Approche globale intégrative pour l'identification de nouvelles cibles moléculaires dans la polyarthrite rhumatoïde

Quentin Miagoux

université PARIS-SACLAY

Approche globale intégrative pour l'identification de nouvelles cibles moléculaires dans la Polyarthrite Rhumatoïde

Integrative, Global approach for the identification of novel biomolecular targets in Rheumatoid Arthritis

Thèse de doctorat de l'Université Paris-Saclay

Ecole Doctorale n°577, Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Sciences de la vie et de la santé
Graduate School : Life Sciences and Health
Référent : Université d'Évry-Val d'Essonne

Thèse préparée dans l'unité de recherche Université Paris-Saclay, Univ Evry, Laboratoire Européen de Recherche pour la Polyarthrite rhumatoïde - Genhotel, 91057, Evry, France, sous la direction d'Élisabeth PETIT-TEIXEIRA, Professeure des universités, Université d'Évry-Val-d'Essonne, le co-encadrement de Valérie CHAUDRU, Maître de conférences, Université d'Évry-Val-d'Essonne et d'Anna NIARAKIS, Maître de conférences, Université d'Évry-Val-d'Essonne

Thèse présentée et soutenue à Evry, le 11/04/2022, par

Quentin MIAGOUX

Composition du jury

Anaïs BAUDOT Chargée de recherche, Université Aix-Marseille	Rapporteure
Fabienne LESUEUR Chargée de recherche, Institut Curie, PSL	Rapporteure
Mohamed ELATI Professeur des Universités, Université de Lille	Examinateur
Fabien FAUCHEREAU Maître de conférences, Université Paris Paris-Diderot	Examinateur
Élisabeth PETIT-TEIXEIRA Professeure des universités, Université d'Évry-Val-d'Essonne	Directrice de thèse
Valérie CHAUDRU Maître de conférences, Université d'Évry-Val-d'Essonne	Co-encadrante de thèse
Anna NIARAKIS Maître de conférences, Université d'Évry-Val-d'Essonne	Co-encadrante de thèse

THESE DE DOCTORAT
NNT : 2022UPASL013

Remerciements

Mes remerciements vont en premier lieu à mes deux encadrantes et à ma directrice de thèse. À Valérie Chaudru mon encadrante, merci tout d'abord de m'avoir donné l'opportunité de réaliser cette thèse, pour nos longues et passionnantes conversations scientifiques autour de la génomique et pour m'avoir poussé à réaliser de l'enseignement à l'université. Merci aussi pour le soutien et tout le temps qu'elle m'a consacré au cours de ces quatre années passées, temps pendant lequel j'ai énormément appris. Ce fût également un réel plaisir d'être encadré par une personne toujours souriante et joviale. À ma seconde encadrante, Anna Niarakis, pour son soutien, son temps consacré et ses encouragements. Je suis extrèmement reconnaissant d'avoir pu apprendre au côté de quelqu'un d'aussi dynamique, efficace et brillante. Enfin à ma directrice, Elisabeth Petit-Teixeira, tout d'abord pour m'avoir accueilli et donné l'opportunité de réaliser cette thèse. Merci pour le temps qu'elle m'a consacré, pour ses précieux conseils autant sur le plan scientifique qu'humain et pour son soutien dans les moments difficiles au cours de ces quatre dernières années.

Je tiens à remercier Mme Anaïs Baudot et Mme Fabienne Lesueur, d'avoir accepté d'être les rapporteures de mon manuscrit de thèse, ainsi que Mr Fabien Fauchereau et Mr Mohamed Elati pour avoir accepté d'être les examinateurs et membres de mon jury de thèse.

Je souhaite remercier également les membres de mon comité de thèse Carène Rizzon, Fabien Fauchereau et Mohamed Elati, pour leurs précieux conseils pendant le déroulement de ma thèse et pour leur sympathie.

Je voudrais remercier le département de Biologie de l'université d'Evry-Val-d'Essonne pour m'avoir donné l'opportunité de réaliser de l'enseignement au cours de mes quatre années de thèse. Merci à tous les enseignants qui m'ont intégré dans leur cours.

Je tiens à remercier Florence Hervy, secrétaire au département de Biologie, pour son aide dans toutes les démarches que j'ai pu réaliser en lien avec l'université et aussi pour sa bonne humeur.

Merci à toutes les personnes de mon laboratoire, à commencer par Maëva Veyssière, pour m'avoir transmis tes précieuses connaissances en bio-informatique avant ton départ du laboratoire et pour les bons moments passés ensemble. Merci également à Dereck de Mezquita pour avoir travailler sur mon projet de thèse et les discussions mémorables que nous avons eues. Je souhaite également remercier mes trois collègues en or, tout d'abord mon amie et collègue, Vidisha Singh, pour son implication dans mes projets de thèse, nos discussions scientifiques, sa bonne humeur et toutes ses petites attentions qui en font une personne si spéciale. À Nawel Zerrouk qui est d'une extrême gentillesse et que je remercie de m'avoir intégré dans l'un de ses projets de recherche. À Sahar Aghakhani avec qui j'ai partagé mon bureau pendant près de 2 ans, merci pour ta bonne humeur et nos conversations inoubliables. Je souhaite également remercier toutes les personnes qui ont partagé mon quotidien au laboratoire, Sara, Pierre, Sacha et Pilar.

Je souhaite remercier mes parents, sans qui, rien n'aurait été possible. Merci pour vos sacrifices qui m'ont permis de réaliser mon parcours et ma thèse. À ma

mère à qui je voudrais dédier ce travail; tu as toujours cru en moi et ce, même aux heures où plus personne n'y croyait. Ta détermination et ta persévérance ont toujours été une source d'inspiration. À mon père également, qui a toujours été un modèle pour moi et dont je me suis toujours inspiré pour me construire.

Je voudrais également remercier ma femme, Romane, d'avoir toujours été à mes côtés au cours de cette thèse. Merci de me donner au quotidien un amour et un bonheur inconditionnel. Merci pour ton soutien et ton courage lors des épreuves que nous avons traversées au cours de ces quatre ans. Merci également de m'avoir offert le plus beau des cadeaux lors de ces derniers mois de thèse, mon petit roi, Arthur. Ce petit cœur m'a donné le courage nécessaire pour aller de l'avant. Je vous aime tous les deux plus que trois fois mille.

Je souhaite remercier ma belle-famille, pour m'avoir soutenu, encouragé et réconforté lors de ces quatre dernières années. Je souhaite tout particulièrement remercier Alexis, mon beau-père et ami, pour son aide dans la relecture de ce manuscrit.

Merci également à chaque membre de ma famille, pour leur soutien et leur amour. J'ai également une pensée particulière aux personnes déjà parties, desquelles je tiens une grande source d'inspiration et qui m'ont permis de devenir la personne que je suis.

Enfin merci à mes amis, Tristan, Brandon, Harry, Abdel, Mathias, Dimitri, Jordan, Pierre, Sulyvan, Allan, William, Irvin et Kunlé, qui pour leur présence et leur bonne humeur, ont rendu cette thèse plus agréable.

Tables des matières

Liste des tableaux . . . . . **i**

Liste des figures . . . . . **iii**

Liste des abbréviations . . . . . **v**

Chapitre 1: La Polyarthrite Rhumatoïde **1**
1.1 Épidémiologie 3
1.1.1 Prévalence 3
1.1.2 Incidence 5
1.1.3 Mortalité 5
1.2 Diagnostic 6
1.2.1 Classification par critères 6
1.2.2 Identification de biomarqueurs 8
1.2.3 Imagerie 8
1.3 Physiopathologie 9
1.3.1 Inflammation non spécifique 10
1.3.2 Amplification de l'inflammation à la membrane synoviale . . 10
1.3.3 Inflammation chronique et destruction articulaire 11
1.4 Étiologie 13
1.4.1 Les facteurs environnementaux et non génétiques 13
1.4.2 Les facteurs génétiques 16
1.5 Traitement 22

Chapitre 2: La biologie des systèmes appliquée aux maladies complexes **27**
2.1 Inférence de réseaux 28
2.2 Cartes moléculaires maladie-spécifique 29
2.3 Modélisation computationnelle 30
2.3.1 Méthodes quantitatives et qualitatives 30
2.3.2 Modélisation booléenne 31

Chapitre 3: Objectifs **33**

Chapitre 4: Caractérisation de CNVs par étude de séquençage d'exomes **35**

4.1 Analyse des performances d'outils de détection de CNVs à partir de données WES . . . 38
4.1.1 Matériel et Méthodes . . . 42
4.1.2 Résultats . . . 52
4.1.3 Discussion . . . 62
4.2 Caractérisation de CNVs rares associés à la Polyarthrite Rhumatoïde 68
4.2.1 Matériel et Méthodes . . . 69
4.2.2 Résultats . . . 76
4.2.3 Discussion . . . 80

Chapitre 5: Caractérisation de variants rares par étude de séquençage génome entier . . . 83
5.1 Matériel et Méthodes . . . 85
5.1.1 Échantillons . . . 85
5.1.2 Séquençage et traitement des données génomiques . . . 88
5.1.3 Recherche des SNVs et indels candidats . . . 90
5.1.4 Recherche des CNVs candidats . . . 98
5.1.5 Typage des allèles *HLA-DRB1* . . . 102
5.2 Résultats . . . 103
5.2.1 Recherche de SNVs et indels candidats . . . 103
5.2.2 Validation des SNVs et indels candidats dans le set de validation 103
5.2.3 Recherche des CNVs candidats . . . 108
5.2.4 Validation des CNVs candidats dans le set de validation . . . 111
5.2.5 Typage de *HLA-DRB1* . . . 112
5.2.6 Variants candidats . . . 115
5.3 Discussion . . . 117

Chapitre 6: Inférence d'un réseau global intégratif spécifique de la Polyarthrite Rhumatoïde . . . 123
6.1 Matériel et Méthodes . . . 125
6.1.1 Description et filtres des données pour l'inférence d'un réseau de co-régulation . . . 125
6.1.2 Inférence d'un réseau de co-régulation . . . 126
6.1.3 Extraction des protéines de la carte moléculaire de la PR . . . 128
6.1.4 Création d'un réseau global intégratif et spécifique de la PR 129
6.1.5 Listes de variants de susceptibilité . . . 130
6.1.6 Analyse d'expression différentielle (DEA) dans deux jeux de données indépendants . . . 131
6.1.7 Extraction d'un sous-réseau pour étudier la réponse au traitement . . . 132
6.1.8 Shiny App . . . 132
6.1.9 Modélisation booléenne . . . 133
6.2 Résultats . . . 136
6.2.1 Inférence d'un réseau de co-régulation . . . 136
6.2.2 Extraction de la carte moléculaire de la PR . . . 139

6.2.3 Réseau global intégratif et spécifique de la PR 139
6.2.4 Superposition de variants génomiques et de DEGs 142
6.2.5 Sous-réseau spécifique de la PR 148
6.2.6 Modélisation booléenne 150
6.3 Discussion 159

Chapitre 7: Conclusion Générale 165

Annexe A: Facteurs de transcription mis en évidence par CoRegNet et leur implication dans la PR selon la litérature tiré de Miagoux et al. 2021 [333] 171

Annexe B: Liste des Communications 175
Écrites 175
Posters 175

Annexe C: Publication 177
Inference of an Integrative, Executable Network for Rheumatoid Arthritis Combining Data-Driven Machine Learning Approaches and a State-of-the-Art Mechanistic Disease Map 177

Bibliographie 197

Liste des tableaux

1.1 Prévalence de la PR dans différents pays en fonction du genre . . . 5
1.2 Critères de classification ACR/EULAR 2010 pour la PR 7
1.3 Facteurs environnementaux et autres associés à la PR 15
1.4 Acides aminés communs sur les positions 70-74 de la chaîne HLA-DRβ1 19
1.5 Les différents traitements de fond pour la PR 25

4.1 Performances de six outils de détection de CNV selon la litérature . 41
4.2 Distribution des CNVs selon leurs tailles (a), fréquences (b) et types (c) 43
4.3 Outils de détection de CNVs basés sur la méthode *read-depth* 50
4.4 Description des métriques de détetion de CNVs par région génomique et individu 51
4.5 Temps de calcul des six outils pour l'identification de CNVs simulés dans des données WES 100x pour 7 chromosomes dans un échantillon de 250 individus 60
4.6 Caractéristiques sérologiques et épidémiologiques pour les 30 individus du set de découverte 70
4.7 Nombre de CNVs et de CNVr détectés par les outils à l'échelle de la région génomique et de l'individu 73
4.8 Classification de l'effet pathogène des variants structuraux par AnnotSV 74
4.9 Résultats de la recherche de CNVs rares associés à la PR dans le set de découverte 78
4.10 Résultats de l'identification des CNVs par ddPCR dans le set de validation 79

5.1 Caractéristiques démographiques des 25 individus 86
5.2 Filtres préconisés selon la méthode GATK *Hard-filter* 91
5.3 Bases de données publiques incluant des fréquences alléliques de référence 92
5.4 Outils d'annotation de variants 94
5.5 Résultats du test d'association-liaison (pVAAST) sur les variants et gènes candidats 106
5.6 Résultats du filtrage des variants candidats dans le set de validation 107
5.7 Résultats des différentes étapes pour l'identification des CNVs dans le set de découverte 110

5.8 Résultats de l'identification des CNVs dans le set de validation . . . 111
5.9 Typage des allèles HLA-DRB1 de 24 individus (15 du set de découverte et 9 individus supplémentaires) 114
5.10 Identification de nouveaux facteurs génétiques rares et délétères, sans phénocopie et avec pénétrance complète associés à la PR. . . . 116

6.1 Conditions initiales pour l'analyse de relation dose-effet 135
6.2 Conditions initiales pour l'analyse des états stables du réseau booléen 135
6.3 Top 5 des facteurs de transcription (TF) identifiés par CoRegNet ayant le plus d'interaction de regulation (interaction TF-gène cible) et interaction de régulation (TF-TF) 138
6.4 Facteur de transcription à partir du réseau global intégratif et spécifique de la PR chevauchant au moins un gène différentiellement exprimé à partir des analyses répondeurs/non-répondeurs à des traitements anti-TNF et avant et après traitement anti-TNF 145
6.5 État stable du réseau booléen sans perturbation 156
6.6 État stable du réseau booléen incluant des perturbations 158

Liste des figures

1.1 Prévalence par pays en 2017 (pour 100,000 individus) 3
1.2 Développement et progression de la Polyarthrite Rhumatoïde 9
1.3 Mécanismes physiopathologiques impliqués dans la polyarthrite rhumatoïde 12
1.4 Les différents types de variants génomiques retrouvés chez l'Homme 17
1.5 Cibles des médicaments utilisés en traitement de fond (DMARDS) . 25

2.1 Exemple d'un réseau de gènes régulateurs (GRN) 29
2.2 Modèle de réseau booléen 32

4.1 Approches de détection de CNVs à partir de séquençage NGS lecture courte 40
4.2 Représentation par boîte à moustaches à partir du nombre de lectures normalisé d'une délétion identifiée par la méthode *Read-depth* tiré G. Povysil et al 2017 [205]. 46
4.3 Performances (sensibilité et précision) des six outils de détection de CNV à partir de données WES simulées, à l'échelle de la région génomique et de l'individu 53
4.4 Performances des six outils de détection selon différentes tailles de CNVs (<5kb, 5-150kb et >150kb). 55
4.5 Performances des six outils de détection selon différentes fréquences de CNVs (<1%, 1-5% et >5-95%). 57
4.6 Performances des six outils de détection selon les types de CNVs (délétion et duplication). 58
4.7 Top 20 des CNVs retrouvés par un ou plusieurs outils à l'échelle de l'individu 59
4.8 Performances des six outils de détection évaluées à partir de données réelles provenant du 1000 Genomes. 61
4.9 Familles étudiées 71
4.10 Répartition moyenne des CNVs identifiés par chromosomes. 77

5.1 Familles multiplexes de PR. 87
5.2 Pipeline de production des SNVs et indels à partir de données de lectures post-séquençage. 89

5.3 Résultats des différentes étapes permettant l'identification des 88 variants obtenus par séquençage de génome entier dans le set de découverte . . . 105
5.4 Résultats des différentes étapes pour l'identification des CNVs obtenus par séquençage de génome entier dans le set de découverte 109

6.1 Analyse en Composante Principale (PCA) réalisée à partir de données d'expression provenant des leucocytes de 95 individus (46 atteints de PR et 49 témoins) (GSE117769). . . . 126
6.2 Représentation d'un système biologique en deux langages de notation graphique de la biologie des systèmes : Flux d'activités et processus de description . . . 130
6.3 Représentation des différentes étapes pour l'obtention d'un sous-réseau contenant des règles booléennes et son analyse par simulation *in silico* . . . 134
6.4 Réseau de co-régulation inféré en utilisant CoRegNet et les données normalisées provenant du jeu de données transcriptomiques (GSE117769) . . . 137
6.5 Réseau de régulation créé à partir des TFs communs entre la carte moléculaire de la PR et le réseau de co-régulation CoRegNet . . . 140
6.6 Réseau global intégratif et spécifique de la PR . . . 141
6.7 Réseau global intégratif et spécifique de la PR et les variants spécifiques de la PR obtenus à partir de DisGeNET . . . 143
6.8 Réseau global intégratif et spécifique de la PR et les DEGs obtenus à partir de patients répondeurs et non-répondeurs à des traitements anti-TNF (37 et 41 patients atteints de PR traités respectivement par adalimumab et etanercept) . . . 146
6.9 Réseau global intégratif et spécifique de la PR et les DEGs obtenus à partir de patients traités et non-traités par traitement anti-TNF (deux cohortes de 40 et 36 patients atteints de PR avant et après une durée de trois mois de traitement avec infliximab ou adalimumab) 147
6.10 Sous-réseau extrait des protéines cibles (TGFB1, IL6, and TNF) du réseau global intégratif et spécifique de la PR. . . . 149
6.11 Simulation en temps réel du sous-réseau avec Cell Collective. . . . 152
6.12 Analyse de la relation dose-effet du sous-réseau . . . 154
6.13 Analyse de sensibilité du sous-réseau . . . 155
6.14 Réseau booléen incluant trois cascades de signalisation pour TNF, TGFB1 et IL6 créé à partir du sous-réseau. . . . 156

Liste des abbréviations

ACP	Analyse en Composante Principale
ACPA	Anti-peptides cycliques citrulinés
ACR	American College of Rheumatology
AF	Flux d'activité
ARA	American Rheumatism Association
BAM	Cartographie d'alignement binaire
BiNoM	Biological Network Manager
cAMP	Adénosine monophosphate cyclique
CLR	Rapport de vraisemblance composite
CPA	Cellules présentatrices d'antigènes
CRP	Protéines C-reactives
DAS	Disease Activity Score
DEA	Analyse d'expression différentielle
DEG	Gène différentiellement exprimé
DECIPHER	DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources
DGV	Database of Genomic Variants
DMARDS	Disease-modifying antirheumatic drugs
boDMARDS	DMARDS biologiques

bsDMARDS	DMARDS biosimilaires
csDMARDS	DMARDS synthétiques conventionnels
tsDMARDS	DMARDS synthétiques ciblés
EI	Index d'évidence
ESR	Taux de sédimentation des éryhtrocytes
EULAR	The European Alliance of Associations for Rheumatology
FDR	False discovery rate
FLS	Synoviocytes de type fibroblastique
FR	Facteur Rhumatoïde
GnomAD	Genome Aggregation Database
GRN	Gene Regulatory Network
GWAS	Genome-Wide Association Studies
GxE	Interactions gène-gène
GxG	Interactions gène-environnement
IgG	Immunoglobulines G
IL6	Interleukine 6
IRM	Imagerie par résonance magnétique
KO	Knock-Out
LDL	Lipoprotéines de basse densité
LDLR	Récepteur des lipoprotéines de basse densité
LED	Lupus érythémateux disséminé
LOD	Logarithme des probabilités
MAI	Maladie auto-immune
MHC	Complexe majeur d'histocompatibilité
NGS	Séquençage de nouvelle génération
NSAIDs	Anti-inflammatoires non stéroïdiens
PD	Processus de description

PPI	Interaction protéine-protéine
PR	Polyarthrite rhumatoïde
SBGN	Notation graphique de la biologie des systèmes
SE	Épitope partagé
SIF	Format d'interaction simple
SNP	Polymorphisme nucléotidique
SNV	Variant nucléotidique
nsSNV	Variant nucléotidique non-synonyme
SSP-PCR	Single Specific Primer-Polymerase Chain Reaction
SV	Variation structurale
TF	Facteur de transcription
TG	Gène cible
TGFB1	Facteur de croissance de transformation beta 1
TNF	Facteur de nécrose tumorale
VDA	Association variant-maladie
WES	Séquençage d'exome entier
WGS	Séquençage du génome entier

Chapitre 1

La Polyarthrite Rhumatoïde

La Polyarthrite Rhumatoïde (PR) est une maladie auto-immune (MAI) inflammatoire chronique et invasive touchant les articulations, entraînant leurs déformations et destructions. C'est une maladie complexe, impliquant des facteurs génétiques, épigénétiques et environnementaux [1].

Le mot Polyarthrite est dérivé de deux mots grecs, "Poly", signifiant plusieurs et "Arthrite" provenant du mot "Arthros" signifiant articulation et suggérant une inflammation. Le mot Rhumatoïde est dérivé du mot Grec "Rheuma" signifiant flux/rhume.

C'est en 1800 avec la thèse du médecin français Augustin-Jacob Landré Beauvais (1772-1840) que sont exposées les premières descriptions des symptômes de la PR sur des patients de l'asile de la Salpêtrière. Le Dr Beauvais choisit alors le terme de Goutte Asthénique Primitive [2]. En 1859, le médecin Anglais Sir Afred Baring Garrod (1819-1907), poursuivant les recherches menées jusqu'alors, écrit pour la première fois le mot Polyarthrite Rhumatoïde pour décrire la maladie [3].

Si, dès le XIXe siècle, il a été possible de nommer et décrire la maladie, l'origine exacte de cette maladie est cependant inconnue. Trois théories permettant d'expliquer son origine s'opposent :

1) La PR est une maladie moderne, et sa pathogénèse est le résultat de la combinaison de facteurs génétiques et environnementaux actuels.
2) La PR est une maladie existant déjà à l'époque de nos ancêtres lointains.
3) La PR était déjà présente dans les populations indigènes et s'est répandue en Europe lors de la découverte de l'Amérique du nord.
Toutefois, au cours de l'histoire, un grand nombre d'indices provenant de la littérature, de l'art et de la paléontologie permettent de soutenir la deuxième théorie [4]. En effet, des études de paléontologie étayent l'hypothèse de l'existence ancienne de la PR, grâce à l'analyse de différents squelettes datés entre 4000 av. J.-C. et 1666 apr. J.-C., provenant de plusieurs endroits de la planète [4, 5]. De par l'art, plusieurs tableaux, notamment Les Trois Grâces (Rubens, 1639) et La Tentation de Saint Anthony (par un auteur anonyme de l'école Flemish-Dutch, 15-16 ième siècle) montrent des personnages ayant des symptômes de la PR au niveau des mains [4, 6, 7]. Enfin, de l'antiquité à la renaissance, des symptômes qui s'apparentent à la PR ont été décrits dans un premier temps par Hippocrate (460-370 av. J.-C.), puis ensuite par de nombreuses personnes au cours de l'histoire telles que Aretaeus, Galien, Soranos d'éphèse, Michel Psellos et Rhazes pour ne citer qu'eux [4, 7]. Malgré ces preuves apportées par notre histoire allant dans le sens de la seconde théorie, elles peuvent facilement être remises en cause. D'une part car de multiples maladies sont similaires à la PR. D'autre part car l'âge d'apparition de la maladie étant d'environ 50 ans, cela ne concorde pas avec l'espérance de vie moyenne d'antan, qui était bien plus faible. Par exemple, la moyenne d'âge en France pour la génération née 1806 est d'environ 37 ans [8].

1.1 Épidémiologie

1.1.1 Prévalence

La prévalence[1] globale de la PR est estimée à 0.46%, selon une méta-analyse incluant 67 études composées de 10 études provenant d'Amérique du nord, 26 d'Europe, 26 d'Asie et 5 d'Afrique [9]. Une seconde étude plus exhaustive utilise des données provenant de 195 pays (soit tous les pays indépendants reconnus par l'organisation des nations unies) et estime cette prévalence à 0.25% [10]. Toutefois, les prévalences entre différents pays et régions du monde sont inégales, comme l'atteste la Figure 1.1. Les États-unis (0.38%) et l'Europe du nord (0.35%) font partie des régions et pays ayant les prévalences les plus élevées. À l'inverse, l'Asie du sud-est (0.10%), l'Océanie (0.14%) et l'Afrique subsaharienne de l'ouest (0.14%) font partie des régions et pays ayant les prévalences les plus faibles [10].

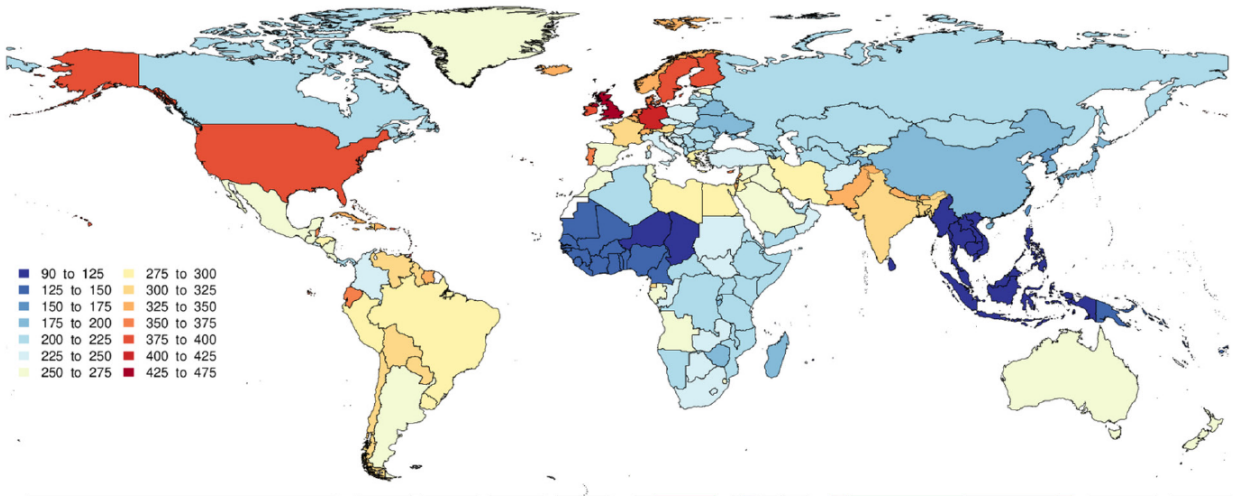

Figure 1.1: Prévalence par pays en 2017 (pour 100,000 individus). Adapté de Safiri et al. 2019 [10].

[1]Nombre de cas d'une maladie dans une population sur une période donnée

D'autres études menées aux États-Unis et dans des pays d'Europe du nord s'accordent sur une prévalence de la maladie dans ces régions comprises entre 0.5 et 1.1% [11, 12]. Les pays de l'Europe du sud ont par ailleurs des prévalences plus faibles comprises entre 0.3 et 0.7%. Enfin, en France, une première étude menée en 2001 à l'échelle de la population estime sa prévalence à 0.31%, tandis qu'une seconde étude basée en Bretagne estime sa prévalence à 0.62% [13, 14].

Les prévalences les plus élevées au monde ont été observées à une échelle locale, dans deux ethnies amérindiennes, les Pima et les Chippewa, avec des prévalences respectivement de 5.3 et 6.8% [15].

Par ailleurs, des études ont également montré que les femmes sont 2 à 3 fois plus susceptibles de développer la maladie que les hommes, ce qui a été vérifié dans plusieurs pays comme le suggère la Table 1.1 [16, 17]. Une hypothèse permettant d'expliquer ce ratio déséquilibré, implique des différences hormonales entre les deux sexes pouvant accroître le risque de développer la maladie chez les femmes. Une association a en effet été démontrée en cas de ménopause précoce ($OR = 2.42\ [1.32 - 4.45]$) [18], au stade post-ménopause ($HR = 2.1\ [1.4 - 3.0]$) [19], lors de la période post-partum (de la fin de l'accouchement jusqu'au retour de couches) ($IRR = 1.73\ [1.11 - 2.70]$) [20] et enfin en cas d'utilisation d'agent anti-oestrogène (e.g. cancer du sein) ($p < 0.0001$) [21]. Ces évènements ont tous en commun une chute des fonctions ovariennes et/ou la biodisponibilité des oestrogènes, qui permettent de stimuler ou inhiber le système immunitaire [22]. Enfin, les changements hormonaux induits par le syndrome de Stein-Leventhal (ou ovaires polykystiques), l'allaitement maternel ou encore le fait de donner naissance plus d'une fois seraient des facteurs de risques controversés [22].

Table 1.1: Prévalence de la PR dans différents pays en fonction du genre. Adapté de Tobón et al. 2010 [16].

Pays	Femmes (%)	Hommes (%)
Etats-unis	1.4	0.74
Royaume-uni	1.16	0.44
Espagne	0.8	0.2
Italie	0.51	0.13
France	0.51	0.09
Grèce	0.45	0.19

1.1.2 Incidence

En 2017, l'incidence[2] annuelle mondiale de la maladie est estimée à 14.9[3], ce qui représente une progression de 8.2% depuis 1990 [10]. Au niveau régional, les régions ayant une incidence élevée sont l'Amérique du nord (22.5), l'Asie du sud (20.7) et l'Europe de l'ouest (20.4). On note parmi les régions ayant une incidence faible l'Asie du sud-est (6.2), l'Océanie (7.9) ainsi que l'Afrique subsaharienne de l'ouest (8.5) [10]. À l'échelle nationale, le Royaume-Uni (27.5), l'Irlande (23.7), et la Suède (23.4), sont les pays ayant l'incidence la plus élevée. Au contraire, l'Indonésie (5.6) et le Sri Lanka (5.9) sont les pays ayant l'incidence la plus faible [10].

1.1.3 Mortalité

Les maladies cardiovasculaires sont la principale cause de mort prématurée chez les individus atteints de PR [23]. Une première étude, réalisée entre 1997 et 2012 sur la mortalité de la PR effectuée sur un échantillon de la population hollandaise (1222 patients dont 72.6% des patients sont des femmes) ayant en moyenne 60 ans, montre une mortalité de 54% supérieure à la population générale, avec une mortalité accrue par des comorbidités telles que les maladies cardiovasculaires, respiratoires, musculosquelettiques et digestives [24]. Une seconde étude réalisée

[2]Nombre de nouveaux cas sur une période donnée
[3]Cas pour 100,000 individus

sur 119 209 femmes infirmières anglaises par Nurses' Health Study entre 1976 et 2012 a confirmé ces résultats [25]. Cette seconde étude démontre que la PR associée à des maladies cardiovasculaires et respiratoires, engendrerait un risque accru de mortalité. En revanche, la présence de cancer chez les femmes PR ne semble pas accroître la mortalité. [25]. De plus, sans autre comorbidité, les femmes atteintes de PR ont un risque de mortalité plus élevé que les femmes sans PR. Cependant, une surveillance des facteurs sérologiques et génétiques, permet d'adapter les thérapies et de prévenir ces risques prématurés de mortalité. Ainsi, lorsque des patients atteints de PR juvéniles sont pris en charges avec les thérapies actuelles, la mortalité prématurée a tendance à diminuer voir ne plus être observée [23, 26].

1.2 Diagnostic

Il n'existe à ce jour aucun diagnostic spécifique de la PR. Cependant, un ensemble de critères permet de définir la maladie.

1.2.1 Classification par critères

Entre 1987 et 2010, les rhumatologues se référaient à des critères de classification établis par l'*American Rheumatism Association* (ARA; renommée depuis en *American College of Rheumatology*). Ces critères ont été jugés depuis inadaptés, ne permettaient pas, par exemple, une détection de la PR précoce [27]. Depuis 2010, l'*American College of Rheumatology* (ACR) et l'*EUropean League Against Rheumatism* (EULAR; renommée depuis en *The European Alliance of Associations for Rheumatology*) ont établi de nouveaux critères de classification pour la Polyarthrite Rhumatoïde [28], détaillés dans la Table 1.2. Ces nouveaux critères de classification ont montré une sensibilité[4] de 21% supérieure aux critères de

[4] $\text{Sensibilité} = \dfrac{\text{vrai positif}}{(\text{vrai positif} + \text{faux positif})}$

classification établis en 1987, malgré une baisse de 16% de spécificité[5], selon une méta-analyse basée sur 6 articles et 4 résumés [29]. Cette nouvelle classification utilise notamment l'étude de caractéristiques cliniques de l'atteinte articulaire, l'identification de facteurs inflammatoires par sérologie, ainsi qu'une étude sur la durée de la synovite par imagerie. Chaque critère est ensuite transformé en score puis additionné. Ce résultat permet alors de poser le diagnostic de PR, si ce score est supérieur ou égal à 6.

Table 1.2: Critères de classification ACR/EULAR 2010 pour la PR

Critère de classification ACR/EULAR pour la PR	Score
Articulation	**0-5**
1 Articulation importante (épaule, coude, hanche, genou, cheville)	0
1-3 Articulations importantes	1
1-3 Petite(s) articulation(s)	2
4-10 petites articulations	3
>10 Articulations	5
Durée des symptômes	**0-1**
<6 Semaines	0
≥6 Semaines	1
Sérologie	**0-3**
FR négatif et ACPA négatif	0
FR positif faible et ACPA positif faible	2
FR positif fort et ACPA positif fort	3
Réactifs de phase aigüe	**0-1**
Niveaux ESR et CRP normaux	0
Niveaux ESR ou CRP anormaux	1

Note:
ACPA : Anti-peptides cycliques citrulinés
CRP : Protéines C-réactives
ESR : Taux de sédimentation des éryhtrocytes
FR : Facteur Rhumatoïde

[5] $\text{Spécificité} = \frac{\text{vrai négatif}}{(\text{vrai négatif} + \text{faux positif})}$

1.2.2 Identification de biomarqueurs

Le Facteur Rhumatoïde (FR) et les anticorps Anti-Peptides Cycliques Citrulinés (ACPA) sont deux anticorps, respectivement produits en réaction aux Immunoglobulines G (IgG) et aux protéines citrulinées, produits au cours de la réaction auto-immune provoquée par la PR. Ces anticorps sont des biomarqueurs utilisés par sérologie lors d'une suspicion de PR chez un patient. Si la présence d'ACPA est vérifiée chez un patient atteint de PR, il s'agit de séropositivité (ACPA+) et dans le cas contraire de séronégativité (ACPA-). Les tests utilisés ont une sensibilité de 69 et 67% et une spécificité de 85 et 95% pour le FR et l'ACPA respectivement [30]. Ces biomarqueurs sont également utilisés dans le cadre d'un pronostic ou d'une PR précoce, car pouvant être présent jusqu'à 10 ans avant l'apparition clinique de la maladie [1, 31–33]. Cependant, ces biomarqueurs sont présents uniquement chez ~80% des personnes ayant une PR avérée, et ~50% des personnes étant au stade précoce de la maladie [34]. D'autre part, ces biomarqueurs ne sont pas spécifiques de la PR : dans le cas du FR, il peut également être retrouvé dans d'autres MAI telles que le rhumatisme psoriasique (<15%), le lupus érythémateux disséminé (15–35%) la sclérodermie systémique (20–30%) et le syndrome de Sjögren (75–95%). Il est également retrouvé dans des maladies infectieuses de type bactérienne, virale et parasitaire (e.g. tuberculose, VIH, hépatite A, B, C et malaria) [35–37]. Enfin, les biomarqueurs ACPA et FR peuvent aussi être retrouvés chez des individus sains, avec une fréquence allant de 2 à 5% [34, 36].

1.2.3 Imagerie

L'imagerie est également utilisée en cas de suspicion de PR chez un patient. Les deux techniques à ce jour les plus utilisées sont l'imagerie par résonance magnétique (IRM) et par ultrason. La première permet de détecter des œdèmes de la moelle

osseuse comme future zone d'érosion, et de différencier une synovite (inflammation de la membrane synoviale) d'autres douleurs et gonflements non-inflammatoires [34, 38]. La seconde permet de quantifier le degré et l'étendue de la synovite inflammatoire [34, 39].

1.3 Physiopathologie

La PR commence à se manifester par une activité anormale du système immunitaire, provoquant plus tard la déformation et la destruction des articulations. Ce phénomène est divisé en trois phases de progression : une phase d'inflammation non spécifique, une phase d'amplification de l'inflammation, puis une inflammation chronique entraînant une destruction et déformation articulaire (Figure 1.2). Ces différentes phases impliquent plusieurs tissus et sollicitent l'immunité innée et adaptative via les lymphocytes B et T, les macrophages et les cellules dendritiques (Figure 1.3).

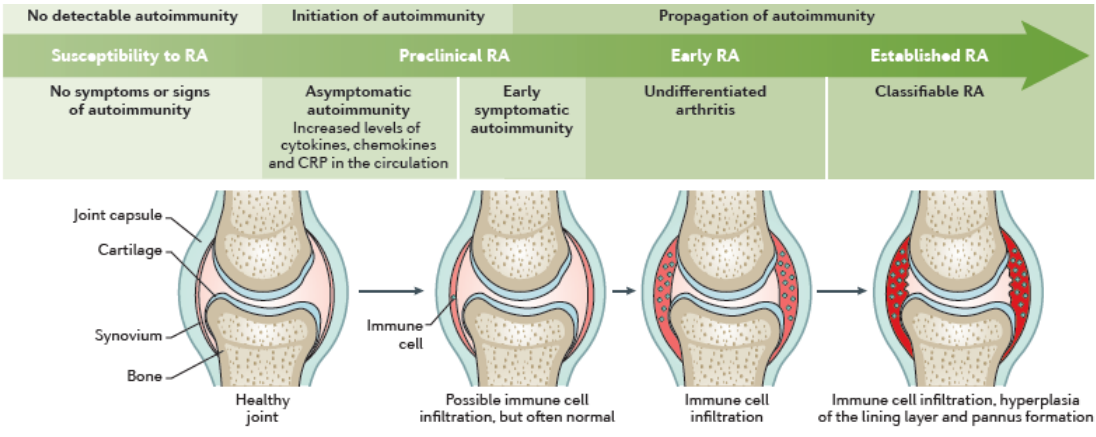

Figure 1.2: Développement et progression de la Polyarthrite Rhumatoïde. Selon Smolen et al. 2018 [1].

1.3.1 Inflammation non spécifique

La première phase de la PR, identifiée comme une inflammation non spécifique, est souvent appelée phase pré-PR et/ou pré-clinique. Celle-ci est initiée par des facteurs environnementaux et/ou génétiques, ce qui provoque un stress dans les muqueuses (bouche, poumon, intestin) [40, 41]. Ce stress entraîne une conversion post-transcriptionnelle d'acides aminés modifiant alors des peptides (e.g. citrullination, acetylation ou carbamylation). Ces peptides altérés sont alors reconnus par le système immunitaire inné et ses cellules présentatrices d'antigènes (CPA) et se fixent alors à leur surface via le complexe majeur d'histocompatibilité (MHC). Ce phénomène de fixation est renforcé dans le cadre d'épitope partagé des allèles du gène *HLA-DRB1* qui est le gène majeur de la PR et sera décrit ultérieurement (voir 5.1.5). Les antigènes fixés alors aux ACP sont ensuite présentés aux lymphocytes T dans les tissus lymphoïdes (Figure 1.3.b) qui en retour stimulent les lymphocytes B afin de synthétiser des anticorps, tels que les FR et ACPA [1, 42, 43]. Ce processus est en apparence le signe d'une réaction immunitaire normale, et ce, malgré l'apparition de biomarqueurs connus dans la PR (ACPA et FR) [1, 44].

1.3.2 Amplification de l'inflammation à la membrane synoviale

La membrane synoviale, aussi appelée synovium est un tissu spécialisé qui lie la surface interne des capsules articulaires synoviales à la gaine du tendon. Le synovium possède deux rôles : 1. La production des lubrifiants permettant de faible frottements sur les surfaces du cartilage et 2. La production des nutriments essentiels au cartilage. L'inflammation de la membrane synoviale, appelée synovite, apparaît lors d'une infiltration de leucocytes dans le compartiment synovial. Cette infiltration est rendue possible via une migration cellulaire par le biais d'une augmentation de

chimiokines, ainsi qu'une augmentation de molécules adhésives (intégrine, sélectine) dans la membrane synoviale [45, 46]. Ces changements entraînent alors une suractivation des deux types cellulaires : les synoviocytes de type macrophage (MLS) et synoviocytes de type fibroblastique (FLS) (aussi appelé respectivement fibroblastes A et B) constituant la membrane synoviale, et qui sont une source conséquente de cytokines et protéases.

1.3.3 Inflammation chronique et destruction articulaire

Une inflammation chronique de la membrane synoviale entraîne une déformation et une destruction des articulations. Cela est largement engendré par les MLS et les FLS. Ces derniers auraient même un comportement invasif, permettant la migration d'articulation en articulation, propageant alors la maladie [1, 47]. Les cytokines, produites par les MLS, activent les FLS adjacents, lymphocytes T ainsi que les cellules dendritiques. Ces cellules produisent à leur tour des cytokines additionnelles, activant d'autres cellules dans le milieu articulaire. Ainsi, le mécanisme d'action des cellules synoviales produisant des cytokines engendre une inflammation perpétuelle de la PR (Figure 1.3.c). Les principaux acteurs de la destruction des cellules du cartilage sont les FLS positifs à la cadhérine, produisant des métalloprotéinases matricielles (MMP) telles que des collagènases et stromélysines [1, 48, 49]. L'érosion des os est, quant à elle, en partie provoquée par les MLS, par un mécanisme impliquant la différentiation et prolifération des ostéoclastes vers la surface du périoste, adjacent aux articulations [46, 50].

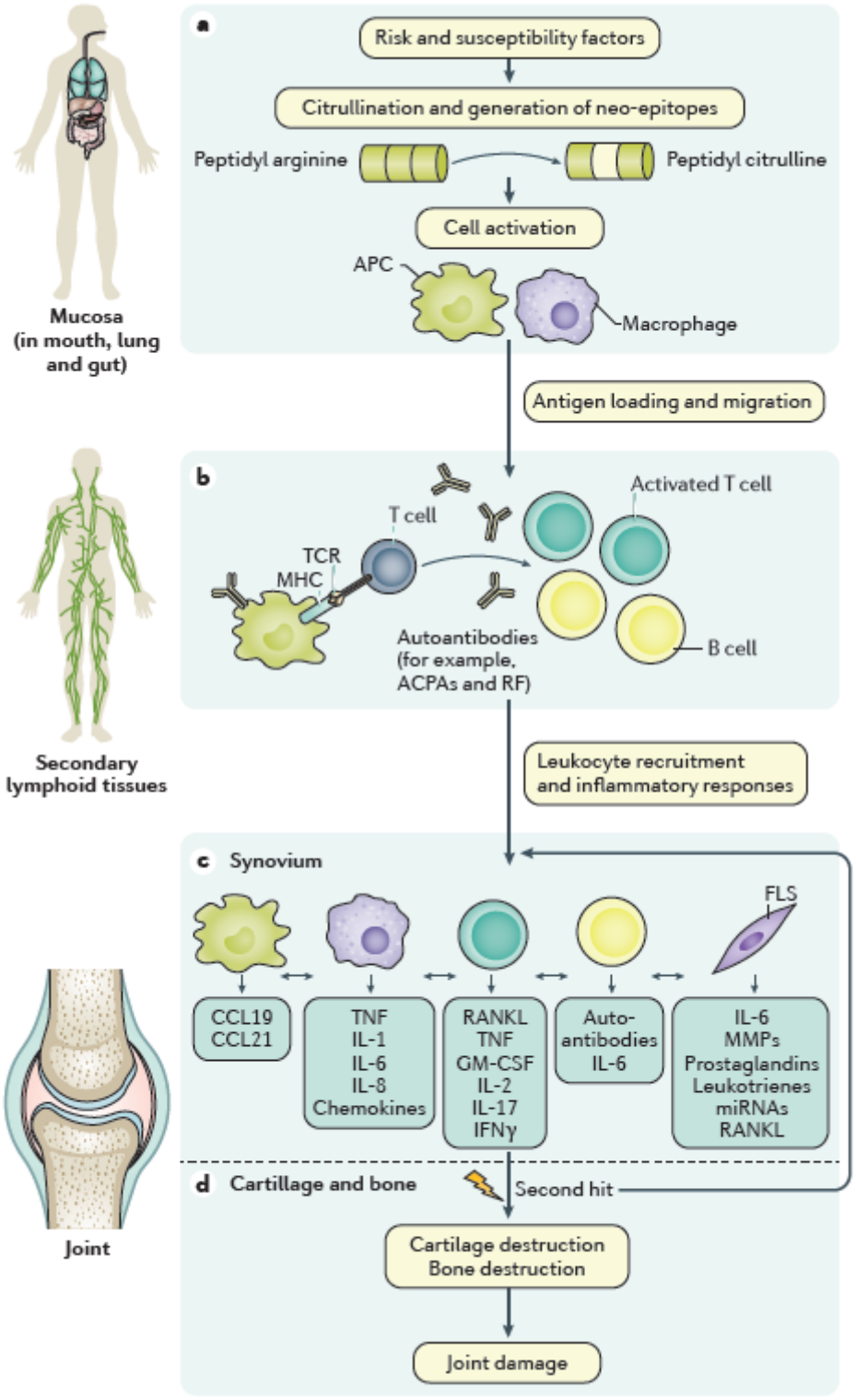

Figure 1.3: Mécanismes physiopathologiques impliqués dans la polyarthrite rhumatoïde. Selon Smolen et al. 2018 [1].

1.4 Étiologie

La polyarthrite rhumatoïde est une maladie complexe et son développement implique plusieurs facteurs de risques tels que des facteurs environnementaux, génétiques et épigénétiques.

1.4.1 Les facteurs environnementaux et non génétiques

1.4.1.1 Tabac

Le tabagisme est le principal facteur environnemental de la PR. Il explique 20 à 30% de la composante environnementale de la maladie [51, 52]. Au cours des quatre dernières décennies, de nombreuses études épidémiologiques ont démontré un risque accru de développer la PR dû à une exposition au tabac [53–62]. Parmi ces études, certaines démontrent un risque renforcé chez les hommes comparés aux femmes [54, 55, 60, 63]. En effet, le risque de PR serait environ deux fois supérieur chez des hommes fumeurs par rapport à des hommes non-fumeurs [60] et approximativement 1.3 fois supérieur chez des femmes fumeuses par rapport à des non-fumeuses [60]. Ces résultats sont confortés par le fait que le tabagisme a un effet direct sur le système immunitaire, causant un stress oxidatif, une apoptose cellulaire, un état systémique inflammatoire, une production d'auto-anticorps et des changements épigénétiques [62]. Certaines études n'ont cependant pas montré d'association entre la PR et le tabac [64, 65].

1.4.1.2 Les autres facteurs environnementaux

Un nombre conséquent d'autres facteurs environnementaux associés à la PR a été démontré par diverses études, augmentant ou diminuant les risques de développer la maladie, référencés dans la Table 1.3. Par exemple, une alimentation saine basée sur la consommation de fruits, de légumes, de céréales complètes, de faible quantité

de sucre, de graisse animale, de poissons et d'acide gras omega-3, diminue les risques de PR. À l'inverse, une consommation accrue de sodium, de viande rouge, de fer et un manque de vitamine D augmentent les risques de PR [52]. L'obésité est également un facteur environnemental aggravant, mais cette association est controversée, car une étude a montré l'augmentation de l'indice de masse corporelle chez des hommes aurait un effet protecteur et serait non associé chez les femmes [66].

D'autre part, une consommation d'alcool modérée (1-2 verres) sur une longue période est associée à une diminution du risque de développer la maladie. D'autres facteurs diminuant le risque de développer la PR ont été identifiés comme la prise de statine ainsi que l'utilisation de contraceptif oral. Ces facteurs sont encore étudiés afin de comprendre leurs mécanismes.

Enfin, la pollution de l'air et l'exposition à la poussière sont des facteurs environnementaux aggravant les risques de développer la maladie, bien que difficiles à analyser. Cependant, ces deux facteurs seraient liés à un faible statut socio-économique. Ainsi les individus avec un faible statut socio-économique seraient alors plus exposés à ces facteurs [67, 68].

1.4.1.3 Microbiote

Les maladies parodontales ont été identifiées comme associées à un risque accru de développer la PR [69]. Le microbiote buccal et notamment les agents pathogènes *Porphyromonas gingivalis* et *Aggregatibacter actinomycetemcomitans* responsables des inflammations parodontales, seraient des candidats potentiels à cette association entre les deux maladies [70, 71]. Ces agents pathogènes induisent une hyper citrullination dans leurs cellules hôtes, les neutrophiles, impliqués dans le développement de la PR [70–74]. Le microbiote intestinal pourrait également jouer un rôle important dans la maladie, sa diversité a été identifiée comme réduite chez

des patients atteints de PR comparé à la population [75]. De plus, une association entre les virus et la PR a également été proposée. Si les rôles des virus Chikungunya et Epstein-Barr avec la PR sont avérés, le rôle du parvovirus B-19 reste à établir [76–78].

Table 1.3: Facteurs environnementaux et autres associés à la PR. tiré de Deane et al. 2017 [52].

Facteurs	Réferences
Augmentation du risque	
Tabagisme	[53–62]
Exposition à la poussière (silice)	[79–82]
Pollution de l'air	[68, 83–88]
Consommation accrue de sodium, viande rouge et fer	[89, 90]
Obésité	[91–93]
Manque de vitamine D	[94]
Diminution du risque	
Consommation de poisson et acide gras omega-3	[95–101]
Consommation modéré d'alcool	[102–104]
Alimentation saine	[105–107]
Prise de statine	[108, 109]
Utilisation de contraceptif oral et hormone	[110, 111]

1.4.2 Les facteurs génétiques

La composante génétique impliquée dans la PR, aussi appelée héritabilité, est estimée à environ 60% d'après une étude réalisée sur des jumeaux [112]. La composante génétique s'explique par les variations génomiques qui existent entre les individus. Il existe de nombreuses variations génomiques, allant d'un simple nucléotide (SNV) à des duplications ou délétions de tout ou partie du chromosome (Copy Number Variations, CNVs), illustrées dans la Figure 1.4. Actuellement, le catalogue des variations génomiques est largement enrichi par les SNPs (SNV commun dont la fréquence est supérieure à 1%) et leur association à des phénotypes via les nombreuses analyses d'associations pan-génomiques (genome-wide association studies) [113]. À noter que le génome humain comporte un SNP tous les 1000 à 2000 bases [114]. Cependant, de plus en plus d'études s'intéressent à d'autres types de variations génomique, tels que les CNVs. Les CNVs sont des variations structurales (SVs) du génome humain et sont définis comme une variation du nombre de copie d'une région génomique d'ADN supérieure à 50pb par rapport à un génome de référence. Ils peuvent être retrouvés sous forme de duplication ou de délétion à une ou plusieurs positions du génome. Si la majorité des variations génomiques sont bénignes et permettent au génome humain d'évoluer et de s'adapter, elles peuvent également être délétères [115, 116]. Celles-ci modifient le fonctionnement d'un gène ou d'une protéine, entraînant une réaction en chaîne ayant un effet direct ou indirect via des voies métaboliques jouant un rôle dans une maladie telle que la PR [117–120]. Cependant, l'ensemble des facteurs génétiques identifiés dans la PR à ce jour explique uniquement environ 50% de la composante génétique impliquée dans cette maladie [121].

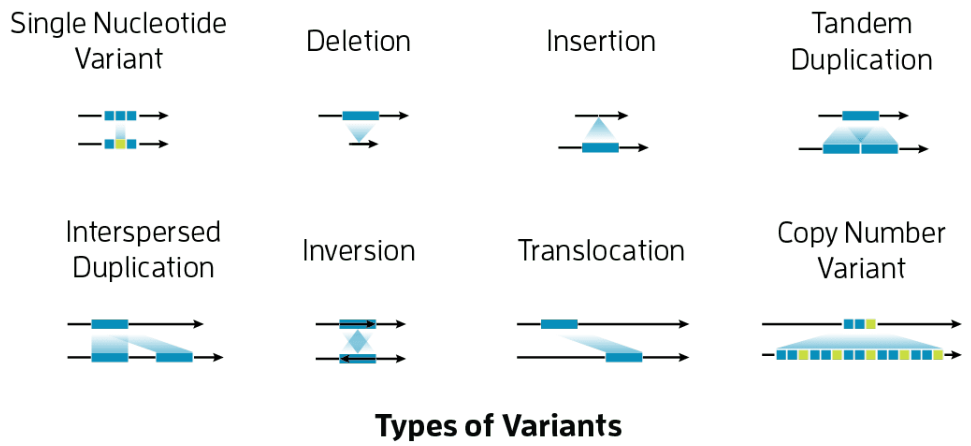

Figure 1.4: Les différents types de variants génomiques retrouvés chez l'Homme.

1.4.2.1 Le locus HLA

Le locus de l'antigène des leucocytes humains (HLA), également appelé complexe majeur d'histocompatibilité (MHC), est situé sur le chromosome 6 (6p21.3) et s'étend sur 3.6Mb. Ce complexe contient environ 220 gènes possédant principalement des fonctions immunorégulatrices et est divisé en trois régions dites classe I, II et III.

Le facteur génétique majeur associé à la PR est le locus HLA de classe II. Celui-ci encode des molécules de surface qui sont retrouvées sur les cellules présentatrices d'antigènes, responsables de la présentation de pathogènes extra-cellulaires aux lymphocytes T, résultant alors en une réponse immunitaire [122]. Cette association est plus précisément due à l'hétérodimère HLA-DR et sa sous-unité *HLA-DRB1* [123] qui se trouve au sein du locus HLA de classe II. Le gène HLA-DRB1 et ses allèles à risque dits *shared epitope* (SE), correspondent à une séquence de cinq acides aminés, des positions 70 à 74 de la chaîne HLA-DRβ. Ces allèles SE constituent à ce jour l'association génétique la plus forte avec la PR, et représente à minima 30% de la composante génétique [124, 125]. Cependant, des résultats publiés ces dernières années divergent légèrement de l'hypothèse des allèles SE associés à la

PR, et montreraient en réalité une association plus forte avec la PR pour les acides aminés en positions 11, 71 et 74 [126, 127].

En 2010, le *WHO Nomenclature Committee for Factors of the HLA System* établit un système d'identifiant unique afin de définir les différents allèles HLA (e.g. HLA-DRB1*XX:XX:XX:XX). Ainsi, un identifiant HLA à 4 chiffres permet de déterminer la structure de la protéine à l'échelle des acides aminés. Par exemple, HLA-DRB1*04:02 donne les acides aminés DERAA.

Parmi les allèles SE les plus communs, nous retrouvons les allèles HLA-DRB1*04:01 et *04:09 pour le motif QKRAA, HLA-DRB1*01:01, *04:08 et *04:04 pour le motif QRRAA et HLA-DRB1*10:01 pour le motif RRRAAA. Ces allèles, ainsi que d'autres allèles SE et non-SE additionnels sont listés dans la Table 1.4, adaptée de Trier et al. 2018 [128]. Les motifs SE sont identiques sur les positions 72-74, avec les acides aminés RAA. Leurs différences se situent sur les positions 70 et 71, où K confère un risque supérieur, R intermédiaire, et enfin A et E un risque faible [128]. Le motif QKRAA encodé par le variant HLA-DRB1*04:01 est le plus commun dans les populations caucasiennes [129]. QRRAA est le deuxième motif SE le plus commun, et selon les populations, l'allèle codant pour ce motif diffère, HLA-DRB1*04:04 est majoritaire chez les populations caucasiennes, HLA-DRB1*04:05 est majoritaire dans la population japonaise, HLA-DRB1*04:08 est majoritaire dans les populations européennes de l'ouest, asiatiques et caucasiennes et HLA-DRB1*14:02 est majoritaire dans la population native américaine. Enfin, RRRAA est le motif le plus rare et son allèle unique, HLA-DRB1*10:01 est notamment retrouvé dans les populations asiatiques, méditerranéennes et africaines [129].

Les individus porteurs de deux allèles SE sont plus exposés à une forme sévère de la PR que les porteurs d'un seul allèle, qui sont également plus exposés à une forme sévère que les individus qui n'en n'ont pas [130]. À l'inverse, il a été démontré

que le motif DERAA avait un rôle protecteur contre les formes sévères de PR [131].

Enfin, est estimé jusqu'à 20% le pourcentage d'individus atteints de la PR n'ayant pas d'allèles SE (SE négatif) [124].

Table 1.4: Acides aminés communs sur les positions 70-74 de la chaîne HLA-DRβ1. Selon Trier et al. 2018 [128].

Acides aminés	**Motif SE**	**Allèles HLA**
QKRAA	+	***04:01**, ***04:09**, *04:13, *04:16, *04:19, *04:21,*14:21
QRRAA	+	***01:01**, ***01:02**, *01:05, ***04:04**, ***04:05**, ***04:08**, *04:10, *04:19, ***14:02**, 14:06, *14:09, *14:13, *14:17, *14:20
RRRAA	+	***10:01**
DKRAA	-	*13:03
DERAA	-	*01:03, *04:02, *11:02, *11:03, *11:16, *11:20, *11:21, *13:01, *13:02, *13:04, *13:08, *13:15, *13:17, *13:19, *13:22, *13:23, *14:16, *15:01
QRRAE	-	*04:03, *04:06, *04:07, *04:11, *04:17, *04:20
RRRAE	-	*09:01, *14:01, *14:04, *14:05, *14:07, *14:08, *14:10, *14:11, *14:14, *14:18
DRRAA	-	*04:15, *08:05, *11:01, *11:04, *11:05, *11:06, *11:09, *11:10, *11:12, *11:15, *11:18, *11:19, *11:22, *12:01, *13:05, *13:06, *13:07, *13:11, *13:12, *13:14, *13:21, *13:25, *14:22, *16:01, *16:05
QARAA	-	*13:09, *15:01
QKRGR	-	*03:01, *04:22, *11:07
DRRGQ	-	*07:01
DRRAL	-	*08:01

Note :
SE : Shared Epitope
Allèles en gras : Allèles les plus fréquemment retrouvés associés à la Polyarthrite Rhumatoïde

1.4.2.2 Les facteurs génétiques non-HLA

Caractérisation de gènes de susceptibilité communs

Au cours des dernières décennies, afin d'identifier des facteurs génétiques communs dans la PR, hors locus HLA, des études d'analyses de liaisons et d'analyses d'association gènes candidats ou pangénomiques (Genome-Wide Association Studies, GWAS) ont été menées.

En 2003, les premières études d'identification de gènes associés à la PR, ont trouvé un variant dans le gène *PADI4*, impliqué dans la citrulination des protéines [132]. Puis en 2004, l'association entre le gène *PTPN22* et la PR est découverte [133]. Ce gène encode une tyrosine phosphatase qui est impliquée dans les voies de signalisations des cellules B et T. Par la suite, les approches gènes candidats ont permis d'identifier les gènes *CTLA4* et le locus *TRAF1/C5* [134, 135]. Plusieurs études ont également utilisé cette approche afin de démontrer l'association des gènes *STAT4*, *IL2*, *IL6* et *NF-kβ* avec la PR, gènes impliqués dans des voies inflammatoires et de réponses auto-immunes [136].

Ces premières découvertes, basées sur des études d'associations génétiques gènes candidats permettaient de cibler un ou quelques gènes seulement. Les puces de génotypage de SNPs, ont ensuite permis de réaliser des études GWAS et de découvrir d'autres loci associés tels que *TRAF1*, *CTLA4*, *IRF5*, *STAT4*, *FCGR3A*, *IL6ST*, *IL2RA*, *IL2RB*, *CCL21*, *CCR6*, *CD40* [137]. D'autre part, certains gènes étaient retrouvés indépendamment du statut sérologique des individus (*PTPN22*, *BLK*, *ANKRD55* et *IL6ST*), alors que d'autres gènes étaient uniquement retrouvés chez des individus ACPA+ (*AFF3*, *CD28* et *TNFAIP3*) ou uniquement chez des individus ACPA- (*PRL* et *NFIA*) [138, 139].

Finalement, l'ensemble de ces études a permis de découvrir plus de 100 loci non-HLA communs associés avec la PR, dont la plupart sont impliqués dans des mécanismes immunitaires [136, 140].

Caractérisation de variants rares par séquençage à haut-débit

Les études GWAS ont mené à la découverte de 100 loci communs associés à la PR. Cependant, ces variants de susceptibilité communs et le gène *HLA-DRB1* ne permettent d'expliquer que 50% de la composante génétique [140, 141]. Par exemple, environ 20% des cas de PR n'ont pas d'allèle *SE*, qui est pourtant le facteur génétique majeur de la PR [124]. Parmi les hypothèses permettant d'expliquer la composante génétique l'héritabilité manquante, l'hypothèse des variants rares est reconnue [142]. Ainsi, des études récentes se sont intéressées aux variants de susceptibilité rares, jusqu'alors non étudiés. Des études d'association cas/témoins [120, 143] et des études utilisant des données familiales [144, 145] visant spécifiquement les variants rares ont permis de caractériser de nouveaux gènes de la PR. Enfin, les interactions gène-gène (GxG) et gène-environnement (GxE) peuvent également jouer un rôle dans la composante génétique. Par exemple, l'effet d'un gène peut être différent selon le génotype d'un deuxième facteur génétique (interaction GxG). Une étude d'interaction entre les gènes *HLA-DRB1* et *PTPN22* a été mise en évidence à partir de 3 cohortes, suédoises (EIRA), nord-américaines (NARAC) et hollandaises (EAC) composées de 1977 cas et 2405 témoins. Cette interaction est précisément due à une interaction entre les allèles *SE HLA-DRB1* et l'allèle R620W du gène *PTPN22* [146].

1.4.2.3 Les autres facteurs

D'autres facteurs sont aussi étudiés afin de mieux connaître l'étiologie de la PR et incluent des facteurs épigénétiques et les micro-ARN.

Les mécanismes épigénétiques

Des études concernant les mécanismes épigénétiques, potentiellement impliqués dans la PR, se multiplient [147]. L'épigénétique est l'étude des mécanismes modifiant

de manière réversible, transmissible et adaptative l'expression des gènes sans en changer la séquence nucléotidique. Parmi ces mécanismes, une hypométhylation de certaine région de l'ADN a été observée par plusieurs études chez des patients atteints de PR, dans les cellules immunitaires et les cellules sanguines périphériques et les fibroblastes synoviaux [147]. Aussi, de récentes études ont montré une modification des histones dans les cellules immunitaires et les FLS d'individus atteints de PR [147].

Les micro-ARN

Des études ont également révélé une régulation anormale des micro-ARN dans la PR. Ce phénomène peut également être inclus dans les mécanismes épigénétiques, car impactant l'expression des gènes sans en modifier la structure ADN. Cette régulation anormale des micro-ARN serait impliquée dans le processus inflammatoire de la PR, incluant le contrôle et la production de cytokine, ainsi que la protection des tissus cartilagineux [147, 148].

1.5 Traitement

Il n'existe actuellement aucun remède à la PR. Cependant, des traitements symptomatiques et de fonds existent.

Les traitements symptomatiques permettent de soulager la douleur et les gonflements, sans pour autant modifier les mécanismes menant à la destruction articulaire [1]. Parmi les médicaments utilisés pour les traitements symptomatiques, nous retrouvons les anti-inflammatoires non stéroïdiens (NSAIDs) avec l'ibuprofène et l'aspirine [149]. Ces traitements symptomatiques peuvent avoir des effets néfastes après une longue utilisation tels que des évènements gastro-intestinaux ainsi qu'un effet sur la coagulation du sang [1]. Il est également possible d'utiliser les glucocor-

ticoïdes (e.g. le prednisolone), qui agissent en modifiant l'activité de la maladie, mais ne peuvent être utilisés que sur une période limitée.

Les traitements de fonds permettent une réduction de l'activité de la maladie et dans les meilleurs cas, une rémission [1]. Ainsi, ils permettent d'interférer avec le processus inflammatoire de la maladie et sont communément appelés en anglais *Disease-modifying antirheumatic drugs* (DMARDs). Les traitements de fond sont divisés en plusieurs catégories : les DMARDS synthétiques conventionnels (csDMARDS), synthétiques ciblés (tsDMARDS), biologiques (boDMARDS) et biosimilaires (bsDMARDS) présentés dans la Table 1.5 [1, 150]. Les cibles des médicaments présentés dans la Table 1.5 sont illustrés dans la Figure 1.5, tirée de Smolen et al. 2018 [1]. Les csDMARDS ont des cibles inconnues bien qu'ils contiennent un des DMARDS les plus utilisés : le Méthotrexate. Les médicaments baricitinib et tofacitinib, tous deux des tsDMARDS, ciblent JAK1 et JAK2, ainsi que JAK1, JAK2 et JAK3 respectivement. Les médicaments adalimumab, certolizumab, etarnercept, golimumab et inflixmab ciblent TNF tandis que les médicaments tocilizumab, sarilumab, clazakizumab olokizumab et siurkumab ciblent l'interleukine IL6. Enfin, d'autres médicaments ciblent et inhibent des spécificités cellulaires comme l'abatacept et le rituximab. Le premier cible CD80 et CD86 impliqués dans la co-stimulation des cellules T tandis que le second cible CD20 exprimé par les cellules B [1, 46].

