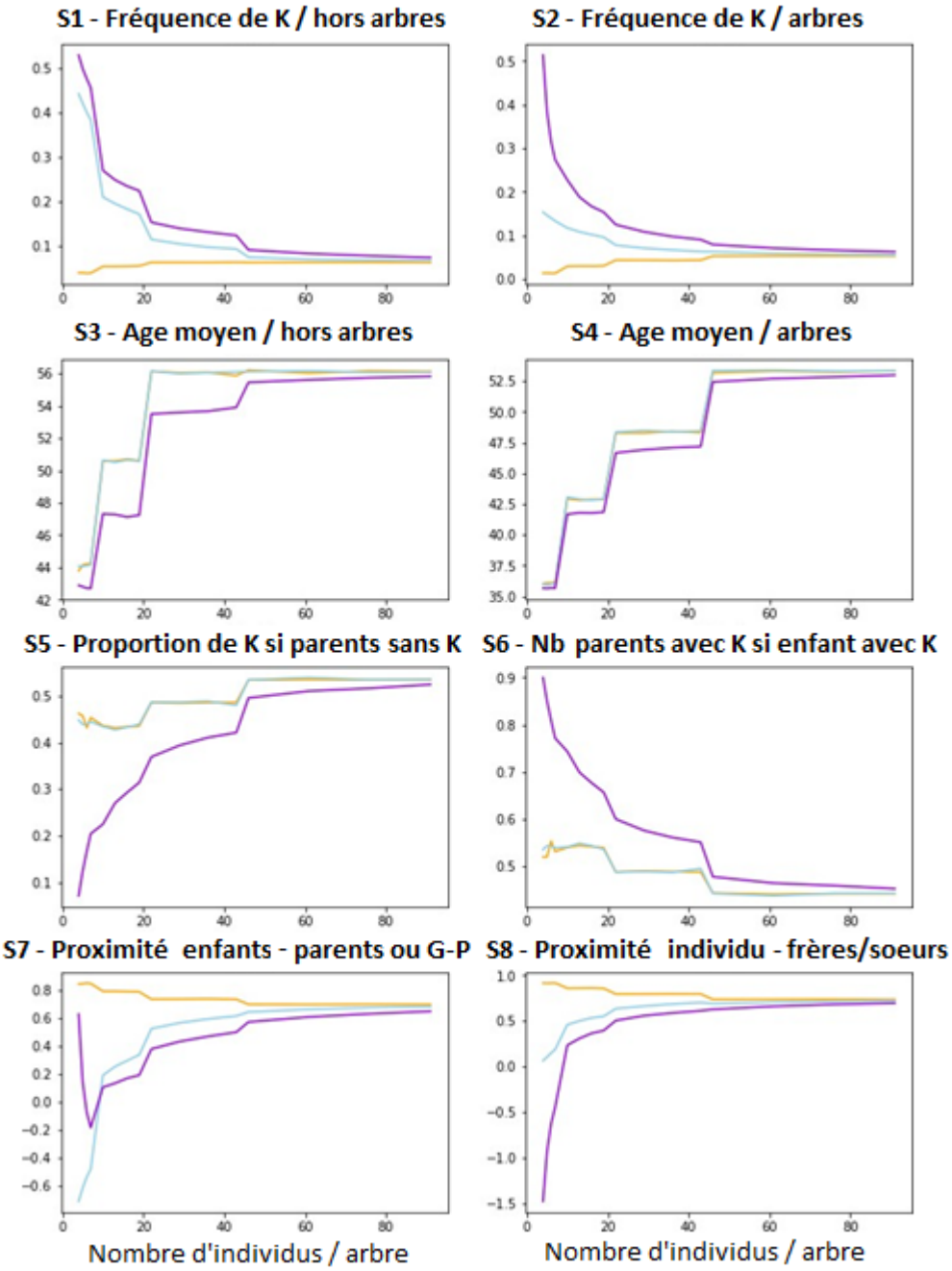


Figure 74 : valeurs des paramètres du résumé statistique en fonction du nombre de générations par arbre, du nombre d'enfants par couple et du conditionnement

On constate de fortes variation quant aux paramètres formant le résumé statistique vis à vis de la taille de l'arbre et du conditionnement. Certains sauts entre paliers apparaissant au niveau des courbes correspondent à des changements de nombre de générations prises en compte. Pour d'autres statistiques, l'effet du nombre de générations n'est pas direct et les variations sont plus lisses. Notons que pour les familles représentatives des arbres testés, les individus ont tous deux parents renseignés, sauf les ancêtres en début d'arbre, et donc, plus il y a de générations et plus ces ancêtres sont âgés, ce qui peut expliquer plusieurs tendances variationnelles.

Au vu de ces graphiques, on peut conclure que la géométrie des arbres généalogiques et le conditionnement adopté pour corriger le biais de sélection sont des facteurs primordiaux qui doivent être pris en compte complètement dans l'algorithme décisionnel et d'estimation.

6.2.9.2 Variabilité des arbres et conséquences

En raison de l'influence de la forme des arbres observée ci-dessus, on peut logiquement penser que sur les données réelles, la variance des statistiques pourrait être plus importante que celle obtenue sur données issues d'arbres fixes. Autrement dit, les statistiques pourraient être plus bruitées et la recherche des paramètres conduisant à ces statistiques réclameraient des échantillons plus grands à performance constante. Pour objectiver visuellement la variabilité des statistiques sur généalogies réelles ou simplifiées, nous avons produit le graphe suivant qui consiste à générer 100 fois des statistiques sur une généalogie fixée. Pour les points rouges, la généalogie est réelle (donc à géométrie variable) et pour les points bleus, la généalogie est simplifiée au sens ou 5 formes d'arbre sont possibles. Le nombre d'individus est le même dans les deux cas soit autour de 10 000. Les autres paramètres sont naturellement identiques.

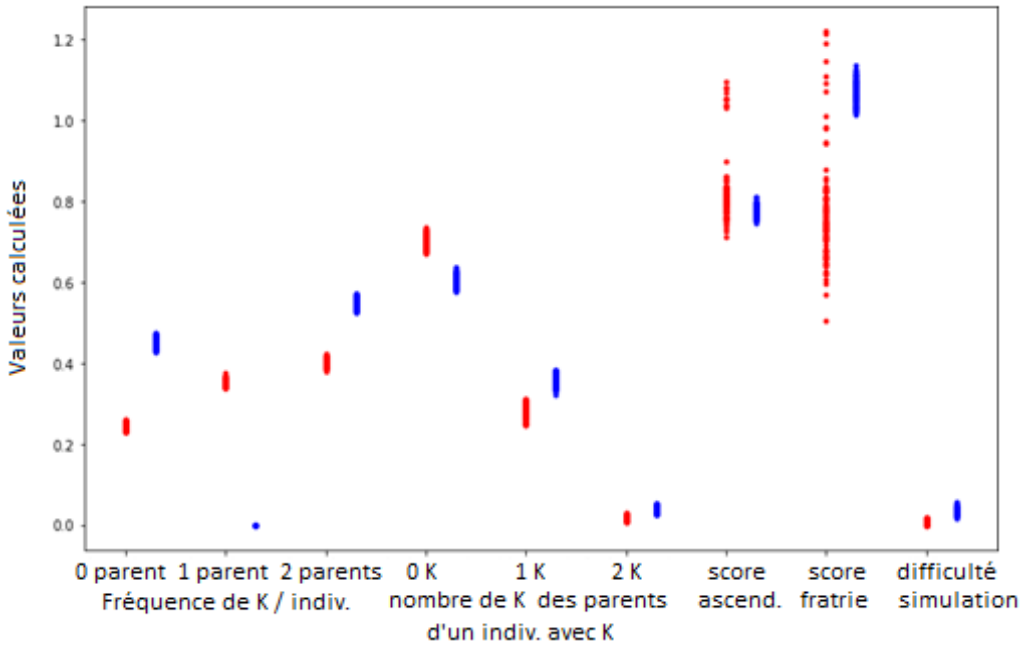


Figure 75 : variabilité des paramètres selon qu'on travaille sur généalogie réelle (rouge) ou préfixée (bleu)

Les trois premières colonnes se lisent : « fréquence de la maladie K chez un individu pour lequel 0, 1 ou 2 parents sont renseignés. On observe que ce sont les statistiques de score de proximité qui montrent la plus grande variabilité entre les généalogies réelles et simplifiées. Pour les autres, hormis le 4ème point qui a une valeur nulle dans ce contexte, il n'y a pas d'effet notable dépendant de la structure des arbres.

6.2.9.3 Quantification du biais de sélection

Nous avons expliqué au § 6.2.2.1, la nécessité du conditionnement pour l'adéquation entre les données générées et les données réelles telles que recueillies dans les bases de données oncogénétiques. Ce faisant, ce conditionnement change radicalement les statistiques de mutation et

de pénétrance pour les individus présents dans l'arbre. Nous disposons d'un programme qui quantifie l'effet du conditionnement en fonction des autres paramètres du modèle et des caractéristiques des arbres. En particulier nous avons trouvé que cet effet variait en fonction des générations incluses dans l'arbre.

Prenons l'exemple d'une mutation simple, de fréquence 5%, une pénétrance de 50% pour les individus mutés et une fréquence de K chez les autres individus de 5%. Utilisons des arbres Reg(4, 2) avec 4 générations et 2 enfants par couple (soit 22 individus par arbre). Nous donnons ci-dessous, en fonction du conditionnement, les facteurs multiplicateurs de la fréquence de mutation, de la fréquence de la maladie chez les mutés et les non-mutés, par rapport aux même paramètres issus de la population générale en prenant pour base un individu appartenant à la génération la plus récente :

Tableau 29 : influence multiplicative du conditionnement sur certains paramètres résultant des tests itératifs

Nombre minimum de malade(s) imposé(s)	**0**	**1**	**2**	**3**
Multiplication de la fréquence de mutation	0.96	1.45	2.56	4.25
Multiplication de la fréquence de K chez les non-mutés	1	1.58	2.25	2.66
Multiplication de la pénétrance (chez les mutés)	1	1.05	1.12	1.31

Il apparaît très clairement que ne considérer que les familles avec un nombre minimal de cas malades induit un biais significatif dans l'estimation des paramètres. Le processus de collecte des données oncogénéalogiques, de par sa recherche la plus exhaustive possible des antécédents familiaux, a donc un impact sur les statistiques populationnelles qui sont déduite des bases de données.

Inversement, ce tableau fournit un outil permettant d'estimer les caractéristiques de la population générale dès lors qu'on connaît les contraintes imposées lors du recueil de la base de données.

6.2.10 Mise en œuvre sur données réelles

6.2.10.1 Données à importer

Compte tenu des constatations faites ci-dessus, la mise en œuvre des calculs sur données réelles requiert l'importation des éléments suivants :

- La forme de l'arbre généalogique de chaque famille
- Le nombre de malades pour chaque famille
- Le résumé statistique calculé sur l'ensemble des familles

La simulation des répliques se fait alors en respectant les contraintes de forme, de nombre de malades et compare les résumés statistiques obtenus à celui des données réelles, dans le but de choisir la modalité et l'estimation des paramètres.

La méthode d'estimation par réseau de neurones doit elle aussi fonctionner sur un échantillon d'apprentissage qui respecte ces contraintes, avec une taille d'échantillon aussi importante que nécessaire. En outre, dans cette phase de production des données d'apprentissage, le bruit de simulation du résumé statistique peut être quasiment éliminé en multipliant les

simulations tout en gardant la structure généalogique. Cela vaut également pour la méthode par minimisation de distance au niveau du calcul des statistiques pour les données répliquées.

6.2.11 Amélioration de la performance

Nous avons déjà remarqué que la phase de validation de nos algorithmes, que ce soit par minimisation de distance ou par réseau de neurones, requiert de travailler sur un grand nombre de jeux de données, qui atteint souvent les limites des capacités des machines ou, à tout le moins, engendre des temps de calcul. Pour l'algorithme par minimisation de distance toutefois, la phase décisionnelle sur données réelles est automatiquement moins exigeante en temps de calcul puisqu'une seule phase de calcul est nécessaire. Cela permet de pousser plus loin la recherche des paramètres de l'algorithme ce qui laisse espérer une performance encore supérieure à celle obtenue durant la phase de validation méthodologique. Pour le réseau de neurones par contre, la phase décisionnelle est presque aussi longue que la validation car la génération des données pour l'apprentissage par le réseau doit être menée spécifiquement en respectant les contraintes généalogiques des données.

Une autre amélioration possible dans l'algorithme par minimisation de distance est d'introduire une nouvelle statistique qui quantifie la difficulté à simuler selon le nombre de malades prescrit. En effet, pour certains jeux de paramètres, simuler les répliques qui satisfont pour chaque famille la contrainte du nombre de malades de l'arbre réel, peut être difficile voire impossible. Comme le choix de la modalité fonctionne par rejet de celles les plus distantes, cela peut même conduire au blocage de l'algorithme. Nous avons donc dû implémenter un dispositif qui permet à l'algorithme de refonctionner en « relaxant » progressivement la contrainte sur le nombre de malades et en comptabilisant ces modifications dans une nouvelle statistique. Cela fournit une information supplémentaire que l'on convertit en donnée quantitative que nous ajoutons au résumé statistique. Quand cette statistique est élevée, cela signifie qu'il a été difficile de respecter les contraintes issues des données réelles et donc que la modalité et/ou le jeu de paramètres envisagés est mauvais. Cette nouvelle statistique se rajoute simplement à celles déjà considérées dans le résumé statistique puis est traitée de façon identique.

6.2.12 Problèmes liés aux données

Quand on utilise des données médicales relatives à des grands-parents ou arrière-grands-parents nés dans la première moitié du XX^{ème} siècle, les données manquantes peuvent être nombreuses. Elles peuvent se révéler invalidantes pour la méthode car les individus ne sont pas considérés individuellement mais comme élément d'un arbre et c'est l'information conjointe à cette structure d'arbre que nous voulons exploiter. Si la proportion d'arbres incluant des données manquantes est très importante, il peut falloir trouver un algorithme pour « réparer » les données. A cet égard, l'appartenance à un arbre généalogique redevient

précieuse car elle permet d'affecter certaines données manquantes, comme l'année de naissance ou le sexe. La réparation fonctionne souvent par passes successives. *In fine*, il arrive souvent que les parties d'arbres qui se trouvent effacées sont celles que n'apportaient en fait guère d'information.

En outre, un certain nombre de situations peuvent compliquer l'exploitation : recomposition des couples parentaux, adoption… Une prise en compte de ces situations par l'outil informatique est alors requise. Une autre source de bruit de fond provient du fait que plus souvent qu'on ne le pense, les pères officiels de certains enfants ne sont pas leurs pères biologiques, ce que montrent les tests ADN. Malheureusement, ce bruit n'est pas contrôlable.

Pour l'application au cancer que nous avons mené, le renseignement du type de cancer et en particulier sa localisation est primordial, mais là encore il constitue souvent une donnée manquante dans les générations les plus anciennes. L'examen des autres membres de l'arbre ne permet malheureusement pas de compléter par la modalité la plus probable, ce qui induit une perte majeure de qualité de l'information généalogique.

Un autre problème pratique est de reconstituer à travers les différentes données de la base les relations de parenté entre les individus, puis d'exprimer cette structure généalogique dans un format compatible avec le logiciel d'analyse. Mais c'est une difficulté purement informatique.

6.2.13 Application à deux types de prédisposition familiale de cancer

Il existe deux prédispositions familiales majeures au cancer : le syndrome sein/ovaire et les cancers colorectaux, ou encore syndrome de Lynch. Dans notre base de données oncogénétique, nous disposons de 5 cohortes d'individus :

- Un groupe de 418 familles chez lesquelles une mutation des gènes BRCA a été détectée pour au moins un des membres
- Un groupe de 1316 familles avec prédisposition sein/ovaire mais pour lesquelles aucune mutation BRCA n'a été retrouvée. Elle sera dénommée « non mutée BRCA ».
- Un groupe de 394 familles où aucune prédisposition héréditaire de cancer n'a été diagnostiquée (principalement des familles consultant à tort)
- Un groupe de 90 familles avec un syndrome de Lynch (côlon) et une mutation MMR trouvée chez au moins un des membres de la famille
- Un groupe de 376 familles avec prédisposition au cancer colorectal mais sans mutation encore diagnostiquée chez un des membres.

L'objectif du travail informatique est de confirmer ou d'infirmer le caractère héréditaire de la prédisposition de cancer pour chacun des groupes cités. Dans le cas des groupes avec mutation BRCA ou MMR avérée, la modalité 1 devrait ressortir. S'il n'y a pas de prédisposition héréditaire, la modalité 0 devrait être sélectionnée. Enfin pour les autres groupes, avec prédisposition mais sans mutation avérée, les modalités 1 ou 2 pourraient également être choisies après calculs. A l'heure où ces lignes sont écrites, les machines n'ont pas encore assez

tourné pour que nous puissions être totalement affirmatifs. Un premier « run » sur deux cohortes a été fait par la méthode de minimisation de distance sans « pousser » au maximum les paramètres de fonctionnement, donc avec un temps de calcul de l'ordre de l'heure. Pour la cohorte dénommée « mutée BRCA », c'est la modalité n°1, donc à mutation unique, qui ressort avec la plus petite distance, environ deux fois moindre que celles correspondant aux deux autres modalités. A titre d'illustration, voici comment les lois de l'âge de déclaration des cancers sont « fittées » dans les 3 modalités :

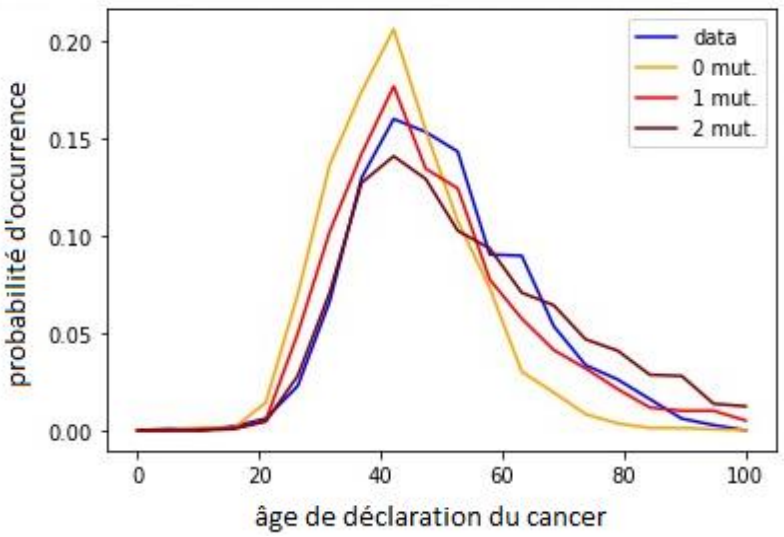


Figure 76 : Distributions de l'âge du cancer obtenues grâce aux simulations selon les trois modalités envisagées

On rappelle que le logiciel cherche à ajuster ces courbes simultanément à bien d'autres statistiques : il doit donc choisir les paramètres qui optimisent le compromis global.

Dans l'évaluation de la cohorte « non mutée BRCA », c'est la modalité 0 qui arrive en tête (distance minimale), suivie par la modalité à une mutation assez proche derrière. On ne peut donc estimer que le logiciel soit très performant dans ces familles où le risque héréditaire n'est pas confirmé par les tests génétiques sur les gènes BRCA, mais où ce risque existe quand même.

6.2.14 Discussion sur les méthodes

6.2.14.1 Comparaison entre les méthodes neuronales ou par minimisation de distance

Comme les tests sur les données réalistes l'ont montré, la méthode par réseau de neurones paraît plus performante tant du point de vue classification qu'estimation des paramètres. Mais la méthode par minimisation de distance conserve les avantages suivants :

- Pour la classification de modalité, les valeurs de distance fournissent une quantification de la proximité des modalités

- La statistique supplémentaire de difficulté de simulation peut être incorporée à la méthode par minimisation mais n'a pas d'équivalent en apprentissage neuronal.
- La méthode neuronale nécessite un apprentissage coûteux en temps machine, qu'il faut refaire sur chaque généalogie avec données réelles
- La méthode par minimisation donne des résultats assez fiables avec des besoins en temps de calcul très acceptables

6.2.14.2 Autres méthodes

On peut s'interroger sur la méthode choisie telle que décrite précédemment et de sa pertinence par rapport à d'autres méthodes classiques, car si ses performance sont bonnes, sont-elles meilleures que celle des autres méthodes ? En l'occurrence ce qui qualifie le mieux nos méthodes est de se dispenser du calcul de vraisemblance. Or la fonction de vraisemblance paraît un outil tout désigné pour quantifier avec le plus de précision les subtilités statistiques de notre échantillon.

Notons tout d'abord que les effets que l'on cherche à détecter sont finalement assez ténus car avec des valeurs de paramètres typiques, les cas de maladie K ne sont pas si fréquents que cela et beaucoup surviennent de manière sporadique au sens où ils ne sont pas dus à une mutation qui se propagerait le long d'une lignée.

Désignons par $L_m(\cdot;\ \theta)$ la fonction de vraisemblance pour la modalité *m* avec les paramètres θ. Cette fonction est définie sur l'espace auquel appartiennent les données, c'est à dire l'espace des phénotypes des individus. Toutefois, dans notre modèle où les phénotypes sont stochastiquement déduits des génotypes, il faut faire intervenir les génotypes sous-jacents des individus comme des variables cachées indispensable à l'écriture d'une fonction de vraisemblance.

Laissons de côté pour l'instant les modalités θ et *m* dans nos notations. La vraisemblance de l'ensemble des génotypes des membres d'une famille $\tilde{L}(g)$ s'exprime facilement en utilisant les lois de Mendel et les fréquences des mutations dans la population générale. Il est aussi facile d'exprimer la vraisemblance conditionnelle $L(f|g)$ des phénotypes f d'une famille sachant le génotype *g* des membres de cette famille. Alors, la fonction de vraisemblance $L(f)$ du modèle est :

$$L(f) = \sum_{g:Géno} L(f|g)\ \tilde{L}(g)$$

Dans la pratique, le calcul exact de cette somme par sommation des contributions associées à tous les génotypes possible devient vite irréalisable car il y a *a priori* 2^n génotypes possibles dans la modalité à une mutation et 4^n dans la modalité à deux mutations en interaction, avec n désignant le nombre d'individus de la famille. En réalité une proportion non négligeable de génotypes est de probabilité nulle mais les identifier revient à calculer leur vraisemblance. Par contre il est bien sûr possible de trouver une valeur approchée de $L(f)$ par simulation :

$$L(f) \approx \frac{1}{k}\sum_{j=1}^{k} L(f|g_j)$$

Où les g_j forment une suite de tirages indépendants de la loi $\tilde{L}(\cdot)$ obtenus soit par simulation directe, soit par l'algorithme de Métropolis. Pour un échantillon $F = \{f_1, \ldots, f_N\}$ formé de N familles, la log-vraisemblance pourra donc être calculée par une expression du type :

$$\log L(f) \approx \sum_{r=1}^{N} \log\left(\frac{1}{k}\sum_{j=1}^{k} L(f_r|g_j)\right)$$

l'égalité étant réalisée pour $k \to +\infty$. Cela signifie que les incertitudes des calculs par simulation de $L(f_r)$ s'ajoutent, bien qu'on puisse raisonnablement espérer une compensation mutuelle au premier ordre si les erreurs sont distribuées de façon centrée et gaussienne et qu'elles sont indépendantes. Toujours est-il que le calcul de vraisemblance se révèle assez long et bruité, ce qui est de mauvais augure pour identifier les paramètres par la méthode du maximum de vraisemblance. Plus grave encore, une étude numérique montre que le profil de la log-vraisemblance, représenté dans ses variations par rapport aux paramètres est assez plat. Au total, l'identification des paramètres par maximum de vraisemblance s'est révélée numériquement impossible.

Pour ce qui est de la sélection de la modalité, on peut imaginer se baser sur un rapport de vraisemblance. Plus précisément, si $f_1, f_2, \ldots$ sont générés selon la modalité 1, on a par loi des grands nombres :

$$lim_{n \to +\infty} \frac{1}{n}\sum_{r=1}^{n}(\log L_1(f_r) - \log L_2(f_r)) = E_1\left(\log\frac{L_1(f)}{L_2(f)}\right) = K(L_1|L_2) > 0$$

Où $K(L_1|L_2)$ est la divergence de Kullback-Liebler de la modalité 2 par rapport à la modalité 1. Par contre si $f_1, f_2, \ldots$ sont générés selon la modalité 2, on a en intervertissant les rôles,

$$lim_{n \to +\infty} \frac{1}{n}\sum_{r=1}^{n}(\log L_1(f_r) - \log L_2(f_r)) = -K(L_2|L_1) < 0$$

Si ce résultat paraît pouvoir conduire à un critère décisionnel, l'expérimentation numérique se révèle décevante, peut-être à cause de la trop faible distance entre modalités pour la divergence Kullback-Leibler, mais peut-être aussi à cause du bruit qui affecte tout calcul de vraisemblance.

Ainsi les expérimentations numériques autour de la vraisemblance se sont révélées négatives tant pour la sélection de modalité que pour l'estimation paramétrique.

6.2.15 Remarques complémentaires sur la transmission génétique

Le modèle d'hérédité que nous utilisons est une simplification de la réalité puisqu'une version de chaque gène figure sur chacun des deux chromosomes appariés soit à l'identique

(mutations homozygotes), soit avec deux expression différentes (mutations hétérozygotes). Avec les lois de Mendel de transmission des gènes, cela assure qu'en l'absence de mutation *de novo*, les proportions des différents allèles d'un gène dans une grande population restent fixes de génération en génération, résultat connu sous le nom d'équilibre de Hardy-Weinberg. Le modèle simplifié que nous utilisons revient à supposer que les cas de mutation sont toujours hétérozygotes. Compte tenu de la faible fréquence f_m des mutations considérées, de l'ordre d'au plus quelques pourcents, c'est une approximation acceptable. Néanmoins, en toute rigueur, ce modèle engendre une diminution de la fréquence de mutation d'une génération à la suivante. En effet, si cette fréquence est f_m dans une génération, elle sera :

$$\frac{3}{4} f_m^2 + \frac{1}{2} 2 f_m(1 - f_m) = f_m \left(1 - \frac{f_m}{4}\right)$$

à la génération suivante, compte tenu des lois de Mendel. Notons toutefois que la diminution est typiquement de l'ordre du pourcent. Cela peut se vérifier numériquement sur les familles simulées. Cela veut dire que la loi régissant le génotype d'un individu dépend, certes assez faiblement, de notre distinction ou non des allèles. Cet effet est néanmoins largement inférieur à l'impact du conditionnement que nous avons dû utiliser pour compenser les biais de sélection.

6.2.16 Bilan des approches par minimisation de distance et réseau de neurones

Le bilan des résultats dont nous disposons pour le moment est que nos algorithmes fonctionnent avec un niveau de performance déjà très satisfaisant sur les données « réalistes » avec les tailles et types de données disponibles et même dans le cas pénalisant d'une maladie réduite à un seul sexe. Les essais que nous avons déjà réalisés sur les données issues de la base de données du Centre Jean Perrin laissent un peu à désirer. Si le distinguo entre les familles sans risque avéré de cancer et celles pour lesquelles ce risque est confirmé par une mutation BRCA, les approches par minimisation ou par réseau de neurones réalisent une bonne classification. Par contre, dans les familles diagnostiquées par l'oncogénéticien comme étant à risque élevé mais où aucune mutation BRCA n'a été trouvée, nos modèles font fausse route, privilégiant la modalité 0, c'est à dire sans mutation délétère. Toutefois pour ce dernier cas, le choix de la modalité par les modèles demeure très proche.

Les performances peuvent et doivent encore progresser avec davantage de données et peut-être un meilleur réglage des paramètres des algorithmes. Malgré tout, on retrouve les conclusions habituelles dans ce genre de situation : quand les familles que l'on peut apparenter à des nuages de points se recouvrent dans une mesure non négligeable, toutes les approches automatiques de classification peinent à atteindre un niveau satisfaisant de performance.

7 Discussion

La problématique de ce travail concernait la manière d'étudier les phénotypes des familles que l'on peut lire grâce à leur arbre généalogique. Deux éclairages ont été utilisés, un premier sur des données simulées et un second grâce à la disposition de données réelles de qualité, celles de la base de données oncogénétique du Centre Jean Perrin.

L'étape préliminaire a donc consisté à produire des données simulées conformes aux données réelles. Ceci a été réalisé en prenant en compte les paramètres populationnels de natalité/fécondité et par ailleurs l'incidence des cancers selon l'âge et les principales mutations délétères connues.

Une première analyse de ces données simulées nous a permis de mettre en évidence les avantages en termes de fertilité que prodiguaient les mutations délétères en compensation du raccourcissement de la durée de vie qu'elles induisent. En parallèles, nous avons pointé un effet secondaire de cet avantage de fertilité qui se traduisait par une augmentation du risque de polymalformation chez les nouveau-nés. Ces résultats ont fait l'objet de deux publications dans des revues de bon impact.

Nous avons ensuite élaboré une nouvelle représentation des arbres généalogiques, appelée sous-arbre, que nous avons déclinée en 2 ou 3 générations. Cette structure devait permettre la comparaison des différents arbres généalogiques et par suite les phénotypes correspondants dans le but de trouver des groupes de familles à risque spécifique. Les sous-arbres ont été comparés aux données moyennées par famille, ces dernières étant utilisées quand on veut calculer des scores de risque mutationnel. Une autre comparaison a été faite avec une méthode de retour aux données individuelles. Deux types d'analyse ont été effectués pour évaluer l'impact sur le pronostic des mutations de ces différentes structures de données.

- Des analyses en composantes principales : elles n'ont pas montré l'utilité des sous-arbres sur les données simulées. Des analyses en clusters (*k-means* clustering) ont alors été réalisées sur les données réelles, au sein des familles sans mutation connue. Les données familiales moyennes ont donc été utilisées. Elles ont permis de caractériser des sous-groupes particuliers. Les numéros des familles ainsi isolées ont ensuite été transmis au laboratoire d'oncologie moléculaire du LOM pour test génétique à l'aide d'un panel de gènes étendu. Nous espérons que des mutations différentes seront trouvées dans chacun de ces sous-groupes.
- Des classifications ascendantes hiérarchiques : Les tests sur données simulées ont semblé montrer que le retour aux données individuelles permettait de mieux différencier que les données familiales moyennées et les données issues des sous-arbres, quand les phénotypes étaient assez similaires. En particulier, il paraissait enfin possible de distinguer les phénotypes issus de mutations délétères seules de ceux produits par des mutations faiblement pénétrantes nécessitant pour devenir délétères des associations à des polymorphismes fréquents. Des tests sur des familles extraites de la base de données oncogénétique n'ont pas confirmé cet avantage en particulier dans le cadre assez pénalisant de la comparaison des mutations BRCA1 et BRCA2. Si ce retour aux données individuelles n'a pas prouvé son intérêt, nous avons par contre élaboré une méthode de comparaison des dendrogrammes générés par les CAH, méthode qu'il nous faudra publier.

Le deuxième type d'approche a consisté à réaliser la classification des familles en les comparant à des groupes de familles simulées selon 3 modalités (0 : pas de mutation, 1 : une seule mutation délétère et 2 : deux mutations/variants en interaction). Pour ce faire, les lois de Mendel ont été utilisées pour gérer à travers les générations la transmission des mutations et les phénotypes ont été calculés,

moyennant un processus stochastique, en respectant les lois de l'âge de déclaration propres à chaque modalité. Deux types de classification ont alors été utilisés : une par minimisation de distance et l'autre en programmant des réseaux multicouches de neurones. Les résultats les plus intéressants mettent en lumière le fait que des arbres de 3 ou 4 générations pourraient constituer un optimum pour la réalisation du diagnostic oncogénétique, les informations supplémentaires semblant constituer un bruit qui détériore les méthodes de classification. On doit toutefois nuancer ces résultats en raison du type de simulation des familles et de leurs arbres : en effet, nous avons supposé, dans tous les chapitres de cette thèse, que les cancers étaient de type et de localisation unique. Or nous avons vu lors de nos approches par clustering que l'information histologique des cancers du sein et la variété des autres localisations cancéreuses permettaient de distinguer les types de risque héréditaire. En outre, la restriction à un nombre limité de générations peut empêcher le recueil des cas de cancer assez rares, mais apportant des précisions importantes pour la typologie du risque génétique en arrière-plan. Ceci nous semble d'autant plus vrai dans le cas où plusieurs mutations ou variants de signification inconnue contribuent au risque de cancer.

Les calculs réalisés à l'aide des deux méthodes précédentes dans des familles exposées à une prédisposition de cancer sein/ovaire issue de la base de données du CJP permettent d'assez bien distinguer les familles avec mutation BRCA et celles sans prédisposition. Par contre, les familles présentant un haut risque mais sans mutation connue sont plus souvent classées dans le groupe sans prédisposition que dans le groupe avec mutation, voire avec des variants en interaction. Ils doivent donc encore être affinés et il semble un peu tôt pour conclure définitivement quant à leur efficacité. Malgré tout, au vu des premiers tests, nous faisons de nouveau face à une problématique bien connue : la plupart des algorithmes fonctionnent bien quand les « paquets » d'information sont spatialement bien éloignés. La difficulté commence quand les nuages se recouvrent dans une large mesure. L'idée d'augmenter la quantité d'information pourrait sembler une bonne solution, mais la pratique des modèles de classification montrent que l'ajout de nouveaux paramètres, partiellement redondants, crée souvent davantage de bruit que d'amélioration de la puissance discriminante : c'est donc un optimum bien délicat qu'il s'agit de trouver. Des recherches théoriques dans ce domaine seraient bien utiles afin d'aider à déterminer où placer la barre entre ces deux directions contradictoires.

Nos modélisations n'ont pas pris en compte les changements environnementaux, comme par exemple les progrès de la médecine qui réduisent la perte d'espérance de vie des femmes mutées BRCA à 3 ans, ce que l'on peut comparer à une perte de 10 ans en cas d'obésité morbide (IMC $\geq$ 40 kg/m²) par exemple[70]. On n'a pas non plus intégré les interactions épigénétiques, qui confèrent une plasticité aux processus purement génétique en permettant une adaptation intergénérationnelle fine à un environnement changeant. On trouve là les limites des modélisations qui ne sont capables, parce qu'elles réduisent drastiquement l'éventail des paramètres biologiques ou autres, de ne capturer que les effets les plus marquants.

Nous n'avons toutefois pas totalement cédé à la facilité en intégrant comme facteur de confusion les cancers sporadiques. Cette introduction, indispensable toutefois pour que les modélisations aient un semblant de réalité, induit deux difficultés : la première est que les cancers sporadiques, pouvant advenir à des âges parfois jeunes, brouillent la frontière entre les prédispositions familiales et les familles sans prédisposition. La seconde est que l'existence de formes familiales de cancers tardifs a été démontrée[71,72]. Ces formes de prédisposition familiale tardive sont rarement diagnostiquées du fait de leur non prise en compte lors du calcul des scores classiques évaluant le risque familial. Elles figurent alors en bonne place au sein des cancers sporadiques ce qui dégrade d'autant la performance

des modèles. Sans doute faudrait-il procéder à une réévaluation des scores pronostiques en réintégrant ces familles exposées à un risque familial réel.

Il serait dommage de ne pas évoquer, dans cette discussion, la problématique de la gestion des données, qu'elles soient simulées ou bien issues de la base de données du Centre Jean Perrin. La mise en forme et la vérification des données constituent une phase souvent silencieuse, invisible, mais pourtant dispendieuse en temps et exigeant une attention de tous les instants. Les données réelles sont quant à elles très souvent lacunaires, ce qui s'avère rédhibitoire quand on veut utiliser des logiciels existant dans R, Matlab ou en Python (ex. les réseaux de neurones). De très importants efforts ont donc été alloués à ces opérations d'arrière-plan, mais aussi pour le développement de certaines parties logicielles (clustering).

Les progrès fantastiques sur le plan technologique permettant aujourd'hui des analyses génétiques pangénomiques pourraient laisser penser que toutes ces recherches méthodologiques sont des combats d'arrière-garde. Loin s'en faut ! En effet, ces avancées technologiques induisent une inflation d'informations sans précédent, ce que l'on nomme le « big-data ». Et aujourd'hui, on sait que quand la production de données biologiques prend 30% de temps-technicien, elle nécessite 70% de temps-biologiste ou ingénieur pour son interprétation. C'est donc un phénomène inverse auquel nous assistons : le développement des moyens d'information va rendre indispensable la création d'outils toujours plus pointus et performants afin d'ordonner, de digérer cette masse d'information croissante et de produire ces synthèses indispensables au diagnostic, à la prise de décision concernant d'éventuelles mesures prophylactiques et au choix des traitements ciblés une fois la maladie déclarée. C'est la condition *sine qua non* de la médecine personnalisée de demain.

Cette thèse, hormis ses apports certes modestes sur le plan méthodologique, aura sans doute permis de mettre en lumière l'effet délétère des interactions génétiques entre variants non reconnus comme délétères à eux seuls. Ce premier résultat doit nécessairement être suivi par de plus amples recherches qui éclaireront à mesure que de nouveaux résultats arrivent, comment et à quel niveau fonctionnent ces interactions.

8 Annexes

8.1 Synoptique du contexte de la thèse

Cette thèse s'insère dans une recherche biologique plus large dons nous regroupons les principales étapes dans le schéma ci-dessous :

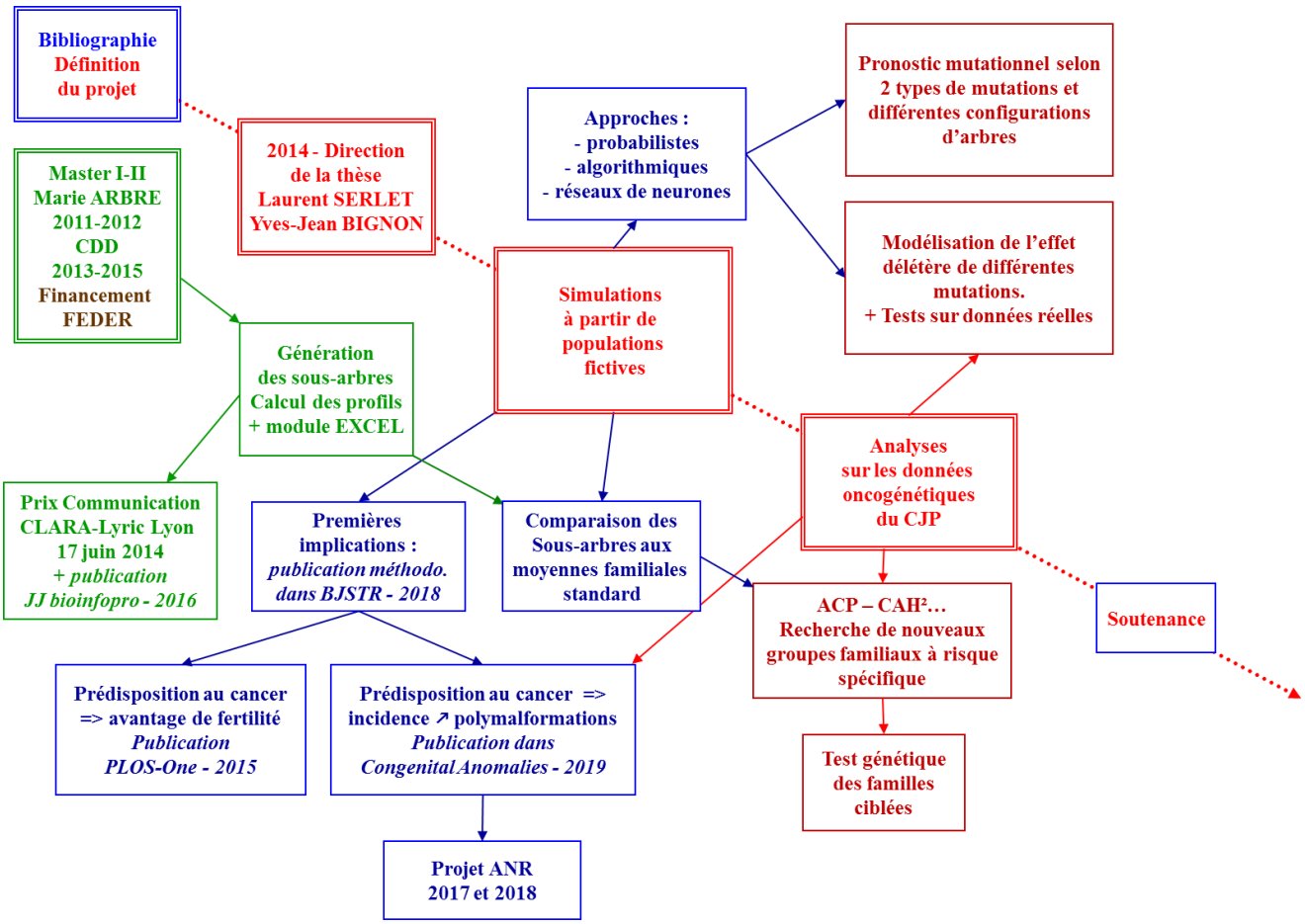


Figure 77 : synoptique de la recherche globale dans laquelle se situe la thèse actuelle

Le projet ANR faisait suite à nos résultats sur la fréquence supérieure de polymalformations congénitales dans les familles mutées BRCA, mais aussi à des recherches non encore publiées sur le lien entre des longueurs un peu plus courtes des télomères (non considérées comme des téloméropathies) et les malformations congénitales menées par le Pr Andreï TCHIRKOV et le Dr Carole GOUMY à l'unité de de cytogénétique de l'hôpital Estaing (Clermont-Ferrand). Les unités de génétique médicale de Lyon et de Clermont étaient associées à ce projet ANR et bien évidemment le LOM du Centre Jean Perrin. Voici un bref résumé de ce projet nommé TREMMA :

« L'étiologie des malformations congénitales reste pour l'essentiel inconnue. Suite à notre mise en évidence d'un risque accru de malformations congénitales multiples chez les individus issus de familles prédisposées au cancer en raison de mutation sur des gènes impliqués dans la réparation de l'ADN, nous avons émis l'hypothèse que les malformations congénitales touchant plusieurs systèmes anatomiques pourraient provenir d'anomalies sporadiques durant l'embryogenèse qui seraient mal réparées en cas de mutation sur

des gènes impliqués dans la réparation de l'ADN ou le maintien du télomère. Pour vérifier cette hypothèse, nous proposons de rechercher ces mutations au sein de 2 groupes de 200 nouveau-nés ou fœtus provenant d'interruption médicale de grossesse, en comparant leur fréquence selon que ces nouveau-nés/fœtus présentent des malformations uniques ou multiples. Un panel d'environ 70 gènes sera analysé en séquençage haut débit à partir à partir de prélèvements sanguins recueillis chez les enfants et les parents. »

Un groupe contrôle sans malformation congénitale était aussi prévu. Ce projet, bien que bien classé, n'a pas été sélectionné par l'ANR.