Il existe plusieurs façons de procéder au traitement des patients. La première, en cas d'activité faible ou modérée de la maladie, consiste à utiliser un des traitements de fond en monothérapie. Tandis que la seconde stratégie de traitement, si l'activité de la maladie est considérée comme modérée ou élevée, consiste à utiliser une combinaison de traitements [151]. Par exemple, l'utilisation de glucocorticoïdes et de methotrexate chez des patients atteints de PR précoce permet une rémission d'environ 25% des patients dans les six premiers mois. Cette combinaison est à ce

jour la plus efficace [1, 152, 153].

Malgré le nombre de traitements développé actuellement, 90% des patients traités par thérapie biologique recevront un traitement anti-TNF, traitement qui peut ne pas être adapté à ces patients [151]. Selon le *Dutch Rheumatoid Arthritis Monitoring* (DREAM), parmi les patients recevant un traitement anti-TNF, seulement 6% des patients parviennent à une rémission selon les critères ACR et EULAR [154]. Une explication à cela est que parmi les personnes traitées, il existe une partie des patients dits non-répondeurs, sur lesquels les traitements ne font pas ou peu d'effets [1]. En effet, 30 à 40% des patients ne répondent pas à un traitement anti-TNF et recevront par la suite d'autres traitements alternatifs [155]. Pour ces raisons, se développe l'utilisation de la médecine personnalisée, permettant d'adapter les traitements en fonctions des patients et aiderait à mieux gérer ces patients non-répondeurs [151].

Table 1.5: Les différents traitements de fond pour la PR. Adapté de Smolen et al. 2018 [1]

Traitements de fond	Médicaments
Synthétiques	
Conventionnels	Methotrexate, sulfasalazine, chloroquine et hydroxychloroquine
Ciblés	Baricitinib et Tofacitinib
Biothérapiques	
Originaux	Adalimumab, certolizumab, etanercept, golimumab, infliximab, tocilizumab, sarilumab, olokizumab, clazakizumab, abatacept et rituximab
Biosimilaires	Infliximab (CT-P13 et SB2) et etanercept (SB4)

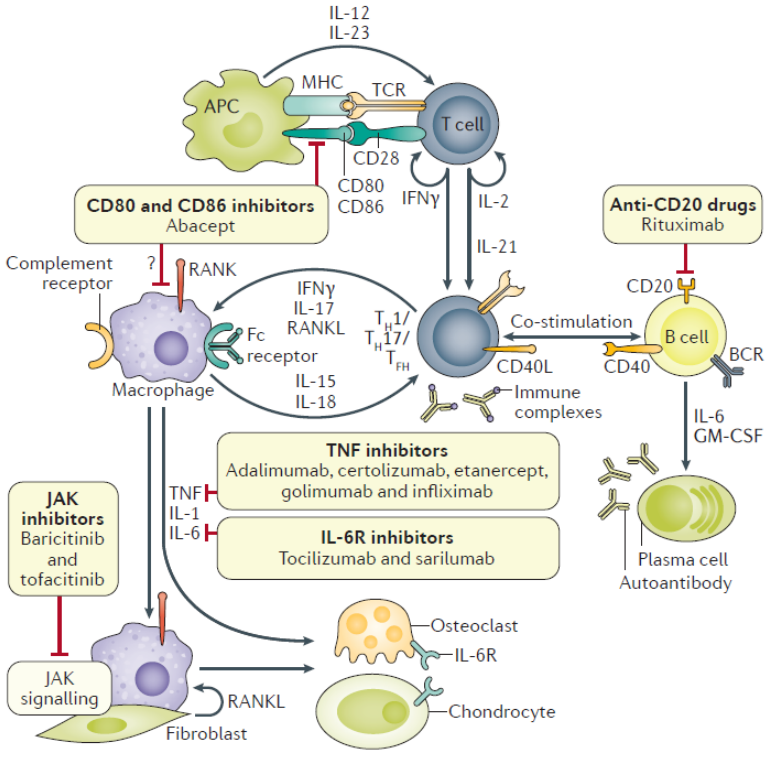

Figure 1.5: Cibles des médicaments utilisés en traitement de fond (DMARDS). Selon Smolen et al. 2018 [1].

Chapitre 2

La biologie des systèmes appliquée aux maladies complexes

Les approches par analyses génomiques et transcriptomiques ont beaucoup apporté à notre compréhension de la PR. Cependant, les facteurs génétiques identifiés par ces approches ne permettent pas d'expliquer entièrement l'étiologie de la PR. Cette dernière implique des processus biologiques complexes. Par exemple, une expression anormale de certains gènes peut entraîner la perturbation d'une ou plusieurs voies métaboliques, impliquant une réaction en chaîne, résultant en des phénotypes cellulaires différents. Les facteurs génétiques représenteraient alors uniquement une première couche de savoir de ce système complexe [156]. Ainsi, d'autres approches permettant l'identification de biomarqueurs, de mécanismes moléculaires et aussi de cibles médicamenteuses potentielles sont requises. Face à ce défi, la biologie des systèmes tente de réduire cette complexité à l'échelle d'un système afin d'en identifier les mécanismes clés. Une stratégie pour l'identification de mécanismes clés consiste à identifier et analyser les facteurs de transcription (TF) contenus dans un système. Ces derniers sont en effet responsables de la régulation de gènes cibles (TG), également appelés target genes et ont ainsi un rôle

central dans un système. Il existerait environ 1600 TF représentant 8% des gènes humains codant pour des protéines [157]. L'ensemble de ces facteurs, TF et TG, forment ce que l'on appelle un réseau de régulation biologique.

2.1 Inférence de réseaux

L'inférence de réseaux est une approche permettant de créer et prédire un réseau de régulation biologique réel, en utilisant des outils bio-informatiques dédiés. Cette dernière exploite des données expérimentales permettant la prédiction d'interactions régulatrices, en créant un réseau de gènes régulateurs (TF et TG) aussi appelé Gene Regulatory Networks (GRN) (illustré en Figure 2.2). Les approches d'inférences de réseaux peuvent utiliser jusqu'à six couches d'informations incluant des données de co-expression, de motifs de séquence, d'interaction ADN et protéine (ChIP, Immunoprécipitation de chromatine), d'orthologie, d'interaction protéine-protéine (PPI) et de la littérature [158]. Des études ont également démontré que l'utilisation d'approche d'inférence de réseaux par intégration de données multi-omiques (protéomique, génomique, transcriptomique et métabolomique) permet de mieux comprendre les mécanismes des maladies complexes [159–162]. Par exemple, l'utilisation du machine learning (apprentissage statistique en français) comme méthode d'inférence permet l'utilisation de ces données massives et complexes [163]. Des études ont également montré que l'intégration de connaissances préalables à des méthodologies basées sur l'inférence de données améliore la qualité et la pertinence biologique des résultats [164–166]. Enfin, l'ensemble de ces connaissances peut être regroupé sous forme de carte moléculaire maladie-spécifique.

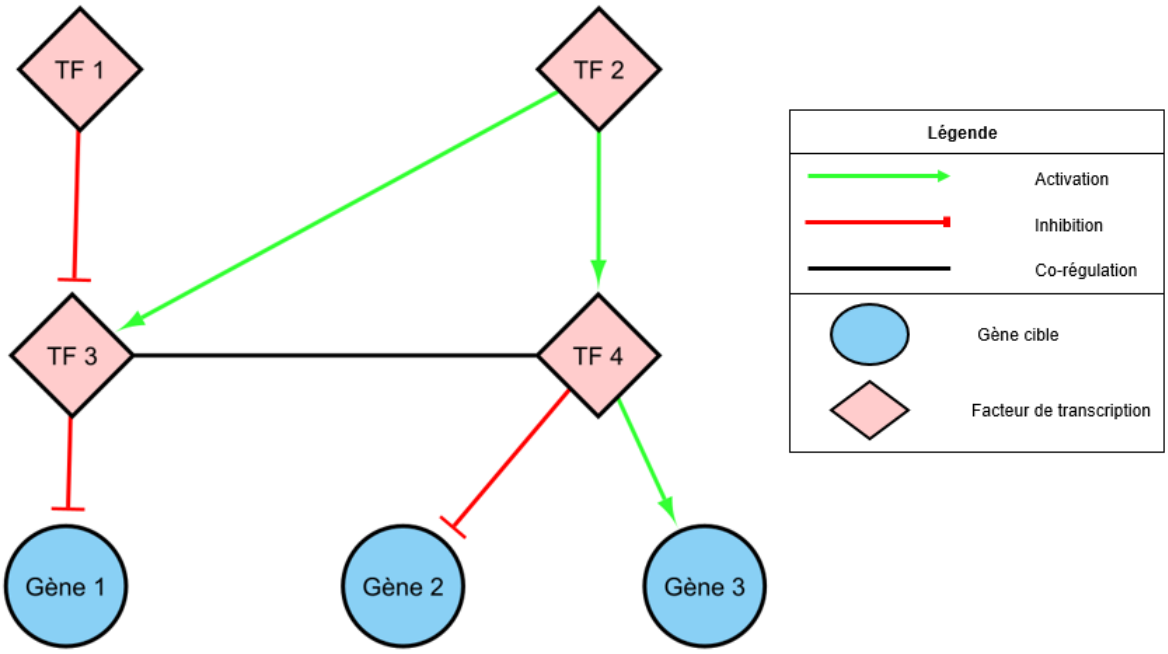

Figure 2.1: Exemple d'un réseau de gènes régulateurs (GRN)

2.2 Cartes moléculaires maladie-spécifique

Dans le but de mieux comprendre les mécanismes liés à une maladie, plusieurs communautés incluant The Cancer Cell Map Initiative (Krogan et al., 2015), l'Atlas of Cancer Signaling Networks (`http://acsn.curie.fr`) et le Disease Map Project (`http://diseasemaps.org`), développent des méthodologies et outils standards, permettant la création de cartes maladie-spécifique reliant à la fois les processus métaboliques, régulateurs et signalétiques des gènes et protéines entre eux [167]. Ainsi, ces communautés ont permis, par l'utilisation de la curation manuelle (étude approfondie de la littérature) ou bien de l'inférence de réseaux à l'aide d'outils bio-informatiques, de produire des cartes moléculaires maladie-spécifique. La maladie de Parkinson [168], l'asthme [167], l'athérosclérose [169] ainsi que la polyarthrite rhumatoïde [170, 171] ont fait l'objet de cartes moléculaires maladie-spécifique. La carte moléculaire de la PR, récemment publiée, est notamment le résultat de l'analyse de 353 publications scientifiques utilisant des données humaines,

manuellement vérifiées et incluant des informations sur des protéines (303), gènes (106), complexes (61), ARN (106) ainsi que des réactions (446) et des phénotypes (8) [170].

2.3 Modélisation computationnelle

Les réseaux de régulations et cartes moléculaires sont par nature statiques. Ils ne permettent pas de comprendre tous les processus biologiques, ni les effets fonctionnels dus à l'activation ou l'inhibition de certains facteurs du réseau. La modélisation computationnelle autorise cela par l'utilisation de réseaux dynamiques, permettant de réaliser des simulations et perturbations *in silico*. Pour cela, le modèle statique est transformé en un modèle mathématique et dynamique, en ajoutant des règles mathématiques, afin de caractériser chacune des relations entre chaque composant et en fixant les paramètres et les conditions initiales. L'ensemble de ces paramètres permet par la suite de réaliser une simulation menant à la prédiction d'un résultat.

2.3.1 Méthodes quantitatives et qualitatives

Bien qu'il existe de nombreuses méthodes de modélisations, elles sont généralement réparties en deux catégories : quantitative et qualitative. L'approche quantitative combine des paramètres cinétiques et mécanistiques ainsi que des équations différentielles, permettant au système dynamique d'être au plus proche possible de la réalité. À l'inverse, l'approche qualitative est basée sur peu de paramètres, fournissant une description plus globale du système [172].

Si la méthode quantitative est une méthode de prédiction beaucoup plus précise grâce à ses nombreux paramètres, son utilisation peut être difficile si l'on ne dispose pas de données cinétiques. Elle peut être applicable principalement à

de petits réseaux préalablement bien caractérisés [172]. L'approche qualitative quant à elle, utilise des modèles logiques tel que le modèle booléen. C'est une méthode de prédiction qui n'implique pas de paramètres cinétiques et de mécanismes moléculaires, en faisant abstraction des niveaux d'activités et de concentrations [173].

2.3.2 Modélisation booléenne

Un modèle booléen est une représentation qualitative d'un système (Figure 2.2.a). Chaque variable booléenne du système, aussi appelée fonction booléenne, est dénotée par 1 (ON) ou 0 (OFF), correspondant aux valeurs logiques binaires VRAI et FAUX (Figure 2.2.b). Les valeurs logiques ON et OFF représentent les états biologiques du modèle correspondant à ces variables binaires, indiquant si un gène est exprimé ou non exprimé ou encore un TF actif ou non actif. L'état de chaque fonction booléenne est défini par l'état logique de ses régulateurs dans le modèle, appelé loi booléenne et exprimé par les opérateurs logiques AND, OR et NOT (Figure 2.2.b). L'ensemble des combinaisons possibles pour chaque variable booléenne est représenté sous forme de table de vérité (Figure 2.2.c). Il est également possible de représenter l'évolution d'un signal dans un réseau dynamique sous la forme d'un graphique de transition d'état (Figure 2.2.d). Enfin, deux types d'attracteurs peuvent être observés dans un modèle booléen, les états stables et les attracteurs cycliques. On nomme un état stable, un état dans lequel le signal a transité et ne peut alors plus évoluer dans le réseau. Ainsi, un état stable dans un système biologique peut refléter un phénotype cellulaire (i.e différenciation et apoptose). Tandis qu'un attracteur cyclique est composé de deux ou plusieurs états à partir desquels le signal du réseau transite, et se répètent de manière ordonnée. Ainsi, un attracteur cyclique dans un système biologique peut quant à lui refléter un comportement d'homéostasie (i.e. cycle de Krebs).

Enfin, la modélisation booléenne a été utilisée avec succès afin de décrire la dynamique des cellules humaines en modélisant : les signaux de transductions et la régulation des gènes [174–179], ainsi que la dérégulation des gènes dans les maladies tels que le cancer [180, 181].

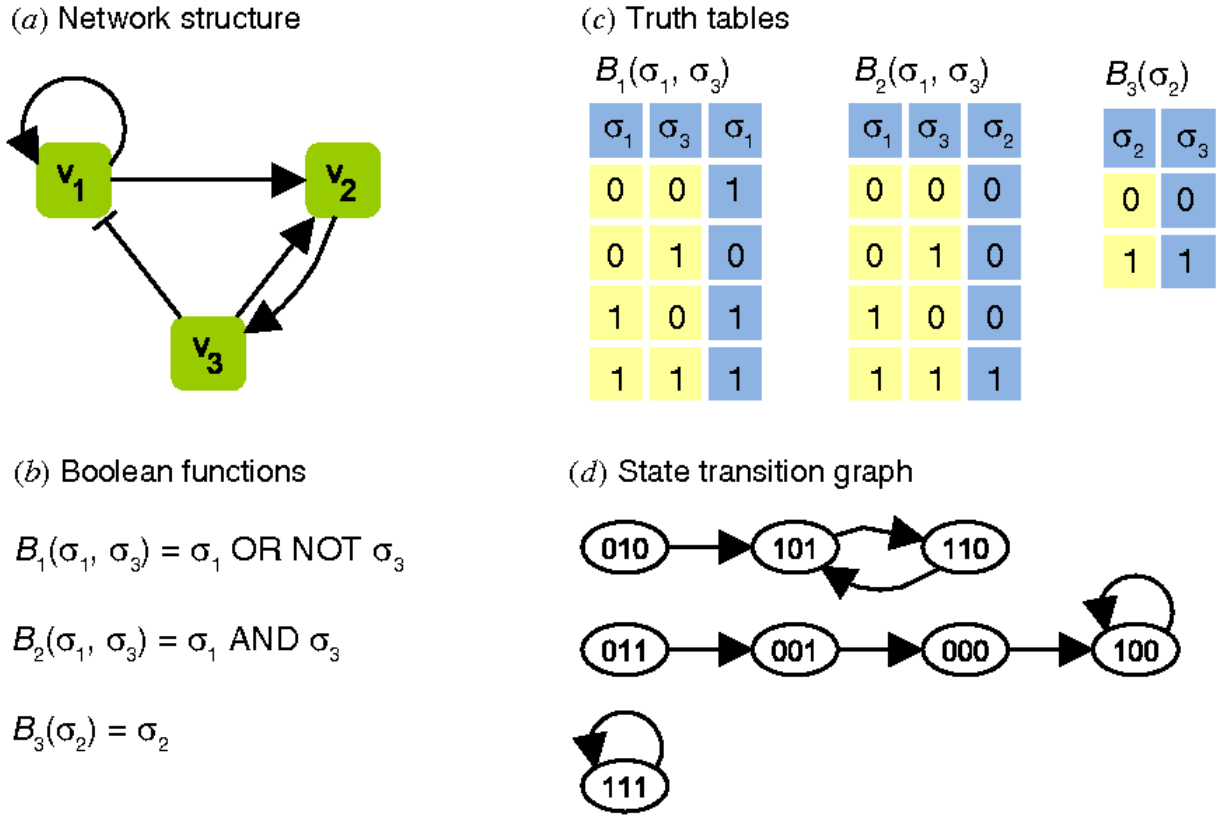

Figure 2.2: Modèle de réseau booléen. Adapté de Wang et al. [182]. Un modèle de réseau booléen est composé de variable booléenne $\{\sigma_1, \sigma_2, ..., \sigma_n\}$ ou chaque valeur est déterminée par une autre variable du réseau via un ensemble de fonction booléenne $B = \{B_1, B_2, ..., B_n\}$, chacune assignée à une variable. Ainsi la valeur de chaque variable σ_i est déterminée par les valeurs actives ou antérieures à ses propres régulateurs. (a) Réseau direct associé au modèle booléen. Chaque flèche pointue représente un effet positif tandis que chaque flèche avec une barre latérale représente un effet négatif. (b) Fonction booléenne du modèle. (c) Table de vérité des fonctions booléennes. (d) État de transition du modèle booléen construit sur les états synchrones successifs ou les états 100 et 111 sont des états stables et les états 101 et 110 sont des attracteurs cycliques

Chapitre 3

Objectifs

La PR est une maladie inflammatoire chronique ayant une prévalence comprise entre 0.25 et 1% dans la population mondiale adulte. La composante génétique impliquée dans la PR est estimée à 60%. Les facteurs génétiques découverts à ce jour incluent le gène majeur HLA-DRB1 ainsi que 100 loci de susceptibilité fréquents identifiés par des études pangénomiques (GWAS). Cependant, ces études ont été principalement focalisées sur la découverte de variants fréquents et ne permettent pas d'expliquer toute la composante génétique (seulement ~50%). Pour caractériser cette part d'héritabilité encore inconnue, une étude proposée par Manolio et al. 2010 [142] porte sur l'étude de variants rares de type SNV et SV, pouvant être mis plus facilement en évidence avec des données familiales. L'étude des variants rares est aujourd'hui possible grâce à l'utilisation de méthodes de séquençage à haut débit telles que le séquençage du génome entier (WGS) et le séquençage d'exome entier (WES), dont le coût a été largement réduit [183, 184].

1. Le premier objectif de cette thèse est d'identifier de nouveaux gènes associés à la PR en analysant des données de séquençage issues d'apparentés atteints et non atteints de familles PR multiplexes. Dans un premier temps, nous étudierons des outils de détection de CNVs dans des données WES simulées,

dont les outils les plus performants seront ensuite utilisés afin de détecter des CNVs rares dans des données WES de familles multiplexes de PR. Dans un second temps, nous étudierons les variants rares issus de données WGS à partir de familles multiplexes de PR. L'ensemble de ces analyses seront mises en place sous forme de pipeline réutilisable.

Les variants génomiques représentent uniquement une première strate d'informations et ne permettent pas de comprendre tous les mécanismes de la PR. Face à la complexité d'une telle maladie, l'utilisation simultanée de plusieurs strates d'informations est essentielle afin d'en mieux comprendre ses mécanismes. La biologie des systèmes permet de réduire la complexité de l'utilisation de ces couches à l'échelle d'un système. Cette approche utilise l'inférence de réseaux par intégration de données multi-omiques [159–162]. D'autre part, l'inférence de réseaux a déjà été utilisé afin de comprendre les mécanismes clés dans des maladies complexes [159–162]. Un réseau, peut être transformé en un réseau booléen en ajoutant des règles logiques, afin d'étudier la régulation des gènes. Ce type d'étude, se faisant par simulation, a par ailleurs montré son efficacité pour décrire des signaux de transductions et la régulation des gènes des cellules humaines [174–179] mais aussi pour mieux comprendre des maladies tels que le cancer [180, 181].

2. Le second objectif de cette thèse est l'étude de mécanismes causaux pouvant lier des perturbations extracellulaires avec des voies de signalisation et l'expression génique dans la PR, en utilisant l'inférence et l'analyse dynamique d'un réseau global et spécifique de la maladie.

Chapitre 4

Caractérisation de CNVs par étude de séquençage d'exomes

Les CNVs sont des variations structurales (SVs) du génome humain. Ils sont définis comme une variation du nombre de copie d'une région génomique d'ADN supérieure à 50pb par rapport à un génome de référence. Les CNVs peuvent être retrouvés sous forme de duplication ou délétion. Environ 4.8 à 9.5% du génome d'un individu est constitué de CNVs [116]. La plupart des CNVs présents dans le génome humain n'ont pas d'effets et résultent de l'évolution et de l'adaptation des espèces [116]. Environ 100 gènes non-essentiels peuvent être totalement délétés par les CNVs sans conséquence phénotypique particulière [116]. Cependant, certains CNVs peuvent altérer la fonction de gènes essentiels, et ainsi être impliqués dans des maladies complexes [185].

La taille des CNVs peut être un indicateur d'un potentiel effet délétère. En effet, lorsque lorsqu'un CNV est visible à l'échelle du caryotype, celui-ci est pratiquement toujours associé à des conséquences phénotypiques [186]. C'est par exemple le cas d'une délétion dans la région 13q, qui a été trouvée délétée chez 14 patients (délétion allant de 4.2 à 75.7 Mb), entraînant des phénotypes incluant des retards

mentaux, des anomalies aux niveaux des yeux, des mains et des pieds [187]. Environ 8% de la population est porteuse de CNVs supérieurs à 500kb contre 25% des patients ayant des déficiences intellectuelles [188–190]. Cependant, les grands CNVs peuvent être également bénins. C'est par exemple le cas d'une délétion de 14.5Mb (13q21.1-13q21.33, représentant 12.5% du chromosome 13) incluant 18 gènes sans conséquence caractérisée dans une famille sur 3 générations [191]. Toutefois, les CNVs de petites tailles peuvent également être la cause de maladie. Par exemple, un retard mental sévère est causé par une délétion de 140 kb sur trois exons du gène MEF2C [192].

Les CNVs sont présents dans la population à des fréquences différentes. Un CNV dont la fréquence est inférieure à 1% est considéré comme rare. À l'inverse, un CNV dont la fréquence est supérieure à 1% est considéré comme commun [193]. Des CNVs rares et communs ont été identifiés comme associés à des maladies humaines. Par exemple, un CNV commun dans le gène *CCL3L1* et un CNV rare dans le gène ITGB8 ont été respectivement retrouvés associés dans la PR et le cancer des ovaires [194, 195].

La détection des CNVs a été facilitée par les nouvelles technologies de séquençage à haut débit (next-generation sequencing, NGS). Avant celles-ci, il n'était pas possible de détecter un CNV d'une taille inférieure à 50kb [196, 197]. Cette détection est maintenant beaucoup plus précise, allant jusqu'à 50pb [198, 199]. Ces technologies NGS incluent le séquençage du génome entier (whole-genome sequencing, WGS) et de l'exome entier (whole-exome sequencing, WES). Ces deux méthodes ont leurs propres avantages et inconvénients. Le WGS couvre le génome entier permettant d'identifier la totalité d'un CNV, cependant la méthode étant coûteuse celle-ci ne permet pas d'obtenir une grande profondeur de séquençage. À l'inverse le WES ne couvre que les régions codantes (1% du génome) ne permettant pas la détection des positions exactes des CNVs, mais il est moins coûteux et

permet alors une plus grande profondeur de séquençage.

Dans ce chapitre, nous allons dans un premier temps étudier la performance de 6 outils de détection de CNVs à partir de données WES simulées. Cette première étude permettra alors de cibler les outils les plus adaptés afin d'identifier dans un second temps des CNVs rares associés à la PR à partir de données WES d'apparentés atteints et non atteints de familles PR multiplexes.

4.1 Analyse des performances d'outils de détection de CNVs à partir de données WES

Les outils permettant la détection de CNV sont basés sur une des quatre principales méthodes d'identification suivantes : *read-pair*, *split-read*, *assembly* et *read-depth* [200, 201]. Il existe également une cinquième méthode, qui est une combinaison des méthodes *read-pair* et *read-depth*, appelée approche combinatoire. La méthode *read-pair* (Figure 4.1.A) utilise les lectures alignées dites discordantes. Celles-ci sont identifiées en comparant la distance entre les deux extrémités d'une lecture appariée à une taille moyenne attendue. La méthode *split-read* (Figure 4.1.B) utilise des lectures alignées incomplètes de chaque lecture appariée afin d'identifier de petits CNVs. La méthode *read-depth* (Figure 4.1.C) est basée sur le nombre de lectures alignées dans chaque région génomique, comparé à un nombre de lectures alignées attendu. Enfin, la méthode *assembly* (Figure 4.1.D) assemble les lectures en contigs. Ceux-ci sont alignés au génome de référence et permettent d'identifier les régions insérées ou délétées. Le séquençage WES ne permettant pas une couverture entière du génome, cela limite les méthodes *read-pair*, *split-read* et *assembly* (Figure 4.1 A-B et D, respectivement). Cependant, le WES offre une grande profondeur de séquençage, ce qui est idéal pour détecter les CNVs avec la méthode *read-depth* (Figure 4.1 C) [201].

Ainsi, la plupart des outils développés pour détecter des CNVs en WES sont basés sur cette méthode. Parmi eux, nous retrouvons des outils tels que ExomeDepth, XHMM, DECoN, panelcn.MOPS, CANOES, CoNIFER et EXCAVATOR, qui ont été développés afin de détecter des CNVs rares [202–208]. Quelques outils ont aussi été développés afin d'identifier à la fois des CNVs rares et communs, comme par exemple CODEX2 et CLAMMS [209, 210]. Le peu d'outils disponibles pour identifier des CNVs communs est notamment dû à la difficulté de

leur identification. En effet, l'identification d'un CNV dans une région génomique donnée est réalisée en comparant la profondeur d'une région pour un individu à une profondeur de référence établie à partir de l'ensemble des individus de l'échantillon. Par exemple, si un CNV fréquent est présent chez 9/10 individus d'un jeu de données, alors celui-ci affectera considérablement le calcul de la référence, ce qui biaisera les résultats. Plusieurs études se sont penchées sur les performances d'outils de détection de CNVs utilisant la méthode *read-depth* afin d'identifier des CNVs rares [211–214], présentés en Table 4.1. Ces études montrent des sensibilités[1] et précisions[2] élevées pour l'ensemble des outils (>70%) à l'exception de XHMM (~50%). Cependant, ces études ont été réalisées dans une configuration particulière où un CNV est inséré chez un seul individu (tous les CNVs avaient la même fréquence dans l'échantillon). Si ces études analysent les caractéristiques des CNVs en faisant varier la taille et le type (délétion et duplication), aucune ne considère les performances des outils selon la fréquence des CNVs (rares et communs). Il est pourtant important de connaître les performances de ces outils lorsque plusieurs individus possèdent un même CNV. Le premier objectif de cette étude est l'analyse de la performance d'outils de détection de CNVs utilisant la méthode *read-depth* à partir d'un jeu de données WES simulés contenant des CNVs incluant différentes tailles, fréquences et types. Par ailleurs, les performances de ces outils ont été étudiées uniquement par rapport à la mise en évidence d'une région génomique incluant un CNV. Or, une région peut être mise en évidence sans être identifiée chez les bons individus, ce qui peut ensuite avoir des conséquences sur les résultats des études d'association. Ainsi, les performances des outils ont aussi été évaluées par rapport à l'identification des CNVs chez les individus.

[1] $\text{Sensibilité} = \dfrac{\text{vrai positif}}{(\text{vrai positif} + \text{faux positif})}$

[2] $\text{Précision} = \dfrac{\text{vrai positif}}{(\text{vrai positif} + \text{faux négatif})}$

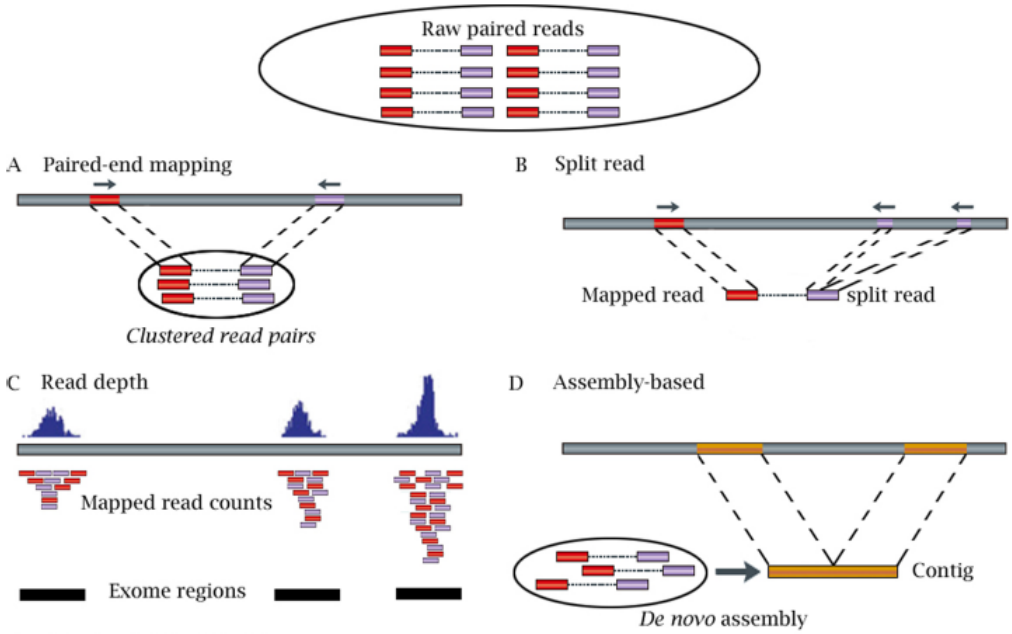

Figure 4.1: Approche de détection de CNVs à partir de séquençage NGS en lecture courte. Adapté de Zhao et al. 2013 [201]. A. La méthode *read-pair* utilise les lectures alignées discordantes. Celles-ci sont identifiées en comparant la distance entre les deux extrémités d'une lecture appariée à une taille moyenne attendue B. La méthode *split-read* utilise des lectures alignées incomplètes de chaque lecture appariée afin d'identifier de petits CNVs. C. La méthode *read-depth* est basée sur le calcul du nombre de lectures alignées dans chaque région génomiquec comparé à un nombre de lectures alignées attendu. D. La méthode *assembly* détecte les CNVs en assemblant les lectures en contigs puis ceux-ci sont alignés au génome de référence et permettent d'identifier les régions insérées ou délétées.

Table 4.1: Performances de six outils de détection de CNV selon la littérature. La sensibilité est indiquée en noir et la précision est indiquée en bleu.

Taille de l'échantillon (nb de sujets)	Mesure	ExomeDepth	CODEX	CODEX2	XHMM	panelcn.MOPS	DECoN	CLAMMS
Sadedin et al. [213]								
38	Sensibilité	0.90	0.96		0.48			
Moreno-Cabrera et al. [215]								
96	Sensibilité	0.96		0.93		0.96	0.97	
	Précision	0.80		0.78		0.45	0.73	
161	Sensibilité	0.98		0.44		0.94	0.98	
	Précision	0.88		0.69		0.86	0.88	
130	Sensibilité	0.84		0.93		0.67	0.93	
	Précision	0.96		0.80		0.80	0.89	
108	Sensibilité	0.79		0.86		0.56	0.89	
	Précision	0.94		0.80		0.76	0.85	
Hong et al. [214]								
167	Précision				0.46			
48	Précision				0.43			
116	Précision				0.50			
54	Précision				0.04			
Roca et al. [212]								
220	Sensibilité	1	0.73				1	0.80
	Précision	~1	0.69				0.97	0
220	Sensibilité	1	0.71				1	0.72
	Précision	~1	0.75				0.99	0.89
220	Sensibilité	1	0.71				1	0.73
	Précision	~1	0.78				~1	0.89

4.1.1 Matériel et Méthodes

4.1.1.1 Simulation d'un jeu de données WES incluant des CNVs

Le jeu de données utilisé dans cette analyse a été obtenu à partir de données WES simulées pour un total de 250 individus. Parmi ces individus, 200 portaient au moins un CNV, tandis que 50 n'en avaient aucun. L'intégralité des étapes évoquées ci-après a été réalisée en utilisant Snakemake, un gestionnaire de workflow [216].

Sélection et intégration de CNVs

La sélection des CNVs à simuler a été effectuée à partir de deux bases de données recensant des CNVs humains : *Database of Genomic Variants* (DGV) et *DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources* (DECIPHER) [217, 218]. DGV est une base de données publique comprenant un catalogue de CNVs identifiés à partir de génomes d'individus ne présentant pas de maladie particulière et appartenant à différentes populations. Celle-ci comprend 983 845 CNVs identifiés chez plus de 54 980 individus. DECIPHER est une base de données publique répertoriant des CNVs identifiés chez des sujets atteints de maladies rares. Cette base de données inclut 39 901 sujets, 172 459 phénotypes observés chez ces patients et 41 169 CNVs.

Au total, sept chromosomes (1,7,11,16,17,19 et 22) ont été sélectionnés afin d'y introduire 1 607 CNVs non-chevauchants extraits des deux bases de données (152 provenant de DECIPHER et 1455 depuis DGV). Les 1607 CNVs sélectionnés correspondent au maximum de CNVs possibles non chevauchants sur les sept chromosomes. De plus, le choix de ne pas inclure tous les chromosomes a été fait afin de réduire le temps de simulation. Les sept chromosomes ont été extraits à partir du génome de référence hg19, puis dupliqués 250 fois afin de correspondre au nombre d'individu de l'étude. Les 1607 CNVs ont alors été attribués aléatoirement

chez 200 individus, en respectant la fréquence de chaque CNV dans les bases de données. RSVSim [219] a été utilisé afin d'insérer chaque CNV à sa position génomique et en respectant son type (duplication ou délétion) dans les fichiers FASTA de chaque individu. Pour les 1607 régions incluant des CNVs, un total de 20595 CNVs a été introduit chez ces 200 individus. La distribution des CNVs en fonction de leurs caractéristiques est présentée en Table 4.2.

Table 4.2: Distribution des CNVs selon leurs tailles (a), fréquences (b) et types (c).

(a)

Taille des CNVs	**<5kb**	**5-150kb**	**>150kb**	**Total**
N (%)	550 (34.2%)	523 (32.5%)	534 (33.2%)	1607

(b)

Fréquences des CNVs	**<1%**	**1-5%**	**>5-95%**	**Total**
N (%)	592 (36.8%)	412 (25.6%)	603 (37.5%)	1607

(c)

Type des CNVs	**Délétion**	**Duplication**	**Duplication et délétion**	**Total**
N (%)	755 (47%)	773 (48.1%)	79 (4.9%)	1607

Simulation de données de séquençage exoniques

Afin de simuler les séquences exoniques pour les 250 individus, il a été utilisé Wessim [220]. Wessim est un programme permettant de générer un séquençage d'exome synthétique. Cet outil réplique des techniques de capture d'exome conventionnelles telles que Agilent's SureSelect et NimbleGen's SeqCap afin de générer des fragments d'ADN dans des régions génomiques cibles. Wessim réalise deux

étapes, détaillées comme suit :

1. Génération de fragments d'ADN

Un fragment d'ADN est défini par sa longueur $L(f)$ et sa séquence $S(f)$ où $f = c_f, s_f, e_f$ avec c_f s_f et e_f, respectivement le chromosome, la position de début et de fin du fragment d'ADN. Wessim réalise la génération de fragments selon deux approches : A. Approche par cible "idéale" B. Approche par hybridation de sonde. La première approche est une approche basée sur des régions exoniques cibles idéales, où les fragments sont automatiquement créés à partir de la taille de la région cible. La seconde approche, que nous avons utilisée car plus réaliste, implémente la capture par hybridation de sonde afin de créer des fragments d'ADN. La séquence de ces sondes sont par exemple gratuitement disponibles sur le site SureDesign, propriété d'Agilent (`https://earray.chem.agilent.com/suredesign/`).
Une sonde oligonucléotidique p est définie par sa séquence $S(p)$, où chaque région p possiblement hybridable est notée $h_i^p \epsilon H^p$ et où chaque séquence $S(h_i^p)$ correspond à $S(p)$ ayant un score élevé (e.g. identité de séquence $\geq 95\%$). La probabilité de sélectionner h_i^p est inversement proportionnelle au nombre de mésappariement entre $S(h_i^p)$ et $S(p)$. Afin de générer un fragment, Wessim choisit aléatoirement une sonde p_x et sélectionne aléatoirement une région hybridable h depuis H^{p_x}. Un fragment f peut alors être généré uniquement si une certaine fraction de f chevauche la région hybridable sélectionnée, ce qui est défini par Wessim comme le rapport de chevauchement minimum b_0. La probabilité de générer un fragment f peut-être calculé comme suit:

$$P(f) = \max\left(P(h|p_x)P(p_x)\frac{b - b_0}{1 - b_0}, 0\right),$$

où $P(p_x)$ est la probabilité de sélectionner une sonde p_x, $P(h|p_x)$ est la probabilité conditionnelle de sélectionner h de H^{p_x} étant donné p_x; et b est la fraction de la

région hybridable de f. Wessim reproduit les biais liés aux fragments tels que le taux de GC et la taille des fragments [221].

Afin de simuler des CNVs, nous avons biaisé la probabilité p_x (sélection aléatoire d'une sonde) dans les régions contenant des CNVs. Ainsi, les sondes sont plus fréquemment choisies si la région contient un CNV de type duplication et inversement, moins sélectionnées si la région contient un CNV de type délétion.

2. Séquençage des fragments

Le séquençage synthétique des fragments est réalisé à partir des fragments générés à l'étape précédente à l'aide d'un simulateur avancé de données NGS, GemSim [222], qui utilise un modèle empirique d'erreur répliquant les erreurs des plate-formes de séquençage [222].

Dans cette étude, nous avons utilisé Wessim et son approche par sonde en utilisant le kit Agilent 4 (`https://earray.chem.agilent.com/suredesign/`) comprenant les régions exoniques et les sondes associées, ainsi qu'un modèle d'erreur reproduisant le séquençage Illimunia HiSeq2000. Wessim a été paramétré afin d'obtenir une couverture de séquençage de 100X ainsi que des lectures de 100bp. Les séquences exoniques des 250 individus obtenues au format fastq ont ensuite été alignées sur le génome de référence hg19 en utilisant BWA [223] puis converties sous format BAM. Ce format permet d'être directement exploité par les outils de détection de CNVs.

4.1.1.2 Détection des CNVs exoniques

Pour identifier les CNVs, nous avons utilisé six outils de détection de CNVs référencés dans la Table 4.3, tous basés sur la méthode *read-depth*. Lors de la sélection de ces outils, nous avons opté pour une stratégie diversifiée en incluant : des outils conçus afin de détecter des CNVs communs (CODEX2 et CLAMMS), des outils ayant globalement de bonnes performances (DECoN, ExomeDepth et

panelcn.MOPS) et une utilisation élevée (XHMM) selon la littérature. La méthode de détection *read-depth* ainsi que les six outils sont brièvement détaillés ci-dessous.

La méthode de détection *read-depth*

La stratégie *read-depth* consiste à poser l'hypothèse selon laquelle la couverture de séquençage dans une région donnée est corrélée au nombre de copie dans cette région. Cette stratégie compare la distribution de la profondeur des lectures d'un individu à la distribution générale de l'échantillon via des modèles mathématiques, stratégie d'autant plus efficace si l'on utilise des individus de référence. Ainsi, une perte de copie résulte en une profondeur plus faible que la profondeur moyenne dans un échantillon de référence [224], comme c'est le cas présenté dans la Figure 4.2 tirée de G. Povysil et al 2017 [205].

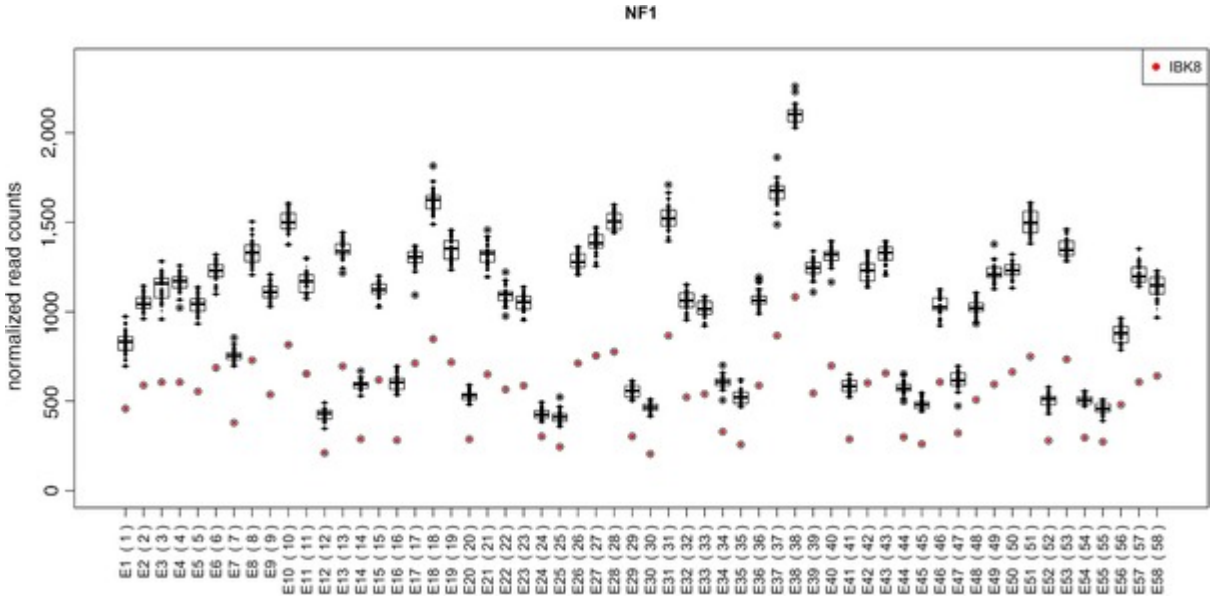

Figure 4.2: Représentation par boîte à moustaches d'une délétion à partir du nombre de lectures par exon du gène NF1 Les boîtes à moustaches représentent le nombre de lectures normalisées des individus testés et témoins. Le nombre de lectures des individus témoins sont représentés par des points noirs, tandis que le nombre de lectures des individus testés sont représentés par des points rouges. La délétion est visible du fait que les points rouges sont distincts et sous les boîtes à moustaches.

La méthode *read-depth* est composée de 4 étapes afin d'identifier des CNVs : 1. l'alignement, 2. la normalisation de la profondeur par exon, 3. l'estimation du nombre de copie par exon et 4. la segmentation. Lors de l'alignement, les lectures sont alignées sur le génome de référence, et la profondeur de lecture est calculée en fonction du nombre de lectures alignées dans une région. La deuxième étape consiste à normaliser la profondeur de lecture et corriger les biais de séquençage tels que le taux de GC, les régions répétées etc. La troisième étape permet d'estimer le nombre de copie afin de déterminer s'il s'agit d'un gain ou une perte. Enfin, la dernière étape consiste à assembler les régions chevauchantes ayant un nombre de copies égal pour identifier les régions contenant un CNV.

La particularité de chaque outil est présentée dans les paragraphes suivants.

CLAMMS corrige le biais du taux de GC en normalisant chaque individu indépendamment par l'utilisation de la profondeur de lecture moyenne des individus. L'outil propose à l'utilisateur de sélectionner un panel d'individus de référence et recherche alors les individus les plus proches pour chaque individu analysé en utilisant sept paramètres de contrôle qualité de séquençage via l'outil Picard [225]. Ensuite, chacune de ces métriques est intégrée dans un arbre k-D, à partir duquel CLAMMS entraîne son modèle via la méthode des k plus proches voisins. Enfin, la détection et l'assemblage des CNVs sont réalisés pour chaque individu en utilisant l'algorithme Hidden Markov Model (HMM).

CODEX2 effectue une série de contrôles qualité des données basée sur la mappabilité par rapport à un génome de référence, la taille des exons et une couverture minimale. Afin de réduire le biais des données WES, CODEX2 normalise les données en comptant le nombre total de lectures, le taux de GC, la qualité d'alignement et la capture d'exon. CODEX2 permet d'utiliser des individus de référence, afin de détecter des CNVs rares et communs. Finalement, l'outil détecte et assemble les

CNVs par segmentation à l'aide d'une vraisemblance de Poisson.

ExomeDepth utilise un modèle bêta-binomial afin de normaliser la profondeur de lecture des individus, basé sur le taux de GC de chaque exon et un paramètre de surdispersion (basé sur le nombre total de lectures). ExomeDepth sélectionne pour chaque individu testé l'individu de l'échantillon ayant la plus forte corrélation avec l'individu testé et l'utilise comme référence. Pour les analyses cas/témoins, il est possible de restreindre ExomeDepth à choisir une référence parmi un set d'individus témoins. Pour chaque exon, une vraisemblance est calculée pour chacun des trois types (délétion, diploïde et duplication). Le type conservé est celui ayant la meilleure vraisemblance. Enfin, pour détecter et assembler les CNVs à partir des exons, l'outil réalise une segmentation en utilisant l'algorithme HMM.

DECoN est basé sur ExomeDepth 1.0.0 avec quelques optimisations. Celui-ci a un mode de fonctionnement différent, car incluant des modifications concernant l'algorithme HMM permettant de détecter les CNVs inclus aux extrémités de chaque chromosome. Cependant, ces modifications ont été implémentées depuis dans la version 1.1.0 d'ExomeDepth. Le mode de fonctionnement de DECoN permet à l'utilisateur de ne pas avoir à utiliser R mais plutôt des lignes de commande, ce qui facilite l'utilisation de l'outil.

Panelcn.MOPS calcule la profondeur de lecture en utilisant ExomeCopy [226]. L'outil réalise deux contrôles qualité, le premier en enlevant les régions ayant une médiane du nombre de lectures trop faibles (<30) chez tous les individus, le second en taguant chaque région comme "low quality" si la région contient des variations très importantes en terme de nombre de lectures entre l'individu testé et les individus de référence. Une normalisation du nombre de lectures est également effectuée en mettant à l'échelle chaque échantillon avec une normalisation

au troisième quartile. Enfin, l'outil utilise un modèle de mélange de Poisson à partir de l'outil cn.MOPS [227] afin de procéder à la détection et l'assemblage des CNVs.

XHMM réalise plusieurs contrôles qualité basés sur le taux de GC, la faible complexité des régions, ainsi que la couverture des régions cibles. Les individus et régions ne remplissant pas ces critères sont jugés aberrants et sont retirés. Ensuite, l'outil utilise une Analyse en Composante Principale (ACP) pour normaliser la profondeur de séquençage, transformant les valeurs de profondeur en Z score pour chaque individu. Ces valeurs sont ensuite utilisées comme entrée pour l'étape de segmentation en utilisant l'algorithme HMM permettant de classifier chaque région en 3 possibilités : diploïde, délétion ou duplication.

Nous avons conservé les paramètres par défaut pour l'ensemble des outils pour la détection des CNVs, excepté pour les outils panelcn.MOPS et XHMM, où il était requis de changer les paramètres, pour que les performances soient comparables à celles des autres outils. Le temps de calcul de chaque outil a été mesuré, en utilisant la fonction "benchmark" du gestionnaire de pipeline Snakemake [216]. Pour chaque outil, nous avons utilisé les 50 individus sans CNV comme références, à l'exception de XHMM qui ne le permettait pas. Afin de considérer un CNV comme détecté par un outil, nous avons appliqué un seuil adapté spécifiquement à chaque outil : Bayes factor > 40 pour DECoN et ExomeDepth, lratio > 40 pour CODEX2, Q_EXACT > 0 pour CLAMMS, Q_SOME > 60 pour XHMM, et non tagué comme "low quality" pour panelcn.MOPS. Tous les résultats des outils de détection de CNVs ont été analysés avec R.

Table 4.3: Outils de détection de CNVs basés sur la méthode *read-depth*

Outil	CNV rare	CNV commun	Langage\interpréteur	Année	Version
CLAMMS [209]	X	X	bash	2016	1.1
CODEX2 [210]	X	X	R	2018	1.3.0
DECoN [202]	X		bash	2016	1.0.2
ExomeDepth [204]	X		R	2012	1.1.15
panelcn.MOPS [205]	X		R	2017	1.8.0
XHMM [203]	X		bash	2012	1.0

4.1.1.3 Analyse des performances d'outils de detection de CNVs

Dans cette étude, nous avons simulé des CNVs ayant des fréquences, types (délétion et duplication) et tailles différentes. La performance de chaque outil de détection a été évaluée par rapport à : 1. La détection de la région génomique d'un CNV; 2. la détection d'un CNV chez les individus. L'analyse par région génomique considère qu'un CNV est vrai positif (TP) lorsqu'il est retrouvé au moins une fois dans l'échantillon sans prendre en considération son type, ni l'individu dans lequel il a été réellement simulé. Tandis que l'analyse par individu considère un CNV comme TP uniquement si celui-ci est retrouvé chez l'individu où il a été simulé et avec le bon type (Voir Table 4.4 pour plus de détails). Enfin, nous avons calculé les performances des outils en utilisant la précision et la sensibilité définies tels que :

$Précision = TP/(TP + FP)$

$Sensibilité = TP/(TP + FN)$

où TP = Vrai positif ; FP = Faux positif et FN = Faux négatif

La précision permet ainsi d'évaluer si un outil détecte beaucoup de faux positifs (plus la précision diminue, plus l'outil identifie des FP). La sensibilité évalue la performance d'un outil à détecter correctement un CNV simulé (plus la sensibilité augmente, moins il y a de CNV non identifiés).

Table 4.4: Description des métriques de détetion de CNVs par région génomique et individu.

Métrique	Région	Individu
Vrai positif (TP)	Région simulée détectée	Région simulée détectée, CNV détécté chez le bon individu et bon type
Faux positif (FP)	Région non-simulée détectée	Région simulée détectée, CNV détécté chez un individu sans CNV ou de mauvais type
Faux négatif (FN)	Région simulée non-detectée	Région simulée détectée, CNV non détecté chez un individu avec CNV

4.1.1.4 Jeu de données 1000 Genomes

Dans cette étude, nous avons également étudié les performances des outils de détection sur des données réelles. Pour cela, deux jeux de données provenant du projet 1000 Genomes [228] ont été utilisés : The Washington University Genome Sequencing Center (WUGSC) et le Baylor College of Medicine (BCM) incluant respectivement 56 et 138 individus. Les mêmes outils et pipelines ont été appliqués sur les fichiers BAM extraits depuis le 1000 Genomes. Le 1000 Genomes met à disposition une liste de CNVs vérifiés et identifiés à partir de l'ensemble des individus séquencés dans leur base de données. Nous avons extrait les CNVs correspondant aux individus des deux jeux de données respectifs, et considéré comme CNV de référence les CNVs de cette liste contenue dans les régions exoniques du kit Agilent 4, régions précédemment utilisées dans les simulations. Ces jeux de données n'incluant pas d'individu sans CNV, aucun individu de référence n'a été inclus dans cette analyse.

4.1.2 Résultats

4.1.2.1 Performances globales des outils de détection de CNVs sur les données simulées

Les performances (sensibilité et précision) des outils de détection sont présentées dans la Figure 4.3.

À l'échelle de la région génomique (Figure 4.3.A), le pourcentage de vrai positif détecté variait en fonction de chaque outil. Nous avons observé une forte sensibilité pour les outils panelcn.MOPS (81.5%) et CODEX2 (69.4%), tandis que XHMM (46.4%), CLAMMS(46%), ExomeDepth (42.4%) et DECoN (41.6%) ont montré des sensibilités similaires mais plus faibles. Les outils ont par ailleurs détecté peu de régions fausses positives, avec une précision variant de 98.7% (CLAMMS) jusqu'à ~100% (XHMM, DECoN, ExomeDepth, CODEX2 et panelcn.MOPS).

À l'échelle de l'individu (Figure 4.3.B), la sensibilité et la précision, calculées pour les régions vraies positives, étaient diminuées par rapport aux performances à l'échelle de la région. L'outil le plus sensible était CODEX2 (67.2%), tandis que l'ensemble des autres outils avait une sensibilité inférieure à 55% (de 54.8% pour DECoN à 16.8% pour XHMM). Les outils CLAMMS (99.7%), CODEX2 (98.6%), DECoN (98.6%), ExomeDepth (95.8%), et panelcn.MOPS (95%) ont montré une très bonne précision, même si légèrement inférieure à la précision observée pour la détection des régions.

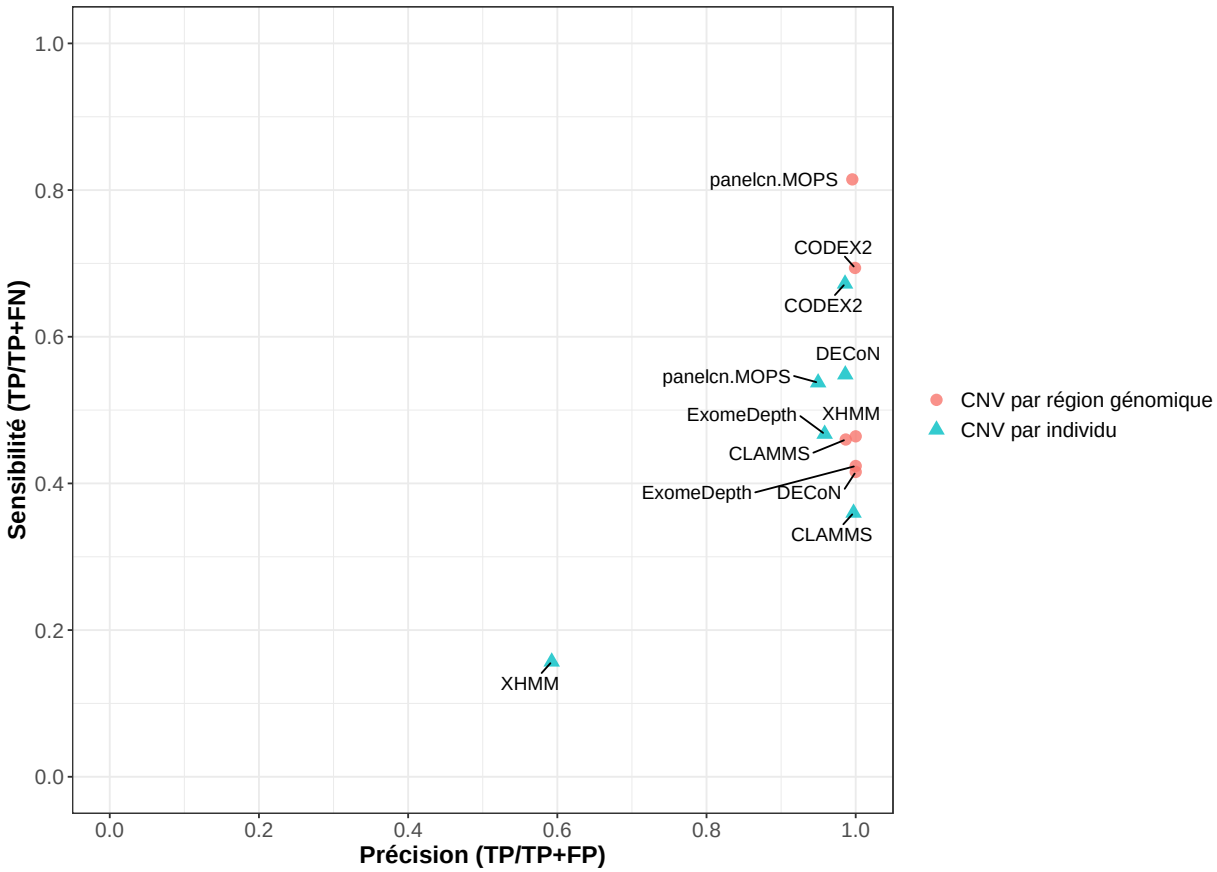

Figure 4.3: Performances (sensibilité et précision) des six outils de détection de CNV à partir de données WES simulées, à l'échelle de la région génomique (point rouge) et de l'individu (triangle bleu)

4.1.2.2 Performances selon la taille des CNVs

Les performances des outils de détection selon la taille des CNVs (<5kb, 5-150kb et >150kb) sont présentées dans la Figure 4.4.

Au niveau de la région génomique (Figure 4.4.A), seule la sensibilité est présentée puisque nous avons observé une précision élevée, toujours supérieure à 95% pour l'ensemble des outils. En revanche, la sensibilité des outils était dépendante de la taille des CNVs. En effet, nous avons observé une sensibilité croissante lorsque la taille des CNVs augmentait. L'outil panelcn.MOPS était le moins influencé par la taille des CNVs, avec une sensibilité ~80% quelque soit la taille des CNVs.

En considérant les grands CNVs (>150kb), CODEX2 était l'outil le plus sensible (90.6%), tandis que pour les CNVs moyens et petits, panelcn.MOPS a montré les meilleures sensibilités (77.6 et 72.9% respectivement). DECoN et ExomeDepth étaient les moins performants afin d'identifier des CNVs avec des tailles $\leq 150kb$ (les sensibilités chutant à 9.8 et 11.6% respectivement pour les CNVs $\leq 5kb$), tandis que CLAMMS était le moins performant pour identifier des CNVs $> 150kb$ (54.5%).

À l'échelle des individus (Figure 4.4.B), CLAMMS, CODEX2, DECoN, ExomeDepth et panelcn.MOPS ont détecté peu de faux positifs puisqu'ils avaient une précision élevée (> 97%), indépendamment de la taille des CNVs. XHMM identifiait le plus de faux positifs (précision entre 55.3 et 63%). L'analyse de la sensibilité a révélé la même tendance que l'analyse par région génomique : plus la taille des CNVs augmente, moins il y a de faux négatifs, ce qui augmente la sensibilité. CODEX2 était l'outil le plus sensible pour les trois tailles de CNVs : 44.59% (<5kb), 73.3% (5-150kb) et 95.46% (>150kb). Globalement, panelcn.MOPS et CODEX2 ont montré la sensibilité la plus élevée (42 à 95%), CLAMMS, DECoN et ExomeDepth ont montré une sensibilité intermédiaire (15.1 à 67.5%) tandis que XHMM a montré la plus faible sensibilité (allant de 10.2 à 29.7%).

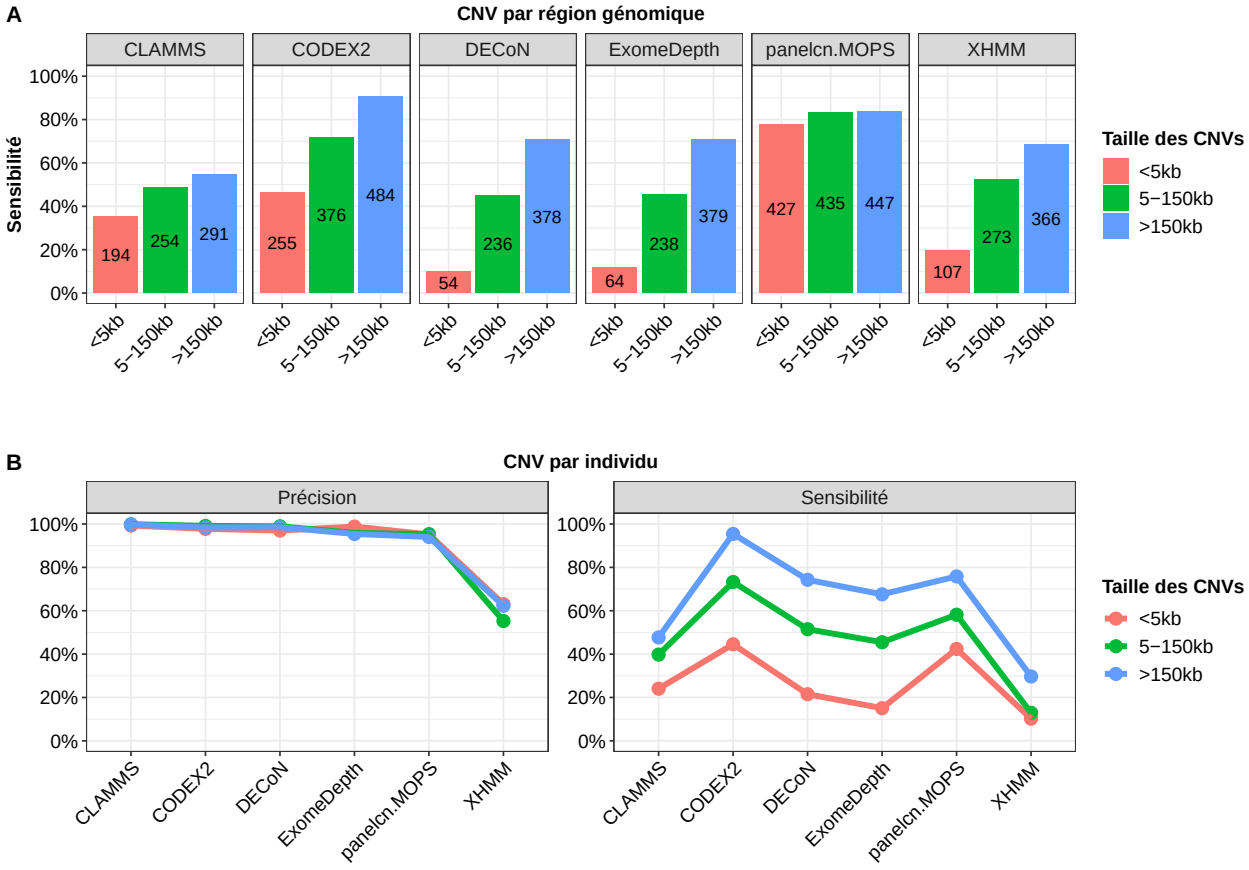

Figure 4.4: Performances des six outils de détection selon les différentes tailles des CNVs (<5kb, 5-150kb et >150kb). A. Sensibilité des outils à l'échelle de la région génomique. **B.** Précision et sensibilité des outils à l'échelle de l'individu.

4.1.2.3 Performances selon la fréquence des CNVs

Les fréquences très rares, rares et communes ($< 1\%$, $1 - 5\%$ et $> 5 - 95\%$ respectivement) des CNVs n'ont pas eu le même effet sur la sensibilité des outils de détection à l'échelle génomique comparé à l'échelle des individus (Figure 4.5).

À l'échelle de la région génomique (Figure 4.5.A), les outils ont montré une meilleure sensibilité pour les CNVs ayant une fréquence entre 1-5% (de 58.3% pour XHMM à 94.5% pour panelcn.MOPS). Les outils panelcn.MOPS et CODEX2 ont obtenu les meilleures performances de sensibilité quelque soient les fréquences des CNVs simulés : panelcn.MOPS était l'outil le plus sensible pour les fréquences

rares (94.5%) et communes (87%), tandis que CODEX2 était légèrement plus sensible pour les fréquences très rares (71.2%). L'analyse des CNVs rares et communs a révélé que CLAMMS et panelcn.MOPS ont eu des performances à l'échelle génomique opposées au reste des outils : ils ont été plus sensibles aux CNVs communs qu'aux CNVs rares. Aussi, ExomeDepth et DECoN ont montré une sensibilité plus faible pour détecter des variants communs (28.4 et 25.7% respectivement).

À l'échelle des individus (Figure 4.5.B), nous avons observé une augmentation de la sensibilité lorsque les fréquences diminuaient. CODEX2 était le meilleur outil en terme de sensibilité pour toutes les catégories de fréquences (95.2, 83.5 et 64.2% pour les CNVs très rares, rares et communs, respectivement). Pour tous les outils, la sensibilité était toujours supérieure à 88.1% pour les variants très rares. Pour les variants rares, la sensibilité variait entre 41.2 et 83.4% (XHMM et CODEX2, respectivement), tandis qu'elle diminuait pour les CNVs communs (elle variait de 8.1 à 64.1% pour XHMM et CODEX2 respectivement).

Contrairement à la sensibilité, la précision ne variait pas de la même manière en fonction des outils selon la fréquence des CNVs. Aucune différence n'était observée pour CLAMMS (précision ~90% quelque soit la fréquence des CNVs), tandis que XHMM détectait plus de faux positifs pour des CNVs fréquents (précision 51.5%) que pour des CNV très rares (précision 75.5%). Enfin, CODEX2 et panelcn.MOPS ont détecté plus de faux positifs pour des CNVs très rares.

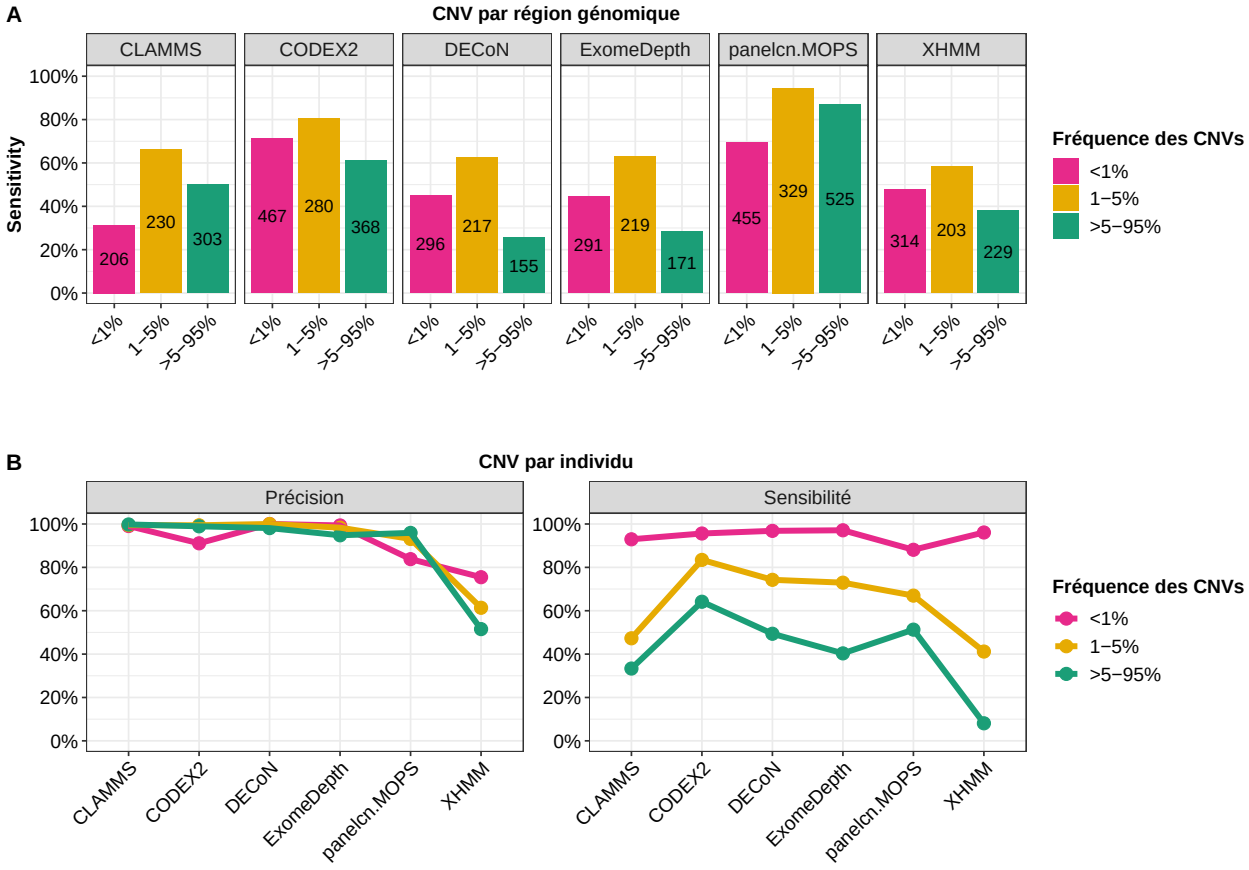

Figure 4.5: Performances des six outils de détection selon différentes fréquences de CNVs (<1%, 1-5% et >5-95%). A. Sensibilité des outils à l'échelle de la région génomique. **B.** Précision et sensibilité des outils à l'échelle de l'individu.

4.1.2.4 Performances selon le type de CNVs

Les performances des outils de détection selon le type de CNV (délétion ou duplication) à l'échelle de la région génomique et de l'individu sont présentées en Figure 4.6.

L'ensemble des outils a eu une sensibilité plus élevée pour les régions génomiques incluant des CNVs présentant à la fois une duplication et une délétion (Figure 4.6.A), variant de 69.6% pour XHMM à 96.2% pour panelcn.MOPS. D'autre part, les outils ont montré une sensibilité plus élevée pour détecter des délétions par rapport aux duplications (à l'exception de XHMM). Panelcn.MOPS était l'outil le

plus sensible sur l'ensemble des catégories (entre 78.7 et 96.2%), suivi de CODEX2 (entre 65.7 et 84.8%).

À l'échelle des individus (Figure 4.6.B), CLAMMS, CODEX2, DECoN, ExomeDepth et panelcn.MOPS ont montré une précision élevée et similaire entre les deux catégories (>93.5%). Cependant, nous avons observé une meilleure sensibilité des outils pour les délétions par rapport aux duplications. Ce différentiel de sensibilité va de 0.84% pour XHMM à 21.9% pour ExomeDepth.

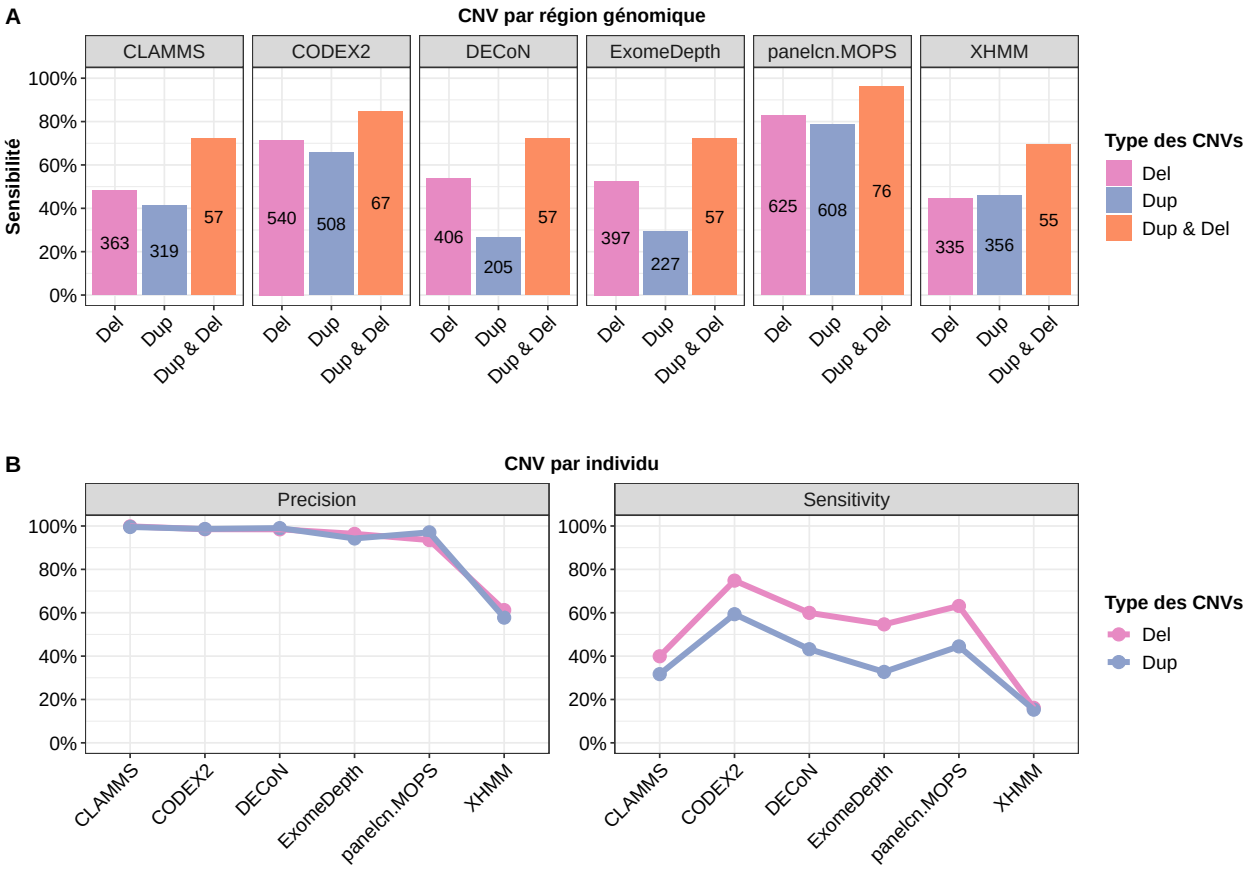

Figure 4.6: Performances des six outils de détection selon différents types de CNVs (délétion et duplication). A. Sensibilité des outils à l'échelle de la région génomique. **B.** Précision et sensibilité des outils à l'échelle de l'individu.

4.1.2.5 Comparaison des CNVs identifiés à l'échelle de l'individu en fonction des outils

Au total, 64.7% des CNVs (13324/20595) ont été identifiés par au moins un outil à l'échelle de l'individu. Le top 20 des CNVs TP retrouvés par un ou plusieurs outils à l'échelle de l'individu est présenté en Figure 4.7. L'analyse du top 20 a révélé qu'uniquement 0.41% (85) des CNVs sont retrouvés par tous les outils et 10.9% des CNVs (2249) sont retrouvés par quatre outils (CODEX2, DECoN, ExomeDepth et panelcn.MOPS). D'autre part, nous avons observé que CODEX2 et panelcn.MOPS étaient les deux outils partageant spécifiquement le plus de CNVs identifiés : 12.26% (2525). Enfin, un nombre non négligeable de CNVs ont été spécifiquement identifiés par un unique outil : panelcn.MOPS (11.5%), CLAMMS (6%) et CODEX2 (2.5%).

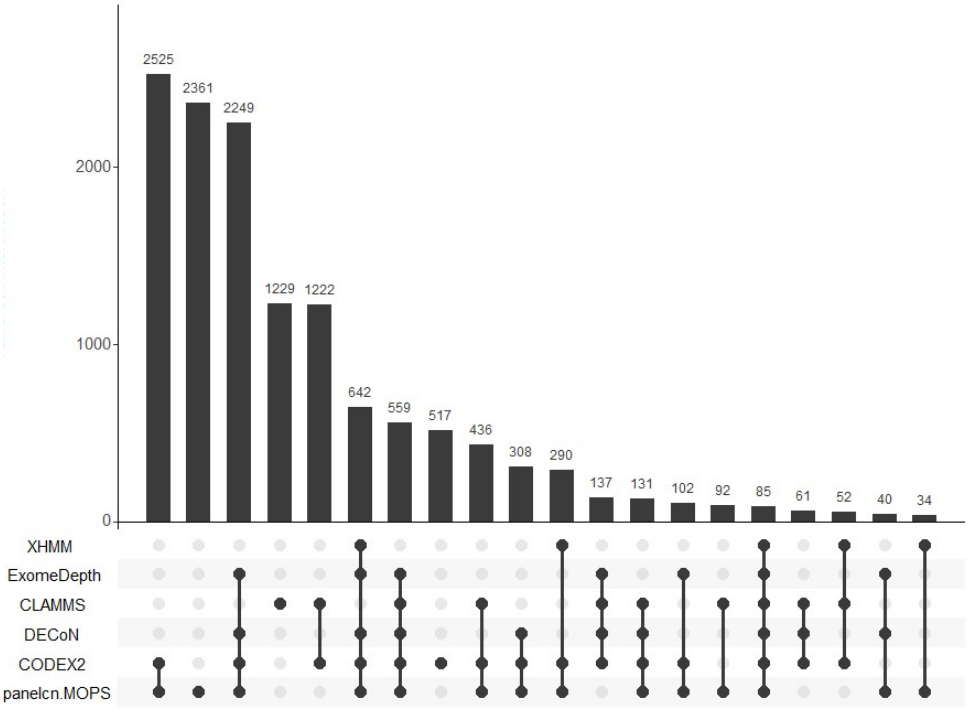

Figure 4.7: Top 20 des CNVs retrouvés par un ou plusieurs outils à l'échelle de l'individu. Le nombre de TP est indiqué au sommet de chaque histogramme.

4.1.2.6 Temps de calcul des outils de détection

Le temps de calcul de chaque outil de détection a également été mesuré pour détecter les CNVs dans nos données WES 100x simulées sur 7 chromosomes dans un échantillon de 250 individus. Ces résultats sont présentés dans la Table 4.5. Les outils DECoN, ExomeDepth et panelcn.MOPS ont été ceux ayant le temps de calcul le plus faible (environ 9H). CODEX2 et CLAMMS ont montré un temps de calcul intermédiaire (avec 14H 54M et 16H 11M, respectivement). Enfin XHMM a été l'outil le plus lent, avec un temps de calcul de 2J 23H 40M.

Table 4.5: Temps de calcul des six outils pour l'identification de CNVs simulés dans des données WES 100x pour 7 chromosomes dans un échantillon de 250 individus.

Outil	Temps de calcul
CLAMMS	16H 11M 10.11S
CODEX2	14H 54M 0.45S
DECoN	9H 29M 45.19S
ExomeDepth	9H 8M 6.34S
panelcn.MOPS	9H 16M 32.89S
XHMM	2J 23H 40M 6.51S

4.1.2.7 Performances des outils de détection dans des données réelles

Nous avons analysé deux jeux de données (BCM et WUGSC) provenant du projet 1000 Genomes [198], afin de comparer les résultats des performances des six outils de détection précédents. Ces résultats sont présentés dans la Figure 4.8.

À l'échelle de la région génomique, DEcoN et XHMM avaient une précision de 100% dans les deux jeux de données. CLAMMS et panelcn.MOPS ont montré la précision la plus basse ($< 25.1\%$) dans les deux jeux de données également. La

précision de CODEX2 était nettement inférieure dans le jeu de données WUGSC (25.1%) par rapport à celui de BCM (67.5%), tandis que la précision d'ExomeDepth était constante (50.3% pour WUGSC et 47.2% pour BCM). L'outil panelcn.MOPS a été le plus sensible (78.8% pour WUGSC et 55.2% pour BCM) tandis que CLAMMS a été l'outil le moins sensible (<12.5%). D'autre part, ExomeDepth et CODEX2 ont eu une sensibilité constante entre les deux jeux de données (~40%).

À l'échelle de l'individu, CLAMMS a montré la plus grande sensibilité (>94.5%) et précision (>57.9%) dans les deux jeux de données. À l'inverse, panelcn.MOPS a montré la plus faible précision (<37.1%) et sensibilité (<30%). Enfin, XHMM, DECoN, ExomeDepth et CODEX2 ont montré des précisions et sensibilités similaires dans les deux jeux de données.

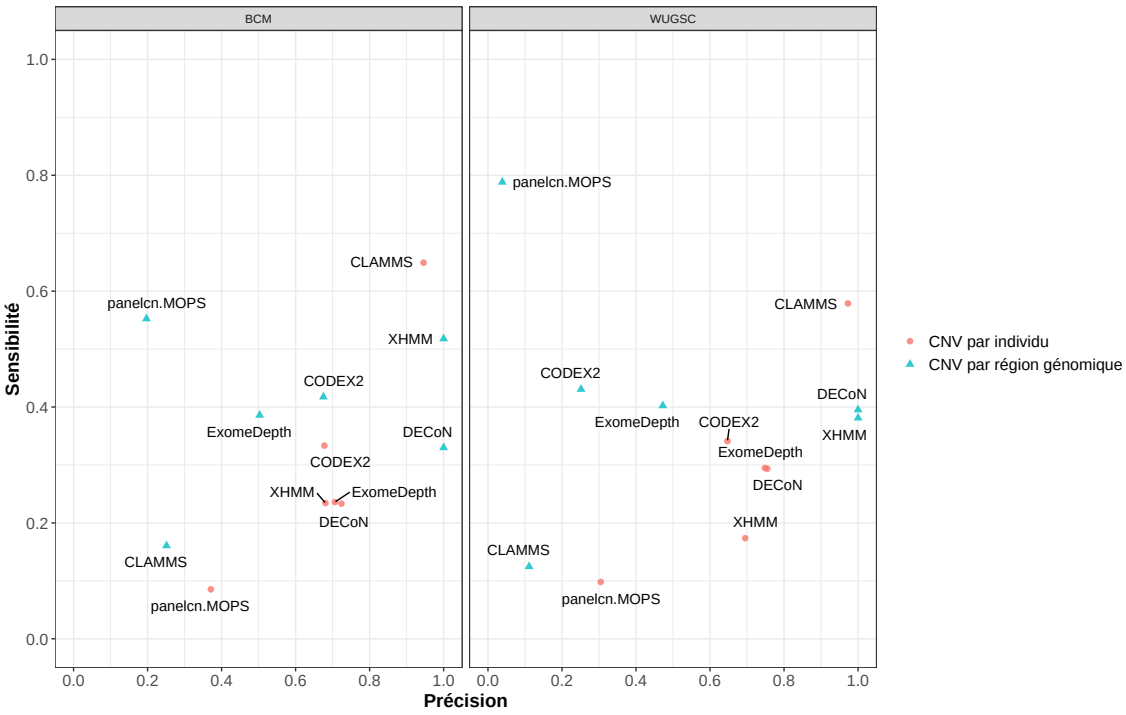

Figure 4.8: Performances des six outils de détection évalués à partir de données réelles provenant du 1000 Genomes. Les deux jeux de données proviennent des deux centres BCM et WUGSC. Les performances des outils sont évaluées à l'échelle de la région génomique ainsi qu'à l'échelle de l'individu.

4.1.3 Discussion

L'identification de CNVs est un domaine de recherche important, car ceux-ci peuvent être impliqués dans des maladies multifactorielles, permettant d'expliquer une part de l'héritabilité inconnue d'une maladie. Actuellement, les technologies NGS permettent de mieux détecter les CNVs que les techniques à puce. De par sa spécificité, la méthode la plus adaptée pour détecter des CNVs à partir de données WES est la méthode *read-depth*. La plupart des outils de détection de CNV tels que CANOES, CoNIFER, DECoN, EXCAVATOR, ExomeDepth, panelcn.MOPS et XHMM [202–208] ont été développés afin de détecter des CNVs rares, tandis que CODEX2 et CLAMMS [209, 210] ont été développés afin de détecter à la fois des CNVs rares et communs. Si la performance de certains de ces outils a été évaluée dans des études incluant des données simulées, elles n'ont pas pris en compte toutes les caractéristiques des CNVs telles que la taille, la fréquence (rare ou commun) et le type (délétion ou duplication). De plus, les outils ont toujours été évalués au niveau de la région génomique et en simulant les CNVs chez un unique individu. Ainsi, ces analyses ne prennent pas en compte plusieurs individus ayant un même CNV simulé pour une région génomique donnée. D'autre part, il est possible qu'une région génomique CNV soit correctement détectée par un outil, tandis qu'à l'échelle des individus, ceux porteurs de CNVs dans cette région ne soient pas correctement identifiés. Or, une mauvaise détection de CNV chez les individus est problématique, car les analyses d'association peuvent en être biaisées. C'est pourquoi, nous avons évalué les performances, à l'échelle génomique et à l'échelle des individus, de six outils (CLAMMS, CODEX2, DECoN, ExomeDepth, panelcn.MOPS et XHMM) les plus utilisés selon la littérature et/ou les mieux mis à jour. Pour cela, nous avons calculé la sensibilité et la précision des outils pour l'identification de CNVs à l'échelle de la région génomique ainsi qu'à l'échelle de

l'individu, selon différentes caractéristiques des CNVs (taille, fréquence et type) dans un jeu de données simulées contenant 250 individus (200 individus avec CNVs et 50 individus sans aucun CNV simulé).

À l'échelle de la région génomique, les outils détectaient très peu de faux positifs (nous avons observé une précision toujours supérieure à 98.6%), et certains outils tels que DECoN, ExomeDepth et XHMM n'ont détecté aucun faux positif (FP). Ces résultats concordent avec ceux observés par Roca et al. [212] où la précision des outils ExomeDepth, DECoN et CLAMMS était supérieure à 89%. Quel que soit les caractéristiques des CNVs, panelcn.MOPS et CODEX2 ont été les outils détectant le mieux les régions avec des CNV simulés (sensibilité variant de 46.4 à 96.2%). L'outil CLAMMS a montré une sensibilité plus faible que panelcn.MOPS pour les CNVs communs (5-95%) alors qu'il est spécifiquement conçu pour les identifier. D'une manière générale, il semblerait que les caractéristiques des CNVs influencent la performance des outils à détecter des régions génomiques puisque tous les outils avaient une meilleure sensibilité à détecter des CNVs très rares (<1%), de grande taille (>150kb) et de type délétion. Ce dernier résultat (meilleure détection des régions génomiques contenant une délétion ou à la fois une délétion et une duplication) est cohérent avec les données de la littérature [211]. De plus, avec l'utilisation de la méthode read-depth, ces résultats étaient attendus. Par exemple, dans le cadre d'une délétion homozygote dans une région donnée, la profondeur de lecture serait nulle ce qui facilite grandement la détection du CNV. Dans notre jeu de données simulées, sur 1607 régions génomiques, 11.8% avaient des CNVs complètement délétés (96.8% de ces régions ont été retrouvées par au moins 1 outil).

À l'échelle de l'individu, en se restreignant aux régions génomiques vraies positives, nous avons également observé une grande précision (>95%) des outils, excepté XHMM (59.2%) qui a détecté beaucoup de faux positifs. Ainsi, pour

ce dernier outil, la détection des régions est en partie basée sur une mauvaise spécification des CNV chez les individus. En particulier, 300 régions (18.7%) ont été détectées par cet outil alors que les CNVs ont tous été détectés chez des individus pour lesquels aucun CNV n'avait été simulé. Comme pour les régions, les outils sont sensibles aux caractéristiques des CNVs pour les identifier correctement chez les individus : une meilleure sensibilité est observée lorsque la taille du CNV simulé augmente, sa fréquence diminue et lorsque son type est une délétion. Pour l'ensemble de ces caractéristiques, CODEX2 a été l'outil le plus sensible (1.04 à 7.33 fois plus sensible que les autres outils) et le plus homogène lors de nos analyses de simulation. À l'inverse, panelcn.MOPS, qui bénéficiait de bonnes performances lors de l'analyse des régions génomiques, a vu ses performances largement réduites à l'échelle des individus, signifiant que cet outil, qui est capable de détecter une région génomique, commet des erreurs non négligeables au niveau de l'identification des individus porteurs d'un CNV. DECoN et ExomeDepth ont montré des résultats similaires malgré une performance légèrement supérieure pour DECoN, qui est une optimisation de l'outil ExomeDepth.

La comparaison des CNV mis en évidence par les différents outils à l'échelle des individus nous montre que seulement 0.41% (n=85/20595) des CNVs sont retrouvés par tous les outils. Après étude de ces 85 CNVs, nous avons constaté que ces CNVs étaient majoritairement très rares, de type délétion et très grands, ce qui conforte l'hypothèse d'un lien entre la facilité d'identification d'un CNV et ses caractéristiques. Un total de 10.9% (n=2249/20595) des CNVs ont été retrouvés en commun par 4 outils parmi les 6 (CODEX2, DECoN, ExomeDepth et panelcn.MOPS). D'autre part, l'outil ayant identifié le plus de CNVs à l'échelle des individus est panelcn.MOPS (48.4% des CNVs simulés), suivi de CODEX2 (46.2% des CNVs simulés). Ces résultats rendent difficile le choix d'un outil comme étant un outil de référence pour la détection des CNV à partir de données WES.

Afin d'évaluer l'impact de la taille de l'échantillon (n=250) sur les performances des outils, nous avons réalisé deux autres analyses où nous avons extrait 62 et 125 individus de nos données simulées. Ces résultats, qui n'ont pas été détaillés ici, ont révélé un faible impact sur la précision et la sensibilité (±5%) pour les analyses à l'échelle de la région génomique ainsi qu'à l'échelle des individus par rapport à l'échantillon initial.

Les résultats obtenus à partir des jeux de données réelles de BCM et WUGSC ont été différents de ceux obtenus à partir de données simulées. La précision et la sensibilité de l'ensemble des outils étaient diminuées par rapport aux données simulées. Globalement, les outils CODEX2, DECoN, ExomeDepth et XHMM ont montré les meilleures performances pour l'analyse par région et par individu entre les deux jeux de données. Cela peut en partie être expliqué par une qualité supérieure des données de simulations (malgré une tentative de reproduction des biais de séquençage), ce qui rend les données plus facile à analyser par les outils de détection. De plus, aucun individu de référence n'a été fourni aux outils pour l'analyse des données réelles. Dans ce cas, les outils doivent sélectionner des individus de référence sur la base de la profondeur moyenne des données de séquence. Il est donc possible que des individus avec CNVs soient sélectionnés comme individus de référence (surtout pour des CNVs communs) et rendant plus difficile la détection des CNVs. A noter que XHMM, outil le plus utilisé dans la littérature, a montré des performances globalement moins bonnes sur les données simulées comparativement aux données réelles.

Le temps de calcul des outils a révélé que DECoN, ExomeDepth et panelcn.MOPS sont 1.5 fois plus rapides que les outils CODEX2 et CLAMMS et 8 fois plus rapides que XHMM. De tels écarts de temps peuvent être expliqués par les choix des languages de programmation des outils, ou encore les modèles mathématiques implémentés par les auteurs. D'autre part, l'utilisation d'outils tels que

CODEX2 et ExomeDepth a été relativement simple, tandis que les outils CLAMMS, panelcn.MOPS, DECoN et XHMM requièrent des compétences bio-informatiques avancées afin de les utiliser.

Dans cette étude, l'analyse des performances des outils a été réalisée en utilisant la sensibilité et la précision. Pourtant, certaines études analysant les performances d'outils utilisent également la spécificité[3] [211, 215]. Le nombre de vrais négatifs (TN) peut être défini par le nombre d'exons et/ou de régions inter-CNV [211, 215]. Cependant, en raison d'un trop grand nombre d'exon et de région inter-CNV dans notre étude, le calcul des CNVs TN a été problématique et cette métrique n'a pas été utilisée.

Actuellement, il n'existe pas d'outil standard établi permettant d'analyser des CNVs à partir de données WES. Par exemple, l'analyse des outils utilisés dans des publications récentes montre qu'il n'y a pas un outil consensus pour identifier des CNV dans des maladies. Ainsi, CODEX2 a par exemple été utilisé pour identifier des CNVs dans la leucémie aiguë [229], ExomeDepth pour le syndrome de Williams [230] et DECoN pour les rétinites pigmentaires [231]. Alors que l'ensemble des outils analysés au cours de cette étude ont montré de bonnes performances afin d'identifier une région génomique avec un CNV, la détection des CNV à l'échelle de l'individu est plus difficile, et ce particulièrement pour des CNVs fréquents. Ainsi, les analyses d'association qui seraient réalisées à partir des CNVs identifiés pourraient être biaisées. Nos travaux suggèrent que CODEX2 a les performances les plus stables, suivis de DECoN, ExomeDepth et CLAMMS. Bien que les performances des outils soient à prendre en compte dans le choix de l'utilisateur, une approche multi-outils semblerait plus adaptée pour palier à la difficulté d'identifier les CNVs chez les individus. Ainsi, nous suggérons l'utilisation de l'outil CODEX2, couplée à un ou

[3] $\text{Spécificité} = \dfrac{\text{vrai négatif}}{(\text{vrai négatif} + \text{faux positif})}$

plusieurs autres outils tels que DECoN, ExomeDepth et CLAMMS.

4.2 Caractérisation de CNVs rares associés à la Polyarthrite Rhumatoïde

Des études disposant d'environ un millier de sujets ont permis d'identifier des CNVs fréquents associés à la PR tels que ceux des gènes *CCL3L1* [194] (populations néozélandaise et britannique avec ancêtres caucasiens) et *VPREB1* (population pakistanaise) [232]. Cependant, peu d'études se sont intéressées jusqu'ici aux variants rares dans la PR. Une étude récente GWAS a identifié 11 CNVs avec des fréquences inférieures à 5% (considérant cette fréquence comme rare) [233]. Enfin, aucune étude à notre connaissance n'a utilisé des données familiales afin d'identifier des CNVs rares à partir de données WES.

Dans cette partie, nous allons utiliser les 4 outils identifiés dans la partie précédente comme étant les plus performants afin d'identifier des CNVs rares associés à la PR, à partir de données WES issues de familles d'origine européenne.

4.2.1 Matériel et Méthodes

4.2.1.1 Échantillons

Tous les individus participant à l'étude ont été informés et ont fourni un consentement éclairé signé. Le recueil des données a été approuvé par le comité d'éthique de l'Hôpital Bicêtre et de l'Hôpital Saint Louis (Paris, France ; CPPRB 94-4).

À partir des données de la thèse de M. Veyssière [234], nous disposions de 9 familles représentées dans la Figure 4.9. Depuis ces familles, un set de découverte de 30 individus a été constitué dans le but de réaliser un séquençage en WES (individus avec un triangle rouge dans la Figure 4.9). Les caractéristiques épidémiologiques et sérologiques des 30 individus du set de découverte sont présentées en Table 4.6. Le set de découverte inclut 19 individus atteints de PR et 11 individus non atteints. Parmi les individus atteints, nous retrouvons un nombre plus élevé de femmes (13) atteintes de PR que d'hommes (6). D'autre part, tous les individus atteints de PR du set de découverte sont porteurs d'au moins un allèle épitope partagé (SE) du gène *HLA-DRB1*. Enfin, ces données sont composées d'au minimum deux individus atteints de PR et un individu non atteint dans chaque famille (à l'exception d'une famille).

Table 4.6: Caractéristiques sérologiques et épidémiologiques pour les 30 individus du set de découverte. Tiré de la thèse de M. Veyssiere [234].

Caractéristiques	**Atteint de PR**	**Non Atteint de PR**
Sexe		
Homme	6 (31.6%)	3 (27.3%)
Femme	13 (68.4%)	8 (72.7%)
Age[a]		
$\geq$ 40 ans	7 (36.8%)	7 (63.6%)
$<$ 40 ans	10 (52.6%)	3 (27.3%)
inconnu	2 (10.5%)	1 (9.1%)
Sérologie		
FR +	16 (84.2%)	0 (0%)
FR -	2 (10.5%)	10 (90.9%)
ACPA +	15 (78.9%)	2 (18.2%)
ACPA -	3 (15.8%)	8 (7.3%)

[a] âge de diagnostic pour les atteitns et âge de prélèvement pour les non atteints.

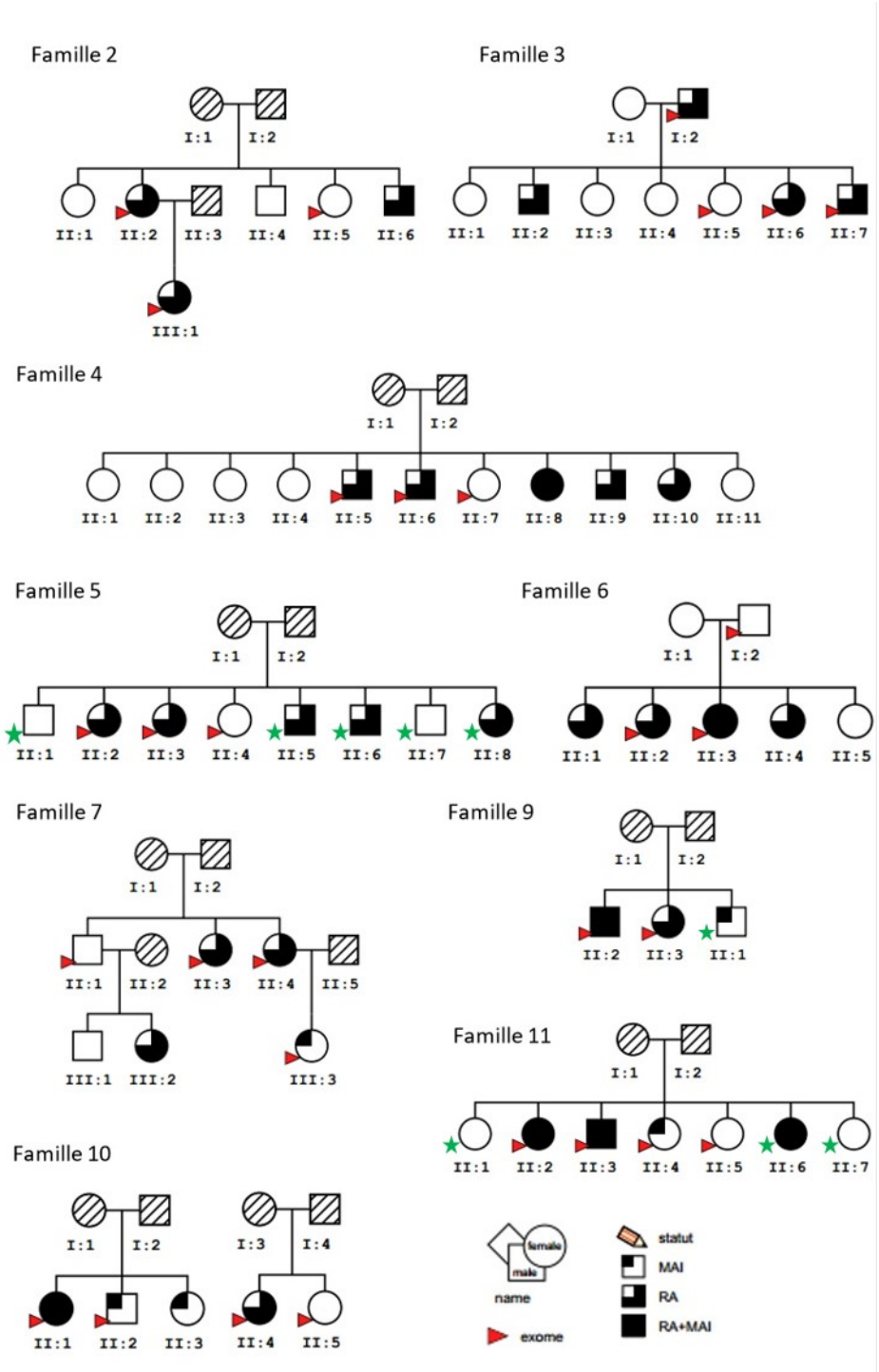

Figure 4.9: Familles étudiées. Tiré de la thèse de M. Veyssiere [234]. Un triangle rouge indique les individus pour lesquels un WES a été éffectué. Une étoile verte indique les individus utilisés pour l'étape de validation.

4.2.1.2 Traitement des données NGS

Le séquençage des données d'exomes des 30 individus a été effectué par le Centre National de Recherche en Génomique Humaine (CNRGH). Les exons ont été capturés en utilisant le kit Agilent SureSelect Human All Exon (V5), ciblant 21 522 gènes incluant 648 740 régions exoniques. Les exons ont ensuite été séquencés avec la plateforme Illumina HiSeq2000. Les lectures obtenues ont été traitées en utilisant un workflow incluant : 1. l'alignement des lectures sur la référence hg19 en utilisant l'algorithme BWA-MEM [235], 2. la compression des données en utilisant SAMtools [236]. 3. Le marquage des duplicats en utilisant PICARD [225]. À l'issue de ce workflow, les fichiers ont été stockés sous le format BAM, qui est un format binaire compressé d'un fichier SAM, directement exploitable par les outils de détection de CNVs exoniques.

4.2.1.3 Détection et sélection des CNVs

Dans cette analyse, quatre outils de détection de CNVs exoniques (CLAMMS, CODEX2, DECoN et ExomeDepth) ainsi que leurs seuils de détection définis dans la section 4.1.1.2, ont été utilisés. Les outils Panelcn.MOPS et XHMM n'ont pas été considérés pour cette étude, panelcn.MOPS à cause de son paramétrage complexe révélé dans notre étude précédente, et XHMM en raison de ses mauvaises performances et d'un temps de calcul trop long.

Un CNV peut être détecté par un outil chez plusieurs individus, sans pour autant que les régions soient identiques mais seulement chevauchantes. Le CNV est alors appelé CNV région (CNVr). Pour déterminer si plusieurs CNVs chevauchants sont en réalité un unique CNVr, nous avons utilisé le coefficient de chevauchement (aussi appelé coefficient Szymkiewicz–Simpson). Cette mesure de similarité calcule le chevauchement entre deux ensembles (*overlap*), et est définie comme la taille de

l'intersection divisée par la plus petite des tailles des deux ensembles :

$$overlap(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

Un CNVr a été construit si le score de chevauchement de deux CNVs était supérieur à 50%. Afin de déterminer les positions de chaque CNVr constitué à partir de deux ou plusieurs CNVs, nous avons considéré les extrémités de la région chevauchante. Le nombre de CNV total ainsi que le nombre de CNVr obtenus pour chaque outil sont référencés dans la Table 4.7.

Table 4.7: Nombre de CNVs et de CNVr détectés par les outils à l'échelle de la région génomique et de l'individu. Les résultats ont été obtenus à partir des données WES des 30 individus.

	CLAMMS	CODEX2	DECoN	ExomeDepth
CNV par individu	514	1309	1552	1654
CNV par région	355	432	315	290

4.2.1.4 Annotation des CNVs

L'annotation des CNVs détectés à l'étape précédente a été réalisée avec l'outil AnnotSV 2.3.2 (Annotation and Ranking of Human Strutural Variations) [237].

AnnotSV annote les variants structuraux à partir de bases de données publiques inclues dans l'outil. À notre connaissance, il s'agit du seul outil d'annotation de variants structuraux permettant d'annoter simultanément les gènes, les éléments régulateurs, la pathogénicité des régions, la fréquence des variants (DGV et GnomAD) et les points de rupture. De plus, l'outil intègre un score permettant la classification en 5 catégories de la pathogénicité potentielle de chaque variant structural (défini dans la Table 4.8) basé sur un consensus recommandé par l'American College of Medical Genetics (ACMG) et ClinGen [238].

Table 4.8: Classification de l'effet pathogène des variants structuraux par AnnotSV

Effet	Rang
Pathogène	5
Probablement pathogène	4
Incertain	3
Probablement bénin	2
Bénin	1

4.2.1.5 Stratégie de sélection des CNVs

Dans le but d'identifier de nouveaux facteurs génétiques à risque, il a été sélectionné uniquement les CNVs présents chez tous les atteints d'une même famille et absents chez l'ensemble des non atteints du set de découverte. Ainsi, les variants sont sans phénocopie au sein d'une famille (les atteints sont porteurs du CNV) et à pénétrance complète (aucun non atteint dans le set de découverte n'est porteur du CNV). D'autre part, nous avons sélectionné uniquement les CNVs trouvés par un minimum de trois outils, incluant obligatoirement CODEX2 (voir section 4.1.3). Enfin, un CNV a été conservé et considéré rare si sa fréquence dans les bases GnomAD ou DGV était inférieur à 1% à partir de l'annotation de l'outil AnnoSV.

4.2.1.6 Validation des CNVs candidats en familles étendues

Les CNVs identifiés à l'étape précédente ont ensuite été génotypés dans le set de découverte ainsi que dans 9 individus supplémentaires des familles où ils ont été identifiés (individus avec une étoile verte dans la Figure 4.9) au laboratoire en utilisant une PCR digitale en gouttelettes (*Droplet Digital PCR*, ddPCR, *QX200 Droplet Digital PCR System, Bio-Rad Laboratories, California, USA*). La technologie ddPCR consiste à amplifier un échantillon d'ADN contenu dans des gouttelettes en suspension. Pour cela, un échantillon d'ADN, ainsi que des sondes TaqMan contenant un fluorophore (FAM et HEX ou VIC), sont introduits dans un milieu

réactionnel contenant de l'huile minérale. Un générateur de gouttelettes (*QX200 Droplet Generator*) effectue une émulsion du milieu contenant l'huile, permettant de répartir l'échantillon d'ADN dans $\simeq$ 20000 gouttelettes. Ces gouttelettes sont ensuite récupérées et amplifiées par PCR. Par la suite, les gouttelettes sont lues par l'automate *QX200 Droplet Reader*, analysant la fluorescence de chaque gouttelette en comptant le nombre d'évènements positifs et négatifs. Un logiciel dédié s'appuie sur une loi de Poisson pour estimer la concentration (copies/µL) pour la région cible et pour un gène de référence diploïde afin de calculer le nombre de copies de cette région.

Pour la validation des CNVs identifiés, des sondes TaqMan ont été spécialement conçues et utilisées selon les positions génomiques de chaque CNV identifié avec le set de découverte. Le gène *RPP30* a été utilisé comme gène de référence.

4.2.2 Résultats

4.2.2.1 Identification de CNVs

À partir des données de séquences exoniques des 30 individus (individus ayant un triangle rouge dans la Figure 4.9), les outils de détection ont identifié entre 290 (ExomeDepth) et 432 (CODEX2) CNVs par région (voir Table 4.7). La répartition moyenne de ces CNVs dans les différents chromosomes est présentée en Figure 4.10. Nous avons observé une distribution inégale des CNVs dans les chromosomes. Les chromosomes 1 et 19 comportaient en moyenne le plus de CNVs (38 et 32 respectivement), tandis que les chromosomes 13, 18, 20 et 21 comportaient en moyenne le moins de CNVs (7, 6, 9 et 6 respectivement). Enfin, nous avons observé en moyenne plus de CNV région de type délétion ou duplication uniquement, que de CNV région de type duplication et délétion.

4.2.2.2 Recherche de CNVs rares associés à la PR

À partir de ces CNVs, il a été établi une liste de 3 CNVr respectant notre stratégie de sélection de CNV (voir 4.2.1.5). Ces trois CNVs région ont été détectés par CODEX2, ainsi que par un minimum de 2 autres outils. Ces CNVs n'étaient pas présents dans la base de données DGV, excepté pour la région *chr19:46623565-46628011*, qui est présente mais avec un type de CNV inverse à celui retrouvé.

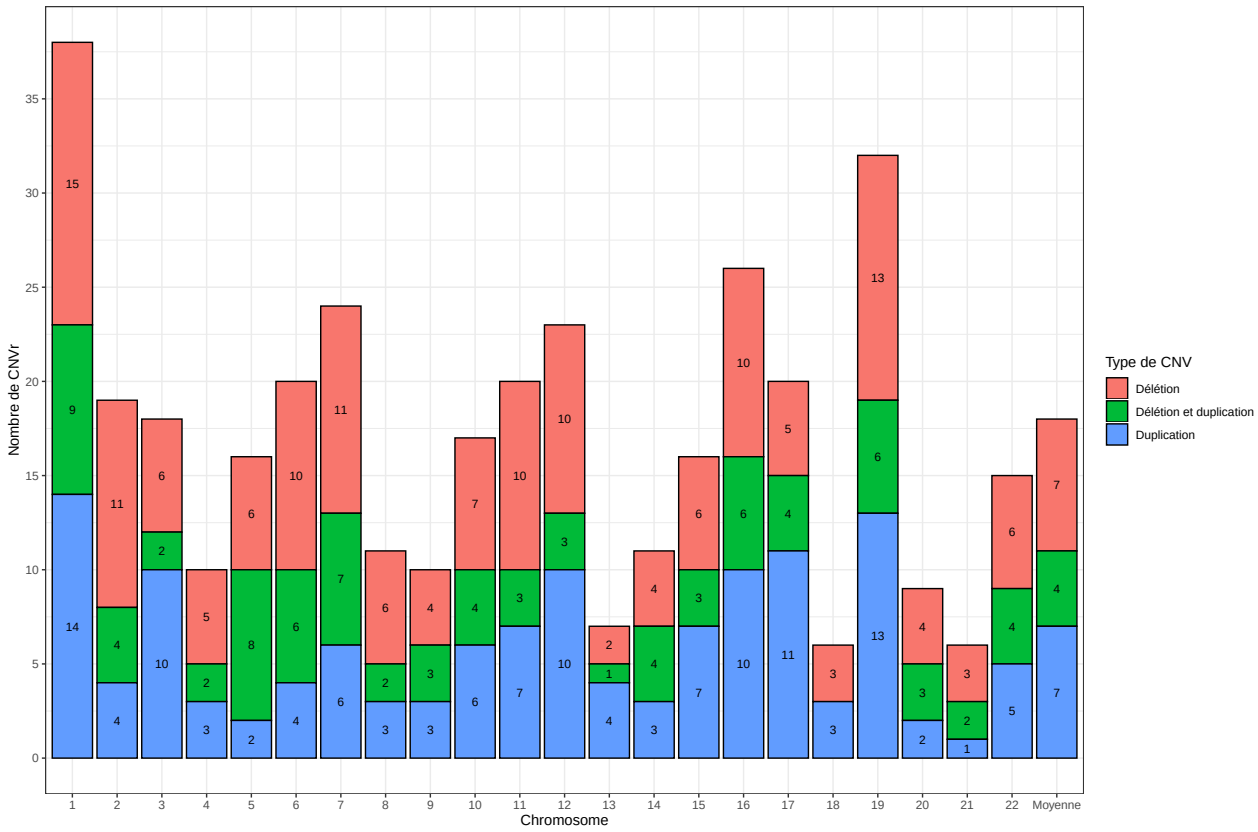

Figure 4.10: Répartition moyenne des CNVs identifiés par chromosomes.

Table 4.9: Résultats de la recherche de CNVs rares associés à la PR dans le set de découverte

Gènes	Position	Outils	Type	n° Famille	# PR	# Non PR
GSDME, *PALS2*	chr7:24727101-24789334	4 (CX, ED, CK, DN)	Délétion	5	2/2	0/1
IGFL3	chr19:46623565-46628011	3 (CX, CS, DN)	Délétion	11	2/2	0/2
CD37, *SLC6A16*, *MIR4324*	chr19:49793390-49840490	4 (CX, ED, CS, DN)	Duplication	9	2/2	NA

Note:
CS : CLAMMS
CX : CODEX2
DN : DECoN
ED : ExomeDepth
NA: Pas de données WES

4.2.2.3 Validation des CNV candidats en familles étendues

La validation des 3 CNVs identifiés précédemment a été réalisée par ddPCR dans les échantillons séquencés des familles concernées. Les résultats de cette validation sont présentés en Table 4.10 (colonne WES) et confirment ceux obtenus par les outils de détection dans le set de découverte.

L'analyse par ddPCR de ces mêmes CNVs pour des individus supplémentaires des familles concernées (individus représentés avec une étoile verte dans la Figure 4.9) a ensuite été réalisée. Ces résultats sont également présentés dans la Table 4.10 (colonnes # PR et # Non PR) et invalident l'hypothèse de variants avec pénétrance complète et sans phénocopie de la PR car ces CNVs ont été identifiés chez des individus non malades et/ou non identifiés chez des individus atteints du set de validation.

Table 4.10: Résultats de l'identification des CNVs par ddPCR dans le set de validation

Gènes	**Position**	**n° Famille**	**WES**	**# PR**	**# Non P**
GSDME, *PALS2*	chr7:24727101-24789334	5	3/3	4/4	2/3
IGFL3	chr19:46623565-46628011	11	4/4	3/3	2/3
CD37, *SLC6A16*, *MIR4324*	chr19:49793390-49840490	9	2/2	4/5	1/3

Note:
Position : Génome GRCh37/hg19
PR, non PR : Nombre d'individus atteints et non atteints avec CNV sur le total d'atteints
non atteints respectivement dans les familles du set de validation

4.2.3 Discussion

Dans cette deuxième partie, nous avons appliqué sur des données réelles 4 outils de détection (CLAMMS, CODEX2, DECoN, ExomeDepth) dont nous connaissions les performances grâce à nos travaux réalisés sur les données de simulation. L'objectif était d'identifier des CNV candidats dans la PR à partir de données de séquence WES de 30 individus appartenant à des familles multiplexes de PR. Puisque les analyses précédentes ont révélé que la caractérisation des CNVs chez les individus n'était pas optimale, nous avons choisi de considérer un CNV s'il était détecté par un minimum de 3 outils, avec, parmi ceux-ci CODEX2, l'outil le plus performant dans notre étude précédente. Trois CNVs remplissaient nos critères de sélection, à savoir des CNVs rares, délétères, une absence de phénocopie au sein d'une famille, ainsi qu'une pénétrance complète. Malheureusement, la validation par ddPCR de ces CNV dans les familles étendues n'a pas permis de valider nos critères de sélection (absence de phénocopie et pénétrance complète). Pour chacun de ces CNVs, la validation par ddPCR a révélé leur présence chez des individus supplémentaires non atteints de PR, indiquant une pénétrance incomplète, ainsi que l'absence d'un CNV chez un individu supplémentaire atteint de PR, indiquant une phénocopie. Toutefois, la ddPCR a permis de confirmer les CNVs chez les individus séquencés dans le set initial. Cela conforte notre choix d'avoir considéré un CNV s'il était trouvé par au moins 3 outils.

De manière intéressante, l'un de ces CNVs, la délétion hétérozygote en position chr19:46623565-46628011[4], sans phénocopie et à pénétrance incomplète dans le set de validation, touche le gène *GSDME*, faisant partie de la famille de la gasdermine (GSDM). Cette famille code pour des protéines responsables de pyroptose qui est un mécanisme de mort cellulaire programmée pro-inflammatoire. Récemment, plusieurs

[4]Position : Génome GRCh37/hg19

études ont montré que la protéine GSDME avait un rôle pathogène dans la PR, en étant impliqué dans la pyroptose des monocytes [239, 240], des macrophages [240] et également dans la modularisation de la prolifération, migration, invasion et de la sécrétion de cytokines inflammatoires dans les FLS de la PR [241]. D'autre part, la duplication hétérozygote en position chr19:49793390-49840490[5] touchant le gène *CD37* avec phénocopie et à pénétrance incomplète dans le set de validation, code pour une protéine qui est un antigène majeur des lymphocytes B, impliquée dans le système immunitaire, incluant l'inhibition de la prolifération des lymphocytes T [242], la régulation négative de l'adhésion et migration des cellules dendritiques et neutrophiles [243, 244] et l'inhibition de la production d'IL6 (protéine hautement impliquée dans la PR) dans les macrophages [245]. Cependant aucun lien direct avec la PR n'a été établi.

Ainsi, certains CNVs identifiés dans le set initial pourraient être de bons candidats dans la PR. En effet, nos critères de sélection étaient très stricts (CNV rare à pénétrance complète et sans phénocopie) alors que les facteurs génétiques impliqués dans les maladies multifactorielles ont souvent une pénétrance incomplète avec phénocopie. Ainsi, un CNV rare pourrait tout à fait ségréger dans une famille sans être présent chez tous les atteints de la famille et en étant présent chez certains non atteints. Les non atteints dans les familles avec les 3 CNVs identifiés dans le set initial avaient en moyenne 61 ans, ce qui réduit la possibilité que ces individus développent la maladie par la suite. À noter que toutes les familles étudiées avaient un allèle *HLA-DRB1* SE qui ségrégeait. Ainsi, les facteurs génétiques identifiés seraient des facteurs modulateurs de l'effet de *HLA-DRB1*.

Une des limites de cette étude est que l'utilisation de données WES ne permet que la détection de CNVs exoniques, représentant 1% du génome, même si ces

[5] 3

derniers restent plus faciles à interpréter concernant leur impact sur la fonctionnalité des protéines. Une autre limite concerne la position exacte des CNVs qui ne peut pas être déterminée en utilisant la méthode *read-depth* puisque limitée aux régions exoniques. Or, les points de cassure peuvent être localisés en dehors des régions exoniques. Face à ces inconvénients, une solution consiste à étudier le génome entier en utilisant le séquençage WGS. L'analyse de données WGS ne se limite pas aux exons et, en utilisant une méthode alternative à la méthode *read-depth*, permet d'identifier les positions exactes des CNVs.

Chapitre 5

Caractérisation de variants rares par étude de séquençage génome entier

Actuellement, la part connue de la composante génétique impliquée dans la PR est estimée à $\simeq$ 60% [121]. Les études GWAS ont largement contribué à mieux caractériser la composante génétique de la PR, en permettant d'identifier des variants génétiques communs ayant un effet faible voir modéré. Cependant, ces approches ont plus de difficultés à évaluer la contribution des variants génétiques rares car la plupart de ces variants sont spécifiques d'une population ou d'une famille [246]. Pourtant, les variants rares permettraient d'expliquer une partie de l'héritabilité manquante dans les maladies complexes [247, 248], incluant la PR [249]. Des approches alternatives, utilisant des échantillons familiaux [250, 251] et des données WGS, permettent de mieux capturer ces variants rares [246].

L'intérêt du WGS, malgré un coût supérieur au WES, est justifié par la position des 100 loci identifiés dans le cadre de la PR par les analyses GWAS. En effet, 80% de ces loci sont localisés dans des régions non codantes, régions ne pouvant

être analysées uniquement par l'utilisation du WGS [252]. D'autre part, si des études familiales portant sur des données WES ont identifié des variants causaux rares touchant les gènes *SUPT20H* [144] et *PLB1* [253] (soit 1 par étude), aucune étude à notre connaissance n'a utilisé des données WGS familiales. Toutefois, cette approche a porté ses fruits dans d'autres maladies complexes telles que la maladie d'Alzheimer (44 variants causaux rares identifiés) [254].

Enfin, l'utilisation de patients atteints de PR non porteurs d'allèle à risque *HLA-DRB1*, facteur génétique majeur dans la PR, pourrait permettre de mettre plus facilement en évidence d'autres facteurs génétiques non découverts.

Dans ce chapitre, nous proposons la première étude à ce jour utilisant des données WGS issues de familles multiplexes de PR d'origine européenne, dans le but d'identifier de nouveaux facteurs génétiques rares associés à cette maladie.

5.1 Matériel et Méthodes

5.1.1 Échantillons

Tous les individus participant à l'étude ont été informés et ont fourni un consentement éclairé signé. Le recueil des données a été approuvé par le comité d'éthique de l'Hôpital Bicêtre et de l'Hôpital Saint Louis (Paris, France ; CPPRB 94-4). Nous avions à notre disposition un échantillon de 25 individus appartenant à 5 familles multiplexes, avec au moins 2 membres atteints de PR par famille, toutes d'origine française (origine des cas index[1] et de leurs 4 grands-parents vérifiée). Dans les 5 familles, certains individus, atteints de PR ou non, avaient développé une autre maladie autoimmune (MAI) : diabète de type I, lupus érythémateux, syndrome de Sjogren, thyroïdite. Sur les 25 individus, la répartition des statuts était la suivante :

- 6 (24%) atteints de PR seulement
- 6 (24%) atteints de PR et d'une autre MAI
- 1 (4%) atteint d'une MAI autre que la PR
- 12 (48%) non atteints

À partir de ces familles, un set de découverte de 15 individus a été constitué. Les individus ont été choisis car ils avaient une quantité d'ADN suffisante pour réaliser un séquençage entier du génome et de manière à avoir deux individus atteints de PR par famille ainsi qu'un individu non atteint (individus représentés avec un triangle rouge dans la Figure 5.1).

Pour valider les variants identifiés dans le set de découverte et analyser leur ségrégation dans les familles, 10 individus supplémentaires (2 atteints et 8 non-atteints) étaient disponibles (individu dont l'identifiant est encadré par des crochets

[1]Individu atteint par lequel la famille a été recueillie

dans la Figure 5.1). Le set de validation comporte donc au total 25 individus.

Les caractéristiques démographiques des 25 individus sont décrits dans la Table 5.1. Dans cet échantillon, bien que non significatif (pvalue du test de Fisher=0.2), le pourcentage de femmes touchées par la PR (58.8%) était environ deux fois plus élevés que le pourcentage des hommes (25%), ce qui concorde avec la littérature. De même, l'âge moyen d'apparition de la PR observé dans la littérature est $\simeq$ 45 ans, ce qui concorde avec l'age moyen observé au diagnostic de 48 ans dans notre échantillon (moyenne calculée avec 9 patients sur 10, l'âge d'un patient est manquant). D'autre part, il n'y a pas de différence significative entre les moyennes d'inclusion des individus atteints et non atteint de PR (p-value du test de Student = 0.34). Enfin, l'âge moyen d'inclusion des individus non atteints de PR est de 61 ans, ce qui limite la possibilité de faux négatifs (individu dont la maladie ne se serait pas encore déclarée).

Table 5.1: Caractéristiques démographiques des 25 individus.

Caractéristiques	**Atteint de PR**	**Non atteint de PR**
Sexe		
Homme	2	6
Femme	10	7
Age		
Moyen d'inclusion	65 ($\pm$14)	61 ($\pm$10)
Moyen d'apparition de la PR	48 ($\pm$11)	

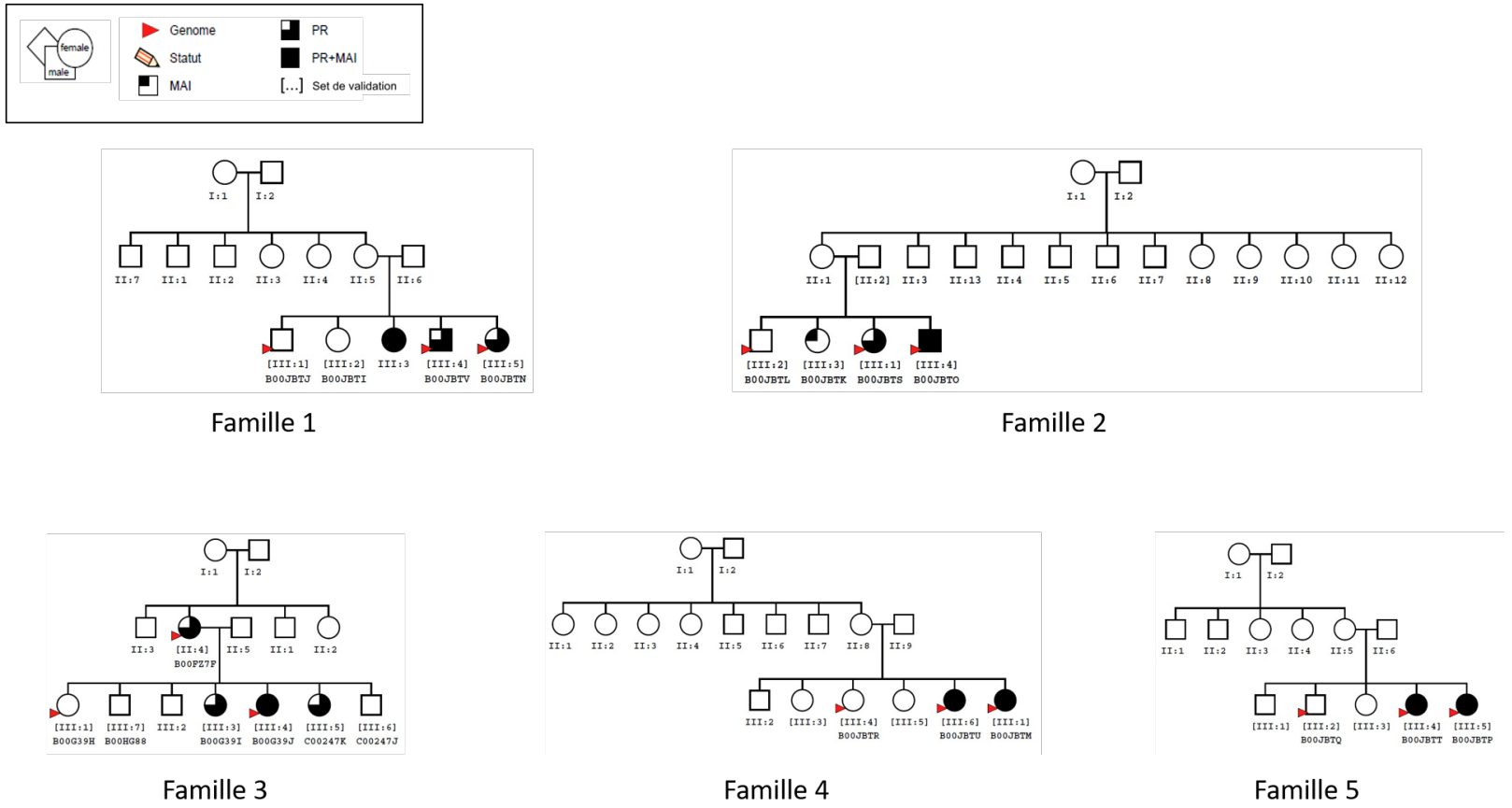

Figure 5.1: Familles multiplexes de PR. Les 15 individus composant le set de découverte sont identifiables par un triangle rouge. Les 25 individus composant le set de validation sont identifiables par des crochets entourant leurs identifiants familiaux.

5.1.2 Séquençage et traitement des données génomiques

Le séquençage des génomes des 15 individus du set de découverte a été effectué par le CNRGH en utilisant la plateforme Illumina HiSeq2000. La couverture de séquençage de ces individus était de 91% (± 1), tandis que 81% (± 10) des génomes avaient une profondeur de séquençage minimum de 30X. Les lectures obtenues et générées sous forme de séquences fastq ont été traitées selon un pipeline d'analyse interne présenté dans la Figure 5.2.A. Tout d'abord, ces lectures ont été alignées à l'aide du programme BWA MEM [235] sur une version du génome issue de la référence humaine HG37. Ces fichiers sont ensuite compressés en fichier BAM (cartographie d'alignement binaire sans perte) en utilisant les programmes Samtools [236] et Picardtools [225]. Les duplicats ont ensuite été référencés avec l'outil Sambamba [255], afin que l'étape de l'appel des variants (variant calling) ne soit pas biaisée. Enfin, pour mieux positionner les indels, les lectures ont été réalignées à l'aide de la suite de programmes GATK [256].

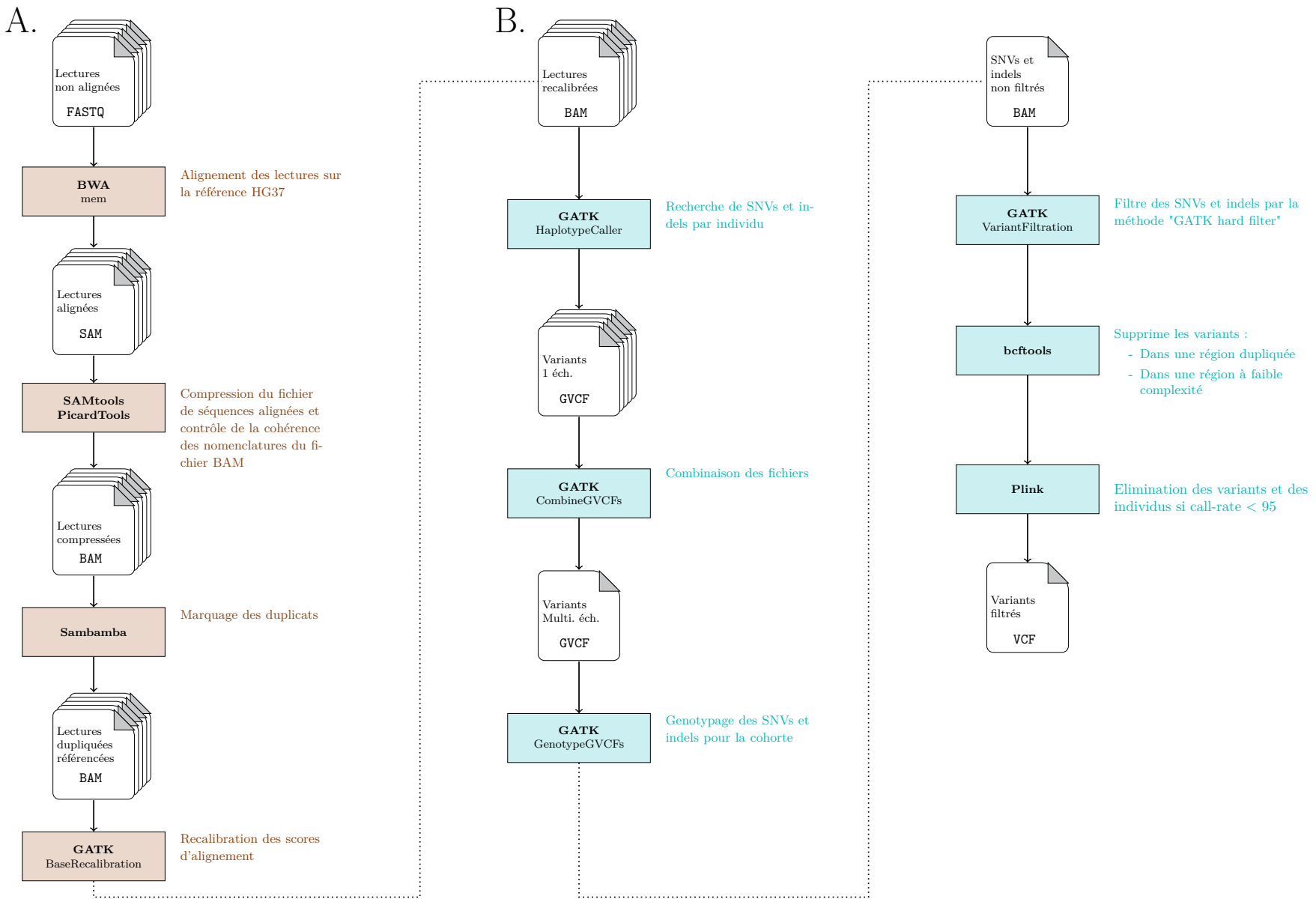

Figure 5.2: Pipeline de production des SNVs et indels à partir de données de lectures post-séquençage. A. Pipeline d'alignement, de recalibrage, de marquage et de compression à partir de lectures non alignées (coloré en beige). **B.** Pipeline de production de SNVs et indels de hautes qualités à partir de lectures recalibrées (coloré en bleu).