8.2 Les articles publiés ou en cours

8.2.1 Plos-One (2015) BRCA Mutations Increase Fertility in Families at Hereditary Breast/Ovarian Cancer Risk

PLOS | ONE

RESEARCH ARTICLE

BRCA Mutations Increase Fertility in Families at Hereditary Breast/Ovarian Cancer Risk

Fabrice Kwiatkowski[1,2]*, Marie Arbre[1], Yannick Bidet[3], Claire Laquet[1], Nancy Uhrhammer[1], Yves-Jean Bignon[1,3]

1 Centre Jean Perrin, Laboratoire d'Oncologie Moléculaire, 63011, Clermont-Ferrand, France, **2** Université Blaise Pascal—Laboratoire de Mathématiques, UMR 6620—CNRS, Campus des Cézeaux—BP, 80026–63171, Aubière cedex, France, **3** Université Clermont Auvergne, Université d'Auvergne, BP 10448, F-63000, Clermont-Ferrand, France

* Fabrice.Kwiatkowski@CJP.fr

OPEN ACCESS

Citation: Kwiatkowski F, Arbre M, Bidet Y, Laquet C, Uhrhammer N, Bignon Y-J (2015) BRCA Mutations Increase Fertility in Families at Hereditary Breast/Ovarian Cancer Risk. PLoS ONE 10(6): e0127363. doi:10.1371/journal.pone.0127363

Academic Editor: Peiwen Fei, University of Hawaii Cancer Center, UNITED STATES

Received: November 20, 2014

Accepted: April 14, 2015

Published: June 5, 2015

Data Availability Statement: All relevant data are in the paper and its Supporting Information file.

Funding: This study received no external funding. This research was totally financed by Centre Jean Perrin. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Background

Deleterious mutations in the BRCA genes are responsible for a small, but significant, proportion of breast and ovarian cancers (5 - 10 %). Proof of *de novo* mutations in hereditary breast/ovarian cancer (HBOC) families is rare, in contrast to founder mutations, thousands of years old, that may be carried by as much as 1 % of a population. Thus, if mutations favoring cancer survive selection pressure through time, they must provide advantages that compensate for the loss of life expectancy.

Method

This hypothesis was tested within 2,150 HBOC families encompassing 96,325 individuals. Parameters included counts of breast/ovarian cancer, age at diagnosis, male breast cancer and other cancer locations. As expected, well-known clinical parameters discriminated between BRCA-mutated families and others: young age at breast cancer, ovarian cancer, pancreatic cancer and male breast cancer. The major fertility differences concerned men in BRCA-mutated families: they had lower first and mean age at paternity, and fewer remained childless. For women in BRCA families, the miscarriage rate was lower. In a logistic regression including clinical factors, the different miscarriage rate and men's mean age at paternity remained significant.

Results

Fertility advantages were confirmed in a subgroup of 746 BRCA mutation carriers and 483 non-carriers from BRCA mutated families. In particular, female carriers were less often nulliparous (9.1 % of carriers versus 16.0 %, $p = 0.003$) and had more children (1.8 ± 1.4 SD versus 1.5 ± 1.3, $p = 0.002$) as well as male carriers (1.7 ± 1.3 versus 1.4 ± 1.3, $p = 0.024$).

Conclusion

Although BRCA mutations shorten the reproductive period due to cancer mortality, they compensate by improving fertility both in male and female carriers.

Introduction

Dominant deleterious mutations in a population should be suppressed unless they have a compensating effect on fertility or are expressed only after the fertile period. Highly penetrant deleterious mutations are subject to selection pressure [1]; those affecting young people without conferring a fertility advantage will therefore often be *de novo* mutations. Mutations in the tumor suppressor genes TP53 or pRb, where cancer often develops in childhood, follow this pattern [2]. Deleterious mutations in the BRCA genes, responsible for hereditary breast/ovarian cancers (HBOC), only partially follow this model. The majority are unique to a family (cf http://research.nhgri.nih.gov/bic/ and http://www.umd.be/BRCA1/, for example), though it is usually not possible to determine in what person it first occurred, and reports of *de novo* mutations are quite rare [3–4]. But this latter point should nowadays be perspectived: frenquency of germline *de novo* mutations has been found higher in recent studies where a rate of 3.5 to 4 mutations per individual is estimated (mostly from male genitors) [5], and with an increasing mutational risk with age [6]. This is in accordance with the French UMD-BRCA1/BRCA2 database where about half of the mutations were unique to a family (thus possibly recent)—respectively 53% and 63% of deleterious BRCA1 and BRCA2 mutations—while 18 other mutations occurred in at least 8 families and two in more than 130 [7]. The haplotypes of many recurring mutations have a common ancestral origin; although a few have occurred more than once, some of those segregating in specific populations are known to be thousands of years old [8–10]. BRCA mutations thus appear to be a mix of rare private mutations, some of which may be recent, and more common mutations passed down through numerous generations.

The age at which BRCA mutation carriers develop cancer overlaps the reproductive period, so there should be some mechanism by which mutations persist in the population. Several studies have noted that BRCA mutation carriers have higher parity than non-carriers, suggesting a positive effect on female fertility [11]. Others, in contrast, have associated BRCA mutation with reduced ovarian function [12], or with voluntarily reduced reproduction [13], and still others have found no effect [14]. We thus decided to evaluate fertility outcomes (resulting from reproduction factors possibly unknown) concurrently with known predictive factors for BRCA mutational status, in a large database of 2,936 BRCA and non-BRCA HBOC families. A two-step analysis was realized: first with families considered as a entity, then by individuals grouped according to their known BRCA mutational status.

Materials and Methods

Pedigrees

Families were accrued at the oncogenetic consultation of the Centre Jean Perrin in central France from 1988 to 2013, and included a huge majority of Caucasians (≈98%). Information collected for pedigrees comprised extensive notation of the proband's relatives without limitation of the number of generations included as long as cause of death was known: median generation count was 4 and interquartile interval [2; 5]). Also included were cancer location and age

at diagnosis, mutations known or discovered subsequently, date of birth, gender, marital status, descendancy, miscarriages, dates of marriage, separation and death.

The database was declared to the French National Informatics & Liberty Committee (CNIL) on May 18th 2011 by the CIL (the local CNIL correspondant) and in accordance with the article R. 1131–2 of the French public health code, counselees signed a special consent enabling the use of their data for reasearch purpose. It was managed using SEM software [15], which also performed statistics and special calculations (age at first/last birth, age at first cancer, rates per family of miscarriage or of childlessness. . .).

When pedigrees with a breast/ovarian cancer risk were extracted from the database, 2,168 families including 96,325 at-risk individuals were selected. Families needed to contain at least 5 female affiliated members (i.e. ≈ 10 members), otherwise we considered that not enough pedigree information was available to be reliable. BRCA1 mutations were found in 10% of families (214 families; 11,349 members), and BRCA2 mutations in 7% (161 families; 8,255 members). In 1,775 families (87,216 members) no mutation was found: this group is identified hereafter as "no mutation". Eighteen families diagnosed with other mutated genes were excluded (Fig 1): 2,150 families were thus statistically analysed.

The average number of family members was 51.2 ± 35.1 (SD), with on average 23.7 ± 15.7 females, and 24.2 ± 16.6 males per pedigree. Cancer in these families included 5,821 breast cancers (5,718 female and 103 male), 631 ovarian cancers, 285 endometrium, 604 colon, 222 pancreatic, 589 prostatic and 3,608 other location.

Prediction parameters

Excepted for fertility, relevant parameters were retrieved from published scoring strategies [16–21]. The following cancer locations were included: breast, ovary, endometrium, pancreas, colon and prostate; all other locations were grouped together. Bilateral breast cancers were counted twice. Female breast cancers were categorized by age at diagnosis in 10-year classes: <30, 30–39, 40–49, 50–59, 60–69, ≥70. Triple-negative breast cancers were not considered because their reporting was too recent and incomplete in our database.

Statistical analysis included the following counts:

- Number of breast cancers in families by class of age
- Number of male breast cancers
- Number of cancers detailed by: ovary, endometrium, colon, pancreas, prostate
- Number of cancers of any other type
- Number of persons with multiple cancers
- Number of members per pedigree and number of males and females
- Number of miscarriages
- Average number of children per potential mother (i.e. women with at least a spouse and/or a child and/or age ≥ 40)
- Average mother's age at first child (if any).

When age at cancer diagnosis was unknown, age at last followup (or of death) was used instead, considering that in older generations, the average time between these events was short. When ages at cancer diagnosis and death were both unknown, we replaced missing values by the average age of family members belonging to the same generation. Age at first child was computed for each woman who had a child, and the average computed by family. Finally, an

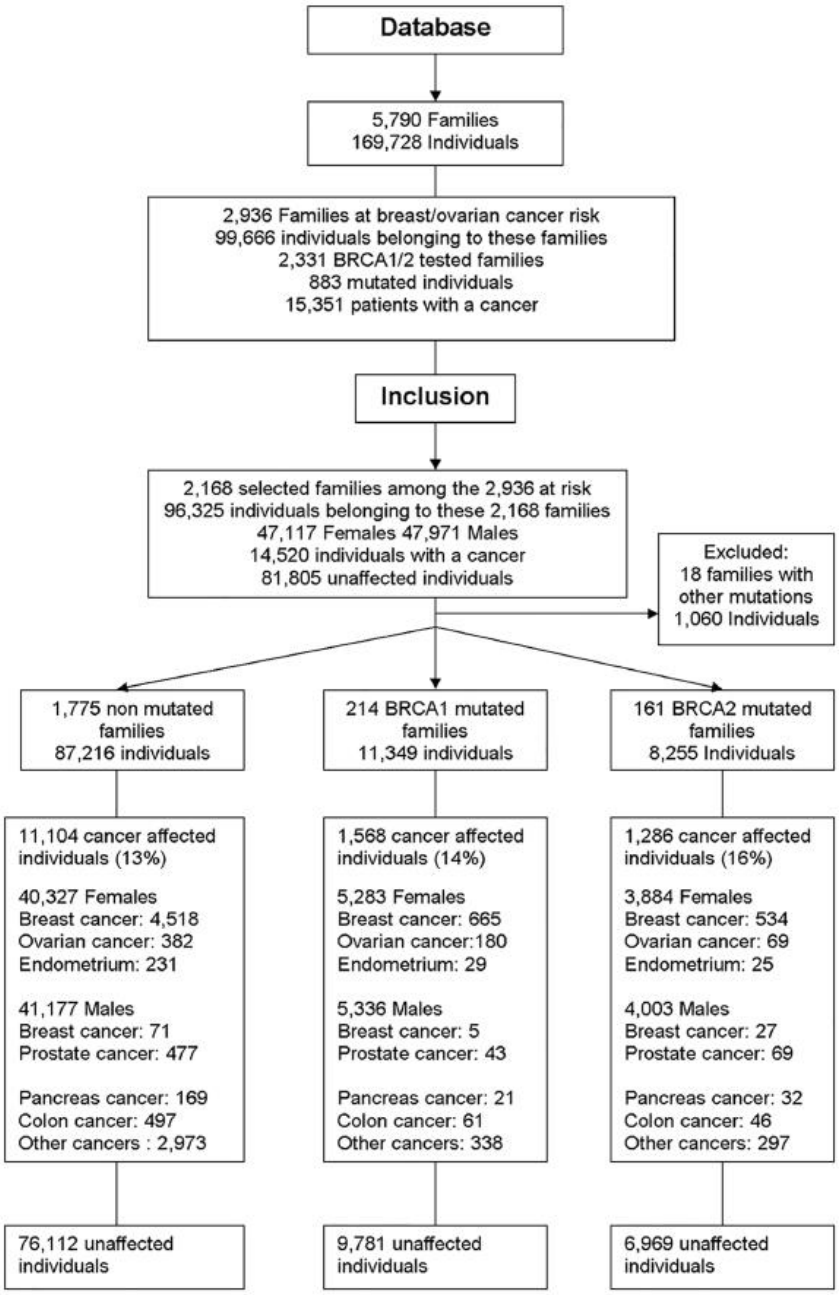


Fig 1. Accrual flowchart of the families and the individuals selected from the Database.

doi:10.1371/journal.pone.0127363.g001

average number of children was calculated per family, taking into account women with at least one child and/or being in a couple at least once, or being older than 40 years (an age *a priori* sufficient to give birth at least once). We used this strategy because the proportion of young single women could differ between groups and counting them as "zero-child mothers" could bias the results. Inherited risk calculations and other statistics were not applied to spouses: only members exposed to the familial cancer-risk were used in calculations.

Statistics

Only the branch(es) with cancer risk was entered into the database for > 90% of pedigrees, so calculations were done by entire pedigree. Univariate comparisons with mutation status were performed using Z-test or H-test depending on homoscedasticity and/or normality of distributions. Tests were two-sided and a p value ≤ 0.05 was considered significant. Multivariate analysis to order covariates consisted of backward logistic regressions. The adequacy of models to the data was evaluated with the Hosmer-Lemeshow test. To test the efficacy of scores to predict mutation status, a ROC analysis was performed, and the area under curve (AUC) compared together.

Significant clinical factors predictive for BRCA-mutated status were selected first using univariate analysis, then classified by logistic regression. Fertility parameters associated to the mutation status by univariate analysis to a p-value ≤ 0.10 were then introduced in the logistic regression model concurrently with significant clinical factors.

Comparisons were performed within following groups:

- BRCA mutation versus no-mutation
- BRCA1 mutation versus no-mutation
- BRCA2 mutation versus no-mutation
- BRCA1 mutation versus BRCA2 mutation

Results

a) role of standard clinical parameters on BRCA mutation risk

Repartition of main cancer locations according to mutation status of families are exhibited in Table 1. Each group corresponds to members of families where a BRCA mutation was found, without testing the mutation status of each individual.

Logistic regression enabled us to estimate the respective weight of each parameter. Standard clinical factors (Fig 2) predicted both BRCA mutation status (BRCA1 and BRCA2), notably the number of breast cancers occurring before 50 years and ovarian cancers. Pancreatic and male breast cancers were significant for BRCA2. The distinction between a BRCA1 and a BRCA2 mutation depended on four factors: ovarian cancers favored BRCA1 mutations (p = 0.00011) while male breast cancers (p = 0.0045), prostatic cancers (p = 0.012) and pancreatic cancer (p = 0.022) were more frequent in BRCA2 mutated families.

In a global comparison by logistic regression of BRCA1 plus BRCA2 mutated families to no-mutation families, the best fitted model performed slightly better (ROC AUC = 0.73 [0.70–0.76]) than the Eisinger and Manchester predictive scores (AUC = 0.70 [0.33–0.73] for both), but this difference was not significant. Also, Eisinger's score incorporating very few elements (ovarian cancers, female breast cancers depending on 6 classes of age and male breast cancers) performed as well as the Manchester score including 15 parameters (16 if triple-negative breast cancer status is included).

Table 1. Number of cancer locations according to diagnosed mutation (% of members(*)).

Cancer location	BRCA1	BRCA2	No mutation	p-value
Women: Breast	502 (12.2%)	419 (13.7%)	3,478 (10.7%)	0.00013
Ovarian	138 (3.3%)	61 (2.0%)	285 (0.9%)	< 10−7
Endometrium	22 (0.5%)	20 (0.7%)	200 (0.6%)	0.91
Men: Breast	2 (0.04%)	14 (0.44%)	53 (0.16%)	0.00053
prostate	36 (0.8%)	57 (1.8%)	386 (1.1%)	0.026
Any sex: Colon	56 (0.7%)	36 (0.6%)	390 (0.9%)	0.59
Pancreas	17 (0.2%)	22 (0.4%)	125 (0.3%)	0.02
Other location	295 (3.6%)	281(4.5%)	2,684 (6.4%)	0.42
Multiple location	76 (1.8%)	70 (2.2%)	570 (1.8%)	0.13
Any cancer	1068 (12.9%)	910 (14.6%)	7,601 (11.7%)	0.0001

P-values are associated to 3-group comparisons.

(*) percentages are calculated on numbers of included female or male individuals concerned by the location (for example, only female for ovarian cancers)

doi:10.1371/journal.pone.0127363.t001

b) Average fertility characteristics per family

The initial analysis compared whole families where at least one individual was diagnosed with a BRCA mutation to other families. A fertility characteristic for a family corresponds to the average of the characteristic calculated across all its affiliated members.

The miscarriage rate in mutated versus non-mutated families was reduced by 35% ($p = 0.015$), while other fertility parameters were similar (Table 2). In men from mutated versus non-mutated families, the average number of offspring was slightly lower ($p = 0.041$), but the reproductive period was advanced by 7 months, with lower overall mean reproduction age ($p = 0.0013$), earlier first child ($p = 0.0023$) and earlier last child ($p = 0.018$).

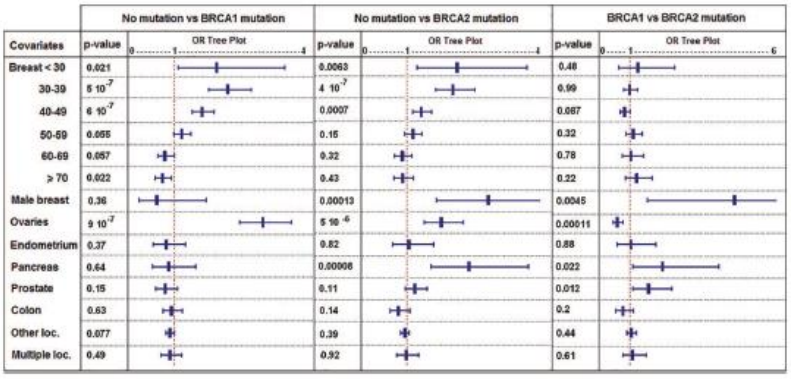


Fig 2. Cancer locations predicting the BRCA mutational risk. (logistic regression: p-values complete the information given by each Odds-Ratio; error bars represent 95%-CI of Odds-Ratios; covariates are cancer locations and "Breast < 30" means female breast cancers occurring before 30 years. . .).

doi:10.1371/journal.pone.0127363.g002

Table 2. Fertility parameters depending on the presence of a BRCA mutation in the family: p-values correspond to tests performed between BRCA1+2 mutated families versus not mutated ones (NM).

Gender	Natality parameter	BRCA1	BRCA2	No mutation (NM)	p-value (BRCA1+2 vs NM)
Affiliated Women	**N =**	3,951	2,978	30,134	
	Number of children	2.6 ± 0.8	2.6 ± 0.8	2.6 ± 0.8	0.61
	Sex ratio of children (M/F)	0.89 ± 0.47	0.89 ± 0.46	0.89 ± 0.53	0.92
	Age at first child	25.1 ± 2.3	24.9 ± 2.4	25.2 ± 2.8	0.38
	Average age at all children	27.5 ± 2.3	27.6 ± 2.4	27.6 ± 2.7	0.55
	Age at last child	30.1 ± 2.8	30.5 ± 3.3	30.3 ± 3.4	0.83
	Miscarriage	0.16 ± 0.62	0.16 ± 0.56	0.25 ± 1.02	**0.015**
	Nulliparous	1,298 (32.9%)	952 (32.0%)	10,819 (35.9%)	0.51
Affiliated Men	**N =**	3,381	2,682	23,357	
	Number of children	2.6 ± 0.9	2.6 ± 1.0	2.7 ± 1.1	**0.041**
	Sex ratio of children (M/F)	0.82 ± 0.51	0.84 ± 0.49	0.82 ± 0.48	0.92
	Age at first child	27.5 ± 3.1	27.5 ± 3.2	28.1 ± 3.7	**0.0023**
	Average age at all children	30.1 ± 2.9	30.1 ± 3.2	30.7 ± 3.7	**0.0013**
	Age at last child	32.7 ± 4.1	32.9 ± 4.1	33.6 ± 4.8	**0.018**
	No child	1 246 (36.9%)	971 (36.2%)	10,133 (43.4%)	0.73

doi:10.1371/journal.pone.0127363.t002

c) Per-family multivariate analysis of natality parameters compared to standard clinical factors predictive for BRCA mutations

We first analysed all natality parameters together in order to extract most important ones. These selected parameters were then added to previous models in order to test if they were independently significant. Two natality factors remained significant when we compared BRCA mutated to non-mutated families: the mean age at fatherhood was lower (p = 0.0028), as was the rate of miscarriages in women (p = 0.021) (Table 3).

The odds ratios given in Table 3 signify that one supplementary year of paternal age diminishes by 6% the chances of belonging to a family with a BRCA mutation, and a 1% increase in the miscarriage rate decreases this probability by 0.2%. Odds ratios for both fertility parameters were significant in the global logistic regression analysis (BRCA1 or BRCA2 vs NM column). However, addition of both parameters to the regression model did not improve the overall predictability, as the area under curve of the associated ROC curve remained almost stable at 0.74 [0.71; 0.77]. Neither parameter differentiated BRCA1 risk from BRCA2 risk.

Table 3. Influence of new natality parameters on the risk for BRCA mutation when analyzed concurrently with cancer locations and age at diagnosis.

	BRCA1 or BRCA2 versus no-mutation	BRCA1 alone versus no-mutation	BRCA2 alone versus no-mutation	BRCA1 versus BRCA2
Men mean age at any birth	0.94 [0.91; 0.98]	0.94 [0.89; 0.99]	0.94 [0.90; 1.00]	1.00 [0.92; 1.08]
	p = 0.0028	p = 0.013	p = 0.038	p = 0.93
Miscarriages	0.80 [0.66; 0.97]	0.81 [0.63; 1.03]	0.81 [0.61; 1.06]	1.12 [0.77; 1.63]
	p = 0.021	p = 0.088	p = 0.12	p = 0.55

First line = Odds-Ratios with 95%-CI; second line = p-value. Usual significant parameters are not reported as they are like in Fig 2.

doi:10.1371/journal.pone.0127363.t003

d) Variation of individual natality characteristics according to known BRCA mutational status

Most members of our pedigrees have not been tested for mutations, notably in oldest generations, although a significant proportion of them must carry the familial mutation. We thus analyzed members with known BRCA mutation status.

Three groups were constituted:

- 583 members tested positive for a BRCA mutation
- 634 members tested negative for a BRCA mutation but belonging to families where a BRCA mutation was found
- 306 members tested negative for a BRCA mutation and belonging to HBOC families where no BRCA mutation was found

The second group is the "ideal" control group (they are very unlikely to carry another mutation favoring cancer) while people belonging to the third group may carry an unknown deleterious mutation (because they were selected for BRCA analysis) that could impact reproductive outcomes. Main differences concerned the average number of children either for men and women (Table 4).

Nulliparity (childlessness) corresponded to individuals without child and aged $\geq$ 40 years, else without child but married/common-law. This age was chosen so that the nulliparity due to young age would not bias outcomes. The rate of childless individuals was reported to the overall number of persons aged over 40 and/or married or common-law with or without children. Childless women were notably rarer among BRCA carriers both when compared to women of second ($p = 0.003$) and of third group ($p = 0.005$)

The reproductive period was slightly longer for women and men carrying a BRCA mutation (respectively 5.9 versus 5.5 years and 6.1 versus 5.6) but without significance. Offspring in females was higher ($p = 0.002$) and was likely related to the excess of nulliparous women (6.7%, $p = 0.019$) among non-carrier family members. Male carriers also had on average more children than non-carrier family members ($p = 0.024$).

Comparison between first and last groups of Table 4 were not detailed: they evidenced no significant difference excepted for the rate of childless women noted above.

Comparisons between groups two and three exhibited almost as many differences as between the first and the second groups. The exception noted above, concerned the rate of childless women that was similar for all non BRCA mutated individuals and close to 16%.

Discussion

BRCA mutations seem to provide fertility advantages that compensate for increased cancer risk and mortality, mainly through an increased number of children, possibly related to a lower rate of childlessness and a longer interval between first and last child. Unexpectedly, fertility differences, calculated on families, were more significant among males than among females. Aside from the childlessness rate, these outcomes confirm those reported by Smith *et al.* in a case-control study of 181 BRCA mutation carriers from 49 kindreds versus 1830 controls [11], all having at least one child, and born before 1930 to avoid the influence of modern birth control. We tested their hypothesis, comparing only persons born before or after 1930, but no major divergence was found between older and more recent cohorts in our population (data not shown). Our results and Smith's partly contradict the conclusions of a study of 96 female mutation carriers, 164 non-carrier cases and 331 controls, which did not show any fertility increase related to BRCA mutations, but which also did not study male fertility [13]. This last

Table 4. Fertility parameters in 1,546 BRCA-tested individuals according to their BRCA mutational status and if they belong or not to a BRCA mutated family.

Gender	Natality parameter	BRCA mutated (1)	p-value (1) vs (2)	Not BRCA mutated but of a mutated family (2)	p-value (2) vs (3)	No known deleterious mutation diagnosed in the family (3)
Women	N tested	583		364		306
	Childless	9.1%	**0.003**	16.0%	0.91	15.7%
	(Nb cases / N')	(46 / 507)		(47 / 293)		(45 / 287)
	Number of children (*)	1.8 ± 1.4	**0.002**	1.5 ± 1.3	**0.0017**	1.8 ± 1.3
	Age at first birth	24.9 ± 4.3	0.97	24.7 ± 4.1	0.94	24.8 ± 4.7
	Mean age at any birth	26.9 ± 4.1	0.32	26.6 ± 4.0	0.57	26.8 ± 4.5
	Age at last birth	29.2 ± 5.2	0.14	28.6 ± 4.9	0.35	29.0 ± 5.3
	Last—first birth (y) (**)	5.9 ± 4.2	0.24	5.5 ± 3.7	0.96	5.6 ± 4.1
	Miscarriages reported	3.1%	0.10	1.4%	0.10	3.4%
Men	N tested	163		119		11
	Childless	11.3%	0.42	14.9%	0.68	9.1%
	(Nb cases / N')	(16 / 141)		(14 / 94)		(1 / 11)
	Number of children (*)	1.7 ± 1.3	**0.024**	1.4 ± 1.3	**0.036**	2.2 ± 1.1
	Age at first birth	26.6 ± 4.0	0.21	27.3 ± 4.2	0.62	26.6 ± 2.4
	Mean age at any birth	28.8 ± 4.1	0.58	29.1 ± 4.0	0.86	29.3 ± 3.5
	Age at last birth	31.3 ± 5.4	0.80	31.1 ± 4.9	0.50	32.2 ± 6.0
	Last—first birth (y) (**)	6.1 ± 4.3	0.22	5.6 ± 4.6	0.59	7.0 ± 5.3
Both gender	**Childless**	9.6%	**0.0029**	18.8%	0.68	15.4%
	(Nb cases / N')	(62 / 648)		(61 / 387)		(46 / 298)
	Number of children (*)	1.7 ± 1.4	**0.00013**	1.5 ± 1.3	**0.00012**	1.8 ± 1.3
	Age at first birth	25.2 ± 4.5	0.49	25.4 ± 4.3	0.20	24.9 ± 4.7
	Mean age at any birth	27.3 ± 4.3	0.70	27.3 ± 4.4	0.71	26.9 ± 4.5
	Age at last birth	29.6 ± 5.3	0.22	29.2 ± 5.0	0.99	29.2 ± 5.4
	Last—first birth (y) (**)	5.9 ± 4.3	0.24	5.5 ± 4.0	0.80	5.7 ± 4.1

p-values correspond to comparisons between columns.

(*) including nulliparous members.

(**) for individuals with at least 2 children with known dates of birth.

N' = number of married/common-law individuals or singles ≥ 40 years old.

(y) years.

doi:10.1371/journal.pone.0127363.t004

study observed a lower male/female ratio for the offspring of female BRCA mutation carriers, which could not be confirmed in our population (data not shown). In a large North American study [14], no fertility differences were found between 2,254 female BRCA carriers and 764 controls from mutated families. But their population was rather recent and young, and the use of contraceptive generalized. The absence of studies concerning male carriers may have hidden their role in the maintenance of deleterious BRCA alleles in the population.

Although observations for tested individuals may appear more reliable than those for whole families, this could be subject to bias. Tested individuals are usually younger than other adults of their families (notably, members of preceding generations are no longer available for direct study). Later generations are more subject to recent birth control measures, as well as social changes in desired family size and delay before having a first child, minimizing small differences in overall reproductive capacity. That is why we also reported statistics based on families including all available generations. The disadvantage of this latter approach is that not all members of BRCA-mutated families are carriers. A similar phenomenon happens in families where no BRCA mutation has been diagnosed. Mutations in other genes are likely to be present in many of these families, and as shown in Table 4, members from the no-familial-mutation group (3) are often closer to BRCA carriers (1) than to non-carrier family members (2): this is in particular true for the average number of children.

Fertility advantages may come from various causes. Some may be strictly biologic, for example a mutation that could play a role on sex hormones production, on an earlier onset of fertility, on the sperm quality, or that could limit the in-utero rejection of a malformed embryo. Apart from the evident influences of the cultural context, it is also possible that some mutated genes could impact behavioral aspects that modify *in fine* the reproductive outcomes. This is why we studied various dimensions of the fertility, in particular the onset and the duration of the reproduction period, the miscarriage rate, and the "celibacy" rate (equivalent to the childlessness rate in our study as 163 of the 169 childless BRCA tested individuals were single). We detail this different factors.

The fertility advantages we observed in mutation carriers occur earlier in life than the age at which breast or ovarian cancer usually develops, and they play a protective role against cancer, since a high number of children and an earlier first child for females are known to reduce breast cancer risk, partly in relation to breast-feeding duration [22–24]. This may explain why differences are stronger in males than in females as these adjustments do not exist for the former. This point is also important for males as later births are more exposed to *de novo* germline mutations, thus probably to congenital malformations [3].

The lower proportion of childless carriers was also an interesting result, and does not seem attributable to a difference in the way pedigrees were collected because we studied only families including at least five at-risk members. If confirmed, this may indicate that genetic profiles, besides cultural environment, could influence reproductive behavior. One might suggest that individual reproductive behaviors may already have changed in our population, with individuals fearing to transmit deleterious mutations to their descendants. Most probands, however, consulted the oncogeneticist because they had cancer and few because of a family history of cancer. At their initial consultation, probands had generally completed their reproductive period (age at consultation 50 ± 13 years). Genetic consultation and mutation testing during the reproductive period was not available to older generations, and thus mutation status could not have conciously influenced their reproductive choices. With the advent of widespread BRCA testing for women and families at risk, current and future generations may however incorporate mutation status into their family planning.

Women affiliated to BRCA mutated families underwent 1/3 fewer miscarriages, in contrast to the observation of a matched case-control study including 3,485 BRCA carriers or non-carriers [25], where no difference in the rate of spontaneous abortions was found. Reporting of miscarriages is surely incomplete in our study, and no distinction was made between spontaneous abortion and stillbirth. In Europe, spontaneous abortion affects about 15 to 20% of all pregnancies [26–27]. In a great majority of cases, this event happens in the first weeks of pregnancy and may often not be perceived by parents. Reporting thus was likely to be heavily weighted toward stillbirths, defined as the death of the fetus during the third trimester of pregnancy [28].

The standard rate of stillbirths is evaluated at around 3% of all pregnancies [27]. We should thus expect about 2,090 stillbirths in our population instead of the 714 reported. Nevertheless, as this reporting takes place during pedigree building, before BRCA status is known, all groups were treated similarly and under-reporting should not be biased toward one or another group. Finally, the observed difference in family miscarriage rate could not be confirmed in directly tested individuals, because of their insufficient numbers.

The standard clinical parameters predicting BRCA mutations were validated in our pedigrees, showing our database and results are consistent with published data. The very large size of the database gave power and accuracy to the study. More than 12% of the regional population is included in the database, with pedigree data collected over decades in the same manner by the same geneticists. The limit imposed for this study of only analysing families with >5 members further contributed to the homogeneity of the study group, as most immigrant families were excluded by this critereon.

To conclude, BRCA mutations that survive selection pressure seem to provide significant fertility advantages. Fertility parameters should thus be considered as a novel source of data for future population research, in particular to shed new light on possible biological mechanisms of reproductive physiology. Also, it could help characterize new subgroups among families at cancer risk but where no BRCA mutation has been diagnosed, in particular distinguishing families where reproduction starts early and those where fertility advantage comes from a lower rate of miscarriage or a higher average number of offspring. These criterea may be useful for stratifying data produced in large-scale genomic analyses of low-penetrance genes.

Supporting Information

S1 Data.
(ZIP)

Author Contributions

Conceived and designed the experiments: FK MA NU. Analyzed the data: MA FK. Contributed reagents/materials/analysis tools: YJB. Wrote the paper: FK MA NU CL YJB YB. Oncogenetics Chief Manager: YJB. Independent local review: YB.

References

1. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant.or not? Hum Mol Genet. 2002; 11: 2417–2423 PMID: 12351577
2. Plon SE, Nathanson K. Inherited susceptibility for pediatric cancer. Cancer J. Sudbury Mass. 2005; 11: 255–267. PMID: 16197716
3. Van der Luijt RB, van Zon PH, Jansen RP, van der Sijs-Bos CJ, Wárlám-Rodenhuis CC, Ausems MG. De novo recurrent germline mutation of the BRCA2 gene in a patient with early onset breast cancer. J Med Genet 2001; 38: 102–105. PMID: 11158174
4. Diez O, Gutiérrez-Enríquez S, Mediano C, Masas M, Saura C, Gadea N, et al. A novel de novo BRCA2 mutation of paternal origin identified in a Spanish woman with early onset bilateral breast cancer. Breast Cancer Res Treat. 2010; 121: 221–225. doi: 10.1007/s10549-009-0494-y PMID: 19649703
5. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012; 488(7412): 471–5 doi: 10.1038/nature11396 PMID: 22914163
6. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci USA. 2010; 107(3): 961–8 doi: 10.1073/pnas.0912629107 PMID: 20080596
7. Caputo S, Benboudjema L, Sinilnikova O, Rouleau1 E, Béroud C and the French BRCA GGC Consortium. Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. Nucleic Acids Res. 2012; 40: D992–1002. doi: 10.1093/nar/gkr1160 PMID: 22144684

8. Im KM, Kirchhoff T, Wang X, Green T, Chow CY, Vijai J, et al. Haplotype structure in Ashkenazi Jewish BRCA1 and BRCA2 mutation carriers. Hum. Genet. 2011; 130: 685–699. doi: 10.1007/s00439-011-1003-z PMID: 21597964
9. Harboe TL, Eiberg H, Kern P, Ejlertsen B, Nedergaard L, Timmermans-Wielenga V, et al. A high frequent BRCA1 founder mutation identified in the Greenlandic population. Fam. Cancer 2009; 8: 413–419. doi: 10.1007/s10689-009-9257-5 PMID: 19504351
10. Laraqui A, Uhrhammer N, Lahlou-Amine I, Rhaffouli H, Baghdadi J, Dehayni M, et al. Mutation screening of the BRCA1 gene in early onset and familial breast/ovarian cancer in Moroccan population. Int. J. Med. Sci. 2013; 10: 60–67. doi: 10.7150/ijms.5014 PMID: 23289006
11. Smith KR, Hanson HA, Mineau GP, Buys SS. Effects of BRCA1 and BRCA2 mutations on female fertility. Proc. Biol. Sci. 2012; 279: 1389–1395. doi: 10.1098/rspb.2011.1697 PMID: 21993507
12. Oktay K, Kim JY, Barad D, Babayev SN. Association of BRCA1 mutations with occult primary ovarian insufficiency: a possible explanation for the link between infertility and breast/ovarian cancer risks. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 2010; 28: 240–244.
13. Moslehi R, Singh R, Lessner L, Friedman JM. Impact of BRCA mutations on female fertility and offspring sex ratio. Am. J. Hum. Biol. Off. J. Hum. Biol. Counc. 2010; 22: 201–205.
14. Pal T, Keefe D, Sun P, Narod SA. and the Hereditary Breast Cancer Clinical Study Group. Fertility in women with BRCA mutations: a case-control study. Fertil. Steril. 2010; 93: 1805–1808. doi: 10.1016/j.fertnstert.2008.12.052 PMID: 19200971
15. Kwiatkowski F, Girard M, Hacene K, Berlie J. Sem: a suitable statistical software adaptated for research in oncology. Bull. Cancer (Paris) 2000; 87: 715–721. PMID: 11084535
16. Parmigiani G, Berry D, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. Am. J. Hum. Genet. 1998; 62: 145–158. PMID: 9443863
17. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat. Med. 2004; 23: 1111–1130. PMID: 15057881
18. Eisinger F, Bressac B, Castaigne D, Cottu PH, Lansac J, Lefranc JP, et al. Identification and management of hereditary predisposition to cancer of the breast and the ovary (update 2004). Bull. Cancer (Paris) 2004; 91: 219–237. PMID: 15171047
19. Bonaïti B, Alarcon F, Bonadona V, Pennec S, Andrieu N, Stoppa-Lyonnet D, et al. A new scoring system for the diagnosis of BRCA1/2 associated breast-ovarian cancer predisposition. Bull. Cancer (Paris) 2011; 98: 779–795. doi: 10.1684/bdc.2011.1397 PMID: 21708517
20. Biswas S, Tankhiwale N, Blackford A, Gutierrez Barrera AM, Ready K, Lu K, et al. Assessing the added value of breast tumor markers in genetic risk prediction model BRCAPRO. Breast Cancer Res. Treat. 2012; 133: 347–355. doi: 10.1007/s10549-012-1958-z PMID: 22270937
21. Fischer C, Kuchenbäcker K, Engel C, Zachariae S, Rhiem K, Meindl A, et al. Evaluating the performance of the breast cancer genetic risk models BOADICEA, IBIS, BRCAPRO and Claus for predicting BRCA1/2 mutation carrier probabilities: a study based on 7352 families from the German Hereditary Breast and Ovarian Cancer Consortium. J. Med. Genet. 2013; 50: 360–367. doi: 10.1136/jmedgenet-2012-101415 PMID: 23564750
22. Layde PM, Webster LA, Baughman AL, Wingo PA, Rubin GL, Ory HW. The independent associations of parity, age at first full term pregnancy, and duration of breastfeeding with the risk of breast cancer. Cancer and Steroid Hormone Study Group. J. Clin. Epidemiol. 1989; 42: 963–973. PMID: 2681548
23. Hinkula M, Pukkala E, Kyyrönen P, Kauppila A. Grand multiparity and incidence of endometrial cancer: a population-based study in Finland. Int. J. Cancer J. Int. Cancer 2002; 98: 912–915 (2002). PMID: 11948472
24. Russo J, Hu YF, Yang X, Russo IH. Developmental, cellular, and molecular basis of human breast cancer. J. Natl. Cancer Inst. Monogr. 2000: 17–37. PMID: 10963618
25. Friedman E, Kotsopoulos J, Lubinski J, Lynch HT, Ghadirian P, Neuhausen SL, et al. Spontaneous and therapeutic abortions and the risk of breast cancer among BRCA mutation carriers. Breast Cancer Res. BCR 2006; 8: R15.
26. Buss L, Tolstrup J, Munk C, Bergholt T, Ottesen B, Grønbaek M, et al. Spontaneous abortion: a prospective cohort study of younger women from the general population in Denmark. Validation, occurrence and risk determinants. Acta Obstet. Gynecol. Scand. 2006; 85: 467–475. PMID: 16612710
27. Zinaman MJ, Clegg ED, Brown CC, O'Connor J, Selevan SG. Estimates of human fertility and pregnancy loss. Fertil. Steril. 1996; 65: 503–509 (1996). PMID: 8774277
28. Lawn JE, Blencowe H, Pattinson R, Cousens S, Kumar R, Ibiebele I, et al. Stillbirths: Where? When? Why? How to make the data count? Lancet 2011; 377: 1448–1463. doi: 10.1016/S0140-6736(10)62187-3 PMID: 21496911

8.2.2 JJBP (2016) From Oncogenetic Pedigrees to Family Profiles: A Necessary Step to Enable Statistics

JACOBS PUBLISHERS

OPEN ACCESS

Jacobs Journal of Bioinformatics and Proteomics

Research Article

From Oncogenetic Pedigrees to Family Profiles: A Necessary Step to Enable Statistics

Arbre M[1], Kwiatkowski F[1,2*], Serlet L[2], Bignon YJ[1*]

[1]Centre Jean Perrin, Laboratoire d'Oncologie Moléculaire, 58, rue Montalembert, 63011 Clermont-Ferrand, France

[2]Université Blaise Pascal, 24, avenue des Landais, BP 80026, 63171 Aubière Cedex, France

*Corresponding author: Dr. Kwiatkowski F, Centre Jean Perrin, Laboratoire d'Oncologie Moléculaire, 58, rue Montalembert, 63011 Clermont-Ferrand, France, Email:Fabrice.Kwiatkowski@CJP.fr

Received: 12-15-2015

Accepted: 05-08-2016

Published:

Background: Cancer has always been a major domain requiring progress in statistics, methodology and bio-informatics. Oncogenetic, focusing on the relationship between genetics and cancer, is particularly concerned with "big data" issues, which includes genealogical pedigrees: their special structure – made of relations between members and possible clinical annotations - is too complex to be directly used for statistical purpose. This article describes a way to condense pedigrees so that they can be handled more easily and compared together.

Method: our approach aggregates the genealogical and clinical information of pedigrees containing many generations. Condensed pedigrees, called "subtrees", are composed of basic 2 or 3-generation pedigrees: for one whole pedigree, a subtree is calculated by the mean of all basic pedigrees it contains. These subtrees can then be grouped together for different subsets of families (for example breast/ovarian cancer families with or without BRCA mutation carrier). Such a grouping named "profile", besides its reduced structure, is particularly interesting because for each studied characteristic, means and standard deviations are available. Moreover, distances between each subtree and various profiles can be calculated and used as a discriminant index.

Results: Subtrees and profiles were validated using a subset of 454 families (22.348 members) with a Lynch syndrome: in 84, at least one member carried an MMR deleterious mutation. Two profiles were computed depending on the presence or the absence of MMR mutation in the families. An ROC analysis showed that distances between each family subtree and both profiles were significant predictors for MMR mutations.

Conclusion: Subtrees and profiles show interesting discriminant properties to study pedigree data. This method seems suitable to search for population differences between monogenic cancer risk models and multigenic ones.

Keywords: Pedigree; Oncogenetic; Genealogy; Modeling; Subtree

Introduction

Currently, with the progress of genetic research, more and more predispositions to hereditary diseases are discovered. As pangenomic analysis (genome-wide screening) cannot be realized routinely - partly for ethical reasons - it is necessary to predict which genes are the most likely to be mutated and then perform targeted genomic analysis.

In the oncogenetic routine, pedigrees are frequently used to diagnose hereditary predispositions. These contain two kinds of data: first the genealogy, i.e. the relations between members from which for example fertility and mortality parameters can be calculated and second, possible clinical information that may characterize the phenotype of a hereditary predisposition to a disease. Overall, both types of information are necessary

Cite this article: Kwiatkowski. From Oncogenetic Pedigrees to Family Profiles: A Necessary Step to Enable Statistics . J J Bioinform Proteom. 2016, 1(1): 005.

for the discovery of new deleterious mutations. Indeed, they enable to isolate pedigrees with special characteristics like the occurrence of typical cancer locations. Once this step is achieved, gene analysis on available DNA samples can be performed with increased chances to point out one or more mutations possibly responsible for the phenotype.

Unfortunately, oncogenetic pedigrees are usually too complex to be analyzed other than visually. To solve this issue, some authors limit their inquiry to smaller pedigrees (only 3 to 4 generations) [1]: although easier, this solution appears in fine deleterious because of the loss of phenotypic information that is spread all over generations. Another way to evaluate cancer or mutational risks of one family is to calculate scores. Concerning breast/ovarian cancer risk, different authors have developed indexes upon pedigrees such as Manchester [2], Eisinger [3] or BRCAPRO [4]. This kind of index combines clinical parameters included in pedigrees that, after a logistic regression, have kept a sufficient significance (i.e. for Manchester score, only breast, ovarian, prostatic, pancreatic cancers reported in the family are used). But these methods only concern a reduced set of familial predispositions and limit the analysis to only a small part of pedigree data, mainly the occurrence of cancers within the family. We have exten ded such a research to uncommon information. It enabled us to conclude that fertility parameters could also help predict these risks [5]. However, much work still remains because first, the indexes available to calculate the familial risk of mutation are only adapted to breast/ovarian cancers and second, because known mutations account for only a minority of cancer predispositions: improving the way to select specific sub-groups of families is thus a necessary step for narrowing the research of new deleterious mutations to a reduced set of genes.

In the literature, if we except indexes, no efficient methodology enables to a group or compare pedigrees. Specialized pedigree software exists but they concern animals [6,7] and they focus on the inbreeding level.

The approach proposed in the next chapter is a modeling of pedigrees into "subtrees", i.e. family representations condensed into two or three generations by aggregating the information of all family members from all generations. How to create profiles (global subtree for several families together presenting similar characteristics) is the matter of another paragraph. Finally, the use of profiles and subtrees will be demonstrated within a sample of 454 families at colon cancer risk extracted from the database of the Oncogenetic Department of the comprehensive anticancer Center Jean Perrin.

Methods

Description of pedigrees

A specialized function of the SEM software [8] has been developed to automatically shape the pedigrees (Figure 1).

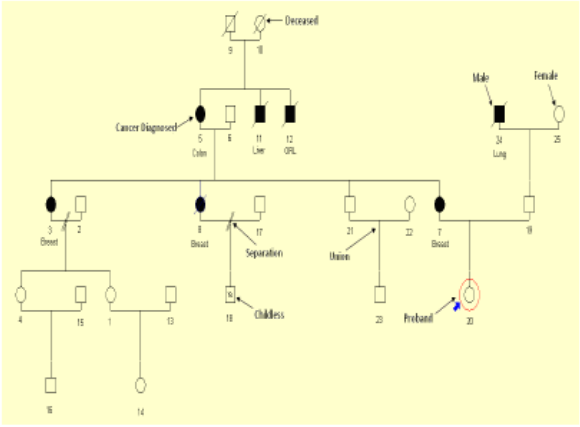


Figure 1. Example of a simple pedigree (males are represented by a square and females by a circle. Every striped symbol represents a deceased individual and ones filled with black indicate a cancer).

Most symbols used to draw pedigree are common and have been recommended in the pedigree standardized nomenclature edited by Bennett et al. [9,10]. The «proband» is the person who requests the creation of the medical file. The proband, in the example (Figure1), is a woman: she is represented by a circle and pointed by a blue arrow. The pedigree is shaped with all the individuals who are related to her. Throughout this article, we will use the family represented in Figure 1 as an example. The Jean Perrin Center database contains families that include sometimes more than 600 members, consequently visual analysis becomes difficult and new representation types are necessary.

Modeling subtree method

The underlying structure of a pedigree is a reduced 2 or 3-generation pedigree that cumulates the information from all family members. These structures must be distinguished from pedigree branches (for example the paternal and maternal branches of a proband): branches try to isolate members of a pedigree carrying a particular genotype while subtrees gather reduced patterns that occur several times within a pedigree or a branch. Three models are considered: a 2-generation subtree, a detailed 2-generation subtree, and a 3-generation subtree.

2-generation subtree

The way to constitute this 2-generation subtree is to find each pair of [mother or father]/ [son and daughter]. Male and female headers (parents) are separated because men and women are not exposed to the same cancer risk. With these pairs, all the information needed for the construction of the 2-generation subtrees is available in the database and can be collected and aggregated as many times as pairs are available.

From the proband, we can find his/her parents, then the parents of proband's parents, and so forth. Once members at the top of the pedigree have been identified (numbered 9, 10, 24 and 25 in Figure 1), we can browse down the pedigree to keep only genetically related members: children of top members can be selected, then children's children and so on until the most recent generation.

This pruning process excludes a few members who are not supposed to bring genetic information about the cancer risk. Else their presence would "feed the background noise", and increase uselessly the overall variability (i.e. lower the precision of estimates). Pruned members are:

- all spouses who do not provide information about their parents (numbers 2, 6, 13, 15, 17 and 22)

- Childless members (numbers 11, 12 and 18)

- Latest generation members, they are usually too young to have children (numbers 14, 16, 20 and 23)

Finally, each pair "parent/child" can be deduced, keeping in mind that one person can be used as a parent as well as the child.

Once the individuals constituting subtrees are identified, all useful information is collected from clinical data registered in the database. The combination of the information for each item of the 2-generation subtree ends up with 6 composite "family members". They cumulate the following information:

- number of female headers
- number of male headers
- number of males from the female header
- number of females from the female header
- number of males from male header
- number of females from the male header

- number and patient's age at the diagnosis of following cancers:
  - breast male and female
  - ovarian
  - endometer
  - prostatic
  - colon
  - pancreatic
  - other cancers (cumulative)
  - number of members without cancers

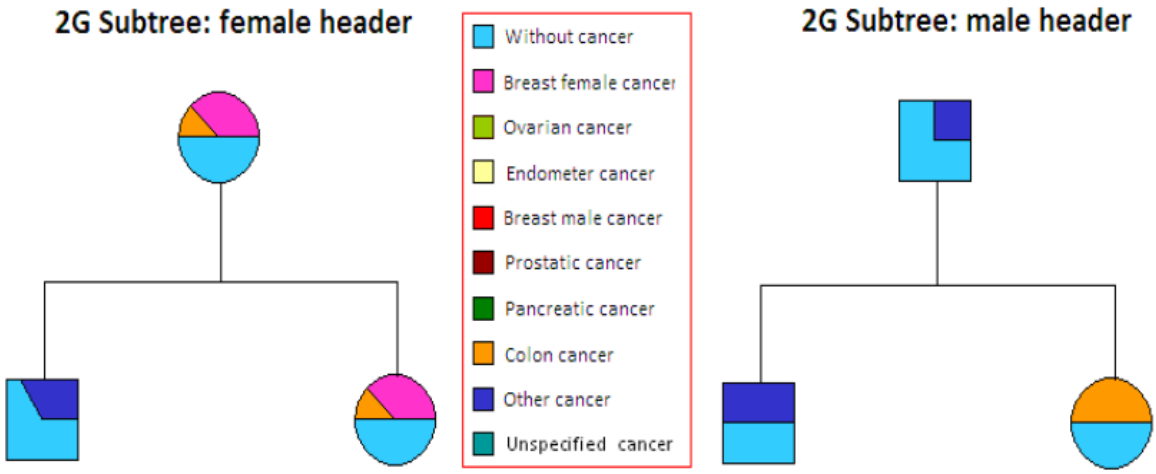


Figure 2. Basic structure of 2-generation subtree built from the pedigree of Figure 1.

With this selection process, four male headers (numbers 9, 19, 21 and 24) and eight female headers (numbers 1, 3, 4, 5, 7, 8, 10, 25) are identified. Twelve 2-generation subtrees are thus available into the pedigree of figure 1 and a resulting subtree can be built. Basic 2-generation subtree is shaped as in figure 2.

Although this list includes already 104 variables, it can be extended if needed.

With this 2-generation subtree, all the information is condensed whatever the size of the family and it becomes easier to compare 2 or more families. The proportion of cancers by location is represented by a pie chart within circles

and squares at each level (Figure 2):

For the pedigree of figure 1, following characteristics are calculated by the software:

- 8 female headers: 3 breast cancer (occurring in average at 49 years), 1 colon cancer (73 years) diagnosed and 50% without cancer

• 6 male children: 1 liver cancer and 1 ORL cancer, so 2 "other cancer" (70 years) diagnosed and 67% without cancer

• 8 female children: 3 breast cancer (49 years), 1 colon cancer (73 years) diagnosed and 50% without cancer

- 4 male headers: 1 lung cancer (48 years) diagnosed and 75% without cancer

• 4 male children: 1 liver cancer and 1 ORL cancer so 2 "other cancer" (70 years) diagnosed and 50% without cancer

• 2 female children: 1 colon cancer(73 years) diagnosed and 50% without cancer

for most families. Another data is also interesting: childless members and miscarriages, which are also included in this detailed representation.

The member selection process does not differ from the one used for the 2-generation subtree. The same exclusions apply here and headers remain unchanged.

Figure 3 exhibits the basic pattern concerning figure-1 pedigree: circles still represent females and squares males. For children, the 4 first vertical lines correspond to children's rank (born first, born second...) and miscarriages (if any) are positioned at the 5th rank using a lozenge (none in fig. 3). The size of squares and circles is proportional to the number of children for each item. The length of the vertical lines connecting parents to children depends on the parental mean age at their children's birth. Proportions of childless adults have represented aside headers using squares and circles with a cross. Proportions of persons diagnosed with a cancer are represented underneath for the children with the same color-code as in figure 2.

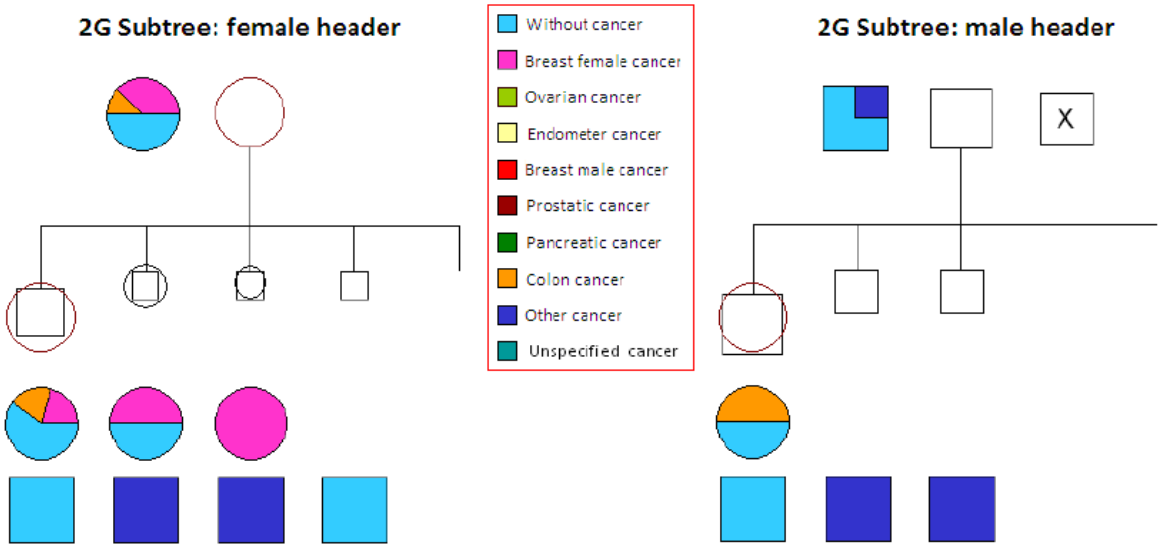


Figure 3. Example of a 2-generation subtree with detailed information by children's rank.

2-generation subtree with details about children

One might wonder if the birth rank may influence the risk for particular events (example, congenital malformations). This rank is available if dates of birth are known and the number of 1st boys, 1st girls, 2nd boys, etc. can be computed per header. Only four children of each gender are retained, this enables to include a maximum of 8 children which is usually enough

3-generation subtree

To highlight possible "variations" of intergenerational cancer transmission, we decided to shape a 3-generation subtree. This 3rd synthetic representation includes 3 generations instead of 2: triplets are now identified, with parents/children/grand-children for both genders at each level. The

same process is applied to the members' selection and the same information as previous 2-generation subtrees are collected. Figure 4 shows the basic structure of such a representation drawn for our family example.

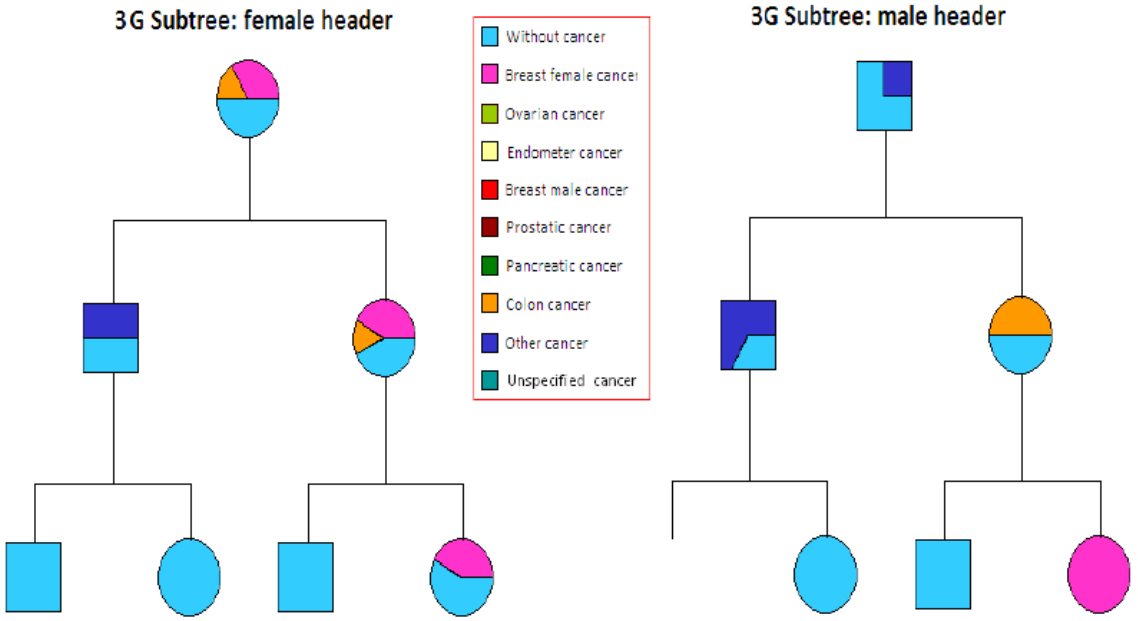


Figure 4. Basic structure of 3-generation subtree from the family of figure 1.

Two recent articles [11,12] have reported that cancers, in mutated families, tended to appear at an earlier age over generations, i.e. daughters had breast cancers sooner than their mothers. Narod [13] suggested this could happen because daughters' exposure time is necessary shorter and late cancers have not enough time to occur. We thus decided to

Figure 3. Example of a 2-generation subtree with detailed information by children's rank.

2-generation subtree with details about children

One might wonder if the birth rank may influence the risk for particular events (example, congenital malformations). This rank is available if dates of birth are known and the number of 1st boys, 1st girls, 2nd boys, etc. can be computed per header. Only four children of each gender are retained, this enables to include a maximum of 8 children which is usually enough

Combining several subtrees to create group profiles

A new concept needs to be introduced if several subtrees are to be grouped together in order to constitute "family profiles" or "group profiles". Concerned families are selected when they present particular characteristics. For example, one might want to design a specific profile for BRCA mutated families, another one focusing families with several lung cancers, or in a completely different domain, families where several suicides are reported, and so on.

3-generation subtree

To highlight possible "variations" of intergenerational cancer transmission, we decided to shape a 3-generation subtree. This 3rd synthetic representation includes 3 generations instead of 2: triplets are now identified, with parents/children/grand-children for both genders at each level. The

viation are calculated for each variable (i.e. personal, familial or clinical characteristic of interest) and registered in a new table of the database. The set of averages and variances per profile corresponds to a multidimensional object which can be represented by a barycenter surrounded by a "cloud of points".

This group of distribution parameters can then be used to realize statistics, to compare several profiles, to calculate the distance between them and a new subtree and also identify particular families' subset.

Statistical considerations

Several statistical tests are used in this study. Distribution parameters (mean and standard deviation) characterize numerical data and numbers / frequencies categorical variables. Best cutoffs optimizing sensitivity and specificity of predictive parameters are calculated using an ROC analysis [14] while the performance of associated ROC curves is evaluated using their area under the curve (AUC) [15]. To build a score predictive for MMR mutation based on standard parameters cited in the literature, a logistic regression model was performed. The corresponding predictive score was calculated using its regression formula.

Results

An example of the use of subtrees and profiles is detailed hereafter. We first describe some characteristics of two profiles and then, we explain how these profiles can help predict the mutational risk.

Description of the family set (454 Lynch syndrome)

The accrual in our pedigree database started in 1988. Today it contains 6,500 families including over 190,000 individuals with clinical information (family diagnosis, mutated gene if any...). Most of these families correspond to a breast/ovarian cancer risk. Another important group represents the Lynch syndrome (or HNPCC = hereditary nonpolyposis colorectal cancer). Less than 20% of families diagnosed with this syndrome will present a mutation in APC gene or one of main MMR genes (MLH1, MSH2, MSH6, PMS2). Thus, even using NGS analyzers, a systematic sequencing of all 5 genes is not relevant. Up to now in Lynch syndrome, no good algorithm developed using pedigree information can predict with a good accuracy the mutation probability [16]. Two main algorithms exist, but they have weak predictive properties: Amsterdam index [17,18] with sensitivity around 80% and specificity of 46% and 68% across studies, and the revised Bethesda index [19] associated with a 89% sensitivity and 58% specificity; to increase the prediction strength of these indexes, two complementary tests can be performed on blood sample: an Immuno-Histo-Chemical test (IHC) and a microsatellite instability test (MSI). They enable to bring up sensitivity and specificity to values close to 100% but they are not cheap. We thus decided to check if profiles could help us develop a new strategy and enable to avoid the intermediate use of IHC/MSI test.

Two profiles have been calculated among families presenting with a Lynch syndrome. The first profile corresponded to 84 mutated families, with at least one member diagnosed with a deleterious mutation in APC gene or on a "mismatch repair" gene. The second included 370 families without any member diagnosed with such mutations. All families needed also to contain at least 10 known members to be sufficiently informative. Respectively 4,218 and 18,130 individuals belonged to these two profiles.

Means and SD were computed for about 100 parameters, respectively 23 and 21 per "synthetic" subtree mother or father and 14 or 12 per "synthetic" daughter or son (mothers' daughters, mothers' sons, fathers' daughters, fathers' sons) and 8 familial fertility scores: the number of features differed by gender because some cancers are gender-specific (prostate, ovaries) and fertility parameters are calculated only for subtree headers. Some of these features are presented in table 1 (mainly of the female header).

Obvious differences can be noticed between both profiles, in particular regarding ages at colon cancer diagnosis for both mothers and fathers (table 1). Cancer frequency is also doubled in fathers if a known deleterious mutation is diagnosed in their family.

Distance calculation between a profile and a subtree

Profiles enable statistical computations. A first method is to calculate distances (Fig. 5) between profiles and a new family (i.e. a subtree), in order to find the nearest one. Profiles can be represented as a cloud with a barycenter (average) and a width (using standard deviations).

The spreading of the cloud can be figured by a disk and its radius by a double arrow between the center of the cloud and the edge. A new subtree corresponds to a new cloud which standard deviation is null, thus a point. Two kinds of distance were envisaged: Euclidean and correlation coefficient.

Several measures are possible:

- D = Distance between the center of the cloud and the new family (Figure 5, double arrow between the center of the cloud for profiles 1 and 2 and Subtree X)

- d = distance between the extremity of the cloud and the new family (Figure 5, the double arrow between the extremity of clouds (profiles 1 and 2) and Subtree X)

- R = ratio between the distance D and the associated cloud spreading (Standard Deviation) = D / SD

	Profile 1 (N = 84) MMR mutated families		Profile 2 (N = 370) Families without MMR mutation	
Label	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Header Female (number per subtree)	7.94	5.10	7.43	5.54
Mothers' sons (N)	1.18	3.76	1.14	3.90
Mothers' daughters (N)	1.08	3.50	1.17	3.98
Childless women (N)	2.45	3.09	2.77	3.01
Mean mothers' age at daughters' birth	27.55	5.93	28.04	6.50
Mean mothers' age at sons' birth	27.36	6.24	28.24	6.57
Mothers' age at 1st daughter's birth	25.11	5.23	25.64	5.79
Mothers' age at 1st son's birth	25.58	5.52	26.09	6.21
Mothers' colon cancer frequency	0.13	0.53	0.09	0.39
Mothers' age at colon cancer	53.92	15.76	58.14	16.23
Mothers' breast cancer frequency	0.07	0.43	0.06	0.35
Mothers' ovarian cancer frequency	0.02	0.14	0.01	0.12
Mothers' pancreatic cancer frequency	0.01	0.09	0.01	0.09
Mothers' endometer cancer frequency	0.07	0.42	0.02	0.22
Other cancer frequency in mothers	0.15	0.68	0.13	0.58
Not documented cancers in mothers	0.08	0.43	0.07	0.37
Fathers' colon cancer frequency	0.21	0.80	0.10	0.44
Fathers' age at colon cancer	50.82	14.30	59.56	13.21

Table 1. Example of characteristics (among 104 available) calculated for 2 groups of families at colorectal cancer risk.

A previous comparison of the predictive values for BRCA mutations of each calculation mode among a very large sample of breast/ovarian cancer-prone families showed that the first Euclidean distance D performed slightly better than other methods (results not shown). We used the ratio D1/D2 to study its discriminant power for MMR mutations. An ROC analysis was performed to compare this result with a logistic regression calculated on best known significant clinical predictors. Figure 6 presents the two results:

The ROC analysis calculated using the ratio of Euclidean distances between subtrees and both profiles is associated with a good AUC (area under the curve) = 0.76 [0.70; 0.81], a 71% sensitivity and a 72% specificity. The positive predictive value (PPV) is limited = 38% while the negative one (NPV) is rather high = 91%. Overall 72% of families are well classified (70% of mutated families and 72% of not mutated ones). Prediction of mutation by the logistic regression selects only 5 clinical parameters calculated per the whole family (independently from filiation): the number of colon cancers, lower age at colon cancer, the number of endometrial cancers, prostatic cancers,and multiple cancers, this latter parameter diminishing the likelihood of an MMR mutation. The regression formula

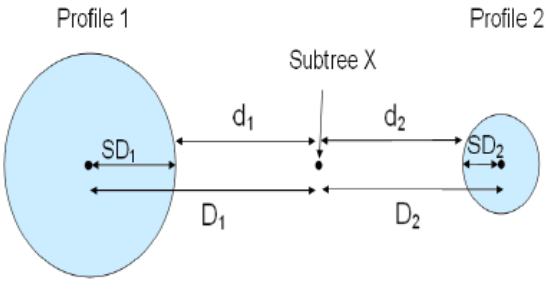


Figure 5. Distances between profiles and a new family X.

associated with these clinical factors yields a slightly better ROC curve (blue curve in figure 6, difference $p < 0.01$): AUC = 8.5 [0.79; 0.90], sensitivity = 80%, specificity = 80%, PPV = 48% and NPV = 95%. Well, the classified rate is 80% overall and for each subgroup. Despite the superiority of the well-adjusted regression model, profiles that require neither selection nor hypothesis on covariates, appear to possess interesting discriminant properties with a fair ROC AUC (> 0.70).

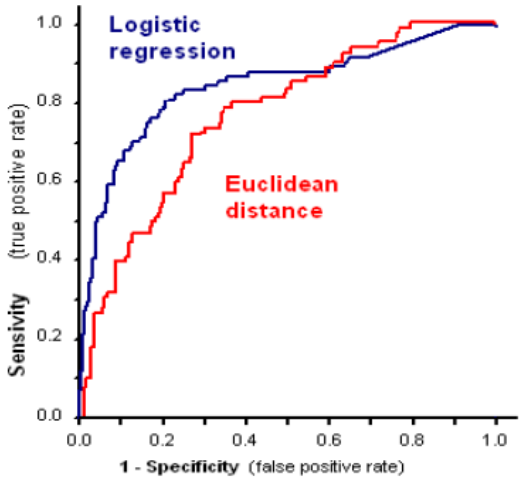


Figure 6. ROC curves comparing the predictive value for MMR mutations of best regression model and the ratio. of Euclidian distance between profiles (difference between curves: $p < 0.01$).

Discussion

Pedigrees used in oncogenetic contain a large amount of clinical and biological information. Besides, large pedigrees provide complementary information, in particular regarding natality/fertility. This approach by "subtrees" represents a helpful solution to use more widely all available data whatever the size of pedigrees. Standardized and synthetic subtrees allow performing statistics on pedigrees, to build standard profiles according to specific characteristics and give indications about familial mutation risk. With the creation of profiles, the comparison between a new family (a single subtree) and various profiles becomes possible. This approach in our example concerning HNPCC predisposition, although not optimal, enabled to well classify most members carrying an MMR mutation without requiring hypothesis and/or restriction about selected criteria.

In the future, geneticists could gain time trying to "categorize" families with this method: they could be more specific when choosing which gene to sequence. We intend to test how the use of subtrees and profiles may help confirm or contradict, within our breast/ovaries cancer-prone families, a hypothesis regarding a monogenic or multigenic etiology.

A current weakness of our computer program is that it is only compatible with SEM software, used almost exclusively in the Jean Perrin Comprehensive Cancer Center. It should be re-developed for different working environments. The Microsoft Visual Basic source code is available on request to the corresponding author.

The purpose of our work was to contribute to the study of familial risks for any type of cancer, in relation to known or unknown deleterious mutations. Of course, such an approach may also be considered for other purposes than mutational risk prediction.

Competing interest

We declare no competing interests: this work was supported by FEDER (European Funding of Regional Development) and Conseil Régional Auvergne (France).

Authors' contributions

Article draft: Marie Arbre, Fabrice Kwiatkowski

Article revising: Pr Yves-Jean Bignon, Pr Laurent Serlet

Project responsible: Pr Yves-Jean Bignon

Subtrees software development: Marie Arbre

Pedigrees software development: Fabrice Kwiatkowski

Mathematical contribution: Pr Laurent Serlet

Statistical analysis: Marie Arbre, Fabrice Kwiatkowski

Acknowledgments

Claire Laquet, the oncogenetic counselor, Laurence Boulègue, Sandrine Casteker, Mélanie Teurio and Sandra Charbonnier, secretaries of Oncogenetics Department.

References

1. Glazner C, Thompson EA. Improving Pedigree-based Linkage Analysis by Estimating Coancestry Among Families. Stat Appl Genet Mol Biol. 2006, 11(2): pii.

2. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med. 2004, 23: 1111-1130.

3. Eisinger F, Bressac B, Castaigne D, Cottu PH, Lansac J et al. Identification and management of hereditary predisposition to cancer of the breast and the ovary (update 2004). Bull Cancer. 2004, 91(3): 219-237.

4. Evans DG, Eccles DM, Rahman N, Young K, Bulman M et al. A new scoring system for the chances of identifying a BRCA1/2 mutation outperforms existing models including BRCAPRO. J Med Genet. 2004, 41: 474-480.

5. Kwiatkowski F, Arbre M, Laquet C, Uhrhammer N, Bignon YJ. BRCA mutations increase fertility in families at hereditary breast/ovarian cancer risk. Plos One. 2015, 10(6): e0127363.

6. Mc Parland S, Kearney JF, Rath M, Berry DP. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. J Anim Sci. 2007, 85(2): 322-331.

7. Cole JB. PyPedal: A computer program for pedigree analysis. Comp Electro. Agric. 2007, 57(1): 107-113.

8. Kwiatkowski F, Girard M, Hacene K, Berlie J. Sem: A suitable statistical software adapted for research in oncology. Bull Cancer. 2000, 87(10): 715-721.

9. Bennett RL, Steinhaus KA, Uhrich SB, O'Sullivan CK, Resta RG et al. Recommendations for Standardized Human, Pedigree Nomenclature. Am J Hum Genet. 1995, 56(3): 745-752.

10. Bennett RL, French KS, Resta RG, Doyle DL. Standardized human pedigree nomenclature: update and assessment of the recommendations of the National Society of Genetic Counselors. J Genet Couns. 2008, 17(5): 424-433.

11. Martinez-Delgado B1, Yanowsky K, Inglada-Perez L, Domingo S, Urioste M et al. Genetic anticipation is associated with telomere shortening in hereditary breast cancer. PLoS Genet. 2011, 7(7): e1002182.

12. Litton JK, Ready K, Chen H, Gutierrez-Barrera A, Etzel CJ et al. Earlier age of onset of BRCA mutation-related cancers in subsequent generations. Cancer. 2011, 118(2): 321-325.

13. Narod SA. Earlier age of onset in BRCA carriers-anticipation or cohort effect? Curr Oncol. 2011, 18(6): 257-258.

14. Bamber D. The area above the ordinal dominance graph and the area below the reciever operating characteristic graph. J Math Psychol. 1975, 12(4): 387-395.

15. Hanley JA, McNeil BJ. The meaning and the use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982, 143(1): 29-36.

16. Terdiman JP, Gum JR, Conrad PG, Miller GA, Weinberg V et al. Efficient detection of hereditary nonpolyposis colorectal cancer gene carriers by screening for tumor microsatellite instability before germline genetic testing. Gastroenterology. 2001, 120(1): 21-30.

17. Vasen HF, Mecklin JP, Khan PM, Lynch HT. The International Collaborative Group on HNPCC. Anticancer Res. 1994, 14(4B): 1661-1664.

18. Vasen HF, Watson P, Mecklin JP, Lynch HT. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. Gastroenterology. 1999, 116(6): 1453-1456.

19. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A et al. Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst. 2004, 96(4): 261-268.

Cite this article: Kwiatkowski. From Oncogenetic Pedigrees to Family Profiles: A Necessary Step to Enable Statistics . J J Bioinform Proteom. 2016, 1(1): 005.

8.2.3 BJSTR (2018) What Selection Pressure Does to Mutations Favoring Cancer? Highlights of A Simulation Approach

ISSN: 2574-1241
DOI: 10.26717/BJSTR.2018.10.001989
Kwiatkowski Fabrice. Biomed J Sci & Tech Res

Research Article Open Access

What Selection Pressure Does to Mutations Favoring Cancer? Highlights of A Simulation Approach

Kwiatkowski Fabrice[1,2]*, Serlet Laurent[2] and Bignon Yves Jean[1]

[1]*Laboratoire d'Oncologie Moléculaire, France*

[2]*Laboratoire de Mathématique: probabilités et statistiques appliquées, France*

Received: October 24, 2018; **Published:** November 01, 2018

***Corresponding author:** Fabrice Kwiatkowski, Centre Jean Perrin, France

Abstract

Context: research in oncogenetics has focused for years on mutations increasing independently the risk of cancer (ex. BRCA mutations for breast/ovarian cancers). Nowadays, interactions between mutated genes are searched for. Besides, because deleterious mutations shorten life and thus reproduction outcomes, why have they not been eliminated by selection pressure?

Methods: we developed software to test various hypotheses about mutations survival among a theoretical population having main demographic characteristics of a primitive population. Various simulations (Monte-Carlo approach) with various genotypes tested how several deleterious mutated genes with various penetrances could interact and possibly survive despite the life shortening they induce.

Results: simulation over millennia showed that deleterious mutations needed to provide evolutionary compensations such as higher fertility and/or earlier onset of reproductive capacities. This las characteristic was a strong factor enabling deleterious mutations to last. Because in female, menopause terminates the fertility period, iterations tended to select spontaneously mutations favoring cancer after menopause, without any consideration about hormonal exposure. Interactions between highly penetrant mutations were very unlikely to last and tended to split apart populations carrying each different mutation. Some results regarding fertility were validated using our database of 9000 pedigrees at high cancer risk.

Keywords: Oncogenetics; Mutation; Cancer; Selection Pressure; Pedigree; Modeling

Introduction

Very penetrant germline (hereditary) mutations in Human are scarce [1] and reports of de novo mutations are quite rare [2,3]. Even if this statistic is not exactly right, this means that mutations favoring cancer like BRCA1 and BRCA2 are very uncommon events in human's history and it is very likely that present known mutations are quite old (several millennia). Other etiologic hypotheses concerning hereditary cancer risk suggest possible interactions between several weakly penetrant mutations. Whatever the penetrance, these deleterious mutations have a direct impact on life expectancy of carriers (it shortens life time because of diseases that they favor at early ages). One might ask why natural selection, generation after generation, did not eliminate them. This is the issue of this article. To answer this question, the first paragraph describes the methods we used, the population parameter that were necessary, and how they were mixed together to provide an acceptable population-size evolution. The characteristics of the mutations that are introduced in the model are in accordance with the knowledge about BRCA mutations. A stochastic computer routine was developed to simulate what happens along generations to these mutations. First tests concerned the introduction of 5 mutations favoring cancer at different ages that do not provide any specific advantage about fertility. Second tests used various but relevant corrective measures to counterbalance the natural decline of mutations rate among studied population. In a second paragraph, these results are discussed in the light of what is known about BRCA mutations. Observational results are then described: they were obtained from a large database of family presenting with a hereditary cancer risk (more than 9 000 families and 190 000 family members recruited since 1985 in Centre Jean Perrin Oncogenetic Department). Limitations and further possibilities are described in the last chapter.

Materials and Methods

Because the economic changes of these last centuries did not concern all continents at the same time, researchers of more advanced countries could collect data that appear relevant to describe primitive population characteristics. Other works have

investigated registries to constitute same king of data. From this data, different kind of pyramid of age can be found in the literature:

a) by INED, the French National Institute of Demography [4] for France and England (1750 and 1850)

b) by GLOBOCAN for Saharan Africa in 2002 for example [5].

c) In PROVIDE survey [6] for South-Africa

Most other parameters could be found in publications of the French National Institute of Demographic Studies (INED) or were extracted from a cohort (COSA) of 1962 women that we constituted in our region (middle France) between 1996 and 2006. Half of the women of this cohort were treated for breast or ovarian cancer and the other half constituted a control group seen for prevention consultation. Their age was comprised between 25 and 89 with an average equal to 58±10 (standard deviation) [7]. To facilitate the reading of following paragraphs, formulas are grouped in **(Appendix 1)**.

Modeling the Risk of Death

Although the simulations reported hereafter are performed in a primitive context, three patterns have been modelled: the mortality risk of primitive populations, the one of under-developed countries and the present one typical of industrial countries. The pyramid of age for England in 1851 was chosen to represent a primitive situation: it corresponds to a rather steady risk of death for males and females all along life, with the known high mortality risk during the first 5 years related to the infant mortality excess. The GLOBOCAN pyramid of age for North Africa that appears similar, already shows the impact of medical progress for the first classes of age, with a reduced probability to die during childhood. The former was used to model primitive context while the latter to estimate life expectancy in under-developed countries. Finally for industrialized countries, the 1989 French pyramid of age was used. After these choices, risk of death according to age was calculated, using a polynomial regression. (Figure 1) shows for females how the mortality risk differs depending on context.

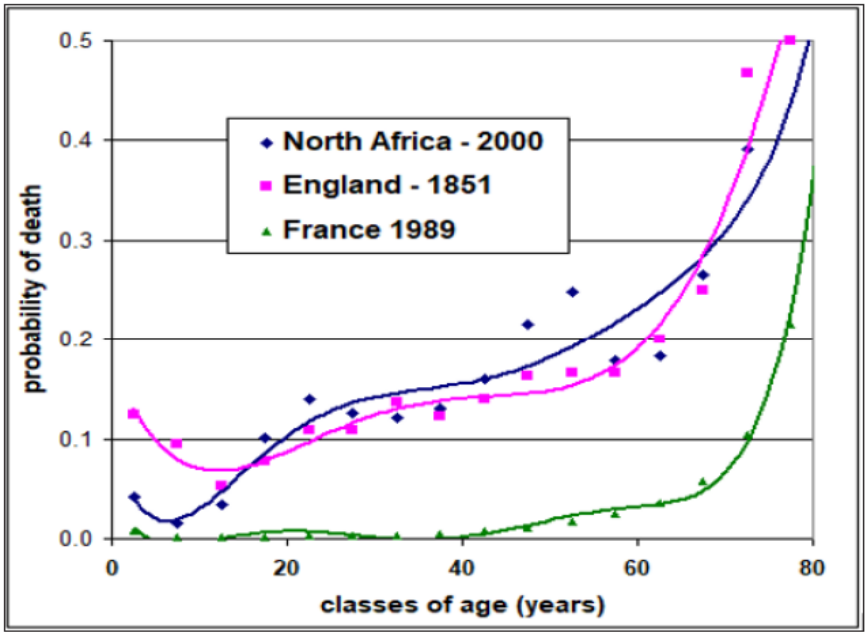


Figure 1: Fitted curves for female mortality by 5-year classes of age and three different contexts.