5.1.3 Recherche des SNVs et indels candidats

5.1.3.1 Contrôle qualité des variants

La recherche des variants de type SNVs et indels a été réalisée en utilisant un pipeline interne présenté dans la Figure 5.2.B. La recherche de SNVs et indels à partir de données de séquence a été effectuée avec l'outil HaplotypeCaller. Nous avons ensuite appliqué un filtre de qualité sur les variants identifiés en utilisant la méthode *hard filter* préconisée par GATK lorsque l'échantillon est inférieur à 30 individus. Cette méthode implique des filtres différents pour les indels et SNVs, comme détaillés dans la Table 5.2. Par la suite, les variants identifiés dans les régions répétées peu complexes et dans des duplications segmentales ont été retirés en utilisant les bases de données RepeatMasker [257] et genomicSuperDups [258, 259] via l'outil BCFtools [236]. Enfin, Plink [260] a été utilisé afin de filtrer les variants et individus ayant un taux de génotypes manquants supérieur à 5%. Les variants ayant passé l'ensemble de ces filtres ont été considérés comme variants de haute qualité.

5.1.3.2 Annotation des variants

Afin d'identifier les variants rares et délétères parmi l'ensemble des variants obtenus à l'étape précédente, nous avons procédé à leur annotation. Pour cela, nous avons utilisé l'annotateur VcfAnno [261], permettant d'annoter un fichier VCF depuis plusieurs sources de données tout en parallélisant les tâches. Les annotations suivantes ont été ajoutées à chaque variant :

1. Fréquences alléliques dans les populations de référence européennes.
2. Prédiction de son effet.

Table 5.2: Filtres préconisés selon la méthode GATK *Hard-filter*.

Filtres	SNV	Indel	Descriptions
QUAL	< 30	< 30	Probabilité à l'échelle Phred que la position ne contienne pas de variant
QD	< 2	< 2	Score QUAL normalisé par la profondeur allélique du variant
SOR	> 3		Estimation du biais de brin par le test odds ratio symétrique (SOR)
FS	> 60	> 200	Score estimant un éventuel biais de brin
MQ	< 40		Qualité d'alignement moyen sur l'ensemble de la lecture
MQRankSum	< -12.5		Test du biais de différence de qualité d'alignement entre les allèles
ReadPosRankSum	< -8	< -20	Test du biais de position relatif aux allèles REF et ALT dans les lectures
DP	< 10	< 10	Profondeur de couverture pour chaque individu

Fréquences alléliques

Au cours des dernières années, plusieurs projets à grande échelle ont été réalisés afin d'estimer la fréquence des variants dans diverses populations. Les populations de référence sont établies en utilisant un échantillon d'individus ayant les mêmes origines ethniques. Cependant, certaines spécificités génétiques au sein d'une population font qu'un pays peut en être exclu, comme c'est le cas avec la Finlande par rapport aux autres pays composant l'Europe. Cette exclusion est historiquement dû à une faible augmentation démographique engendrant des dérives génétiques [262, 263]. Parmi les études à grande échelle menées, nous retrouvons le projet 1000 Genomes [198], le NHLBI Exome Sequencing Project (ESP6500) [264], le UK10K [265] et l'Exome Aggregation Consortium (ExAC) [266] maintenant fusionné avec le Genome Aggregation Database (GnomAD) [267]. Ces bases de données, présentées dans la Table 5.3, contiennent des séquences WGS et/ou WES

d'individus à partir desquelles ont été établies les fréquences des variants connus. Ainsi, l'ensemble des fréquences européennes (incluant notamment des pays de l'Europe du sud et de l'Europe du nord) de ces bases de données ont été téléchargées puis utilisées afin d'annoter les SNVs et les indels.

Table 5.3: Bases de données publiques contenant des fréquences alléliques de référence.

	1000 Genomes	**UK10K**	**ESP6500**	**ExAC**	**GnomAD**
Nb d'individus	3 115	10 000	6 503	60 706	141 456
Nb de populations	26	1	2	7	17
SNVs et indels	✓	✓	✓	✓	✓
Variants structuraux	✓			✓	✓
Type de données	WGS	WGS	WES	WES	WES & WGS

Note:
✓: Disponible dans la base de données
WGS : Whole Genome Sequencing
WES : Whole Exome Sequencing

Effet des variants

Il existe de nombreux outils permettant d'annoter l'effet d'un variant. Ces outils sont basés sur des méthodes utilisant l'information des SNV non-synonymes (nsSNV) afin de déterminer l'impact de ces variants sur la protéine produite (SIFT [268], Polyphen2 [269], LRT [270], et MutationTaster [271] étant les pionniers) ou encore la conservation phylogénétique du site de variation (PhyloP [272], phastCons [273] et GERP++ [274]). Plus récemment, d'autres outils ont été implémentés pour utiliser des sources d'informations multiples grâce au *machine learning* afin de prédire l'effet d'un variant (CADD [275], DANN [276], MetaLR [277], MetaSVM [277]).

Afin de faciliter l'accessibilité à ces outils, la méta-base de données dbNSFP [278] a été développée, permettant d'annoter l'ensemble des SNV non-synonymes (nsSNV)

du génome. La méta-base dbNSFP contient un total de 37 outils de prédiction d'effets de variants, ainsi que 5 outils de prédiction de score de conservation. L'équipe ayant développé cette méta-base a également créé une base de donnée dédiée au génome entier, nommée WGSA [279], intégrant les outils implémentés dans dbNSFP ainsi que d'autres outils permettant l'interprétation de la partie non-exonique du génome (FATHMM [280], Eigen [281], CADD [275], LINSIGHT [282]).

En ce qui concerne les indels, peu d'outils existent. Cependant, l'outil CADD propose l'annotation des indels. À notre connaissance, il s'agit du seul outil permettant l'annotation à la fois des SNV non-synonymes, des SNVs non-exoniques et des indels. Pour ces raisons, cet outil a été utilisé pour l'annotation de tous les variants. Cependant, il est nécessaire d'utiliser un ensemble d'outil afin d'avoir une prédiction robuste. En ce sens, basé sur les outils disponibles dans les bases de données WGSA et dbNSFP et après étude de la littérature [283, 284], nous avons établi la liste des outils à utiliser afin d'annoter séparément les indels, les SNVs exoniques et les SNV non-exoniques. Ces trois ensembles d'outils, ainsi que les scores utilisés dans cette étude, sont référencés dans la Table 5.4.

5.1.3.3 Recherche de SNVs et indels rares et délétères

Un variant a été sélectionné et considéré comme rare si sa fréquence était inférieure à 1% dans une population européenne (hors population finlandaise) dans au moins une des bases de données présentées en Table 5.3. Par la suite, nous avons filtré les trois sous catégories de variants (SNV exonique, SNV non-exonique et indels) selon les scores de la Table 5.4, en considérant comme délétère les SNVs exoniques avec au moins 3 scores supérieurs aux seuils, les SNVs non-exoniques avec 4 scores supérieurs aux seuils et les indels dont le score Phred CADD était supérieur à 20. Enfin, nous avons uniquement conservé les variants présents chez

Table 5.4: Outils d'annotation de variants.

Outil	Seuil	Types de score
SNV exonique		
CADD [275]	≥ 20	Phred
Eigen [281]	≥ 20	Phred
MetaLR [277]	≥ 0.85	Rang
MetaSVM [277]	≥ 0.85	Rang
REVEL[286]	≥ 0.85	Rang
VEST4 [285]	≥ 0.85	Rang
SNV non-exonique		
CADD [275]	≥ 20	Phred
DANN [276]	≥ 0.99	Rang
Eigen_PC [281]	≥ 0.85	Rang
Fantom5 [288]	identifié	Annotation
FATHMM-XF [280]	≥ 0.99	Rang
Gerp++ [274]	≥ 4	Score
LINSIGHT [282]	≥ 90	Score
Phastcons46 [273]	≥ 0.99	Score
RegulomeDB [287]	= 2a,1a	Rang
Indels		
CADD [275]	≥ 20	Phred

tous les individus atteints d'une même famille, et absents des individus non atteints (seuls les variants sans phénocopie et avec une pénétrance complète dans les familles ont été sélectionnés).

Dans la mesure où nous disposions de données familiales, une analyse combinée association-liaison a été réalisée avec le programme pVAAST [289] afin de déterminer si les variants rares restants après les différentes étapes de filtres étaient significativement associés et/ou liés à la PR. Cette analyse a également été réalisée pour les gènes candidats (gènes incluant les variants rares). Ainsi, pour chaque variant rare ou chaque gène candidat, un score global $CLRp$ est obtenu en additionnant un score d'association (CLRv) et un score de liaison obtenu à l'aide de la méthode du lod-score (LOD) :

$$CLR_p = CLR_v + c \sum_{i=1}^{N} LOD_i$$

Score d'association CLRv

L'association des variants rares ou des gènes candidats a été évaluée à partir de tous les atteints de nos familles et de 98 individus de la population CEU du projet 1000 Genomes (individus utilisés comme témoins). Pour un gène candidat, les variants du gène sont divisés en K groupes : un groupe incluant tous les variants rares à risque, un groupe incluant tous les variants rares protecteurs et autant de groupes que de variants fréquents. Les variants rares sont inclus dans l'un des deux groupes sur la base des observations dans l'échantillon. Par exemple, si le variant rare est plus souvent observé chez les cas que chez les témoins, il sera classé dans la catégorie variant rare à risque. A l'inverse, si le variant rare est plus souvent observé chez les témoins que chez les cas, il sera classé dans la catégorie variant protecteur. Dans notre étude, tous les variants rares appartenaient à la catégorie à risque. Pour n individus analysés (nU non atteints et nA atteints), le score d'association $CLRv$ est calculé selon la formule suivante qui est basée sur un rapport de vraisemblance composite :

$$\lambda = \sum_{j=1}^{K} ln\left(w_j \times \frac{L_{Nulle}}{L_{Alt}}\right) = \sum_{j=1}^{K} ln\left[\frac{h_j}{a_j} \times \frac{(p_j)^{X_j}(1-p_j)^{2l_j n - X_j}}{(p_j^U)^{X_j^U}(1-p_j^U)^{2l_j n^U - X_j^U}(p_j^A)^{X_j^A}(1-p_j^A)^{2l_j n^A - X_j^A}}\right]$$

Ainsi, pour un variant ou groupe de variants j parmi les K possibles, les variables X_j , X_j^U et X_j^A représentent le nombre de copies de l'allèle mineur chez tous les individus, les non-atteints et les atteints respectivement, p_j , p_j^U et p_j^A est la fréquence de l'allèle mineur dans l'échantillon total, chez les non atteints et chez les atteints respectivement. Enfin, l_j est égal au nombre de variants dans le groupe j. Une pondération w_j, (égale à $\frac{h_j}{a_j}$), reflétant l'ampleur de l'effet du variant sur la fonction de la protéine, est appliquée au rapport de vraisemblance. Pour les groupes avec un seul variant, la variable h_j correspond à la probabilité que le changement d'acide aminé pour le variant j ne contribue pas au risque de développer la maladie ; pour les groupes avec plusieurs variants, h_j est la moyenne

des probabilités des variants du groupe j. De même, pour les groupes incluant un seul variant, la variable a_j est la probabilité que ce changement contribue à ce risque ; pour un groupe incluant plusieurs variants, c'est la moyenne des probabilités des variants du groupe j. Ces probabilités sont calculées en se basant sur la fréquence du changement d'acide aminé observé (AAS) et sur le score PhastCons du locus (mesure de la conservation phylogénétique du site de variation). La fréquence du AAS est déterminée à partir de la base de données OMIM pour a_j et à partir des témoins fournis à pVAAST pour h_j.

Score de liaison LOD

Le LOD score est un score de liaison qui permet d'évaluer s'il existe une liaison génétique entre 2 loci à l'aide de l'estimation du taux de recombinaison entre ces loci (il est alors possible de dire si ces loci sont proches l'un de l'autre sur le génome ou s'ils sont indépendants). Dans notre étude, le LOD score est calculé entre un de nos variants (dont la localisation est déjà connue) et le locus d'un gène de maladie (dont on cherche à déterminer la localisation par rapport à ce variant). Pour estimer le taux de recombinaison, les génotypes pour les 2 loci doivent être connus. Pour le variant d'intérêt, le génotype est connu grâce aux données de séquençage. Pour le gène de maladie, les génotypes des individus sont déterminés à partir d'un modèle génétique[2] prédéfini. Ici, un modèle dominant a été considéré et les valeurs de pénétrance et de phénocopie ont été estimées à partir de l'échantillon familial analysé. Pour un échantillon de n familles, le LOD est calculé de la manière suivante :

$$LOD = \sum_{i=1}^{n} log_{10}\left(\frac{max_{0\leq\theta\leq0.5}L_i(\theta)}{L_i(0.5)}\right)$$

Où la vraisemblance calculée pour la famille i $L_i(\theta) = \sum_{phases} P(phases) \times [(\theta)/2]^r \times [(1-\theta)/2]^{t-r}$ où r est le nombre de gamètes recombinés transmis par

[2]Le modèle génétique inclut le mode de transmission et les valeurs de pénétrance et de phénocopie

les parents double hétérozygotes sous une phase donnée et t est le nombre total de gamètes transmis par les parents double hétérozygotes. Et avec $L_i(\theta = 0.5)$ la vraisemblance attendue sous l'hypothèse nulle d'absence de liaison génétique ($\theta = 0.5$). Pour chaque gène, le LOD score est calculé à partir du variant ayant le CLRTv le plus grand.

Significativité du score CLRTp

La significativité du score CLRTp est évaluée à l'aide d'une procédure de ré-échantillonnage par randomisation [93]. Pour chaque échantillon familial, des individus sont sélectionnés au hasard parmi l'ensemble des individus disponibles (atteints et non-atteints des familles et témoins externes) pour servir de membres fondateurs aux familles (les structures familiales étant conservées). Puis, pour chaque membre de chaque famille, les génotypes sont générés via une procédure de gene-drop, i.e. en sélectionnant au hasard les allèles parmi ceux des parents sachant que l'on considère que chaque allèle peut être transmis de manière équiprobable. Bien que le calcul des scores ne permet pas d'ajuster sur des co-variables telles que le sexe des individus, il est possible d'introduire un biais dans la permutation afin de contrôler ces co-variables pour que leur répartition dans les jeux de permutations reflète la répartition de ces mêmes co-variables dans le jeu de données réelles. Ainsi, les analyses ont été réalisées en prenant en compte l'effet de la variable sexe. Un million de permutations a été réalisé pour évaluer la significativité du score.

5.1.3.4 Validation des SNVs et indels rares et délétères

Dans le but de valider les SNVs et indels identifiés, les régions exoniques des 15 individus du set de découverte et de 9 individus supplémentaires ont été séquencés, à l'exception d'un individu qui ne présentait pas une quantité d'ADN suffisante. Cette étape a un intérêt double, elle nous permet de valider la présence des variants

précédemment identifiés en re-séquençant les individus du set de découverte, ainsi que de caractériser les variants chez les autres individus des familles étudiées pour analyser leur ségrégation. Le séquençage des exomes a été effectué par le CNRGH en utilisant la plateforme Illumina HiSeq2000. L'individu C00247L avait une faible qualité d'ADN et l'individu B00JBTP une faible qualité de séquence (voir section 5.1). Ces deux individus ont donc été retirés de nos analyses réduisant le nombre d'individus total à 22 (13 du set de découverte et 9 supplémentaires). Ces 22 individus sont représentés avec des crochets entourant leur ID familial dans la Figure 5.1.

Les traitements et filtres appliqués sur les données post-séquençages ont été les mêmes que ceux présentés en Table 5.2. Les nouveaux résultats de génotypage des variants ont ensuite été croisés avec les résultats précédant afin de conserver les variants d'intérêt à l'aide de BCFtools.

Les variants présents chez tous les individus atteints d'une même famille, et absents des individus non atteints dans les familles étendues ont été conservés. L'outil pVAAST a ensuite été utilisé afin de vérifier l'association-liaison pour : 1. chaque gène incluant les variants rares, 2. chaque variant rare séparément et 3. chaque gène sans tenir compte du variant rare. Ainsi, il a été conservé les variants rares dont le score d'association était supérieur à 0 dans les analyses 1 et 2.

5.1.4 Recherche des CNVs candidats

Cette analyse a été réalisée à partir des lectures alignées, compressées et référencées au format BAM selon le pipeline détaillé en Figure 5.2.A. Afin d'identifier des CNVs, qui sont une sous-classe des variants structuraux (SVs) à partir de données WGS et après étude de la littérature [290–292], nous avons sélectionné trois outils : Manta [75], Lumpy [293] et CNVpytor [294] (extension de CNVnator dans le langage python [295]). Ces outils sont brièvement décrits ci-dessous.

5.1.4.1 Manta

Manta effectue une détection de variants structuraux en deux phases majeures en utilisant à la fois les lectures *paired* et *split* (Voir Figure 4.1). Lors de la première phase, le génome est analysé afin de créer un graphe intégrant toutes les cassures du génome à partir des lectures alignées des individus. Ce graphe est composé de nœuds représentant une région du génome où sont présentes une à plusieurs cassures et d'arêtes représentant l'évidence d'une jonction entre cassures. Ces arêtes ne sont pas spécifiques d'un hypothétique SV, mais représentent toutes les potentielles jonctions entre les différentes régions (faible spécificité) et permet à Manta la découverte via le graphe de tous les types de variants, tout en étant contenu dans une faible quantité de mémoire. Lors de la deuxième phase, chaque ensemble d'arêtes du graphe est considéré indépendamment et en parallèle pour la découverte de SVs. La découverte de SVs est réalisée en générant un SV candidat à partir des arêtes et nœuds du graphe précédemment créé (via les évidences), et des lectures associées à ces régions. Les SVs découverts sont assemblés puis alignés contre le génome de référence. Plusieurs filtres sont appliqués en interne, et se reflètent dans un score de qualité du variant fourni dans le fichier final.

5.1.4.2 Lumpy

Lumpy détecte des variants structuraux en utilisant plusieurs méthodes d'analyses incluant des méthodes *read-pair*, *split-read*, *read-depth* à partir d'un même jeu de données WGS ainsi qu'à partir de SVs connus (SVs provenant du 1000 Genomes). Pour l'utilisation des méthodes *split-read* et *read-pair*, Lumpy va extraire les lectures *split-read* et *paired-end* et appliquer un module spécifique à chaque type de lecture afin de détecter les points de rupture. La méthode de détection *read-depth* sera mise en œuvre via un outil de détection de CNV dédié et un module générique permettant également de détecter les points de rupture. Le

module générique sera également utilisé sur les SVs connus, permettant également d'établir des points de rupture. Enfin, Lumpy effectue un clustering des points de rupture identifiés par les différentes méthodes, résultant en une liste de régions des points de rupture prédites.

5.1.4.3 CNVpytor

CNVpytor est une implémentation en python de l'outil CNVnator, lui permettant d'être 2 à 20 fois plus rapide. CNVnator détecte les variants structuraux en utilisant la méthode *read-depth*. Cet outil effectue plusieurs étapes incluant l'analyse des fichiers d'alignements, le calcul et stockage interne de la profondeur de lecture par intervalles de 100pb, le pool des lectures par fenêtre d'intervalle (défini par l'utilisateur), la correction des données pour les biais en GC, la segmentation des données par la méthode *mean-shift* ainsi que la découverte des CNVs.

Les outils Manta 1.6.0, CNVpytor 1.0 et Lumpy via la suite Smoove 0.2.5 (outil simplifiant l'utilisation de Lumpy) ont été intégrés dans un pipeline Snakemake pour la découverte de variants structuraux. Les trois outils ont été utilisés dans leur configuration par défaut. La taille des fenêtres de lectures requises par l'outil CNVpytor a été définie sur 10000 pb (valeur intermédiaire proposée par l'outil). Les outils Manta et Lumpy étant également en capacité d'identifier tous les types de SVs, uniquement les CNVs ont été considérés pour ces outils.

5.1.4.4 Annotation des CNVs

L'annotation des CNVs a été réalisée avec l'outil AnnoSV 2.3.2, dont la méthodologie est détaillée dans la section 4.2.1.4.

5.1.4.5 Recherche de CNVs rares et délétères

Nous avons sélectionné et considéré un CNV rare si sa fréquence dans les bases GnomAD ou DGV était inférieure à 1% à partir de l'annotation de l'outil AnnoSV. De même, un variant a été considéré délétère et conservé si son score d'effet pathogène établi par AnnoSV était supérieur ou égal à 4 (Voir Table 4.8). Nous avons ensuite conservé les CNVs présents uniquement chez tous les individus atteints d'une même famille, et absents des individus non atteints du set de découverte. Les méthodes de détection des outils Lumpy et Manta étant similaires et le nombre de résultats trouvé par l'outil Manta étant conséquents, nous avons décidé de conserver uniquement les résultats communs entre ces deux outils. À l'inverse, CNVpytor utilisant une méthode différente (*read-depth* uniquement), cet outil a été filtré indépendamment. Deux filtres ont été appliqués sur les régions identifiées par cet outil : 1. Un score de qualité q0[3] pour une région identifiée inférieur à 50% et 2. Une pvalue inférieure à 5% à partir d'un test de Student par région identifiée en comparant la profondeur de lecture de la région d'intérêt et la profondeur de lecture moyenne des régions de tous les autres génomes de l'échantillon.

5.1.4.6 Validation des CNVs rares et délétères

La validation des CNVs identifiés à l'étape précédente a été réalisée dans le set de validation (n=25) par ddPCR (voir 4.2.1.6). Les sondes TaqMan utilisés en 3' et 5' de chaque CNV lors de la ddPCR ont été conçues en fonction des positions génomiques des CNVs.

[3]Pourcentage des lectures alignées avec une qualité nulle dans la région identifiée

5.1.5 Typage des allèles *HLA-DRB1*

Dans le but de connaître le statut *HLA-DRB1* des individus chez lesquels nous avons identifiés des variants candidats et dont nous possédions les séquences, leurs allèles ont été typés en utilisant l'outil HLA-HD [296]. Cet outil est brièvement décrit ci-dessous.

HLA-HD utilise une approche consistant en plusieurs phases. Les lectures sont dans un premier temps alignées à l'aide de Bowtie2 [297] contre les exons et introns d'un dictionnaire HLA composés de 31 675 allèles HLA à partir de la base de données IPD-IMGT/HLA (`http://www.ebi.ac.uk/ipd/imgt/hla/`), version 3.46.0 (2021-10) [298]. Par la suite, les lectures alignées remplissant les critères suivants sont conservées : au moins 50% de la taille de la lecture est alignée avec la cible et aucun mésappariement pour un exon ou maximum 2 mésappariement pour un intron. Afin de calculer un score pour chaque paire d'allèles, des scores pondérés sont attribués à chaque lecture en fonction du nombre d'alignement contre un allèle. Par la suite, pour chaque paire d'allèles est réalisé la somme pondérée des lectures dans le domaine G (domaine responsable de la présentation d'antigène). La paire d'allèles ayant le score le plus élevé est conservé. Le domaine G peut cependant correspondre à plusieurs combinaisons d'allèles dans certains cas, ne permettant donc pas de déterminer la paire d'allèles. Pour cela, l'outil ajoute des exons additionnels permettant de prédire la paire d'allèles avec plus de précision.

Les 15 individus du set de découverte ont été typés à partir des données de séquences WGS, tandis que les 22 individus du set de validation ont été typés à partir de données de séquence WES. Ainsi, 13 individus ont été typés deux fois.

5.2 Résultats

5.2.1 Recherche de SNVs et indels candidats

À partir des données WGS obtenues par le CNRGH pour 15 individus appartenant à 5 familles, nous disposions de 10 353 918 variants non filtrés incluant 8 365 829 SNVs et 1 988 089 indels. Nous avons ensuite procédé à des filtres successifs incluant des contrôles qualités (5.1.3.1), une fréquence faible (<1%), l'effet délétère des variants, la présence chez tous les individus atteints d'une même famille et absents des individus non atteints. L'effet des variants a ensuite été testé avec pVAAST. Les résultats après chaque étape du processus d'identification de variants de susceptibilité est détaillé dans la Figure 5.3. Au final, un total de 88 variants hétérozygotes incluant 37 SNVs non-exoniques, 47 SNVs exoniques et 4 indels exoniques a été identifié.

5.2.2 Validation des SNVs et indels candidats dans le set de validation

Dans le but de valider les variants dans les séquences codantes (ou proches des exons), un re-séquençage par WES dans le set de validation a été effectué. Les 47 SNVs exoniques les 4 indels identifiés précédemment ont été retrouvés. Cependant, uniquement 15 SNVs non-exoniques proches des exons ont été retrouvés dû à l'utilisation de la méthode WES. Après contrôle qualité des variants, 26 variants incluant 17 SNVs exoniques, 5 SNVs non-exoniques et 4 indels ont été conservés. Ensuite, la présence des variants chez tous les individus atteints d'une même famille et absents des individus non atteints a été vérifiée, et 15 variants répondaient toujours à ces critères incluant 9 SNVs exoniques, 4 SNVs non-exoniques et 2 indels. Ces 15 variants étaient répartis dans 15 gènes, soit un variant par gène.

Par la suite nous avons utilisé pVAAST afin de tester l'association-liaison des variants/gènes avec la PR et procédé à trois tests : 1. En incluant tous les variants dans le gène, 2. En incluant uniquement le variant d'intérêt dans le gène et 3. En excluant le variant d'intérêt dans le gène. Ainsi, 12 gènes significatifs (p-value $\leq$ 0.05) ont été identifiés en respectant ces critères détaillés en Table 5.5. Tous les variants seuls étaient associés à la PR (p-value ≤ 0.05) à l'exception de la délétion chr5:141024690 (*FCHSD1*) pour laquelle, sans celle-ci, le gène était toujours associé avec la PR. Ces 12 variants de susceptibilité validés par nos analyses sont répartis dans 4 des 5 familles, détaillés en Table 5.6.

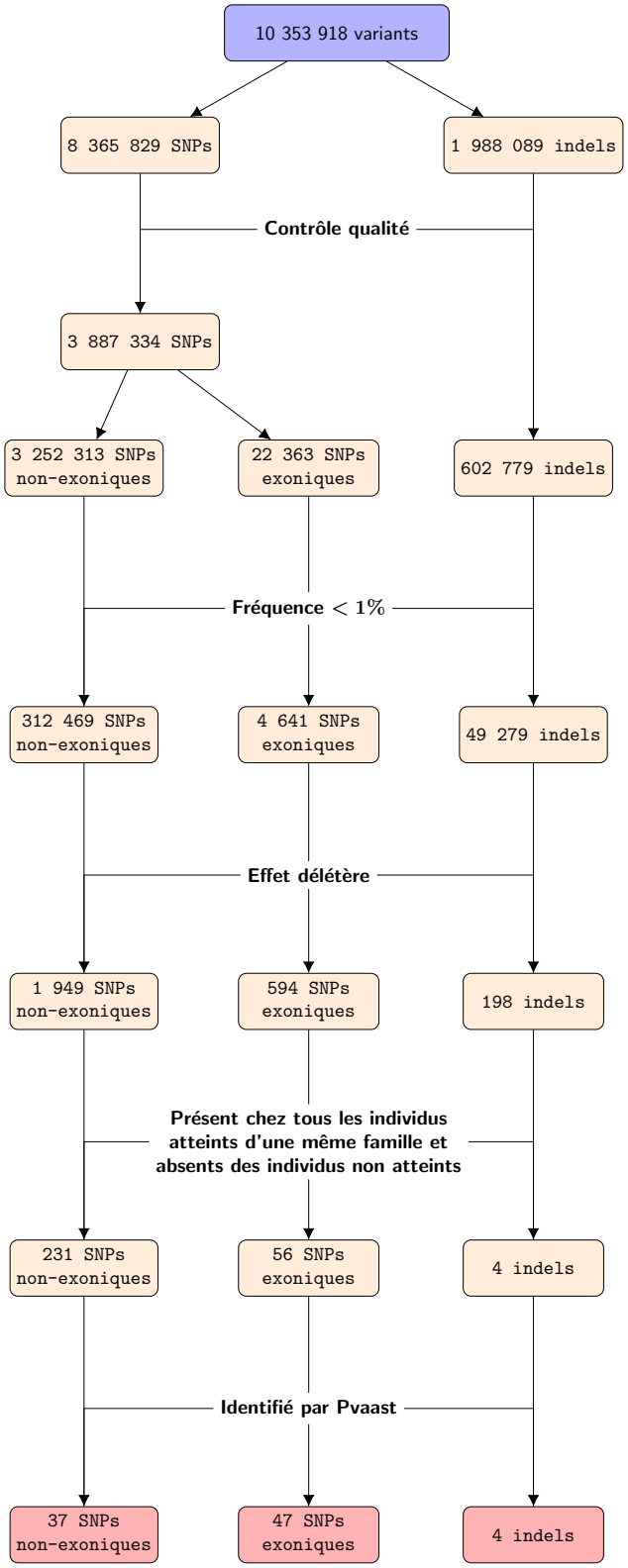

Figure 5.3: Résultats des différentes étapes permettant l'identification des 88 variants obtenus par séquençage de génome entier dans le set de découverte.

Table 5.5: Résultats du test d'association-liaison (pVAAST) sur les variants et gènes candidats.

Gène		Variant majeur (VM)				Hors VM
Nom	**p-value (score)**	**Position**	**Ref**	**Alt**	**p-value (score)**	**p-value (score)**
SRRM4	0.00583 (2.77)	chr12:119419656	G	A	0.00646 (2.77)	1 (0)
NAGK	0.0022 (9.91)	chr2:71298831	G	A	0.00102 (16.05)	1 (0)
KCNMA1	0.0052 (2.77)	chr10:78644841	G	A	0.00528 (2.77)	1 (0)
PRKAR1B	0.0074 (5.714)	chr7:635796	C	T	0.00716 (5.714)	1 (0)
PCSK9	0.00661 (10.83)	chr1:55518374	C	T	0.00457 (10.83)	1 (0)
PROC	0.0048 (7.54)	chr2:128180699	T	C	0.00528 (7.54)	1 (0)
MYH1	0.000948 (12.43)	chr17:10412807	C	T	0.000968 (12.43)	1 (0)
FGD6	0.00208 (10.48)	chr12:95603246	G	A	0.00216 (10.48)	1 (0)
RYR1	0.000948 (9.4)	chr19:38987562	G	A	0.00102 (9.4)	1 (0)
QARS	0.00551 (10.39)	chr3:49141888	C	A	0.00638 (10.39)	1 (0)
FCHSD1	0.000998 (17.876)	chr5:141024690	AAGTC	A	0.000998 (16.05)	0.0473 (5.472)
CAPN2	0.00102 (16.05)	chr1:223900575	C	CCACGGTAGGAAGCG	0.000938 (9.41)	1 (0)

Note:

Position: Génome GRCh37/hg19

Table 5.6: Résultats du filtrage des variants de susceptibilité dans les individus supplémentaires.

Gène	Position	Ref	Alt	Localisation	nº Famille	# PR	# Non PR
PCSK9	chr1:55518374	C	T	Exonique	2	2/2	0/1
CAPN2	chr1:223900575	C	CCACGGTAGGAAGCG	Exonique	4	2/2	0/3
NAGK	chr2:71298831	G	A	Exonique	1	2/2	0/2
PROC	chr2:128180699	T	C	Exonique	2	2/2	0/1
QARS	chr3:49141888	C	A	Exonique	2	2/2	0/1
FCHSD1	chr5:141024690	AAGTC	A	Exonique	4	2/2	0/3
PRKAR1B	chr7:635796	C	T	Exonique	5	1/1	0/3
KCNMA1	chr10:78644841	G	A	Intronique	2	2/2	0/1
FGD6	chr12:95603246	G	A	Exonique	1	2/2	0/2
SRRM4	chr12:119419656	G	A	Intronique	2	2/2	0/1
MYH1	chr17:10412807	C	T	Exonique	4	2/2	0/3
RYR1	chr19:38987562	G	A	Exonique	4	2/2	0/3

Note:
Position: Génome GRCh37/hg19
PR, non PR : Nombre d'individus atteints et non atteints avec variant sur le total d'atteints et non atteints respectivement dans les familles du set de validation

5.2.3 Recherche des CNVs candidats

Les outils Lumpy, Manta et CNVpytor ont respectivement identifié 8 737, 9 996 et 14 574 CNVs en utilisant les données de séquence au format BAM. Nous avons ensuite effectué plusieurs étapes successives pour l'identification de CNV candidats, détaillées dans la Figure 5.4. Ces étapes ont abouti à l'identification de 8 CNVs communs des outils Manta et Lumpy ainsi qu'à l'identification de 2 CNVs par l'outil CNVpytor. Ces CNVs, incluent 1 CNV commun aux trois outils, portant le nombre de CNVs candidats identifiés à 9, présentés dans la Table 5.7.

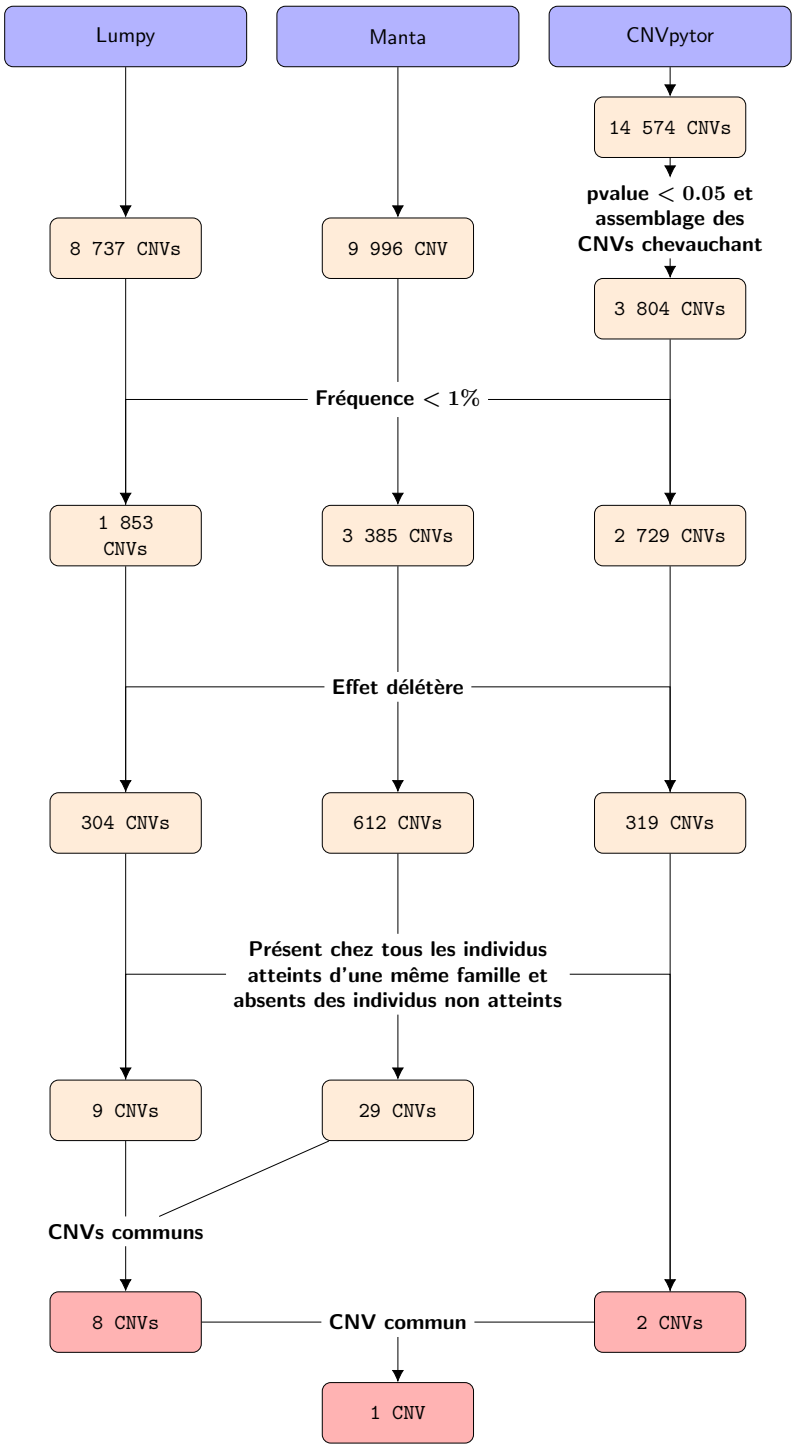

Figure 5.4: Résultats des différentes étapes pour l'identification des CNVs obtenus par séquençage de génome entier dans le set de découverte.

Table 5.7: Résultats des différentes étapes pour l'identification des CNVs dans le set de découverte.

Gène	Position	Type	Localisation	Outils	n° Familles	# PR	# Non PR
BCL11A	chr2:60700139-60700185	Del	Intron	MA & LU	3 & 5	4/4	0/2
FER	chr5:108153984-108162362	Del	Intron	MA, LU & CN	5	2/2	0/1
NKAIN2	chr6:124432501-124471000	Dup	Intron & Exon	CN	3	2/2	0/1
DGKZ	chr11:46380878-46382112	Del	Intron	MA & LU	4	2/2	0/1
SIPA1L1	chr14:71935690-71935954	Del	Intron	MA & LU	1	2/2	0/1
NUP93	chr16:56831102-56831377	Del	Intron	MA & LU	2	2/2	0/1
STRADA	chr17:61803133-61807340	Dup	Intron & Exon	MA & LU	4	2/2	0/1
QRICH2	chr17:74280757-74282070	Del	Intron	MA & LU	1	2/2	0/1
PCDH19	chrX:99567671-99567741	Del	Intron	MA & LU	5	2/2	0/1

Note:
Position : Génome GRCh37/hg19
Dup : Duplication
Del : Délétion
CN : CNVpytor
MA : Manta
LU : Lumpy

5.2.4 Validation des CNVs candidats dans le set de validation

À ce jour, la validation de 8 des 9 CNVs a été réalisée dans le set de validation par ddPCR (le CNV inclus dans le gènes *QRICH2* reste à valider). Les résultats de cette validation sont présentés dans la Table 5.8.

La totalité des CNVs a été validée chez les individus où ils avaient été identifiés par les outils de détection de CNVs. Les résultats ont montré que 5 CNVs avaient une pénétrance incomplète et/ou présence de phénocopie. Cependant, parmi ces 5 CNVs, celui touchant le gène *BCL11A* et présent dans deux familles, montre une absence de phénocopie dans les deux familles et une pénétrance complète dans l'une d'entre elles. Enfin, 3 CNVs (*DGKZ*, *SIPA1L1* et *NUP93*) avaient une pénétrance complète et n'ont pas montré de phénocopie.

Table 5.8: Résultats de l'identification des CNVs dans le set de validation

Gène	Position	nº Familles	# PR	# Non PR
BCL11A	chr2:60700139-60700185	3 & 5	6/6	2/5
FER	chr5:108153984-108162362	5	2/2	2/3
NKAIN2	chr6:124432501-124471000	3	3/4	2/3
DGKZ	chr11:46380878-46382112	4	2/2	0/3
SIPA1L1	chr14:71935690-71935954	1	2/2	0/3
NUP93	chr16:56831102-56831377	2	2/2	0/3
STRADA	chr17:61803133-61807340	4	1/2	1/3
PCDH19	chrX:99567671-99567741	5	2/2	1/3

Note :
Position : Génome GRCh37/hg19
PR, non PR : Nombre d'individus atteints et non atteints avec CNV sur le total d'atteints et non atteints respectivement dans les familles du set de validation

5.2.5 Typage de *HLA-DRB1*

Le typage et la connaissance des allèles *HLA-DRB1* chez les individus inclus dans notre étude est une étape importante. En effet, si l'allèle à risque SE est absent de nos individus, nous nous attendons à pouvoir mettre en évidence d'autres facteurs génétiques impliqués dans la PR (et non facteurs modulant l'effet de HLA-DRB1).

Table 1.4 permet d'avoir la liste de tous les allèles SE de HLA-DRB1 sur les positions 70-74, allèles à risque dans la PR.

Préalablement au séquençage, nous disposions des génotypes de *HLA-DRB1* pour 5 individus. Ces génotypes ont été obtenues par une méthode d'amplification avec amorce séquence spécifique (SSP-PCR). Parmi ces individus, 4 étaient présents dans le set de découverte et n'avaient pas d'allèle *HLA-DRB1* (B00JBTM, B00JBTN, B00JBTS et B00JBTT) et 1 individu du set de validation (B00G39I) avait deux allèles à risque *HLA-DRB1* (allèles 04:01 et 04:05).

L'outil HLA-HD [296] a été utilisé dans un premier temps sur les séquences génomes-entiers de 15 individus du set de découverte afin de confirmer les allèles *HLA-DRB1* des 4 individus préalablement connus ainsi que d'identifier le statut *HLA-DRB1* des 10 autres individus dont les résultats sont présentés dans la Table 5.9 (analyse WGS). Les résultats obtenus confirment une absence d'allèles SE pour les 4 individus (B00JBTM, B00JBTN, B00JBTO et B00JBTP). Le typage de novo des 11 autres individus a révélé un allèle à risque SE chez 2 individus (B00FZ7F et B00JBTR), deux allèles à risque SE chez 2 individus (B00G39H et B00G39J) tandis que les 7 autres individus n'étaient pas porteurs d'allèle à risque SE.

Dans un second temps, cet outil a été utilisé sur les séquences exoniques des 22 individus du set de validation (composés des 13 individus du set de découverte et 9 individus supplémentaires). Les résultats de cette analyse sont présentés en

Table 5.9 (analyse WES). Cette analyse nous a permis de confirmer les deux allèles des 13 individus (individus ayant deux allèles surlignés en gras dans la Table 5.9). Parmi ces 13 individus, 2 présentaient une différence de typage entre l'analyse des séquences WES et WGS (15:01:01 et 15:01:44 respectivement pour B00JBTU et 03:01:01 et 03:147 respectivement pour B00JBTV). Ces résultats ont cependant été considérés comme validés, la différence entre ces allèles respectifs est due à la substitution d'un nucléotide, identifiées récemment [299, 300].
D'autre part, parmi les 9 individus supplémentaires, 3 individus avait un allèle SE (C00247K, C00247M, C00247N) et 2 individus avait deux allèles SE (B00G39I, C00247J) tandis que 4 individus n'avaient aucun allèle SE. L'individu B00G39I est le cinquième individu pour lequel nous disposions de l'information des génotypes, où un allèle a été confirmé (04:01:01) tandis que le second était différent (04:05 par génotypage moléculaire et 04:04:01 avec l'outil HLA-HD).

Au final, parmi les 5 familles inclues dans notre étude, 2 présentaient une ségrégation d'allèles SE du gène *HLA-DRB1*.

Table 5.9: Typage des allèles HLA-DRB1 de 24 individus (15 du set de découverte et 9 individus supplémentaires).

Famille	Analyse	Individu	Allèle 1	Allèle 2	Motif SE	Typage moléculaire connu
1	WGS	B00JBTJ	***03:01:01**	***13:02:01**	-	
		B00JBTN	***03:01:01**	***13:02:01**	-	✓
		B00JBTV	***03:01:01**	***15:01:01**	-	
	WES	B00JBTI	*03:01:01	*15:01:01	-	
2	WGS	B00JBTL	*07:01:01	*09:01:02	-	
		B00JBTO	***09:01:02**	***15:01:01**	-	✓
		B00JBTS	***07:01:01**	***09:01:02**	-	
	WES	B00JBTK	*07:01:01	*15:01:01	-	
3	WGS	B00FZ7F	***04:04:01**	***13:01:01**	+	
		B00G39H	***04:04:01**	***04:04:01**	+	
		B00G39J	***04:01:01**	***04:04:01**	+	
	WES	B00G39I	*04:01:01	*04:04:01	+	✓
		C00247J	*04:04:01	*04:04:01	+	
		C00247K	*04:04:01	*13:01:01	+	
4	WGS	B00JBTM	***07:01:01**	***07:01:01**	-	✓
		B00JBTR	***01:02:01**	***07:01:01**	+	
		B00JBTU	***07:01:01**	***15:01:01**	-	
	WES	C00247M	*01:02:01	*07:01:01	+	
		C00247N	*01:02:01	*15:01:01	+	
5	WGS	B00JBTP	*03:01:01	*11:04:01	-	✓
		B00JBTQ	***03:01:01**	***11:04:01**	-	
		B00JBTT	***03:01:01**	***13:03:01**	-	
	WES	C00247I	*03:01:01	*13:03:01	-	
		C00247L	*11:04:01	*15:01:01	-	

Note :
SE : Shared Epitope
WES : Whole Exome Sequencing
WGS : Whole Genome Sequencing
Allèles en gras : Allèles identifiés en WGS puis confirmés en WES

5.2.6 Variants candidats

Le bilan des résultats incluant les variants candidats rares, délétères, sans phénocopie et avec pénétrance complète de type SNV, indel et CNV, ainsi que le statut HLA-DRB1 des familles dans lequel chaque variant a été identifié est présenté dans la table 5.10. Ces résultats incluent 10 SNVs (8 exoniques, 2 introniques), 2 indels exoniques et 3 CNVs introniques, soit un total de 15 variants. Cinq variants étaient trouvés dans la famille 4, où des allèles *HLA-DRB1 SE* ségrégeaient, tandis que les 10 autres variants étaient identifiés dans des familles où aucun allèle *HLA-DRB1 SE* n'a été observé.

Table 5.10: Identification de nouveaux facteurs génétiques rares et délétères, sans phénocopie et avec pénétrance complète associés à la PR

Gène	Position	Localisation	nº Familles	# PR	# Non PR	Motif SE
SNV						
PCSK9	chr1:55518374	Exonique	2	2/2	0/1	-
NAGK	chr2:71298831	Exonique	1	2/2	0/2	-
PROC	chr2:128180699	Exonique	2	2/2	0/1	-
QARS	chr3:49141888	Exonique	2	2/2	0/1	-
PRKAR1B	chr7:635796	Exonique	5	1/1	0/3	-
KCNMA1	chr10:78644841	Intronique	2	2/2	0/1	-
FGD6	chr12:95603246	Exonique	1	2/2	0/2	-
SRRM4	chr12:119419656	Intronique	2	2/2	0/1	-
MYH1	chr17:10412807	Exonique	4	2/2	0/3	+
RYR1	chr19:38987562	Exonique	4	2/2	0/3	+
Indel						
CAPN2	chr1:223900575	Exonique	4	2/2	0/3	+
FCHSD1	chr5:141024690	Exonique	4	2/2	0/3	+
CNV						
DGKZ	chr11:46380878-46382112	Intronique	4	2/2	0/3	+
SIPA1L1	chr14:71935690-71935954	Intronique	1	2/2	0/3	-
NUP93	chr16:56831102-56831377	Intronique	2	2/2	0/3	-

Note :
Position : Génome GRCh37/hg19
PR, # non PR : Nombre d'individus atteints et non-atteints dans les familles où les variants ont été identifiés
Motif SE : un + indique la présence d'allèle *HLA-DRB1 SE* qui ségrège dans la famille, tandis qu'un - indique qu'aucun allèle *HLA-DRB1 SE* ne ségrège dans la famille

5.3 Discussion

Au cours de cette étude, nous avons identifié 15 variants hétérozygotes délétères rares et spécifiques de la PR (10 SNVs, 2 indels et 3 CNVs), présentés en Table 5.10. Ces 15 variants présentent une pénétrance complète et pas de phénocopie dans les familles étudiées.

Une étude de la littérature des gènes incluant ces variants a montré que 7 gènes (4 avec SNVs, 1 avec indel et 2 avec CNVs) pouvaient avoir une implication dans la physiopathologie de la PR. L'effet de ces 7 gènes et leur potentielle implication dans la PR sont brièvement détaillés ci-dessous.

1. Le gène *KCNMA1* (localisé en 10:78629359-79398353[4]), aussi appelé KCa1.1, pour lequel un SNV a été identifié en 3' UTR, code pour une protéine exprimée au niveau de la membrane cellulaire. Celle-ci est responsable de l'activation de canaux potassiques lors de la dépolarisation de la membrane ou d'une augmentation de Ca2+ cytosolique permettant l'exportation de K(+) [301, 302]. Plusieurs études ont démontré que cette protéine était un régulateur essentiel des synoviocytes de type fibroblastique (FLS), notamment sur leur interaction avec les lymphocytes T dans la PR [303], et, d'autre part sur leur propriété migratoire invasive dans la PR [304, 305]. En effet, une inhibition de KCNMA1 augmenterait l'adhésion des FLS aux ligands via les intégrines β1 exprimées à leurs surfaces. Cette inhibition induirait alors une augmentation de l'expression des intégrines β1 [305]. D'autre part, la cascade de signalisation impliquant la régulation des intégrines β1 impliquerait AKT et ca2+, également impliqué dans la PR [305].

[4]Position sur le génome GRCh37/hg19

2. Le gène *PCSK9* (localisé en 1:55505221-55530525[5]), pour lequel une substitution non synonyme a été identifiée dans l'exon 5, code pour une protéine ayant un rôle crucial dans la régulation de l'homéostasie du cholestérol plasmatique, ciblant et dégradant le récepteur des lipoprotéines de basse densité (LDLR), ce qui entraîne une augmentation du niveau de LDL (lipoprotéine de basse densité) [306]. Ce gène est ainsi indirectement lié à l'immunité innée, puisque les lipoprotéines plasmatiques participent à la défense contre une infection bactérienne [306]. Cette protéine est également responsable d'effets pro-inflammatoires (notamment dans les macrophages) [307, 308]. Celle-ci a été retrouvée impliquée dans une maladie chronique auto-immune, le lupus érythémateux disséminé (LED) [310]. Enfin, une récente étude a identifié qu'un niveau d'expression faible du gène PCSK9 était observé chez des patients en rémission de la PR après traitement anti-TNF-α. PCSK9 stimulerait alors une production de cytokines pro-inflammatoires dans les macrophages et les FLS [311].

3. Le gène *CAPN2* (localisé en 1:223889295-223963720[5]), aussi connu sous les noms de m-calpain ou calpain-2, pour lequel une insertion a été identifiée dans l'exon 1, code pour une protéine présente dans une variété de cellules, incluant les macrophages, monocytes et fibroblastes. Cette protéine appartient à un groupe de protéases à cystéine sensible au calcium, exprimées de manière ubiquitaire. Cette protéine a été retrouvée 3.5 fois plus présente dans les fluides articulaires de patients atteints de PR que dans ceux de témoins [312]. D'autre part, elle serait secrétée à partir des FLS et responsable de la destruction de la matrice extracellulaire du cartilage dans la PR [313, 314].

[5]Position sur le génome GRCh37/hg19

4. Le gène *PROC* (localisé en 2:128176003-128186822[6]), aussi connu sous le nom de protéine C activée, pour lequel une substitution non synonyme a été identifiée dans l'exon 5, code pour une protéine ayant un rôle clé dans la régulation de la coagulation du sang [315]. Indépendamment de son action sur la coagulation, cette protéine exerce un large éventail d'actions cytoprotectrices incluant la stabilisation et la suppression de l'inflammation des barrières épithéliales et endothéliales [316]. D'autre part, l'utilisation à des fins thérapeutiques de cette protéine sur des modèles murins a démontré son efficacité sur des maladies immunitaires inflammatoires telles que le diabète de type 1 [317], le lupus [318] et la PR [316]. Dans le cadre de la PR, la protéine avait alors une action réduisant le développement et l'apparition de la maladie en supprimant l'inflammation et l'invasion des FLS.

5. Le gène *DGKZ* (localisé en 11:46354455-46402104[6]), pour lequel une délétion de 1.235kb a été identifiée dans l'intron 1, code pour une protéine contrôlant une variété de processus cellulaires et agit en tant que substrat dans la synthèse de nombreuses molécules lipidiques [319, 320]. L'un des processus dans lequel cette protéine est impliquée inclut l'activation de cellules T en régulant négativement les voies de signalisations des récepteurs T, qui sont en partie médiées par le diacyglycerol [321]. Ces résultats suggèrent que l'activation de DGKZ permet de réduire la réponse pro-inflammatoire [322]. Une étude de modèle murin a montré que DGKZ permettait de réguler la réponse des macrophages dans l'arthrite chronique juvénile [322]. Une autre étude a identifié ce gène différentiellement exprimé chez des patients atteints de PR par rapport à des témoins [323], tandis qu'une seconde étude a identifié ce gène différentiellement méthylé en comparant des jumeaux (monozygotes) atteints de PR et des jumeaux témoins non atteints de PR [324]. D'autre

[6]Position sur le génome GRCh37/hg19

part, nous avons également identifié un variant rare de type SNV dans le gène DGKZ (chr11:46396584[7]) au cours de l'étude du WGS, qui n'a pas été retenu car uniquement identifié comme délétère par 2 outils (notre seuil étant à 3). Ce SNV a également été validé lors de l'étude de validation des variants par WES dans le set de validation. Ainsi, ce SNV est présent chez les mêmes individus que la délétion identifiée. Néanmoins, nous n'avons pas la possibilité d'évaluer le déséquilibre de liaison entre le CNV et le SNV de manière aisée.

6. Le gène *PRKAR1B* (localisé en 7:588834-767287[7]), pour lequel une substitution non synonyme a été identifiée dans l'exon 7, code pour une protéine kinase qui est impliquée dans la voie de signalisation de l'adénosine monophosphate cyclique (cAMP). La protéine PRKAR1B est responsable de la régulation des lipides et du métabolisme du glucose. Les sites CpGs du gène PRKAR1B ont été retrouvés différentiellement méthylés chez des patients atteints de PR par étude de cellules lymphocytes T CD4+ et de FLS [325], indiquant une régulation spécifique associée à la PR.

7. Le gène *NUP93* (localisé en 16:56764017-56878797[7]), pour lequel une délétion de 276b a été identifiée dans l'intron 2, code pour une protéine nucléoporine, qui est un composant principal du pore nucléaire. Cette protéine cible des caspases qui ont un rôle central dans la mort cellulaire programmée par apoptose. Des travaux antérieurs au laboratoire ont montré que ce gène est différentiellement exprimé chez des patients atteints de PR [326]. D'autre part, il est à noter que, dans le cadre d'une analyse sur l'implication de NUP93 dans la hyalinose segmentaire et focale (maladie rénale rare) une équipe a mis en évidence une double mutation sur le gène NUP93 pour un

[7]Position sur le génome GRCh37/hg19

individu qui avait aussi développé la PR [327].

D'autre part, parmi les variants ne faisant pas partie de la liste finale des 15 variants, le variant touchant le gène BCL11A a retenu notre attention, car, celui-ci est présent chez deux familles différentes, malgré une pénétrance incomplète dans l'une d'entre elles. Ce gène code pour un facteur de transcription, qui est un composant essentiel dans la régulation du devenir des lymphocytes B [328]. Des travaux récents effectués au laboratoire ont par ailleurs identifié BCL11A comme un facteur de transcription (TF) clé, en utilisant l'inférence de réseaux sur des données NGS de FLS issues de patients atteints de PR [329].

Ainsi, malgré la taille limitée de l'échantillon analysé (pour le set de découverte et pour le set de validation), les critères très stricts de sélection des variants (variants rares délétères avec pénétrance complète et absence de phénocopie) ont permis d'identifier des variants potentiellement candidats. Des études supplémentaires seraient cependant nécessaires pour conclure sur l'effet de ces variants dans le développement de la PR.

Si 10 SNVs et 2 indels ont été caractérisés dans l'échantillon de validation, il faut noter que nous disposions au départ d'une liste de 88 variants d'intérêts rares et délétères après séquençage du génome entier du set de découverte. Le choix d'utiliser la méthode WES afin de confirmer les SNVs et indels d'intérêts nous a limité dans la validation de SNVs non-exoniques (37 SNVs étaient localisés dans les introns ou régions intergéniques et uniquement 15 SNVs proches des exons ont pu être identifiés dans les données WES). Par ailleurs 30 variants exoniques et 10 variants non-exoniques n'ont pas passé le contrôle qualité, rejeté par le critère MQ (Qualité d'alignement moyen sur l'ensemble de la lecture) sur les données WES de validation. Une des perspectives de cette étude serait de valider ces autres variants par une autre approche.

Nous avons également identifié 3 CNVs qui respectaient nos critères de sélection après validation par ddPCR dans le set de validation. La validation par ddPCR des CNVs a également permis de confirmer les résultats trouvés par les outils bio-informatiques utilisés (Manta, Lumpy et CNVpytor). Dans cette étude, il a été décidé de se focaliser sur les SVs de type CNVs. Néanmoins, les outils Manta et Lumpy sont capables de détecter tous les types de SVs et une inversion touchant le gène NRXN1, non encore validée, a été identifiée par les outils Manta et Lumpy et respectait tous les critères de sélection.

En conclusion, malgré les effectifs limités des sets de découverte et de validation, des variants intéressants ont pu être mis en évidence. Des études fonctionnelles seraient cependant nécessaires pour confirmer l'effet de ces variants.

Chapitre 6

Inférence d'un réseau global intégratif spécifique de la Polyarthrite Rhumatoïde

Alors qu'un grand nombre de facteurs génétiques ont été identifiés par analyse de données génomiques, cela ne permet pas d'expliquer entièrement l'étiologie de la PR, qui implique des processus biologiques complexes. Par exemple, les thérapies actuelles pour la PR incluent l'utilisation de Disease-modifying antirheumatic drugs (DMARDs), ciblant les protéines du système immunitaire. Environ 90% des patients traités reçoivent un traitement anti-TNF [151] et 30 à 40% des patients recevant ce traitement n'y répondent pas [155]. Face à la complexité d'une telle maladie, l'utilisation simultanée de plusieurs couches d'informations est essentielle afin d'en mieux comprendre ses mécanismes. La biologie des systèmes permet de réduire la complexité de l'utilisation de ces couches à l'échelle d'un système. En utilisant des approches d'inférence de réseaux par intégration de données multi-omiques [159–162]. Cela permet alors d'obtenir une vue globale des voies de signalisation moléculaires impliquées dans la maladie et pourrait permettre de mieux comprendre

l'inefficacité de ces traitements chez certains patients.

Par exemple, un nombre conséquent d'approches de biologie computationnelle, principalement basées sur l'inférence de réseaux intégrant des données multi-omiques (protéomiques, génomiques, transcriptomiques et métabolomiques), ont permis de mieux comprendre des mécanismes clés dans des maladies complexes [159–162]. Ces données peuvent aussi être intégrées par des outils utilisant des méthodes de machine learning [163, 330]. D'autre part, ces méthodologies, associées à l'utilisation de connaissances biologiques préalables (i.e carte moléculaire, littérature...), permettraient d'obtenir des résultats plus pertinents [164–166].
Enfin, l'utilisation de la modélisation booléenne sur des réseaux inférés a permis de décrire la dynamique des cellules humaines telles que les signaux de transductions et la régulation des gènes [174–179], mais aussi la dérégulation des gènes dans les maladies tels que le cancer [180, 181].

Dans ce chapitre, nous proposons l'étude de mécanismes potentiellement impliqués dans la PR, en utilisant l'inférence et l'analyse dynamique d'un réseau global, spécifique de la PR. Pour cela, nous avons construit dans un premier temps un réseau, en utilisant des données transcriptomiques publiques de patients atteints de PR combiné avec une carte moléculaire de la PR précédemment construite au laboratoire [170]. Ce réseau a été notamment étudié afin d'identifier des facteurs de transcription clés impliqués dans la PR. Dans un deuxième temps, nous avons mis en évidence des molécules clés dans la PR, en utilisant le réseau inféré et en y associant : les variants génomiques identifiés dans le chapitre précédent (Chapitre 5), des variants génomiques publiques et des données transcriptomiques publiques provenant de patients ayant reçu des traitements anti-TNF. Ces résultats ont ensuite été analysés via des simulations *in silico*, permettant de mettre en évidence des mécanismes clés dans la PR.

6.1 Matériel et Méthodes

6.1.1 Description et filtres des données pour l'inférence d'un réseau de co-régulation

Dans cette étude, nous avons utilisé un jeu de données publiques obtenu à partir de l'analyse de leucocyte provenant de patients atteints de PR et d'individus témoins (GSE117769), [séquencé par Commonwealth Serum Laboratories (CSL) Limited/bio21 Institute (30 Flemington Rd, Parkville, Australia) via la plateforme Illumina HiSeq 2500 (Illumina, Inc)]. Ce jeu de données est composé de 120 individus (51 atteints de PR, 50 témoins, 19 atteints de spondylarthrite ankylosante ou rhumatisme psoriasique). À partir de celui-ci il a été extrait 46 individus atteints de PR (5 individus avaient des origines asiatiques et n'ont pas été inclus) et 50 témoins, soit un total de 96 individus. Parmi les individus atteints de PR, 43 avaient des ancêtres d'origines caucasiennes tandis que 3 individus avaient des ancêtres d'origine inconnue. D'autre part, le jeu de données contenait un réplicat pour un individu témoin, dont nous avons conservé l'expression moyenne. Une analyse préliminaire a été conduite avec la matrice d'expression des individus et l'outil DESeq2 1.32.0 [331], en utilisant une normalisation ainsi qu'une transformation de la variance [332]. Cette transformation logarithmique permet notamment d'atténuer l'effet de la variance sur la moyenne des données d'expression. À partir des données transformées, une Analyse en Composante Principale (PCA) a été réalisée, révélant cinq individus (4 atteints et 1 non-atteint) avec des valeurs aberrantes, qui ont été retirés du jeu de données (voir Figure 6.1). Enfin, nous avons conservé uniquement les gènes dont le nombre de lectures étaient supérieur à 10.

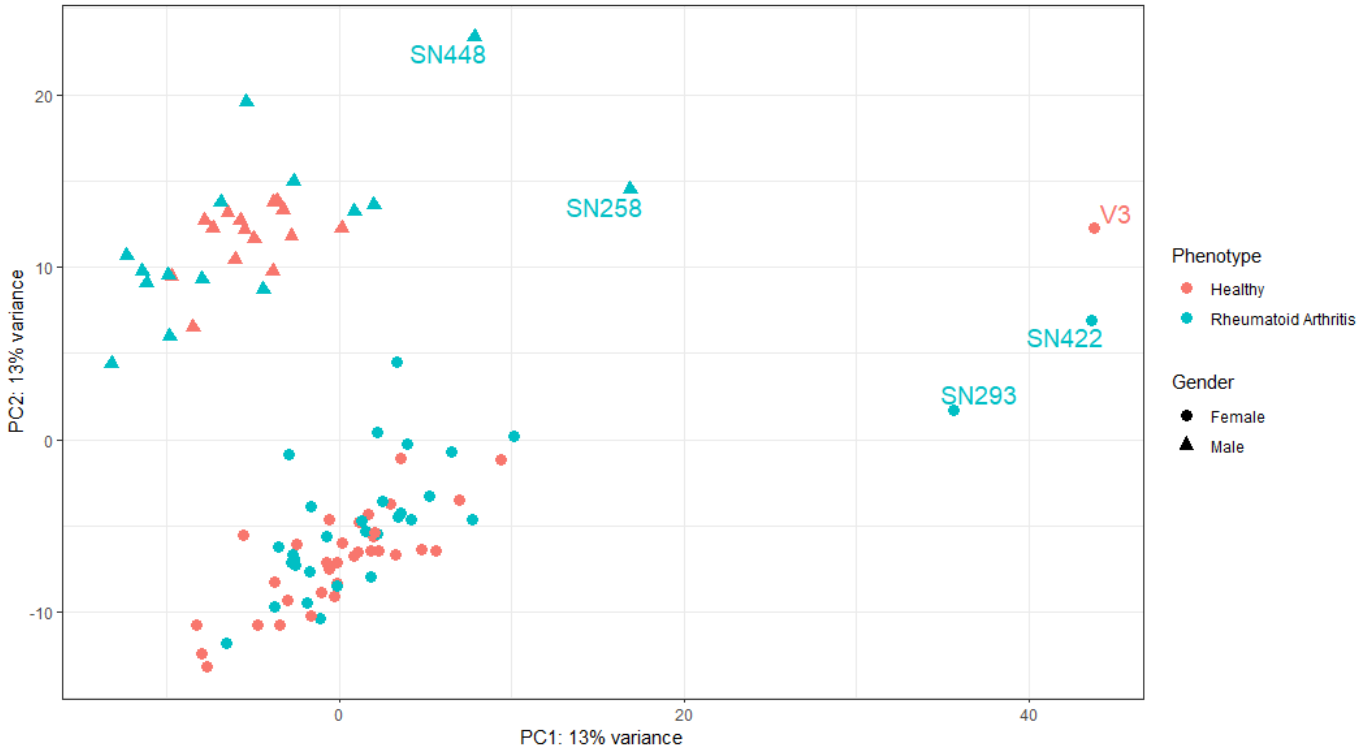

Figure 6.1: Analyse en Composante Principale (PCA) réalisée à partir de données d'expression provenant des leucocytes de 95 individus (46 atteints de PR et 49 témoins). Tiré de Miagoux et al. 2021 [333]. Une transformation stabilisante de la variance [332] a été réalisée sur les données d'expression.