Marriage

Marriage (rate and age) has social determinants whatever the civilization. For example, at the end of the 18th century in France, celibacy was enjoined to 12% of women when they got nun and about 15% in main cities because of domesticity; earlier, at the end of the 17th, childless women were estimated according to church registers between 6% and 7% [8]. In our COSA cohort, which represents modern occidental societies, 3.5% of women never got married or lived in couple less than 3 years and were nulliparous. A 4% celibacy rate was introduced in our model, but we had no means to check if this value was relevant in a primitive context. Somehow, a celibacy rate seems inevitable because of congenital abnormalities or other possible social rules, even in primitive groups that may forbid marriage to some individuals. The same celibacy rate was attributed to men, although it might sometimes be much higher, for instance if polygamy is permitted: females are

therefore less available for bachelors and men remain more often alone. Polygamy could be set on in our modelling, but monogamy was the most tested situation. [9] found a mean age at first child of 25 with a standard deviation equal to 5 years.

We decided that this age was also relevant for wedding although one year could have been retrieved (at least 9 months before birth of first child). But in our modelling, it was possible to give birth to a baby just after the wedding, so this correction was not necessary. For primitive populations, we fixed the mean age of first wedding at 18±5: this agrees with the mean age at first birth that was 18.7 in rural women of Zimbabwe in 1980 [10]. Similar figures were reported by [11] in rural populations of eight countries of Southern and Eastern Africa. The minimum fertility age for females (menarche) was fixed at 13 years, under the null hypothesis. Men were considered fertile at the same age but because average wedding age for males is usually greater than females' one [9], we added 2 years in contrast to women's wedding mean age. Another reason for the older marriage age for men comes from the high maternal mortality rate causing men getting re-married at later ages. The mortality risk at each pregnancy reached 1.5% during Middle-Age [12] and such figures are still observed in some sub-Saharan countries conversely to developed countries where a 10/100 000 rate is common [13].

Natality Parameters: Number of Children

Natality corresponds to several parameters, but main statistics evaluate natality rate using fecundity rate per women. INED edited in 2004 a population survey comparing fecundity of French native women to North-African migrant women [14]. Another survey concerned households of South Africa in 1995 [6]. Both sources found a females' fecundity peak at 28-year. Data from Bangladesh contrasted from these previous outcomes [15]: fecundity peak of women between 1960 and 1976 was close to age of 22 years. The 28 years peak was nevertheless chosen although this could be discussed. An earlier fecundity period would shorten intergeneration intervals and would accelerate the modeled process. Our choice is thus more conservative (it does not interfere with events that we search for: it only delays them). We also considered that fecundity in women was a stable variable and neither social nor medical factors could let final fecundity rates vary. These parameters were used to compute the fecundity curve for primitive population. Another parameter, that does not seem to depend on social factors, is the miscarriage rate. This rate reduces significantly the final number of viable children. In our COSA cohort, this rate almost reached 20% of all pregnancies. Other population surveys have reported similar statistics [16-18]. Owing to the anti-apoptotic properties of BRCA genes, we made this rate variable to evaluate possible compensations for the deleterious role of tested mutations. The maximum absolute reduction of this rate was set to 10% although in a population-based analysis, this diminution reached one third [19].

Gender of Children and Mutation Transmission

Normally, to compensate for the higher male infant mortality, more males get born than female. In France, 105 males get born for 100 females, thus a probability to be a male of 51.2%. This percent was kept in our modeling. The chance for a child to receive a mutated gene from one of his parents is 50%. In our modeling, this value was set for each possible mutation.

Age of Menopause

In our COSA cohort, mean age of menopause was 51±3.8 years and its distribution was Gaussian. This age is rather stable worldwide: for example the same age was reported in a large Norway cohort [20] and a slightly younger age 49.3±3.4 in a large Japanese cohort [21]. This parameter is very important because reproduction period length (time from menarche to menopause) has a direct impact on natality outputs. In our model, we fixed the mean menopause age at 50±5 years. Of course, because of usual early mortality in primitive population, most of simulated women died before menopause. Of course, because of usual early mortality in primitive population, most of simulated women died before menopause.

A limitation of male fertility must be questioned. A Canadian study based on ancient family registries [22] showed that male widows continued to get married and to have children even when they were 60 years old or more. In this survey, chosen old men were married with women younger than 30. They contributed to an average of 2.2 children which was compared to 2.8 children for men married before they were 30. This small decline with age can be justified by studies of semen quality among men of various ages [23], that show a slight but constant yearly reduction after 30 years of sperm motility and concentration. However, this decrease happens to be inferior to the one due to environmental changes (nutrition, life habits, pollution...): for instance between 1973 and 1992 the decrease of motile and normal spermatozoa in fertile French men was higher among young adults of each period than the decrease caused by age. Other reproductive problems may occur also for older men such as congenital anomalies whose rate increases with age [24]. Considering also that a reduced correction of male fertility rate would be of insignificant value as it would apply to a very limited population (most of men die before 50 in a primitive context) in our modeling, no upper limit was fixed to male fertility age.

Cancer Risk and Incidence

Incidence and mortality of cancer statistics were available in FRANCIM report [25]. For women older than 20, the risk to develop a breast cancer according to age exhibited a peak between the ages of 60 and 64. In the COSA study, the peak was 58±12 years and distribution was almost gaussian. The cancer incidence curve to simulate cancer risk was set gaussian with a 5-year standard deviation, considering cancers caused by mutation happen much earlier and in a more grouped fashion. Breast cancer is supposed to induce death in the following 5 years in 30% of cases. This cancer stops rapidly fecundity. In our modeling, the end of fecundity (for both gender) was set at disease onset and we made the average delay till death equaled to 5±5 years although this is a very optimistic hypothesis. All cancers prognosis was supposed to follow the same gaussian shape. Penetrance is the cumulative

rate of disease occurring in a population of mutation carriers. For BRCA mutations, penetrance is supposed to vary according to environmental conditions: for example BRCA penetrance is close to 70% today which means that 70% of mutated women will have a breast or ovarian cancer during their life-time.

Some life habits can change this penetrance (breast feeding, parity, nutrition, physical activity) and the BRCA penetrance was supposed to be much lower, around 50% one century ago. In our modeling, this parameter could be changed, but calculations reported hereafter were performed using a 50% penetrance. Sporadic cancers have also been considered, to better fit reality: indeed, main difficulty in population survey is often to distinguish between sporadic cancers and those favored by hereditary factors. An interesting work of [26] showed that most of cancers happen randomly at a frequency that increases proportionally with the number of cell divisions in tissues during life (correlation r = 0.81), therefore with age too. To implement the risk for sporadic cancer, we used epidemiological data reported by the department of UK cancer research 2012-2014 [27]. In case of competing cancer risk, that is a familial one associated to a sporadic one, the first occurring – most of the time the familial cancer – was kept. There is no biological reason to believe that this proportionality was different in the ancient times.

Mutations Geographic Spread Map

Simulations have also investigated how deleterious mutations dissemination could be influenced by geographical aspects. A virtual map was created, and individuals were randomly spread on it at the beginning. Travels across the map were limited by generation and distances were limiting factors for individuals to find a spouse. Different conditions were tested as for example mutations interacting in synergy or not regarding the cancer risk.

Complementary Parameters

We supposed that marriage lasted until death (no divorce), that marriage was not possible between brothers and sisters, parents and children, while more distant filiations were admitted forming new couples. After the death of a spouse, the widow could get married again, except menopausal women in order not to reduce men's descendance (as a man getting married with a menopaused woman would not expect a child). Polyandry was not implemented as it seems to be a very rare social rule in Man's history.

Randomization for the Monte-Carlo Simulations

Two algorithms were implemented:

a) the standard random function of Microsoft Visual Basic (a random table) with a pointer re-initialized by the timer function at the beginning of main loops.

b) Mersenne-Twister algorithm [28]: it is a validated algorithm that generates pseudo-random numbers.

Tests performed using both algorithms yield similar statistics, but the second algorithm was 10 times slower than the first one. For the calculations reported hereafter, the first algorithm was thus used.

Interface to Enter Parameters

A user-friendly interface enables to introduce calculation parameters. It includes:

a) Iterations number (default = 100)

b) Size of the initial population (2000 = 1000 females and 1000 males)

c) Length of the period during which the test is to be carried on (2000 years)

d) Time slice for the statistics (100 years), i.e. 20 checkpoints if period lasts to 2 millennia as above.

e) Type of mortality curve, i.e. age pyramid. We used here the primitive context.

f) Average fecundity per women (ex. 2.1 for a developed context)

g) Polygamy permitted or not (no)

h) Number of mutations to test (5)

i) Initial frequency of mutation carriers per mutated gene (5%)

j) Impact of these mutations on fecundity:

a. Miscarriage risk set to 20% (and decrease to 5%)

b. Marriage precocity (±2 years, default = 0)

c. Synergy between mutations: yes/no and number of mutations necessary for this synergy to work

k) Mutations penetrance (50%)

l) Synergy between mutations: yes/no and number of mutations necessary to make penetrance cumulative.

m) Gender concerned by the cancer risk (males, females or both)

n) Age peak for penetrance: the age when most cancers related to the mutation occur (ex. 40 years)

o) Progressive penetrance precocity (years): if 5 mutations are studied with a median peak for cancer susceptibility = 40 years, a 10-year precocity, would attribute to the first mutation a cancer risk peak at (40 - 2 x 10) = 20 years, and 30, 40, 50, 60 respectively to the four other mutations.

Verification

An EXCEL routine was developed to check how families were generated by our software: this routine enables to draw for each family the resulting pedigrees (Figure 2). The legend of the figure describes main symbols signification. Mutations are indicated using a lowercase letter added to the subject number when needed. Statistics consist in means and standard deviations calculated for each parameter over the total number of iterations. They are registered at each time slice, as for example every century if check points are scheduled by intervals of 100 years (Appendix 2).

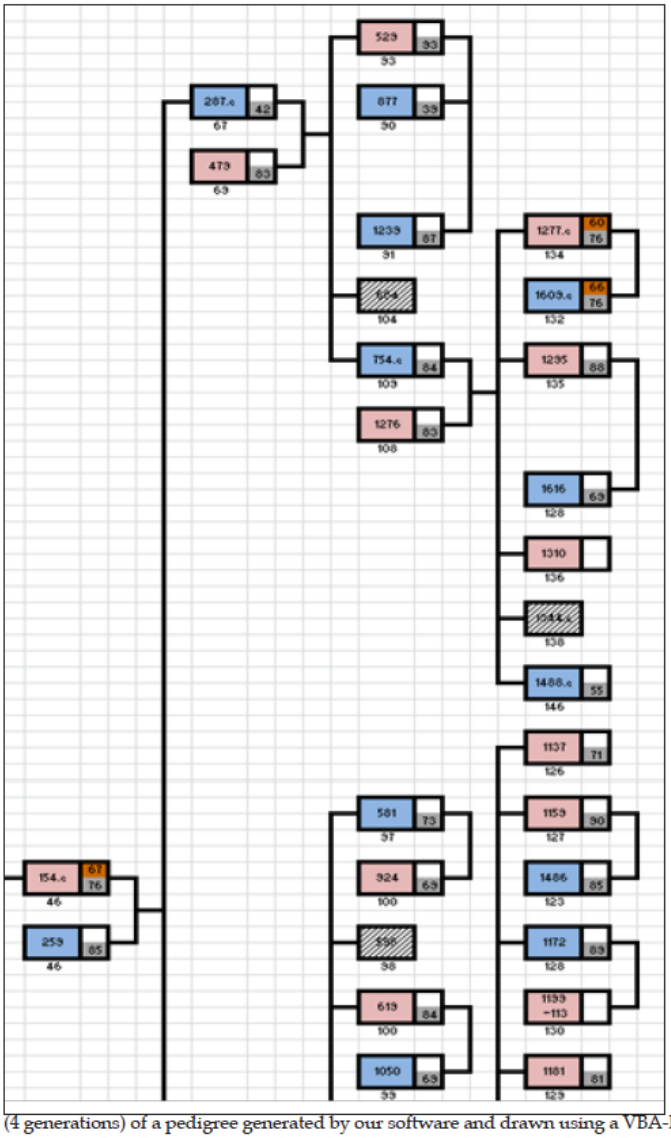


Figure 2: Partial view (4 generations) of a pedigree generated by our software and drawn using a VBA-EXCEL routine to verify the shape of resulting families: female = pink rectangle, male = blue rectangle, orange = familial cancer + age at occurrence, yellow = sporadic cancer (none here), grey age at death and hatched = miscarriage.

Results

Results presented hereafter were obtained using simulations within a primitive context, considering this was the case most of the time homo-sapiens existed (>50 millennia). Also, only two millennia of evolution were evaluated by calculation, enough duration to notice already significant evolutions.

Tests with No Deleterious Mutation

Under the null hypothesis, i.e. no deleterious mutations introduced into the model, final pyramid of age (Figure 3) kept the shape corresponding to the type of context selected by the user (developed, under-developed or primitive).

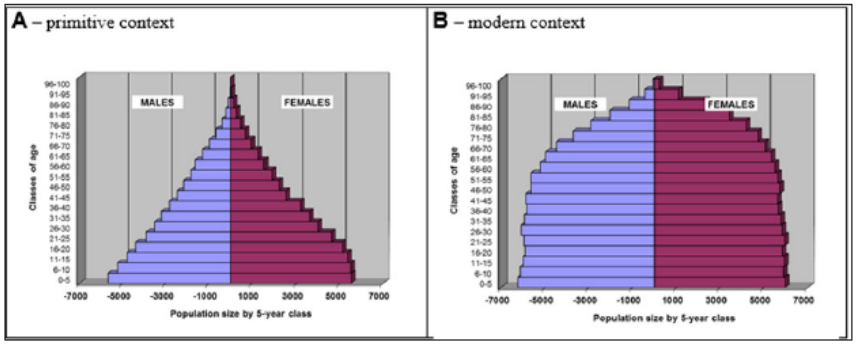


Figure 3: Pyramid of age of the last century after 2 millennia of follow-up resulting from simulations according to two contexts: primitive and modern.

Tests with Five Deleterious Mutations but without Impact on Fertility

Five deleterious mutations were added to the model. Each mutation was supposed to favor cancer at different ages and concerned both men and women. Mutations were then randomly assigned to 5% of the initial population (the 20-year old couples). So, about 25% of the population carried at least one mutation. After all iterations were performed by the program, a longitudinal diagram was drawn (Figure 4A). Mutations that favor a cancer before 30 years were rapidly eliminated, while the two mutations inducing cancer around or after menopause tended to "survive". Other statistics were calculated: figure 4B represents the evolution of underlying population characteristics, still in a primitive context. Except for cancer age, ages at each event were stable: ages of women at their children births. Mean age of deaths, if we exclude the artifact of first century, exhibited no evolution and stayed close to 35 years as well as mean menopause age around 50, that is for women that did not die before. This is true also for cancers: average age at cancer onset is higher than average age of death because cancer concerned only individuals that reached the age of each cancer risk.

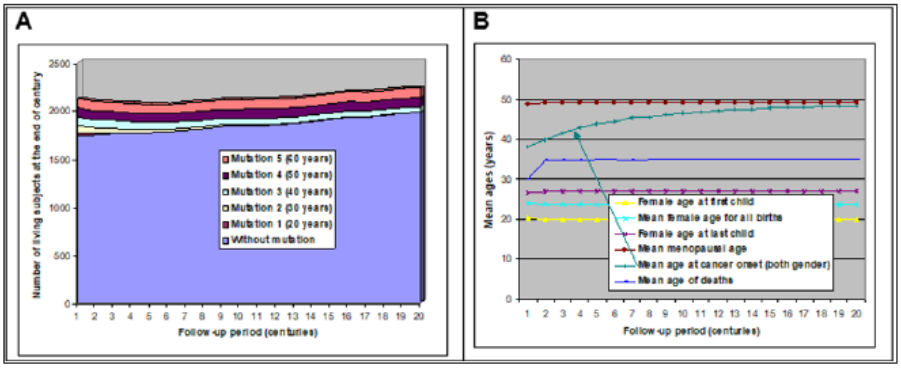


Figure 4: Population evolution according to 5 types of mutation inducing a 50% cancer risk occurring in average at 5 ages [20-60] in men and women (follow-up: 2 millennia).
A-Evolution of population size.
B-Evolution of ages for different events.

The most interesting trend concerned the mean age of cancer onset (marked by an arrow in (Figure 4B)). Year after year, this mean age increased until it almost reached menopause mean age. This highlights how selection pressure works: because carriers of mutations favoring the most precocious cancers die rapidly, their reproduction period is shortened, and their offspring is reduced. Thus, generation after generation, these carriers become less, and less numerous and worst mutations disappear. Precocious cancers then become rarer and mean age at cancer onset increases. This phenomenon tends to reduce the incidence of all major diseases happening before menopause. Hence, without any hypothesis about the etiology of cancers, it evidences that onset of cancers is likely to happen naturally near menopause or after. This selection happens also for men if partners' ages are often similar: but this relationship is indirect.

Tests with Five Deleterious Mutations Providing a Fertility Advantage

As concluded above, ancient mutations favoring early cancer in human should have disappeared with time. But it is not the case. Thus these mutations must have provided some evolutionary advantages. Two possibilities were tested in our simulations: they permitted to compensate for the loss of reproductive chances. The first one was a reduced miscarriage rate, and the second an earlier onset of reproduction period (i.e. by reducing first marriage age). As in (Figure 4), the two most deleterious mutations were eliminated during the first millennium, although penetrance was set at 50% (half of mutation carriers would have a cancer if they did not die before from another cause). First difference concerned the overall size of the population which grew steadily, especially when mutated women could get pregnant one year earlier: the global increase was 11.3% for (Figure 5A) and 20.4% for (Figure 5B) versus 5.6% in figure 4 when no compensation was introduced. Interestingly, non-mutated individuals seemed to benefit as others from the reproductive advantage of slow penetrant mutations. Final proportion of mutation carriers were respectively 15.5% when the advantage was a 5% miscarriage risk reduction and 17.3% if reproduction begun one year earlier. As 3 of the 5 mutations remained at the end of the follow-up period, it suggests the proportion of mutation carriers should increase in the following millennia, especially for the latter advantage.

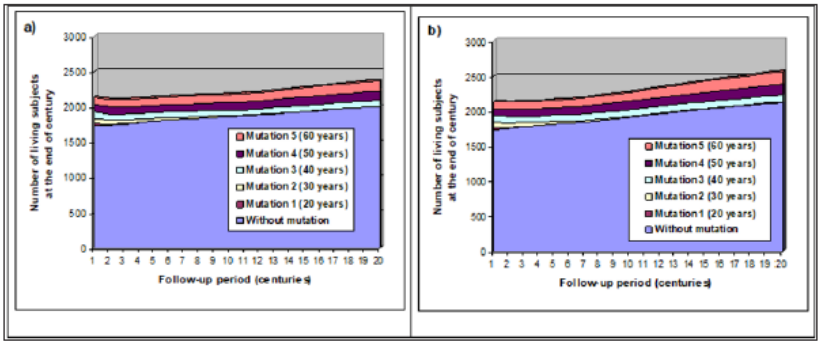


Figure 5: Impact of two types of reproduction advantages.
Note:
a) with a 5% absolute decrease of miscarriage risk (15% instead of 20%).
b) with a 1-year advance of reproduction onset.

Tests With 5 Mutated Genes that May have a Synergy Together

Results did not differ much from previous situations. The best condition for the mutations to spread was when:

a) They gave reproductive advantages but did not favor cancer when they were alone.

b) Cancer risk was heightened only when two or more mutations were carried by individuals.

Geographic Dissemination of Synergic Deleterious Mutations

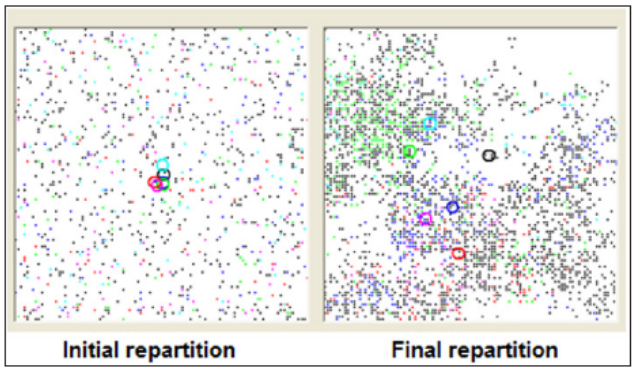


Figure 6: Spontaneous geographic dissemination of 5 mutations increasing cancer risk when individuals carry more than one of them: black points and circle correspond to individual carrying no deleterious mutation. Other circles locate the centroid of mutation carriers.

When deleterious mutations do not interact, location of carriers covers the map at random. Centroids of mutation carriers remain close to the center of the map. Frequently though, the map gradually exhibits empty zones that individuals leave because of the absence of enough potential partners. Polygamy only seemed to slow down a little this evolution. When mutations have a synergic influence on cancer occurrence, populations of carriers spread significantly across the map and centroids deviate from each other (Figure 6).

Discussion

Because genetic research concerning species evolution is long, difficult and expensive, it is tempting today to perform simulations to limit ground investigations. The same approach exists in other scientific fields such as astronomy and climate, and they have already brought much essential knowledge. The simulations described here in the field of oncogenetics, required a much simpler software and could be performed on a desktop computer. Calculations testing each situation lasted a few hours (at most a few days), enough to obtain stable statistics using one hundred iterations per tested condition. Larger numbers of iterations (500 and 1000) were tested without gain in accuracy and/or differences between resulting evolutions. An initial population size of 2000 persons (1000 males and 1000 females) seemed to be the minimum necessary to yield a good stability of outcomes. Although initial population contained only members of identical age (20 years), the ending pyramid of age (built using data of last century) was correctly matched with the theoretical one, whatever the context used (developed, under-developed or primitive). Finally, the type of generator of pseudo-random numbers did not alter the global performance of the modeling. The most difficult part of the modeling was to find reliable data (age pyramids) likely to fit with life conditions of our ancestors. Because modern way of life did not concern all populations on earth during the past century, it was possible to obtain good estimators of age-related survival probabilities of primitive populations. We considered other biologic parameters (fecundity, menopause…) stable since long, and assumed that they were same as now. Some hypothesis was made concerning ancient marriage habits, and when this was not relevant (for example polygamy), we chose to add parameters that could enable to test various situations.

Overall, population evolutions were compared with or without mutation carriers but under the same context conditions (fertility…). We could question if altered conditions would not yield major changes in our conclusions. Many other tests have been realized: main changes only modified the period length necessary to evidence the differences that we reported. Insignificant results were not described above in particular the impact of polygamy, despite age matching between spouses was deactivated. This could be expected because the several wives of each polygamous man were no more available for other bachelors. The fact that cancers would concern a gender rather the other one, was neither of major importance. But this could come from the preference that was computed: partners of similar ages were first sought for marriages, and because this was usually possible, spouses tended to be well age-matched. Therefore, cancers impacting exclusively males or females, although new marriages were possible, had a global identical effect on reproduction outputs.

Reproductive advantages can result from various biological parameters and because of behavioral changes: only a fertility increase was tested (by reducing the rate of miscarriages), and no other possibility was implemented since they all would end up with a similar percent increase. An earlier onset of the reproduction period was also envisaged because this has a different impact on offspring: the number of children increases mainly because parents that begin earlier reproduction escape from causes of death they are supposed to meet later. In our modeling, because menopause age was not reduced by the same advance duration, the enlargement of reproduction period could also play a role, although minor as life expectancy usually limits this advantage. Our modeling is nevertheless compatible with Bjelland's findings that evidence only a modest reduction of menopause age when menarche occurs earlier [20].

The geographic simulations brought interesting conclusions: when mutations have a synergic impact on cancer risk, they tend to diverge from each other. This reflects the disappearance along time of individuals carrying several mutations and/or polymorphisms that alone would have no effect on cancer occurrence. This is of major importance: it is possible that today, an increased incidence of cancer may be caused by the population melting. Indeed, because it is so easy to travel over long distances, inhabitants of different countries/even regions - carrying "safe" polymorphisms resulting from an environment-related selection pressure lasting for centuries - meet "again" together and exchange through their descends mutations that were and still are deleterious together. Some population study should investigate this issue to confirm or infirm this hypothesis.

Main weakness of our modeling concerned the uniformity of the context. When a "primitive" context was set at the beginning of simulations, it was alike at the end of simulations, that is after two (or more) millennia. This cannot adequately fit the two last millennia of mankind. Evidently, this suggests that mutations that could provide an evolutionary advantage in the past, may on the contrary become nocuous today as considerable changes have altered man's environment (birth control). Another weakness may come from the complexity of the genetic biology: we tested synergies between only 5 mutations without real improvement of reported results. But the relationships between the ±30 000 genes of our genome may be able to produce completely different adaptation features. The debate thus remains opened, especially for low penetrance mutations that can mix by hundred their interactions. Conclusions resulting from our simulations have been tested among a very large database of pedigrees (±9 000 families including more than 190 000 members). They have shown that mutated families had indeed more children and less miscarriages. They also use to have they first child earlier than non-mutated families. Surprisingly, this was also true when mutated members were males instead of females [19,29].

References

1. Thomtson, Thomtson (2004) Genetics in medicine. In: Robert Nussbaum (Eds.). (6th Edn), Saunders, Philadephia, p. 84-85.
2. Van der Luijt RB, Van Zon PH, Jansen RP, van der Sijs Bos CJ, Wárlám Rodenhuis CC, et al. (2001) De novo recurrent germline mutation of the BRCA2 gene in a patient with early onset breast cancer. J Med Genet 38(2): 102-105.
3. Diez O, Gutiérrez Enríquez S, Mediano C, Masas M, Saura C, et al. (2010) A novel de novo BRCA2 mutation of paternal origin identified in a Spanish woman with early onset bilateral breast cancer. Breast Cancer Res Treat 121(1): 221-225.
4. Vallin J, Caselli G (1999) Quand l'Angleterre rattrapait la France. INED-Population et sociétés 346: 1-4.
5. IARC Descriptive Epidemiology Group (2002) GLOBOCAN -cancer incidence, mortality and prevalence worldwide.
6. Punt C, Pauwn K, Mohube E, Gilimani B, Rantho L, et al. (2003) Demographics of South African Households – 1995. PROVIDE background paper 3: 1-22.
7. Delort L, Kwiatkowski F, Chalabi N, Satih S, Bignon YJ, et al. (2007) Risk factors for early age at breast cancer onset - The "COSA program" population-based study. Anticancer Res 27(2): 1087-1094.
8. Henry L, Houdaille J (1978) Célibat et âge au mariage aux XVIIIe et XIXe siècles en France I. Célibat définitif. In: Population, 33e année, n°1: 43-84.
9. Levy ML, Sardon JP (1982) L'écart d'âge entre époux. INED-Population et sociétés 162: 1-2.
10. Thomas D, Maluccio J (1995) Contraceptive choice, fertility, and public policy in Zimbabwe. Living Standards Measurement Study (LSMS) working paper 10(1): 1-60.
11. Harwood Lejeune A (2000) Rizing age at marriage and Fertility in Southern and Eastern Africa. European Journal of Population 17: 261-280.
12. Matart BY (1994) Approche de la mortalité maternelle au Moyen-Age en Provence. Actes des 6e Journées Anthropologiques. Dossier de documentation archéologique n°17. CNRS Editions, Paris, France.
13. World Health Organization (2014) Trends in maternal mortality: 1990 to 2013. Estimates by WHO, UNICEF, UNFPA, The World Bank and the United Nations Population Division. ISBN 978 92 4 1507226.
14. Toulemon L (2004) La fécondité des immigrés: nouvelles données, nouvelle approche. INED-Population et sociétés 400: 1-4.
15. Amin R, Farukee R (1980) Fertility and its regulation in Bangladesh. Bank Staff Working Paper n°383. The World Bank, Washington, USA.
16. Zinman MJ, Clegg DE, Brown CC, O Connor J, Selvan SG (1996) Estimates of human fertility and pregnancy loss. Fertil Steril 65(3): 503-509.
17. Buss L, Tolstrup J, Munk C, Bergholt T, Ottesen B, et al. (2006) Spontaneous abortion: a prospective cohort study of younger women from the general population in Denmark. Validation, occurrence and risk determinants. Acta Obstet Gynecol Scand 85(4): 467-475.
18. Katz VL (2007) Spontaneous and recurrent abortion: etiology, diagnosis, treatment. In: Katz VL, Lentz GM, Lobo RA, Gershenson DM (Eds.) Comprehensive Gynecology (5th Edn) Philadelphia, PA: Mosby Elsevier
19. Kwiatkowski F, Arbre M, Bidet Y, Laquet C, Uhrhammer N, et al. (2015) BRCA Mutations Increase Fertility in Families at Hereditary Breast/Ovarian Cancer Risk. PLoS One 10(6): e0127363.
20. Bjelland EK, Hofvind S, Byberg L, Eskild A (2018) The relation of age at menarche with age at natural menopause: a population study of 336 788 women in Norway. Hum Reprod 33(6): 1149-1157.
21. Otsuki S, Saito E, Sawada N, Abe SK, Hidaka A, et al. (2018) Female reproductive factors and risk of all-cause and cause-specific mortality among women: The Japan Public Health Center-based Prospective Study (JPHC study). Ann Epidemiol 28(9): 597-604.
22. Payeur F (2008) Les registres démographiques renseignent sur la fertilité de l'homme en fonction de l'âge. Bulletin 335 de l'ambassage de France au Canada.
23. Auger J, Kunstmann JM, Czyglik F, Jouannet P (1995) Decline in semen quality among fertile men in Paris during the past 20 years. NEJM 332(5): 281-285.
24. Wagner L (2004) Fertilité de l'homme vieillissant. Prog Urol 14(4): 577-582.
25. Menegoz F, Cherie Challine L (1999) Le cancer en France: incidence et mortalité. Situation en 1995-evolution entre 1975 et 1995. Editions INSERM, Paris, France.
26. Tomasetti C, Vogelstein B (2015) Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science 347(6217): 78-81.
27. Cancer Research UK (2016) URL. Access: September 2018.
28. Matsumoto M, Nishimura T (1998) Mersenne-Twister, a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transaction on Modeling and Computer Simulation 8(1): 3-30.
29. Guay JH (2013) Perspective Monde-outil pédagogique des grandes tendances mondiales depuis 1945. Sherbrooke University, Canada.

ISSN: 2574-1241

DOI: *10.26717/BJSTR.2018.10.001989*

Kwiatkowski Fabrice. Biomed J Sci & Tech Res

Submission Link: https://biomedres.us/submit-manuscript.php

8.2.4 Congenital anomalies (2019) Association between hereditary predisposition to common cancers and congenital multimalformations

Received: 5 November 2018 | Revised: 28 December 2018 | Accepted: 17 February 2019

DOI: 10.1111/cga.12329

WILEY Congenital Anomalies

ORIGINAL ARTICLE

Association between hereditary predisposition to common cancers and congenital multimalformations

Fabrice Kwiatkowski[1,2] | Isabelle Perthus[3] | Nancy Uhrhammer[1] | Christine Francannet[3] | Marie Arbre[1] | Yannick Bidet[1] | Yves-Jean Bignon[1]

[1]Oncogenetics Department, Centre Jean Perri (Comprehensive Cancer Center), Clermont-Ferrand, France

[2]Laboratory of Mathematics: Probabilities and Applied Statistics, Clermont-Auvergne University, Clermont-Ferrand, France

[3]Medical Genetics Department, Study Center of Congenital Malformations in Auvergne (Centre d'Etude des Malformations Congénitales en Auvergne), Clermont-Ferrand, France

Correspondence
Fabrice Kwiatkowski, Oncogenetics Department, Centre Jean Perrin (Comprehensive Cancer Center), 58, rue Montalembert, 63011 Clermont-Ferrand, France.
Email: fabrice.kwiatkowski@clermont.unicancer.fr

Funding information
FEDER (European Funding of Regional Development) and Conseil Régional Auvergne, Grant/Award Number: convention n°36760 (2013); Conseil Régional Auvergne; FEDER (European Funding of Regional Development)

In a previous article we reported that mutations favoring cancer at adulthood seemed to improve fertility and limit miscarriages. Because spontaneous abortion may result from anomalies in embryo, we questioned if an increased frequency of congenital malformation could be evidenced among cancer-prone families. Oncogenetics database (≈193 000 members) of the comprehensive cancer center Jean Perrin was crossed with regional registry of congenital malformations (≈10 000). Among children born between 1986 and 2011, 176 children with malformation matched in both databases. In breast/ovaries cancer-prone families, the risk for malformations was multiplied by 2.4 [1.2-4.5] in case of a BRCA1 mutation. Frequencies of malformation in BRCA2 and MMR mutated families were similar to families without a cancer syndrome. In comparison to malformations concerning a unique anatomical system, multimalformations were significantly more frequent in case of BRCA or MMR mutations: compared to families without cancer syndrome, the risk of multimalformations was multiplied by 4.1 [0.8-21.7] for cancer-prone families but with no known deleterious mutation, by 6.9 [1.2-38.6] in families with a known mutation but an unknown parental mutational status and by 10.4 [2.3-46.0] when one parent carried the familial mutation. No association with the type of anatomical system was found, nor with multiple births. These results suggest that BRCA and MMR genes play an important role in human embryogenesis and that if their function is lowered because of heterozygote mutations, congenital malformations are either more likely (BRCA1 mutations) and/or more susceptible to concern several anatomical systems.

KEYWORDS
BRCA, cancer syndrome, congenital malformation, HBOC, oncogenetics

1 | INTRODUCTION

Predisposition to cancer in early adulthood exerts selective pressure on predisposed individuals by reducing life expectancy and consequently the length of the reproductive period. Mutations in one of the BRCA genes, with an associated risk of cancer starting as early as age 30, are relatively frequent in spite of this selective pressure. The persistence of such mutations in the population is demonstrated with founder mutations known to be thousands of years old.[1–3] In a large retrospective survey of our oncogenetic database, we found that in families predisposed to breast/ovarian cancer, women from BRCA mutated families had a 36% lower miscarriage frequency ($P = 0.015$) than those from families with no known deleterious mutation; among individuals of both genders tested for a BRCA mutation, childless individuals were 22% less frequent for carriers ($P = 0.0022$) and the interval between first and last child was 16% longer ($P = 0.042$).[4,5] Although the underlying biological mechanisms are not yet known, this finding could suggest two opposing hypotheses:

1. Natural miscarriage triggers could be inhibited by defective genetic pathways that predispose to cancer. Considering that about 50%-70% of miscarriages are caused by cytogenetic abnormalities[4–8] of which a majority are induced by de novo

aneuploidy,[9] a higher frequency of congenital anomalies might be seen among carriers of mutations in these pathways.
2. Mutations predisposing to cancer at adulthood could reduce the risk for embryonic or fetal malformation—by an unknown genetic mechanism—and thus diminish the resulting frequency of miscarriage in the mutated population.

To the best of our knowledge, no research has been published on the incidence of congenital malformations in the offspring of parents with hereditary cancer risk. Sources of data on both conditions were available in the Auvergne region of France and were extensive enough to enable research regarding this issue. We crossed the information available in two large databases, the first used by the oncogenetic service of Centre Jean Perrin Comprehensive Cancer Center containing pedigrees of cancer-prone families and the second from the regional register of congenital malformations that includes all children born with a congenital malformation. The study period corresponded to 26 years, from January 1986 to December 2011.

2 | MATERIALS AND METHODS

The Auvergne region contains about 2% of the national population.[10] Because of its distance from ports and major trade routes and its traditional rural character, migration inflows have always been limited. This makes Auvergne a good model for long-term oncogenetic studies.

2.1 | Regional registry of congenital abnormalities

The Centre d'Etude des Malformations Congénitales en Auvergne (CEMC-Auvergne) is one of seven regional registries of congenital malformations in France. Launched in 1983, it is certified by the National Committee of Registries. It concerns all mothers giving birth in either public or private clinics in the region, about 14 000 births per year, including stillbirths and therapeutic abortions. All malformed newborns are registered as they are born after a pregnancy of at least 22 weeks of amenorrhea or if pregnancy was interrupted for congenital malformation regardless of the duration of amenorrhea. For livebirths, the diagnosis of malformation must be made during the first year of life. Exhaustivity is ensured because each health institution must declare cases of malformation and more than 99% of women give birth in public or private maternities.

All types of malformations are registered, including single or associated malformations, multiple syndromes (identified or not), with or without abnormal karyotype. Excluded are inborn errors of metabolism and minor deformations (hip shift without luxation, foot deformation, angioma or naevi smaller than 4 cm^2, umbilical hernia with no need for surgery). Statistical analyses were performed either using the four categories of malformation (unique, multiple, syndromal, and karyotypic) or using only two groups, unique malformations vs all other three categories that are more extended.

We selected the 10 026 livebirth cases in the registry. Considering about 364 000 children were born in Auvergne during the 25-year study period, this corresponds to a malformation frequency of 2.8%, a rate compatible to the nation-wide estimation of 2.4% in 2011 to 2012.[11]

2.2 | Centre Jean Perrin Oncogenetics database

Created in 1988, the oncogenetics department of the regional comprehensive anticancer Centre Jean Perrin is the only center in the region offering evaluation of hereditary cancer risk. A large majority of local individuals belonging to high-risk families seeking oncogenetics advice address this department. At the time of this study, the database included about 190 000 family members from 6600 families. Based on the population size of Auvergne and the expected prevalence of women with BRCA mutations in France, about 45% of women with familial breast/ovarian cancer risk in the region are included in the database.

Types of cancer risk were grouped by geneticists into several categories. It was based on their expertise in the early days and later on the calculation of scores like Eisinger[12] or Manchester[13] for hereditary breast and ovarian syndrome or, for hereditary colon cancer syndrome (HNPCC), using Amsterdam[14] and/or Bethesda criteria.[15] Breast/ovarian families were divided into families with BRCA1 or BRCA2 mutations and families with breast/ovarian cancer predisposition but no diagnosed BRCA mutation. PALB2 mutations could not be studied because of the too recent discovery of the implication of this gene in HBOC. Lynch syndrome families were grouped with colon cancer syndrome and again split into two classes: HNPCC/colon syndrome either with or without a known mutation in the mismatch repair genes (MLH1, MSH2, MSH6, or PMS2). When a family was diagnosed with both breast/ovarian and colon syndromes, it was placed within the group corresponding to the first indication given by the oncogeneticist. All other cancer risks (prostate, kidney, thyroid, hematological, digestive tract other than HNPCC/colon) were excluded from our analysis because of the wide heterogeneity of syndromes. A reference group included all individuals and their family members who consulted at the oncogenetics department but for who no cancer risk was diagnosed and therefore no genetic test was performed. A priori, this group was not exposed to a higher cancer risk than the general population, and was assumed to carry the same congenital malformation risk. Another control group was composed of individuals testing negative for a known familial mutation.

2.3 | Database matching

A temporary mixed database was constituted to evaluate the frequencies of congenital abnormalities: children born alive between 1986 and 2011, regardless of parental mutation status and type of cancer predisposition were extracted from our oncogenetic database. A computer "robot" was developed to match its records with those of the malformations register, based on similar/close children names, date of birth, mother's names, and age of parents; a visual control list permitted to validated manually each proposed match. To limit bias, that is, artificially increase the risk for malformation in particular subgroups of cancer-prone families because of a syndrome combining both cancers and malformations, four cases of microcephaly were excluded from the analysis: these cases were referred for genetic analysis on the

basis of specific congenital abnormalities suggestive of Nijmegen breakage syndrome (NBS), a syndrome that includes strong predisposition to lymphoid malignancy.[16] Their corresponding families were thus excluded from the data-matching diagram (Figure 1). Fanconi anemia was considered for exclusion, but none of our cancer-prone families presenting with this syndrome had children born after 1985.

One hundred and seventy-seven children with congenital anomalies corresponded to children in our oncogenetic database. The 33 abnormalities found for the 3244 children belonging to families with cancer predispositions other than Lynch and HBOC were excluded from the analysis because of the wide heterogeneity of this population.

NB. A public repository has been created in order to make data available. It contains several Excel sheets with first, an explanation on how are organized other sheets, second the results sheets and finally detailed information by subgroup. The address of the repository is: http://www.cjp.fr/fichiers/Anoccan_data.zip

2.4 | Ethics

CEMC-Auvergne is certified by the National Committee of Registries (2012-2015) and the database was authorized by the French Authority for personal data protection (CNIL no 1387396). The consent signed by parents to enter this registry, allows the use of clinical data for research purpose. Centre Jean Perrin Database was declared by the CNIL correspondent under the number 1621407V0 on January first 2001. This certificate was renewed by CNIL on May 17th 2017 (no 2030983V 1). Counselees signed an informed consent that enables the use of data for research purpose. Another special authorization was requested because of the French regulation regarding the merge of databases coming from different entities (here CCC Jean Perrin and CEMC belonging to the Regional University Hospital). This authorization was granted by CEERES (national expertize committee for research, studies and evaluation regarding health) on March 15th 2018 (no TPS 37636) and then by CNIL on May 18th 2018 (DR 2018-108) which permitted us to perform our study without any special information to families members about it. Finally, study ethics approval was obtained on 25 July 2018 (CECIC Rhône-Alpes-Auvergne, Grenoble, IRB 5921, file number CE-CIC-GREN-17-13).

2.5 | Statistical analysis

Each family where a deleterious mutation was diagnosed yields three sets of children: those born to carriers of the familial mutation, those born to parents of unknown mutation status, and children born to non-carriers. In the first set, 682 children were born to parents

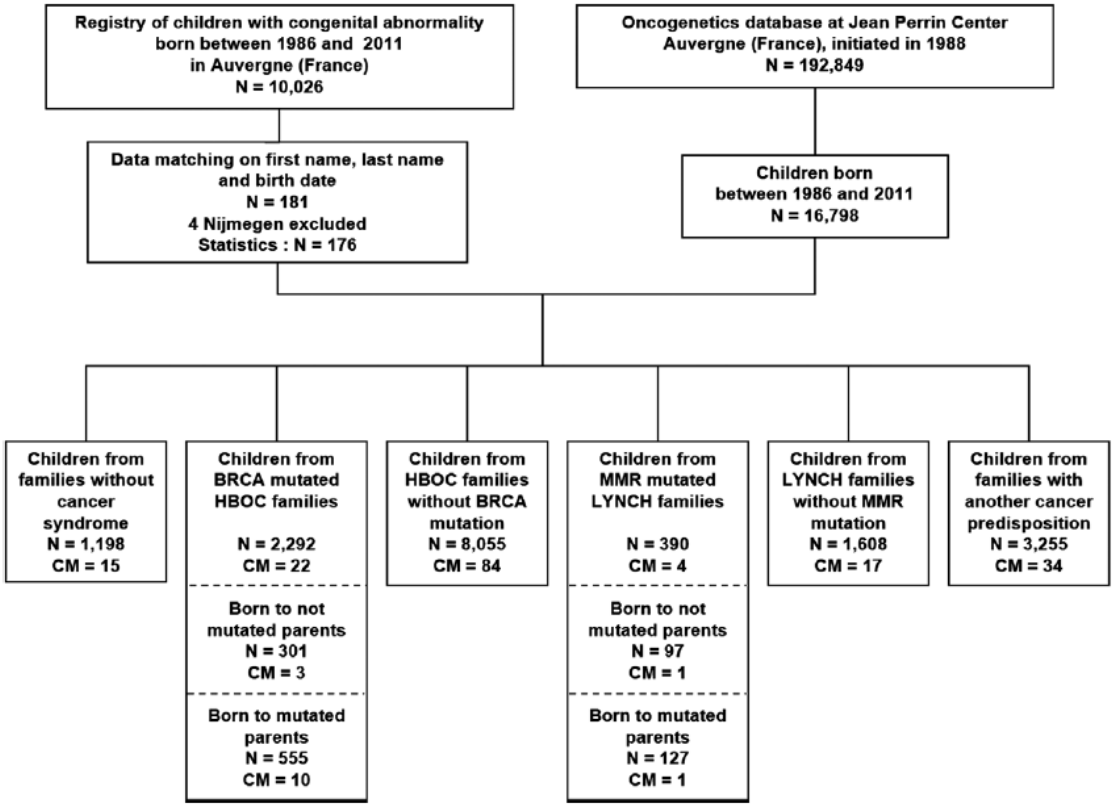


FIGURE 1 Data-matching flowchart of the regional registry of congenital malformations and CJP oncogenetics database (CM, congenital malformation; HBOC, hereditary breast/ovarian cancer; Lynch, Lynch syndrome)

carrying the familial mutation (555 BRCA+ and 127 MMR+). The second set comprises 1602 children born to families with a known cancer predisposing mutation but parental carrier status was unknown. Finally, the 398 children born to known non-carriers (301 BRCA- and 97 MMR-) form a new control group, as their parents do not carry any known mutation favoring cancer. Congenital anomaly frequencies were compared for children of parents positive for a mutation or of unknown status belonging to a mutated family, vs other situations.

Because the risk for congenital malformation increases in cases of multiple births,[17–21] the frequency of these events was computed per subgroup. These frequencies were compared to frequencies of congenital anomaly both to verify that our statistics were not biased because of this, and also to control that these frequencies were in accordance with national figures.

To compare proportions of children with abnormalities between subsets, χ^2 test were used, or Fisher's exact test if needed. 95% confidence intervals were calculated assuming the number of observed cases followed the Poisson distribution of rare events. Cochran-Armitage test for trend was used to evaluate the proportion increase of syndromal, chromosomal, and multiple malformations across mutational groups. The relation between proportions of twins and congenital abnormalities were tested using standard Pearson correlation. R version 3.0 and SEM software[22] were used for statistics and data-management.

3 | RESULTS

In the oncogenetic database, 16 798 children with a known date of birth were born between 1986 and 2011, of whom 2292 belonged to families with a BRCA mutation and 390 to families with an MMR mutation. The registry contained 11 234 cases diagnosed during the same period, with 10 026 corresponding to the selection criteria. Considering about 364 000 children were born in Auvergne during the 26-year study period, this corresponds to a malformation rate of 2.8%, a rate compatible to the nation-wide estimation equal to 2.4% in 2011 to 2012.[11] Overall, 176 children with congenital anomalies corresponded to children in our oncogenetic database (the matching process is described in the material and methods section).

3.1 | Frequencies of congenital malformation

Because most families recruited at our center consult for breast/ovarian cancer syndrome, 62% of children belonged to HBOC families, that is, exposed to Hereditary Breast or Ovarian Cancer susceptibility (Table 1). The remaining corresponded to Lynch families (12%), families without any cancer syndrome (7%), and families consulting for various other cancer syndromes (19%).

The overall malformation frequency was 1.05% [0.90-1.22]. This frequency did not vary significantly according to the presence or the absence of familial cancer predisposition (1.03% vs 1.19% for both control groups together, $P = 0.55$) (Figure 2).

For children born to a BRCA1 mutated parent, the frequency of malformation (2.47%) was slightly higher than the frequency of both control groups (1.19%, $P = 0.13$), and when compared to the HBOC group without known deleterious mutation (1.04%), the risk for malformation was increased by RR = 2.37 [1.18-4.78], $P = 0.025$. A comparison with all other groups together yielded the same risk increase: 2.42 [1.22-4.81] ($P = 0.025$). This was not the case for BRCA2 (0.84%) or MMR mutations (0.79%).

3.2 | Analysis according to categories of malformation

Malformations were classified in four categories according to extent. The unique type (including single malformations or multiple malformations of a single organ) was the most frequent (75.6%), followed by multiple, that is, malformations concerning several organs (14.4%), chromosomal (7.8%), and syndromal (2.2%). These figures significantly differed from the repartition calculated over the whole registry, respectively 68.0%, 10.0%, 13.7%, and 8.3% ($P = 0.00077$), with an under-representation of syndromal and chromosomal malformations in our families. Distribution varied significantly according to subgroups of cancer risk and parental mutation status (Figure 3).

If syndromal, multiple and chromosomal abnormalities were grouped together under the label "extended," their proportion compared to unique malformations yield significant differences ($P = 0.014$) according to parental mutation status and a significant trend was objectivized ($P = 0.0011$) from the first to the fourth group (Table 2):

Children of deleterious mutation carriers had a 10-fold risk of "extended" malformation in comparison to the control group. A same trend was found if we only considered multi-malformations ($P = 0.0013$). But no significant trend was found for syndromes alone ($P = 0.13$) and chromosomal anomalies ($P = 0.18$).

The frequencies of "extended" malformation per group increased proportionally to the probability that children might carry a deleterious mutation (ie, 100% divided by two in the group of carriers parents, 25% when the mutational status of the parent was unknown but from a mutated family, and so on): the proportion of "extended" malformations (Table 2) varied accordingly if it was calculated either over the total number of malformations or over the number of children per group (respectively $P = 0.008$ and 0.003). Meanwhile in these four groups, the global frequencies of malformation whatever the type (unique or "extended") did not differ ($P = 0.25$).

Finally, as the risk of malformation tends to increase with father's age, due to an increased frequency of de novo mutations,[23,24] fathers' age was compared according to malformation types. Overall, no difference was found between the four groups of parents ($P = 0.28$) although the association between chromosomal malformations and older fathers was close to significance ($P = 0.06$). Mean fathers' age was respectively 31.2 ± 5.7 for unique malformations, multiple 31.4 ± 5.9, syndrome 31.3 ± 12.3, and chromosomal 36.7 ± 10.1. If age of fathers was split into two classes (<45 vs ≥45 years), the relative risk for chromosomal malformation with older fathers was multiplied by 9.6 [3.1-29.4] ($P = 0.01$). Mean mothers' age was similar in all groups ($P = 0.77$), respectively 29.0 ± 5.0, 29.6 ± 4.6, 27.5 ± 3.1, and 29.9 ± 4.9.

TABLE 1 Repartition of abnormalities according to cancer predisposition and parental or familial mutation status

Children's origin	Children number	Families	Malformations	Malf. rate (%)	95%-CI Poisson
Control group 1: from families where no hereditary cancer risk has been diagnosed	**1198**	**291**	**15**	**1.25**	**[0.70%-2.07%]**
from family where a BRCA1 mutation has been diagnosed, including those born to mutated or non-mutated parents	**1353**	**228**	**15**	**1.11**	**[0.58%-1.77%]**
from families with a BRCA1 mutation but unknow status of parents	858		4	0.47	[0.07%-1.05%]
children fathered by a parent carrier of a BRCA1 mutation	324	138	8	2.47	[1.09%-4.99%]
from family where a BRCA1 mutation but born to non-mutated parents (ie not exposed to the familial risk)	171	70	3	1.75	[0.36%-5.13%]
from family where a BRCA2 mutation has been diagnosed, including those born to mutated or non-mutated parents	**961**	**178**	**7**	**0.73**	**[0.36%-1.63%]**
from families with a BRCA2 mutation but unknow status of parents	592		5	0.84	[0.37%-2.18%]
children with a parent carrier of a BRCA2 mutation	239	105	2	0.84	[0.10%-3.02%]
from family where a BRCA2 mutation has been diagnosed but born to non-mutated parents (ie not exposed to the familial risk)	130	51	0	0.00	[0.00%-2.84%]
from family at hereditary breast/ovarian cancer risk, but without any known deleterious mutation diagnosed	**8062**	**1699**	**84**	**1.04**	**[0.83%-1.29%]**
from family with a Lynch syndrome where a MMR mutation has been diagnosed, including those born to mutated or non-mutated parents	**390**	**83**	**4**	**1.03**	**[0.28%-2.63%]**
from families with a MMR mutation but an unknown status of parents	166		2	1.20	[0.15%-4.35%]
children with a parent carrier of a MMR mutation	127	55	1	0.79	[0.02%-4.39%]
from family where a MMR mutation has been diagnosed but born to non-mutated parents (ie not exposed to the familial risk)	97	37	1	1.03	[0.03%-5.74%]
from family at hereditary colon cancer risk, but without any MMR mutation diagnosed in the family	**1613**	**338**	**17**	**1.05**	**[0.61%-1.69%]**
Control group 2: total from families where a mutation has been diagnosed but born to non-mutated parents	**398**	**158**	**4**	**1.01**	**[0.27%-2.57%]**
Total from families where a mutation has been diagnosed and born to mutated parents	**690**	**298**	**11**	**1.59**	**[0.80%-2.85%]**
Excluded group: Children not included because they belong to families with other cancer syndrome	**3255**	**782**	**34**	**1.04**	**[0.72%-1.46%]**
TOTAL	**16798**	**3599**	**176**	**1.05**	**[0.90%-1.22%]**

3.3 | Analysis according to anatomical system concerned by malformations

Malformations were reported for the following anatomical systems: heart in 33% of cases, skeleton 25%, genital 25%, digestive 14%, CNS 11%, face 10%, Down syndrome 9%, and skin 5% (Figure 4).

In our population, systems concerned by malformations did not significantly vary according to cancer risk groups ($P \approx 0.90$).

3.4 | Analysis of multiple births frequency

Because associations have been described between multiple births and congenital malformation incidence, we investigated if frequencies of abnormalities and twinning were associated in our population. Two families registered triple births: they were grouped with twins. The overall frequency of multiple births was 2.47%. No association was found between frequencies of twins and frequencies of malformations ($r = 0.07$, $P = 0.81$). Seven malformations were reported among our 341 twins/triplets, all unique except one, multiple, malformations in one twin born to a BRCA2-mutated parent. Five concerned heart anomalies and one skeletal anomaly, while the only multiple case associated heart, digestive and CNS malformations. Surprisingly, no twins were observed in families with MMR mutations: $P = 0.00023$ vs 3.38% of twins in the control families (Figure 5).

When all cancer syndromes together were compared to the control group, the frequencies of multiple births were similar (respectively 2.95% and 3.38%, $P = 0.32$). No interaction regarding congenital malformations was observed between multiple births and BRCA mutations ($P = 0.89$).

4 | DISCUSSION

In our previous study,[4] we observed fewer miscarriages in cancer-prone families and suspected a possible increase of congenital

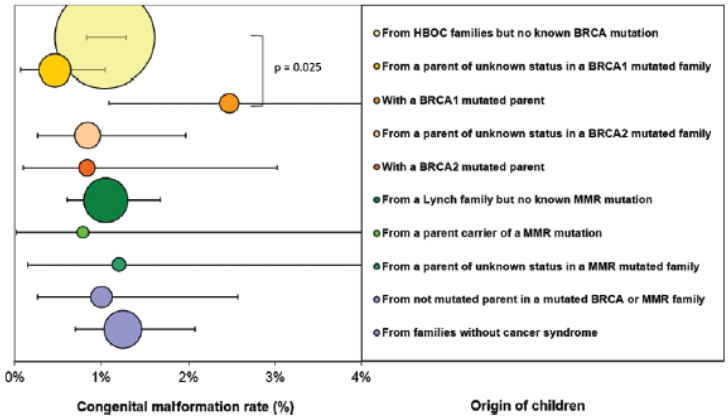


FIGURE 2 Frequency of congenital anomalies per sub-group. Area of circles corresponds to group sample size. Error bars represent Poisson 95%-CI of frequencies. The two bottom groups (no cancer syndrome and children born to non-mutated parents) constitute the "normal" reference frequency

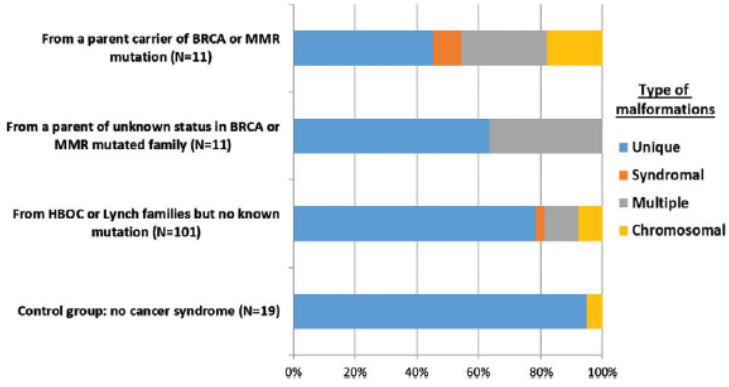


FIGURE 3 proportion of congenital anomalies per sub-group according to anomaly type

malformations. In the present study, cancer predisposition was weakly associated with a higher incidence of congenital abnormalities in offspring. This association was entirely attributable to BRCA1 mutated parents who were associated with an increased risk of 2.37 [95% CI 1.18-4.78] (P = 0.025) compared to HBOC families with no known BRCA mutation and 2.42 [1.22-4.81] (P = 0.025) when compared to anyone. However, the level of evidence of this result is mild and the possibility that the higher frequency of congenital abnormalities in descendants of BRCA1-mutated parents might be due to chance cannot be discarded. BRCA2 as well as MMR mutations did not seem to produce an increase in overall congenital malformation risk.

The main conclusion of our study concerns the type of congenital malformation: the incidence of multimalformations significantly increased with the probability of BRCA or MMR mutation in one of the parents. This trend argues in favor of a direct biological impact of these mutations on embryonic development. These genes are

TABLE 2 Distribution of unique abnormalities vs syndromal, multiple and chromosomal abnormalities together ("Extended") according to parental mutation status

	Malformations		Rate of extended malformation in malformed children			Number of children	Rate of extended malformation in all children		
Parental group	Unique	Extended	Rate (%)	RR	95%-CI		Rate (%)	RR	95% CI
Control groups (no cancer syndrome)	18	1	5.3%	1		1596	0.06%	1	
No known mutation in the family	79	22	21.8%	4.1	[0.8 à 21.7]	9675	0.23%	3.6	[0.6-23.5]
Unknown status in a mutated family	7	4	36.4%	6.9	[1.2 à 38.6]	1602	0.25%	4.0	[0.5-30.2]
Parent a known mutation carrier	5	6	54.5%	10.4	[2.3 à 46.0]	682	0.88%	14.0	[2.8-69.9]

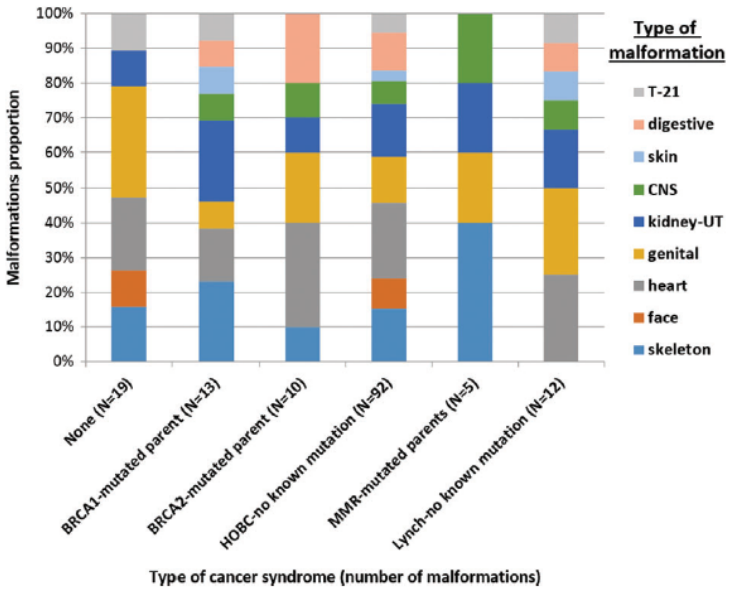


FIGURE 4 anatomical system concerned by malformation according to group of cancer risk (T-21, Down syndrome; CNS, central nervous system; UT, urinary tract)

involved in DNA repair, and it is likely that when repair is less efficient, early genetic anomalies causing malformation may not be corrected and consequences concern several anatomical systems. This also suggests that some congenital malformations result from a reduced capacity of the embryo to repair DNA anomalies, whatever their cause (either spontaneous or related to an inherited anomaly). Such a hypothesis has been proposed in xeroderma pigmentosum in relation to disorders of DNA repair and transcription gene.[25,26] The association of these disorders and the risk for pre-eclampsia[27] is also interesting and confirms the relevance of our working hypothesis: mutations affecting DNA repair and transcription have an impact on both congenital malformations and miscarriage mechanisms.

The slight association of BRCA1 mutations with the risk for congenital malformation may be linked to the interaction between BRCA1 and the Notch signaling pathway. This interaction is a key regulator of breast cell differentiation[28]: dysregulations are associated with basal-like tumors. Shifley et al reported the implication of this pathway in the development of the vertebral column in embryos and the occurrence of congenital skeletal defects when mutations disrupt the segmentation clock function controlled by Notch pathway.[29] Although skeleton malformations were slightly more frequent (23%) when one parent carried a BRCA1 mutation than in other groups (15%), our study is not powered enough to confirm this hypothesis. In rodents, BRCA1 has proven to play an important role in the early development of embryos.[30] Homozygous BRCA1 mutations are lethal: mutant mice die before 2 weeks of embryogenesis.[31] Abnormalities often concern the neural tube, with 40% of the embryos presenting with varying degrees of spina bifida and anencephaly.[32] For Hakem et al,[33] the death of mutant embryos "*may be due to a failure of the proliferative burst required for the development of the different germ layers.*" In

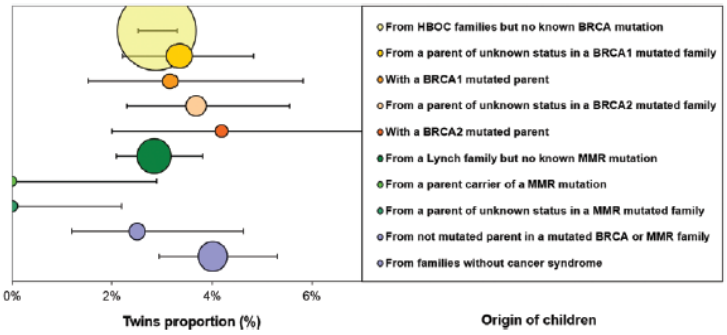


FIGURE 5 Proportion of twins among children registered in our study (error bars correspond to Poisson 95%CI)

humans, no study of the impact of homozygote mutations on BRCA genes is available, likely because embryos are not viable. This is supported by a study of families where both parents carried a BRCA1 mutation[34]: because none of the children carried both mutated alleles, the authors concluded that the most likely reason was that BRCA homozygotes were not viable. Heterozygous mutations have been proposed to expose embryo to a lethal risk unless compensated by other particular genotypes, such as specific alleles of FMR1.[35] This assertion was questioned in a study of Ashkenazi BRCA1/2 mutation carriers where FMR1 sub-types were only slightly unbalanced.[36] The hypothesis of possible genotypic compensations remains relevant and heterozygous mutations could very well favor multimalformations following a similar pattern as described in mice, but depending on a genomic context which remains to be defined.

BRCA2 interacts with a signaling pathway that regulates fibroblast growth factors, and mutations are associated with a wider variety of breast cancer phenotypes than BRCA1 mutations.[37] Double heterozygosity involving BRCA1 and BRCA2 mutations have been reported in the literature,[38] though was not more penetrant than BRCA1 mutation alone. It is unknown if double heterozygosity further increases the risk for congenital malformation.

MMR deficiencies have been associated to agenesis of corpus callosum and gray matter heterotopia[39] and to neural tube defects.[40] One of the three malformed children from MMR mutated families suffered from a CNS anomaly, while the others presented malformations of the kidneys and skeleton. This small population is of course insufficient to confirm any trend.

The rate of congenital abnormalities in Europe was determined in 2010 at 2.3% of all births, among which, 80% of newborns survive.[41] More recent data estimated to be around 3% of livebirths and 15% to 20% of stillbirths in the state of Utah-USA.[42] Higher rates of congenital abnormalities have been reported in France: 3.3% in 2007 in Paris[43] after an 85% increase from 1981 to 2007 of total abnormalities (stillbirths and livebirths together). For the period 1986 to 2011, an overall 2.8% malformation rate was estimated from Auvergne registry (including medical abortions for intra-uterin malformations). French national estimations by National Health Institute (INVS) reported a 2.4% malformation rate for livebirths in 2011 to 2012.[11] Our no cancer-syndrome control group exhibited a 1.4% congenital abnormality frequency, suggesting that the database matching process was only able to flag about half of the expected congenital abnormalities for the children in the oncogenetics database. The origins of this discrepancy may include the way pedigrees are built and how they are updated. Oncogenetics mainly targets cancer in adulthood: for the most recent generation, dates of birth as well as the first and/or last names of children may be omitted. Secondly, probands are asked to update their pedigree when a new cancer is diagnosed, but not for the birth of a child. Therefore, many children born after the pedigree was built are missing. This lack is not related to the familial cancer risk and there is no reason why this would induce any bias in our results. But this limits the accuracy of our estimates and reduces the power of our study. Another reason could be responsible for this low frequency of malformations in our cohort: a younger age of mothers: mothers' age at their children birth was in average 29.5 ± 4.9 years and ranged between 16 and 59. These figures are in accordance with French national statistics (from 29.8 in 2007 to 30.7 in 2017).[43] Finally, the frequency of syndromal and chromosomal malformations in our sample was low when compared to our registry (2% vs 8% in the registry for syndromes, $P = 0.009$ and 8% vs 14%, $P = 0.023$ for karyotype defects): indeed, genetic anomalies corresponding to labeled syndromes rarely induce cancer and those favoring cancer (Nijmegen for example) were excluded. This is also true for chromosomal anomalies (trisomy for example) that highly reduce the risk for solid tumors while it increases the risk for leukemia.[44] Conversely multimalformations were more frequent in our sample (14% vs 10%, $P = 0.05$).

One source of bias may be related to the evolution in recent decades in prenatal diagnosis, where sensitivity of screening has strongly improved and malformations can be diagnosed earlier.[45] For example, in our congenital malformation registry, the frequency of prenatal diagnosis among malformed children doubled from 26.4% before 2000 to 46.9% in recent years. Currently, the most severe malformations are ended by medical termination of pregnancy, while previously these pregnancies resulted in spontaneous abortion or stillbirth and would appear in the registry of congential malformations. Because we included only livebirths, we likely underestimated the frequency of severe malformation. To check if fetal deaths and/or medical abortions in recent years are correlated to BRCA mutations will be the subject of a further study.

Multiple births have been described to increase the risk for congenital anomalies[17–21]: the 3% twins' frequency in our global population was equal to that cited by Boyle et al (2013) which confirms the fitness of our pedigree registration to expected figures. Overall, multiple births doubled the risk for congenital malformation. This increase was not significant ($P = 0.12$), but was similar to the risk of 1.71 [1.43-2.12] found by Glinianaia et al[20] in an English cohort. So, BRCA mutations did not seem to increase the risk for malformation in case of multiple births ($P = 0.80$), but our study was not powered to investigate this issue. We may however conclude that multiple births per category are not likely to bias our statistics. The only particularity found in our pedigrees is the absence of twins in the MMR families. The associated probability when compared to our reference group ($P = 0.00023$) let us suggest this might not be an artifact and MMR mutations may indeed interfere with the mechanisms favoring multiple births.

5 | CONCLUSION

In our study, BRCA or MMR mutations significantly increased the risk for congenital multimalformations. This suggests that DNA repair genes play a role in embryonic development and that some congenital malformations may result from either less efficient repair, or from non-repair functions of these genes. This agrees with the recent review on this issue by Terabayashi et al.[46] This study partially confirms one of our working hypotheses: BRCA1 mutation carriers seem more likely to give birth to children with malformations. Further study is necessary to evaluate the influence of BRCA and MMR mutations on fetal deaths and/or medical abortions in case of fetal anomalies.

ACKNOWLEDGMENTS

This work was supported by FEDER (European Funding of Regional Development) and Conseil Régional Auvergne: convention n°36 760 (2013). These public funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors thank Laurence, Mélanie, Sandra, and Sandrine, secretaries of oncogenetics department, for their constant attention and care with pedigree constitution. They also thank Claire Laquet, genetic counselor nowadays retired, that collected clinical information of family members during decades.

DISCLOSURE OF INTEREST

None.

ORCID

Fabrice Kwiatkowski https://orcid.org/0000-0002-4041-3999

REFERENCES

1. Im KM, Kirchhoff T, Wang X, et al. Haplotype structure in Ashkenazi Jewish BRCA1 and BRCA2 mutation carriers. *Hum Genet*. 2011;130:685-699.
2. Harboe TL, Eiberg H, Kern P, et al. A high frequent BRCA1 founder mutation identified in the Greenlandic population. *Familial Cancer*. 2009;8:413-419.
3. Laraqui A, Uhrhammer N, Lahlou-Amine I, et al. Mutation screening of the BRCA1 gene in early onset and familial breast/ovarian cancer in Moroccan population. *Int J Med Sci*. 2013;10:60-67.
4. Bernirschke K. Chromosomal errors and reproductive failure. *Basic Life Sci*. 1974;4:73-90.
5. Kwiatkowski F, Arbre M, Bidet Y, Laquet C, Uhrhammer N, Bignon Y-J BRCA. Mutations increase fertility in families at hereditary breast/-ovarian cancer risk. *PLoS One*. 2015;10(6):e0127363. https://doi.org/10.1371/journal.pone.0127363.eCollection.
6. Boue J, Boue A, Lazar P. Retrospective and prospective epidemiological studies of 1,500 karyotyped spontaneous abortions. *Teratology*. 1975;12:11-26.
7. Boue A, Boue J, Gropp A. Cytogenetics of pregnancy wastage. *Adv Hum Genet*. 1985;14:1-57.
8. Hassold T, Chen N, Funkhouser J, et al. A cytogenetic study of 1000 spontaneous abortions. *Ann Hum Genet*. 1980;44(Pt 2):151-178.
9. Romero ST, Geiersbach KB, Paxton CN, et al. Differentiation of genetic abnormalities in early pregnancy loss. *Ultrasound Obstet Gynecol*. 2015;45(1):89-94.
10. INSEE – French National Institue for Statistics and Economic Studies. 2010. https://www.insee.fr/fr/statistiques/1291295. Accessed Feburary, 2019.
11. INVS, French National Institute of Health Surveillance. 2016. http://invs.santepubliquefrance.fr/Dossiers-thematiques/Maladies-chroniques-et-traumatismes/Malformations-congenitales-et-anomalies-chromosomiques/Donnees. Accessed Feburary, 2019.
12. Eisinger F, Bressac B, Castaigne D, et al. Identification and management of hereditary predisposition to cancer of the breast and the ovary (update 2004). *Bull Cancer*. 2004;91:219-237.
13. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med*. 2004;23:1111-1130.
14. Vasen HF, Watson P, Mecklin JP, Lynch HT. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the international collaborative group on HNPCC. *Gastroenterology*. 1999;116:1453-1456.
15. Umar A, Boland CR, Terdiman JP, et al. Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J Natl Cancer Inst*. 2004;96(4):261-268.
16. Digweed M, Sperling K. Nijmegen breakage syndrome: clinical manifestation of defective response to DNA double-strand breaks. *DNA Repair*. 2004;3(8–9):1207-1217.
17. Hay S, Wehrung DA. Congenital malformations in twins. *Am J Hum Genet*. 1970;22:662-678.
18. Chen CJ, Wang CJ, Yu MW, Lee TK. Perinatal mortality and prevalence of major congenital malformations of twins in Taipei city. *Acta Genet Med Gemellol*. 1992;41:197-203.
19. Mastroiacovo P, Castilla EE, Arpino C, et al. Congenital malformations in twins: an international study. *Am J Med Genet*. 1999;83:117-124.
20. Glinianaia SV, Rankin J, Wright C. Congenital anomalies in twins: a register-based study. *Hum Reprod*. 2008;23:1306-1311.
21. Boyle B, McConkey R, Garne E, et al. Trends in the prevalence, risk and pregnancy outcome of multiple births with congenital anomaly: a registry-based study in 14 European countries 1984-2007. *BJOG*. 2013;120(6):707-716.
22. Kwiatkowski F, Girard M, Hacene K, Sem BJ. A suitable statistical software adapted for research in oncology. *Bull Cancer*. 2000;87(10):715-721.
23. Kong A, Frigge ML, Masson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*. 2012;488(7412):471-475.
24. Girard SL, Bourassa CV, Lemieux Perreault LP, et al. Paternal age explains a major portion of De novo germline mutation rate variability in healthy individuals. *PLoS One*. 2016;11(10):e0164212. https://doi.org/10.1371/journal.pone.0164212 eCollection 2016.
25. Lehmann AR. DNA repair-deficient diseases, xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy. *Biochimie*. 2003;85(11):1101-1111.
26. Moslehi R, Signore C, Tamura D, et al. Adverse effects of trichothiodystrophy DNA repair and transcription gene disorder on human fetal development. *Clin Genet*. 2010;77(4):365-373.
27. Moslehi R, Ambroggio X, Nagarajan V, Kumar A, Dzutsev A. Nucleotide excision repair/transcription gene defects in the fetus and impaired TFIIH-mediated function in transcription in placenta leading to preeclampsia. *BMC Genomics*. 2014;15:373. https://doi.org/10.1186/1471-2164-15-373.
28. Buckley NE, Nic an tSaoir CB, Blayney JK, et al. BRCA1 is a key regulator of breast differentiation through activation of notch signalling with implications for anti-endocrine treatment of breast cancers. *Nucleic Acids Res*. 2013;41(18):8601-8614.
29. Shifley ET, Cole SE. The vertebrate segmentation clock and its role in skeletal birth defects. *Birth Defects Res*. 2007;81:121-133.
30. Liu CY, Flesken-Nikitin A, Li S, Zeng Y, Lee WH. Inactivation of the mouse BRCA1 gene leads to failure in the morphogenesis of the egg cylinder in early postimplantation development. *Genes Dev*. 1996;10:1835-1843.
31. Evers B, Jonkers J. Mouse models of BRCA1 and BRCA2 deficiency: past lessons, current understanding and future prospects. *Oncogene*. 2006;25(43):5885-5897.
32. Gowen LC, Johnson BL, Latour AM, Sulik KK, Koller BH. BRCA1 deficiency results in early embryonic lethality characterized by neuroepithelial abnormalities. *Nat Genet*. 1996;12:191-194.
33. Hakem R, de la Pompa JL, Sirard C, et al. The tumor suppressor gene Brca1 is required for embryonic cellular proliferation in the mouse. *Cell*. 1996;85:1009-1023.
34. Friedman E, Bar-Sade Bruchim R, Kruglikova A, et al. Double heterozygotes for the Ashkenazi founder mutations in BRCA1 and BRCA2 genes. *Am J Hum Genet*. 1998;63(4):1224-1227.
35. Weghofer A, Tea M-F, Barad DH, et al. BRCA1/2 mutations appear embryo-lethal unless rescued by low (CGG n<26) FMR1 sub-genotypes: explanation for the "BRCA paradox"? *PLoS One*. 2012;7(9):e44753. https://doi.org/10.1371/journal.pone.0044753.
36. Dagan E, Cohen Y, Mory A, et al. BRCA1/2 mutations and FMR1 alleles are randomly distributed: a case control study. *Eur J Hum Genet*. 2014;22(2):277-279.
37. Bane AL, Pinnaduwage D, Colby S, et al. Expression profiling of familial breast cancers demonstrates higher expression of FGFR2 in

BRCA2-associated tumors. *Breast Cancer Res Treat*. 2009;117(1):183-191.
38. Lavie O, Narod S, Lejbkowicz F, et al. Double heterozygosity in the BRCA1 and BRCA2 genes in the Jewish population. *Ann Oncol*. 2011;22(4):964-966.
39. Baas AF, Gabbett M, Rimac M, et al. Agenesis of the corpus callosum and gray matter heterotopia in three patients with constitutional mismatch repair deficiency syndrome. *Eur J Hum Genet*. 2013;21(1):55-61.
40. Liu Z, Wang Z, Li Y, et al. Association of genomic instability, and the methylation status of imprinted genes and mismatch-repair genes, with neural tube defects. *Eur J Hum Genet*. 2012;20(5):516-520.
41. Dolk H, Loane M, Garne E. The prevalence of congenital anomalies in Europe. *Adv Exp Med Biol*. 2010;686:349-364.
42. Feldkamp ML, Carey JC, Byrne JLB, et al. Etiology and clinical presentation of birth defects: population based study. *BMJ*. 2017;357:j2249.
43. INSEE, Bilan démographique; INED, L'évolution démographique en France, Population. 2017; 3. https://www.ined.fr/fr/tout-savoir-population/chiffres/france/naissance-fecondite/age-moyen-maternite/. Accessed Feburary, 2019.
44. Yang Q, Rasmussen SA, Friedman JM. Mortality associated with Down's syndrome in the USA from 1983 to 1997: a population-based study. *Lancet*. 2002;359(9311):1019-1025.
45. Lelong N, Thieulin AC, Vodovar V, Goffinet F, Khoshnood B. Epidemiological surveillance and prenatal diagnosis of congenital anomalies in the Parisian population, 1981–2007. *Arch Pediatr*. 2012;19(10):1030-1038.
46. Terabayashi T, Hanada K. Genome instability syndromes caused by impaired DNA repair and aberrant DNA damage responses. *Cell Biol Toxicol*. 2018;34:337-350.

How to cite this article: Kwiatkowski F, Perthus I, Uhrhammer N, et al. Association between hereditary predisposition to common cancers and congenital multimalformations. *Congenit Anom*. 2019;1–10. https://doi.org/10.1111/cga.12329

8.2.5 Design et performance du pronostic de mutation délétère basé sur les ATCD familiaux

FABRICE KWIATKOWSKI, LAURENT SERLET, ANDRZEJ STOS

A mettre en version anglaise et soumettre

CENTRE JEAN PERRIN CLERMONT-FERRAND ET UNIVERSITÉ CLERMONT AU- VERGNE, LABORATOIRE DE MATHÉMATIQUES (CNRS UMR 6620)

E-mail address: Laurent.Serlet@uca.fr

8.3 Autres métriques utilisées pour la méthode de Ward

8.3.1 Introduction

Ce chapitre a été placé en annexe afin de simplifier la lecture des principaux résultats dans le corps de la thèse. Pour décrire ces métriques, nous allons nous servir de 7 familles générées par le programme POLYGENE tel que décrit dans le chapitre 2.1. Elles ont été choisies de manière à présenter des caractéristiques bien tranchées : 3 sont sans mutation délétère (n°A1-0, A2-0 et A6-0) tandis que les 4 dernières ont des mutations favorisant le cancer soit à 30 ans (n°A43-1, A80-1), soit à 50 ans (A40-3, A54-3), chaque fois avec une pénétrance à 80%.

Dans la figure ci-après, deux familles sont décrites par leurs sous-arbres synthétiques avec la proportion de cancers apparaissant en jaune pour les cancers sporadiques et en rouge pour les cancers familiaux. Les courbes de survie sans cancer associées générées par notre routine Excel sont aussi représentées. Attention, sur ces courbes de survie, l'axe des abscisses commence à 20 ans, ce qui fait que l'occurrence d'un cancer avant 20 ans fait apparaître des courbes débutant plus bas que 100% (ex. famille 43.1).