6.1.2 Inférence d'un réseau de co-régulation

Afin d'inférer un réseau de co-régulation, nous avons utilisé le package R CoRegNet (1.26.0) [163] sur les données obtenues à l'étape précédente. Cet outil est brièvement décrit ci-dessous.

CoRegNet implémente une méthode hybride d'inférence nommée H-LICORN [334]. Celle-ci repose sur l'algorithme LICORN [335] et vise à identifier deux types de régulations : 1) Les régulations entre les facteurs de transcriptions (TFs) et les gènes. 2) les régulations coopératrices entre les TFs. Dans un premier

temps, l'algorithme LICORN est utilisé par CoRegNet sur les données d'expression transcriptomiques normalisées, où chaque gène g est discrétisé en trois valeurs $S_g \in \{-1, 0, 1\}$ représentant respectivement le statut sous-, normalement ou sur-exprimé du gène. Plusieurs réseaux de régulations (GRNs) candidats sont établis à partir de ces données en utilisant dans un premier temps l'algorithme *apriori* [336] afin d'identifier des potentiels co-régulateurs (e.g. TFs). Puis, dans un second temps, l'algorithme LICORN établira des co-activateurs et co-inhibiteurs candidats. CoRegNet utilise ensuite l'algorithme H-LICORN afin de sélectionner un GRN à partir des GRNs candidats obtenus par l'algorithme LICORN. Pour cela, chaque GRN est soumis à une régression linéaire où le réseau ayant la plus faible erreur prédite sera sélectionné. Ce GRN peut être enrichi par CoRegNet, en ajoutant de l'information provenant de connexions établies entre TFs à partir de données ChIP-Seq et d'interactions protéine-protéine depuis les bases de données CHEA [337], ENCODE [338], HIPPIE [339] et STRING [340]. À partir de ces informations ajoutées, CoRegNet permet d'affiner le réseau en utilisant un algorithme de sélection intégratif [341]. Enfin, la significativité de la régulation coopératrice entre deux TFs est testée en utilisant un test exact de Fisher basé sur le nombre de gènes communs régulés par la paire de TFs.

Le réseau de co-régulation inféré avec CoRegNet a été enrichi puis affiné par une méthode non supervisée utilisant une moyenne non pondérée. Enfin, à partir du réseau de co-régulation, nous avons conservé les régulations coopératrices significatives entre les TFs ($False\ Discovery\ Rate\ (FDR) < 5\%$).

6.1.3 Extraction des protéines de la carte moléculaire de la PR

Dans le but de créer un réseau global pour la PR, il est essentiel d'ajouter de l'information spécifique de la PR à notre réseau de co-régulation. Dans ce but, une carte moléculaire de la PR [170] construite au sein de notre laboratoire a été utilisée.

Cette carte moléculaire de la PR est une base de connaissances interactive de la maladie disponible sur la plateforme MINERVA [342]. Cette carte est représentée sous forme de cellule, où nous retrouvons un flux d'informations parcourant les divers composants de la cellule : espace extracellulaire, membrane plasmique, cytoplasme, noyau. À partir de ce flux d'informations, il est possible d'identifier les phénotypes cellulaires associés à la PR. Notre objectif était d'identifier les TFs communs entre la carte moléculaire de la PR et notre réseau de co-régulation. Pour cela, nous avons superposé les TFs identifiés par CoRegNet sur la carte moléculaire de la PR et ainsi identifié ceux chevauchants. Notre analyse s'est ensuite concentrée sur les régulateurs en amont des TFs chevauchants dans la carte moléculaire de la PR. À cet égard, un plugin MINERVA a été utilisé [342] permettant d'extraire les TFs chevauchants ainsi que l'ensemble des protéines en amont. Le fichier obtenu depuis MINERVA a ensuite été exporté au format XML CellDesigner [343]. Ce fichier a été vérifié via l'outil CellDesigner, où les réactions de traduction liant des ARNm du noyau à des protéines dans le cytoplasme et la membrane ont été retirées, notre objectif étant de se concentrer uniquement sur les protéines en amont des TFs chevauchants et localisées dans le cytoplasme et l'espace extra-cellulaire.

Le fichier extrait à partir de la carte moléculaire de la PR était dans un langage nommé *Processus de Description* (*Process Description*, PD), langage créé par la communauté scientifique *Notation graphique de la biologie des systèmes*

(*Systems Biology Graphical Notation*, SBGN) [344] (voir Figure 6.2). Ce langage très descriptif est idéal afin d'avoir une représentation schématique proche des voies métaboliques et des voies de régulation retrouvées dans la littérature. Cependant, il était essentiel d'avoir un langage simplifié afin de réaliser nos analyses. Pour cela, le langage *Flux d'activité* (*Activity Flow* AF) est plus adapté, plus simpliste (Voir Figure 6.2) et également développé par le SBGN. Dans ce but, nous avons utilisé l'outil *CaSQ* v0.9.11 [345], permettant de passer d'un langage PD à un langage AF, en convertissant le format XML du fichier en un format SBML-qual et permettant aussi d'ajouter des règles booléennes au fichier. *CaSQ* produit également un fichier de type *format d'interaction simple* (*Simple Interaction Format*, SIF).

Le fichier au format SIF produit par *CaSQ* a été récupéré puis ajusté en utilisant un programme R créé pour l'occasion, afin de simplifier les complexes obtenus via la carte moléculaire. Pour cela, le réseau au format SIF contenait des complexes incluant plusieurs molécules ne pouvant pas être retrouvés dans notre réseau de co-régulation ou aucun complexe n'était présent. Cependant, les molécules contenues dans certains complexes l'étaient. Nous avons alors créé un parent et une interaction vers le complexe pour chaque molécule constituant un complexe, seulement si ce parent n'existait pas déjà. Ainsi, un chevauchement entre les gènes/protéines des deux réseaux a été possible. Afin de simplifier notre réseau nous avons ensuite conservé uniquement une entité (gène/protéine) lorsque celle-ci était présente au moins deux fois et toutes ses interactions.

6.1.4 Création d'un réseau global intégratif et spécifique de la PR

Nous avons utilisé le package R *igraph* [347] afin de convertir le réseau de co-régulation *CoRegNet* ainsi que le réseau issu de l'extraction de la carte moléculaire de la PR en deux réseaux distincts mais uniformes. Ensuite, nous avons fusionné

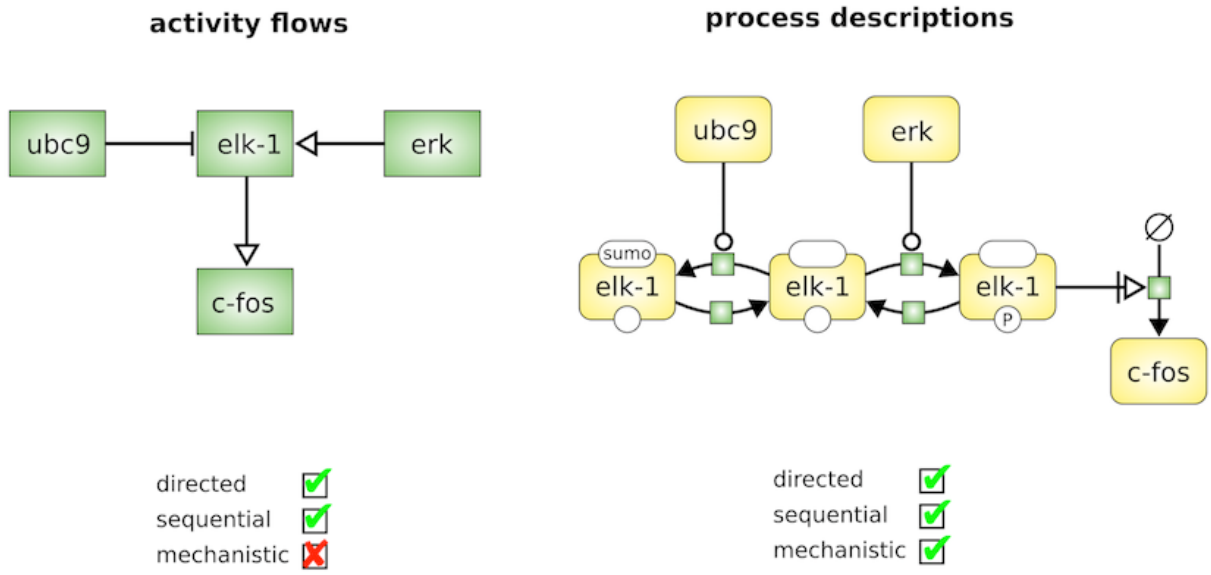

Figure 6.2: Représentation d'un système biologique en deux langages de notation graphique de la biologie des systèmes : Flux d'activités et processus de description. Adapté de Le Novere et al. 2015 [346]

les deux réseaux à l'aide d'*igraph*, puis les avons importé dans *Cytoscape* [348] en utilisant le package Bioconductor RCy3[1] [349], formant le réseau global et spécifique de la PR.

6.1.5 Listes de variants de susceptibilité

Dans un premier temps, les 15 variants rares et délétères associés à des gènes identifiés dans le Chapitre 5 ont été utilisés. Ces variants ont été superposés avec le réseau global et spécifique de la PR en utilisant Cytoscape.

Dans un second temps, une liste de variants publiques provenant de la base de données DisGeNET [350] a été utilisée. Cette base de données recense plus de 24 000 maladies associées à des gènes ou variants identifiés à partir de source multiple incluant la littérature. À partir de DisGeNET, nous avons extrait les variants ayant une association variant-maladie (VDA) et un index d'évidence (EI) supérieurs à

[1]Plugin dédié à l'importation de réseau depuis R vers *Cytoscape*

0.7. Le VDA est calculé à partir de publications vérifiées et non vérifiées associant le variant avec la maladie. L'EI permet, quant à lui, de mettre en évidence les publications contradictoires concernant une association du variant/gène avec la maladie. En choisissant un seuil de 0.7 pour le VDA et l'EI, chaque variant avait au moins une publication supportant l'association du variant avec la maladie, ainsi qu'un faible nombre de publications contradictoires. Au total, 1635 variants répondaient à ces critères. Parmi ces variants, 731 étaient associés avec un gène et ont été conservés. Ces 731 gènes ont ensuite été superposés avec la carte spécifique de la PR en utilisant Cytoscape.

6.1.6 Analyse d'expression différentielle (DEA) dans deux jeux de données indépendants

Deux analyses d'expression différentielle (DEA) ont été réalisées en utilisant deux jeux de données différents. Le premier est composé de données d'expression ARN normalisées provenant de cellules lymphocytaires T CD4+ incluant des patients atteints de PR selon leurs réponses à des traitements anti-TNF (GSE138747). Ce jeu de données est composé de deux cohortes analysées indépendamment de patients atteints de PR traités avec de l'adalimumab (37 patients) ou étanercept (41 patients). Le second jeu de données comprend des données d'expression non normalisées d'ARN provenant d'hématocytes à partir de patients atteints de PR patients n'ayant jamais reçu de traitement, prélevés à T0 ainsi qu'après trois mois de traitements à l'infliximab ou adalimumab. Ce jeu de données est composé de deux cohortes de 40 et 36 patients atteints de PR qui ont été analysés indépendamment.

En utilisant DESeq2, nous avons conduit deux DEA sur les niveaux d'expression des gènes en comparant les individus répondeurs et non-répondeurs pour les deux médicaments (adalimumab et etanercept). Puis deux autres DEA sur les niveaux d'expression des gènes normalisés avec DESeq2 en comparant les patients à T0 et

après trois mois de traitement. Nous avons considéré un gène différentiellement exprimé (DEG) lorsque sa p-value corrigée (FDR) était inférieure à 0.1. Ces listes de DEG furent par la suite superposées sur le réseau global et spécifique de la PR en utilisant Cytoscape.

6.1.7 Extraction d'un sous-réseau pour étudier la réponse au traitement

Le sous-réseau a été créé à partir de trois molécules hautement impliquées dans la PR :

1. Le facteur de nécrose tumorale (TNF)
2. L'interleukine 6 (IL6)
3. Le facteur de croissance de transformation beta 1 (TGFB1)

Ces trois molécules ont été extraites via le réseau ainsi que l'ensemble des protéines impliquées en aval jusqu'au premier TF, afin de réduire la complexité du réseau et d'axer notre analyse sur les régulateurs en amont des TFs identifiés. Pour l'extraction, nous avons utilisé le plugin *Biological Network Manager* (BiNoM) [351] et sélectionné de manière progressive les voisins en aval des molécules TGFB1, IL6 et TNF jusqu'au premier TF(s) rencontré(s).

6.1.8 Shiny App

L'ensemble des réseaux (réseau de co-régulation, l'extraction de la carte moléculaire de la PR au format AF, le réseau global et spécifique de la PR et le sous-réseau) présenté dans ce chapitre ainsi que la superposition de gènes provenant des différents jeux de données transcriptomiques et de la liste de variants génomiques publique ont été intégrés dans une application en ligne en utilisant Shiny app [352] et R. L'application en ligne intègre une visualisation *Cytoscape* via le package

R *cyjShiny* [353]. Cette application est librement accessible (`https://quentin-miagoux.shinyapps.io/global_ra_network`).

6.1.9 Modélisation booléenne

À partir du sous-réseau et des perturbations identifiées via la superposition de gènes provenant des différents jeux de données transcriptomiques et de la liste de variants génomiques publique, nous avons recherché à évaluer l'impact de ces perturbations de façon isolée et combinée, sur l'expression des TFs en sortie de notre réseau (JUN, JUND, FOS et NFKBIA). Pour cela, le sous-réseau a alors été exporté, depuis *Cytoscape* dans un format SBML en utilisant le plugin *BiNoM* et sa fonction "export to SBML". Ensuite, les interactions de co-régulation entre les TFs ont été retirées via *CellDesigner*, étant donné que ce ne sont pas des interactions complexes, celles-ci ne peuvent pas être utilisées dans un modèle booléen. Afin d'ajouter des règles booléennes à notre réseau pour les simulations, le sous-réseau a été transformé en un format SBML-Qual en utilisant *CaSQ*. À partir de ce réseau, quatre analyses ont été réalisées : 1. Analyse par simulation en temps réel, 2. Analyse de sensibilité, 3. Analyse dose-effet et 4. Le calcul des états stables et la simulation *in silico* par KO. L'ensemble des étapes mentionnées ci-dessus est résumé Figure 6.3. Pour réaliser les trois premières analyses (analyse par simulation en temps réel, analyse de sensibilité et analyse dose-effet), représentées en bleu dans la Figure 6.3, le fichier SBML-Qual a été importé sur *Cell Collective* [354]. L'analyse de simulation en temps réel permet d'analyser distinctement l'impact des molécules en entrée (IL6, TGFB1 et TNF) sur le niveau d'activité des molécules en sortie (JUN, JUND, FOS, NFKBIA et NFKB1) en fonction du temps et des différentes perturbations précédemment identifiées (avant/après traitement, répondeurs non-répondeurs et gènes mutés). Cette analyse a été réalisée en utilisant l'onglet *Simulation* de Cell Collective où l'activation des trois molécules (activité

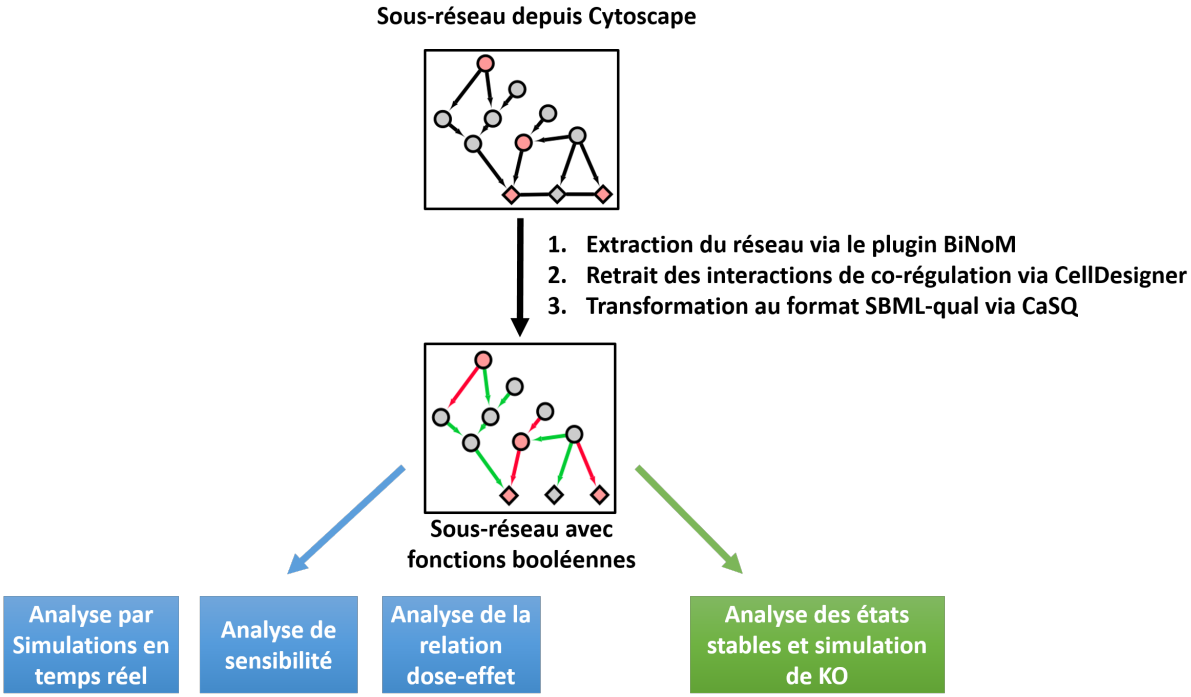

Figure 6.3: Représentation des différentes étapes pour l'obtention d'un sous-réseau contenant des règles booléennes et son analyse par simulation *in silico*

à 100%) a été étudiée séparément, et où les perturbations ont été reproduites séparément (activité à 0%). Au total, 9 analyses de simulation ont été réalisées.

L'analyse dose-effet, permet de répliquer l'injection d'un médicament sur le niveau d'activité du sous-réseau. Dans notre cas, les molécules IL6, TGFB1 et TNF, connues pour être des cibles thérapeutiques, ont été considérées comme médicament afin d'analyser l'effet sur les TFs en sortie de notre réseau. Cette analyse de relation dose-effet a été réalisée en utilisant cinq conditions initiales différentes, décrites dans la Table 6.1.

L'analyse de sensibilité permet d'étudier directement l'activité des molécules en sortie du réseau à partir de changement d'activité des molécules en entrée. Cette analyse nous permet d'étudier l'impact d'autres voies n'impliquant pas TNF, afin d'observer si un mécanisme de compensation se met en place permettant d'expliquer pourquoi des patients ne répondent pas à des traitements anti-TNF. Pour cela,

les molécules en sortie du réseau ont été testées (JUN, JUND, FOS, NFKBIA et NFKB1). Enfin, l'analyse des états stables et simulation de KO, permet d'analyser

Table 6.1: Conditions initiales pour l'analyse de relation dose-effet.

Conditions initiales	État
1	Tous inactifs
2	IL6 actif
3	TGFB1 actif
4	IL6 + TGFB1 actifs
5	TNF actif

l'activité des molécules en entrée (IL6, TGFB1 et TNF) et le cheminement de cette activation dans le sous-réseau dans sa globalité, tout en prennant en compte les perturbations précédemment identifiées (avant/après traitement, répondeurs non-répondeurs et gènes mutés). Dans ce but, Le fichier SBML-Qual a été également utilisé avec l'outil *GINsim* 3.0.0b-SNAPSHOT [355], présenté en vert dans la Figure 6.3. Afin de permettre à *GINsim* de reconnaître les noms des molécules contenues dans notre fichier, nous avons utilisé le mode *recognition* de GINsim. Ensuite, GINsim a été utilisé afin de réduire la complexité du réseau via *Reduce model*. Les états stables de notre réseau ont été calculés en utilisant la fonction *Compute stable states*. Nous avons ensuite calculé les états stables en reproduisant les perturbations potentiellement induites par les composants des jeux de données (avant/après traitement, répondeurs non-répondeurs et gènes mutés). Pour cela, la fonction *Run simulation* a été utilisée en configurant les perturbations comme des Knock-Out (KO) *in silico* (voir détail Table 6.2).

Table 6.2: Conditions initiales pour l'analyse des états stables du réseau booléen.

Conditions initiales	Knock-Out (KO)
1	MAPK1 et MAPK14
2	DAXX, ILK et MAP2K1
3	DAXX et NFKB1

6.2 Résultats

6.2.1 Inférence d'un réseau de co-régulation

Un jeu de données transcriptomique obtenu à partir de la base de données GEO (GSE117769) a été utilisé afin d'inférer le réseau de co-régulation. Ce jeu de données inclut 120 individus (51 atteints de PR, 50 témoins et 19 atteints de spondylarthrite ankylosante ou rhumatisme psoriasique). Après plusieurs étapes de filtre incluant les origines des individus, les duplicats, et la qualité des données, nous avons conservé 90 individus (48 témoins et 42 patients atteints de PR). À partir de ces individus, nous avons normalisé les données en utilisant DESeq2 à partir desquelles CoRegNet a été utilisé pour inférer le réseau de co-régulation, présenté en Figure 6.4.

Le réseau de co-régulation inféré inclut un total de 19 TFs, 14 interactions de co-régulation ainsi que de 373 gènes cibles des TFs. Après analyse des TFs et leurs interactions, nous avons établi le top 5 des TFs ayant le plus d'interactions de régulation et/ou de co-régulation, présenté en Table 6.3. Par la suite, il a été réalisé une étude de la littérature à partir des 19 TFs du réseau de co-régulation pour mettre en évidence leur potentielle implication avec la PR. Les résultats de cette étude sont présentés en Annexe A résumant le rôle clé de chaque TF ainsi que les études correspondantes où ils ont été identifiés. L'étude de la littérature pour ces 19 TFs a notamment révélé que la totalité de ces TFs ont été retrouvés impliqués dans la PR ou dans des voies impliquant le système immunitaire.

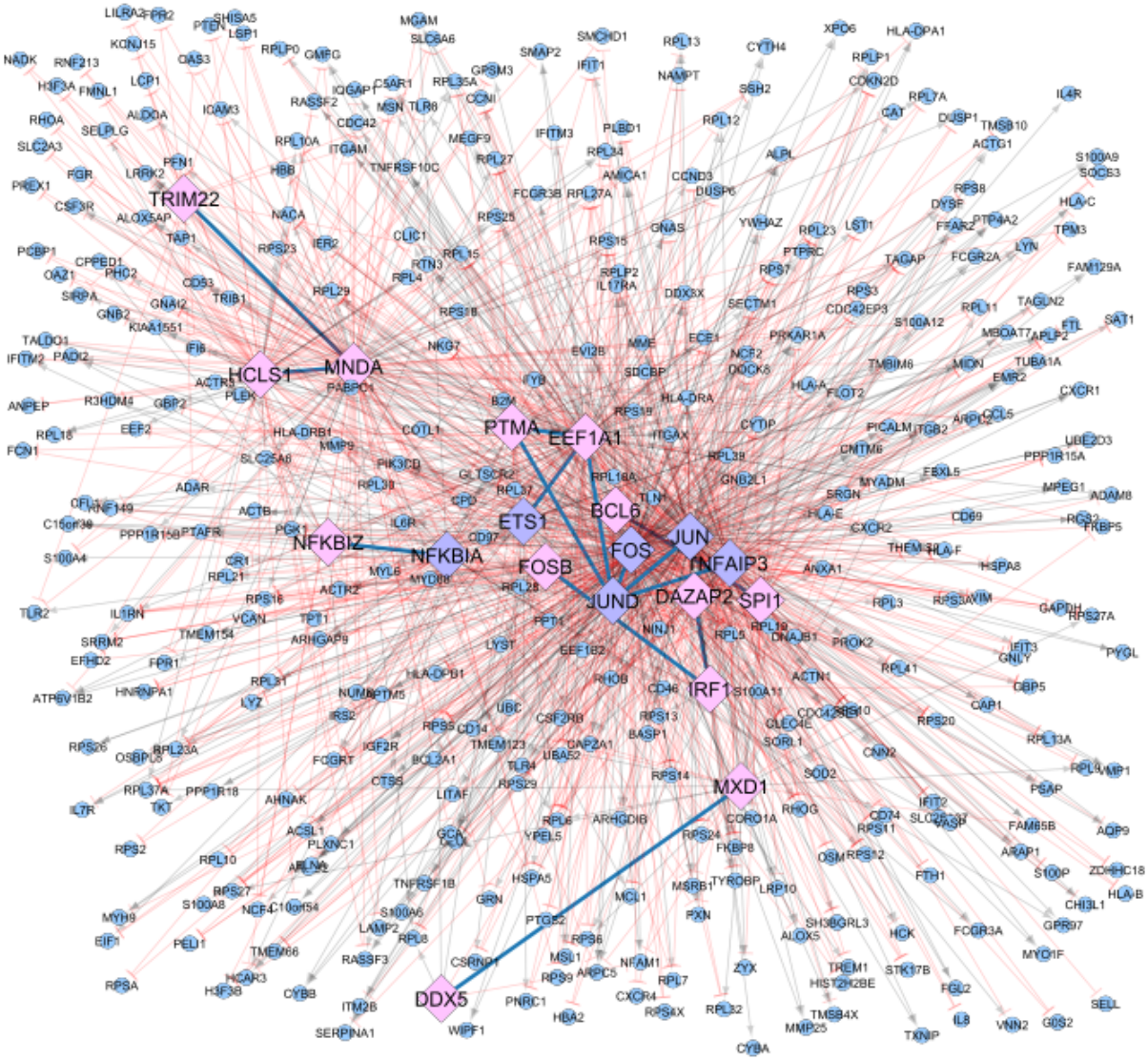

Figure 6.4: Réseau de co-régulation inféré en utilisant CoRegNet et les données normalisées provenant du jeu de données transcriptomiques (GSE117769). Le jeu de données inclut 90 individus (48 témoins et 42 atteints de PR). Les facteurs de transcription communs avec la carte moléculaire de la PR sont représentés par un losange et une couleur violette (6) tandis que les facteurs de transcription non communs sont représentés par un losange et une couleur rose (13). Les gènes cibles des facteurs de transcription sont représentés par un cercle et de forme bleue (373). Une inhibition est représentée par une flèche plate (⟞) rouge, une activation est représentée par une flèche grise (⟶) et une interaction de co-régulation est représentée par une barre sans flèche (—)

Table 6.3: Top 5 des facteurs de transcription (TF) identifiés par CoRegNet ayant le plus d'interaction de regulation (interaction TF-gène cible) et interaction de régulation (TF-TF).

Top 5 TFs	**Nombre d'interactions**
FOS	288
JUN	211
EEF1A1	155
MNDA	136
TNFAIP3	125

Top 5 TFs	**Nombre d'interactions de co-régulation**
JUND	5 *(EEF1A1, FOS, JUN, PTMA, TNFAIP3)*
EEF1A1	3 *(ETS1, JUND, PTMA)*
IRF1	2 *(DAZAP2, FOSB)*
MNDA	2 *(HCLS1)*
PTMA	2 *(EEF1A1 JUND)*

6.2.2 Extraction de la carte moléculaire de la PR

À partir des 19 TFs identifiés précédemment dans le réseau de co-régulation, 6 TFs ont été retrouvés dans la carte moléculaire de la PR : ETS1, FOS, JUN, JUND, NFKBIA, TNFAIP3. Les 6 TFs ont été utilisés comme points d'entrées pour l'extraction de la carte moléculaire de la PR. À partir de ces TFs, il a été extrait tous les régulateurs en amont connectés à ces TFs. Le réseau extrait de la carte moléculaire de la PR à partir des TFs inclut 244 molécules est présenté en Figure 6.5.

6.2.3 Réseau global intégratif et spécifique de la PR

Le réseau global intégratif et spécifique de la PR est le résultat du couplage de deux réseaux précédemment créés, le réseau de co-régulation obtenu avec CoRegNet ainsi que l'extraction de la carte moléculaire de la PR réalisée à partir des TFs communs avec le réseau de co-régulation. Ce réseau comprend 614 molécules et 1736 interactions (848 inhibitions, 874 activations et 14 interactions de co-régulation) incluant des gènes, protéines, complexes protéiques et des molécules simples, présenté dans la Figure 6.6. Le réseau global et spécifique de la PR inclut 6 TFs communs entre les réseaux CoRegNet et la carte moléculaire de la PR. De plus, 16 gènes cibles des TFs indentifiés par l'outil CoRegNet sont également retrouvés dans la carte moléculaire de la PR extraite.

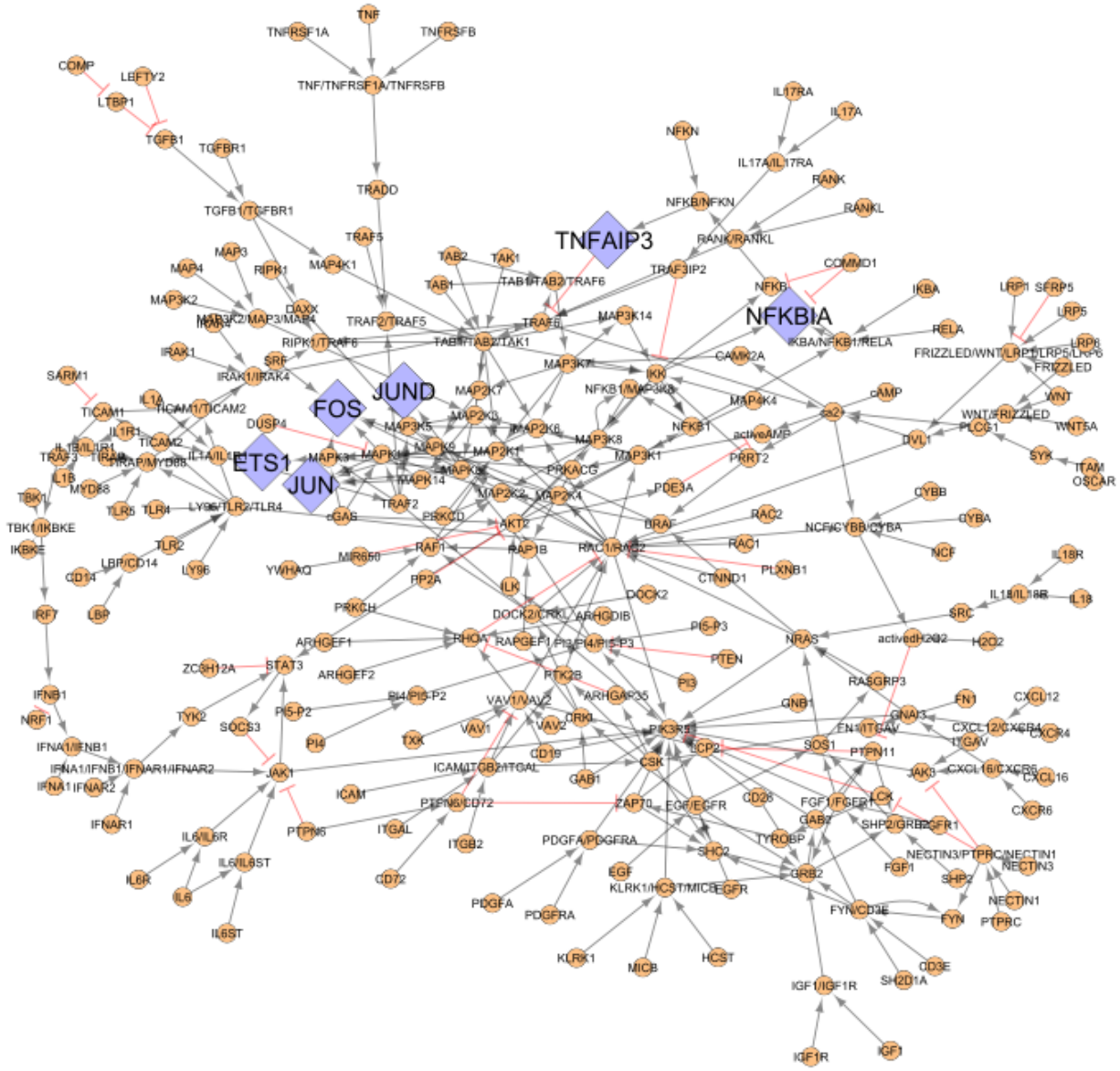

Figure 6.5: Réseau de régulation créé à partir des TFs communs entre la carte moléculaire de la PR et le réseau de co-régulation CoRegNet
Les 6 TFs communs avec CoRegNet sont représentés en violet (forme losange), tandis que les autres régulateurs sont représentés en orange (forme circulaire). Une inhibition est représentée par une flèche plate (—⊣) rouge tandis qu'une activation est représentée par une flèche grise (⟶).

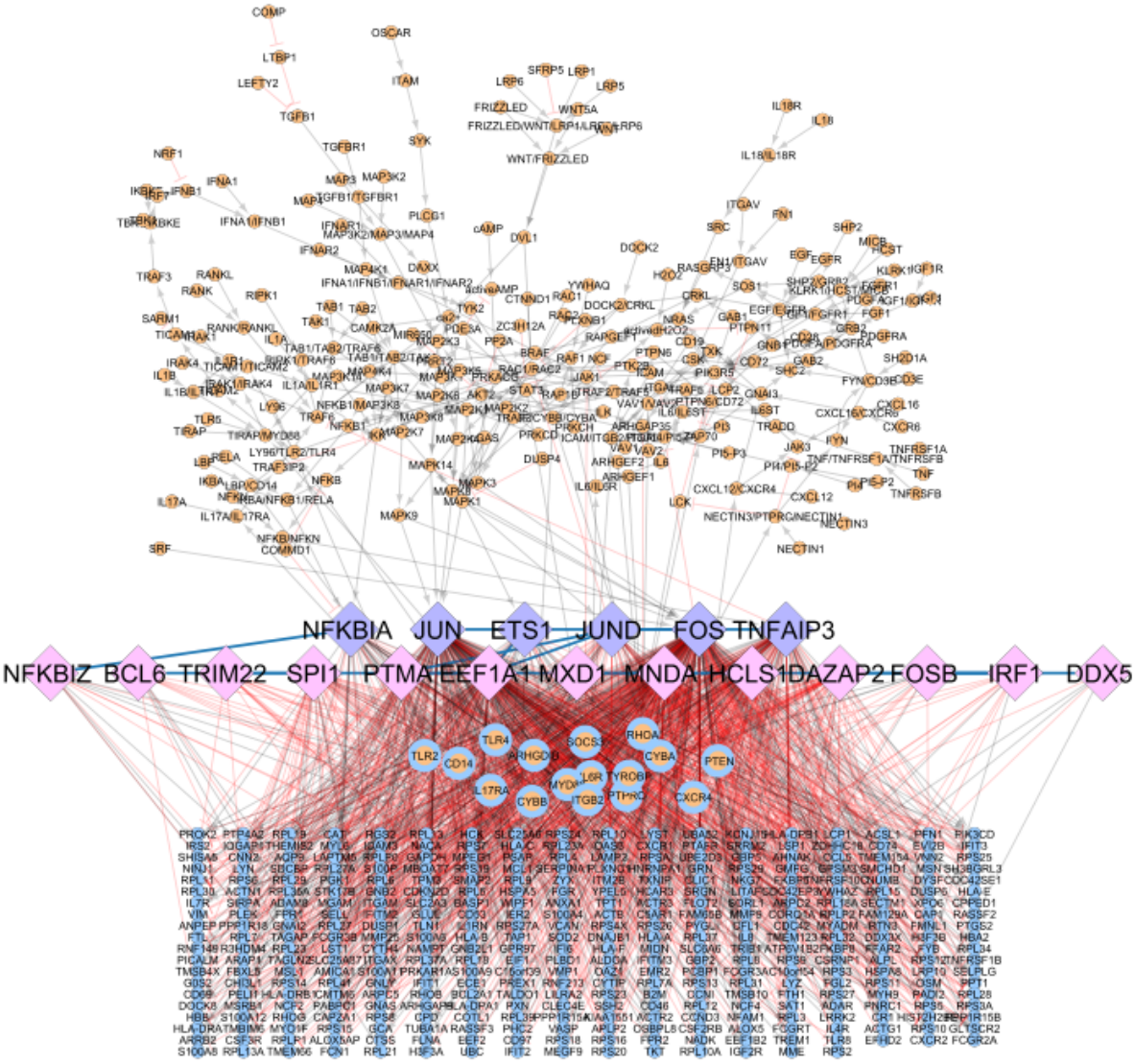

Figure 6.6: Réseau global intégratif et spécifique de la PR] Les 6 TFs communs sont représentés en violet (forme losange), les 16 gènes/protéines cibles sont représentés en bleu et en orange, les 222 protéines provenant de la carte moléculaire de la PR sont représentés en orange, les 13 TFs spécifique du réseau CoRegNet sont représentés en rose et les 357 gènes cibles spécifiques du réseau CoRegNet sont représentés en bleu. Les facteurs de transcriptions sont représentés par une forme de losange tandis que les protéines provenant de la carte moléculaire de la PR ainsi que les gènes cibles provenant du réseau CoRegNet sont représentés par une forme circulaire. Une inhibition est représentée par une flèche plate (—⊣) rouge tandis qu'une activation est représentée par une flèche grise (⟶).

6.2.4 Superposition de variants génomiques et de DEGs

En premier lieu, nous avons recherché à superposer les 15 variants identifiés dans le Chapitre 5. La superposition des gènes contenant ces variants n'a montré aucun gène commun avec le réseau global intégratif et spécifique de la PR. Ces variants ont par la suite été comparés à la carte moléculaire de la PR (`https://ramap.uni.lu/minerva/index.xhtml?id=ra_map_20avril_2021_pmaip1_corrected`), où aucun gène commun n'a également été identifié.

Il ensuite a été décidé d'utiliser une liste de variants publique provenant de DisGeNET à partir de laquelle nous avons extrait 1635 variants correspondant à 731 gènes. À partir des 731 gènes provenant de DisGeNET, 61 gènes ont été retrouvés communs avec le réseau global intégratif et spécifique de la PR. Parmi ces gènes communs, nous avons identifié TNFAIP3 qui est un TF commun entre le réseau CoRegNet et la carte moléculaire extraite de la PR.

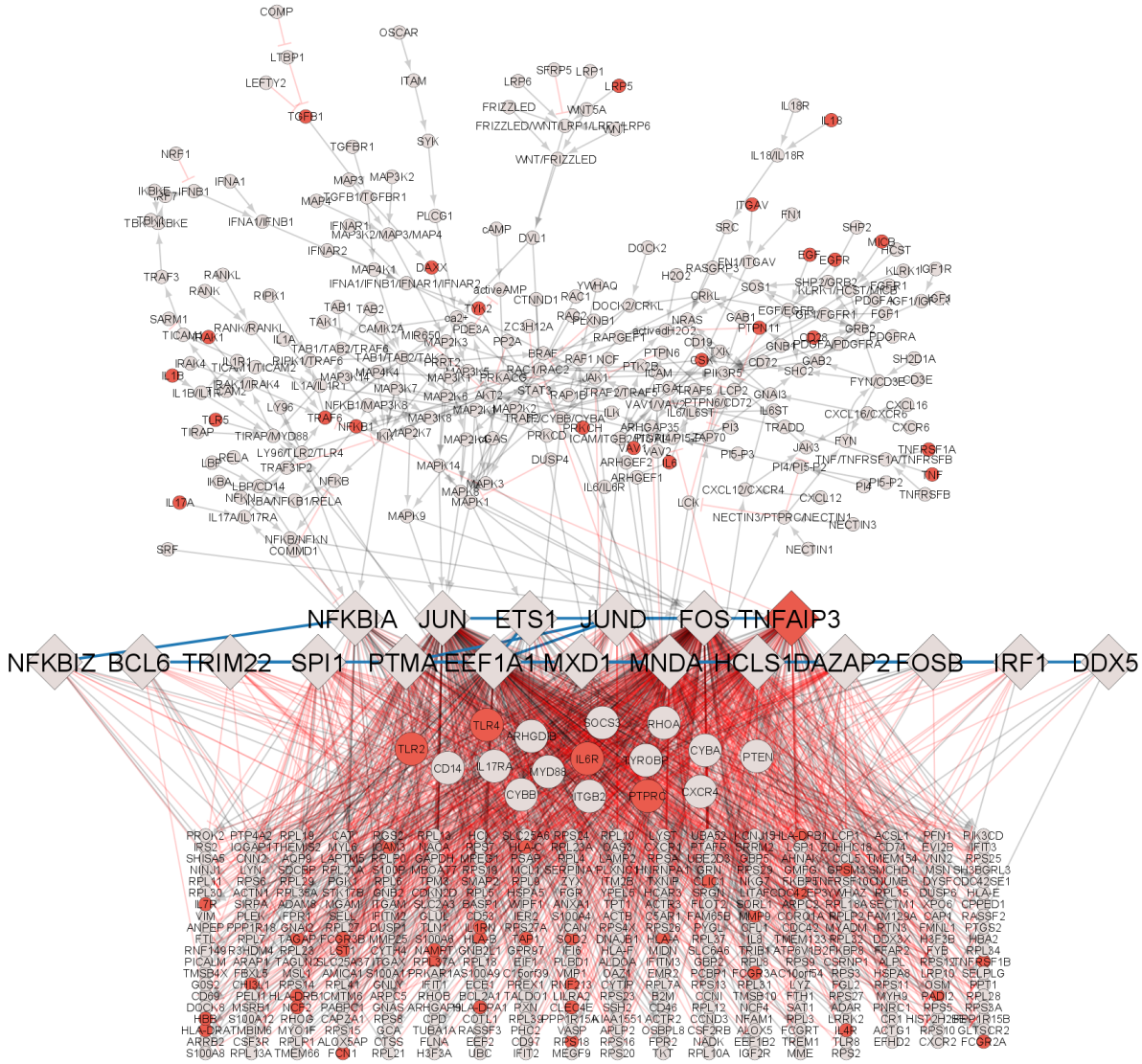

Figure 6.7: Réseau global intégratif et spécifique de la PR et les variants spécifique de la PR obtenus à partir de DisGeNET Les variants chevauchants sont représentés en rouge (61), tandis que les gènes/protéines non-chevauchants sont représentés en gris (553). Les facteurs de transcriptions provenant de CoRegNet sont représentés par une forme de losange, tandis que les protéines provenant de la carte moléculaire de la PR ainsi que les gènes cibles provenant du réseau CoRegNet sont représentés par une forme circulaire. Une inhibition est représentée par une flèche plate (—⊣) rouge tandis qu'une activation est représentée par une flèche grise (⟶).

Dans un second temps, nous avons réalisé la superposition des DEGs issus de 2 jeux de données d'expression transcriptomique RNAseq. Le premier contient 37 et 41 patients atteints de PR ayant reçu des traitements anti-TNF (adalimumab et etanercept respectivement) dans le but d'étudier les patients répondeurs et non-répondeurs. Le second implique des patients ayant reçu des traitements anti-TNF ainsi que des patients n'ayant reçu aucun traitement anti-TNF. Les DEGs obtenus à partir de ces deux jeux de données ont été superposés avec le réseau global intégratif et spécifique de la Polyarthrite Rhumatoïde (présenté en Figure 6.8 et 6.9 respectivement).

La superposition des DEGs provenant des données de patients atteints de PR répondeurs et non-répondeurs a révélé 15 molécules chevauchantes incluant 4 DEGs etanercept et 11 DEGs adalimumab. Parmi ces DEGs, nous retrouvons les 4 TFs clés retrouvés entre CoRegNet et la carte moléculaire de la PR (NFKBIA, JUN, FOS et TNFAIP3) et 1 TF identifié par CoRegNet uniquement (FOSB). La superposition des DEGs provenant des patients avant et après traitement a révélé 101 molécules chevauchantes, incluant 2 TFs clés retrouvés entre CoRegNet et la carte moléculaire de la PR (NFKBIA et FOS) et 4 TFs identifié par CoRegNet uniquement *(BCL6, MXD1, MNDA et DAZAP2)*.

Enfin, l'analyse combinée des DEGs provenant des deux jeux de données de patients avec traitement et sans traitement anti-TNF a montré un total de 101 molécules chevauchantes, incluant 9 des 19 TFs inclus dans le réseau global intégratif et spécifique de la PR (présentés en Table 6.4. Parmi ces 9 TFs, deux ont été retrouvés dans les deux analyses (NFKBIA et FOS).

Table 6.4: Facteur de transcription à partir du réseau global intégratif et spécifique de la PR chevauchant au moins un gène différentiellement exprimé à partir des analyses répondeurs/non-répondeurs à des traitements anti-TNF et avant et après traitement anti-TNF.

Source	TF	Répondeurs/non-répondeurs	Avant/après traitement
CoRegNet et carte moléculaire de la PR	NFKBIA	↓	↓
	JUN		↓
	FOS	↓	↓
CoRegNet	*TNFAIP3*		↓
	BCL6	↓	
	MXD1	↓	
	MNDA	↓	
	DAZAP2	↓	
	FOSB		↓
	NFKBIA	↓	↓
	JUN		↓

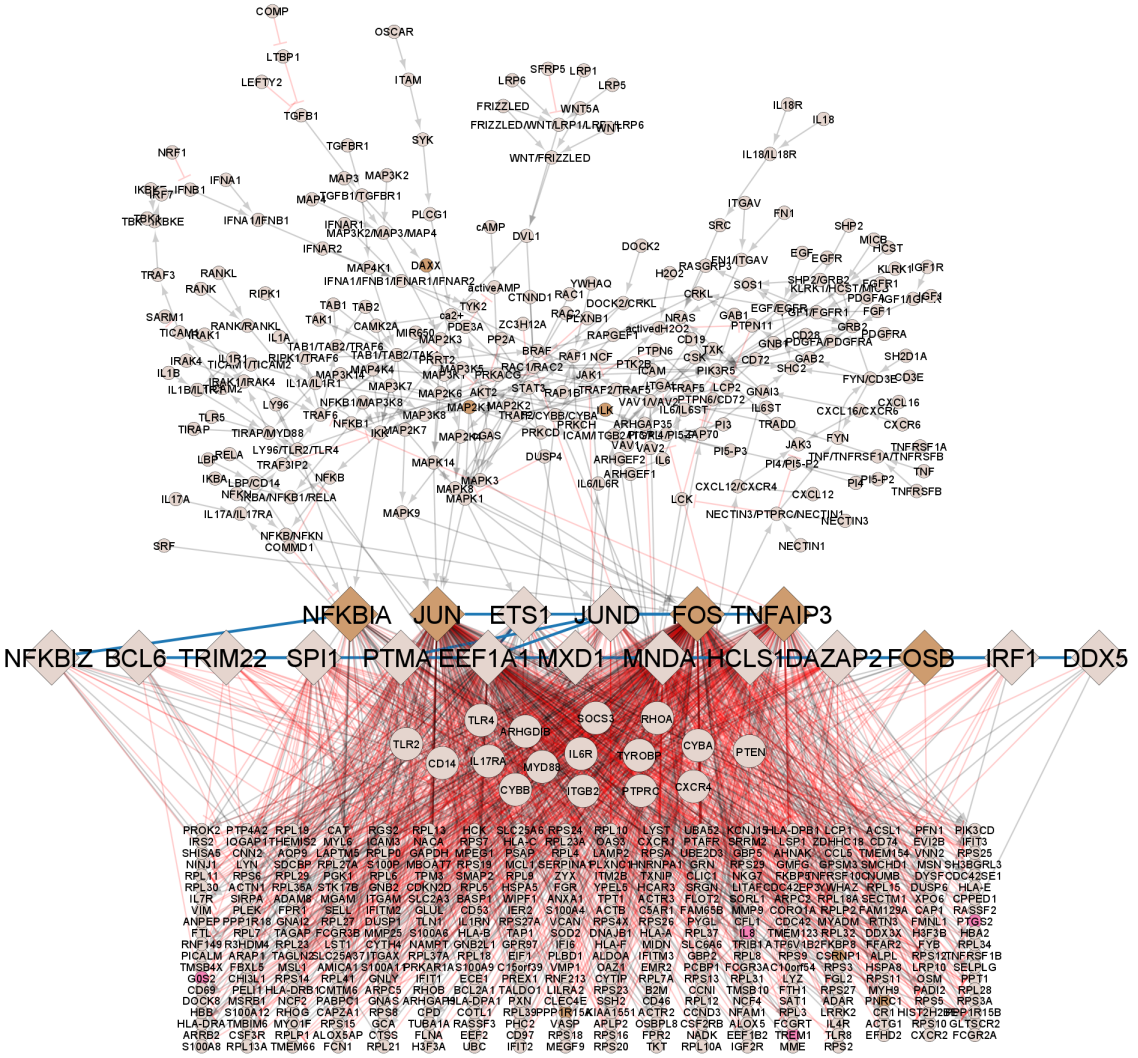

Figure 6.8: Réseau global intégratif et spécifique de la PR et les DEGs obtenus à partir de patients répondeurs et non-répondeurs à des traitements anti-TNF (37 et 41 patients atteints de PR traités respectivement par Adalimumab et Etanercept) Les DEGs chevauchants à partir des traitements adalimumab et etanercept sont respectivement représentés en marron (11) et rose (4) tandis que les gènes/protéines non chevauchants sont représentés en gris (599). Les facteurs de transcriptions provenant de CoRegNet sont représentés par une forme de losange, tandis que les protéines provenant de la carte moléculaire de la PR ainsi que les gènes cibles provenant du réseau CoRegNet sont représentés par une forme circulaire. Une inhibition est représentée par une flèche plate (—⊣) rouge tandis qu'une activation est représentée par une flèche grise (⟶).

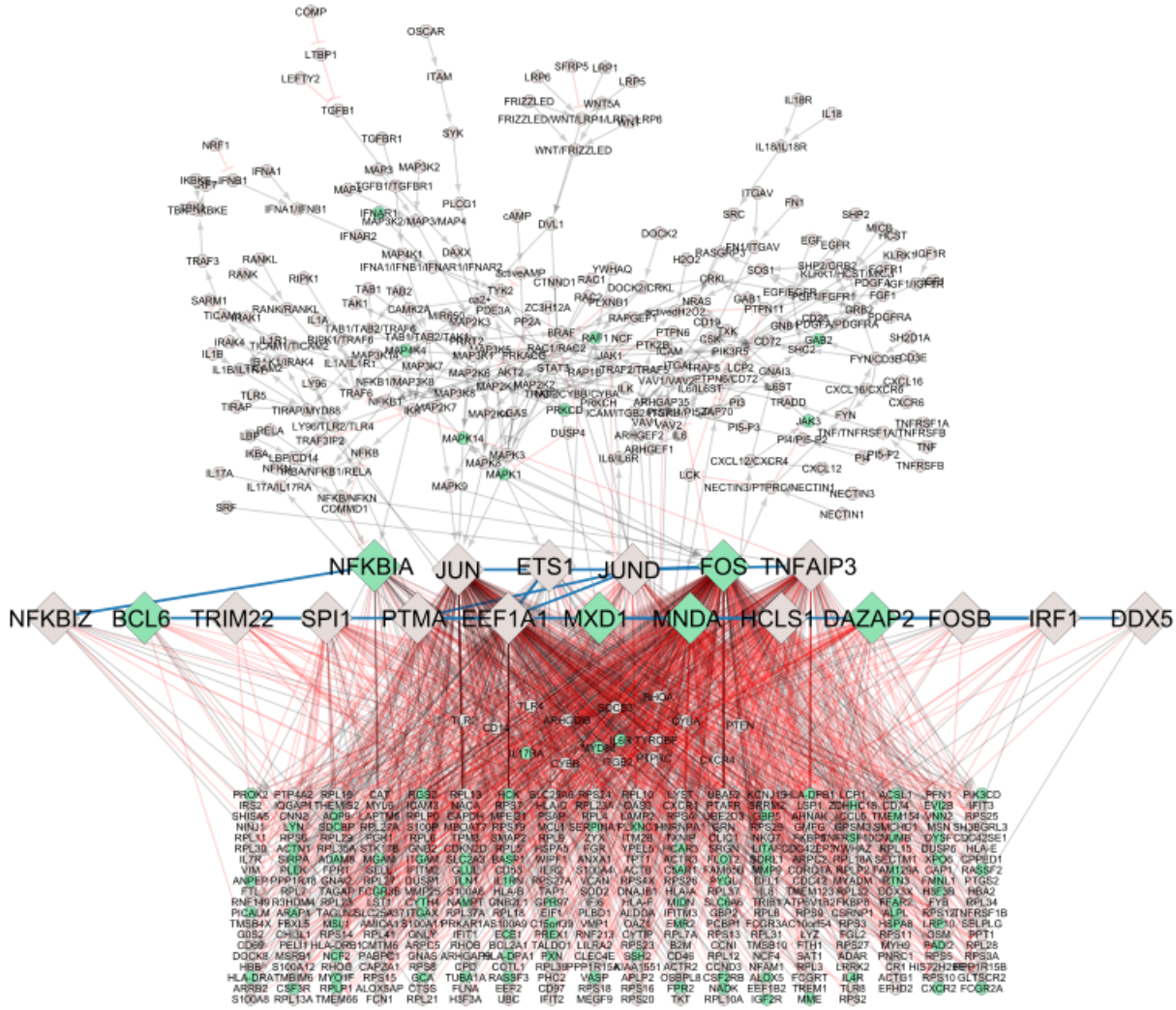

Figure 6.9: Réseau global intégratif et spécifique de la PR et les DEGs obtenus à partir de patients traités et non-traités par traitement anti-TNF (deux cohortes de 40 et 36 patients atteints de PR avant et après une durée de trois mois de traitement avec infliximab ou adalimumab) Les DEGs chevauchants sont représentés en vert (101), tandis que les gènes/protéines non-chevauchants sont représentés en gris (513). Les facteurs de transcriptions provenant de CoRegNet sont représentés par une forme de losange, tandis que les protéines provenant de la carte moléculaire de la PR ainsi que les gènes cibles provenant du réseau CoRegNet sont représentés par une forme circulaire. Une inhibition est représentée par une flèche plate (—⊣) rouge tandis qu'une activation est représentée par une flèche grise (⟶).

6.2.5 Sous-réseau spécifique de la PR

Les analyses précédentes ont permis d'identifier des TFs clés différentiellement sous-exprimés après traitements selon la réponse à des traitements anti-TNF. Ainsi, il est évident que la régulation de ces TFs clés est due à une régulation en cascade, en plus de la voie de signalisation de TNF.

Afin d'analyser ces régulations en cascade potentiellement interconnectées, nous avons décidé de regarder en détail les voies d'IL6, de TGF-β (TGFB1 dans notre réseau) ainsi que celle de TNF. IL6 est une cible du tocilizumab, qui est un inhibiteur fréquemment utilisé dans les traitements contre la PR [356]. La voie de signalisation TGF-β a été retrouvée activée dans la membrane synoviale, cependant son blocage ne semble pas affecter une arthrite expérimentalement induite [357]. Dans le but d'étudier ces voies de signalisation et leurs impacts sur les TFs clés, nous avons extrait un sous-réseau en sélectionnant les protéines TGFB1, IL6 et TNF ainsi que l'ensemble des protéines impliquées en aval jusqu'au premier TF à partir du réseau global intégratif et spécifique de la PR.

Le sous-réseau contient 38 molécules, incluant 4 TFs (FOS, JUN, JUND et NFKBIA), présenté en Figure 6.10. En projetant une nouvelle fois les DEGs, et gènes associés à des variants génomiques provenant de DisGeNET, nous observons que les molécules intermédiaires et TFs de ce réseau sont communes avec des DEGs (Figure 6.10 (a) et (b)), tandis que les trois protéines à partir desquelles le réseau a été extrait (IL6, TGFB1 et TNF) et quelques molécules intermédiaires sont communes avec les variants génomiques provenant de DisGeNET (Figure 6.10 (c)).

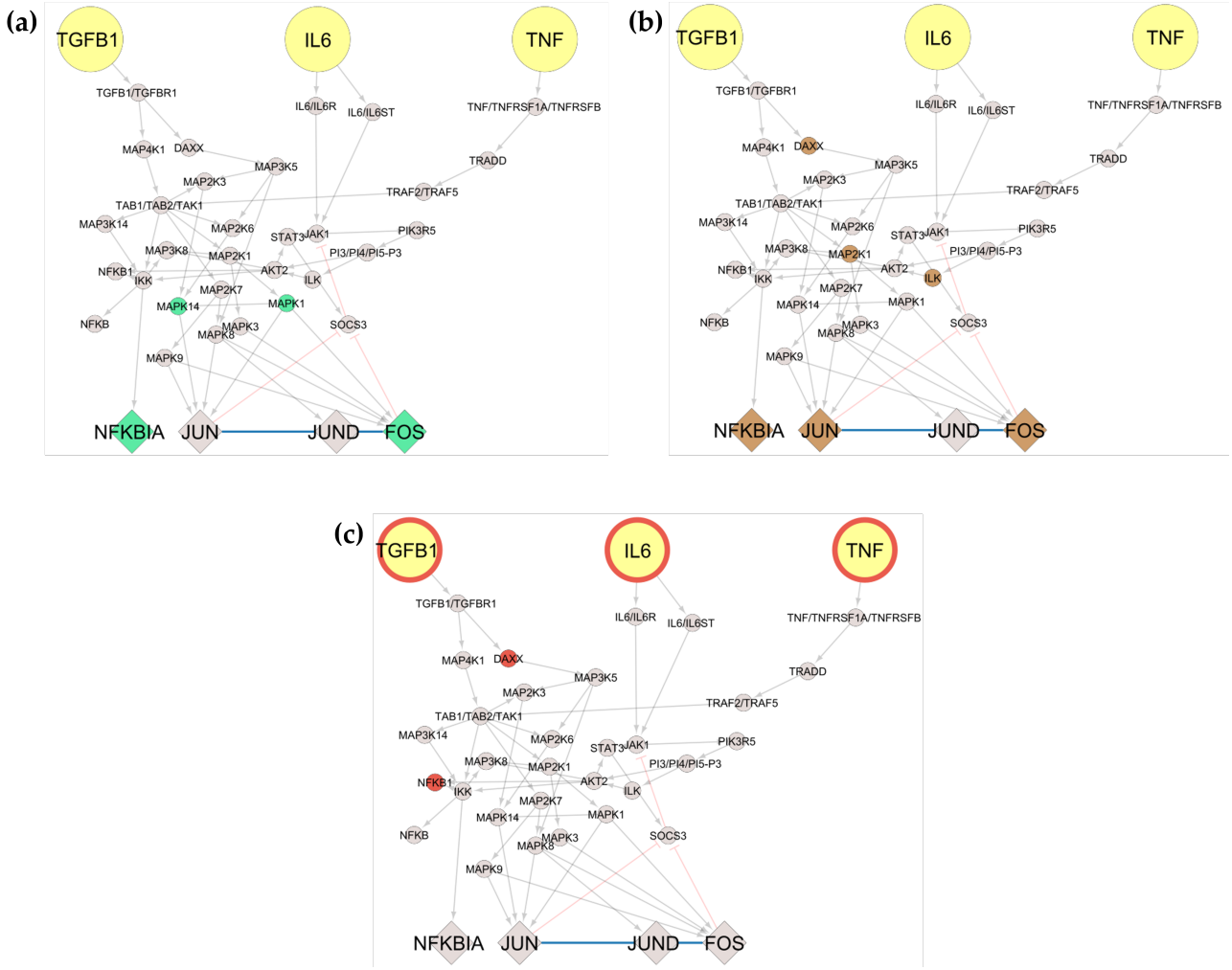

Figure 6.10: Sous-réseau extrait des protéines cibles (TGFB1, IL6, and TNF) du réseau global intégratif et spécifique de la PR. Ce sous-réseau est basé sur les protéines ciblées: TGFB1, IL6 et TNF (coloré en jaune), ainsi que l'ensemble des protéines impliquées en aval (forme circulaire) jusqu'au premier TF (forme de losange). Nous avons par la suite projeté **(a)** les DEGs obtenus à partir de patients atteints de PR répondeurs et non-répondeurs à des traitements anti-TNF (vert), **(b)** les DEGs obtenus à partir de patients atteints de PR avant et après traitement anti-TNF montré en marron (6) et **(c)** les variants génomiques obtenus à partir de la base de données DisGeNET en rouge (5).

À partir de ce sous-réseau, nous avons voulu évaluer par simulation l'impact de perturbations isolées et combinées sur l'expression des TFs en sortie de notre réseau (FOS, JUN, JUND et NFKBIA). Pour cela nous avons transformé le sous-réseau en un réseau booléen, en ajoutant des règles booléennes par l'utilisation de l'outil

CaSQ [345]. Cet outil utilise un fichier SBML CellDesigner [343] et produit un réseau booléen contenant des règles logiques. Le réseau booléen obtenu contenait alors 59 interactions reliant 38 molécules dont 3 en entrée (IL6, TGFB1 et TNF), 6 en sortie (FOS, JUN, JUND, NFKB, NFKB1 et NFKBIA) et 29 intermédiaires.

Le réseau booléen obtenu sous un format SBML-qual a ensuite été importé dans un premier temps sur Cell Collective [354] afin de réaliser des analyses de simulation en temps réel ainsi qu'une analyse de sensibilité. Dans un second temps, ce fichier a été importé sur GINsim [355] afin de calculer les états stables et de réaliser des simulations *in silico* par KO (pour ces simulations, une version réduite du réseau a été utilisée).

6.2.6 Modélisation booléenne

6.2.6.1 Simulations en temps réel

À partir du sous-réseau nous avons recherché à évaluer l'impact des molécules identifiées par les traitements ou porteurs de mutation sur les TFs clés (FOS, JUN, JUND et NFKBIA) en utilisant Cell Collective. Pour l'analyse du jeu de données avant et après traitement anti-TNF, l'analyse a révélé que MAPK14 et MAPK1 étaient sous-exprimés. Pour l'analyse du jeu de données des patients répondeurs et non-répondeurs, MAP2K1, ILK et DAXX ont également été identifiés comme sous-exprimés. Enfin DAXX et NFKB1 ont été identifiés comme porteur d'une mutation selon l'analyse des variants de la base de données DisGeNET. Afin de reproduire les effets d'une sous-expression sur ces molécules, nous avons réalisé une simulation *in silico* afin de réduire leur niveau d'expression à 0.

Pour le jeu de données avant et après traitement anti-TNF des patients atteints de PR, MAPK14 et MAPK1 ont été mis sur zéro et les protéines en entrée du réseau (TNF, TGFB1 ou IL6) ont été séquentiellement activées, révélant que tous

les TFs en sortie sont exprimés (Figure 6.11 a-c). Lorsque nous avons répliqué les sous-expressions des protéines MAP21K1, ILK et DAXX pour le jeu de données contenant des patients répondeurs/non-répondeurs à des traitements anti-TNF, nous avons observés que l'activation de TNF et TGFB1 permettait à tous les TFs d'être exprimés (Figure 6.11 d et e). Cependant, lorsque IL6 était activé, seulement NFKBIA était exprimé (Figure 6.11 f). Enfin, la réplication de la sous-expression des protéines DAXX et NFKB1 pour le jeu de données de variants provenant de DisGeNET, a révélé que l'activation de TNF, TGFB1 ou IL6, active l'ensemble des TFs (Figure 6.11 g-i).