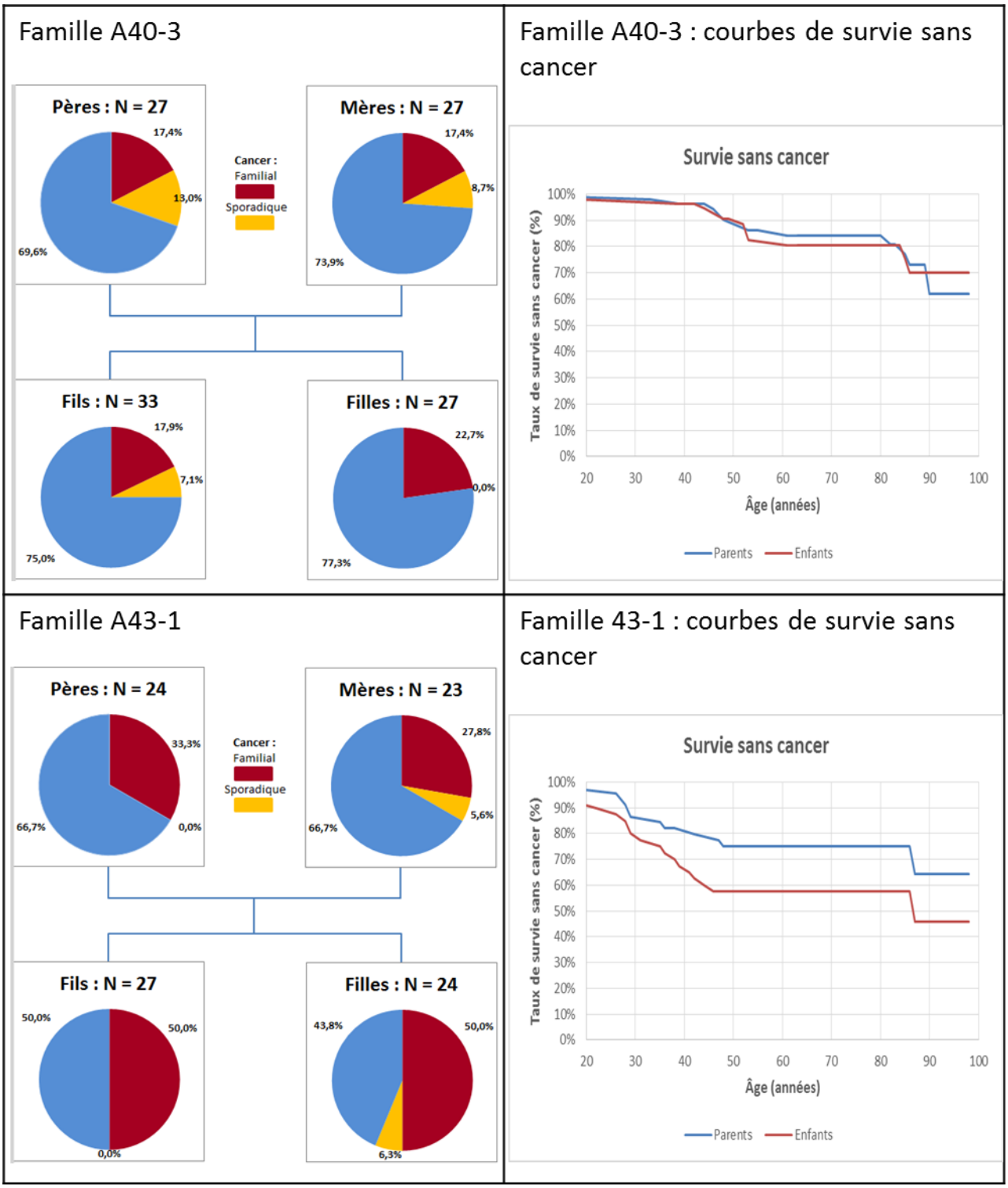


Figure 78 : caractéristiques des familles 40 et 43 mutées

D'autres caractéristiques de ces familles sont données dans le Tableau 30 : la somme des carrés des écarts à la moyenne (SCE) est calculée séparément par sexe pour la partie intra-variable (ou intra-classe) et pour celle inter-variables. Pour vérifier la justesse des calculs, la SCE inter-sexes, c'est à dire la distance entre les centres de gravité par sexe et le centre de gravité global (le point origine en fait puisque les données sont centrées). La SCE totale est aussi calculée : elle cumule pour chaque famille l'inertie due à la dispersion des points pour chaque caractéristique, celle lié à la dispersion des

individus autour de chaque sexe et enfin celle de chaque sexe autour de l'origine. Pour ces calculs, nous avons utilisé le théorème de Huygens qui stipule que l'on peut morceler le calcul des inerties, ce que l'on peut schématiser ainsi (Figure 79) :

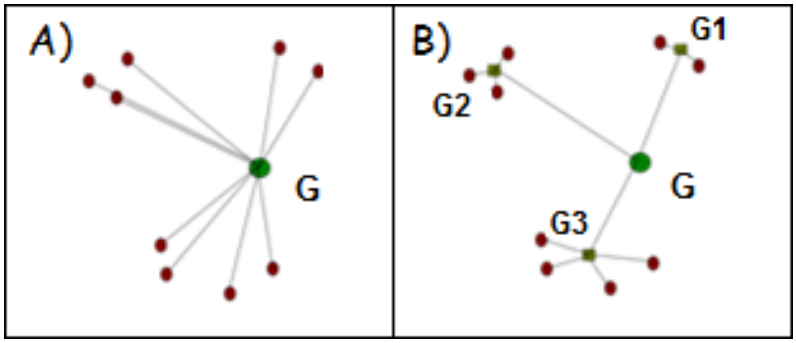


Figure 79 : deux manières de calculer l'inertie du nuage de 9 points e_i de centre de gravité G

L'inertie globale de l'ensemble des 9 points égale $\sum_{i=1}^{9} d^2(e_i, G)$, où d² exprime la distance au carré entre chaque point e_i et G (Figure 79-A), mais elle est aussi égale au cumul des inerties autour de chaque centre de gravité local (G1, G2 et G3) avec les distances au carré entre ces centroïdes locaux et le centre de gravité global G multipliées par le nombre de points par sous-groupe (Figure 79-B). Appliqué à nos 7 familles, l'inertie peut être répartie entre la SCE intra-variables (ou intra-classe), l'inertie inter-variables et enfin l'inertie inter-sexes. Pour le clustering, il existe un large éventail de critères ainsi que le suggèrent Whishart[73] ou encore Murtagh et Legendre[74]. Quant à nous, c'est certaines parties de l'inertie qui nous semblent les plus intéressantes. Ces inerties sont résumées dans le Tableau 30.

Tableau 30 : calcul des différentes inerties à partir des items (et non des individus)

Familles :	**A1-0**	**A2-0**	**A6-0**	**A40-3**	**A43-1**	**A54-3**	**A80-1**
Nombre de membres :	123	135	77	112	92	40	88
Nombre d'hommes :	65	68	40	58	47	21	42
Nombre de femmes :	58	67	37	54	45	19	46
Nb valeurs Hommes :	518	542	325	474	388	180	341
Nb valeurs Femmes :	527	615	328	492	412	163	424
SCE intra-var. hommes :	397	494	137	378	274	134	253
SCE intra-var. femmes :	470	523	193	476	542	122	371
SCE inter-var. hommes :	65	67	24	58	64	29	57
SCE inter-var. femmes :	66	74	54	62	74	22	54
SCE inter-sexes :	29	38	21	37	17	3	7
SCE totale :	1027	1194	429	1011	970	310	743

Comme on peut le constater, les inerties intra-variables sont de loin majoritaires. Viennent en second les inerties inter-variables. Les deux lignes "cumuls SCE précédentes" et "SCE totale" sont le résultat de l'inertie toutes composantes confondues, l'une calculée en sommant les inerties intermédiaires et la dernière à partir de l'ensemble des points directement. Cela permet de vérifier la justesse des calculs. Les variables utilisées incluaient le sexe, l'âge des parents à la naissance, la présence de malformations à la naissance (aucune, unique ou multiples), l'occurrence d'un cancer (oui/non) et l'âge d'arrivée de ce cancer, pour les femmes l'âge de la ménopause et enfin des données liées à la natalité : l'âge de mariage, le nombre d'enfants et le nombre de fausses couches.

8.3.2 Quelle métrique utiliser pour distinguer les familles les plus proches ?

8.3.2.1 Utilisation de l'inertie (somme des carrés des écarts aux centroïdes)

Quand on recherche les familles les plus proches en utilisant l'algorithme utilisé ci-dessus, c'est à dire en moyennant les distances interindividuelles (Tableau 31), la paire de familles A1-0 et A6-0 (carré gris foncé) est classée en premier, puis celle des familles A6-0 et A54-3 (gris moyen) et enfin la paire A2-0 et A6-0 (carré gris clair).

Tableau 31 : tableau croisé des distances interfamiliales calculées à l'aide des distances interindividuelles

Familles	A2-0	A6-0	A40-3	A43-1	A54-3	A80-1
A1-0	1,13	1,00	1,15	1,21	1,11	1,16
A2-0		1,04	1,18	1,24	1,14	1,19
A6-0			1,07	1,13	1,04	1,09
A40-3				1,24	1,16	1,20
A43-1					1,21	1,24
A54-3						1,15

Tableau 32 : tableau croisé des distances interfamiliales calculées à l'aide des distances interindividuelles mais après correction par la variance commune

Familles	A2-0	A6-0	A40-3	A43-1	A54-3	A80-1
A1-0	26,73	22,08	25,95	25,02	18,34	25,02
A2-0		22,85	27,11	26,02	18,83	25,76
A6-0			22,84	22,14	18,38	22,70
A40-3				24,91	18,67	25,04
A43-1					17,87	23,61
A54-3						18,23

La correction change de manière importante la classification des distances interfamiliales, avec dans le premier cas les paires A1-0 et A6-0, puis la paire A6-0 et A54-3 et en 3ème la paire A2-0 et A6-0. Avec les distances corrigées, sortent dans l'ordre décroissant la paire A43-1 et A54-3, la paire A54-3 et A80-1 et enfin la paire A1-0 avec A54-3. Les distances brutes favorisent le regroupement des familles à faible effectif ce que font moins les valeurs corrigées. Avec ces dernières, la 3ème place associe une famille sans risque de cancer (A1-0) avec une autre à risque intermédiaire A54-3, ce qui n'est pas optimal non plus.

Les inerties familiales résultant du regroupement deux à deux des 7 familles, en distinguant le sexe des individus, sont les suivantes :

Tableau 33 : inertie (ou SCE) intra-variables des paires de familles regroupées

Familles	A2-0	A6-0	A40-3	A43-1	A54-3	A80-1
A1-0	1892	1222	1737	1729	1135	1519
A2-0		1371	1890	1887	1288	1674
A6-0			1215	1194	609	998
A40-3				1696	1127	1497

A43-1					1091	1452
A54-3						893

Comment s'évalue le gain d'inertie quand on regroupe deux familles ? Par simple soustraction : ainsi pour le regroupement des familles 1 et 2, l'inertie intra-classes est 1892 tandis que les inerties initiales (hors celles inter-sexes du Tableau 30) étaient (397 + 494) pour les hommes et (470 + 523) pour les femmes soit environ 8 (en ignorant les arrondis), ce que l'on retrouve dans le Tableau 34-A.

On a toutefois le choix quant au type d'inertie que l'on considère. En effet, si l'on suppose que plus des familles seront distinctes et plus l'inertie inter-variables variera, on doit considérer cette dernière dans les calculs et omettre les autres. Voici les variations d'inertie selon le point de vue considéré :

Tableau 34 : gains/pertes d'inertie selon la composante inertielle par paire de familles (les cellules en gris correspondent aux meilleures valeurs : plus le gris est foncé et plus l'association est forte

A)

	Familles	A2-0	A6-0	A40-3	A43-1	A54-3	A80-1
Inertie intra-variables	A1-0	8,40	25,25	15,60	45,71	12,09	28,02
	A2-0		24,36	19,18	54,04	14,89	32,78
	A6-0			30,91	47,95	22,74	44,32
	A40-3				26,42	16,42	18,99
	A43-1					18,86	12,49
	A54-3						12,95

B)

	Familles	A2-0	A6-0	A40-3	A43-1	A54-3	A80-1
Inertie inter-variables	A1-0	-7,49	-18,50	-14,58	-45,49	-10,29	-26,00
	A2-0		-12,65	-19,06	-52,61	-13,15	-29,21
	A6-0			-20,21	-41,93	-14,20	-35,89
	A40-3				-24,66	-14,08	-14,66
	A43-1					-17,72	-11,63
	A54-3						-12,61

C)

	Familles	A2-0	A6-0	A40-3	A43-1	A54-3	A80-1
Inertie inter-sexes	A1-0	0,91	6,76	1,02	0,22	1,81	2,02
	A2-0		11,71	0,12	1,43	1,73	3,58
	A6-0			10,70	6,01	8,54	8,43
	A40-3				1,76	2,34	4,33
	A43-1					1,15	0,86
	A54-3						0,34

D)

	Familles	A2-0	A6-0	A40-3	A43-1	A54-3	A80-1
inertie globale	A1-0	1,28	6,77	0,72	0,000	0,40	0,13
	A2-0		13,23	0,06	1,14	0,020	0,45
	A6-0			10,96	5,97	6,44	7,58
	A40-3				0,66	0,001	0,19
	A43-1					0,39	0,12
	A54-3						0,12

En comparaison du Tableau 31, l'utilisation comme critère de regroupement le gain d'inertie minimal produit des classements assez différents : seule la variation d'inertie intra-variables arrive à regrouper préférentiellement des familles à risque similaire (2 sur 3). Il en est de même pour l'inertie inter-variables qui varie en miroir de la précédente : cela veut dire que lorsque l'on regroupe les familles, les distances inter-variables diminuent alors que la variance autour de ces mêmes variables augmente. Cela paraît intéressant car les différences entre familles sont principalement localisées sur les paramètres liés au cancer. A mesure qu'on considère l'inertie à une échelle plus globale, ces différences relativement ténues disparaissent alors dans la globalité.

Avec les inerties inter-sexes et globales, les familles les plus proches sont alors la A1-0 et la A43-1 soit une famille sans risque héréditaire et l'autre avec le plus haut risque. La seconde paire regroupe deux familles mutées (A40-3 et A54-3), ce qui est acceptable et enfin la troisième position relie la famille A2-0 et A54-3 avec la même disparité que pour la première paire. On peut néanmoins remarquer que ces deux métriques ont tendance à privilégier le regroupement des familles à faible effectif. Ici, la famille A54-3 ne comprend que 40 membres. De fait, son inertie globale est d'emblée faible et son association à d'autres familles induit naturellement le moins de changement. Pour remédier à ce défaut, on peut être tenté par une métrique basée sur la variance.

8.3.2.2 Utilisation des modifications de variance pour la proximité des familles

L'inertie et la variance sont étroitement liées la seconde étant l'inertie divisée par l'effectif. Tout comme l'inertie, la variance est cumulative quand on associe k ensembles de points, avec la formule suivante :

$V = \frac{1}{N}\sum_{i=1}^{k} n_i(V_i + (M - M_i)^2)$

Où $N = \sum_{i=1}^{k} N_i$, les effectifs par ensemble, M_i et V_i leur moyenne et leur variance respectivement, et M la moyenne globale.

L'intérêt de cette formule est qu'elle suggère un moyen d'obtenir la distance entre deux groupes : dans notre contexte, prenons deux familles A et B d'effectif n_A et n_B avec $N = n_A + n_B$. La variance globale V_{AB} devient :

$V_{AB} = \frac{1}{N}\ [\ n_A V_A + n_B V_B + n_A(M - M_A)^2 + n_B(M - M_B)^2\ ]$

d'où $\quad n_A(M - M_A)^2 + n_B(M - M_B)^2 = N\,V_{AB} - n_A V_A - n_B V_B$

Les moyennes M, M_A et M_B correspondent aux centroïdes des nuages de points respectifs $A \cup B$, A et B. Ils sont alignés. Soit G le centre de gravité global. Les distances d_A de A à G et d_B de B à G respectent les propriétés suivantes par rapport à d_{AB}, la distance de A à B :

$d_{AB} = d_A + d_B$ et $n_A d_A = n_B d_B$ (définition du centre de gravité)

d'où $d_A = (\frac{n_B}{n_A+n_B})\, d_{AB}$ et $d_B = (\frac{n_A}{n_A+n_B})\, d_{AB}$

$n_A(M - M_A)^2 + n_B(M - M_B)^2 \iff n_A\,[\frac{n_B}{n_A+n_B} d_{AB}]^2 + n_B\,[\frac{n_A}{n_A+n_B} d_{AB}]^2 = (\frac{n_A\, n_B\,(}{n_A+n_B})\, d_{AB}^2$

=> $V_{AB} - \frac{1}{N}(n_A V_A + n_B V_B)$ est bien une fonction strictement monotone croissante de d_{AB}, ce qui était nécessaire pour pouvoir l'utiliser pour le classement des distances inter-familles

La minimisation de la seconde partie de l'équation, $V_{AB} - \frac{1}{N}(n_A V_A + n_B V_B)$ correspondant à la différence entre la variance globale et la variance commune, est donc un bon moyen pour trouver les familles les plus proches. Elle a un gros avantage par rapport au calcul de la moyenne des distances interindividuelles : elle est beaucoup plus rapide à calculer requérant $n_A + n_B$ opérations quand la dernière en nécessite $n_A \times n_B$. Cela n'est pas sans conséquence quand on analyse les données de milliers de familles contenant parfois des centaines de membres, eux-mêmes caractérisés par plusieurs dizaines de paramètres.

Comme pour l'utilisation de l'inertie précédemment, deux approches ont été implémentées : la première avec un calcul des moyennes et variances en distinguant les hommes et les femmes et l'autre en ignorant cette caractéristique.

Tableau 35 : classement des paires de familles en utilisant la différence entre la variance commune et les variances de chaque famille (en séparant les sexes)

Familles	A2-0	A6-0	A40-3	A43-1	A54-3	A80-1
A1-0	0,00041	0,00398	0,00051	0,00012	0,00130	0,00112
A2-0		0,00647	0,00006	0,00073	0,00115	0,00186
A6-0			0,00661	0,00414	0,00857	0,00594
A40-3				0,00100	0,00179	0,00250
A43-1					0,00100	0,00055
A54-3						0,00031

Tableau 36 : classement des paires de familles en utilisant la différence entre la variance commune et les variances de chaque famille (en ignorant les sexes)

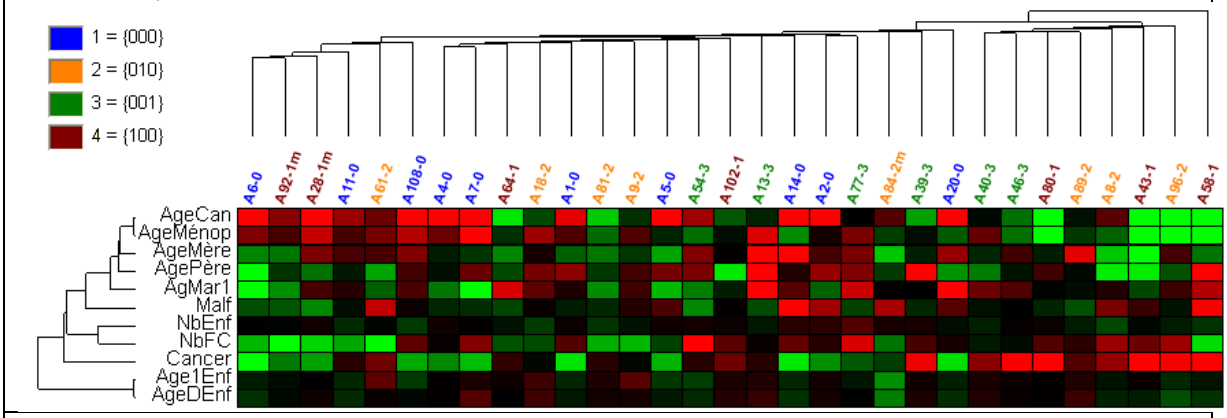
Familles	A2-0	A6-0	A40-3	A43-1	A54-3	A80-1
A1-0	0,00058	0,00398	0,00036	0,00000	0,00029	0,00007
A2-0		0,00731	0,00003	0,00058	0,00001	0,00024
A6-0			0,00677	0,00411	0,00647	0,00535
A40-3				0,00037	0,00000	0,00011
A43-1					0,00034	0,00008
A54-3						0,00011

L'approche en séparant les hommes et les femmes n'est guère meilleure (Tableau 35) que quand on ignore le genre (Tableau 36). Mais dans les deux cas, le classement apparaît peu efficace : en particulier la paires de familles A0-1 et A43-1 se trouve associée alors qu'elle ne le devrait pas.

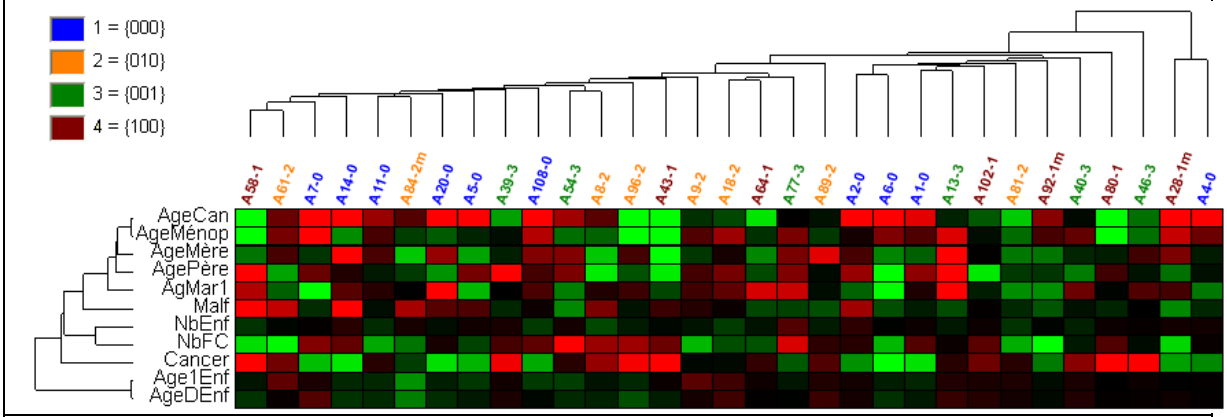
Nous avons corrélé les résultats de ces calculs et les corrélations se sont avérées médiocres. Ces divers indicateurs semblent donc refléter des aspects différents de la distance interfamiliale. Il pourra être intéressant d'évaluer la performance de tous ces critères sur un jeu d'essai conséquent. Sans doute certains critères s'avéreront nettement plus efficaces que d'autres dans le clustering.

En attendant, voici le résultat obtenu en clustering à l'aide de ces 4 indicateurs sur les 31 familles tests :

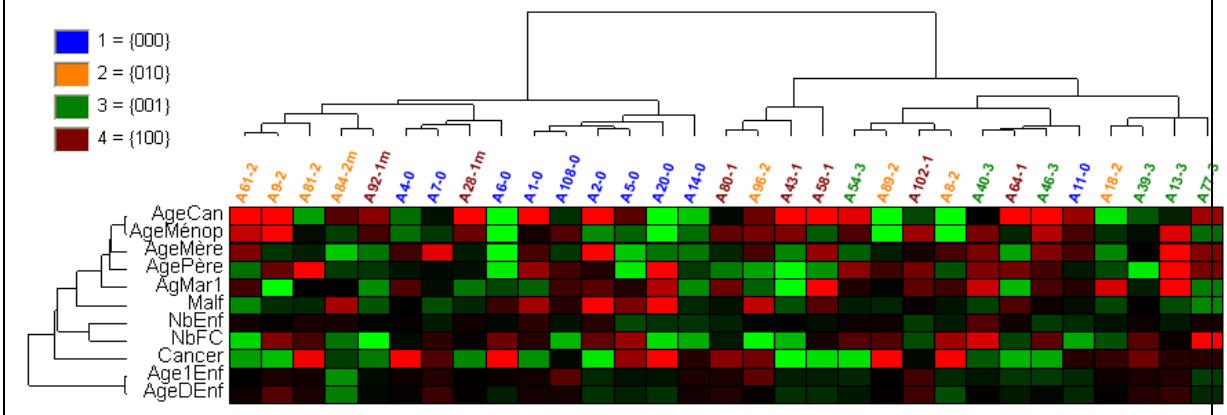
1) Moyenne des distances inter-individuelles



2) Moyenne des distances inter-individuelles corrigées



3) Variation minimale de l'inertie intra-variables



4) Variance globale – variance commune

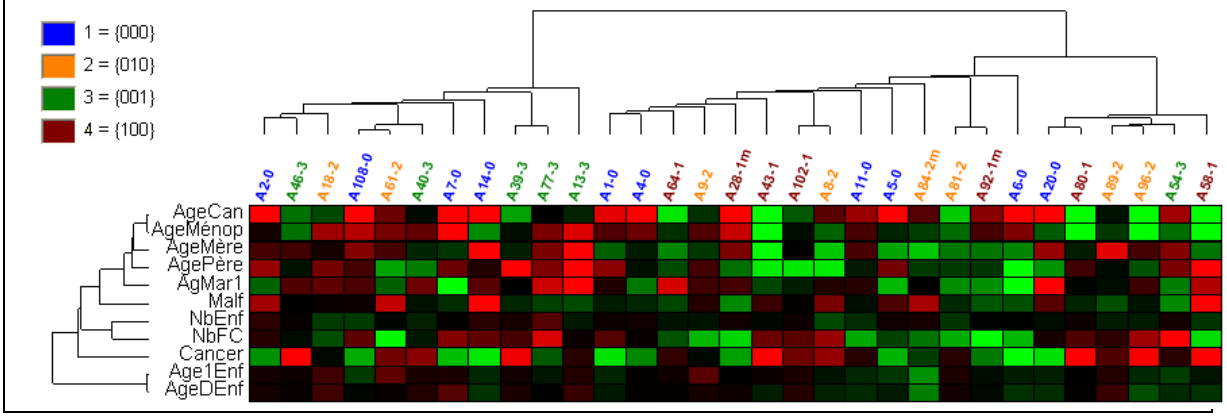


Figure 80 : Classements des 31 familles tests obtenus à l'aide des 4 indicateurs étudiés

On observe que le classement fourni par l'approche avec la variation minimale de SCE est le plus intéressant puisque les 10 familles non mutées sont regroupées dans un seul et unique cluster, le 2ème, et que seulement deux « intrus » y sont inclus. Toutes les autres approches fournissent des classements peu intéressants, au moins en ce qui concerne la méthode de Ward. Il est clair qu'avec les distances inter-individuelles moyennes, voire la différence entre variances globale et commune, on pourrait aussi utiliser un groupement par sauts (moyen, minimum ou maximum).

Dernière remarque, si l'on recalcule le clustering de la Figure 80-3 sans stratifier sur le sexe, on obtient un dendrogramme nettement moins performant quant au classement des familles :

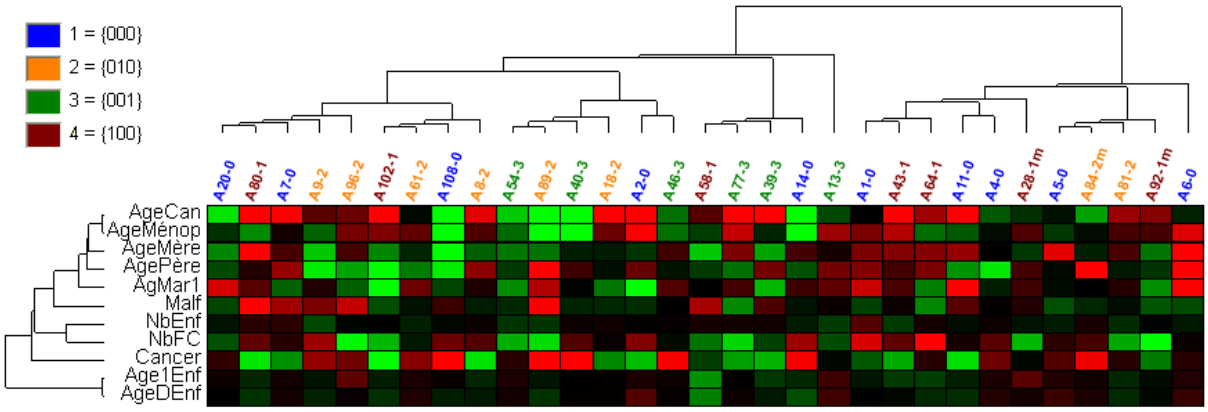


Figure 81 : clustering basée sur la perte minimale d'inertie, mais sans tenir compte du sexe

Les regroupements sont rares et les familles semblent réparties indépendamment du risque de cancer associé. Ces premières analyses confirment que la prise en compte séparément des composantes de l'inertie est indispensable pour que le clustering réalise les regroupements attendus. Les tests de validation de ces nouvelles approches se feront donc en tenant compte de cette conclusion.

8.3.2.3 Quelques remarques sur la programmation des calculs

Ce genre de calculs reposant sur une double hiérarchie n'est pas simple à mettre en œuvre, en particulier si l'on veut que les temps de calcul soient acceptables. Plusieurs tables ont donc été utilisées :

- TTd : La table des données brutes indexée sur le n° d'individu (autant de données par individu que de variables à analyser dans le clustering)
- Mx : La table des données centrées et réduites (avec la valeur "Null" pour les données manquantes) : une ligne par individu et autant de colonnes que de variables
- Trgr : Une table d'association individu-famille : cette table contient une ligne par individu. Elle contient en outre l'adresse de l'individu dans Mx mais aussi la valeur que prend la variable "sexe" par individu. Ceci est nécessaire pour effectuer les calculs en stratifiant sur le sexe. La table est triée par n° de famille.

- TFam : La table des familles : en plus du n° de famille, elle contient les rangs début/fin des individus de la famille dans Trgr, ce qui permet ensuite un pointage vers Mx.
- TDc_F : au fur et à mesure que les familles sont rassemblées, elle mémorise le rang de début de de fin des familles dans TFam. Elle contient aussi quelques résultats intermédiaires nécessaires aux calculs.
- D'autres tables ne sont pas décrites : elles ne servent que pour les calculs internes du clustering.

Quel est l'avantage de cette organisation ? C'est de ne pas réitérer à chaque fois que l'on regroupe des familles le tri des données de base (Mx ou Trgr) qui sont les plus nombreuses: cela nécessiterait du temps machine en quantité. Ce qui est effectué en lieu et place de ce tri, c'est le déplacement unique des enregistrements dans TFam. Quand deux familles (ou plus) sont regroupées, on descend dans TFam l'enregistrement de rang le plus élevé vers celui le plus bas. Dans TDc_F, on ajoute 1 dans le rang de fin du groupe de familles correspondant. Ensuite on renumérote les rangs suivants de TDc_F qui doivent pointer vers des familles décalées d'une position supplémentaire, etc. Lors des calculs d'inertie, il n'y a plus qu'à faire une routine en lui précisant de calculer de la famille de rang x à celle de rang y (rangs mémorisés dans TDc_F), ces familles pointant vers autant d'individus dans Trgr, eux-mêmes associés à autant de valeurs dans Mx (Figure 82) :

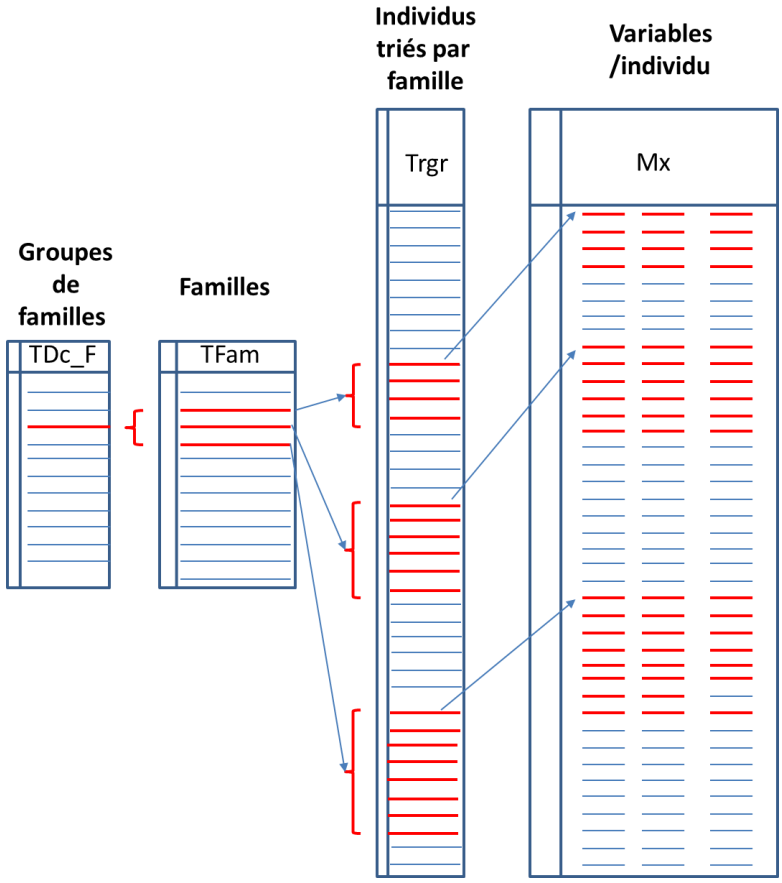


Figure 82 : tables permettant de gérer pendant le clustering l'accès aux données individuelles, même après regroupement de plusieurs familles.

Les calculs de base ont été faits concurremment et dans EXCEL et dans le logiciel SEM sans copie des lignes de code, ceci afin de vérifier les calculs : ce n'était pas du luxe. Pour les 31 familles tests, cela

constituait quand même le maniement de 2 968 individus soit environ 30 000 données. Le débogage des programmes s'en est trouvé fortement ralenti, mais il était difficile de tester le clustering sur un plus petit échantillon de familles.

Dans le chapitre suivant, la performance des diverses approches vont être comparées sur plusieurs centaines de familles afin de mettre en évidence d'une part la meilleure approche en clustering hiérarchique hiérarchique, mais aussi de les comparer à l'utilisation de variables synthétiques par famille (c'est à dire sans revenir aux individus).

8.4 Liste des publications auxquelles j'ai collaboré

1 VERRELLE P., MEISSONNIER F., FONCK Y., FEILLEL V., DIONET C., KWIATKOWSKI F., PLAGNE R., CHASSAGNE J._Clinical relevance of immunohistochemical detection of multidrug resistance P-glycoprotein in breast carcinoma., *J Natl Cancer Inst ; (*1991). **83(2) :** 111-116 - (184 cit.)

2 BELEMBAOGO E., FEILLEL V., CHOLLET P., CURE H., VERRELLE P., KWIATKOWSKI F., ACHARD J.L., LE BOUEDEC G., CHASSAGNE J., BIGNON Y.J., DE LATOUR M., LAFAYE C., DAUPLAT J._Neoadjuvant chemotherapy in 126 operable breast cancers., *Eur J Cancer ; (*1992). **28A(4-5) :** 896-900 - I.F. 1,300 (73 cit.)

3 LORIMIER G., DAVER A., BASUYAU J.P., BRUNELLE P., GAILLARD G., BESSE G., KWIATKOWSKI F., WAFFLART J., ANGIBEAU R.M., AUVRAY E., BENOIST L. Prognostic value of total cathepsin D in breast cancer patients undergoing surgery directly. A multicenter study., *15th Annual San Antonio Breast Cancer Symposium. San Antonio (Texas, U.S.A.), december 8-9, 1992. ; (*1992)

4 SOUBRIER M., BITAR S., KWIATKOWSKI F., DUBOST J.J., BUSSIERE J.L., SAUVEZIE B. Localisation du 1er tassement vertébral porotique, *Rachis ; (*1993). **5 :** 21-24 .

5 TAILLANDIER J., PERISSEL B., KWIATKOWSKI F., BOITEUX J.P., MALET P., GIRAUD B. Analyse chromosomique des tumeurs de vessie. Aspects techniques, corrélations anatomocliniques et perspectives. A propos de 18 cas. *Progrès en Urologie : journal de l'Association Française d'Urologie, de la Société Française d'Urologie et de l'Association des Urologues du Québec ; (*1993). **3 :** 395-405 .

6 ATTIA-SOBOL J., FERRIERE J.P., CURE H., KWIATKOWSKI F., ACHARD J.L., VERRELLE P., FEILLEL V., DE LATOUR M., LAFAYE C., DELOCHE C., DAUPLAT J., DOLY A., ROZAN R., CHOLLET P. Treatment results, survival and prognostic factors in 109 inflammatory breast cancers : univariate and multivariate analysis., *Eur J Cancer ; (*1993). **29A(8) :** 1081-1088 - I.F. 1,300 (38 cit.)

7 JANNY P., CURE H., MOHR M., HELDT N., KWIATKOWSKI F., LEMAIRE J.J., PLAGNE R., ROZAN R. Low grade supratentorial astrocytomas. Management and prognostic factors. *Cancer ; (*1994). **73 :** 1937-1945 - I.F. 3,000 (126 cit.)

8 FLEURY J., LEGROS M., CURE H., TORTOCHAUX J., CONDAT P., DIONET C., TRAVADE P., BELEMBAOGO E., TAVERNIER F., KWIATKOWSKI F., CHOLLET P., PLAGNE R. The hematopoietic stem cell transplantation in Hodgkin's disease : questions and controversies., *Leuk Lymphoma ; (*1994). **15 :** 419-432 - I.F. 1,559 (3 cit.)

9 FLEURY J., TORTOCHAUX J., LEGROS M., CURE H., KWIATKOWSKI F., FERRIERE J.P., TRAVADE P., DIONET C., GAILLARD G., CHASSAGNE J., BIGNON Y.J., VAN PRAAGH I., BELEMBAOGO E., CHOLLET P., PLAGNE R. Valeur pronostique de la bêta-2-microglobuline dans la maladie de Hodgkin de l'adulte jeune., *Bull Cancer ; (*1994). **81 :** 625-631 - I.F. 1,000 (1 cit.)

10 BESSE G., KWIATKOWSKI F., GAILLARD G., DAVER A., DALIFARD I., BASUYAU J.P., BRUNELLE P., WAFFLART J., ANGIBEAU R.M., AUVRAY E., GOUSSARD J. Rôle pronostique de la protéine p52 dans 1065 cas de cancers du sein. Etude multicentrique., *Bull Cancer ; (*1994). **81 :** 289-296 - (7 cit.)

11 LE BOUEDEC G., ACHARD J.L., KAUFFMANN P., KWIATKOWSKI F., GIOANNI G., DAUPLAT J. Traitement radio-chirurgical des cancers du col utérin supérieurs ou égaux à 4 cm., *JOBGYN ; (*1995). **3 :** 35-43 .

12 MARTINEAU-PIVOTEAU N., CUSSAC-BUCHDAHL C., CHOLLET P., ROLHION C., DEBITON E., RAPP M., KWIATKOWSKI F., MADELMONT J.C., LEVI F. Circadian variation in O6-methylguanine-DNA methyltransferase activity in mouse liver., *Anticancer Drugs ; (*1996). **7(6) :** 703-709 - (4 cit.)

13 MARTINEAU-PIVOTEAU N., LEVI F., ROLHION C., KWIATKOWSKI F., LEMAIGRE G., FILIPSKI E., CHOLLET P. Circadian rhythm in toxic effects of cystemustine in mice : relevance for chronomodulated delivery., *Int J Cancer ; (*1996). **68(5) :** 669-674 - I.F. 3,534 (5 cit.)

14 SLIM K., BOUSQUET J., KWIATKOWSKI F., PEZET D., CHIPPONI J. Analysis of randomized controlled trials in laparoscopic surgery., *Br J Surg ; (*1997). **84 :** 610-614 - I.F. 2,429 (45 cit.)

15 TCHIRKOV A., BAY J.O., PERNIN D., BIGNON Y.J., RIO P., GRANCHO M., KWIATKOWSKI F., GIOLLANT M., MALET P., VERRELLE P. Detection of heterozygous carriers of the ataxia-telangiectasia (ATM) gene by G2 phase chromosomal radiosensitivity of peripheral blood lymphocytes., *Hum Genet ; (*1997). **101(3) :** 312-316 - I.F. 2,445 (27 cit.)

16 KASIA J.M., RAIGA J., DOH A.S., BIOUELE J.M., POULY J.L., KWIATKOWSKI F., EDZOA T., BRUHAT M.A. Laparoscopic fimbrioplasty and neosalpingostomy. Experience of the Yaoundé General Hospital, Cameroon (report of 194 cases)., *Eur J Obstet Gynecol Reprod Biol ; (*1997). **73(1) :** 71-77 - (10 cit.)

17 FERRIERE J.P., CHARRIER S., CURE H., KWIATKOWSKI F., COURTADON M., BELEMBAOGO E., DE LATOUR M., ACHARD J.L., DAUPLAT J., CHOLLET P. Adjuvant chemotherapy with doxorubicin-containing regimen for 326 stage II breast cancers : 15-year results., *Am J Clin Oncol ; (*1997). **20 :** 219-225 - (1 cit.)

18 BACIN F., KWIATKOWSKI F., DALENS H., ROZAN R., GAGYI S., DONNARIEIX D., BARD J.J., ROBERT J.L. Résultats à long terme de la curiethérapie par le cobalt 60 des mélanomes de l'uvée., *J Fr Ophtalmol ; (*1998). **21 :** 333-344 - I.F. 0,162 (6 cit.)

19 FERRIERE J.P., ASSIER I., CURE H., CHARRIER S., KWIATKOWSKI F., ACHARD J.L., DAUPLAT J., CHOLLET P. Primary chemotherapy in breast cancer : correlation between tumor response and patient outcome., *Am J Clin Oncol ; (*1998). **21 :** 117-120 - (45 cit.)

20 TCHIRKOV A., GIOLLANT M., TAVERNIER F., BRIANCON G., TOURNILHAC O., KWIATKOWSKI F., PHILIPPE P., CHOUFI B., DEMEOCQ F., TRAVADE P., MALET P. Interphase cytogenetics and competitive RT-PCR for residual disease monitoring in patients with chronic myeloid leukaemia during interferon - alpha therapy, *Br J Haematol ; (*1998). **101 :** 552-557 - I.F. 3,209 (23 cit.)

21 GOSSE-BRUN S., SAUVAIGO S., DAVER A., LARRA F., KWIATKOWSKI F., BIGNON Y.J., BERNARD-GALLON D. Association between H-ras minisatellite and colorectal cancer risk., *Anticancer Res ; (*1998). **18 :** 2611-2616 - I.F. 1,236 (8 cit.)

22 VAURS-BARRIERE C., VIDAL V., PENAULT-LLORCA F., KWIATKOWSKI F., MAUGARD C., BIGNON Y-J. Pathology of sporadic breast tumors with LOH at the BRCA1 locus : correlation with histopathological features specific to familial BRCA1 tumors and absence of microsatellite instability., *Int J Oncol ; (*1998). **12 :** 1373-1378 - I.F. 1,040 (9 cit.)

23 RIO P.G., PERNIN D., BAY J.O., ALBUISSON E., KWIATKOWSKI F., DE LATOUR M., BERNARD-GALLON D.J., BIGNON Y.J. Loss of heterozygosity of BRCA1, BRCA2 and ATM genes in sporadic invasive ductal breast carcinoma., *Int J Oncol ; (*1998). **13(4) :** 849-853 - I.F. 1,040 (41 cit.)

24 HENOU C., DIDIER E., de LATOUR M., KWIATKOWSKI F., COMMUNAL Y., BESSE G., BIGNON Y.J. Association of N-acetyllactosamine with tumor progression in human breast cancer : a study using a 16 kDa chick embryo lectin., *Int J Mol Med ; (*1998). **1 :** 771-775 .

25 BERLIE J., STEVENS D., STALAIN I., ETTER E., GIRARD M., ROUESSE J., KWIATKOWSKI F. Changes in breast cancer patient characteristics at the Rene-Huguenin Center from 1956 to 1996., *Sem Hop ; (*1998). **74 :** 1170-1174 .