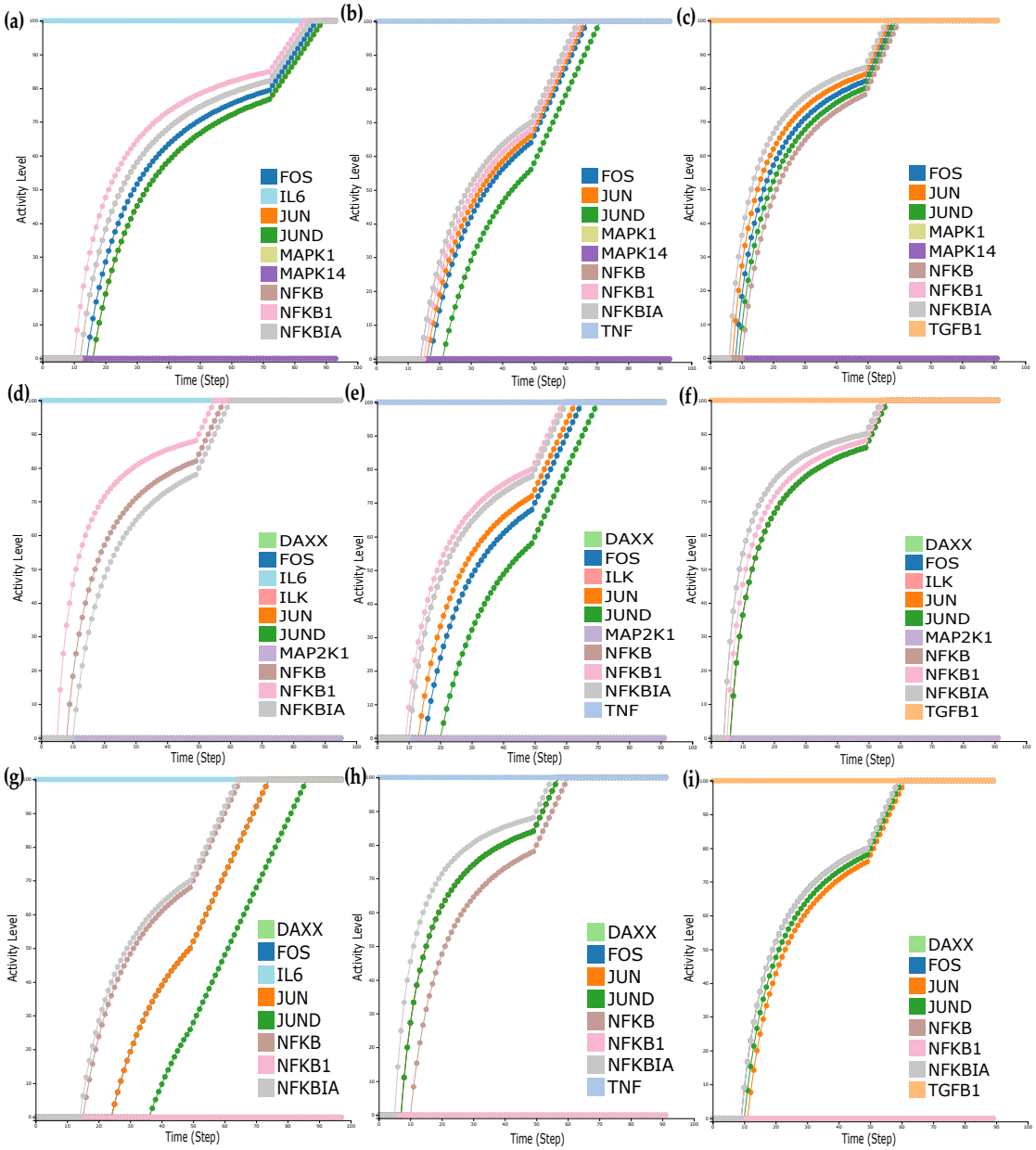

Figure 6.11: Simulation en temps réel du sous-réseau avec Cell Collective. Les simulations ont été réalisées en répliquant : La sous expression des protéines MAPK14 et MAPK1 pour le jeu de données avant et après traitement anti-TNF pour des patients atteints de PR **(a-c)**; **(a)** Simulation avec l'activité d'IL6 fixée à 100%. **(b)** Simulation avec l'activité de TNF fixée à 100%. **(c)** Simulation avec l'activité de TGFB1 fixée à 100%. La sous expression des protéines MAP2K1, ILK et DAXX pour le jeu de données avant et après traitement anti-TNF pour des patients atteints de PR **(d-f)**; **(d)** Simulation avec l'activité d'IL6 fixée à 100%. **(e)** Simulation avec l'activité de TNF fixée à 100%. **(f)** Simulation avec l'activité de TGFB1 fixée à 100%. **(g-i)** Les mutations observées dans la base de données DisGeNET. **(g)** Simulation avec l'activité d'IL6 fixée à 100%. **(h)** Simulation avec l'activité de TNF fixée à 100%. **(i)** Simulation avec l'activité de TGFB1 fixée à 100%.

6.2.6.2 Relation dose-effet et analyse de sensibilité

L'étude de la relation dose-effet a été réalisée selon cinq conditions initiales, présentées en Table 6.1, afin de reproduire les différents scénarios possibles. La première condition correspond à un blocage simultané de TNF, IL6 et TGFB1 ; la seconde correspond à l'activation d'IL6 ; la troisième de TGFB1 actif et la quatrième à l'activation d'IL6 et de TGFB1 et la cinquième condition à l'activation de TNF.

L'analyse de la relation dose-effet pour chaque condition nous a permis d'observer une expression dose-dépendante pour TNF, TGFB1 et IL6 (Figure 6.12 b,e et f), tandis que l'activation simultanée d'IL6 et de TGFB1 a entraîné une activation supérieure pour les TFs, et ce même lorsque les doses d'IL6 et de TGFB1 étaient faibles (Figure 6.12 c et d).

Par la suite, nous avons regardé si la sous expression des TFs observés après traitements anti-TNF pouvait être contre-balancée par d'autres voies moléculaires. Pour cela nous avons réalisé une analyse de sensibilité afin d'identifier si les deux autres protéines ciblées (IL6 et TGFB1) avaient un impact significatif sur la régulation des TFs lorsque l'activité de TNF est bloquée. Ces résultats, présentés en Figure 6.13 ont montré que les TFs peuvent être activés et ce, en l'absence d'activité de TNF et lorsque IL6 et TGFB1 sont activés simultanément.

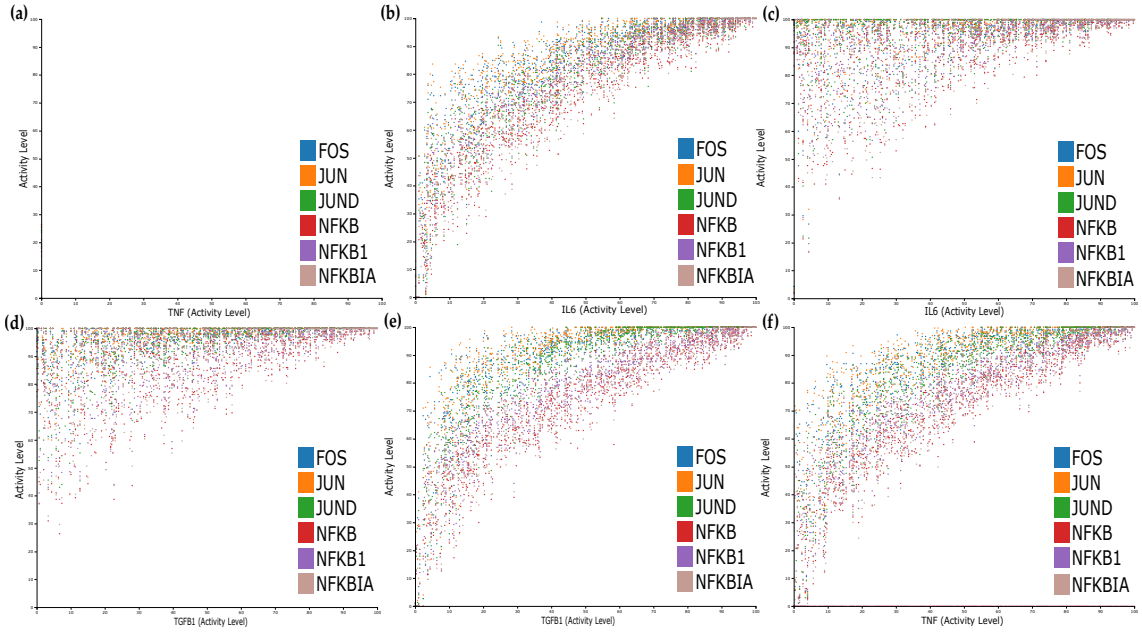

Figure 6.12: Analyse de la relation dose-effet du sous-réseau. **(a)** TNF, TGFB1 et IL6 actifs. **(b)** IL6 actif. **(c)** IL6 et TGFB1 actifs (vue TGFB1). **(d)** IL6 et TGFB1 actif (vue IL6). **(e)** TGFB1 actif. **(f)** TNF actif.

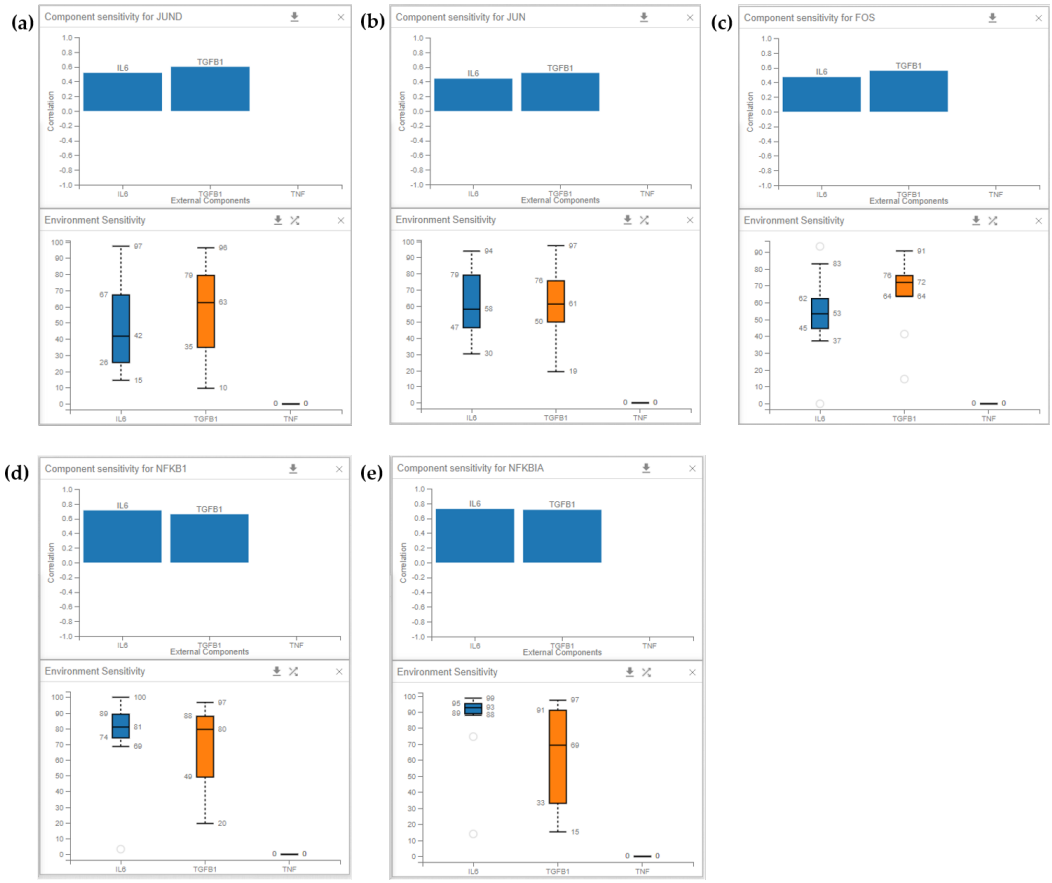

Figure 6.13: Analyse de sensibilité du sous-réseau. Les facteurs de transcription en sortie peuvent être activés en l'absence d'activité de TNF et lorsque les deux protéines IL6 et TGFB1 sont activées simultanément. La partie haute de chaque sous figure démontre l'impact des composants externes sur l'activité des TFs sélectionnés tandis que la partie basse de chaque sous figure représente l'étendue du pourcentage d'activité des composants externes afin d'obtenir l'optimisation de l'activité des TFs sélectionnés. Analyse de sensibilité pour **(a)** JUND; **(b)** JUN; **(c)** FOS; **(d)** NFKB1; **(e)** NFKBIA; les boîtes à moustache représentent: IL6 en bleu et TGFB1 en orange. TNF est représenté par une ligne noire car fixé sur off.

6.2.6.3 Analyse d'état stable et simulation de KO

L'analyse des états stables a été réalisée avec le modèle présenté en Figure 6.14 en utilisant l'outil GINsim. L'analyse sans perturbation a révélé cinq états stables et aucun attracteur complexe. La configuration de ces cinq états stables est présentée dans la Table 6.5. Cette analyse a révélé que l'activation d'IL6 ou IL6 et TGFB1 peut réguler positivement l'expression des TFs en sortie et ce même

en présence de traitement anti-TNF ($TNF = 0$). L'expression des TFs peut être complètement stoppée uniquement si TNF est bloqué ainsi que IL6 et TGFB1. D'autre part, DAXX était exprimé uniquement lorsque TGFB1 était activé, et ILK dépendait de l'activation d'IL6. Les molécules MAPK du sous-réseau dépendaient de l'activation d'IL6 et de TGFB1 mais ne semblait pas impactées par le blocage de TNF.

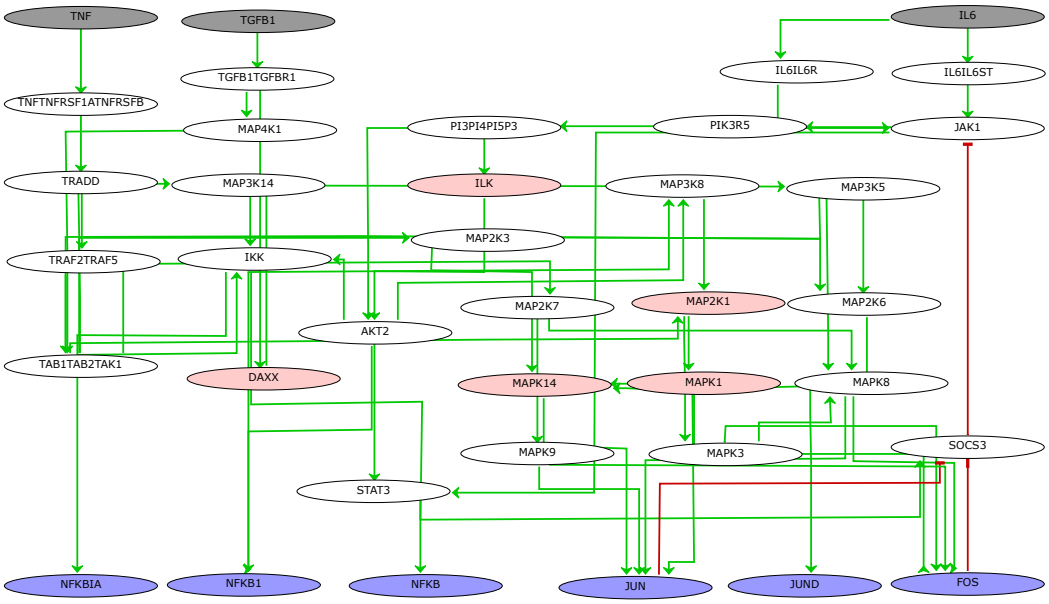

Figure 6.14: Réseau booléen incluant trois cascades de signalisation pour TNF, TGFB1 et IL6. Les règles booléennes de ce réseau ont été inférées à partir de l'outil CaSQ (Voir la section 6.1). Les trois cibles de ce sous-réseau sont représentées en gris, les TFs d'intérêts sont représentés en violet, les molécules intermédiaires affectées par un traitement anti-TNF, ou identifiées comme porteuses d'une mutation sont représentées en rose. Une inhibition est représentée par une flèche plate (—⊣) rouge tandis qu'une activation est représentée par une flèche grise (⟶).

Table 6.5: État stable du réseau booléen sans perturbation.

État stable	TNF	IL6	TGFB1	JUN	FOS	JUND	NFKBIA	DAXX	ILK	NFKB1	MAP2K1	MAPK1	MAPK14
ss1	0	0	0	0	0	0	0	0	0	0	0	0	0
ss2	0	1	0	1	1	1	1	0	1	1	1	1	1
ss3	0	1	1	1	1	1	1	1	1	1	1	1	1
ss4	1	1	0	1	1	1	1	0	1	1	1	1	1
ss5	1	1	1	1	1	1	1	1	1	1	1	1	1

Le réseau booléen qui contient les 38 molécules du sous-réseau initial, a par la suite été réduit en utilisant la fonction de réduction de GINsim afin de réaliser des expériences *in silico* pour les molécules d'intérêts. Le modèle booléen après réduction était composé de 23 molécules, à partir duquel nous avons reproduit via des KO les effets des traitements anti-TNF ainsi que les molécules porteuses de mutations qui avaient été précédemment identifiées. Ces conditions sont présentées dans la Table 6.2. Afin de réaliser ces simulations, nous avons fixé les conditions initiales de TNF sur zéro, laissé les autres molécules d'intérêts (IL6 et TGFB1) libres et fixé les conditions initiales de toutes les molécules intermédiaires du sous-réseau sur zéro.

Les résultats de ces expériences *in silico* pour les molécules d'intérêts et les trois conditions sont présentés en Table 6.6. Pour chaque condition, nous avons obtenu trois états stables. Pour la première condition, nous avons observé que DAXX est strictement "TGFB1 dépendant" (Table 6.6 **a**, ss2 et ss3), cependant, l'ensemble des TFs peut être activé par la présence d'IL6 ou d'IL6 et TGFB1, et ce malgré le blocage de TNF ainsi que le KO des molécules MAPK14 et MAPK1.

La deuxième condition a montré que lorsque DAXX, ILK et MAP2K1 sont bloqués, IL6 seul ne permettait pas d'activer les TFs FOS, JUN et JUND ainsi que les kinases MAPK14 et MAPK1 (Table 6.6 **b**, ss2). Cependant, IL6 et TGFB1 tous deux activés permettait d'activer l'ensemble des TFs ainsi que la kinase MAPK14 (Table 6.6 **b**, ss3).

Enfin, l'étude de la dernière condition sur les effets des porteurs de mutation a montré que malgré le blocage de TNF et les KO de DAXX et NFKB1, tous les TFs, ainsi que les kinases était activées et ce, en présence d'IL6 ou d'IL6 et TGFB1 combinés (Table 6.6 **c**).

Table 6.6: État stable du réseau booléen incluant des perturbations. **(a)** MAPK1, MAPK14 KO et TNF = 0. **(b)** DAXX, ILK et MAP2K1 KO et TNF = 0. **(c)** DAXX, NFKB1 KO et TNF = 0.

(a)

État stable	TNF	IL6	TGFB1	JUN	FOS	JUND	NFKBIA	DAXX	ILK	NFKB1	MAP2K1	MAPK1	MAPK14
ss1	0	0	0	0	0	0	0	0	0	0	0	0	0
ss2	0	1	0	1	1	1	1	0	1	1	1	0	0
ss3	0	1	1	1	1	1	1	1	1	1	1	0	0

(b)

État stable	TNF	IL6	TGFB1	JUN	FOS	JUND	NFKBIA	DAXX	ILK	NFKB1	MAP2K1	MAPK1	MAPK14
ss1	0	0	0	0	0	0	0	0	0	0	0	0	0
ss2	0	1	0	0	0	0	1	0	0	1	0	0	0
ss3	0	1	1	1	1	1	1	0	0	1	0	0	1

(c)

État stable	TNF	IL6	TGFB1	JUN	FOS	JUND	NFKBIA	DAXX	ILK	NFKB1	MAP2K1	MAPK1	MAPK14
ss1	0	0	0	0	0	0	0	0	0	0	0	0	0
ss2	0	1	0	1	1	1	1	0	1	0	1	1	1
ss3	0	1	1	1	1	1	1	0	1	0	1	1	1

6.3 Discussion

Dans ce chapitre, nous avons proposé l'étude de mécanismes potentiellement impliqués dans la PR, en utilisant l'inférence et l'analyse dynamique d'un réseau global, spécifique de la PR. Pour cela, nous avons construit dans un premier temps un réseau, en utilisant des données transcriptomiques publiques de patients atteints de PR combiné avec une carte moléculaire de la PR précédemment construite au laboratoire [170]. Ce réseau a notamment été étudié pour l'identification de facteurs de transcription clés impliqués dans la PR. Dans un deuxième temps, nous mettons en évidence des molécules clés dans la PR, en utilisant le réseau inféré et en y associant : les variants génomiques identifiés dans le chapitre précédent (Chapitre 5), des variants génomiques publiques et des données transcriptomiques publiques provenant de patients ayant reçu des traitements anti-TNF. Ces résultats sont ensuite analysés via des simulations *in silico* et permettent de mettre en évidence de nouveaux mécanismes dans la PR.

Au cours de ce travail, nous avons recherché à étudier les mécanismes clés de la régulation dans la PR. Pour cela, nous avons inféré un réseau global et spécifique de la PR en inférant un réseau de co-régulation de gène combiné via les facteurs de transcription communs à une carte moléculaire de la PR. L'étude du réseau de co-régulation nous a permis d'identifier 19 TFs clés, et l'étude de la littérature a montré que ces TFs étaient tous impliqués dans la PR et l'auto-immunité. Six d'entre eux sont par ailleurs présents dans la carte moléculaire de la PR utilisée lors de l'inférence du réseau, qui est un réseau mécanistique exhaustif de la maladie créée manuellement. Ces 6 TFs, ETS1, FOS, JUN, JUND, NFKBIA et TNFAIP3 ont été utilisés afin de créer le lien entre notre réseau de co-régulation inféré et la carte moléculaire de la PR. Ainsi, nous avons créé un réseau global et intégratif de la PR incluant les voies de signalisations de la carte moléculaire, TFs et gènes

cibles.

Par la suite, nous avons voulu étudier les mécanismes causaux potentiellement impliqués dans la PR et ses voies de signalisation à partir de notre réseau. Pour cela, nous avons intégré des données à ce réseau en incluant des variants génomiques issus du Chapitre 5 et de bases de génomiques publiques ainsi que des données transcriptomiques issus de la réponse à des traitements anti-TNF de patients atteints de PR.

À partir des variants identifiés dans le Chapitre 5, aucun résultat commun n'en est ressorti avec notre réseau. Il est possible d'expliquer ce résultat par l'une des caractéristiques de ces variants : ils sont rares et par conséquent ont peu de chance d'être retrouvés dans notre réseau. En effet, la caractéristique de notre réseau est qu'il est global car inféré à partir de données d'expression de patients atteints de PR ainsi que via une carte moléculaire de la PR constituée à partir de la littérature. Par conséquence, le réseau est constitué principalement de molécules communes dans la PR ce qui n'est potentiellement pas compatible. Une solution serait d'ajouter ces variants rares et non de les superposer à partir d'interactions protéine-protéine via des voisins éloignés, issues de bases de données protéiques telles que STRING [340].

Cependant, l'utilisation de variants génomiques publiques impliqués dans la PR a montré que 61 gènes contenant un variant étaient communs avec notre réseau. De même, l'étude des données transcriptomiques de patients atteints de PR a permis d'identifier 15 DEGs (données répondeurs/non-répondeurs à des traitements anti-TNF) et 101 DEGs (données traitement/sans traitement anti-TNF) communs avec notre réseau.

A partir de ces résultats, notre étude s'est concentrée sur l'impact des traitements sur l'activité des voies de signalisations incluses dans notre réseau et sur les différents états possibles des TFs identifiés. Pour cela, nous avons extrait un sous-réseau à

partir des cascades de signalisations de TNF, IL6 et TGFB1 jusqu'au premier TF affecté afin de réduire la complexité de notre réseau, et de concentrer notre travail sur les régulateurs en amont de ces TF clés. IL6 a été choisi car il est une cible par traitement biothérapique original (boDMARDS) dans la PR [358–361], ainsi que TGF-β, car c'est une cytokine immunomodulatrice hautement exprimée chez des patients atteints de PR, dont le rôle est encore à clarifier [124, 247, 362].

Afin d'étudier ces mécanismes, nous avons décidé de transformer notre sous-réseau en un modèle booléen en utilisant CaSQ [345] afin de réaliser des simulations *in silico*. Les simulations en temps réels et l'analyses dose-réponse ont démontré que les cascades impliquant IL6 et TGFB1 peuvent affecter l'expression des TFs et peuvent même contre-balancer la sous-régulation des TFs causée par le blocage de TNF, comme l'a montré l'analyse de sensibilité. L'analyse des états stables a confirmé les résultats des analyses de simulation en temps réel, montrant que les TFs identifiés comme TFs clés, l'activation des cascades d'IL6 ou d'IL6 et de TGFB1 permettaient de réguler positivement leur expression. Bloquer la cascade TNF permet de bloquer complètement l'expression de ces TFs, uniquement si combiné avec le blocage d'IL6 et TGFB1. Ces résultats laissent envisager que le blocage de plusieurs cibles pourrait être une piste, ce qui a par ailleurs été mis en avant par des études, comme l'utilisation de thérapies bi-spécifiques permettant par exemple de bloquer IL6 et TNF simultanément [363], ou en administrant deux thérapies spécifiques ciblées en même temps. Il faut cependant garder en tête que l'administration de thérapie biologique combinée a été associée à une augmentation d'effets indésirables [364, 365].

Les simulations réalisées en activant les molécules d'intérêts en entrée (IL6, TGFB1 et TNF) ainsi que les KO combinés répliquant l'effet des traitements anti-TNF en combinaison avec la sous-régulation des gènes perturbés identifiés (avant/après traitement, répondeurs non-répondeurs et gènes mutés) ont permis

de confirmer la dépendance des TFs en sortie du sous-réseau dans des conditions spécifiques. En effet, lorsque TNF est inactivé et DAXX, ILK et MAP2K1 sont sous-régulés, l'activation d'IL6 n'est pas suffisante pour activer les TFs en sortie et les kinases MAPK14 et MAPK1 (Table 6.6 **b**, ss2). Cependant, lorsque IL6 et TNF sont tous deux activés, MAPK14 est à nouveau activé (Table 6.6 **b**, ss3). MAPK14 (P38a kinase) et MAPK1 (anciennement ERK2) sont notamment deux protéines connues pour jouer un rôle important dans la PR, en activant une variété de signaux incluant des cytokines telles que TNF et IL6 mais aussi TGF-β [366]. D'autre part, P38 a été proposé afin comme cible thérapeutique potentielle, afin de réduire la destruction des os et du cartilage, les inhibiteurs de cette protéine ayant montré des résultats thérapeutiques insuffisants [367, 368].

La suppression de MAPK14 ((p38a), présentée en Table 6.6 **b**)), ne permet pas d'inhiber l'activation les TFs en sortie, et ce même en présence de traitement anti-TNF car les molécules IL6 et TGFB1 permettent de contre-balancer ce processus. Concernant l'inhibition de MAPK1, peu d'informations existent dû à un manque d'efficacité des inhibiteurs pharmacologiques [369]. Dans une étude moins récente, un inhibiteur de MAPK1 a montré son efficacité dans une arthrite expérimentale au collagène chez des souris [370], cependant il n'y a pas eu de suivi significatif sur celui-ci. Dans notre modèle, l'inhibition de MAPK1 ne semble pas affecter l'activation des TFs en sortie du réseau, étant donné que FOS, JUN et d'autres molécules de notre réseau peuvent les activer.

En conclusion, un réseau global intégratif et spécifique de la PR a été créé, en utilisant une méthode innovatrice et combinant les outils adaptés. À l'aide de celle-ci nous avons pu identifier les TFs clés de ce réseau, tous impliqués dans la PR ou l'auto-immunité. D'autre part, il a été montré qu'il était possible d'utiliser des données multi-omiques, afin d'étudier des potentiels mécanismes clés de la PR. L'utilisation de la modélisation booléenne permettant par la suite d'étudier

par simulation les voies de signalisation clés d'un système biologique extrêmement complexe.

Chapitre 7

Conclusion Générale

La polyarthrite rhumatoïde (PR) est une maladie inflammatoire auto-immune multifactorielle et complexe, dont la composante génétique est estimée à environ 60%. Le facteur génétique majeur, découvert dès la fin des années 1980 est le gène HLA-DRB1 dont les allèles épitopes partagés SE sont les allèles à risque. Depuis, des études d'analyses de liaisons et des analyses d'association de gènes candidats ou pangénomiques ont permis d'identifier plus de 100 loci communs non-HLA associés à la PR, dont la plupart jouent un rôle dans des mécanismes immunitaires [136]. Cependant, tous ces facteurs génétiques n'expliquent environ que 50% de la composante génétique de la PR [121]. Dans le but de caractériser cette part d'héritabilité manquante, l'étude de variants rares (<1%) a été proposée par Manolio et al. 2010 [142], même si la caractérisation d'interactions (gène-gène et/ou gène-environnement) ou de sites méthylés pourrait aussi expliquer une partie de la composante génétique non identifiée. Les techniques NGS apportent une aide certaine dans l'étude des variants rares puisqu'elles permettent d'identifier de manière exhaustive tous les variants présents chez des individus (contrairement aux puces de génotypage qui ciblent des variants particuliers et en général fréquents). Cependant, les études d'association cas-témoins de variants rares restent difficiles

car peu puissantes puisque ces variants sont présents chez très peu d'individus. Des tests d'agrégation de variants rares (tests de Burden) sont ainsi privilégiés mais l'étude de données familiales est une bonne alternative pour observer la ségrégation d'un variant rare [144, 253].

Ainsi, le premier objectif de cette thèse était d'identifier de nouveaux gènes porteurs de variants rares associés à la PR en analysant des données génomiques (WES et WGS) issues d'apparentés atteints et non atteints de familles PR multiplexes d'origine française. Des outils d'appel de variants (variants callers) adaptés ont été appliqués aux données de séquence afin d'identifier des SNVs/indels d'une part et des CNVs d'autre part. Après un traitement bio-informatique standard pour éliminer les variants de mauvaise qualité, l'annotation des variants a été réalisée afin de filtrer uniquement les variants rares (<1%) et avec un effet délétère. Seuls les variants avec une pénétrance complète et une absence de phénocopie ont été génotypés et validés par PCR digitale dans un set de validation. Ainsi, les variants sélectionnés lors de cette dernière étape ont des caractéristiques très particulières. Or, dans le cas de maladies multifactorielles, des variants avec pénétrance incomplète et de la phénocopie peuvent jouer un rôle. Il serait donc intéressant d'élargir la recherche de variants rares à des variants ayant des caractéristiques moins strictes.

Pour les données WES, les SNVs/indels candidats ont été mis en évidence avant mon arrivée au laboratoire. En particulier, le gène SUPT20H était un bon candidat [144]. Pour l'identification de CNV, une étude des performances de 6 outils a d'abord été réalisée sur données simulées avant d'appliquer les plus performants à nos données WES.

L'étude des outils de détection de CNVs à partir de données simulées WES utilisant la méthode *read-depth* a montré qu'il n'existe pas d'outil consensus pour ce type d'analyse. Cette étude a par ailleurs montré que certaines caractéristiques

des CNVs facilitaient leurs détections : le fait qu'ils soient rares, grands et de type délétion. Si les outils détectent assez facilement les régions génomiques avec CNV, ils ont plus de difficulté à identifier correctement des CNVs à l'échelle des individus. Ces derniers résultats montrent qu'il faut donc prendre l'identification des CNV (surtout les CNV fréquents) avec précaution pour ensuite éffectuer des analyses d'association. Nos conclusions suggèrent une utilisation de plusieurs outils afin de palier à cela. Ainsi, pour l'analyse de données réelles, nous suggérons l'utilisation de l'outil le plus performant dans notre analyse, CODEX2, dont les résultats peuvent être croisés avec ceux des outils tels que DECoN, ExomeDepth et CLAMMS.

Ainsi, l'analyse de données WES avec ces 4 outils a permis de caractériser 3 CNVs dans le set de découverte, ne respectant cependant pas l'ensemble de nos critères de sélection avec une présence phénocopie et/ou une pénétrance incomplète dans le set de validation. Néanmoins, l'étude de la littérature a révélé que deux des gènes impactés, IGFL3 et GSDME, sont impliqués dans le système immunitaire. La protéine produite par le gène IGFL3 est un antigène majeur des lymphocytes B et est impliqué dans la régulation des lymphocytes T et B [242], des cellules dendritiques et neutrophiles [243, 244] et de la production d'IL6 dans les macrophages [245]. Le gène GSDME, lui, a été retrouvé associé à la PR par diverses études récentes, impliquant les monocytes [239, 240], les macrophages [240] et la prolifération et invasion des FLS dans la PR [241].

Par la suite, 15 variants délétères rares et spécifiques de la PR (10 SNVs, 2 indels et 3 CNVs) identifiés à partir de données WGS, ont été confirmés dans un set de validation (PCR digitale et données WES). Ces 15 variants respectaient tous nos critères de sélection. L'étude des gènes incluant ces variants par la littérature a montré que 7 gènes étaient impliqués dans la physiopathologie de la PR. Il est à noter que parmi l'ensemble des SNVs/indels identifiés dans le set initial, certains n'ont pas pu être retrouvés dans le set de validation puisque cette étape est réalisée à

partir de données WES. Ainsi, les variants localisés en dehors des régions exoniques ne pouvaient pas être validés.

Avec les critères très restrictifs de sélection des variants et parfois certaines contraintes des données du set de validation, des analyses supplémentaires seraient à effectuer pour caractériser au mieux les variants rares pouvant être impliqués dans la PR. Par exemple, le CNV identifié pour le gène BCL11A présent dans 2 familles et chez 6 patients (sur 6) et 2 non malades (sur 5), pourrait se révéler intéressant. L'analyse de ce variant se poursuit actuellement dans des familles trio de PR.

Si l'on compare les gènes mis en évidence par les 2 analyses initiales (WES et WGS), aucun gène en commun n'a pu être caractérisé. Ceci peut être expliqué par l'hétérogénéité des échantillons étudiés, tant par la structure familiale que par la présence ou non d'allèles SE du gène HLA-DRB1.

En conclusion, les travaux réalisés dans cette première partie de thèse ont permis de mettre en place plusieurs protocoles d'analyse de données WGS et de mettre en évidence des variants intéressants malgré un nombre de familles étudiées limité. Enfin, des études fonctionnelles seraient nécessaires pour valider l'effet de ces variants. Des analyses d'enrichissement et d'interactions gène-gène pourraient aussi être effectuées à partir de gènes incluant ces variants rares.

L'identification de variants génomiques associés à des gènes ne permet cependant pas de comprendre tous les mécanismes de la PR. Ils représentent uniquement une première couche d'information. Face à la complexité des maladies multifactorielles, il est essentiel d'utiliser plusieurs types de données, afin de comprendre leurs mécanismes. La biologie des systèmes le permet, en réduisant la complexité de l'utilisation de données multi-omiques à l'échelle d'un système [159–162]. Cela a déjà été réalisé, notamment en utilisant l'inférence de réseaux, pour l'étude de maladies complexes [159–162]. D'autre part, l'étude de ces réseaux par simulation,

en ajoutant des règles booléennes a montré son efficacité pour comprendre la régulation des gènes dans les cellules humaines [174–179], mais aussi de mieux comprendre les mécanismes impliquées dans des maladies tels que le cancer [180, 181].

Ainsi, le second objectif de cette thèse a été l'étude de mécanismes causaux pouvant lier des perturbations extracellulaires avec des voies de signalisation et l'expression génique dans la PR, en utilisant l'inférence et l'analyse dynamique d'un réseau global et spécifique à la maladie.

Dans ce but, nous avons créé un réseau global intégratif et spécifique de la PR, en utilisant une méthode innovatrice et en combinant des outils adaptés. Par la suite, son étude nous a permis d'identifier les TFs clés de ce réseau, tous impliqués dans la PR ou l'auto-immunité. Notre tentative de lier les deux parties de cette thèse, en ajoutant les variants rares identifiés dans le Chapitre 5 au réseau n'a pas pu être réalisé, ceci étant dû au caractère essentiellement rare des variants identifiés précédemment. Cependant, l'ajout de variants génomiques associées à la PR a pu être fait en utilisant des données publiques. Nous avons donc démontré qu'il était possible d'ajouter des données multi-omiques afin d'étudier des potentiels mécanismes clés de la PR. Il a été aussi démontré qu'il était possible d'utiliser un réseau global, afin d'étudier des voies de signalisation clés de la PR, par modélisation booléenne. Pour cela, nous avons recherché à analyser l'implication d'IL6, TGFB1 et TNF dans la réponse à des traitements anti-TNF, à l'aide de l'extraction d'une sous partie de notre réseau, des données multi-omiques et de simulation *in silico*. Ces résultats suggèrent qu'un blocage de TNF est insuffisant pour bloquer les facteurs de transcription en sortie de notre réseau, car ces derniers pouvaient toujours être activés par la combinaison d'IL6 et TGFB1. Dans ce sens, des études ont proposé l'utilisation de thérapies bi-spécifiques ou l'administration de deux thérapies spécifiques ciblées. Finalement, nos résultats ont montré que l'utilisation

de la biologie des systèmes, de données multi-omiques, de l'inférence de réseau, de la modélisation booléenne et de l'analyse par simulation *in silico* permet de mieux comprendre les mécanismes impliqués dans une maladies multifactorielle complexe comme la PR.

Cependant, l'utilisation de patients à plus large échelle, de méthodes computationnelles plus efficaces et de données multi-omiques à l'échelle d'un patient permettraient de réaliser une analyse et des simulations plus robustes et renforceraient la puissance prédictive des modèles inférés dans cette étude, tout en ouvrant la voie à la médecine personnalisée. Cela permettrait de mieux comprendre les mécanismes impliqués dans la PR, ainsi que la réponse, à l'échelle d'un patient, à des traitements anti-TNF.

[333] [333]

Annexe A

Facteurs de transcription mis en évidence par CoRegNet et leur implication dans la PR selon la litérature tiré de Miagoux et al. 2021 [333]

Transcription factor	**Role in RA**	**References (PMID)**
TNFAIP3	NF-kB target gene, also involved in negative-feedback mechanism to block NF-kB activation through its ubiquitin-editing function in response to various inflammatory signaling, including TNF, IL-1ß	20822710, 22402800, 20852893, 26405544
IRF1	IRF1 is critical for the TNF-driven interferon response in rheumatoid fibroblast-like synoviocytes	31285419, 21834067, 32765497
ETS1	Factor involved in the cytokine-mediated inflammatory and destructive cascade which is a characteristic of RA	23101665, 11976735, 11229456
FOS	Subunit of AP1 transcription factor which is involved in the transcriptional regulation of many pro inflammatory genes in RA	8660103, 9153554, 19395871
NFKBIA	Involved in different pathways and cellular processes such as TNFa signalling via NFkB	30468518, 18454843
JUND	Subunit of AP1 transcription factor which is involved in the transcriptional regulation of many pro inflammatory genes	9764613, 17515956
HCLS1	Dysregulated in RA synovial tissue	12905466, 19563633
SPI1	Essential for the expression of gliostatin/thymidine phosphorylase in RA which has angiogenic and arthritogenic activities	22534375, 28192374
MXD1	Expressed in RA peripheral blood cells, RA synovium	22753658, 10568429
JUN	Subunit of AP1 transcription factor which is involved in the transcriptional regulation of many pro inflammatory genes in RA	18454843

(Suite)

(Suite)

Transcription factor	**Role in RA**	**References (PMID)**
NFKBIZ	Involved in TNF and IL-17 mediated signaling	32079724
TRIM22	Expressed in RA peripheral blood	24756903
FOSB	Subunit of AP1 transcription factor which is involved in the transcriptional regulation of many pro inflammatory genes in RA	29326694
DDX5	DDX5 is required for the transcription of key Th17 genes involved in Th17-mediated autoimmune inflammation in RA	29254845
BCL6	Interleukin-29 regulates T follicular helper cells by repressing BCL6 in RA	16508929, 28150777, 32468318
MNDA	Citrullinated protein identified in RA synovial fluid; Interferon induced nuclear and cytoplasmic protein	23044660, 15158620
EEF1A1	Expressed in RA peripheral blood	21444302
PTMA	Regulated by c-Myc, an oncoprotein overexpressed in synovium of RA, and is associated with cell proliferation	17372028 (mice)
DAZAP2	Expressed in RA peripheral blood mononuclear cells (PBMCs)	26352601

Annexe B

Liste des Communications

Écrites

Miagoux Q, Singh V, de Mézquita D, Chaudru V, Elati M, Petit-Teixeira E, Niarakis A. Inference of an Integrative, Executable Network for Rheumatoid Arthritis Combining Data-Driven Machine Learning Approaches and a State-of-the-Art Mechanistic Disease Map. Journal of Personalized Medicine. 2021; 11(8):785. https://doi.org/10.3390/jpm11080785

Zerrouk N, Miagoux Q, Dispot A, Elati M, Niarakis A. Identification of putative master regulators in rheumatoid arthritis synovial fibroblasts using gene expression data and network inference. Sci Rep. 2020;10(1):16236. Published 2020 Oct 1. doi:10.1038/s41598-020-73147-4

Posters

Miagoux Q, de Mezquita D, Singh V, Chalabi S, Petit-Teixeira E, & Niarakis A. (2020). Combining bottom-up and top-down systems biology methods to obtain an integrative, global RA-specific network (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.4266124

Miagoux Q, Veyssiere M, Niarakis A, Petit-Teixeira E, Chaudru V. Caractérisation de CNV (variants de nombre de copies) à partir de données de séquences exoniques simulées. JOBIM 2019, Nantes, 2-5 juillet, Cité des congrès.

Miagoux Q, Zerrouk N, Singh V, Petit-Teixeira E, Niarakis A. Gene co-regulatory

inference and modelling of perturbations effects on Rheumatoid Arthritis phenotypes. Genome Campus Advanced Course: Systems Biology: From large datasets to biological insight 08 Jul 2019 - 12 Jul 2019. Wellcome Trust Advanced Courses Wellcome, Cambridge, England.

Annexe C

Publication

Inference of an Integrative, Executable Network for Rheumatoid Arthritis Combining Data-Driven Machine Learning Approaches and a State-of-the-Art Mechanistic Disease Map

Journal of *Personalized Medicine*

MDPI

Article

Inference of an Integrative, Executable Network for Rheumatoid Arthritis Combining Data-Driven Machine Learning Approaches and a State-of-the-Art Mechanistic Disease Map

Quentin Miagoux [1], Vidisha Singh [1], Dereck de Mézquita [1], Valerie Chaudru [1], Mohamed Elati [2], Elisabeth Petit-Teixeira [1] and Anna Niarakis [1,3,*]

[1] Université Paris-Saclay, Univ Evry, Laboratoire Européen de Recherche pour la Polyarthrite rhumatoïde-Genhotel, 91057 Evry, France; quentin.miagoux@univ-evry.fr (Q.M.); vidisha.kumar@univ-evry.fr (V.S.); dereckdemezquita@gmail.com (D.d.M.); valerie.chaudru@univ-evry.fr (V.C.); elisabeth.teixeira@univ-evry.fr (E.P.-T.)
[2] CANTHER, University of Lille, CNRS UMR 1277, Inserm U9020, 59045 Lille, France; mohamed.elati@univ-lille.fr
[3] Lifeware Group, Inria, Saclay-île de France, 91120 Palaiseau, France
* Correspondence: anna.niaraki@univ-evry.fr

Abstract: Rheumatoid arthritis (RA) is a multifactorial, complex autoimmune disease that involves various genetic, environmental, and epigenetic factors. Systems biology approaches provide the means to study complex diseases by integrating different layers of biological information. Combining multiple data types can help compensate for missing or conflicting information and limit the possibility of false positives. In this work, we aim to unravel mechanisms governing the regulation of key transcription factors in RA and derive patient-specific models to gain more insights into the disease heterogeneity and the response to treatment. We first use publicly available transcriptomic datasets (peripheral blood) relative to RA and machine learning to create an RA-specific transcription factor (TF) co-regulatory network. The TF cooperativity network is subsequently enriched in signalling cascades and upstream regulators using a state-of-the-art, RA-specific molecular map. Then, the integrative network is used as a template to analyse patients' data regarding their response to anti-TNF treatment and identify master regulators and upstream cascades affected by the treatment. Finally, we use the Boolean formalism to simulate *in silico* subparts of the integrated network and identify combinations and conditions that can switch on or off the identified TFs, mimicking the effects of single and combined perturbations.

Keywords: network inference; integrative biology; rheumatoid arthritis; signaling cascades; gene regulation; transcription factors; Boolean simulations; systems biology

Citation: Miagoux, Q.; Singh, V.; de Mézquita, D.; Chaudru, V.; Elati, M.; Petit-Teixeira, E.; Niarakis, A. Inference of an Integrative, Executable Network for Rheumatoid Arthritis Combining Data-Driven Machine Learning Approaches and a State-of-the-Art Mechanistic Disease Map. *J. Pers. Med.* **2021**, *11*, 785. https://doi.org/10.3390/jpm11080785

Academic Editor: Hatem A. Elshabrawy

Received: 10 July 2021
Accepted: 10 August 2021
Published: 12 August 2021

1. Introduction

Rheumatoid arthritis (RA) is an inflammatory, autoimmune disease that affects the joints of the body. While the exact aetiology is unknown, it involves a combination of environmental and genetic factors such as smoking and susceptibility genes, along with sex and age factors. RA affects 0.5–1% of the world population, with women three times more susceptible to developing RA than men [1,2]. The onset of the disease is set around the fourth to fifth decade of one's life [3] and, if left untreated, it can be debilitating for the individual. Symptoms of RA include synovial inflammation, joint stiffness and pain, cartilage destruction, and bone erosion. In early RA, leukocytes invade the synovial joints, followed by other pro-inflammatory mediators, instigating an inflammatory cascade and provoking synovitis [2]. In addition, activated monocytes and T cells, both a source of pro-inflammatory cytokines such as TNF-a, can be found in peripheral blood [4], and many RA studies have used peripheral blood cells to identify disease-related genes [5–8].

The typical therapy for RA includes the use of disease-modifying anti-rheumatic drugs (DMARDs). Conventional DMARDs include drugs that target the entire immune system, whereas biologic DMARDs are monoclonal antibodies (mAbs) and soluble receptors that target protein messenger molecules or cells. Patients who do not respond to conventional DMARDs usually initiate therapy with TNF inhibitors. However, approximately 30–40% of RA patients fail to respond to anti-TNF therapy and are usually obliged to undergo several rounds of drug combinations [9]. Due to the complex nature of RA, systems biology and integrative approaches are needed to gain insight into the disease pathogenesis and progression. In addition, focusing only on one aspect of the disease provides a limited understanding of the multifactorial nature of RA.

Recently, many computational approaches, mainly network-based, which rely on integrating multi-omics data (proteomics, genomics, transcriptomics, and metabolomics), have succeeded in unravelling key mechanisms in complex diseases [10–13]. In this direction, machine learning is a promising bioinformatics field that allows the use and integration of various biomedical data with inherent complexity and large size. Furthermore, studies have shown that incorporating prior knowledge to data-driven methodologies improves the quality and the biological relevance of the outcome [14–16]. One such machine learning tool is CoRegNet, which is an R/Bioconductor package that infers co-regulatory networks of transcription factors (TFs) and target genes by analysing transcriptomic data and estimating TFs activity profiles. Moreover, the software also allows for network enrichment by integrating regulation evidence for TF binding sites, protein–protein interaction data, and chromatin immunoprecipitation (ChIP) data from various databases to support cooperative TFs [17]. In this work, we present a framework for integrating signalling and transcriptional regulation cascades with genomic mutations, combining data-driven approaches with prior knowledge in the form of an integrative RA-specific network. To do so, we use publicly available transcriptomic data of white blood cells from patients suffering from RA and the tool CoRegNet to infer a co-regulatory network.

Next, we develop an integration pairing method to couple the RA co-regulatory network with a state-of-the-art disease map for RA [18] to enrich the cooperativity network with upstream signalling regulators. Disease maps are comprehensive, knowledge-based representations of disease mechanisms, including disease-related molecular interactions supported by literature-based evidence [19,20]. Next, we project on the integrative RA network public genomic data and transcriptomic data from treated RA patients, highlighting key mutation carriers and differentially expressed genes associated with the response to anti-TNF treatment (Figure 1). The goal is to unravel mechanisms governing the regulation of key transcription factors and genes identified as mutation carriers or DEGs in RA patients undergoing anti-TNF treatment.

Lastly, we study the system's dynamic behaviour using Boolean formalism to simulate subparts of the integrated network [21,22]. We perform real-time simulations, sensitivity analysis, and dose–response analyses to study the impact of other signalling cascades on the expression of the identified TFs, and steady-state analysis revealing combinations and conditions that can switch on or off the identified TFs, mimicking the effects of the treatment [23].

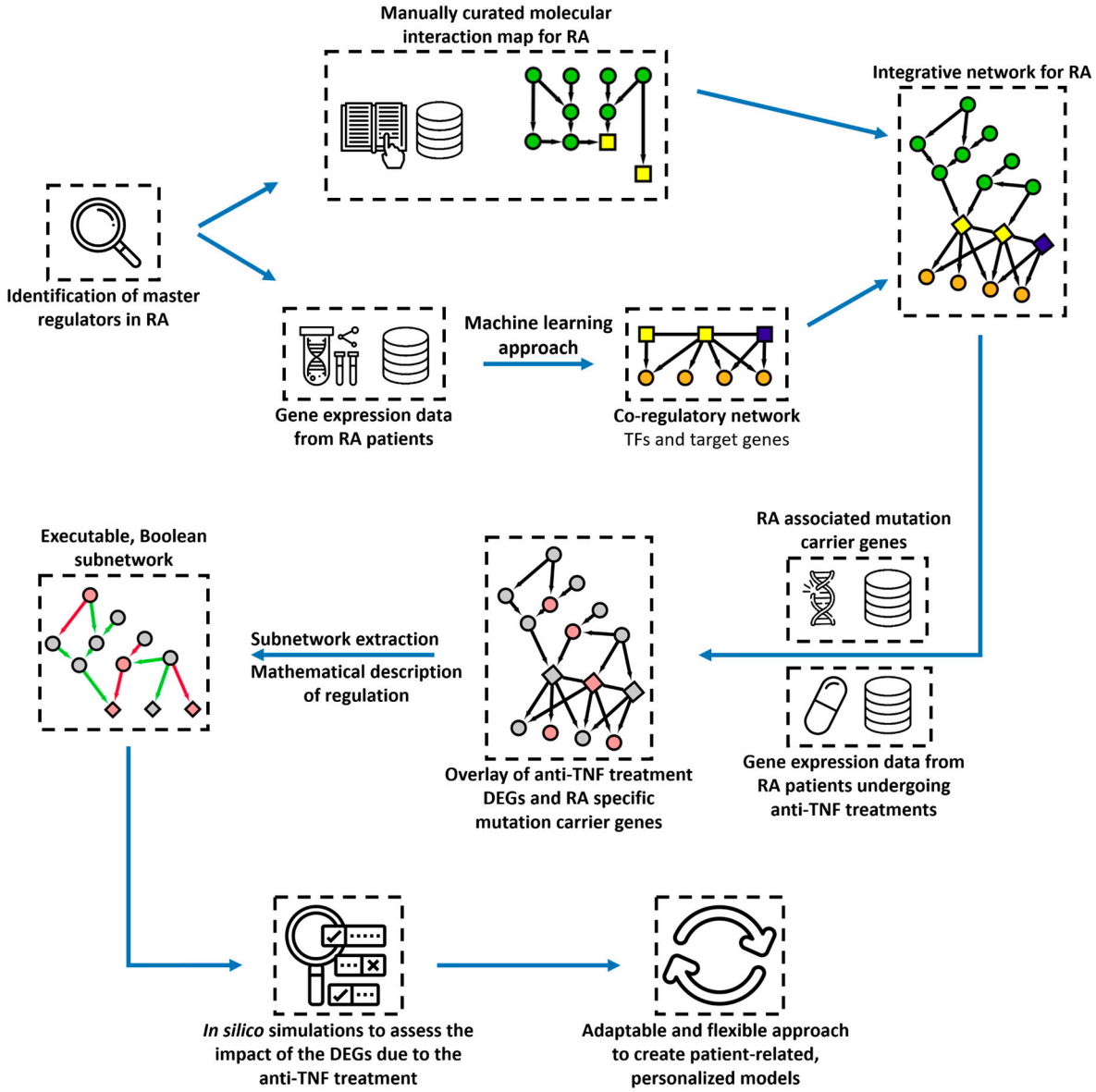

Figure 1. Workflow for creating an integrative and executable network for rheumatoid arthritis (RA). The steps include network inference and identification of master regulators, the combination of the co-regulatory network with curated signalling cascades, the analysis of omics data of anti-TNF treatment, and the addition of Boolean rules to create adaptable and flexible personalised models for *in silico* simulations.

2. Materials and Methods

2.1. Data Description and Pre-Processing

We used a transcriptomic dataset of white blood cells from RA patients and healthy donors (GSE117769) sequenced at CSL Limited/bio21 Institute (30 Flemington Rd, Parkville, Australia) using the Illumina HiSeq 2500 platform (Illumina, Inc.). The dataset describes 120 subjects in total (51 RA and 50 controls, and 19 patients with either ankylosing spondylitis or psoriatic arthritis). From this dataset, we extracted 96 samples (46 RA and 50 controls). For the 46 RA samples, 43 were of Caucasian ancestry and 3 were of unknown ancestry. Moreover, two samples consisted of duplicates of the exact origin (female, control, unknown ancestry), and their expression matrix average was used for the analysis. We conducted a preliminary analysis on the expression matrix data with DESeq2 version 1.32.0 [24], using normalisation and variance stabilising transformation on the matrix expression. Principal Component Analysis (PCA) revealed five outliers (4 RA and one control) that were removed from the dataset (Figure S1). The final dataset used for further analysis comprised

90 samples (42 RA and 48 controls). Finally, we performed a normalisation on the raw expression matrix after removing low read counts (>10).

2.2. Inference of the Co-Regulatory Network (CoRegNet)

We used CoRegNet [17] R package version 1.26.0 to infer the co-regulatory network with the normalised gene expression matrix of the pre-processing step. The CoRegNet package implements the H-LICORN algorithm, allowing identifying cooperative gene regulators [25]. We enriched the inferred network with protein–protein and regulation evidence and further refined it with an unsupervised method using the unweighted mean. The co-regulatory network inferred with CoRegNet is composed of the significant edges between TF with a False Discovery Rate (FDR) [26] of 5%.

2.3. RA Map Upstream Protein Extraction

The RA map is a state-of-the-art interactive knowledge base for the RA disease [18]. The RA map is organised in the form of a cell representing the flow of information from the extracellular space to the plasma membrane and then to the cytoplasm, the nucleus, and the secreted compartment or cellular phenotypes. For our analysis, as we were focused on upstream regulators of identified TFs, we mainly used the RA map's signalling part. More specifically, the list of TFs was uploaded as an overlay to the RA map, and the matching TFs were identified. The matching TFs were subsequently used as seeding nodes for the upstream plugin [27], setting the mode of extraction as upstream and selecting non-blocking modifiers. Finally, the obtained file was extracted as an XML CellDesigner file.

With this method, we extracted upstream signalling cascades and seven translation reactions for which the mRNA was directly linked with the protein in the cytoplasm or the membrane. The RA map, and consequently the extracted network, is written in the Process Description Systems Biology Graphical Notation scheme [28]. To obtain a more simplified representation of the network, the CellDesigner XML file of the previous step was used as an input to the tool CaSQ [29] to create an Activity Flow (AF)-like executable network. CaSQ provides SBML-qual files for performing *in silico* simulations, but in our case, we used only the SIF file that contains information about the source, interaction type, and target of the Boolean network.

The obtained SIF file was further modified to address the issue of complexes. First, we recreated the reactants for every complex represented as a single node in the AF network. This way, we would not miss interactions and overlaps between nodes existing inside complexes. On the other hand, regarding entities represented multiple times (as genes, proteins, or mRNAs), we kept only one entity for simplification purposes and merged the corresponding interactions.

2.4. Global RA Network Inference

We used the R package igraph [30] to convert the CoRegNet object (co-regulatory network) and the RA map SIF file into separate graphs. Then, we merged both networks using igraph and imported the network into Cytoscape using the RCy3 R/Bioconductor package [31], forming the global RA-specific network.

2.5. Differential Expression Analysis (DEA) Using Independent Datasets

We conducted multiple DEA using two different datasets. One dataset contained normalised counts from RNA sequencing data of CD4+ T cells, including different responses of RA patients to anti-TNF treatment (GSE138747). The dataset comprises two cohorts of RA patients treated with adalimumab (37 patients) and etanercept (41 patients), which were analysed independently. The second dataset comprises raw counts from RNA sequencing data of whole blood cells of biologic naive RA patients from baseline and after three months of treatment with infliximab or adalimumab (GSE129705). This dataset contains two different cohorts of 40 and 36 RA patients, which were also analysed independently. Thus, using DESeq2, we conducted two DEA on comparing responders and non-responders'

gene expression levels for both drugs (adalimumab and etanercept) and two DEA on the comparison of baseline and after three months of anti-TNF treatment gene expression level. We considered as differentially expressed genes (DEG) the ones with a corrected *p*-value (FDR) < 0.1 for all performed analyses. The DEG lists were used as an overlay for the global RA-specific network.

2.6. List of Variants

DisGeNET [32] contains the most extensive publicly available collection of genes and variants associated with human diseases. From this database, we extracted 2387 variants associated with RA. Then, we filtered out variants with a variant disease association (VDA) and evidence index (EI) score lower than 0.7. The VDA score is computed using the number of curated and non-curated publications supporting the variant disease association, while the EI score is computed using contradictory results in publications supporting the variant. A 0.7 threshold gives us at least one curated publication supporting the variant and the disease association resulting in 1635 variants. Within these 1635 variants, we identified 731 associated genes that were subsequently used as an overlay for the global RA-specific network.

2.7. Subnetwork Extraction

The subnetwork, based on the global network for RA, is focused on Tumor Necrosis Factor (TNF), Interleukin 6 (IL6), and Transforming Growth Factor Beta 1 (TGFB1), which are three molecules highly implicated in RA (see Section 4). These three proteins were extracted with their downstream cascades up to the first TF to reduce complexity and focus on the upstream regulators. From the global network for RA displayed in Cytoscape, we selected TNF, IL6, and TGFB1 simultaneously and using the Biological Network Manager (BiNoM) plugin, we selected in a stepwise manner the downstream neighbours of TGFB1, IL6, and TNF up to the first affected TF(s).

2.8. Shiny App

The co-regulatory network inferred with CoRegNet, the RA map Activity Flow extracted network, and the merged global RA network, along with their overlays, were integrated into a web-based Shiny application using R [33]. The web application uses a Cytoscape viewer based on the R package cyjShiny [34] and is freely available (https://quentin-miagoux.shinyapps.io/global_ra_network (accessed on 1 July 2021)).

2.9. Inference of a Boolean Network for In Silico Simulations

The subnetwork obtained in Section 2.7 was imported in Cytoscape and exported in SBML format using the BiNoM plugin and its function "export to SBML". The file was subsequently imported in CellDesigner to adjust the layout and remove co-regulatory interactions between TFs to focus only on upstream regulators. Finally, the CellDesigner SBML file was used to infer a Boolean model in an SMBL-Qual format using CaSQ (CellDesigner as SBML-Qual) v0.9.11.

Then, the SMBL-Qual file was imported into Cell Collective to perform real-time simulation experiments. Using the "*Simulation*" tab on Cell Collective, we mimicked the down-regulation of the components in the datasets used (before/after treatment, responders/non-responders, mutation carriers) under different initial conditions for each input (TNF, IL6, and TGFB1). Furthermore, we performed sensitivity and dose–response analyses using five different initial conditions described in Table 3.

The same SBML-Qual file was also used for analysis with the software GINsim [35] after a post-processing modification step for node name recognition. We used the nightly build version 3.0.0b-SNAPSHOT, and the functions *Reduce model* for the reduced version, *Compute stable states* to obtain the stable states of the model and *Run simulation* with configurations of perturbations for the *in silico* knock out (KO) experiments.

3. Results

3.1. Inference of the Co-Regulatory Network

We selected a transcriptomic dataset from the GEO database (GSE117769) to infer the co-regulatory network, including 120 samples (51 RA and 50 control, and 19 patients with either ankylosing spondylitis or psoriatic arthritis). After a series of pre-processing checks, including the sample origins, duplicates, and quality of the data using a PCA on the matrix expression with normalisation and variance stabilising transformation (shown in Figure S1), we kept for further analysis a total of 90 samples (48 Controls and 42 RA patients). Then, of the remaining samples, we obtained normalised counts using DESeq2, on which we finally applied CoRegNet to infer the co-regulatory network, which is presented in Figure 2.

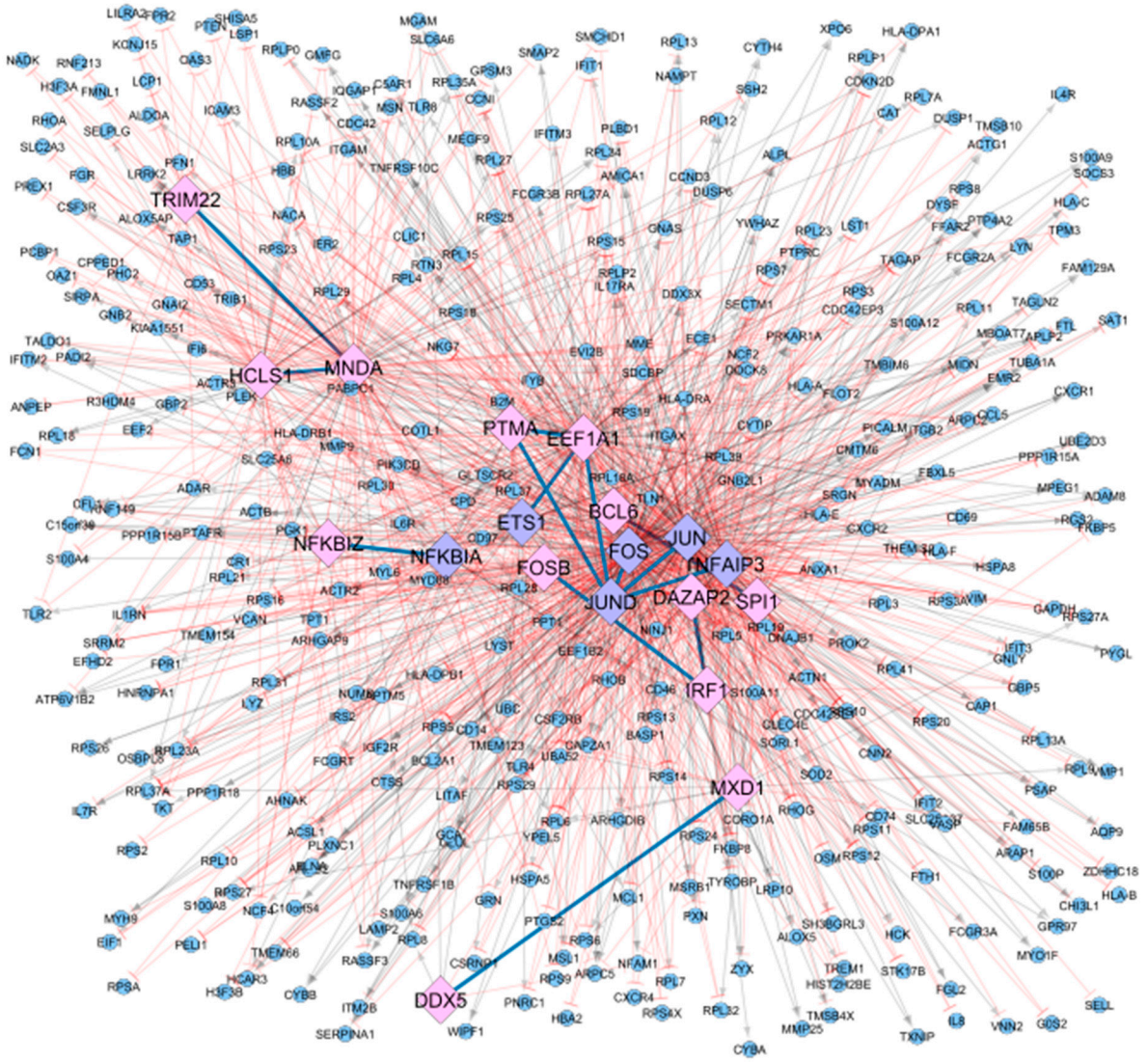

Figure 2. Co-regulatory network inferred using the tool CoRegNet and the matrix of normalised counts from the transcriptomic dataset (GSE117769). The dataset used included data from 90 samples (48 Controls and 42 RA patients). Matching TFs from CoRegNet with the RA map are depicted using a diamond shape and coloured in purple (6), while non-matching TFs are depicted in a diamond shape and coloured in pink (13). CoRegNet target genes are coloured in blue (373) and have round shapes. Inhibitions are represented with blunt red arrows, and activations are represented with grey arrows.

This network includes a total of 19 TFs, 14 co-regulatory interactions, and a total of 373 regulated target genes. Table 1 summarises the top five TFs with the highest number of regulatory and co-regulatory interactions. The literature search for the nineteen TFs identified from CoRegNet as the master regulators in the dataset showed their potential

implication to RA. Supplementary Table S1 summarises key roles of the TFs and the corresponding literature reference.

Table 1. Top 5 of the CoRegNet identified transcription factors (TFs) with the highest number of regulatory interactions (TF–target gene) and co-regulatory interactions (TF–TF). The analysis was performed using data from 90 samples, 48 Controls, and 42 RA patients.

Top 5 TFs	No. of Regulatory Interactions
FOS	288
JUN	211
EEF1A1	155
MNDA	136
TNFAIP3	125
Top 5 TFs	**No. of Co-Regulatory Interaction(s)**
JUND	5 (EEF1A1, FOS, JUN, PTMA, TNFAIP3)
EEF1A1	3 (ETS1, JUND, PTMA)
IRF1	2 (DAZAP2, FOSB)
MNDA	2 (HCLS1, TRIM22)
PTMA	2 (EEF1A1, JUND)

3.2. RA Map Upstream Regulators of the TFs Identified from CoRegNet

Six out of the 19 TFs, namely ETS Proto-Oncogene 1, Transcription Factor (ETS1), Fos Proto-Oncogene, AP-1 Transcription Factor Subunit (FOS), Jun proto-oncogene, AP-1 transcription factor subunit (JUN), JunD Proto-Oncogene, AP-1 Transcription Factor Subunit (JUND), NFKB Inhibitor Alpha (NFKBIA), and TNF Alpha Induced Protein 3 (TNFAIP3) from the co-regulatory network are present in the RA map. Therefore, they were used as seeds to extract their upstream regulators. The extracted network comprising the RA map upstream regulators of the matching TFs includes 244 nodes, as shown in Figure 3.