26 RIO P., PERNIN D., KWIATKOWSKI F., DE LATOUR M., BERNARD-GALLON D., BAY J.O., BIGNON Y.J. Analysis of allelic losses of BRCA1, BRCA2 and ATM genes in sporadic invasive ductal breast carcinoma., *Dis Markers ; (*1998). **14 :** 60-61 - I.F. 0,250

27 SLIM K., BOUSQUET J., KWIATKOWSKI F., LESCURE G., PEZET D., CHIPPONI J. Première validation de la version française de l'index de qualité de vie pour les maladies digerstives., *Gastroenterol Clin Biol ; (*1999). **23 :** 25-31 - (73 cit.)

28 LAPLACE-MARIEZE V., PRESNEAU N., SYLVAIN V., KWIATKOWSKI F., LORTHOLARY A., HARDOUIN A., BIGNON Y.J. Systematic sequencing of the BRCA-1 coding region for germ-line mutation detection in 70 French high-risk families., *Int J Oncol ; (*1999). **14 :** 971-977 - I.F. 1,381 (5 cit.)

29 UHRHAMMER N., BAY J., PERNIN D., RIO P., GRANCHO M., KWIATKOWSKI F., GOSSE-BRUN S., DAVER A., BIGNON Y.J. Loss of heterozygosity at the ATM locus in colorectal carcinoma., *Oncol Rep ; (*1999). **6 :** 655-658 - I.F. 0,659 (9 cit.)

30 KWIATKOWSKI F., CHEVRIER R., DE RENZIS J.P., CHARRIER S., CURE H., BARGNOUX P.J., LEGER-ENREILLE A., CHOLLET P. Chrono-pharmacocinétique du 5-fluoro-uracile dans le traitement des cancers colorectaux métastatiques et indicateurs de réponse métabolique., *J Pharm Clin ; (*1999). **18 :** 138-143

31 KWIATKOWSKI F., GACHON F., CHARRIER S., SANCHO-GARNIER H., KRAMAR A. Stratégie de décision en biologie : l'analyse ROC, *Bull Cancer ; (*1999). **86 :** 787-789 - (2 cit.)

32 ROLHION C., PENAULT-LLORCA F., KEMENY J.L., KWIATKOWSKI F., LEMAIRE J.J., CHOLLET P., FINAT-DUCLOS F., VERRELLE P. O6-methylguanine-DNA methyltransferase gene (MGMT) expression in human glioblastomas in relation to patient charactheristics and p53 accumulation., *Int J Cancer ; (*1999). **84 :** 416-420 - I.F. 4,700 (44 cit.)

33 SLIM K., BOUSQUET J., KWIATKOWSKI F., LESCURE G., PEZET D., CHIPPONI J. Effectg of C02 gas warming on pain after laparoscopic surgery. A randomized double-blind controlled trial., *Surg Endosc ; (*1999). **13 :** 1110-1114 - (31 cit.)

34 UHRHAMMER N., BAY J.O., GOSSE-BRUN S., KWIATKOWSKI F., RIO P., DAVER A., BIGNON Y.J. Allelic imbalance at NBS1 is frequent in both proximal and distal colorectal carcinoma., *Oncol Rep ; (*2000). **7(2) :** 427-431 - I.F. 0,659 (9 cit.)

35 CHARRIER S., CHOLLET P., BAY J.O., CURE H., KWIATKOWSKI F., PORTEFAIX G., COMMUNAL Y., BETAIL G., PLAGNE R., CHASSAGNE J. Hematological recovery and peripheral blood progenitor cell mobilization after induction chemotherapy and GM-CSF plus G-CSF in breast cancer., *Bone Marrow Transplant ; (*2000). **25 :** 705-710 - I.F. 2,277 (1 cit.)

36 DIONET C., TCHIRKOV A., ALARD J.P., ARNOLD J., DHERMAIN J., RAPP M., BODEZ V., TAMAIN J.C., MONBEL I., MALET P., KWIATKOWSKI F., DONNARIEIX D., VEYRE A., VERRELLE P. Effects of low dose neutrons applied at reduced dose rate on human melanoma cells., *Radiat Res ; (*2000). **154 :** 406-411 - I.F. 2,807 (16 cit.)

37 KWIATKOWSKI F., GIRARD M., HACENE K., BERLIE J. [Sem: a suitable statistical software adaptated for research in oncology], *Bull Cancer ; (*2000). **87 :** 715-721 - (106 cit.)

38 PENAULT-LLORCA F., LEVREL O., CLEMENSON A., KWIATKOWSKI F., POMEL CH., FOUILHOUX G., DE LATOUR M., CURE H., DECHELOTTE P., FONCK Y., DAUPLAT J. Evaluation de différents systèmes de grade histopronostique des cancers ovariens primitifs. Série de cent patientes traitées de façon homogène pour un adénocarcinome ovarien primitif (Travail original présenté lors de la Société Française d'oncologie gynécologique des 5 et 6 novembre 1999)., *J Gynecol Obstet Biol Reprod ; (*2000). **29 :** 548-554 - (1 cit.)

39 SLIM K., BOUSQUET J., KWIATKOWSKI F., LESCURE G., PEZET D., CHIPPONI J. Quality of life before and after laparoscopic fundoplication (presented in part at the Seventh International Congress of the European Association for Endoscopic Surgery, Linz (Austria), June 1999)., *Laparoscopy ; (*2000). **180 :** 41-45 - (49 cit.)

40 CHARRIER S., PORTEFAIX G., CHOLLET P., KWIATKOWSKI F., COMMUNAL Y., CURE H., BAY J.O., BETAIL G., PLAGNE R., CHASSAGNE J. Red blood cells (RBC) and high fluorescence reticulocytes (HFR) production increased by induction chemotherapy and GM-CSF plus G-CSF in peripheral blood of breast cancer patients., *Hematol Cell Ther ; (*2000). **42 :** 165-170 - I.F. 0,907

41 GERAADS A., PETIT F.X., KWIATKOWSKI F., POUR LE COLLÈGE DES PNEUMOLOGUES DES HÔPITAUX

GÉNÉRAUX. Le syndrome d'apnées du sommeil en pratique clinique : l'étude SASOM du CPHG., *Rev Mal Respir ; (*2001). **18 :** 49-56 - I.F. 0,451

42 MARCHENAY C., CELLARIER E., LEVI F., ROLHION C., KWIATKOWSKI F., CLAUSTRAT B., MADELMONT J.C., CHOLLET P. Circadian variation in 06-alkylguanine-DNA alkyltransferase activity in circulating blood mononuclear cells of healthy human subjects., *Int J Cancer ; (*2001). **91 :** 60-66 - I.F. 3,545 (14 cit.)

43 CURE H., CHEVALIER V., ADENIS A., TUBIANA-MATHIEU N., NIEZGODZKI G., KWIATKOWSKI F., PEZET D., PERPOINT B., COUDERT C., FOCAN C., LEVI F., CHIPPONI J., CHOLLET P. Phase II trial of chronomodulated infusion of high-dose 5-fluorouracil and l-folinic acid in previously untreated patients with metastatic colorectal cancer., *J Clin Oncol ; (*2002). **20 :** 1175-1181 - I.F. 9,868 (22 cit.)

44 CAYRE A., CACHIN F., MAUBLANT J., MESTAS D., FEILLEL V., FERRIERE J.P., KWIATKOWSKI F., CHEVILLARD S., FINAT-DUCLOS F., VERRELLE P., PENAULT-LLORCA F. Single static view 99mTc-sestamibi scintimammography predicts response to neoadjuvant chemotherapy and is related to MDR expression., *Int J Oncol ; (*2002). **20 :** 1049-1055 - I.F. 2,931 (25 cit.)

45 CAYRE A., PENAULT-LLORCA F., DE LATOUR M., ROHLION C., FEILLEL V., FERRIERE J.P., KWIATKOWSKI F., FINAT-DUCLOS F., VERRELLE P. O6-Methylguanine-DNA methyl transferase gene expression and prognosis in breast carcinoma., *Int J Oncol ; (*2002). **21 :** 1125-1131 - I.F. 2,931 (16 cit.)

46 CURE H., AMAT S., PENAULT-LLORCA F., LE BOUEDEC G., FERRIERE J.P., MOURET-REYNIER M.A., KWIATKOWSKI F., FEILLEL V., DAUPLAT J., CHOLLET P. Prognostic value of residual node involvement in operable breast cancer after induction chemotherapy., *Breast Cancer Res Treat ; (*2002). **76(1) :** 37-45 - I.F. 3,132 (24 cit.)

47 AMAT S., BOUGNOUX P., PENAULT-LLORCA F., FETISSOF F., CURE H., KWIATKOWSKI F., ACHARD J.L., BODY G., DAUPLAT J., CHOLLET P. Neoadjuvant docetaxel for operable breast cancer induces a high pathological response and breast-conservation rate., *Br J Cancer ; (*2003). **88 :** 1339-1345 - I.F. 3,639 (62 cit.)

48 DURANDO X., LEMAIRE J.J., TORTOCHAUX J., VAN PRAAGH I., KWIATKOWSKI F., VINCENT C., BAILLY C., VERRELLE P., IRTHUM B., CHAZAL J., BAY J.O. High-dose BCNU followed by autologous hematopoietic stem cell transplantation in supratentorial high-grade malignant gliomas : a retrospective analysis of 114 patients., *Bone Marrow Transplant ; (*2003). **31 :** 559-564 - I.F. 2,378 (15 cit.)

49 TCHIRKOV A., ROLHION C., KEMENY J.L., IRTHUM B., PUGET S., KHALIL T., CHINOT O., KWIATKOWSKI F., PERISSEL B., VAGO P., VERRELLE P. Clinical implications of quantitative real-time RT-PCR analysis of hTERT gene expression in human gliomas., *Br J Cancer ; (*2003). **88 :** 516-520 - I.F. 3,639 (39 cit.)

50 KWIATKOWSKI F., LEVI F. Vers le développement des applications médicales de la chronobiologie., *Pathol Biol ; (*2003). **51 :** 185-190 - I.F. 0,953 (3 cit.)

51 CHEVALIER V., GACHON F., KWIATKOWSKI F., PAPON J., CURE H., DOLY M., MADELMONT J.C., CHOLLET P. Existe-t-il un rythme circadien de l'interleukine 15 chez l'homme ?, *Pathol Biol ; (*2003). **51 :** 194-196 - I.F. 0,751 (1 cit.)

52 KWIATKOWSKI F., CHEVALIER V., CHEVRIER R., RICHARD D., CURE H., CHOLLET P. Modélisation de la clairance du 5-FU pendant une perfusion chronomodulée., *Pathol Biol ; (*2003). **51 :** 231-233 - I.F. 0,751

53 SLIM K., NINI E., FORESTIER D., KWIATKOWSKI F., PANIS Y., CHIPPONI J. Methodological index for non-randomized studies (minors): development and validation of a new instrument., *ANZ J Surg ; (*2003). **73 :** 712-716 - I.F. 0,599 (176 cit.)

54 MANTION G., DELROEUX D., DENUÉ P.O., KWIATKOWSKI F., SLIM K., EUVRARD P., MATHIEU P. Réhospitalisation après chirurgie colo-rectale : définitions, facteurs de risque et incidence, *Monographie de l'Association Française de Chirurgie lors du 105ème Congrès Français de Chirurgie ; (*2003). **1 :** 57-70 .

55 ALVES A., PANIS G., MANTION G., SLIM K., KWIATKOWSKI F., HEYD B. Mortalité et morbidité en chirurgie colorectale, *Monographie de l'Association Française de Chirurgie lors du 105ème Congrès Français de Chirurgie ; (*2003). **1 :** 3-28 .

56 SLIM K., KWIATKOWSKI F., BRUGERE C., DA COSTA V., MEZOUGHI S., PANIS Y. MANTION G, CHIPPONI J. Facteurs et scores prédictifs de la mortalité et la morbidité après chirurgie colorectale., *Monographie de l'Association Française de Chirurgie lors du 105ème Congrès Français de Chirurgie ; (*2003). **1 :** 29-41 .

57 ALVES A., SLIM K., KWIATKOWSKI F., PANIS Y. Mortalité et morbidité après exérèse du cancer du rectum sous-péritonéal., *Monographie de l'Association Française de Chirurgie lors du 105ème Congrès Français de Chirurgie ; (*2003). **1 :** 83-91 - (63 cit.)

58 MANTION G., MATHIEU P., SLIM K., PASSEBOIS L., WEN H., KWIATKOWSKI F., LANDECY G. Mortalité et morbidité en chirurgie colo-rectale d'urgence., *Monographie de l'Association Française de Chirurgie lors du 105ème Congrès Français de Chirurgie ; (*2003). **1 :** 93-107 .

59 GLEHEN O., OSINSKY D., COTTE E., KWIATKOWSKI F., FREYER G., ISAAC S., TRILLET-LENOIR V., SAYAG-BEAUJARD A.C., FRANCOIS Y., VIGNAL J., GILLY F.N. Intraperitoneal chemohyperthermia using a closed abdominal procedure and cytoreductive surgery for the treatment of peritoneal carcinomatosis : morbidity and mortality analysis of 216 consecutive procedures., *Ann Surg Oncol ; (*2003). **10 :** 863-869 - I.F. 3,308 (148 cit.)

60 SLIM K., NINI E., FORESTIER D., BRUGERE C., KWIATKOWSKI F., PANIS Y., CHIPPONI J. Validation d'un index méthodologique (MINORS) pour les études non-randomisées., *Ann Chir ; (*2003). **128 :** 688-693 - I.F. 0,578 (1 cit.)

61 MERLE P., JANICOT H., FILAIRE M., ROUX D., BAILLY C., VINCENT C., GACHON F., TCHIRKOV A., KWIATKOWSKI F., NAAME A., ESCANDE G., CAILLAUD D., VERRELLE P. Early CYFRA 21-1 variation predicts tumor response to chemotherapy and survival in locally advanced non-small cell lung cancer patients., *Int J Biol Markers ; (*2004). **19 :** 310-315 - I.F. 1,092 (10 cit.)

62 MINET-QUINARD R., VAN PRAAGH I., KWIATKOWSKI F., BEAUJON G., FEILLEL V., BEAUFRERE B., BARGNOUX P.J., CYNOBER L., VASSON M.P. Pre- and postoperative aminoacidemia in breast cancer: a study vs matched healthy subjects, *Cancer Invest ; (*2004). **22 :** 203-210 - (5 cit.)

63 GLEHEN O., KWIATKOWSKI F., SUGARBAKER P.H., ELIAS D., LEVINE E.A., DE SIMONE M., BARONE R., YONEMURA Y., CAVALIERE F., QUENET F., GUTMAN M., TENTES A.A.K., LORIMIER G., BERNARD J.L., BEREDER J.M., PORCHERON J., GOMEZ-PORTILLA A., SHEN P., DERACO M., RAT P. Cytoreductive surgery combined with perioperative intraperitoneal chemotherapy for the management of peritoneal carcinomatosis from colorectal cancer : a multi-institutional study., *J Clin Oncol ; (*2004). **22 :** 3284-3292 - I.F. 10,864 (409 cit.)

64 MOURET-REYNIER M.A., ABRIAL C., FERRIERE J.P., AMAT S., CURE H., KWIATKOWSKI F., FEILLEL V., LE BOUEDEC G., PENAULT-LLORCA F., CHOLLET P. Neoadjuvant FEC 100 for operable breast cancer : eigh-year experience at Centre Jean Perrin., *Clin Breast Cancer ; (*2004). **5 :** 303-307 - (8 cit.)

65 TCHIRKOV A., CHALETEIX C., MAGNAC C., VASCONCELOS Y., DAVI F., MICHEL A., KWIATKOWSKI F., TOURNILHAC O., DIGHIERO G., TRAVADE P. hTERT expression and prognosis in B-chronic lymphocytic leukemia., *Ann Oncol ; (*2004). **15 :** 1476-1480 - I.F. 3,605 (30 cit.)

66 ALVES A., PANIS Y., MATHIEU P., MANTION G., KWIATKOWSKI F., SLIM K. FOR THE ASSOCIATION FRANÇAISE DE CHIRURGIE. Postoperative mortality and morbidity in French patients undergoing colorectal surgery., *Arch Surg ; (*2005). **140 :** 278-283 - I.F. 3,076 (139 cit.)

67 DEJAX C., VENNAT J.C., DE FREITAS D., KWIATKOWSKI F., LEROUX M.A., AUBERT B. Traitement de l'hyperthyroïdie du sujet âgé par l'iode 131. A propos de 180 patients. Problèmes de radioprotection et de gestion des déchets liés au traitement par l'iode 131., *Médecine nucléaire. Imagerie fonctionnelle et*

*métabolique ; (*2005). **29 :** 609-619 .

68 ALVES A., PANIS Y., MATHIEU P., KWIATKOWSKI F., SLIM K., MANTION G., THE ASSOCIATION FRANÇAISE DE CHIRURGIE (AFC). Mortality and morbidity after surgery of mid and low rectal cancer. Results of a French prospective multicentric study., *Gastroenterol Clin Biol ; (*2005). **29 :** 509-514 - I.F. 0,816 (28 cit.)

69 ABRIAL C., KWIATKOWSKI F., CHEVRIER R., GACHON F., CURE H., CHOLLET P. Potentiel thérapeutique de la mélatonine dans la prise en charge de la pathologie cancéreuse., *Pathol Biol ; (*2005). **53 :** 265-268 - I.F. 0,543 (5 cit.)

70 KWIATKOWSKI F., ABRIAL C., GACHON F., CHEVRIER R., CURE H., CHOLLET P. Stress, cancer et rythme circadien de la mélatonine., *Pathol Biol ; (*2005). **53 :** 269-272 - I.F. 0,543 (5 cit.)

71 KWIATKOWSKI F., DALLE C., AUVRAY H. Le cancer indépendant du psychisme : un manque d'interprétation créatrice ?, *Rev Francoph Psycho-Oncologie ; (*2005). **4 :** 105-116 .

72 KWIATKOWSKI F., LEVI F. Chronobiologie et immunité (éditorial)., *Pathol Biol ; (*2005). **53 :** 251-254 - I.F. 0,953 (2 cit.)

73 BAY J.O., DHEDIN N., GOERNER M., VANNIER J.P., MARIE-CARDINE A., STAMATOULLAS A., JOUET J.P., YAKOUB-AGHA I., TABRIZI R., FAUCHER C., DIEZ-MARTIN J.L., NUNEZ G., PARODY R., MILPIED N., ESPEROU H., GARBAN F., GALAMBRUN C., KWIATKOWSKI F., DARLAVOIX I., ZINAI A., FISCHER A., MICHALLET M., VERNANT J.P. Inolimomab in steroid-refractory acute graft-versus-host disease following allogeneic hematopoietic stem cell transplantation: retrospective analysis and comparison with other interleukin-2 receptor antibodies., *Transplantation ; (*2005). **80 :** 782-788 - I.F. 3,568 (25 cit.)

74 ALVES A., AVIT-MIOSSEC S., SLIM K., KWIATKOWSKI F., PANIS Y. Traitement chirurgical électif des diverticulites sigmoïdiennes., *Monographie de l'Association Française de Chirurgie lors du 105ème Congrès Français de Chirurgie ; (*2005). **1 :** 71-82 .

75 ALVES A., PANIS Y., SLIM K., HEYD B., KWIATKOWSKI F., MANTION G., ASSOCIATION FRANCAISE DE CHIRURGIE. French multicentre prospective observational study of laparoscopic versus open colectomy for sigmoid diverticular disease., *Br J Surg ; (*2005). **92 :** 1520-1525 - I.F. 3,580 (65 cit.)

76 ABRIAL C., LEHEURTEUR M., CABRESPINE A., MOURET-REYNIER M.A., DURANDO X., FERRIERE J.P., KWIATKOWSKI F., PENAULT-LLORCA F., CURE H., CHOLLET P. Does survival increase in metastatic breast cancer with recently available anticancer drugs?, *Oncol Res ; (*2006). **15 :** 431-439 - I.F. 1,844 (6 cit.)

77 SLIM K., PANIS Y., ALVES A., KWIATKOWSKI F., MATHIEU P., MANTION G. ASSOCIATION FRANÇAISE DE CHIRURGIE. Predicting postoperative mortality in patients undergoing colorectal surgery., *World J Surg ; (*2006). **30 :** 100-106 - I.F. 1,601 (35 cit.)

78 CABRESPINE A., GUY L., KHENIFAR E., CURE H., FLEURY J., PENAULT-LLORCA F., KWIATKOWSKI F., BARTHOMEUF C., CHOLLET P., BAY J.O. Randomized phase II study comparing paclitaxel and carboplatin versus mitoxantrone in patients with hormone-refractory prostate cancer., *Urology ; (*2006). **67 :** 354-359 - I.F. 2,139 (9 cit.)

79 LE BOUEDEC G., GEISSLER B., GIMBERGUES P., CACHIN F., PENAULT-LLORCA F., KWIATKOWSKI F., DAUPLAT J., MAUBLANT J. Lymphadénectomie axillaire sentinelle après chimiothérapie néoadjuvante pour cancer du sein : influence du statut ganglionnaire initial., *Bull Cancer ; (*2006). **93 :** 415-419 - I.F. 0,753 (2 cit.)

80 LE PAGE S., KWIATKOWSKI F., PAULIN C., MOHAMED F., PEZET D., CHIPPONI J., BENHAMED M., GILLY F.N., GLEHEN O. In vitro thermochemotherapy of colon cancer cell lines with irinotecan alone and combined with mitomycin C., *Hepatogastroenterology ; (*2006). **53 :** 693-697 - I.F. 0,699 (4 cit.)

81 BUC E., KWIATKOWSKI F., ALVES A., PANIS Y., MANTION G., SLIM K. Tobacco smoking : a factor of early onset of colorectal cancer., *Dis Colon Rectum ; (*2006). **49 :** 1-4 - I.F. 2,264 (20 cit.)

82 BAY J.O., RAY-COQUARD I., FAYETTE J., LEYVRAZ S., CHERIX S., PIPERNO-NEUMANN S., CHEVREAU C., ISAMBERT N., BRAIN E., EMILE E., LE CESNE A., CIOFFI A., KWIATKOWSKI F., COINDRE J.M., BINH BUI N., PEYRADE F., BLAY J.Y., FOR THE GROUPE SARCOME FRANÇAIS. Docetaxel and gemcitabine combination in 133 advanced soft-tissue sarcomas : a retrospective analysis., *Int J Cancer ; (*2006). **119 :** 706-711 - I.F. 4,700 (73 cit.)

83 GIMBERGUES P., DAUPLAT M.M., CAYRE A., DURANDO X., LE BOUEDEC G., FINAT-DUCLOS F., PORTEFAIX G., KWIATKOWSKI F., DAUPLAT J., PENAULT-LLORCA F., TCHIRKOV A. Correlation between molecular metastases in sentinel lymph nodes of breast cancer patients and St Gallen risk category., *EJSO. Eur J Surg Oncol ; (*2007). **33 :** 16-22 - I.F. 1,873 (12 cit.)

84 DELORT L., KWIATKOWSKI F., CHALABI N., SATIH S., BIGNON Y.J., BERNARD-GALLON D.J. Risk factors for early age at breast cancer onset - The "COSA program" population-based study., *Anticancer Res ; (*2007). **27 :** 1087-1094 - I.F. 1,479 (6 cit.)

85 KWIATKOWSKI F., BIGNON Y.J. Tester l'impact de l'hypnothérapie sur l'immunité et les rythmes biologiques chez des patients cancéreux traités en palliatif : une cible prometteuse ?, *Pathol Biol ; (*2007). **55 :** 186-193 - I.F. 0,667 (1 cit.)

86 BAY J.O., LINASSIER C., BIRON P., DURANDO X., VERRELLE P., KWIATKOWSKI F., ROSTI G., DEMIRER T. FOR THE EBMT SOLID TUMORS WORKING PARTY. Does high-dose carmustine increase overall survival in supratentorial high-grade malignant glioma ? An EBMT retrospective study., *Int J Cancer ; (*2007). **120 :** 1782-1786 - I.F. 4,693 (6 cit.)

87 ALVES A., PANIS Y., MANTION G., SLIM K., KWIATKOWSKI F., VICAUT E. The AFC Score : validation of a 4-item predicting score of postoperative mortality after colorectal resection for cancer or diverticulitis : results of a prospective multicenter study in 1049 patients., *Ann Surg ; (*2007). **246 :** 91-96 - I.F. 7,678 (42 cit.)

88 KWIATKOWSKI F., SLIM K., VERRELLE P., CHAMOREY E., KRAMAR A. Le score de propension : intérêt et limites., *Bull Cancer ; (*2007). **94 :** 680-686 - I.F. 0,906 (7 cit.)

89 GUINIER D., MANTION G.A., ALVES A., KWIATKOWSKI F., SLIM K., PANIS Y. Risk factors of unplanned readmission after colorectal surgery : a prospective, multicenter study., *Dis Colon Rectum ; (*2007). **50 :** 1316-1323 - I.F. 2,442 (25 cit.)

90 LE PAGE S., CAPUTO S., KWIATKOWSKI F., BERARD P., GOUILLAT C. Séquelles fonctionnelles et qualité de vie après duodénopancréatectomie céphalique, *J Chir ; (*2008). **145 :** 32-36 - I.F. 0,509 (5 cit.)

91 CHOLLET P., ABRIAL C., DURANDO X., THIVAT E., TACCA O., MOURET-REYNIER M.A., LEHEURTEUR M., KWIATKOWSKI F., DAUPLAT J., PENAULT-LLORCA F. A new prognostic classification after primary chemotherapy for breast cancer: residual disease in breast and nodes (RDBN), *Cancer J ; (*2008). **14 :** 128-132 - I.F. 2,769 (12 cit.)

92 GIMBERGUES P., ABRIAL C., DURANDO X., LE BOUEDEC G., CACHIN F., PENAULT-LLORCA F., MOURET-REYNIER M.A., KWIATKOWSKI F., MAUBLANT J., TCHIRKOV A., DAUPLAT J. Sentinel lymph node biopsy after neoadjuvant chemotherapy is accurate in breast cancer patients with a clinically negative axillary nodal status at presentation, *Ann Surg Oncol ; (*2008). **15 :** 1316-1321 - I.F. 3,898 (38 cit.)

93 FONTANA L., BOSVIEL R., DELORT L., GUY L., CHALABI N., KWIATKOWSKI F., SATIH S., RABIAU N., BOITEUX J.P., CHAMOUX A., BIGNON Y.J., BERNARD-GALLON D.J. DNA repair gene ERCC2, XPC, XRCC1, XRCC3 polymorphisms and associations with bladder cancer risk in a French cohort, *Anticancer Res ; (*2008). **28(3B) :** 1853-1856 - I.F. 1,390 (37 cit.)

94 BERNARD-GALLON D., BOSVIEL R., DELORT L., FONTANA L., CHAMOUX A., RABIAU N., KWIATKOWSKI F., CHALABI N., SATIH S., BIGNON Y.J. DNA repair gene ERCC2 polymorphisms and associations with breast and ovarian cancer risk, *Mol Cancer ; (*2008). **7 :** 36-0 - (20 cit.)

95 VERONESE L., TOURNILHAC O., VERRELLE P., DAVI F., DIGHIERO G., CHAUTARD E., VEYRAT-MASSON R., KWIATKOWSKI F., GOUMY C., VAGO P., TRAVADE P., TCHIRKOV A. Low MCL-1 mRNA expression correlates with prolonged survival in B-cell chronic lymphocytic leukemia (letter to the editor), *Leukemia ; (*2008). **22 :** 1291-1293 - I.F. 8,634 (11 cit.)

96 DELORT L., CHALABI N., SATIH S., RABIAU N., KWIATKOWSKI F., BIGNON Y.J., BERNARD-GALLON D.J. Association between genetic polymorphisms and ovarian cancer risk, *Anticancer Res ; (*2008). **28((5B)) :** 3079-3081 - I.F. 1,390 (18 cit.)

97 CHALABI N., BERNARD-GALLON D.J., BIGNON Y.J., BREAST MED CONSORTIUM, KWIATKOWSKI F., AGIER M., VIDAL V., LAPLACE-CHABAUD V., SYLVAIN-VIDAL V., BERTHOLET V., DE LONGUEVILLE F., LACROIX M., LECLERCQ G., REMACLE J., SIBILLE C., ZAMMATEO N., BEN JAAFAR N., SEFIANI A., OULDIM K., MEGARBANE K., JALKH N., MAHFOUDH W., TROUDI W., BEN AMMAR-EL GAIED A., CHOUCHANE L. Comparative clinical and transcriptomal profiles of breast cancer between French and South Mediterranean patients show minor but significative biological differences, *Cancer Genomics Proteomics ; (*2008). **5 :** 253-261 - (15 cit.)

98 DELORT L., KWIATKOWSKI F., CHALABI N., SATIH S., BIGNON Y.J., BERNARD-GALLON D.J. Central adiposity as a major risk factor of ovarian cancer, *Anticancer Res ; (*2009). **29 :** 5229-5234 - I.F. 1,428 (1 cit.)

99 SANTARPIA A., BLANCHET A., MININNI G., KWIATKOWSKI F., LINDEMAN L., LAMBERT J.F. The "weight" of words on the forearms during relaxation, *Appl Psychophysiol Biofeedback ; (*2009). **34 :** 105-111 - I.F. 1,175 (1 cit.)

100 VUILLAUME M.L., UHRHAMMER N., VIDAL V., VIDAL V.S., CHABAUD V., JESSON B., KWIATKOWSKI F., BIGNON Y.J. Use of gene expression profiles of peripheral blood lymphocytes to distinguish BRCA1 mutation carriers in high risk breast cancer families, *Cancer Inform ; (*2009). **7 :** 41-56 .

101 GIMBERGUES P., ABRIAL C., DURANDO X., LE BOUEDEC G., CACHIN F., PENAULT-LLORCA F., MOURET-REYNIER M.A., KWIATKOWSKI F., MAUBLANT J., TCHIRKOV A., DAUPLAT J. Clinicopathological factors and nomograms predicting nonsentinel lymph node metastases after neoadjuvant chemotherapy in breast cancer patients, *Ann Surg Oncol ; (*2009). **16 :** 1946-1951 - I.F. 4,130 (3 cit.)

102 DUCHER J.L., KWIATKOWSKI F., DUPUY C., MANGEON J.P., FERAL A., SCHMIDT J., LLORCA P.M. Validation de l'échelle d'évaluation du risque suicidaire RSD après tentative de suicide chez 320 patients hospitalisés dans un service d'urgence, *Journal de Thérapie Comportementale et Cognitive ; (*2009). **19 :** 47-52 .

103 VERONESE L., TOURNILHAC O., VERRELLE P., DAVI F., DIGHIERO G., CHAUTARD E., VEYRAT-MASSON R., KWIATKOWSKI F., GOUMY C., GOUAS L., BAY J.O., VAGO P., TCHIRKOV A. Strong correlation between VEGF and MCL-1 mRNA expression levels in B-cell chronic lymphocytic leukemia, *Leukemia Res ; (*2009). **3 :** 1623-1626 - I.F. 2,358 (8 cit.)

104 DELORT L., SATIH S., KWIATKOWSKI F., BIGNON Y.J., BERNARD-GALLON D.J. Evaluation of breast cancer risk in a multigenic model including low penetrance genes involved in xenobiotic and estrogen metabolisms, *Nutr Cancer ; (*2010). **62 :** 243-251 - I.F. 2,553 (15 cit.)

105 SANTARPIA A., BLANCHET A., MININNI G., ANDRASIK F., KWIATKOWSKI F., LAMBERT J.F. Effects of weight-related literal and metaphorical suggestions about the forearms during hypnosis, *Int J Clin Exp Hypn ; (*2010). **58 :** 350-365 - I.F. 1,246 (2 cit.)

106 BAYET-ROBERT M., KWIATKOWSKI F., LEHEURTEUR M., GACHON F., PLANCHAT E., ABRIAL C., MOURET-REYNIER M.A., DURANDO X., BARTHOMEUF C., CHOLLET P. Phase I dose escalation trial of docetaxel plus curcumin in patients with advanced and metastatic breast cancer, *Cancer Biol Ther ; (*2010). **9(1) :** 8-14 - I.F. 2,907 (59 cit.)

107 GILLIOT O., DURANDO X., ABRIAL C., BELLIERE A., GIMBERGUES P., THIVAT E., PLANCHAT E., LAPEYRE

M., KWIATKOWSKI F., TOLEDANO I., CHOLLET P., NABHOLTZ J.M., VERRELLE P. Does regional lymph node irradiation improve the outcome of N0 and pN0 breast cancer ?, *Cancer Invest ; (*2010). **28 :** 195-200 - I.F. 2,390 (4 cit.)

108 ROCHE B., LARROUMETS G., DEJAX C., KWIATKOWSKI F., DESBIEZ F., THIEBLOT P., TAUVERNON I. Epidémiologie, présentation clinique, modalités thérapeutiques et facteurs pronostiques d'une série régionale de 26 cancers anaplasiques de la thyroïde. Comparaison par rapport aux données de la littérature, *Ann Endocrinol ; (*2010). **71 :** 38-45 - I.F. 0,583

109 THIVAT E., THERONDEL S., LAPIROT O., ABRIAL C., GIMBERGUES P., GADEA E., PLANCHAT E., KWIATKOWSKI F., MOURET-REYNIER M.A - , CHOLLET P., DURANDO X. Weight change during chemotherapy changes the prognosis in non metastatic breast cancer for the worse, *BMC Cancer ; (*2010). **10 :** 648 - I.F. 3,153 (59 cit.)

110 FUTIER E., CONSTANTIN J.M., PETIT A., JUNG B., KWIATKOWSKI F., DUCLOS M., JABER S., BAZIN J.E. Positive end-expiratory pressure improves end-expiratory lung volume but not oxygenation after induction of anaesthesia, *Eur J Anaesthesiol ; (*2010). **27 :** 508-513 - I.F. 1,679 (16 cit.)

111 GIMBERGUES P., DAUPLAT M.M., DURANDO X., ABRIAL C., LE BOUEDEC G., MOURET-REYNIER M.A., CACHIN F., KWIATKOWSKI F., TCHIRKOV A., DAUPLAT J., PENAULT-LLORCA F. Intraoperative imprint cytology examination of sentinel lymph nodes after neoadjuvant chemotherapy in breast cancer patients, *Ann Surg Oncol ; (*2010). **17 :** 2132-2137 - I.F. 4,182 (3 cit.)

112 POIRIER A.L., COMMER J.M., KWIATKOWSKI F., MERCIER M., BONNETAIN F. Qualité de vie et transfusion en cancérologie : revue de la littérature, *Transfus Clin Biol ; (*2010). **17 :** 357-361 - I.F. 0,782 (2 cit.)

113 FUTIER E., CONSTANTIN J.M., PETIT A., CHANQUES G., KWIATKOWSKI F., FLAMEIN R., SLIM K., SAPIN V., JABER S., BAZIN J.E. Conservative vs restrictive individualized goal-directed fluid replacement strategy in major abdominal surgery, *Arch Surg ; (*2010). **145 :** 1193-1200 - I.F. 4,500 (71 cit.)

114 BIGNON Y.J., LEGER-ENREILLE A., BERAUD J.F., ACHARD J.L., BEZY O., BRIDON F., CARDINAUD S., CHAPIER R., DUCLOS M., HAHN T., KWIATKOWSKI F., JOUVENCY S., MOURET M.A., PAUL E., SOBKOWICZ M., VAN PRAAGH I., VASSON M.P., VIGIER M., TRAVADE A. "PACThe" : programme of Accompanying women after breast Cancer treatment completion in Thermal resorts : intermediate results on 122 patients at 6 months follow-up and 83 patients at 1 year, *Press Therm Climat ; (*2010). **147 :** 47-49 .

115 DUMAY-LEVESQUE T., LEMERY S., DAUPLAT M.M., BOUSSION V., DIEU V., BAILLY A., KWIATKOWSKI F., BOYER L. Evaluation des macrobiopsies mammaires stéréotaxiques par système Vacora® 10-gauge (541 procédures), *J Radiol ; (*2011). **92 :** 226-235 - I.F. 0,567

116 PIRLET I.A., SLIM K., KWIATKOWSKI F., MICHOT F., MILLAT B.L. Emergency preoperative stenting versus surgery for acute left-sided malignant colonic obstruction: a multicenter randomized controlled trial, *Surg Endosc ; (*2011). **25 :** 1814-1821 - I.F. 4,013 (140 cit.)

117 BOSVIEL R., MICHARD E., LAVEDIAUX G., KWIATKOWSKI F., BIGNON Y.J., BERNARD-GALLON D.J. Peripheral blood DNA methylation detected in the BRCA1 or BRCA2 promoter for sporadic ovarian cancer patients and controls, *Clin Chim Acta ; (*2011). **412 :** 1472-1475 - I.F. 2,535 (13 cit.)

118 PLANCHAT E., ABRIAL C., THIVAT E., MOURET-REYNIER M.A., KWIATKOWSKI F., POMEL C., WANG-LOPEZ Q., CHOLLET P., NABHOLTZ J.M., DURANDO X. Late lines of treatment benefit survival in metastatic breast cancer in current practice ?, *Breast ; (*2011). **20 :** 574-578 - I.F. 2,491 (13 cit.)

119 RABIAU N., DECHELOTTE P., ADJAKLY M., KEMENY J.L., GUY L., BOITEUX J.P., KWIATKOWSKI F., BIGNON Y.J., BERNARD-GALLON D. BRCA1, BRCA2, AR and IGF-I expression in prostate cancer: correlation between RT-qPCR and immunohistochemical detection, *Oncol Rep ; (*2011). **26 :** 695-702 - I.F. 1,835 (20 cit.)

120 RABISCHONG B., LARRAIN D., CANIS M., LE BOUEDEC G., POMEL C., JARDON K., KWIATKOWSKI F., BOURDEL N., ACHARD J.L., DAUPLAT J., MAGE G. Long-term follow-up after laparoscopic management

of endometrial cancer in the obese : a fifteen-year cohort study, *J Minim Invasive Gynecol ; (*2011). **18 :** 589-596 - I.F. 1,556 (18 cit.)

121 FUTIER E., CONSTANTIN J.M., PELOSI P., CHANQUES G., MASSONE A., PETIT A., KWIATKOWSKI F., BAZIN J.E., JABER S. Noninvasive ventilation and alveolar recruitment maneuver improve respiratory function during and after intubation of morbidly obese patients : a randomized controlled study, *Anesthesiology ; (*2011). **114 :** 1354-1363 - I.F. 5,359 (68 cit.)

122 PLANCHAT E., DURANDO X., ABRIAL C., THIVAT E., MOURET-REYNIER M.A., FERRIERE J.P., POMEL C., KWIATKOWSKI F., CHOLLET P., NABHOLTZ J.M. Prognostic value of initial tumor parameters after metastatic relapse, *Cancer Invest ; (*2011). **29 :** 635-643 - I.F. 1,847 (10 cit.)

123 KWIATKOWSKI F., DESSENNE P., LAQUET C., PETIT M.F., BIGNON Y.J. Permanence of the information given during oncogenetic counseling to persons at familial risk of breast/ovarian and/or colon cancer, *Eur J Hum Genet ; (*2012). **20 :** 141-147 - I.F. 4,319 (1 cit.)

124 POIRIER A.L., KWIATKOWSKI F., COMMER J.M., D'AILLIERES B., BERGER V., MERCIER M., BONNETAIN F. Health-related quality of life in cancer patients at the end of life, translation, validation, and longitudinal analysis of specific tools : study protocol for a randomized controlled trial, *Trials ; (*2012). **13 :** 39-0 - I.F. 2,206 (4 cit.)

125 BOSVIEL R., GARCIA S., LAVEDIAUX G., MICHARD E., DRAVERS M., KWIATKOWSKI F., BIGNON Y.J., BERNARD-GALLON D.J. BRCA1 promoter methylation in peripheral blood DNA was identified in sporadic breast cancer and controls, *Cancer Epidemiology ; (*2012). **36 :** 177-182 - I.F. 2,232 (28 cit.)

126 BOSVIEL R., DURIF J., GUO J., MEBREK M., KWIATKOWSKI F ., BIGNON Y.J., BERNARD-GALLON D.J. BRCA2 promoter hypermethylation in sporadic breast cancer (letter to the editor), *OMICS ; (*2012). **16 :** 707-710 - I.F. 2,730 (5 cit.)

127 VERONESE L., TOURNILHAC O., CALLANAN M., PRIE N., KWIATKOWSKI F., COMBES P., CHAUVET M., DAVI F., GOUAS L., VERRELLE P., GUIEZE R., VAGO P., BAY J.O., TCHIRKOV A. Telomeres and chromosomal instability in chronic lymphocytic leukemia (letter to the editor), *Leukemia ; (*2013). **27 :** 490-493 - I.F. 9,379 (10 cit.)

128 THIVAT E., VAN PRAAGH I., BELLIERE A., MOURET-REYNIER M.A., KWIATKOWSKI F., DURANDO X., MAHAMMEDI H., DILLIES A.F., CHOLLET P., CHEVRIER R. Adherence with oral oncologic treatment in cancer patients: interest of an adherence score of all dosing errors, *Oncology; (*2013). **84 :** 67-74 - I.F. 2,613 (10 cit.)

129 JOUVE P., BAZIN J.E., PETIT A., MINVILLE V., GERARD A., BUC E., DUPRE A., KWIATKOWSKI F., CONSTANTIN J.M., FUTIER E. Epidural versus continuous preperitoneal analgesia during fast-track open colorectal surgery : a randomized controlled trial, *Anesthesiology ; (*2013). **118 :** 622-630 - I.F. 6,168 (29 cit.)

130 KWIATKOWSKI F., MOURET-REYNIER M.A., DUCLOS M., LEGER-ENREILLE A., BRIDON F., HAHN T., VAN PRAAGH-DOREAU I., TRAVADE A., GIRONDE M., BEZY O., LECADET J., VASSON M.P., JOUVENCY S., CARDINAUD S., ROQUES C., BIGNON Y.J. Long term improved quality of life by a 2-week group physical and educational intervention shortly after breast cancer chemotherapy completion. Results of the 'Programme of Accompanying women after breast Cancer treatment completion in Thermal resorts' (PACThe) randomised clinical trial of 251 patients, *Eur J Cancer ; (*2013). **49 :** 1530-1538 - I.F. 4,819 (21 cit.)

131 VUILLAUME M.L., KWIATKOWSKI F., UHRHAMMER N., BIDET Y., BIGNON Y.J. Analyse de données d'expression transcriptomiques rythmées par des gènes-horloge : approche méthodologique et optimisation, *Pathol Biol ; (*2013). **61 :** 89-95 - I.F. 1,074

132 WANG-LOPEZ Q., ABRIAL C., PLANCHAT E., MOURET-REYNIER M.A., CURE H., GIMBERGUES P., DUBRAY-LONGERAS P., GADEA E., KWIATKOWSKI F., PENAULT-LLORCA F., CHOLLET P., DURANDO X. Long-term

significance (15 years) of pathological complete response after dose-dense neoadjuvant chemotherapy in breast cancer (letter to the editor), *Breast J ; (*2013). **19 :** 448-450 - I.F. 1,433

133 KWIATKOWSKI F., FRANCIS L., FOCAN C. Les grandes questions de la chronobiologie médicale : pourquoi, comment, quand ?, *Pathol Biol ; (*2013). **61 :** 175-177 - I.F. 1,074 (3 cit.)

134 FUTIER E., CONSTANTIN J.M., PAUGAM-BURTZ C., PASCAL J., EURIN M., NEUSCHWANDER A., MARRET E., BEAUSSIER M., GUTTON C., LEFRANT J.Y., ALLAOUCHICHE B., VERZILLI D., LEONE M., DE JONG A., BAZIN J.E., PEREIRA B., JABER S. ; IMPROVE STUDY GROUP INCLUDING, KWIATKOWSKI F. A trial of intraoperative low-tidal-volume ventilation in abdominal surgery, *N Engl J Med ; (*2013). **369 :** 428-437 - I.F. 54,420 (218 cit.)

135 BUC E., COUVELARD A., KWIATKOWSKI F., DOKMAK S., RUSZNIEWSKI P., HAMMEL P., BELGHITI J., SAUVANET A. Adenocarcinoma of the pancreas : does prognosis depend on mode of lymph node invasion ?, *Eur J Surg Oncol ; (*2014). **40 :** 1578-1585 - I.F. 3,009 (16 cit.)

136 KLOTZ T., BOUSSION V., KWIATKOWSKI F., DIEU-DE FRAISSINETTE V., BAILLY-GLATRE A., LEMERY S., BOYER L. Shear wave elastography contribution in ultrasound diagnosis management of breast lesions, *Diagn Interv Imaging ; (*2014). **95 :** 813-824 - (21 cit.)

137 NABHOLTZ J.M., ABRIAL C., MOURET-REYNIER M.A., DAUPLAT M.M., WEBER B., GLIGOROV J., FOREST A.M., TREDAN O., VANLEMMENS L., PETIT T., GUIU S., VAN PRAAGH I., JOUANNAUD C., DUBRAY-LONGERAS P., TUBIANA-MATHIEU N., BENMAMMAR K.E., KULLAB S., BAHADOOR M.R.K., RADOSEVIC-ROBIN N., KWIATKOWSKI F., DESRICHARD A., CAYRE A., UHRHAMMER N., CHALABI N., CHOLLET P., PENAULT-LLORCA F. Multicentric neoadjuvant phase II study of panitumumab combined with an anthracycline/taxane-based chemotherapy in operable triple-negative breast cancer : identification of biologically defined signatures predicting treatment impact, *Ann Oncol ; (*2014). **25 :** 1570-1577 - I.F. 7,040 (49 cit.)

138 GADEA E., THIVAT E., MERLIN C., PAULON R., KWIATKOWSKI F., CHADEYRAS J.B., COUDERT B., BOIRIE Y., MORIO B., DURANDO X. Brown adipose tissue activity in relation to weight gain during chemotherapy in breast cancer patients : a pilot study, *Nutr Cancer ; (*2014). **66 :** 1092-1096 - I.F. 2,322 (4 cit.)

139 MOURGUES C., GERBAUD L., LEGER S., AUCLAIR C., PEYROL F., BLANQUET M., KWIATKOWSKI F., LEGER-ENREILLE A., BIGNON Y.J. Positive and cost-effectiveness effect of spa therapy on the resumption of occupational and non-occupational activities in women in breast cancer remission : a French multicentre randomised controlled trial, *Eur J Oncol Nurs ; (*2014). **18 :** 505-511 - I.F. 1,426 (6 cit.)

140 VASSON M.P., TALVAS J., PERCHE O., DILLIES A.F., BACHMANN P., PEZET D., ACHIM A.C., POMMIER P., RACADOT S., WEBER A., RAMDANI M., KWIATKOWSKI F., BOUTELOUP C. Immunonutrition improves functional capacities in head and neck and esophageal cancer patients undergoing radiochemotherapy : a randomized clinical trial, *Clin Nutr ; (*2014). **33 :** 204-210 - I.F. 4,476 (25 cit.)

141 POIRIER A.L., KWIATKOWSKI F., COMMER J.M., SWAINE-VERDIER A. MONTEL S., CHARPY J.P., MAUCOURT F., LE PAPE E., BAIZE N., VILLET S., GAMBLIN V., FAVIER L., BERGER V., MERCIER M., BONNETAIN F. Report on the first stages in the translation of measures of health-related quality of life at the end of life, *J Palliat Care Med ; (*2014). **4(3) :** 178 .

142 DUFOUR R., DAUMAR P., MOUNETOU E., AUBEL C., KWIATKOWSKI F., ABRIAL C., VATOUX C., PENAULT-LLORCA F., BAMDAD M. BCRP and P-gp relay overexpression in triple negative basal-like breast cancer cell line : a prospective role in resistance to Olaparib, *Sci Rep ; (*2015). **5(112670) :** I.F. 5,228 (9 cit.)

143 WANG-LOPEZ Q., MOURET-REYNIER M.A., SAVOYE A.M., ABRIAL C., KWIATKOWSKI F., GARBAR C., DUBRAY-LONGERAS P., EYMARD J.C., LE BOUEDEC G., VAN PRAAGH I., PENAULT-LLORCA F., CHOLLET P., CURE H. Is it important to adapt neoadjuvant chemotherapy to the visible clinical response ? An open randomized phase II study comparing response-guided and standard treatments in HER2-negative operable breast cancer, *Oncologist ; (*2015). **20 :** 243-244 - I.F. 4,789 (2 cit.)

144 OLLIER M., RADOSEVIC-ROBIN N., KWIATKOWSKI F., PONELLE F., VIALA S., PRIVAT M., UHRHAMMER N., BERNARD-GALLON D., PENAULT-LLORCA F., BIGNON Y.J., BIDET Y. DNA repair genes implicated in triple negative familial non-BRCA1/2 breast cancer predisposition, *Am J Cancer Res ; (*2015). **5 :** 2113-2126 - I.F. 3,425 (17 cit.)

145 BASU S., COMBE K., KWIATKOWSKI F., CALDEFIE-CHEZET F., PENAULT-LLORCA F., BIGNON Y.J., VASSON M.P. Cellular expression of cyclooxygenase, aromatase, adipokines, inflammation and cell proliferation markers in breast cancer specimen, *PLoS One ; (*2015). **10((10):e0138443) :** I.F. 3,057 (12 cit.)

146 MOMBELLI S., KWIATKOWSKI F., ABRIAL C., WANG-LOPEZ Q., DE BOISSIEU P., GARBAR C., BENSUSSAN A., CURE H. Prognostic factors in operable breast cancer treated with neoadjuvant chemotherapy : towards a quantification of residual disease, *Oncology ; (*2015). **88 :** 261-272 - I.F. 2,152 (4 cit.)

147 KWIATKOWSKI F., LAQUET C., DESSENNE P., BIGNON Y.J. Informer la famille : émotions et attitudes du consultant en oncogénétique pour risque familial de cancer sein/ovaire ou côlon, *Bull Cancer ; (*2015). **102 :** 162-73 - I.F. 0,706

148 DURANDO X., DALENC F., ABRIAL C., MOURET-REYNIER M.A., HERVIOU P., KWIATKOWSKI F., CHOLLET P., ROCHE H., THIVAT E. Neurotoxicity as a prognostic factor in patients with metastatic breast cancer treated with ixabepilone as a first-line therapy, *Oncology ; (*2015). **88 :** 180-188 - I.F. 2,152

149 MERCIER J., KWIATKOWSKI F., ABRIAL C., BOUSSION V., DIEU-DE FRAISSINETTE V., MARRAOUI W., PETITCOLIN-BIDET V., LEMERY S. The role of tomosynthesis in breast cancer staging in 75 patients, *Diagn Interv Imaging ; (*2015). **96 :** 27-35 - (11 cit.)

150 TALVAS J., GARRAIT G., GONCALVES-MENDES N., ROUANET J., VERGNAUD-GAUDUCHON J., KWIATKOWSKI F., BACHMANN P., BOUTELOUP C., BIENVENU J., VASSON M.P. Immunonutrition stimulates immune functions and antioxidant defense capacities of leukocytes in radiochemotherapy-treated head & neck and esophageal cancer patients : a double-blind randomized clinical trial, *Clin Nutr ; (*2015). **34 :** 810-817 - I.F. 4,487 (6 cit.)

151 POUGET M., ABRIAL C., PLANCHAT E., VAN PRAAGH I., ARBRE M., KWIATKOWSKI F., DUBRAY-LONGERAS P., DEVAUD H., DOHOU J., HERVIOU P., MAHAMMEDI H., DURANDO X., CHOLLET P., MOURET-REYNIER M.A. Everolimus in metastatic breast bancer : clinical experience as a late treatment line, *Oncology ; (*2015). **89 :** 319-331 - I.F. 2,152 (2 cit.)

152 KWIATKOWSKI F., ARBRE M., BIDET Y., LAQUET C., UHRHAMMER N., BIGNON Y.J. BRCA mutations increase fertility in families at hereditary breast/ovarian cancer risk, *PLos One ; (*2015). **10(6:e0127363) :** I.F. 3,057 (13 cit.)

153 TARDY-MEDOUS M., FILAIRE M., PATOIR A., GAUTIER-PIGNONBLANC P., GALVAING G., KWIATKOWSKI F., COSTES F., RUDDY R. Exercise cardiac output limitation in pectus excavatum (letter), *J Am Coll Cardiol ; (*2015). **66 :** 976-977 - I.F. 17,759 (4 cit.)

154 JEANNIN G., MERLE P., JANICOT H., THIBONNIER L., KWIATKOWSKI F., NAAME A., CHADEYRAS J.B., GALVAING G., BELLIERE A., FILAIRE M., VERRELLE P. Combined treatment modalities in Pancoast tumor: results of a monocentric retrospective study, *Chin Clin Oncol ; (*2015). **4(4) :** 39-0 - (2 cit.)

155 FILLERON T., KWIATKOWSKI F. Le score de propension, une alternative crédible à la randomisation ? (revue), *Bull Cancer ; (*2016). **103 :** 113-122 - I.F. 0,853

156 NABHOLTZ J.M., CHALABI N., RADOSEVIC-ROBIN N., DAUPLAT M.M., MOURET-REYNIER M.A., VAN PRAAGH I., SERVENT V., JACQUIN J.P., BENMAMMAR K.E., KULLAB S., BAHADOOR M.R., KWIATKOWSKI F., CAYRE A., ABRIAL C., DURANDO X., BIGNON Y.J., CHOLLET P., PENAULT-LLORCA F. Multicentric neoadjuvant pilot phase II study of cetuximab combined with docetaxel in operable triple negative breast cancer, *Int J Cancer ; (*2016). **138 :** 2274-2280 - I.F. 6,513 (13 cit.)

157 GAY-BELLILE M., ROMERO P., CAYRE A., VERONESE L., PRIVAT M., SINGH S., COMBES P., KWIATKOWSKI F., ABRIAL C., BIGNON Y.J., VAGO P., PENAULT-LLORCA F., TCHIRKOV A. *J Pathol Clin Res ; (*2016). **2 :** 234-

246 - (5 cit.)

158 KWIATKOWSKI F., LAPORTE S., SLIM K. Le score de propension : partie I - présentation de la méthode, *La Lettre du Pharmacologue ; (*2016). **30 (3-4) :** 71-72 .

159 KWIATKOWSKI F., LAPORTE S., SLIM K. Le score de propension : partie II - exemple pour valider ou non l'utilité de la préparation colique, *La Lettre du Pharmacologue ; (*2016). **30 (3-4) :** 73-74 .

160 EL GUERRAB A., BAMDAD M., KWIATKOWSKI F., BIGNON Y.J., PENAULT-LLORCA F., AUBEL C. Anti-EGFR monoclonal antibodies and EGFR tyrosine kinase inhibitors as combination therapy for triple-negative breast cancer, *Oncotarget ; (*2016). **7 :** 73618-73637 - I.F. 5,168 (11 cit.)

161 JACOMET C., ILLES G., KWIATKOWSKI F., VIDAL M., MROZEK N., AUMERAN C., CORBIN V., LESENS O., LAURICHESSE H., BAILLY P. Prevalence of aortic valve dystrophy and insufficiency in a cohort of 255 HIV-positive patients followed-up in a cardiology department between 2012 and 2014, *Int J Cardiol ; (*2016). **220 :** 82-86 - I.F. 6,189

162 GUIEZE R., PAGES M., VERONESE L., COMBES P., LEMAL R., GAY-BELLILE M., CHAUVET M., CALLANAN M., KWIATKOWSKI F., PEREIRA B., VAGO P., BAY J.O., TOURNILHAC O., TCHIRKOV A. Telomere status in chronic lymphocytic leukemia with TP53 disruption, *Oncotarget ; (*2016). **7 :** 56976-56985 - I.F. 5,168 (3 cit.)

163 KWIATKOWSKI F., DESSENNE P., LAQUET C., DAURES J.P., GAY-BELLILE M., BIGNON Y.J. BRACAVENIR - impact of a psychoeducational intervention on expectations and coping in young women (aged 18-30 years) exposed to a high familial breast/ovarian cancer risk: study protocol for a randomized controlled trial (study protocol), *Trials ; (*2016). **17 :** 509-0 - I.F. 1,969

164 GAY-BELLILE M., ROMERO P., CAYRE A., VERONESE L., PRIVAT M., SINGH S., COMBES P., KWIATKOWSKI F., ABRIAL C., BIGNON Y.J., VAGO P., PENAULT-LLORCA F., TCHIRKOV A. ERCC1 and telomere status in breast tumours treated with neoadjuvant chemotherapy and their association with patient prognosis, *J Pathol Clin Res ; (*2016). **2 :** 234-246 .

165 DOHOU J., MOURET-REYNIER M.A., KWIATKOWSKI F., ARBRE M., HERVIOU P., POUGET M., ABRIAL C., PENAULT-LLORCA F. A retrospective study on the onset of menopause after chemotherapy: analysis of data extracted from the Jean Perrin Comprehensive Cancer Center database concerning 345 young breast cancer patients diagnosed between 1994 and 2012, *Oncology ; (*2017). **92 :** 255-263 - (2 cit.)

166 GAY-BELLILE M., VERONESE L., COMBES P., EYMARD-PIERRE E., KWIATKOWSKI F., DAUPLAT M.M., CAYRE A., PRIVAT M., ABRIAL C., BIGNON Y.J., MOURET-REYNIER M.A., VAGO P., PENAULT-LLORCA F., TCHIRKOV A. TERT promoter status and gene copy number gains: effect on TERT expression and association with prognosis in breast cancer, *Oncotarget ; (*2017). **8 :** 77540-77551 - (3 cit.)

167 KWIATKOWSKI F., MOURET-REYNIER M.A., DUCLOS M., BRIDON F., HANH T., VAN PRAAGH-DOREAU I., TRAVADE A., VASSON M.P., JOUVENCY S., ROQUES C., BIGNON Y.J. Long-term improvement of breast cancer survivors' quality of life by a 2-week group physical and educational intervention: 5-year update of the 'PACThe' trial., *Br J Cancer ; (*2017). **116 :** 1389-1393 - I.F. 5,922 (1 cit.)

168 EL GUERRAB A., CAYRE A., KWIATKOWSKI F., PRIVAT M., ROSSIGNOL J., ROSSIGNOL F., PENAULT-LLORCA F., BIGNON Y.J. Quantification of hypoxia-related gene expression as a potential approach for clinical outcome prediction in breast cancer, *PLOS ONE ; (*2017). **12(e0175960.) :** I.F. 2,766 (1 cit.)

169 VALTON E., WAWRZYNIAK I., AMBLARD C., COMBOURIEU B., BAYLE M.L., DESMOLLES F., KWIATKOWSKI F., PENAULT-LLORCA F., BAMDAD M. P-gp expression levels in the erythrocytes of brown trout: a new tool for aquatic sentinel biomarker development, *Biomarkers ; (*2017). **22 :** 566-574 - I.F. 1,976

170 DAUPLAT J., KWIATKOWSKI F., ROUANET P., DELAY E., CLOUGH K., VERHAEGHE J.L., RAOUST I., HOUVENAEGHEL G., LEMASURIER P., THIVAT E., POMEL C. ; STIC-RMI WORKING GROUP ; 46 COLL. INCLUDING, ABRIAL C., DURANDO X., GIMBERGUES P., LE BOUEDEC G. Quality of life after mastectomy with or without immediate breast reconstruction, *Br J Surg ; (*2017). **104 :** 1197-1206 - I.F. 5,433 (10 cit.)

171 BIAU J., MIROIR J., MILLARDET C., SAROUL N., PHAM-DANG N., RACADOT S., HUGUET F., KWIATKOWSKI F., PEREIRA B., BOURHIS J., LAPEYRE M. Présentation de l'étude GORTEC 2017-03 : radiothérapie en conditions stéréotaxiques postopératoire des cancers localisés de l'oropharynx et de la cavité buccale avec marges à risque (PHRC-K-16-164) (mise au point), *Cancer Radiother ; (*2017). **21 :** 527-532 - I.F. 1,128 (1 cit.)

172 MOREAU J., BIAU J., ACHARD J.L., TOLEDANO I., BENHAIM C., KWIATKOWSKI F., LOOS G., LAPEYRE M. Intraprostatic fiducials compared with bony anatomy and skin marks for image-guided radiation therapy of prostate cancer, *Cureus ; (*2017). **9(10) :** e1769 - (2 cit.)

173 LAURENT H., GALVAING G., THIVAT E., COUDEYRE E., AUBRETON S., RICHARD R., KWIATKOWSKI F., COSTES F., FILAIRE M. Effect of an intensive 3-week preoperative home rehabilitation programme in patients with chronic obstructive pulmonary disease eligible for lung cancer surgery: a multicentre randomised controlled trial, *BMJ Open ; (*2017). **7(e017307) :** I.F. 2,413 (1 cit.)

174 BEGUINOT M., DAUPLAT M.M., KWIATKOWSKI F., LE BOUEDEC G., TIXIER L., POMEL C., PENAULT-LLORCA F., RADOSEVIC-ROBIN N. Analysis of tumour-infiltrating lymphocytes reveals two new biologically different subgroups of breast ductal carcinoma in situ, *BMC Cancer ; (*2018). **18 :** 129-0 - I.F. 3,288

175 TCHIRKOV A., GAY-BELLILE M., ROMERO P., CAYRE A., VÉRONÈSE L., PRIVAT M., SINGH S., COMBES P., KWIATKOWSKI F., ABRIAL C., BIGNON Y-J, VAGOA P., PENAULT-LLORCA F. Valeur pronostique d'ERCC1 et du statut des télomères dans le cancer du sein traité par chimiothérapie néoadjuvante, *morphologie ; (*2018). **102 (338):** 150-151 .

176 GINZAC A., THIVAT E., MOURET-REYNIER M.A., DUBRAY-LONGERAS P., VAN PRAAGH I., PASSILDAS J., ABRIAL C., KWIATKOWSKI F., BOIRIE Y., DUCLOS M., MORIO B., GADEA E., DURANDO X. Weight evolution during endocrine therapy for breast cancer in postmenopausal patients : effect of initial fat mass percentage and previous adjuvant treatments, *Clin Breast Cancer ; (*2018). **18(5:e1093-1102) :** I.F. 2,703

177 PASSILDAS J., COLLARD O., SAVOYE A.M., DOHOU J., GINZAC A., THIVAT E., DURANDO X., KWIATKOWSKI F., PENAULT-LLORCA F., ABRIAL C., MOURET-REYNIER M.A. Impact of chemotherapy-induced menopause in women of childbearing age with non-metastatic breast cancer - Preliminary results from the MENOCOR study (AOP), *Clin Breast Cancer ; (*2018). **:** I.F. 2,703

178 KWIATKOWSKI F., SERLET L., BIGNON Y.J. What selection pressure does to mutations favoring cancer ? Highlights of a stimulation approach, *Biomed J Sci & Tech Res ; (*2018). **10(4)**

179 BINGULA R., FILAIRE M., RADOSEVIC-ROBIN N., BERTHON J.Y., BERNALIER-DONADILLE A., VASSON M.P., THIVAT E., KWIATKOWSKI F., FILAIRE E. Characterisation of gut, lung, and upper airways microbiota in patients with non-small cell lung carcinoma: Study protocol for case-control observational trial, *Medicine ; (*2018). **97(e13676) :** I.F. 2,028

180 KWIATKOWSKI F., DUCHER J.L. Evaluation of the risk for suicide with the RSD scale of Ducher (review), *BJSTR. Biomed J Sci & Tech Res ; (*2018). **12(2) :** I.F. 0,548

181 BARRES B., KELLY A., KWIATKOWSKI F., BATISSE-LIGNIER M., FOUILHOUX G., AUBERT B., DUTHEIL F., TAUVERON I., CACHIN F., MAQDASY S. Stimulated Thyroglobulin and Thyroglobulin Reduction Index Predicts Excellent Response in Differentiated Thyroid Cancers, *J Clin Endocrinol Metab ; (*2019). **104(8):** 3462-6472 - I.F. 5,789

182 KWIATKOWSKI F., PERTHUS I., UHRHAMMER N., FRANCANNET C., ARBRE M., BIDET Y., BIGNON J.Y. Association between hereditary predisposition to common cancers and congenital multimalformations, *Congenital Anomalies ; (*2019). 1-10 . I.F. 1,149

183 YAKHNI M., BRIAT A., EL GUERRAB A., FURTADO L., KWIATKOWSKI F., MIOT-NOIRAULT E., CACHIN F., PENAULT-LLORCA F., RADOSEVIC-ROBIN N. Homoharringtonine, an approved anti-leukemia drug,

suppresses triple negative breast cancer growth through a rapid reduction of anti-apoptotic protein abundance, *am j cancer res ; (*2019). **9(5):** 1043-1060 - I.F. 3,425

184 BERNADACH M., LAPEYRE M., DILLIES AF, MIROIR J., MOREAU J., KWIATKOWSKI F., PHAM-DANG N., SAROUL N., DURANDO X., BIAU J. Toxicity of docetaxel, platine, 5-fluorouracil-based induction chemotherapy for locally advanced head and neck cancer: The importance of nutritional status, *cancer radiother ; (*2019). **23(4):** 273-280 - I.F. 1,128

185 KWIATKOWSKI1 F., GAY-BELLILE M., DESSENNE P., LAQUET C. BRACAVENIR: an observational study of expectations and coping in young women with high hereditary risk of breast and ovarian cancer, *Hereditary Cancer in Clinical Practice; (*2019). **17:** 1-9

186 LE BON M., LAPEYRE M., MOREAU J., BELLIERE-CALANDRY A., KWIATKOWSKI F., MARTIN F., BENOIT C. Tolerance of hypofractionated stereotactic radiotherapy for hepatic tumours, *cancer radiother ; (*2019). **23(5):** 385-394 - I.F. 1,128

187 KELLY A., BARRES B., KWIATKOWSKI F., BATISSE-LIGNIER M., AUBERT B., VALLA C., SOMDA F., CACHIN F., TAUVERON I., MAQDASY S. Age, thyroglobulin levels and ATA risk stratification predict 10-year survival rate of differentiated thyroid cancer patients, *plos one ; (* 2019). **14:** I.F. 2,766

188 BOUVET C., BARRES B., KWIATKOWSKI F., BATISSE-LIGNIER M., CHAFAI EL ALAOUI M., KAUFFMANN P., CACHIN F., TAUVERON I., KELLY A., MAQDASY S. Re-treatment With Adjuvant Radioactive Iodine Does Not Improve Recurrence-Free Survival of Patients With Differentiated Thyroid Cancer., *Front Endocrinol (Lausanne) ; (* 2019). **10:** 671 .

189 PENAULT-LLORCA F., KWIATKOWSKI F., ARNAUD A., LEVY C., LEHEURTEUR M., UWER L., DERBEL O., LE ROL A., JACQUIN JP, JOUANNAUD C., QUENEL-TUEUX N., GIRRE V., FOA C., GUARDIOLA E., LORTHOLARY A., CATALA S., GUIU S., VALENT A., BOINON D., LEMONNIER J., DELALOGE S. Decision of adjuvant chemotherapy in intermediate risk luminal breast cancer patients: A prospective multicenter trial assessing the clinical and psychological impact of EndoPredict® (EpClin) use (UCBG 2-14)., *breast ; (* 2019). **49:** 132-140. I.F. 2,951

190 VACHER L., THIVAT E., POIRIER C., MOURET-REYNIER MA, CHOLLET P., DEVAUD H., DUBRAY-LONGERAS P., KWIATKOWSKI F., DURANDO X., VAN PRAAGH-DOREAU I., CHEVRIER R. Improvement in adherence to Capecitabine and Lapatinib by way of a therapeutic education program., *Support Care Cancer ; (* 2019). I.F. 2,698

191 VASSON MP, KWIATKOWSKI F., ROSSARY A., JOUVENCY S., MOURET-REYNIER MA, DUCLOS M., VAN PRAAGH-DOREAU I., TRAVADE A., BIGNON YJ (2020) Effectiveness of a global multidisciplinary supportive and educational intervention in thermal resort on anthropometric, biological status and disease-free survival after breast cancer treatment completion (PACThe), *j oncol* ; (sous presse). I.F. 4,528

8.5 Références bibliographiques

[1] Bourchis D (2015) unité Inserm 934/CNRS UMR 3215/Université Pierre et Marie Curie, Institut Curie, Paris. URL : https://www.inserm.fr/thematiques/genetique-genomique-et-bioinformatique/dossiers-d-information/epigenetique
Accès juillet 2017

[2] Haupt A., Kane T.T. (2004) Guide de démographie du Population Reference Bureau (quatrième édition). Population Reference Bureau, Washington, USA

[3] Dictionnaire Larousse : URL : www.larousse.fr/dictionnaires/francais/polymorphisme/62352
Accès : juillet 2017

[4] King, MC, Marks JH, Mandell JB (2003) Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* 302: 643–6.

[5] Stoppa-Lyonnet D, Jeanpierre M (2004) BRCA1 : de l'identification du gène à l'estimation des risques tumoraux. *Med Sci* (Paris) 20(3): 262–263.

[6] Grice EA, Segre JA (2012) The human microbiom: our second genome. *Annu Rev Hum Genet* 13: 151-70

[7] de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, Plummer M (2012). Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol*. 13(6): 607-15.

[8] Bonnet M, Buc E, Sauvanet P, Darcha C, Dubois D, Pereira B, Déchelotte P, Bonnet R, Pezet D, Darfeuille-Michaud A (2014) Colonization of the human gut by E. coli and colorectal cancer risk. *Clin Cancer Res* 20(4): 859-67

[9] Kwiatkowski F, Arbre M, Perthus I, Bignon Y-J (2014) Comment les prédispositions familiales de cancer ont-elles échappé à la sélection naturelle : de l'utilité des grandes bases de données. 8ème Congrès EPICLIN & 21èmes Journées des statisticiens des CLCC. Bordeaux – 14-16 mai 2014

[10] Andrieux V, Chantel C (2013) Espérance de vie, durée passée à la retraite. *Dossiers Solidarité et Santé* ; N°40 : 1-35

[11] Harboe TL, Eiberg H, Kern P, Ejlertsen B, Nedergaard L, Timmermans-Wielenga V, et al. (2009) A high frequent BRCA1 founder mutation identified in the Greenlandic population. *Fam. Cancer* 8: 413-9.

12 Im KM, Kirchhoff T, Wang X, Green T, Chow CY, Vijai J, et al. (2011) Haplotype structure in Ashkenazi Jewish BRCA1 and BRCA2 mutation carriers. *Hum. Genet*. 130: 685–699

[13] Laraqui A, Uhrhammer N, Lahlou-Amine I, Rhaffouli H, Baghdadi J, Dehayni M, et al. (2013) Mutation screening of the BRCA1 gene in early onset and familial breast/ovarian cancer in Moroccan population. *Int. J. Med. Sci.* 10: 60–67.

[14] Vallin J, Caselli G (1999) Quand l'Angleterre rattrapait la France. *INED - Population et Sociétés* n° 346

[15] Henry L (1978) Démographie historique : l'enquête de l'I.N.E.D. sur la population de la France avant 1800 Journal de la société statistique de Paris, 119(3) : 229-33

[16] IARC, Descriptive Epidemiology Group (2002) GLOBOCAN – cancer incidence, mortality and prevalence worldwide. URL http://www-dep.iarc.fr, accès : juillet 2017

[17] Punt C, Pauwn K, Mohube E, Gilimani B, Rantho L, Leaver R, McDonald S, Chant L, Valente C. (2003) Demographics of South African Households – 1995. *PROVIDE background paper* 3: 1-22

[18] Toulemon L (2004). La fécondité des immigrés: nouvelles données, nouvelle approche. *INED-Population et sociétés*, 400:1-4

[19] Amin R, Farukee R (1980) Fertility and its regulation in Bangladesh. *Bank Staff Working Paper* n°383. The World Bank, Washington, USA

[20] Jones GW, Gubhaju B (2009) Trends in marriage in the low fertility countries of East and Southeast Asia, *Asian Population Studies* 5(3) : 237-65.

[21] Dommaraju P (2009) Instruction des femmes et évolution du mariage en Inde. *Population* 4(64) : 757-75.

[22] Henry L, Houdaille J (1979) Célibat et âge au mariage aux XVIIIe et XIXe siècles en France. II – âge au premier mariage. *Population* 2, 34e année : 403-42

[23] Levy ML, Sardon JP (1982) L'écart d'âge entre époux. *Population et sociétés* 162 : 1-3

[24] Dumas J, Peron Y (1992) Marriage and conjugal life in Canada. Statistics Canada – demographic division. Toronto. ISSN 0827-0392, ISBN 0-660-14418-2

[25] Deparcieux A (1746) Essai sur les probabilités de durée de la vie humaine, Paris. Cité par Henry L, Houdaille J (1978)

[26] Henry L, Houdaille J (1978). Célibat et âge au mariage aux XVIIIe et XIXe siècles en France. I. Célibat définitif. In: *Population*, 33e année, n°1 : 43-84 ; http://www.persee.fr/doc/pop_0032-4663_1978_num_33_1_16693

[27] Delort L, Kwiatkowski F, Chalabi N, Satih S, Bignon YJ, Bernard-Gallon DJ. (2007) Risk factors for early age at breast cancer onset--the "COSA program" population-based study. *Anticancer Res*; 27(2): 1087-94

[28] Matart BY (1994) Approche de la mortalité maternelle au Moyen-Age en Provence. Actes des 6e Journées Anthropologiques. Dossier de documentation archéologique n°17. CNRS Editions, Paris, France

[29] World Health Organization (2014) Trends in maternal mortality: 1990 to 2013. Estimates by WHO, UNICEF, UNFPA, The World Bank and the United Nations Population Division.
ISBN 978 92 4 150722 6.
URL : http://apps.who.int/iris/bitstream/10665/112682/2/9789241507226_eng.pdf?ua=1
accès : juillet 2017

[30] Zinman MJ, Clegg DE, Brown CC, O'Connor J, Selvan SG (1996) Estimates of human fertility and pregnancy loss. *Fertil Steril* 65: 503-9

[31] Buss L, Tolstrup J, Munk C, Bergholt T, Ottesen B, Gronbaek M, Kjaer SK (2006) Spontaneous abortion: a prospective cohort study of younger women from the general population in Denmark. Validation, occurrence and risk determinants. *Acta Obstet Gynecol Scand* 85(4): 467-75

[32] Katz VL. (2007) Spontaneous and recurrent abortion: etiology, diagnosis, treatment. In Katz VL, Lentz GM, Lobo RA, Gershenson DM. Eds Comprehensive Gynecology. 5th ed. Philadelphia, PA: Mosby Elsevier

[33] Kwiatkowski F [1,2], Arbre M[1], Bidet Y[3], Laquet C[1], Uhrhammer N[1], Bignon Y-J[1,3] (2015) BRCA Mutations Increase Fertility in Families at Hereditary Breast/Ovarian Cancer Risk. *PLOS-One*. DOI:10.1371/journal.pone.0127363. [1] Centre Jean Perrin, Laboratoire d'Oncologie Moléculaire ; [2] Université Blaise Pascal—Laboratoire de Mathématiques, UMR 6620—CNRS, Campus des Cézeaux ; [3] Université Clermont Auvergne, Université d'Auvergne

[34] Léridon H (1992) L'âge de la ménopause. *Gyn Obst* n° spécial mars, 4.

[35] Treloar AE (1981) Menstrual cyclicity and the pre-menaupause. *Maturitas* 3 : 249-64

[36] Daan NM, Fauser BC (2015) Menopause prediction and potential implications. *Maturitas* 82(3): 257-65

[37] Payeur F. (2008) Âge et fertilité masculine : une analyse biodémographique. Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de Maîtrise ès sciences (M.Sc.) en démographie. Faculté des arts et des sciences, Université de Montréal, Canada

[38] Auger J, Kunstmann JM, Czyglik F, Jouannet P. (1995) Decline in semen quality among fertile men in Paris during the past 20 years. *NEJM* 332(5): 281-5

[39] Wagner L (2004) Fertilité de l'homme vieillissant. *Prog Urol* 14(4): 577-82

[40] Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL et al. (2003) Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet*. 72(5) : 1117-30

[41] Duclos M (2017) Cancer: physical activity matters (prevention, treatment and survival). Fourth international congress of translational research in human nutrition – Clermont-Ferrand. June 22-23 2017.

[42] Franberg M, Gertow K, Hamsten A, PROCARDIS consortium, Lagergren J, Sennblad B (2015)Discovering Genetic Interactions in Large-Scale Association Studies by Stage-wise Likelihood Ratio Tests. *PLOS Genetics*, DOI:10.1371/journal.pgen.1005502

[43] Tomasetti C, Vogelstein B (2015) , Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* 347(6217) : 78-81.

[44] Cancer Research UK, 2016, URL http://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/age#heading-Zero
accès : juillet 2017

[45] Box GE, Muller ME (1958), A Note on the Generation of Random Normal Deviates, *The Annals of Mathematical Statistics* 29(2) : 610–11

[46] The World Bank – Living Standards Measurement Study. South Africa (2013)
URL http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTLSMS/0,,contentMDK:21371306~pagePK:64168445~piPK:64168309~theSitePK:3358997,00.html
accès : juillet 2017

[47] Kwiatkowski F, Girard M, Hacene K, Berlie J. (2000) Sem : un outil de gestion informatique et statistique adapté à la recherche en cancérologie. *Bull Cancer* 87(10), 715-21

[48] Kwiatkowski F [1,2], Serlet L[2] and Bignon Y-J[1] (2018) What Selection Pressure Does to Mutations Favoring Cancer? Highlights of A Simulation Approach. *Biomedical Journal of Scientific & Technical Research*. DOI: 10.26717/BJSTR.2018.10.001989. [1]Laboratoire d'Oncologie Moléculaire, [2]Laboratoire de Mathématique: probabilités et statistiques appliquées

[49] Kaplan EL, Meier P (1958) Non parametric estimation from incomplete observations. *Journal of the American Statistical Association*; 53: 457–81

[50] Autier P, Boniol M, Koechlin A, Pizot C, Boniol M. (2017) Effectiveness of and overdiagnosis from mammography screening in the Netherlands: population based study. *BMJ*; 359: j5224. doi: 10.1136/bmj.j5224.

[51] Gordon A (1999) Classification, 2ème édition. Londres : Chapman and Hall-CRC

[52] Milligan GW, Cooper MC (1985) An examination for determining the number of clusters in a data set. *Psychometrika*; 50(2): 159-79

[53] Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a dataset via the gap *statistic. J. R. Statist. Soc. B*; 63, part 3:411-23

[54] Milligan GW, Sokol LM (1980) A two-stage Clustering algorithm with robust recovery characteristics *Educational and Psychological measurement*; 40: 755-9

[55] Davies DL, Bouldin DW. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227. doi:10.1109/TPAMI.1979.4766909.

[56] Dunn JC (1974) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*; 4: 95-104

[57] Maulik U, Bandyopadhyay S (2000) Genetic algorithm based clustering technique. *Pattern recognition*; 33: 1455-65

[58] Bayati H, Davoudi H, Fatemizadeh E (2008) A heuristic method for finding the optimal number of clusters with application in medical data. 30th annual international IEEE EMBS conference. Vancouver, Canada, August 20-24.

[59] Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*; 20: 53-65.

[60] Domingues-Silva B, Silva B, Azzalin CM (2019) Alternative Functions for Human FANCM at Telomeres. *Front Mol Biosci*. 6: 84. doi: 10.3389/fmolb.2019.00084.

[61] Stoppa-Lyonnet D, Buecher B, Gauthier-Villars M, Houdayer C, de Pauw A, de la Rochefordière A, This P, Asselain B, AndrieuN (2008) Comment prendre en compte le risque génétique ? Gènes impliqués et risques tumoraux associés. 30èmes journées de la SFSPM, La Baule, France

[62] Ward JH. (1963), "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association*, 58, 236–244.

[63] James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer Science+Business Media, New-York, USA.

[64] Eisinger F, Bressac B, Castaigne D, Cottu PH, Lansac J, Lefranc JP, et al. (2004) Identification and management of hereditary predisposition to cancer of the breast and the ovary (update 2004). *Bull. Cancer* (Paris); 91: 219-37. PMID : 15171047

[65] Oh M, McBride A, Yun S, Bhattacharjee S, Slack M, Martin JR, Jeter J, Abraham I (2018) BRCA1 and BRCA2 Gene Mutations and Colorectal Cancer Risk: Systematic Review and Meta-analysis. *J Natl Cancer Inst.*; 110(11): 1178-89. doi: 10.1093/jnci/djy148.

[66] Frånberg M, Gertow K, Hamsten A; PROCARDIS consortium, Lagergren J, Sennblad B. (2015) Discovering Genetic Interactions in Large-Scale Association Studies by Stage-wise Likelihood Ratio Tests. PLoS Genet; 11(9): e1005502. doi: 10.1371/journal.pgen.1005502

[67] Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Human Hereditary; 21: 523-42

[68] Bonaïti B, Alarcon F, Andrieu N, Bonadona V, Dondon MG, Pennec S, Stoppa-Lyonnet D, Bonaïti-Pellié C, Perdry H (2014) A new scoring system in cancer genetics: application to criteria for BRCA1 and BRCA2 mutation screening. *J Med Genet*; 51(2): 114-21. doi: 10.1136/jmedgenet)-2013-101674.

[69] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21(6); 1087

[70] The Global BMI Mortality Collaboration (2016) Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *Lancet* 388(10046): 776-86

[71] Pavard S, Metcalf CJ (2007) Negative selection on BRCA1 susceptibility alleles sheds light on the population genetics of late-onset diseases and aging theory. *PLoS One* 2(11): e1206.

[72] Ruijs MW1, Verhoef S, Wigbout G, Pruntel R, Floore AN, de Jong D, van T Veer LJ, Menko FH.(2006) Late-onset common cancers in a kindred with an Arg213Gln TP53 germline mutation. *Fam Cancer* 5(2): 169-74.

[73] Wishart D (1969) An algorithm for hierarchical classification. *Biometrics*; 25 : 165-170

[74] Murtagh F, Legendre P (2014) Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. *Journal of Classification*; 31(3): 274–295