3.3. Coupling Gene Co-Regulation with Signalling Cascades to Obtain a Global, Integrative RA Network

The global, integrative RA network results from merging the RA map signalling cascades and the CoRegNet object, using as an interface the matching TFs. It comprises 614 nodes and 1736 interactions (848 inhibitions, 874 activations, and 14 co-regulatory interactions shared among TFs), including genes, proteins, complexes, and simple molecules shown in Figure 4. In this network, six TFs were shared between the CoRegNet network and the RA map (seeding TFs). In addition, 16 target genes identified with CoRegNet overlapped with the RA map upstream regulators.

3.4. Two Use Cases: Identification of Key TFs Using DEG from RA Patients Undergoing Anti-TNF Treatment

Two datasets of RNAseq expression data coming from RA patients undergoing anti-TNF treatment were analysed to obtain DEG. The first dataset focuses on responders and non-responders to RA treatment, including 37 and 41 RA patients treated with adalimumab and etanercept, respectively. The second one involves untreated and treated (infliximab or adalimumab) RA patients, including two cohorts of 40 and 36 RA patients. DEGs from these analyses were mapped to the global network for RA (presented in Figures S2 and S3, respectively).

DEGs from responders/non-responders RA patients data mapping shows a total of 15 matching nodes, including 4 etanercept DEGs and 11 adalimumab DEGs. In addition, four matching nodes are CoRegNet and RA map TFs (NFKBIA, JUN, FOS, and TNFAIP3) and 1 CoRegNet TF only (FOSB), in the global network for RA DEGs from untreated and treated RA patients mapping show a total of 101 matching nodes, including 2 CoRegNet and RA map TFs (NFKBIA and FOS) and 4 CoRegNet TF only (BCL6 Transcription Repres-

sor (BCL6), MAX Dimerisation Protein 1 (MXD1), Myeloid Cell Nuclear Differentiation Antigen (MNDA), and DAZ-Associated Protein 2 (DAZAP2)).

Finally, cross-analysis revealed that a total of 9 over 19 TFs included in the global network for RA overlapped with a DEG from at least one analysis (presented in Table 2). Among these 9 TFs, two of them, NFKBIA and FOS, overlapped with a DEG in both analyses.

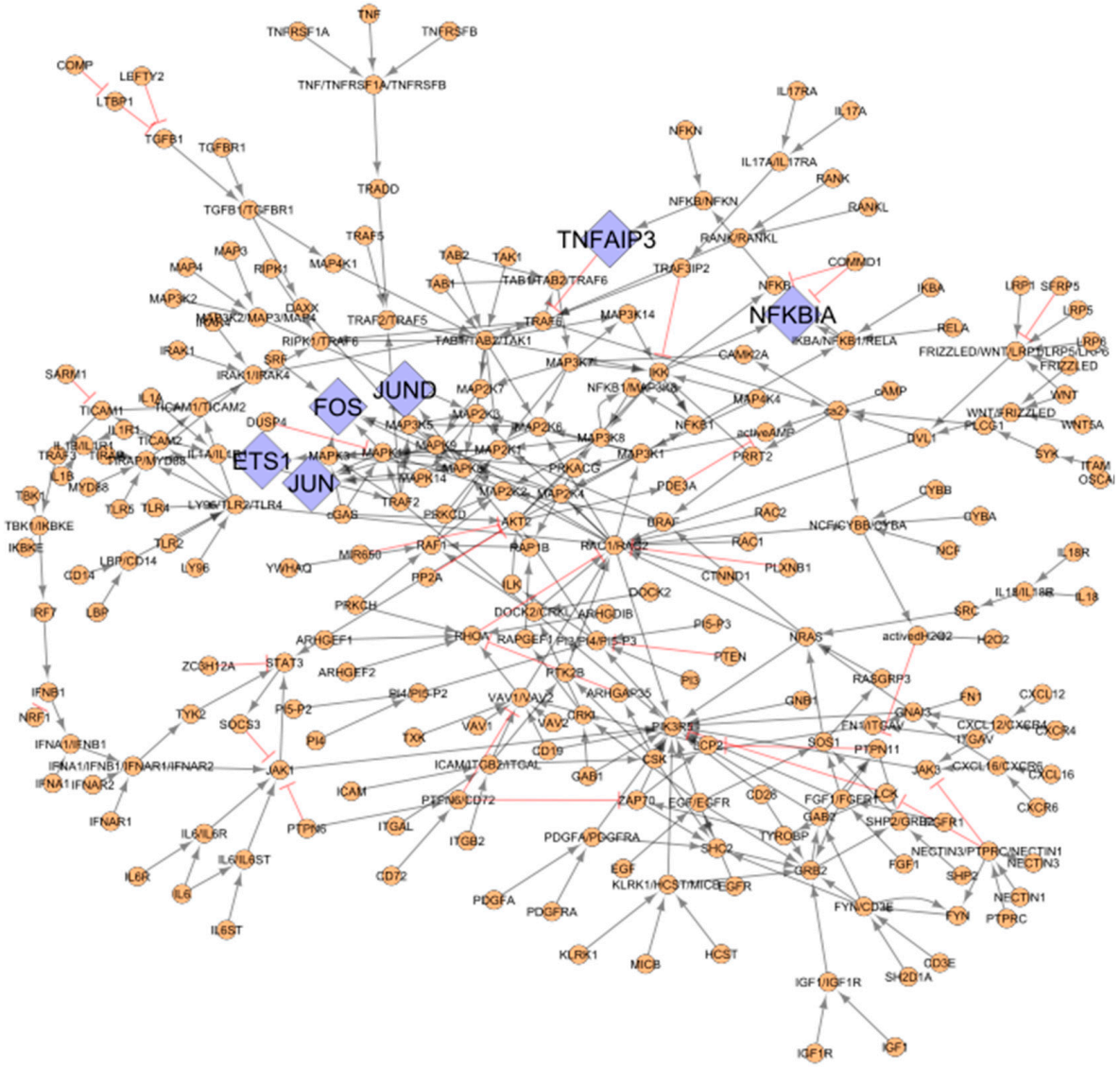

Figure 3. The upstream regulators of the matching TFs between the RA map and the CoRegNet co-regulatory network. Matching TFs from the inferred network with CoRegNet are coloured in purple (diamond shape) (6), and upstream regulators (round shape) are coloured in orange (238). Inhibitions are represented with blunt red arrows and activations with grey arrows.

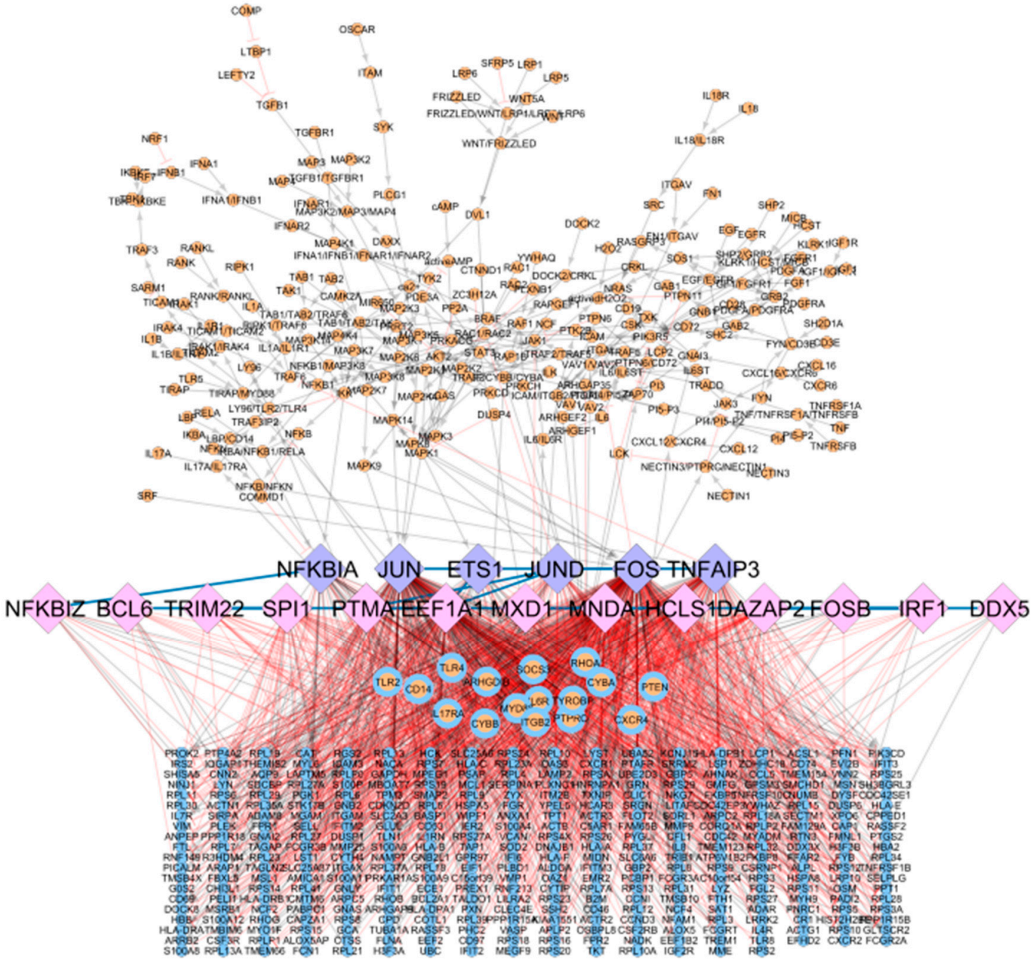

Figure 4. Integrative global network for rheumatoid arthritis. RA map upstream regulators are bound to TFs identified in the CoRegNet co-regulatory network. Matching TFs are coloured in purple (6), matching genes/proteins are coloured in blue and orange (16), upstream regulators from the RA map are coloured in orange (222), CoRegNet TFs are coloured in pink (13), and CoRegNet target genes are coloured in blue (357). Inhibitions are represented with blunt red arrows, and activations are represented with grey arrows. Transcription factors are depicted in diamond shapes, while upstream regulators and target genes are depicted using round shapes.

Table 2. Transcription factors from the global network for RA with at least one differentially expressed gene overlapping from the analysis of (a) responders/non-responders RA patients to anti-TNF treatment (37 and 41 RA patients treated with adalimumab and etanercept, respectively) and (b) after/before anti-TNF treatment of RA patients (two different cohorts of 40 and 36 RA patients from baseline and after three months treatment with Infliximab or Adalimumab). ↓ denotes downregulation.

Source	TF	Responders/Non-Responders	After Treatment/Before Treatment
CoRegNet and RA map	NFKBIA	↓	↓
	JUN		↓
	FOS	↓	↓
CoRegNet	TNFAIP3		↓
	BCL6	↓	
	MXD1	↓	
	MNDA	↓	
	DAZAP2	↓	
	FOSB		↓
	NFKBIA	↓	↓
	JUN		↓

3.5. Logic-Based Dynamical Analysis of the Subnetwork

While the analyses highlight TFs differentially expressed (downregulated) after treatment or response to treatment with anti-TNF drugs, it is evident from the global network that the identified TFs can be regulated by a variety of other upstream cascades, besides those implicated in the TNF signalling.

To study further the interconnections with other pathways, we focused on IL6 and TGF-beta signalling (mentioned as TGFB1 in the network). IL6 is the target of tocilizumab (TCZ), which is an IL6 inhibitor frequently used in the treatment of RA. The inhibitor was developed in 2008, and its therapeutic efficacy is quite similar to those of TNF inhibitors [36]. TGF-beta signalling is activated in RA synovium; however, TGF-beta blockade did not seem to affect experimental arthritis [37]. To study further the impact of these cascades on the expression of the identified TFs, we constructed a subnetwork by selecting the molecules TGFB1, IL6, and TNF in the global network along with their downstream neighbours until reaching an identified TF.

The subnetwork contains 38 nodes, including 4 TFs and is highly enriched in MAPKs, as seen in Figure 5. By projecting the DEGs and known genomic variants associated with the disease, we can see that intermediate nodes and TFs are downregulated by the anti-TNF treatment, while the inputs (IL6, TNF, and TGFB1) along with a few intermediate nodes are characterised as mutation carriers.

In the next step, we wanted to evaluate the impact of single and combined perturbations of the network inputs on the expression of the TFs. We used the possibility of adding Boolean rules to the network with the tool CaSQ [29]. Boolean models have been long used to describe biological mechanisms in health and disease [38], and they are an optimal approach for modelling signalling and gene regulation when kinetic parameters are scarce. Boolean models use binary values and logical operators (AND, OR, and NOT) to describe the regulation of all molecules in the system [21]. The CaSQ tool receives an SBML CellDesigner file [39] and produces a Boolean network with preliminary logical rules.

The Boolean model produced from our subnetwork has 38 nodes (three inputs, six outputs and 29 intermediate nodes) and 59 interactions. The SBML qual file was imported to Cell Collective [40] to perform real-time simulations and sensitivity analysis and GINsim [35] to calculate stable states and perform *in silico* KO simulations. For the *in silico* KO simulations, a reduced version of 23 nodes was used.

3.6. Real-Time Simulations Using the Cell Collective Platform

First, we wanted to see the impact of the molecules, either affected by the treatment or identified as mutation carriers, on the model outputs. Before and after anti-TNF treatment, the analysis showed that Mitogen-Activated Protein Kinases such as MAPK14 and MAPK1 were downregulated. Accordingly, for the responders and non-responders' analysis, Mitogen-Activated Protein Kinase Kinase 1 (MAP2K1), Integrin Linked Kinase (ILK), and Death Domain-Associated Protein (DAXX) were also identified as downregulated. Lastly, DAXX and Nuclear Factor Kappa B Subunit 1 (NFKB1) were identified as mutation carriers. To mimic the effects of the downregulation of these molecules on the model outputs, we performed *in silico* simulations setting their activation level to zero.

For the dataset of before and after anti-TNF treatment of RA patients, MAPK14 and MAPK1 activity levels were set to zero, and simulations turning the inputs sequentially active revealed that when setting either TNF, TGFB1, or IL6 on, all TFs are expressed (Figure 6a–c). Furthermore, when mimicking the downregulation of MAP2K1, ILK, and DAXX for the dataset of responders/non-responders to anti-TNF treatment, we observed that when setting TNF and TGFB1 on, all TFs are expressed (Figure 6d,e). However, when IL6 is set on, only the TF NFKBIA is expressed (Figure 6f). Finally, when mimicking the downregulation of DAXX and NFKB1 for the mutation carrier from DisGeNET, we observe that when setting either TNF, TGFB1, or IL6 on, all TFs are expressed (Figure 6g–i).

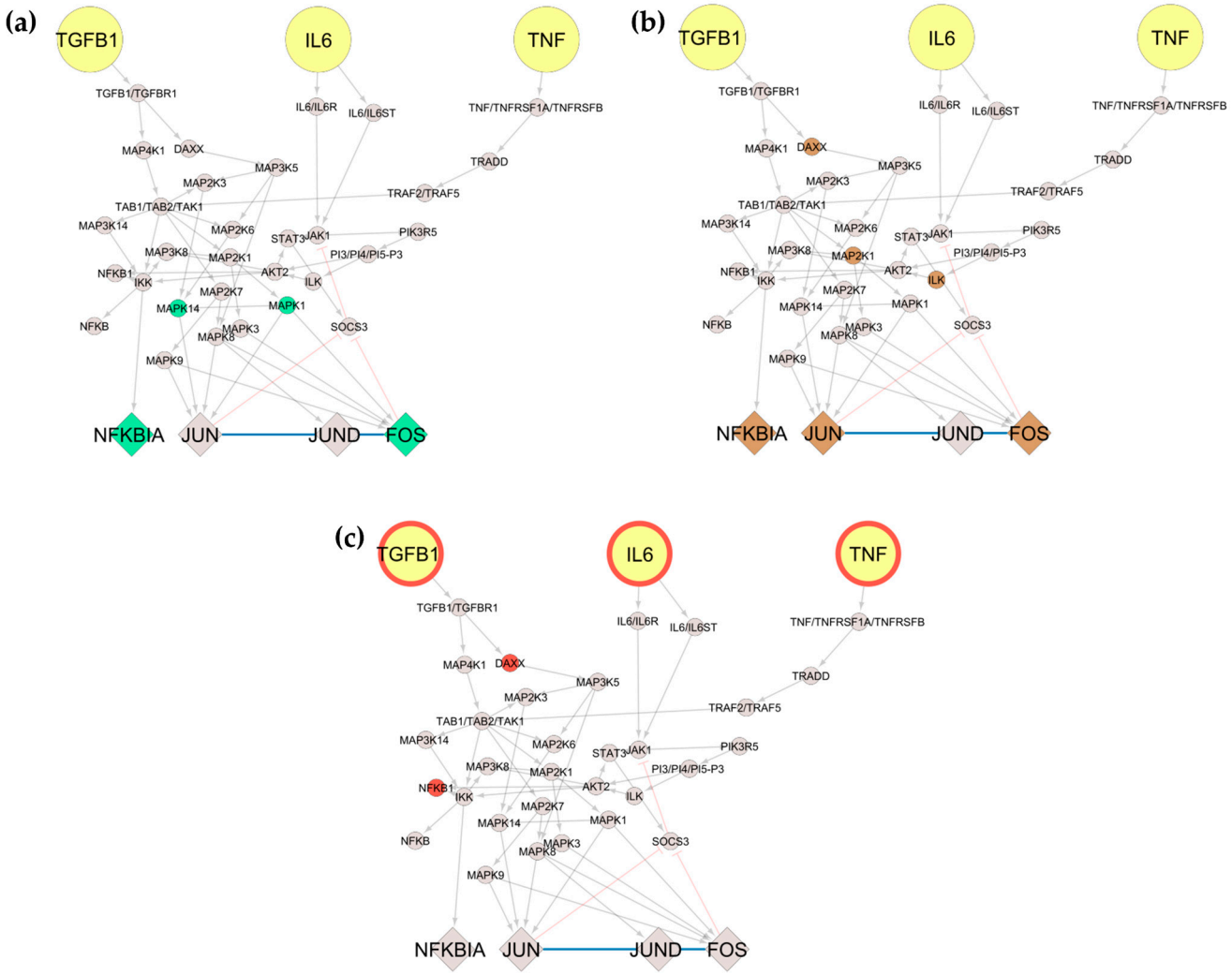

Figure 5. Subnetwork extraction of target molecules (TGFB1, IL6, and TNF) from the global network for RA. We focused on TGFB1, IL6, and TNF (coloured in yellow) and their downstream neighbours (depicted in round shapes) until reaching the first transcription factor (depicted in diamond shapes). Then, we projected on the network (**a**) DEG from responders/non-responders RA patients to anti-TNF treatment, as shown in green (4); (**b**) DEG after/before anti-TNF treatment of RA patients is shown in brown (6); and (**c**) DisGeNET variants are shown in red (5).

3.7. Dose–Response and Sensitivity Analysis

For the dose–response, we studied five different initial conditions shown in Table 3 that mimic different scenarios' effects in combination with TNF activity status. The first condition corresponds to TNF blockade and simultaneous impairment of IL6 and TGFB1 signalling; the second corresponds to having the IL6 cascade active, the third to having the TGFB1 active, the fourth to having both IL6 and TGFB1 active, and lastly, the fifth condition mimics what happens to the system when only the TNF input is active.

Table 3. Initial conditions for dose–response analysis.

Initial Conditions	Input State
1	All inputs inactive
2	IL6 active
3	TGFB1 active
4	IL6 + TGFB1 active
5	TNF active

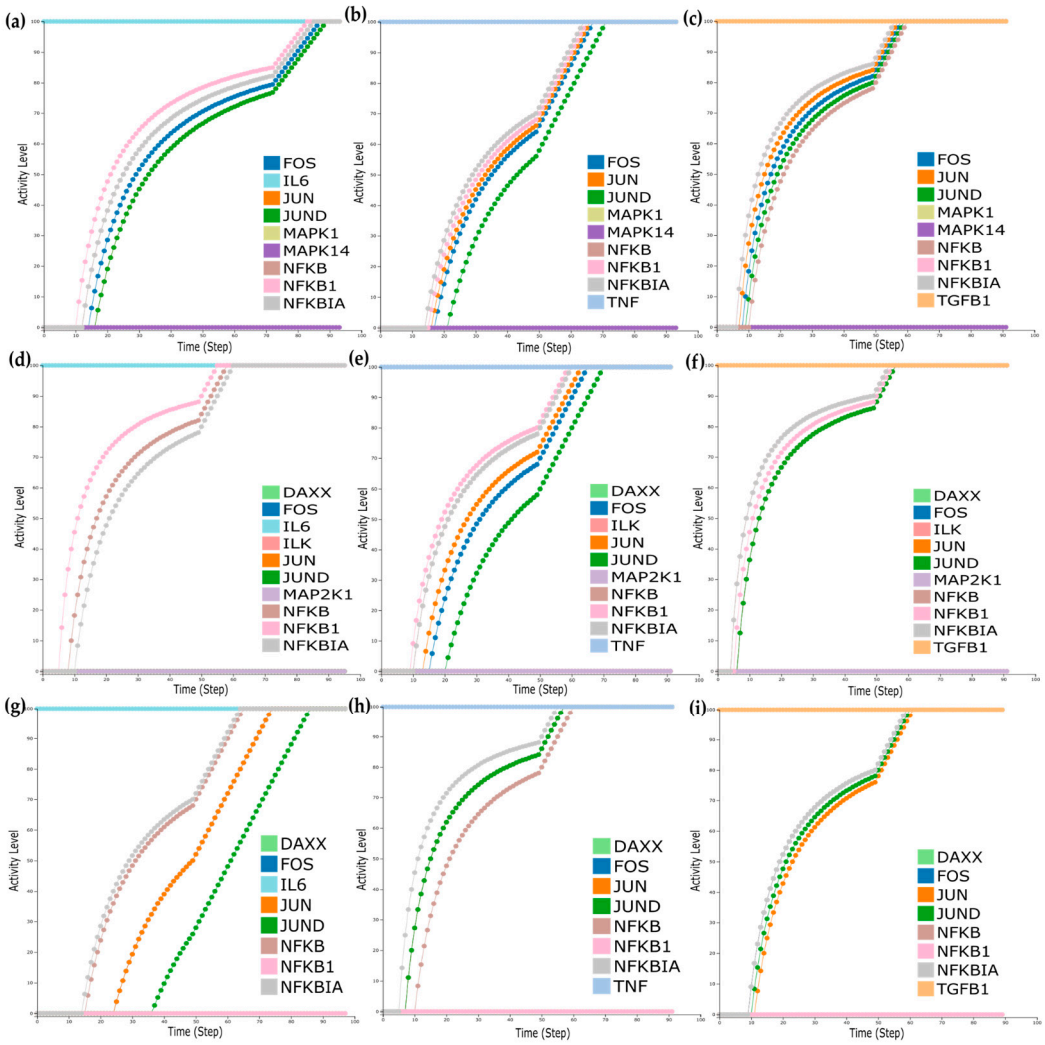

Figure 6. Real-time simulations with Cell Collective for the dataset of before and after anti-TNF treatment of RA patients, where MAPK14 and MAPK1 were found to be downregulated (**a**–**c**); (**a**) simulations with IL6 activity set to 100%, (**b**) simulations with TNF activity set to 100%, (**c**) simulations with TGFB1 activity set to 100%. Real-time simulations for the dataset of responders/non-responders where MAP2K1, ILK, and DAXX were found downregulated (**d**–**f**); (**d**) simulations with IL6 activity set to 100%, (**e**) simulations with TNF activity set to 100%, (**f**) simulations with TGFB1 activity set to 100%. Real-time simulations for the dataset of the mutation carrier. (**g**–**i**); (**g**) simulations with IL6 activity set to 100%, (**h**) simulations with TNF activity set to 100%, (**i**) simulations with TGFB1 activity set to 100%.

We performed dose–response analysis for all conditions and observed that the expression of the TFs is dose-dependent for TNF, TGFB1, and IL6 (Figure 7b,e,f), while the simultaneous activation of IL6 and TGFB1 cascades has a synergistic effect causing an increase on the activation levels of the TFs even for lower doses of IL6 and TGFB1 (Figure 7c,d).

Next, we wanted to see how the downregulation of the TFs observed after the anti-*TNF* treatment could be counterbalanced by the other pathways, given that in non-responders, the expression of these TFs was kept intact, despite the administered treatment. Therefore, we performed an environment sensitivity analysis to identify which of the two model inputs (IL6 and TGFB1) has the most significant impact on the up-regulation of the TFs included in the model when TNF activity is blocked. The results showed that the TFs could be upregulated in the absence of TNF activity for a combination of activity ranges of the other two inputs (Supplementary Materials: Figure S4).

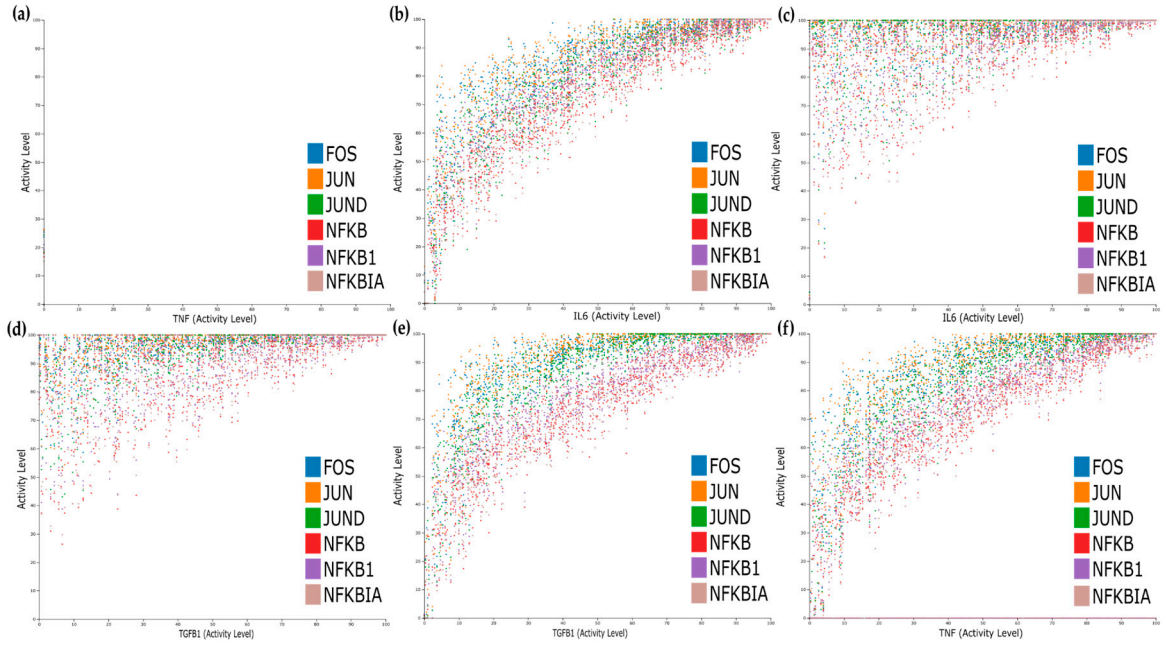

Figure 7. Dose–response analysis. (**a**) All inputs inactive. (**b**) IL6 active. (**c**) IL6 and TGFB1 active (TGFB1 view). (**d**) IL6 and TGFB1 active (IL6 view). (**e**) TGFB1 active. (**f**) TNF active.

3.8. Wild-Type Stable-State Analysis and KO Simulations

We performed stable-state analysis for the model using the software GINsim (Figure 8). The analysis for the wild type (no perturbations) revealed five steady states (fixed points) and no complex attractor. The configurations of these five stable states as far as the molecules of interest are concerned (grey, blue, and pink nodes of Figure 8) are shown in Table 4.

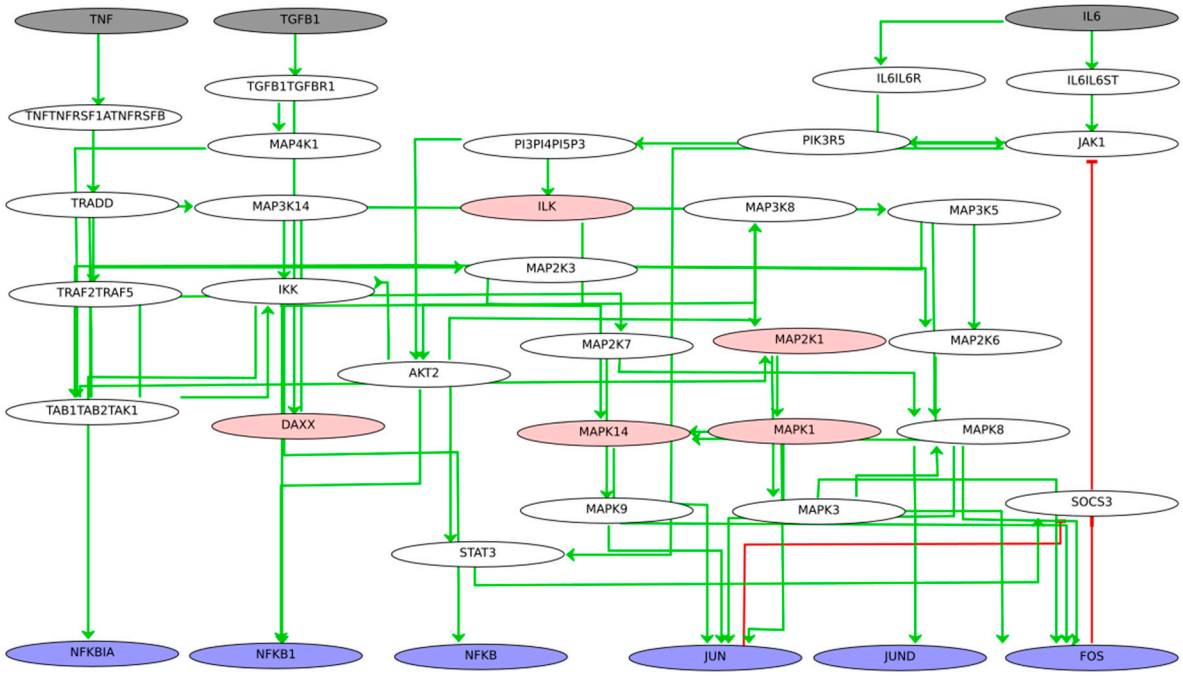

Figure 8. The Boolean network comprising the three signalling cascades for TNF, TGFB1, and IL6. Boolean rules were inferred using the tool CaSQ (see Section 2 for a step-by-step model inference). Inputs are depicted in grey, TFs of interest are depicted in purple, and intermediate nodes affected by the drug treatment or identified as mutation carriers are depicted in pink. Green arrows denote activation, and blunt red arrows denote inhibition.

Table 4. Stable states of the Boolean network (wild type).

Steady States	TNF	IL6	TGFB1	JUN	FOS	JUND	NFKBIA	DAXX	ILK	NFKB1	MAP2K1	MAPK1	MAPK14
ss1	0	0	0	0	0	0	0	0	0	0	0	0	0
ss2	0	1	0	1	1	1	1	0	1	1	1	1	1
ss3	0	1	1	1	1	1	1	1	1	1	1	1	1
ss4	1	1	0	1	1	1	1	0	1	1	1	1	1
ss5	1	1	1	1	1	1	1	1	1	1	1	1	1

The analysis shows that for the TFs identified as master regulators, the activation of IL6 or IL6 and TGFB1 can positively regulate their expression, even in the presence of the anti-TNF treatment (TNF = 0). Blocking the TNF cascade can completely shut down their expression only if combined with the blocking of IL6 and TGFB1 cascades. Regarding DAXX, it is actively expressed only when TGFB1 is activated, and ILK is dependent on the activation of IL6. The MAPK molecules are dependent on the activation of IL6 and TGFB1 and do not seem to be impacted by TNF blocking.

Next, we created a reduced version of the Boolean modelling using the reduction function of the software GINsim to perform *in silico* experiments with combined perturbations. The reduced Boolean model comprised 23 nodes, and for the analysis, we created virtual KOs for (a) MAPK14 and MAPK1, (b) DAXX, ILK, and MAP2K1, and (c) DAXX and NFKB1 to mimic the effects of the anti-TNF treatment and the mutation carriers, which were identified previously. For the simulations, we set the initial conditions for the TNF to zero and let the other inputs free, while setting the initial condition for all intermediate nodes to zero.

The results of the *in silico* experiments for the molecules of interest and the three conditions are shown in Tables 5–7. For each set of conditions, the system was able to reach three steady states. For the first set of conditions, we observe that besides DAXX that is strictly TGFB1 dependent (Table 5, ss2 and ss3), all TFs can get activated with the presence of IL6 or IL6 and TGFB1, despite the TNF blockade and the downregulation of MAPK14 and MAPK1.

Table 5. Stable states of the Boolean network (MAPK1, MAPK14 KO, input TNF = 0).

Steady States	TNF	IL6	TGFB1	JUN	FOS	JUND	NFKBIA	DAXX	ILK	NFKB1	MAP2K1	MAPK1	MAPK14
ss1	0	0	0	0	0	0	0	0	0	0	0	0	0
ss2	0	1	0	1	1	1	1	0	1	1	1	0	0
ss3	0	1	1	1	1	1	1	1	1	1	1	0	0

Table 6. Stable states of the Boolean network (DAXX, ILK and MAP2K1 KO, input TNF = 0).

Steady States	TNF	IL6	TGFB1	JUN	FOS	JUND	NFKBIA	DAXX	ILK	NFKB1	MAP2K1	MAPK1	MAPK14
ss1	0	0	0	0	0	0	0	0	0	0	0	0	0
ss2	0	1	0	0	0	0	1	0	0	1	0	0	0
ss3	0	1	1	1	1	1	1	0	0	1	0	0	1

Table 7. Stable states of the Boolean network (DAXX and NFKB1 KO, input *TNF* = 0).

Steady States	TNF	IL6	TGFB1	JUN	FOS	JUND	NFKBIA	DAXX	ILK	NFKB1	MAP2K1	MAPK1	MAPK14
ss1	0	0	0	0	0	0	0	0	0	0	0	0	0
ss2	0	1	0	1	1	1	1	0	1	0	1	1	1
ss3	0	1	1	1	1	1	1	0	1	0	1	1	1

For the second set of conditions, we observe that when TNF is blocked and DAXX, ILK, and MAP2K1 are downregulated, the IL6 signal alone is not enough to activate the TFs JUN, FOS, and JUND and the kinases MAPK14 and MAPK1 (Table 6, ss2). However, when both IL6 and TGFB1 signals are on, all TFs are activated, and the activity level of kinase MAPK14 is restored (Table 6, ss3).

Lastly, we simulated the effects of the mutation carriers, as identified by DisGeNet, and the TNF blockade on the activity of the TFs and the kinases in our network. In Table 7, we observe that despite TNF blockade and DAXX and NFKB1 downregulation, all identified TFs and kinases are activated in the presence of IL6 or for IL6 and TGFB1 combined activity.

4. Discussion

In the present work, we combine gene co-regulation with mechanistic signalling cascades to provide information about upstream regulation. Furthermore, we use the integrative RA network to analyse transcriptomic data regarding anti-TNF treatment and map information about known disease-associated mutation carriers. Lastly, we use the tool CaSQ to add Boolean dynamics to a subnetwork of interest to mimic the effects of the anti-TNF treatment and estimate the impact of IL6 and TGFB1 and the downregulated genes on the activation profile of the identified TFs.

The nineteen TFs identified as master regulators have been implicated in RA and autoimmunity, as the literature evidence supports. Six out of the nineteen TFs were also present in the RA map, which is a state-of-the-art mechanistic network for the disease built using manual curation. These six TFs, namely JUN, JUND, FOS, NFKBIA, ETS1, and TNFAIP3, were used as a functional overlap between the co-regulation and the signalling events, enabling us to obtain a network comprising upstream cascades, active TFs, and target genes.

We used the integrative network as a template to analyse two independent datasets regarding anti-TNF treatment. First, we observed the downregulation of some of the TFs previously identified as master regulators. Second, to study the impact of the treatment in parallel with the activity of other signalling cascades, we extracted subgraphs from the integrative network. Finally, we selected the cascades of TNF, IL6, and TGFB1, up to the first affected TF to reduce complexity and focus on the upstream regulators.

We selected IL6, as it is one of the targets of the biologic treatment in RA [36,41–43] and TGF-beta because it is an immunomodulatory cytokine highly expressed in RA patients, with a role that is yet to be determined [44–46]. We adjusted the map-to-model framework described in Aghamiri et al. [29] to obtain an executable Boolean subnetwork to perform *in silico* analysis. As demonstrated from the real-time simulations and the dose–response analysis, both IL6 and TGFB1 cascades could affect the expression of the TFs, and as seen from the component sensitivity analysis, IL6 and TGFB1 could even counterbalance the downregulation of the studied TFs caused by the TNF blockade.

The steady-state analysis confirmed the real-time simulation results showing that for the TFs identified as master regulators, the activation of IL6 or IL6 and TGFB1 cascades can positively regulate their expression. Blocking the TNF cascade can completely shut down the expression of these TFs only if combined with the blocking of IL6 and TGFB1 cascades. Towards this direction, dual-targeted therapies have been proposed, either with the development of dual-target agents, blocking IL6 and TNF simultaneously, for example [47], or by administering two biologics at the same time. However, the administration of combined biologics has been linked to increased adverse effects, and it is currently under study to evaluate better dosage schemes [48,49].

Simulations with combined KOs mimicking the effect of anti-TNF treatment in combination with the downregulation of genes observed in the analysed datasets confirmed the dependency of the TFs activation state on the presence of inputs and further highlighted specific conditions. For example, when TNF is blocked, and DAXX, ILK and MAP2K1 are downregulated, the IL6 signal alone is not enough to activate the TFs JUN, FOS, JUND, and the kinases MAPK14 and MAPK1 (Table 6, ss2). However, when both IL6 and TGFB1

signals are on, all TFs are activated, and the activity level of kinase MAPK14 is restored (Table 6, ss3). MAPK14 (p38a kinase) and MAPK1 (ERK2) are two proteins known to play a pivotal role in RA and are activated by a variety of signals, including cytokines such as TNF and IL6 but also TGF-beta [50]. While p38 had been proposed as a potential target to reduce the destruction of bone and cartilage, p38 inhibitors have given disappointing results regarding therapeutic efficacy [51,52].

As seen in Table 5, the suppression of MAPK14 (p38a) does not inhibit the activation of the identified as master regulators TFs, even in the presence of anti-TNF treatment, as other inputs, such as IL6 or TGF-beta, can counterbalance the effects. Regarding ERK inhibitors, limited data are available, which may be due to a lack of efficient pharmacological inhibitors [53]. In older studies, FR180204, an ERK inhibitor, had demonstrated effectiveness against mouse collagen-induced arthritis [54], but there was no significant follow-up. In our model, ERK2 inhibition (MAPK1) does not seem to significantly impact the activation of its downstream target TFs, JUN, and FOS, as other regulators can also activate them.

5. Conclusions

While resistance to TNF therapy is a common event in the treatment of RA, the reasons behind its mechanisms are still unclear [55]. In addition, currently, there is no way of predicting which patient will respond or not to targeted therapy [56].

While the heterogeneity in RA is evident as manifested by the different patient profiles, the affected molecular pathways involved in the disease and autoimmunity are well known and studied. Therefore, a way to address the heterogeneity is to create backbone models comprising all affected pathways derived from the literature and big data to obtain global blueprints of the perturbed cascades. Then, patient-specific data can be used to contextualise each model by highlighting affected biomolecules (genes, proteins, metabolites, etc., depending on the type of data). In this way, patient-specific models could be created based on integrated, personalised data, such as clinical information, comorbidities, and genetic factors (mutations in specific genes). In addition, single-cell datasets could also provide insights into the disease heterogeneity at the cellular level.

Executable, integrated networks can accelerate the building of personalised models, as mapping dysregulated genes could reveal potentially impacted pathways shedding light on therapy response. Dynamic analysis and *in silico* simulations can also inform about the outcome of combined perturbations, predicting the emergent behaviour of the system. Integrating multi-omics data is a key step in understanding pathogenetic mechanisms of multifactorial diseases, where one level of information does not suffice to explain the complex phenotypic traits. Integrative networks allow for patient-level analysis by using patient-specific data and analysing the effects of patient-specific mutations and DEGs, combined with treatment effects. Such approaches could inform on the possibilities of success of a given therapy. For example, one could test the effects of mono or combined therapy, such as methotrexate (MTX) and anti-TNF, to better evaluate possible responses. Larger patient cohorts and more efficient computational techniques that would allow simulations on a larger scale could enhance the robustness and the predictive power of such models, helping to understand the response or non-response to a given therapy at a patient level.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/jpm11080785/s1, Figure S1: Principal component analysis (PCA) in samples of human blood cells from RA patients and healthy controls. The PCA shows 95 samples from the GSE117769 dataset (46 RA samples and 49 controls). In addition, a variance stabilising transformation was carried on the matrix expression data. Figure S2: Global RA network and DEG from responders/non-responders to anti-TNF treatment (37 and 41 RA patients treated with adalimumab and etanercept, respectively). Overlapping DEG from adalimumab treatment and etanercept treatment data are shown in brown (11) and pink (4), respectively, while non-overlapping genes/proteins are shown in grey (599). Transcription factors are depicted in diamond shapes, while upstream regulators and target genes are depicted using round shapes. Figure S3: Global RA network and DEG from untreated and treated RA patients with anti-TNF treatment (two different cohorts of 40 and 36 RA patients from baseline and after three months treatment with infliximab or adalimumab). Overlapping DEG are shown in green (101), while non-overlapping genes/proteins are shown in grey (513). Transcription factors are depicted in diamond shapes, while upstream regulators and target genes are depicted using round shapes. Figure S4: Environment sensitivity analysis. Transcription factors (TFs) could be upregulated in the absence of TNF activity for a combination of activity ranges of the other two inputs (IL6 and TGFB1). The upper part of each subfigure shows the impact of the external components on the activity state of the selected TFs, and the lower part of the image shows the range of activity percentage of the external components to achieve the optimisation of the activity state of the selected TFs. Sensitivity analysis for (**a**) JUND; (**b**) JUN; (**c**) FOS; (**d**) NFKB1; (**e**) NFKBIA; boxplots of IL6 in blue, TGFB1 in orange, and TNF as black line as it is set to off. Table S1: Transcription factors identified from CoRegNet and their involvement in RA based on literature evidence.

Author Contributions: Conceptualisation, A.N.; methodology, A.N. and M.E.; software, Q.M., D.d.M., V.S., M.E. and A.N.; validation, A.N., Q.M. and V.S.; formal analysis, Q.M., D.d.M., M.E. and A.N.; investigation, Q.M., D.d.M., M.E. and A.N.; resources, V.C., E.P.-T. and A.N.; data curation, Q.M., D.d.M. and V.S.; writing—original draft preparation, Q.M., D.d.M., V.S. and A.N.; writing—review and editing, Q.M., D.d.M., V.S., V.C., M.E., E.P.-T. and A.N.; visualisation, Q.M. and V.S.; supervision, M.E. and A.N.; project administration, A.N.; funding acquisition, V.C., E.P.-T. and A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the doctorate program of the University of Paris Saclay, France.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets (GSE117769, GSE129705, GSE138747, and DisGeNET variants) used for the analysis are publicly available. All data and code used to generate results, including networks inference, differential expression analysis, network visualisation, and Boolean model simulation, are available on a GitLab repository at https://gitlab.com/genhotel/inference-of-a-global-integrative-network-for-rheumatoid-arthritis (accessed on 1 July 2021). The Shiny app is freely available at https://quentin-miagoux.shinyapps.io/global_ra_network (accessed on 1 July 2021).

Acknowledgments: We would like to thank Smahane Chalabi, GenHotel, UEVE for her comments on the statistical treatment of the dataset used to infer the CoRegNet object. We would also like to acknowledge Fondagen, Genopole, for providing a training scholarship to QM.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ngo, S.T.; Steyn, F.J.; McCombe, P.A. Gender Differences in Autoimmune Disease. *Front. Neuroendocrinol.* **2014**, *35*, 347–369. [CrossRef]
2. Smolen, J.S.; Aletaha, D.; Barton, A.; Burmester, G.R.; Emery, P.; Firestein, G.S.; Kavanaugh, A.; McInnes, I.B.; Solomon, D.H.; Strand, V.; et al. Rheumatoid Arthritis. *Nat. Rev. Dis. Primers* **2018**, *4*, 18001. [CrossRef] [PubMed]
3. Deane, K.D.; Demoruelle, M.K.; Kelmenson, L.B.; Kuhn, K.A.; Norris, J.M.; Holers, V.M. Genetic and Environmental Risk Factors for Rheumatoid Arthritis. *Best Pract. Res. Clin. Rheumatol.* **2017**, *31*, 3–18. [CrossRef] [PubMed]
4. Batliwalla, F.M.; Baechler, E.C.; Xiao, X.; Li, W.; Balasubramanian, S.; Khalili, H.; Damle, A.; Ortmann, W.A.; Perrone, A.; Kantor, A.B.; et al. Peripheral Blood Gene Expression Profiling in Rheumatoid Arthritis. *Genes Immun.* **2005**, *6*, 388–397. [CrossRef] [PubMed]

5. Edwards, C.J.; Feldman, J.L.; Beech, J.; Shields, K.M.; Stover, J.A.; Trepicchio, W.L.; Larsen, G.; Foxwell, B.M.; Brennan, F.M.; Feldmann, M.; et al. Molecular Profile of Peripheral Blood Mononuclear Cells from Patients with Rheumatoid Arthritis. *Mol. Med.* **2007**, *13*, 40–58. [CrossRef]
6. Micsik, T.; Lőrincz, A.; Gál, J.; Schwab, R.; Peták, I. MDR-1 and MRP-1 Activity in Peripheral Blood Leukocytes of Rheumatoid Arthritis Patients. *Diagn. Pathol.* **2015**, *10*, 216. [CrossRef]
7. Kuuliala, K.; Kuuliala, A.; Koivuniemi, R.; Kautiainen, H.; Repo, H.; Leirisalo-Repo, M. Baseline, J.A.K. Phosphorylation Profile of Peripheral Blood Leukocytes, Studied by Whole Blood Phosphospecific Flow Cytometry, Is Associated with 1-Year Treatment Response in Early Rheumatoid Arthritis. *Arthritis Res. Ther.* **2017**, *19*. [CrossRef]
8. Li, X.; Lei, Y.; Gao, Z.; Zhang, B.; Xia, L.; Lu, J.; Shen, H. Effect of IL-34 on T Helper 17 Cell Proliferation and IL-17 Secretion by Peripheral Blood Mononuclear Cells from Rheumatoid Arthritis Patients. *Sci. Rep.* **2020**, *10*. [CrossRef]
9. Farutin, V.; Prod'homme, T.; McConnell, K.; Washburn, N.; Halvey, P.; Etzel, C.J.; Guess, J.; Duffner, J.; Getchell, K.; Meccariello, R.; et al. Molecular Profiling of Rheumatoid Arthritis Patients Reveals an Association between Innate and Adaptive Cell Populations and Response to Anti-Tumor Necrosis Factor. *Arthritis Res. Ther.* **2019**, *21*, 216. [CrossRef]
10. Eguchi, R.; Karim, M.B.; Hu, P.; Sato, T.; Ono, N.; Kanaya, S.; Altaf-Ul-Amin, M. An Integrative Network-Based Approach to Identify Novel Disease Genes and Pathways: A Case Study in the Context of Inflammatory Bowel Disease. *BMC Bioinform.* **2018**, *19*, 264. [CrossRef]
11. Karimizadeh, E.; Sharifi-Zarchi, A.; Nikaein, H.; Salehi, S.; Salamatian, B.; Elmi, N.; Gharibdoost, F.; Mahmoudi, M. Analysis of Gene Expression Profiles and Protein-Protein Interaction Networks in Multiple Tissues of Systemic Sclerosis. *BMC Med. Genom.* **2019**, *12*, 199. [CrossRef] [PubMed]
12. Haghjoo, N.; Moeini, A.; Masoudi-Nejad, A. Introducing a Panel for Early Detection of Lung Adenocarcinoma by Using Data Integration of Genomics, Epigenomics, Transcriptomics and Proteomics. *Exp. Mol. Pathol.* **2020**, *112*, 104360. [CrossRef] [PubMed]
13. Sahu, A.; Chowdhury, H.A.; Gaikwad, M.; Chongtham, C.; Talukdar, U.; Phukan, J.K.; Bhattacharyya, D.K.; Barah, P. Integrative Network Analysis Identifies Differential Regulation of Neuroimmune System in Schizophrenia and Bipolar Disorder. *Brain Behav. Immun. Health* **2020**, *2*, 100023. [CrossRef]
14. Greenfield, A.; Hafemeister, C.; Bonneau, R. Robust Data-Driven Incorporation of Prior Knowledge into the Inference of Dynamic Regulatory Networks. *Bioinformatics* **2013**, *29*, 1060–1067. [CrossRef] [PubMed]
15. Zuo, Y.; Cui, Y.; Yu, G.; Li, R.; Ressom, H.W. Incorporating Prior Biological Knowledge for Network-Based Differential Gene Expression Analysis Using Differentially Weighted Graphical LASSO. *BMC Bioinform.* **2017**, *18*, 99. [CrossRef]
16. Benedetti, E.; Pučić-Baković, M.; Keser, T.; Gerstner, N.; Büyüközkan, M.; Štambuk, T.; Selman, M.H.J.; Rudan, I.; Polašek, O.; Hayward, C.; et al. A Strategy to Incorporate Prior Knowledge into Correlation Network Cutoff Selection. *Nat. Commun.* **2020**, *11*, 5153. [CrossRef]
17. Nicolle, R.; Radvanyi, F.; Elati, M. CoRegNet: Reconstruction and Integrated Analysis of Co-Regulatory Networks. *Bioinformatics* **2015**, *31*, 3066–3068. [CrossRef]
18. Singh, V.; Kalliolias, G.D.; Ostaszewski, M.; Veyssiere, M.; Pilalis, E.; Gawron, P.; Mazein, A.; Bonnet, E.; Petit-Teixeira, E.; Niarakis, A. RA-Map: Building a State-of-the-Art Interactive Knowledge Base for Rheumatoid Arthritis. *Database* **2020**, *2020*. [CrossRef]
19. Mazein, A.; Ostaszewski, M.; Kuperstein, I.; Watterson, S.; Le Novère, N.; Lefaudeux, D.; De Meulder, B.; Pellet, J.; Balaur, I.; Saqi, M.; et al. Systems Medicine Disease Maps: Community-Driven Comprehensive Representation of Disease Mechanisms. *NPJ Syst. Biol. Appl.* **2018**, *4*, 21. [CrossRef] [PubMed]
20. Ostaszewski, M.; Gebel, S.; Kuperstein, I.; Mazein, A.; Zinovyev, A.; Dogrusoz, U.; Hasenauer, J.; Fleming, R.M.T.; Le Novère, N.; Gawron, P.; et al. Community-Driven Roadmap for Integrated Disease Maps. *Brief. Bioinform.* **2019**, *20*, 659–670. [CrossRef]
21. Niarakis, A.; Helikar, T. A Practical Guide to Mechanistic Systems Modeling in Biology Using a Logic-Based Approach. *Brief. Bioinform.* **2020**. [CrossRef] [PubMed]
22. Schwab, J.D.; Kühlwein, S.D.; Ikonomi, N.; Kühl, M.; Kestler, H.A. Concepts in Boolean Network Modeling: What Do They All Mean? *Comput. Struct. Biotechnol. J.* **2020**, *18*, 571–582. [CrossRef]
23. Abou-Jaoudé, W.; Traynard, P.; Monteiro, P.T.; Saez-Rodriguez, J.; Helikar, T.; Thieffry, D.; Chaouiya, C. Logical Modeling and Dynamical Analysis of Cellular Networks. *Front. Genet.* **2016**, *7*. [CrossRef] [PubMed]
24. Love, M.I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* **2014**, *15*. [CrossRef] [PubMed]
25. Chebil, I.; Nicolle, R.; Santini, G.; Rouveirol, C.; Elati, M. Hybrid Method Inference for the Construction of Cooperative Regulatory Network in Human. *IEEE Trans. Nanobiosci.* **2014**, *13*, 97–103. [CrossRef]
26. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Society. Ser. B* **1995**, *57*, 289–300. [CrossRef]
27. Hoksza, D.; Gawron, P.; Ostaszewski, M.; Smula, E.; Schneider, R. MINERVA API and Plugins: Opening Molecular Network Analysis and Visualization to the Community. *Bioinformatics* **2019**, *35*, 4496–4498. [CrossRef]
28. Le Novère, N.; Hucka, M.; Mi, H.; Moodie, S.; Schreiber, F.; Sorokin, A.; Demir, E.; Wegner, K.; Aladjem, M.I.; Wimalaratne, S.M.; et al. The Systems Biology Graphical Notation. *Nat. Biotechnol.* **2009**, *27*, 735–741. [CrossRef]
29. Aghamiri, S.S.; Singh, V.; Naldi, A.; Helikar, T.; Soliman, S.; Niarakis, A. Automated Inference of Boolean Models from Molecular Interaction Maps Using CaSQ. *Bioinformatics* **2020**, *36*, 4473–4482. [CrossRef]
30. Csardi, G.; Nepusz, T. The Igraph Software Package for Complex Network Research. *InterJournal Complex Syst.* **2006**, *1695*, 1–9.

31. Gustavsen, A.J.; Pai, S.; Isserlin, R.; Demchak, B.; Pico, A.R. RCy3: Network Biology Using Cytoscape from within R. *F1000Research* **2019**. [CrossRef]
32. Piñero, J.; Ramírez-Anguita, J.M.; Saüch-Pitarch, J.; Ronzano, F.; Centeno, E.; Sanz, F.; Furlong, L.I. The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* **2020**, *48*, D845–D855. [CrossRef]
33. Chang, W.; Cheng, J.; Allaire, J.J.; Xie, Y.; McPherson, J. *Shiny: Web Application Framework for R*; RStudio Inc.: Boston, MA, USA, 2020.
34. Shah, O.; Shannon, P. *CyjShiny: CyjShiny*. Available online: https://github.com/paul-shannon/cyjShiny (accessed on 1 July 2021).
35. Chaouiya, C.; Naldi, A.; Thieffry, D. Logical Modelling of Gene Regulatory Networks with GINsim. *Methods Mol. Biol.* **2012**, *804*, 463–479. [CrossRef]
36. Ogata, A.; Kato, Y.; Higa, S.; Yoshizaki, K. IL-6 Inhibitor for the Treatment of Rheumatoid Arthritis: A Comprehensive Review. *Mod. Rheumatol.* **2019**, *29*, 258–267. [CrossRef]
37. Gonzalo-Gil, E.; Criado, G.; Santiago, B.; Dotor, J.; Pablos, J.L.; Galindo, M. Transforming Growth Factor (TGF)-β Signalling is Increased in Rheumatoid Synovium but TGF-β Blockade does not Modify Experimental Arthritis. *Clin. Exp. Immunol.* **2013**, *174*, 245–255. [CrossRef]
38. Hall, B.; Niarakis, A. Data Integration in Logic-Based Models of Biological Mechanisms. *Preprints* **2021**. [CrossRef]
39. Funahashi, A.; Matsuoka, Y.; Jouraku, A.; Morohashi, M.; Kikuchi, N.; Kitano, H. CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proc. IEEE* **2008**, *96*, 1254–1265. [CrossRef]
40. Helikar, T.; Kowal, B.; McClenathan, S.; Bruckner, M.; Rowley, T.; Madrahimov, A.; Wicks, B.; Shrestha, M.; Limbu, K.; Rogers, J.A. The Cell Collective: Toward an Open and Collaborative Approach to Systems Biology. *BMC Syst. Biol.* **2012**, *6*, 96. [CrossRef]
41. Hennigan, S.; Kavanaugh, A. Interleukin-6 Inhibitors in the Treatment of Rheumatoid Arthritis. *Ther. Clin. Risk Manag.* **2008**, *4*, 767–775.
42. Woodrick, R.; Ruderman, E.M. Anti-Interleukin-6 Therapy in Rheumatoid Arthritis. *Bull. NYU Hosp. Jt. Dis.* **2010**, *68*, 211–217.
43. Choy, E.H.; De Benedetti, F.; Takeuchi, T.; Hashizume, M.; John, M.R.; Kishimoto, T. Translating IL-6 Biology into Effective Treatments. *Nat. Rev. Rheumatol.* **2020**, *16*, 335–345. [CrossRef]
44. Sakuma, M.; Hatsushika, K.; Koyama, K.; Katoh, R.; Ando, T.; Watanabe, Y.; Wako, M.; Kanzaki, M.; Takano, S.; Sugiyama, H.; et al. TGF-β Type I Receptor Kinase Inhibitor down-Regulates Rheumatoid Synoviocytes and Prevents the Arthritis Induced by Type II Collagen Antibody. *Int. Immunol.* **2007**, *19*, 117–126. [CrossRef] [PubMed]
45. Guo, L. TGF Beta in the Rheumatoid Arthritis Research. *Eur. J. Biomed. Res.* **2017**, *3*, 5–8. [CrossRef]
46. Zhou, G.; Sun, X.; Qin, Q.; Lv, J.; Cai, Y.; Wang, M.; Mu, R.; Lan, H.; Wang, Q.-W. Loss of Smad7 Promotes Inflammation in Rheumatoid Arthritis. *Front. Immunol.* **2018**, *9*. [CrossRef] [PubMed]
47. A Dual Target-Directed Agent against Interleukin-6 Receptor and Tumor Necrosis Factor α Ameliorates Experimental Arthritis | Scientific Reports. Available online: https://www.nature.com/articles/srep20150 (accessed on 23 June 2021).
48. Boleto, G.; Kanagaratnam, L.; Dramé, M.; Salmon, J.-H. Safety of Combination Therapy with Two BDMARDs in Patients with Rheumatoid Arthritis: A Systematic Review and Meta-Analysis. *Semin. Arthritis Rheum.* **2019**, *49*, 35–42. [CrossRef]
49. Biologic Combination Therapy for RA May Increase Risk for Side Effects. Available online: https://rheumatology.medicinematters.com/rheumatoid-arthritis-/biologics/biologic-combination-therapy-for-ra-may-increase-risk-for-side-e/16387312 (accessed on 23 June 2021).
50. Canovas, B.; Nebreda, A.R. Diversity and Versatility of P38 Kinase Signalling in Health and Disease. *Nat. Rev. Mol. Cell Biol.* **2021**, *22*, 346–366. [CrossRef]
51. Clark, A.R.; Dean, J.L. The P38 MAPK Pathway in Rheumatoid Arthritis: A Sideways Look. *Open Rheumatol. J.* **2012**, *6*. [CrossRef]
52. Haller, V.; Nahidino, P.; Forster, M.; Laufer, S.A. An Updated Patent Review of P38 MAP Kinase Inhibitors (2014–2019). *Expert Opin. Ther. Pat.* **2020**, *30*, 453–466. [CrossRef]
53. Bonilla-Hernán, M.G.; Miranda-Carús, M.E.; Martin-Mola, E. New Drugs beyond Biologics in Rheumatoid Arthritis: The Kinase Inhibitors. *Rheumatology* **2011**, *50*, 1542–1550. [CrossRef]
54. Ohori, M. ERK Inhibitors as a Potential New Therapy for Rheumatoid Arthritis. *Drug News Perspect.* **2008**, *21*, 245–250. [CrossRef]
55. Sidiropoulos, P.I.; Boumpas, D.T. Differential Drug Resistance to Anti-tumour Necrosis Factor Agents in Rheumatoid Arthritis. *Ann. Rheum Dis.* **2006**, *65*, 701–703. [CrossRef]
56. Rheumatoid Arthritis: A Case for Personalized Health Care. Available online: https://onlinelibrary.wiley.com/doi/full/10.1002/acr.22289 (accessed on 1 July 2021).

Bibliographie

1. Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS, et al. Rheumatoid arthritis. Nat Rev Dis Primers. 2018;4:18001.

2. Landré-Beauvais AJ. The first description of rheumatoid arthritis. Unabridged text of the doctoral dissertation presented in 1800. Joint Bone Spine. 2001;68:130–43.

3. Garrod AB. The nature and treatment of gout and rheumatic gout. London : Walton; Maberly; 1859.

4. Entezami P, Fox DA, Clapham PJ, Chung KC. Historical perspective on the etiology of rheumatoid arthritis. Hand Clin. 2011;27:1–10.

5. Rothschild BM, Turner KR, DeLuca MA. Symmetrical erosive peripheral polyarthritis in the late archaic period of alabama. Science. 1988;241:1498–501.

6. Appelboom T. Rubens and the question of antiquity of rheumatoid arthritis. JAMA. 1981;245:483.

7. Dequeker J, Rico H. Rheumatoid arthritis-like deformities in an early 16th-century painting of the flemish-dutch school. JAMA. 1992;268:249–51.

8. Vallin J, Meslé F. Tables de mortalité françaises pour les XIXe et XXe siècles et projections pour le XXIe siècle. Paris: Institut National d'Études Démographiques; 2001.

9. Almutairi K, Nossent J, Preen D, Keen H, Inderjeeth C. The global prevalence of rheumatoid arthritis: A meta-analysis based on a systematic review. Rheumatol Int. 2020. `https://doi.org/10.1007/s00296-020-04731-0`.

10. Safiri S, Kolahi AA, Hoy D, Smith E, Bettampadi D, Mansournia MA, et al. Global, regional and national burden of rheumatoid arthritis 1990–2017: A

systematic analysis of the global burden of disease study 2017. Annals of the Rheumatic Diseases. 2019;78:1463–71.

11. Myasoedova E, Crowson CS, Kremers HM, Therneau TM, Gabriel SE. Is the incidence of rheumatoid arthritis rising?: Results from olmsted county, minnesota, 1955-2007. Arthritis Rheum. 2010;62:1576–82.

12. Hunter TM, Boytsov NN, Zhang X, Schroeder K, Michaud K, Araujo AB. Prevalence of rheumatoid arthritis in the united states adult population in healthcare claims databases, 2004-2014. Rheumatol Int. 2017;37:1551–7.

13. Guillemin F, Saraux A, Guggenbuhl P, Roux C, Fardellone P, Le Bihan E, et al. Prevalence of rheumatoid arthritis in france: 2001. Ann Rheum Dis. 2005;64:1427–30.

14. Saraux A, Guedes C, Allain J, Devauchelle V, Valls I, Lamour A, et al. Prevalence of rheumatoid arthritis and spondyloarthropathy in brittany, france. Société de rhumatologie de l'Ouest. J Rheumatol. 1999;26:2622–7.

15. Shapira Y, Agmon-Levin N, Shoenfeld Y. Geoepidemiology of autoimmune rheumatic diseases. Nat Rev Rheumatol. 2010;6:468–76.

16. Tobón GJ, Youinou P, Saraux A. The environment, geo-epidemiology, and autoimmune disease: Rheumatoid arthritis. Autoimmunity Reviews. 2010;9:A288–92.

17. Ngo ST, Steyn FJ, McCombe PA. Gender differences in autoimmune disease. Frontiers in Neuroendocrinology. 2014;35:347–69.

18. Pikwer M, Bergström U, Nilsson J-Å, Jacobsson L, Turesson C. Early menopause is an independent predictor of rheumatoid arthritis. Ann Rheum Dis. 2012;71:378–81.

19. Bengtsson C, Malspeis S, Orellana C, Sparks JA, Costenbader KH, Karlson EW. Menopausal factors are associated with seronegative RA in large prospective cohorts: Results from the nurses' health studies. Arthritis Care Res (Hoboken). 2017;69:1676–84.

20. Wallenius M, Skomsvoll JF, Irgens LM, Salvesen KA, Koldingsnes W, Mikkelsen K, et al. Postpartum onset of rheumatoid arthritis and other chronic arthritides: Results from a patient register linked to a medical birth registry. Ann Rheum

Dis. 2010;69:332–6.

21. Chen JY, Ballou SP. The effect of antiestrogen agents on risk of autoimmune disorders in patients with breast cancer. J Rheumatol. 2015;42:55–9.

22. Alpízar-Rodríguez D, Pluchino N, Canny G, Gabay C, Finckh A. The role of female hormonal factors in the development of rheumatoid arthritis. Rheumatology. 2017;56:1254–63.

23. López-Mejías R, Castañeda S, González-Juanatey C, Corrales A, Ferraz-Amaro I, Genre F, et al. Cardiovascular risk assessment in patients with rheumatoid arthritis: The relevance of clinical, genetic and serological markers. Autoimmun Rev. 2016;15:1013–30.

24. Hoek J van den, Boshuizen HC, Roorda LD, Tijhuis GJ, Nurmohamed MT, Bos GAM van den, et al. Mortality in patients with rheumatoid arthritis: A 15-year prospective cohort study. Rheumatol Int. 2017;37:487–93.

25. SPARKS JA, CHANG S-C, LIAO KP, LU B, FINE AR, SOLOMON DH, et al. Rheumatoid arthritis and mortality among women during 36 years of prospective follow-up: Results from the nurses' health study. Arthritis Care Res (Hoboken). 2016;68:753–62.

26. Markusse IM, Akdemir G, Dirven L, Goekoop-Ruiterman YPM, Groenendael JHLM van, Han KH, et al. Long-term outcomes of patients with recent-onset rheumatoid arthritis after 10 years of tight controlled treatment: A randomized trial. Ann Intern Med. 2016;164:523–31.

27. Aletaha D, Breedveld FC, Smolen JS. The need for new classification criteria for rheumatoid arthritis. Arthritis Rheum. 2005;52:3333–6.

28. Kay J, Upchurch KS. ACR/EULAR 2010 rheumatoid arthritis classification criteria. Rheumatology. 2012;51 suppl_6:vi5–9.

29. Radner H, Neogi T, Smolen JS, Aletaha D. Performance of the 2010 ACR/EULAR classification criteria for rheumatoid arthritis: A systematic literature review. Ann Rheum Dis. 2014;73:114–23.

30. Nishimura K, Sugiyama D, Kogata Y, Tsuji G, Nakazawa T, Kawano S, et al. Meta-analysis: Diagnostic accuracy of anti–cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. Ann Intern Med.

2007;146:797–808.

31. Nielen MMJ, Schaardenburg D van, Reesink HW, Stadt RJ van de, Horst-Bruinsma IE van der, Koning MHMT de, et al. Specific autoantibodies precede the symptoms of rheumatoid arthritis: A study of serial measurements in blood donors. Arthritis Rheum. 2004;50:380–6.

32. Gossec L, Combescure C, Rincheval N, Saraux A, Combe B, Dougados M. Relative clinical influence of clinical, laboratory, and radiological investigations in early arthritis on the diagnosis of rheumatoid arthritis. Data from the french early arthritis cohort ESPOIR. J Rheumatol. 2010;37:2486–92.

33. Heidari B. Rheumatoid arthritis: Early diagnosis and treatment outcomes. Caspian J Intern Med. 2011;2:161–70.

34. Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. Lancet. 2016;388:2023–38.

35. Ingegnoli F, Castelli R, Gualtierotti R. Rheumatoid factors: Clinical applications. Dis Markers. 2013;35:727–34.

36. Shmerling RH, Delbanco TL. The rheumatoid factor: An analysis of clinical utility. Am J Med. 1991;91:528–34.

37. Newkirk MM. Rheumatoid factors: Host resistance or autoimmunity? Clin Immunol. 2002;104:1–13.

38. Jimenez-Boj E, Nöbauer-Huhmann I, Hanslik-Schnabel B, Dorotka R, Wanivenhaus A-H, Kainberger F, et al. Bone erosions and bone marrow edema as defined by magnetic resonance imaging reflect true bone marrow inflammation in rheumatoid arthritis. Arthritis Rheum. 2007;56:1118–24.

39. Mandl P, Balint PV, Brault Y, Backhaus M, D'Agostino M-A, Grassi W, et al. Metrologic properties of ultrasound versus clinical evaluation of synovitis in rheumatoid arthritis: Results of a multicenter, randomized study. Arthritis Rheum. 2012;64:1272–82.

40. Klareskog L, Catrina AI. Autoimmunity: Lungs and citrullination. Nat Rev Rheumatol. 2015;11:261–2.

41. Reynisdottir G, Olsen H, Joshua V, Engström M, Forsslund H, Karimi R, et al. Signs of immune activation and local inflammation are present in the bronchial

tissue of patients with untreated early rheumatoid arthritis. Ann Rheum Dis. 2016;75:1722–7.

42. Muller S, Radic M. Citrullinated autoantigens: From diagnostic markers to pathogenetic mechanisms. Clin Rev Allergy Immunol. 2015;49:232–9.

43. Holers VM. Autoimmunity to citrullinated proteins and the initiation of rheumatoid arthritis. Curr Opin Immunol. 2013;25:728–35.

44. Beers JJBC van, Willemze A, Jansen JJ, Engbers GHM, Salden M, Raats J, et al. ACPA fine-specificity profiles in early rheumatoid arthritis patients do not correlate with clinical features at baseline or with disease progression. Arthritis Res Ther. 2013;15:R140.

45. Szekanecz Z, Pakozdi A, Szentpetery A, Besenyei T, Koch AE. Chemokines and angiogenesis in rheumatoid arthritis. Front Biosci (Elite Ed). 2009;1:44–51.

46. McInnes IB, Schett G. The pathogenesis of rheumatoid arthritis. N Engl J Med. 2011;365:2205–19.

47. Lefèvre S, Knedla A, Tennie C, Kampmann A, Wunrau C, Dinser R, et al. Synovial fibroblasts spread rheumatoid arthritis to unaffected joints. Nat Med. 2009;15:1414–20.

48. Keyszer G, Redlich A, Häupl T, Zacher J, Sparmann M, Engethüm U, et al. Differential expression of cathepsins b and l compared with matrix metalloproteinases and their respective inhibitors in rheumatoid arthritis and osteoarthritis: A parallel investigation by semiquantitative reverse transcriptase-polymerase chain reaction and immunohistochemistry. Arthritis Rheum. 1998;41:1378–87.

49. Kiener HP, Niederreiter B, Lee DM, Jimenez-Boj E, Smolen JS, Brenner MB. Cadherin 11 promotes invasive behavior of fibroblast-like synoviocytes. Arthritis Rheum. 2009;60:1305–10.

50. Gravallese EM, Harada Y, Wang JT, Gorn AH, Thornhill TS, Goldring SR. Identification of cell types responsible for bone resorption in rheumatoid arthritis and juvenile rheumatoid arthritis. Am J Pathol. 1998;152:943–51.

51. Klareskog L, Gregersen PK, Huizinga TWJ. Prevention of autoimmune rheumatic disease: State of the art and future perspectives. Ann Rheum Dis. 2010;69:2062–6.

52. Deane KD, Demoruelle MK, Kelmenson LB, Kuhn KA, Norris JM, Holers VM. Genetic and environmental risk factors for rheumatoid arthritis. Best Practice & Research Clinical Rheumatology. 2017;31:3–18.

53. Vessey MP, Villard-Mackintosh L, Yeates D. Oral contraceptives, cigarette smoking and other factors in relation to arthritis. Contraception. 1987;35:457–64.

54. Heliövaara M, Aho K, Aromaa A, Knekt P, Reunanen A. Smoking and risk of rheumatoid arthritis. J Rheumatol. 1993;20:1830–5.

55. Uhlig T, Hagen KB, Kvien TK. Current tobacco smoking, formal education, and the risk of rheumatoid arthritis. J Rheumatol. 1999;26:47–54.

56. Karlson EW, Lee IM, Cook NR, Manson JE, Buring JE, Hennekens CH. A retrospective cohort study of cigarette smoking and risk of rheumatoid arthritis in female health professionals. Arthritis Rheum. 1999;42:910–7.

57. Criswell LA, Merlino LA, Cerhan JR, Mikuls TR, Mudano AS, Burma M, et al. Cigarette smoking and the risk of rheumatoid arthritis among postmenopausal women: Results from the iowa women's health study. Am J Med. 2002;112:465–71.

58. Padyukov L, Silva C, Stolt P, Alfredsson L, Klareskog L. A gene-environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. Arthritis Rheum. 2004;50:3085–92.

59. Costenbader KH, Feskanich D, Mandl LA, Karlson EW. Smoking intensity, duration, and cessation, and the risk of rheumatoid arthritis in women. Am J Med. 2006;119:503.e1–9.

60. Sugiyama D, Nishimura K, Tamaki K, Tsuji G, Nakazawa T, Morinobu A, et al. Impact of smoking as a risk factor for developing rheumatoid arthritis: A meta-analysis of observational studies. Ann Rheum Dis. 2010;69:70–81.

61. Di Giuseppe D, Discacciati A, Orsini N, Wolk A. Cigarette smoking and risk of rheumatoid arthritis: A dose-response meta-analysis. Arthritis Res Ther. 2014;16:R61.

62. Chang K, Yang SM, Kim SH, Han KH, Park SJ, Shin JI. Smoking and rheumatoid arthritis. Int J Mol Sci. 2014;15:22279–95.

63. Krishnan E, Sokka T, Hannonen P. Smoking-gender interaction and risk for rheumatoid arthritis. Arthritis Res Ther. 2003;5:R158–162.

64. Naranjo A, Toloza S, Guimaraes da Silveira I, Lazovskis J, Hetland ML, Hamoud H, et al. Smokers and non smokers with rheumatoid arthritis have similar clinical status: Data from the multinational QUEST-RA database. Clin Exp Rheumatol. 2010;28:820–7.

65. Vesperini V, Lukas C, Fautrel B, Le Loet X, Rincheval N, Combe B. Association of tobacco exposure and reduction of radiographic progression in early rheumatoid arthritis: Results from a french multicenter cohort. Arthritis Care Res (Hoboken). 2013;65:1899–906.

66. Turesson C, Bergström U, Pikwer M, Nilsson J-Å, Jacobsson LTH. A high body mass index is associated with reduced risk of rheumatoid arthritis in men, but not in women. Rheumatology (Oxford). 2016;55:307–14.

67. Camacho EM, Verstappen SMM, Symmons DPM. Association between socioeconomic status, learned helplessness, and disease outcome in patients with inflammatory polyarthritis. Arthritis Care Res (Hoboken). 2012;64:1225–32.

68. De Roos AJ, Koehoorn M, Tamburic L, Davies HW, Brauer M. Proximity to traffic, ambient air pollution, and community noise in relation to incident rheumatoid arthritis. Environ Health Perspect. 2014;122:1075–80.

69. Hajishengallis G. Periodontitis: From microbial immune subversion to systemic inflammation. Nat Rev Immunol. 2015;15:30–44.

70. Kharlamova N, Jiang X, Sherina N, Potempa B, Israelsson L, Quirke A-M, et al. Antibodies to porphyromonas gingivalis indicate interaction between oral infection, smoking, and risk genes in rheumatoid arthritis etiology. Arthritis Rheumatol. 2016;68:604–13.

71. Konig MF, Abusleme L, Reinholdt J, Palmer RJ, Teles RP, Sampson K, et al. Aggregatibacter actinomycetemcomitans-induced hypercitrullination links periodontal infection to autoimmunity in rheumatoid arthritis. Sci Transl Med. 2016;8:369ra176.

72. Wegner N, Wait R, Sroka A, Eick S, Nguyen K-A, Lundberg K, et al. Peptidylarginine deiminase from porphyromonas gingivalis citrullinates human fibrinogen and a-enolase: Implications for autoimmunity in rheumatoid arthritis. Arthritis

Rheum. 2010;62:2662–72.

73. Konig MF, Paracha AS, Moni M, Bingham CO, Andrade F. Defining the role of porphyromonas gingivalis peptidylarginine deiminase (PPAD) in rheumatoid arthritis through the study of PPAD biology. Ann Rheum Dis. 2015;74:2054–61.

74. Laugisch O, Wong A, Sroka A, Kantyka T, Koziel J, Neuhaus K, et al. Citrullination in the periodontium–a possible link between periodontitis and rheumatoid arthritis. Clin Oral Investig. 2016;20:675–83.

75. Chen J, Wright K, Davis JM, Jeraldo P, Marietta EV, Murray J, et al. An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. Genome Med. 2016;8:43.

76. Gasque P, Bandjee MCJ, Reyes MM, Viasus D. Chikungunya pathogenesis: From the clinics to the bench. J Infect Dis. 2016;214 suppl 5:S446–8.

77. Naciute M, Mieliauskaite D, Rugiene R, Nikitenkiene R, Jancoriene L, Mauricas M, et al. Frequency and significance of parvovirus b19 infection in patients with rheumatoid arthritis. J Gen Virol. 2016;97:3302–12.

78. Tan EM, Smolen JS. Historical observations contributing insights on etiopathogenesis of rheumatoid arthritis and role of rheumatoid factor. J Exp Med. 2016;213:1937–50.

79. Klockars M, Koskela RS, Järvinen E, Kolari PJ, Rossi A. Silica exposure and rheumatoid arthritis: A follow up study of granite workers 1940-81. Br Med J (Clin Res Ed). 1987;294:997–1000.

80. Sluis-Cremer GK, Hessel PA, Hnizdo E, Churchill AR. Relationship between silicosis and rheumatoid arthritis. Thorax. 1986;41:596–601.

81. Stolt P, Yahya A, Bengtsson C, Källberg H, Rönnelid J, Lundberg I, et al. Silica exposure among male current smokers is associated with a high risk of developing ACPA-positive rheumatoid arthritis. Ann Rheum Dis. 2010;69:1072–6.

82. Turner S, Cherry N. Rheumatoid arthritis in workers exposed to silica in the pottery industry. Occup Environ Med. 2000;57:443–7.

83. Gan RW, Deane KD, Zerbe GO, Demoruelle MK, Weisman MH, Buckner JH, et al. Relationship between air pollution and positivity of RA-related autoantibodies in individuals without established RA: A report on SERA. Annals

of the Rheumatic Diseases. 2013;72:2002–5.

84. Hart JE, Källberg H, Laden F, Costenbader KH, Yanosky JD, Klareskog L, et al. Ambient air pollution exposures and risk of rheumatoid arthritis. Arthritis Care Res (Hoboken). 2013;65:1190–6.

85. Hart JE, Källberg H, Laden F, Bellander T, Costenbader KH, Holmqvist M, et al. Ambient air pollution exposures and risk of rheumatoid arthritis: Results from the swedish EIRA case-control study. Ann Rheum Dis. 2013;72:888–94.

86. Essouma M, Noubiap JJN. Is air pollution a risk factor for rheumatoid arthritis? J Inflamm (Lond). 2015;12:48.

87. Chang K-H, Hsu C-C, Muo C-H, Hsu CY, Liu H-C, Kao C-H, et al. Air pollution exposure increases the risk of rheumatoid arthritis: A longitudinal and nationwide study. Environ Int. 2016;94:495–9.

88. Sun G, Hazlewood G, Bernatsky S, Kaplan GG, Eksteen B, Barnabe C. Association between air pollution and the development of rheumatic disease: A systematic review. Int J Rheumatol. 2016;2016:5356307.

89. Pattison DJ, Symmons DPM, Lunt M, Welch A, Luben R, Bingham SA, et al. Dietary risk factors for the development of inflammatory polyarthritis: Evidence for a role of high level of red meat consumption. Arthritis Rheum. 2004;50:3804–12.

90. Sundström B, Johansson I, Rantapää-Dahlqvist S. Interaction between dietary sodium and smoking increases the risk for rheumatoid arthritis: Results from a nested case-control study. Rheumatology (Oxford). 2015;54:487–93.

91. Hair MJH de, Landewé RBM, Sande MGH van de, Schaardenburg D van, Baarsen LGM van, Gerlag DM, et al. Smoking and overweight determine the likelihood of developing rheumatoid arthritis. Ann Rheum Dis. 2013;72:1654–8.

92. Rakieh C, Nam JL, Hunt L, Hensor EMA, Das S, Bissell L-A, et al. Predicting the development of clinical arthritis in anti-CCP positive individuals with non-specific musculoskeletal symptoms: A prospective observational cohort study. Ann Rheum Dis. 2015;74:1659–66.

93. Ljung L, Rantapää-Dahlqvist S. Abdominal obesity, gender and the risk of rheumatoid arthritis - a nested case-control study. Arthritis Res Ther.

2016;18:277.

94. Merlino LA, Curtis J, Mikuls TR, Cerhan JR, Criswell LA, Saag KG, et al. Vitamin d intake is inversely associated with rheumatoid arthritis: Results from the iowa women's health study. Arthritis Rheum. 2004;50:72–7.

95. Linos A, Kaklamanis E, Kontomerkos A, Koumantaki Y, Gazi S, Vaiopoulos G, et al. The effect of olive oil and fish consumption on rheumatoid arthritis–a case control study. Scand J Rheumatol. 1991;20:419–26.

96. Shapiro JA, Koepsell TD, Voigt LF, Dugowson CE, Kestin M, Nelson JL. Diet and rheumatoid arthritis in women: A possible protective effect of fish consumption. Epidemiology. 1996;7:256–63.

97. Linos A, Kaklamani VG, Kaklamani E, Koumantaki Y, Giziaki E, Papazoglou S, et al. Dietary factors in relation to rheumatoid arthritis: A role for olive oil and cooked vegetables? Am J Clin Nutr. 1999;70:1077–82.

98. Di Giuseppe D, Crippa A, Orsini N, Wolk A. Fish consumption and risk of rheumatoid arthritis: A dose-response meta-analysis. Arthritis Res Ther. 2014;16:446.

99. Gan RW, Young KA, Zerbe GO, Demoruelle MK, Weisman MH, Buckner JH, et al. Lower omega-3 fatty acids are associated with the presence of anti-cyclic citrullinated peptide autoantibodies in a population at risk for future rheumatoid arthritis: A nested case-control study. Rheumatology (Oxford). 2016;55:367–76.

100. He J, Wang Y, Feng M, Zhang X, Jin Y-B, Li X, et al. Dietary intake and risk of rheumatoid arthritis-a cross section multicenter study. Clin Rheumatol. 2016;35:2901–8.

101. Gan RW, Demoruelle MK, Deane KD, Weisman MH, Buckner JH, Gregersen PK, et al. Omega-3 fatty acids are associated with a lower prevalence of autoantibodies in shared epitope-positive subjects at risk for rheumatoid arthritis. Ann Rheum Dis. 2017;76:147–52.

102. Pedersen M, Jacobsen S, Klarlund M, Pedersen BV, Wiik A, Wohlfahrt J, et al. Environmental risk factors differ between rheumatoid arthritis with and without auto-antibodies against cyclic citrullinated peptides. Arthritis Res Ther. 2006;8:R133.

103. Källberg H, Jacobsen S, Bengtsson C, Pedersen M, Padyukov L, Garred P, et al. Alcohol consumption is associated with decreased risk of rheumatoid arthritis: Results from two scandinavian case-control studies. Ann Rheum Dis. 2009;68:222–7.

104. Jin Z, Xiang C, Cai Q, Wei X, He J. Alcohol consumption as a preventive factor for developing rheumatoid arthritis: A dose-response meta-analysis of prospective studies. Ann Rheum Dis. 2014;73:1962–7.

105. He J, Wang Y, Feng M, Zhang X, Jin Y-B, Li X, et al. Dietary intake and risk of rheumatoid arthritis-a cross section multicenter study. Clin Rheumatol. 2016;35:2901–8.

106. Hu Y, Sparks JA, Malspeis S, Costenbader KH, Hu FB, Karlson EW, et al. Long-term dietary quality and risk of developing rheumatoid arthritis in women. Ann Rheum Dis. 2017;76:1357–64.

107. Hu Y, Cui J, Sparks JA, Malspeis S, Costenbader KH, Karlson EW, et al. Circulating carotenoids and subsequent risk of rheumatoid arthritis in women. Clin Exp Rheumatol. 2017;35:309–12.

108. Chodick G, Amital H, Shalem Y, Kokia E, Heymann AD, Porath A, et al. Persistence with statins and onset of rheumatoid arthritis: A population-based cohort study. PLoS Med. 2010;7:e1000336.

109. Tascilar K, Dell'Aniello S, Hudson M, Suissa S. Statins and risk of rheumatoid arthritis: A nested case-control study. Arthritis Rheumatol. 2016;68:2603–11.

110. Doran MF, Crowson CS, O'Fallon WM, Gabriel SE. The effect of oral contraceptives and estrogen replacement therapy on the risk of rheumatoid arthritis: A population based study. J Rheumatol. 2004;31:207–13.

111. Spector TD, Roman E, Silman AJ. The pill, parity, and rheumatoid arthritis. Arthritis Rheum. 1990;33:782–9.

112. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum. 2000;43:30–7.

113. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009;10:241–51.

114. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409:928–33.

115. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39:1256–60.

116. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16:172–83.

117. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451–81.

118. Shastry BS. SNPs in disease gene mapping, medicinal drug development and evolution. J Hum Genet. 2007;52:871–80.

119. Cirillo E, Kutmon M, Gonzalez Hernandez M, Hooimeijer T, Adriaens ME, Eijssen LMT, et al. From SNPs to pathways: Biological interpretation of type 2 diabetes (t2dm) genome wide association study (GWAS) results. PLoS One. 2018;13.

120. Li Y, Lai-Han Leung E, Pan H, Yao X, Huang Q, Wu M, et al. Identification of potential genetic causal variants for rheumatoid arthritis by whole-exome sequencing. Oncotarget. 2017;8:111119–29.

121. Vries R de. Genetics of rheumatoid arthritis: Time for a change! Curr Opin Rheumatol. 2011;23:227–32.

122. Yarwood A, Huizinga TWJ, Worthington J. The genetics of rheumatoid arthritis: Risk and protection in different stages of the evolution of RA. Rheumatology (Oxford). 2016;55:199–209.

123. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum. 1987;30:1205–13.

124. Holoshitz J. The rheumatoid arthritis HLA-DRB1 shared epitope. Curr Opin Rheumatol. 2010;22:293–8.

125. Helm-van Mil AHM van der, Huizinga TWJ, Schreuder GMT, Breedveld FC, Vries RRP de, Toes REM. An independent role of protective HLA class

II alleles in rheumatoid arthritis severity and susceptibility. Arthritis Rheum. 2005;52:2637–44.

126. Viatte S, Plant D, Han B, Fu B, Yarwood A, Thomson W, et al. Association of HLA-DRB1 haplotypes with rheumatoid arthritis severity, mortality, and treatment response. JAMA. 2015;313:1645–56.

127. Viatte S, Barton A. Genetics of rheumatoid arthritis susceptibility, severity, and treatment response. Semin Immunopathol. 2017;39:395–408.

128. Trier N, Izarzugaza J, Chailyan A, Marcatili P, Houen G. Human MHC-II with shared epitope motifs are optimal epstein-barr virus glycoprotein 42 ligands—relation to rheumatoid arthritis. Int J Mol Sci. 2018;19:317.

129. Almeida DE de, Ling S, Holoshitz J. New insights into the functional role of the rheumatoid arthritis shared epitope. FEBS Lett. 2011;585:3619–26.

130. Mewar D, Marinou I, Coote AL, Moore DJ, Akil M, Smillie D, et al. Association between radiographic severity of rheumatoid arthritis and shared epitope alleles: Differing mechanisms of susceptibility and protection. Ann Rheum Dis. 2008;67:980–3.

131. Carrier N, Cossette P, Daniel C, Brum-Fernandes A de, Liang P, Ménard HA, et al. The DERAA HLA-DR alleles in patients with early polyarthritis: Protection against severe disease and lack of association with rheumatoid arthritis autoantibodies. Arthritis Rheum. 2009;60:698–707.

132. Suzuki A, Yamada R, Chang X, Tokuhiro S, Sawada T, Suzuki M, et al. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet. 2003;34:395–402.

133. Begovich AB, Carlton VEH, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am J Hum Genet. 2004;75:330–7.

134. Kurreeman FAS, Padyukov L, Marques RB, Schrodi SJ, Seddighzadeh M, Stoeken-Rijsbergen G, et al. A candidate gene approach identifies the TRAF1/c5 region as a risk factor for rheumatoid arthritis. PLoS Med. 2007;4:e278.

135. Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT, Karlson EW, et al. Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from north america and sweden: Association of susceptibility with PTPN22, CTLA4, and PADI4. Am J Hum Genet. 2005;77:1044–60.

136. Zamanpoor M. The genetic pathogenesis, diagnosis and therapeutic insight of rheumatoid arthritis. Clin Genet. 2019;95:547–57.

137. Kurkó J, Besenyei T, Laki J, Glant TT, Mikecz K, Szekanecz Z. Genetics of rheumatoid arthritis — a comprehensive review. Clin Rev Allergy Immunol. 2013;45:170–9.

138. Viatte S, Massey J, Bowes J, Duffus K, arcOGEN Consortium, Eyre S, et al. Replication of associations of genetic loci outside the HLA region with susceptibility to anti-cyclic citrullinated peptide-negative rheumatoid arthritis. Arthritis Rheumatol. 2016;68:1603–13.

139. Viatte S, Plant D, Bowes J, Lunt M, Eyre S, Barton A, et al. Genetic markers of rheumatoid arthritis susceptibility in anti-citrullinated peptide antibody negative patients. Ann Rheum Dis. 2012;71:1984–90.

140. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506:376–81.

141. Okada Y, Eyre S, Suzuki A, Kochi Y, Yamamoto K. Genetics of rheumatoid arthritis: 2018 status. Ann Rheum Dis. 2019;78:446–53.

142. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.

143. Motegi T, Kochi Y, Matsuda K, Kubo M, Yamamoto K, Momozawa Y. Identification of rare coding variants in TYK2 protective for rheumatoid arthritis in the japanese population and their effects on cytokine signalling. Ann Rheum Dis. 2019;78:1062–9.

144. Veyssiere M, Perea J, Michou L, Boland A, Caloustian C, Olaso R, et al. A novel nonsense variant in SUPT20H gene associated with rheumatoid arthritis identified by whole exome sequencing of multiplex families. PLoS One. 2019;14:e0213387.

145. Wang Y, Chen S, Chen J, Xie X, Gao S, Zhang C, et al. Germline genetic patterns underlying familial rheumatoid arthritis, systemic lupus erythematosus and primary sjögren's syndrome highlight t cell-initiated autoimmunity. Ann Rheum Dis. 2020;79:268–75.

146. Kallberg H, Padyukov L, Plenge RM, Ronnelid J, Gregersen PK, Helm-van Mil AHM van der, et al. Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. Am J Hum Genet. 2007;80:867–75.

147. Nemtsova MV, Zaletaev DV, Bure IV, Mikhaylenko DS, Kuznetsova EB, Alekseeva EA, et al. Epigenetic changes in the pathogenesis of rheumatoid arthritis. Front Genet. 2019;10.

148. Tavasolian F, Abdollahi E, Rezaei R, Momtazi-Borojeni AA, Henrotin Y, Sahebkar A. Altered expression of MicroRNAs in rheumatoid arthritis. J Cell Biochem. 2018;119:478–87.

149. Crofford LJ. Use of NSAIDs in treating patients with arthritis. Arthritis Res Ther. 2013;15 Suppl 3:S2.

150. Schulze-Koops H, Skapenko A. Biosimilars in rheumatology: A review of the evidence and their place in the treatment algorithm. Rheumatology (Oxford). 2017;56 Suppl 4:iv30–48.

151. Johnson KJ, Sanchez HN, Schoenbrunner N. Defining response to TNF-inhibitors in rheumatoid arthritis: The negative impact of anti-TNF cycling and the need for a personalized medicine approach to identify primary non-responders. Clin Rheumatol. 2019;38:2967–76.

152. Jong PH de, Hazes JM, Han HK, Huisman M, Zeben D van, Lubbe PA van der, et al. Randomised comparison of initial triple DMARD therapy with methotrexate monotherapy in combination with low-dose glucocorticoid bridging therapy; 1-year data of the tREACH trial. Ann Rheum Dis. 2014;73:1331–9.

153. Verschueren P, De Cock D, Corluy L, Joos R, Langenaken C, Taelman V, et al. Methotrexate in combination with other DMARDs is not superior to methotrexate alone for remission induction with moderate-to-high-dose glucocorticoid bridging in early rheumatoid arthritis after 16 weeks of treatment: The CareRA trial. Ann Rheum Dis. 2015;74:27–34.

154. Punder YMR de, Fransen J, Kievit W, Houtman PM, Visser H, Laar MAFJ van de, et al. The prevalence of clinical remission in RA patients treated with anti-TNF: Results from the dutch rheumatoid arthritis monitoring (DREAM) registry. Rheumatology (Oxford). 2012;51:1610–7.

155. Farutin V, Prod'homme T, McConnell K, Washburn N, Halvey P, Etzel CJ, et al. Molecular profiling of rheumatoid arthritis patients reveals an association between innate and adaptive cell populations and response to anti-tumor necrosis factor. Arthritis Research & Therapy. 2019;21:216.

156. Urbanski AH, Araujo JD, Creighton R, Nakaya HI. Integrative biology approaches applied to human diseases. In: Husi H, editor. Computational biology. Brisbane (AU): Codon Publications; 2019.

157. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell. 2018;172:650–65.

158. Mercatelli D, Scalambra L, Triboli L, Ray F, Giorgi FM. Gene regulatory network inference resources: A practical overview. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2020;1863:194430.

159. Eguchi R, Karim MB, Hu P, Sato T, Ono N, Kanaya S, et al. An integrative network-based approach to identify novel disease genes and pathways: A case study in the context of inflammatory bowel disease. BMC Bioinformatics. 2018;19:264.

160. Karimizadeh E, Sharifi-Zarchi A, Nikaein H, Salehi S, Salamatian B, Elmi N, et al. Analysis of gene expression profiles and protein-protein interaction networks in multiple tissues of systemic sclerosis. BMC Medical Genomics. 2019;12:199.

161. Sahu A, Chowdhury HA, Gaikwad M, Chongtham C, Talukdar U, Phukan JK, et al. Integrative network analysis identifies differential regulation of neuroimmune system in schizophrenia and bipolar disorder. Brain, Behavior, & Immunity - Health. 2020;2:100023.

162. Haghjoo N, Moeini A, Masoudi-Nejad A. Introducing a panel for early detection of lung adenocarcinoma by using data integration of genomics, epigenomics, transcriptomics and proteomics. Exp Mol Pathol. 2020;112:104360.

163. Nicolle R, Radvanyi F, Elati M. CoRegNet: Reconstruction and integrated

analysis of co-regulatory networks. Bioinformatics. 2015;31:3066–8.

164. Benedetti E, Pučić-Baković M, Keser T, Gerstner N, Büyüközkan M, Štambuk T, et al. A strategy to incorporate prior knowledge into correlation network cutoff selection. Nat Commun. 2020;11:5153.

165. Zuo Y, Cui Y, Yu G, Li R, Ressom HW. Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. BMC Bioinformatics. 2017;18:99.

166. Greenfield A, Hafemeister C, Bonneau R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics. 2013;29:1060–7.

167. Mazein A, Knowles RG, Adcock I, Chung KF, Wheelock CE, Maitland-van der Zee AH, et al. AsthmaMap: An expert-driven computational representation of disease mechanisms. Clin Exp Allergy. 2018;48:916–8.

168. Fujita KA, Ostaszewski M, Matsuoka Y, Ghosh S, Glaab E, Trefois C, et al. Integrating pathways of parkinson's disease in a molecular interaction map. Mol Neurobiol. 2014;49:88–102.

169. Parton A, McGilligan V, Chemaly M, O'Kane M, Watterson S. New models of atherosclerosis and multi-drug therapeutic interventions. Bioinformatics. 2019;35:2449–57.

170. Singh V, Kalliolias GD, Ostaszewski M, Veyssiere M, Pilalis E, Gawron P, et al. RA-map: Building a state-of-the-art interactive knowledge base for rheumatoid arthritis. Database. 2020;2020.

171. Wu G, Zhu L, Dent JE, Nardini C. A comprehensive molecular interaction map for rheumatoid arthritis. PLOS ONE. 2010;5:e10137.

172. Saadatpour A, Albert R. A comparative study of qualitative and quantitative dynamic models of biological regulatory networks. EPJ Nonlinear Biomed Phys. 2016;4:1–13.

173. Abou-Jaoudé W, Traynard P, Monteiro PT, Saez-Rodriguez J, Helikar T, Thieffry D, et al. Logical modeling and dynamical analysis of cellular networks. Front Genet. 2016;7:94.

174. Helikar T, Konvalina J, Heidel J, Rogers JA. Emergent decision-making in

biological signal transduction networks. Proc Natl Acad Sci U S A. 2008;105:1913–8.

175. Calzone L, Tournier L, Fourquet S, Thieffry D, Zhivotovsky B, Barillot E, et al. Mathematical modelling of cell-fate decision in response to death receptor engagement. PLoS Comput Biol. 2010;6:e1000702.

176. Grieco L, Calzone L, Bernard-Pierrot I, Radvanyi F, Kahn-Perlès B, Thieffry D. Integrative modelling of the influence of MAPK network on cancer cell fate decision. PLoS Comput Biol. 2013;9:e1003286.

177. Flobak Å, Baudot A, Remy E, Thommesen L, Thieffry D, Kuiper M, et al. Discovery of drug synergies in gastric cancer cells predicted by logical modeling. PLOS Computational Biology. 2015;11:e1004426.

178. Cho S-H, Park S-M, Lee H-S, Lee H-Y, Cho K-H. Attractor landscape analysis of colorectal tumorigenesis and its reversion. BMC Systems Biology. 2016;10:96.

179. Traynard P, Fauré A, Fages F, Thieffry D. Logical model specification aided by model-checking techniques: Application to the mammalian cell cycle regulation. Bioinformatics. 2016;32:i772–80.

180. Fumiã HF, Martins ML. Boolean network model for cancer pathways: Predicting carcinogenesis and targeted therapy outcomes. PLoS One. 2013;8:e69008.

181. Hu Y, Gu Y, Wang H, Huang Y, Zou YM. Integrated network model provides new insights into castration-resistant prostate cancer. Sci Rep. 2015;5:17280.

182. Wang R-S, Saadatpour A, Albert R. Boolean modeling in systems biology: An overview of methodology and applications. Phys Biol. 2012;9:055001.

183. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biology. 2017;18:77.

184. The cost of sequencing a human genome. Genome.gov. `https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost`. Accessed 7 Sep 2021.

185. Schaschl H, Aitman TJ, Vyse TJ. Copy number variation in the human genome and its implication in autoimmunity. Clin Exp Immunol. 2009;156:12–6.

186. Nowakowska B. Clinical interpretation of copy number variants in the human

genome. J Appl Genet. 2017;58:449–57.

187. Ballarati L, Rossi E, Bonati MT, Gimelli S, Maraschio P, Finelli P, et al. 13q deletion and central nervous system anomalies: Further insights from karyotype–phenotype analyses of 14 patients. J Med Genet. 2007;44:e60.

188. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2009;84:148–61.

189. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nat Genet. 2011;43:838–46.

190. Coe BP, Girirajan S, Eichler EE. The genetic variability and commonality of neurodevelopmental disease. Am J Med Genet C Semin Med Genet. 2012;160C:118–29.

191. Filges I, Röthlisberger B, Noppen C, Boesch N, Wenzel F, Necker J, et al. Familial 14.5 mb interstitial deletion 13q21.1-13q21.33: Clinical and array-CGH study of a benign phenotype in a three-generation family. Am J Med Genet A. 2009;149A:237–41.

192. Nowakowska BA, Obersztyn E, Szymańska K, Bekiesińska-Figatowska M, Xia Z, Ricks CB, et al. Severe mental retardation, seizures, and hypotonia due to deletions of MEF2C. Am J Med Genet B Neuropsychiatr Genet. 2010;153B:1042–51.

193. Valsesia A, Macé A, Jacquemont S, Beckmann JS, Kutalik Z. The growing importance of CNVs: New insights for detection and clinical interpretation. Front Genet. 2013;4:92.

194. McKinney C, Merriman ME, Chapman PT, Gow PJ, Harrison AA, Highton J, et al. Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. Ann Rheum Dis. 2008;67:409–13.

195. Li YR, Glessner JT, Coe BP, Li J, Mohebnasab M, Chang X, et al. Rare copy number variants in over 100,000 european ancestry subjects reveal multiple disease associations. Nat Commun. 2020;11:255.

196. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet. 2007;39:S7–15.

197. Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: Past, present and future. J Hum Genet. 2010;55:403–15.

198. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.

199. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12:363–76.

200. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12:363–76.

201. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. BMC Bioinformatics. 2013;14 Suppl 11:S1.

202. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. ExomeDepth: A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics (Oxford, England). 2012;28:2747–54.

203. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, et al. Xhmm: Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. American Journal of Human Genetics. 2012;91:597–607.

204. Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Research. 2016;1:20.

205. Povysil G, Tzika A, Vogt J, Haunschmid V, Messiaen L, Zschocke J, et al. Panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. Human Mutation. 2017;38:889–97.

206. CANOES: Detecting rare copy number variants from whole exome sequencing

data.

207. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. CoNiFeR: Copy number variation detection and genotyping from exome sequence data. Genome Research. 2012;22:1525–32.

208. D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. Nucleic Acids Research. 2016;44:e154.

209. Jiang Y, Wang R, Urrutia E, Anastopoulos IN, Nathanson KL, Zhang NR. CODEX2: Full-spectrum copy number variation detection by high-throughput DNA sequencing. Genome Biology. 2018;19:202.

210. Packer JS, Maxwell EK, O'Dushlaine C, Lopez AE, Dewey FE, Chernomorsky R, et al. CLAMMS: A scalable algorithm for calling common and rare copy number variants from exome sequencing data. Bioinformatics. 2016;32:133–5.

211. Zhao L, Liu H, Yuan X, Gao K, Duan J. Comparative study of whole exome sequencing-based copy number variation detection tools. BMC Bioinformatics. 2020;21:97.

212. Roca I, González-Castro L, Fernández H, Couce ML, Fernández-Marmiesse A. Free-access copy-number variant detection tools for targeted next-generation sequencing data. Mutat Res Rev Mutat Res. 2019;779:114–25.

213. Sadedin SP, Ellis JA, Masters SL, Oshlack A. Ximmer: A system for improving accuracy and consistency of CNV calling from exome data. Gigascience. 2018;7.

214. Hong CS, Singh LN, Mullikin JC, Biesecker LG. Assessing the reproducibility of exome copy number variations predictions. Genome Med. 2016;8:82.

215. Moreno-Cabrera JM, Del Valle J, Castellanos E, Feliubadaló L, Pineda M, Brunet J, et al. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. Eur J Hum Genet. 2020. `https://doi.org/10.1038/s41431-020-0675-z`.

216. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with snakemake. F1000Res. 2021;10:33.

217. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of chromosomal imbalance and phenotype in humans

using ensembl resources. Am J Hum Genet. 2009;84:524–33.

218. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The database of genomic variants: A curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42 Database issue:D986–992.

219. Bartenhagen C, Dugas M. RSVSim: An r/bioconductor package for the simulation of structural variations. Bioinformatics. 2013;29:1679–81.

220. Kim S, Jeong K, Bafna V. Wessim: A whole-exome sequencing simulator based on in silico exome capture. Bioinformatics. 2013;29:1076–7.

221. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.

222. McElroy KE, Luciani F, Thomas T. GemSIM: General, error-model based simulator of next-generation sequencing data. BMC Genomics. 2012;13:74.

223. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25:1754–60.

224. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012;28:2711–8.

225. Picard toolkit. 2019.

226. Love M. exomeCopy: Copy number variant detection from exome sequencing read depth. 2021.

227. Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, et al. Cn.MOPS: Mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012;40:e69.

228. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

229. Huang J, Zhou J, Xiao M, Mao X, Zhu L, Liu S, et al. The association of complex genetic background with the prognosis of acute leukemia with ambiguous lineage. Sci Rep. 2021;11:24290.

230. Ripen AM, Chiow MY, Rama Rao PR, Mohamad SB. Revealing chronic granulomatous disease in a patient with williams-beuren syndrome using whole exome sequencing. Front Immunol. 2021;12:778133.

231. García Bohórquez B, Aller E, Rodríguez Muñoz A, Jaijo T, García García G, Millán JM. Updating the genetic landscape of inherited retinal dystrophies. Front Cell Dev Biol. 2021;9:645600.

232. Aslam MM, John P, Fan K-H, Bhatti A, Feingold E, Demirci FY, et al. Association of VPREB1 gene copy number variation and rheumatoid arthritis susceptibility. Dis Markers. 2020;2020:7189626.

233. Uddin M, Sturge M, Rahman P, Woods MO. Autosome-wide copy number variation association analysis for rheumatoid arthritis using the WTCCC high-density SNP genotype data. J Rheumatol. 2011;38:797–801.

234. Veyssiere M. Etude de la composante génétique de la polyarthrite rhumatoïde par séquençage d'exomes : Contribution des variants rares. These de doctorat. Université Paris-Saclay (ComUE); 2019.

235. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio]. 2013.

236. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10:giab008.

237. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, et al. AnnotSV: An integrated tool for structural variations annotation. Bioinformatics. 2018;34:3572–4.

238. Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: A joint consensus recommendation of the american college of medical genetics and genomics (ACMG) and the clinical genome resource (ClinGen). Genet Med. 2020;22:245–57.

239. Wu X-Y, Li K-T, Yang H-X, Yang B, Lu X, Zhao L-D, et al. Complement c1q synergizes with PTX3 in promoting NLRP3 inflammasome over-activation and pyroptosis in rheumatoid arthritis. J Autoimmun. 2020;106:102336.

240. Zhai Z, Yang F, Xu W, Han J, Luo G, Li Y, et al. Attenuation of rheuma-

toid arthritis through the inhibition of tumor necrosis factor-induced caspase 3/gasdermin e-mediated pyroptosis. Arthritis Rheumatol. 2021. `https://doi.org/10.1002/art.41963`.

241. Wu T, Zhang X-P, Zhang Q, Zou Y-Y, Ma J-D, Chen L-F, et al. Gasdermin-e mediated pyroptosis—a novel mechanism regulating migration, invasion and release of inflammatory cytokines in rheumatoid arthritis fibroblast-like synoviocytes. Frontiers in Cell and Developmental Biology. 2022;9.

242. Spriel AB van, Puls KL, Sofi M, Pouniotis D, Hochrein H, Orinska Z, et al. A regulatory role for CD37 in t cell proliferation. J Immunol. 2004;172:2953–61.

243. Gartlan KH, Wee JL, Demaria MC, Nastovska R, Chang TM, Jones EL, et al. Tetraspanin CD37 contributes to the initiation of cellular immunity by promoting dendritic cell migration. European Journal of Immunology. 2013;43:1208–19.

244. Wee JL, Schulze KE, Jones EL, Yeung L, Cheng Q, Pereira CF, et al. Tetraspanin CD37 regulates b2 integrin-mediated adhesion and migration in neutrophils. J Immunol. 2015;195:5770–9.

245. Meyer-Wentrup F, Figdor CG, Ansems M, Brossart P, Wright MD, Adema GJ, et al. Dectin-1 interaction with tetraspanin CD37 inhibits IL-6 production. The Journal of Immunology. 2007;178:154–62.

246. Höglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, et al. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. Sci Rep. 2019;9:16844.

247. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11:446–50.

248. Saint Pierre A, Génin E. How important are rare variants in common disease? Brief Funct Genomics. 2014;13:353–61.

249. Chung SA, Shum AK. Rare variants, autoimmune disease, and arthritis. Curr Opin Rheumatol. 2016;28:346–51.

250. Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. Nat Rev Genet. 2000;1:182–90.

251. Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and

rare variants affecting complex traits. Hum Mol Genet. 2013;22:R16–21.

252. Suzuki A, Terao C, Yamamoto K. Linking of genetic risk variants to disease-specific gene expression via multi-omics studies in rheumatoid arthritis. Semin Arthritis Rheum. 2019;49:S49–53.

253. Okada Y, Diogo D, Greenberg JD, Mouassess F, Achkar WAL, Fulton RS, et al. Integration of sequence data from a consanguineous family with genetic data from an outbred population identifies PLB1 as a candidate rheumatoid arthritis risk gene. PLoS One. 2014;9:e87645.

254. Beecham GW, Vardarajan B, Blue E, Bush W, Jaworski J, Barral S, et al. Rare genetic variation implicated in non-hispanic white families with alzheimer disease. Neurol Genet. 2018;4:e286.

255. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.

256. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

257. Smit A, Hublet R, Green P. RepeatMasker open-4.0. 2013.

258. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: Organization and impact within the current human genome project assembly. Genome Res. 2001;11:1005–17.

259. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. Science. 2002;297:1003–7.

260. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.

261. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: Fast, flexible annotation of genetic variants. Genome Biology. 2016;17:118.

262. Nevanlinna HR. The finnish population structure. A genetic and genealogical study. Hereditas. 1972;71:195–236.

263. Norio R. Finnish disease heritage i: Characteristics, causes, background. Hum

Genet. 2003;112:441–56.

264. Exome variant server, NHLBI GO exome sequencing project (ESP). Seattle, WA.

265. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526:82–90.

266. Walsh R, Thomson KL, Ware JS, Funke BH, Woodley J, McGuire KJ, et al. Reassessment of mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genet Med. 2017;19:192–203.

267. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.

268. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: Predicting effects of amino acid substitutions on proteins. Nucleic Acids Research. 2012;40:W452–7.

269. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.

270. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61.

271. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7:575–6.

272. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.

273. Felsenstein J, Churchill GA. A hidden markov model approach to variation among sites in rate of evolution. Mol Biol Evol. 1996;13:93–104.

274. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6:e1001025.

275. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-splice-improving

genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021;13:31.

276. Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31:761–3.

277. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24:2125–37.

278. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Human Mutation. 2016;37:235–41.

279. Liu X, White S, Peng B, Johnson AD, Brody JA, Li AH, et al. WGSA: An annotation pipeline for human genome sequencing studies. Journal of Medical Genetics. 2016;53:111–2.

280. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. Hum Mutat. 2013;34:57–65.

281. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48:214–20.

282. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017;49:618–24.

283. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017;18.

284. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res. 2018;46:7793–804.

285. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3:S3.

286. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti

S, et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. The American Journal of Human Genetics. 2016;99:877–85.

287. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22:1790–7.

288. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology. 2015;16:22.

289. Hu H, Roach JC, Coon H, Guthery SL, Voelkerding KV, Margraf RL, et al. A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat Biotechnol. 2014;32:663–9.

290. Whitford W, Lehnert K, Snell RG, Jacobsen JC. Evaluation of the performance of copy number variant prediction tools for the detection of deletions from whole genome sequencing data. J Biomed Inform. 2019;94:103174.

291. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 2019;20.

292. Cameron DL, Stefano LD, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun. 2019;10:1–11.

293. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. Genome Biol. 2014;15:R84.

294. Suvakov M, Panda A, Diesh C, Holmes I, Abyzov A. CNVpytor: A tool for CNV/CNA detection and analysis from read depth and allele imbalance in whole genome sequencing. 2021.

295. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974–84.

296. Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. Hum Mutat. 2017;38:788–97.

297. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.

298. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA database. Nucleic Acids Res. 2020;48:D948–55.

299. He Y, You X, Tao S, He J, Zhu F. HLA-DRB1*15:01:43 and HLA-DRB1*15:01:44 alleles were identified by next-generation sequencing. HLA. 2022. https://doi.org/10.1111/tan.14535.

300. Ralazamahaleo M, Elsermans V, Top I, Guidicelli G, Visentin J. Characterization of the novel HLA-DRB1*03:147 allele by sequencing-based typing. HLA. 2019;93:53–4.

301. Li X, Poschmann S, Chen Q, Fazeli W, Oundjian NJ, Snoeijen-Schouwenaars FM, et al. De novo BK channel variant causes epilepsy by affecting voltage gating but not ca2+ sensitivity. Eur J Hum Genet. 2018;26:220–9.

302. Liang L, Li X, Moutton S, Schrier Vergano SA, Cogné B, Saint-Martin A, et al. De novo loss-of-function KCNMA1 variants are associated with a new multiple malformation syndrome and a broad spectrum of developmental and neurological phenotypes. Hum Mol Genet. 2019;28:2937–51.

303. Tanner MR, Pennington MW, Chauhan SS, Laragione T, Gulko PS, Beeton C. KCa1.1 and kv1.3 channels regulate the interactions between fibroblast-like synoviocytes and t lymphocytes during rheumatoid arthritis. Arthritis Res Ther. 2019;21:6.

304. Hu X, Laragione T, Sun L, Koshy S, Jones KR, Ismailov II, et al. KCa1.1 potassium channels regulate key proinflammatory and invasive properties of fibroblast-like synoviocytes in rheumatoid arthritis. J Biol Chem. 2012;287:4014–22.

305. Tanner MR, Pennington MW, Laragione T, Gulko PS, Beeton C. KCa1.1 channels regulate b1-integrin function and cell adhesion in rheumatoid arthritis fibroblast-like synoviocytes. FASEB J. 2017;31:3309–20.

306. Lagace TA. PCSK9 and LDLR degradation: Regulatory mechanisms in circulation and in cells. Curr Opin Lipidol. 2014;25:387–93.

307. Ricci C, Ruscica M, Camera M, Rossetti L, Macchi C, Colciago A, et al. PCSK9

induces a pro-inflammatory response in macrophages. Sci Rep. 2018;8:2267.

308. Momtazi-Borojeni AA, Sabouri-Rad S, Gotto AM Jr, Pirro M, Banach M, Awan Z, et al. PCSK9 and inflammation: A review of experimental and clinical evidence. European Heart Journal - Cardiovascular Pharmacotherapy. 2019;5:237–45.

309. Brown M, Ahmed S. Emerging role of proprotein convertase subtilisin/kexin type-9 (PCSK-9) in inflammation and diseases. Toxicology and Applied Pharmacology. 2019;370:170–7.

310. Liu A, Rahman M, Hafström I, Ajeganova S, Frostegård J. Proprotein convertase subtilisin kexin 9 is associated with disease activity and is implicated in immune activation in systemic lupus erythematosus. Lupus. 2020;29:825–35.

311. Frostegård J, Ahmed S, Hafström I, Ajeganova S, Rahman M. Low levels of PCSK9 are associated with remission in patients with rheumatoid arthritis treated with anti-TNF-a: Potential underlying mechanisms. Arthritis Research & Therapy. 2021;23:32.

312. Fukui I, Tanaka K, Murachi T. Extracellular appearance of calpain and calpastatin in the synovial fluid of the knee joint. Biochem Biophys Res Commun. 1989;162:559–66.

313. Yamamoto S, Shimizu K, Shimizu K, Suzuki K, Nakagawa Y, Yamamuro T. Calcium-dependent cysteine proteinase (calpain) in human arthritic synovial joints. Arthritis Rheum. 1992;35:1309–17.

314. Ishikawa H, Nakagawa Y, Shimizu K, Nishihara H, Matsusue Y, Nakamura T. Inflammatory cytokines induced down-regulation of m-calpain mRNA expression in fibroblastic synoviocytes from patients with osteoarthritis and rheumatoid arthritis. Biochem Biophys Res Commun. 1999;266:341–6.

315. Esmon CT, Schwarz HP. An update on clinical and basic aspects of the protein c anticoagulant pathway. Trends Cardiovasc Med. 1995;5:141–8.

316. Xue M, Dervish S, McKelvey KJ, March L, Wang F, Little CB, et al. Activated protein c targets immune cells and rheumatoid synovial fibroblasts to prevent inflammatory arthritis in mice. Rheumatology. 2019;58:1850–60.

317. Xue M, Dervish S, Harrison LC, Fulcher G, Jackson CJ. Activated protein c

inhibits pancreatic islet inflammation, stimulates t regulatory cells, and prevents diabetes in non-obese diabetic (NOD) mice *. Journal of Biological Chemistry. 2012;287:16356–64.

318. Lichtnekert J, Rupanagudi KV, Kulkarni OP, Darisipudi MN, Allam R, Anders H-J. Activated protein c attenuates systemic lupus erythematosus and lupus nephritis in MRL-fas(lpr) mice. J Immunol. 2011;187:3413–21.

319. Roberts MF. First thoughts on lipid second messengers. Trends Cell Biol. 1994;4:219–23.

320. Hodgkin MN, Pettitt TR, Martin A, Michell RH, Pemberton AJ, Wakelam MJ. Diacylglycerols and phosphatidates: Which molecular species are intracellular messengers? Trends Biochem Sci. 1998;23:200–4.

321. Zhong X-P, Hainey EA, Olenchock BA, Jordan MS, Maltzman JS, Nichols KE, et al. Enhanced t cell responses due to diacylglycerol kinase zeta deficiency. Nat Immunol. 2003;4:882–90.

322. Mahajan S, Mellins ED, Faccio R. Diacylglycerol kinase z regulates macrophage responses in juvenile arthritis and cytokine storm syndrome mouse models. J Immunol. 2020;204:137–46.

323. Edwards CJ, Feldman JL, Beech J, Shields KM, Stover JA, Trepicchio WL, et al. Molecular profile of peripheral blood mononuclear cells from patients with rheumatoid arthritis. Mol Med. 2007;13:40–58.

324. Webster AP, Plant D, Ecker S, Zufferey F, Bell JT, Feber A, et al. Increased DNA methylation variability in rheumatoid arthritis-discordant monozygotic twins. Genome Med. 2018;10:64.

325. Rhead B, Holingue C, Cole M, Shao X, Quach HL, Quach D, et al. Rheumatoid arthritis naive t cells share hypermethylation sites with synoviocytes. Arthritis Rheumatol. 2017;69:550–9.

326. Teixeira VH, Olaso R, Martin-Magniette M-L, Lasbleiz S, Jacq L, Oliveira CR, et al. Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. PLoS One. 2009;4:e6803.

327. Hashimoto T, Harita Y, Takizawa K, Urae S, Ishizuka K, Miura K, et al.

In vivo expression of NUP93 and its alteration by NUP93 mutations causing focal segmental glomerulosclerosis. Kidney Int Rep. 2019;4:1312–22.

328. Lee B-S, Lee B-K, Iyer VR, Sleckman BP, Shaffer AL, Ippolito GC, et al. Corrected and republished from: BCL11A is a critical component of a transcriptional network that activates recombinase activating gene expression and v(D)J recombination. Mol Cell Biol. 2017;38:e00362–17.

329. Zerrouk N, Miagoux Q, Dispot A, Elati M, Niarakis A. Identification of putative master regulators in rheumatoid arthritis synovial fibroblasts using gene expression data and network inference. Scientific Reports. 2020;10:16236.

330. Lecca P. Machine learning for causal inference in biological networks: Perspectives of this challenge. Frontiers in Bioinformatics. 2021;1:45.

331. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

332. Tibshirani R. Variance stabilization and the bootstrap. Biometrika. 1988;75:433–44.

333. Miagoux Q, Singh V, Mézquita D de, Chaudru V, Elati M, Petit-Teixeira E, et al. Inference of an integrative, executable network for rheumatoid arthritis combining data-driven machine learning approaches and a state-of-the-art mechanistic disease map. Journal of Personalized Medicine. 2021;11:785.

334. Chebil I, Nicolle R, Santini G, Rouveirol C, Elati M. Hybrid method inference for the construction of cooperative regulatory network in human. IEEE Trans Nanobioscience. 2014;13:97–103.

335. Elati M, Neuvial P, Bolotin-Fukuhara M, Barillot E, Radvanyi F, Rouveirol C. LICORN: Learning cooperative regulation networks from gene expression data. Bioinformatics. 2007;23:2407–14.

336. Agrawal R, Srikant R. Fast algorithms for mining association rules.:13.

337. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: Transcription factor regulation inferred from integrating genome-wide ChIP-x experiments. Bioinformatics. 2010;26:2438–44.

338. ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004;306:636–40.

339. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res. 2017;45:D408–14.

340. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47:D607–13.

341. Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, et al. Predictive regulatory models in drosophila melanogaster by integrative inference of transcriptional networks. Genome Res. 2012;22:1334–49.

342. Hoksza D, Gawron P, Ostaszewski M, Smula E, Schneider R. MINERVA API and plugins: Opening molecular network analysis and visualization to the community. Bioinformatics. 2019;35:4496–8.

343. Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H. CellDesigner 3.5: A versatile modeling tool for biochemical networks. Proceedings of the IEEE. 2008;96:1254–65.

344. Rougny A, Touré V, Moodie S, Balaur I, Czauderna T, Borlinghaus H, et al. Systems biology graphical notation: Process description language level 1 version 2.0. J Integr Bioinform. 2019;16:/j/jib.2019.16.issue–2/jib–2019–0022/jib–2019–0022.xml.

345. Aghamiri SS, Singh V, Naldi A, Helikar T, Soliman S, Niarakis A. Automated inference of boolean models from molecular interaction maps using CaSQ. Bioinformatics. 2020;36:4473–82.

346. Le Novère N. Quantitative and logic modelling of molecular and gene networks. Nat Rev Genet. 2015;16:146–58.

347. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695.

348. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

349. Gustavsen JA, Pai S, Isserlin R, Demchak B, Pico AR. RCy3: Network biology

using cytoscape from within r. F1000Res. 2019;8:1774.

350. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48:D845–55.

351. Bonnet E, Calzone L, Rovera D, Stoll G, Barillot E, Zinovyev A. BiNoM 2.0, a cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. BMC Systems Biology. 2013;7:18.

352. Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, et al. Shiny: Web application framework for r. 2021.

353. Shah O, Shannon P, Chandrakar D. cyjShiny: cyjShiny. 2021.

354. Helikar T, Kowal B, McClenathan S, Bruckner M, Rowley T, Madrahimov A, et al. The cell collective: Toward an open and collaborative approach to systems biology. BMC Syst Biol. 2012;6:96.

355. Chaouiya C, Naldi A, Thieffry D. Logical modelling of gene regulatory networks with GINsim. In: Helden J van, Toussaint A, Thieffry D, editors. Bacterial molecular networks: Methods and protocols. New York, NY: Springer; 2012. pp. 463–79.

356. Ogata A, Kato Y, Higa S, Yoshizaki K. IL-6 inhibitor for the treatment of rheumatoid arthritis: A comprehensive review. Modern Rheumatology. 2019;29:258–67.

357. Gonzalo-Gil E, Criado G, Santiago B, Dotor J, Pablos JL, Galindo M. Transforming growth factor (TGF)-b signalling is increased in rheumatoid synovium but TGF-b blockade does not modify experimental arthritis. Clinical and Experimental Immunology. 2013;174:245–55.

358. Hennigan S, Kavanaugh A. Interleukin-6 inhibitors in the treatment of rheumatoid arthritis. Ther Clin Risk Manag. 2008;4:767–75.

359. Woodrick R, Ruderman EM. Anti-interleukin-6 therapy in rheumatoid arthritis. Bulletin of the NYU Hospital for Joint Diseases. 2010;68:211–1.

360. Ogata A, Kato Y, Higa S, Yoshizaki K. IL-6 inhibitor for the treatment of rheumatoid arthritis: A comprehensive review. Modern Rheumatology. 2019;29:258–67.

361. Choy EH, De Benedetti F, Takeuchi T, Hashizume M, John MR, Kishimoto T. Translating IL-6 biology into effective treatments. Nat Rev Rheumatol. 2020;16:335–45.

362. Sakuma M, Hatsushika K, Koyama K, Katoh R, Ando T, Watanabe Y, et al. TGF-beta type i receptor kinase inhibitor down-regulates rheumatoid synoviocytes and prevents the arthritis induced by type II collagen antibody. Int Immunol. 2007;19:117–26.

363. Kim Y, Yi H, Jung H, Rim YA, Park N, Kim J, et al. A dual target-directed agent against interleukin-6 receptor and tumor necrosis factor a ameliorates experimental arthritis. Sci Rep. 2016;6:20150.

364. Boleto G, Kanagaratnam L, Dramé M, Salmon J-H. Safety of combination therapy with two bDMARDs in patients with rheumatoid arthritis: A systematic review and meta-analysis. Seminars in Arthritis and Rheumatism. 2019;49:35–42.

365. Biologic combination therapy for RA may increase risk for side effects. Rheumatology.medicinematters.com. 2019. https://rheumatology.medicinematters.com/rheumatoid-arthritis-/biologics/biologic-combination-therapy-for-ra-may-increase-risk-for-side-e/16387312. Accessed 7 Feb 2022.

366. Canovas B, Nebreda AR. Diversity and versatility of p38 kinase signalling in health and disease. Nat Rev Mol Cell Biol. 2021;22:346–66.

367. Clark AR, Dean JL. The p38 MAPK pathway in rheumatoid arthritis: A sideways look. Open Rheumatol J. 2012;6:209–19.

368. Haller V, Nahidino P, Forster M, Laufer SA. An updated patent review of p38 MAP kinase inhibitors (2014-2019). Expert Opinion on Therapeutic Patents. 2020;30:453–66.

369. Bonilla-Hernán MG, Miranda-Carús ME, Martin-Mola E. New drugs beyond biologics in rheumatoid arthritis: The kinase inhibitors. Rheumatology. 2011;50:1542–50.

370. Ohori M. ERK inhibitors as a potential new therapy for rheumatoid arthritis. Drug News Perspect. 2008;21:245–50.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre: Approche globale intégrative pour l'identification de nouvelles cibles moléculaires dans la Polyarthrite Rhumatoïde

Mots clés: Polyarthrite rhumatoïde - Variants rares - Génomique humaine - Inférence de réseaux - Biologie des systèmes

Résumé: La polyarthrite rhumatoïde (PR) est une maladie multifactorielle complexe et autoimmune, impliquant des facteurs génétiques, épigénétiques et environnementaux. Elle touche en moyenne 0.25 à 0.46% de la population mondiale. Aujourd'hui, notre connaissance de la composante génétique de la maladie est estimée à uniquement 50%. Ainsi, afin d'expliquer la part d'héritabilité manquante, nous avons en premier lieu focalisé nos recherches sur l'identification de nouveaux facteurs génétiques de la PR en analysant les variants rares à partir de données exoniques et pangénomiques de familles de patients atteints de PR dans deux cohortes différentes. À partir de simulation de données exoniques, nous avons identifié des outils performants pour l'identification de CNVs, incluant CODEX2. Ces outils, utilisés par la suite sur les données exoniques, ont permis d'identifier 3 CNVs rares qui ont cependant montré une pénétrance incomplète et/ou la présence de phénocopie, malgré une potentielle implication pour deux d'entre eux dans la PR ou dans le système immunitaire. L'étude des variants rares à partir de données pangénomiques nous a permis d'identifier 15 variants (10 SNVs, 2 indels et 3 CNVs) rares spécifiques de la PR. Sept gènes impactés par ces variants (incluant 4 SNVs, 1 indel et 2 CNVs) ont montré une implication dans la physiopathologie de la PR selon la littérature. Dans la seconde partie de cette thèse, nous avons utilisé la biologie computationnelle des systèmes afin d'étudier des mécanismes complexes impliqués dans la pathologie de la PR tels que les facteurs de transcription (TF), régulateurs clés dans les maladies, ne pouvant être étudiés uniquement à l'aide de données génomiques. Pour cela, un réseau global et spécifique de la PR a été créé, en combinant des données multi-omiques et des méthodes d'inférence et une carte d'interaction moléculaire de la PR. Ce réseau a ensuite été utilisé comme base afin d'analyser et d'identifier les voies de signalisation affectées par la réponse à des traitements anti-TNF de patients atteints de PR. Dans ce but nous avons utilisé le formalisme et transformé notre réseau en modèle booléen afin de réaliser des simulations in silico, dans le but de répliquer des perturbations combinées et isolées induites par des thérapies et des prédispositions génétiques. Ces résultats montrent que le blocage de TNF n'est pas suffisant pour stopper l'activité des TFs liés à une inflammation, suggérant que l'utilisation de thérapies combinées et ciblées est un scénario plausible pour des patients non-répondeurs à des traitements anti-TNF.

Title: Integrative, Global approach for the identification of novel biomolecular targets in Rheumatoid Arthritis

Keywords: Rheumatoid arthritis - Rare variants - Human genomics - Network inference - Systems biology

Abstract: Rheumatoid arthritis (RA) is a multifactorial, complex autoimmune disease that involves various genetic, environmental, and epigenetic factors. It affects in average 0.25 to 0.46% of world population. Nowadays, we understand only 50 % of RA genetic component. Genetic factors involved in RA include a major genetic factor, the HLA-DRB1 gene, and about one hundred of susceptibility factors. In the first part of this thesis, we aimed to explain the missing heritability of RA, by looking for new rare variants with exonic and pangenomic data from relatives of RA patients of two samples. We first conducted an analysis on simulated exonic data where we identify performant detection tools, including CODEX2. These tools were then used to analyse our real exonic dataset, where 3 CNVs were identified, but unfortunately showed an incomplete penetrance and/or a presence of phenocopy, despite that two of them concern genes involved in RA or immune system. The study of pangenomic data showed 15 rare variants (10 SNVs, 1 indel and 2 CNVs) that were RA specific. Seven genes impacted by these variants (4 SNVs, 1 indel and 2 CNVs) were implicated in RA physiopathology, according to the literature. In the second part of this thesis, we used computational system biology approaches to study the complex mechanisms involved in RA pathology, including the regulation of key disease –related transcription factors, which cannot be fully understood with the use of only genomic data. For this purpose, we built an integrative global network for RA, using multi-omic data and statistical inference along with prior knowledge encoded in the form of a molecular interaction map. The network is used as a template to study RA patient response to anti-TNF treatment and to identify master regulators and upstream cascades affected by the treatment. Finally, we employ the logical formalism and transform a subnetwork into a Boolean model to perform in silico simulations mimicking the effects of single and combined perturbations induced by therapies and genetic predisposition. The results show that TNF blockage was not sufficient to downregulate the activity of TFs linked to inflammation, suggesting that combined targeted therapies might be a plausible scenario for the non-responders to anti-TNF therapy.