Une approche de modélisation de biologie des systèmes sur la spondylarthrite

Emmanuel Chaplais

UNIVERSITÉ DE VERSAILLES SAINT-QUENTIN-EN-YVELINES
UFR DES SCIENCES DE LA SANTÉ – SIMONE VEIL
ECOLE DOCTORALE GAO

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE
VERSAILLES SAINT-QUENTIN-EN-YVELINES

Sciences de la vie et de la santé

Spécialité : Génétique Humaine

Présentée et soutenue publiquement par

Emmanuel Chaplais

Le 28 Septembre 2015

UNE APPROCHE DE MODÉLISATION DE BIOLOGIE DES SYSTÈMES SUR LA SPONDYLARTHRITE

JURY

Mme le Dr. Valentina BOEVA	Rapporteur
Mr le Pr. Philippe Broet	Rapporteur
Mr le Pr. Christophe AMBROISE	Examinateur
Mr le Pr. Henri-Jean GARCHON	Directeur de thèse

Résumé

La Spondyloarthrite (SpA) est un rhumatisme inflammatoire chronique fréquent, avec une prévalence de 0,43 % en France. Elle consiste en une atteinte prédominante du squelette axial, mais aussi des articulations périphériques, et peut conduire à une immobilité du rachis et des articulations sacro-iliaques. Des atteintes extra-articulaires sont fréquentes, telles qu'une uvéite, un psoriasis ou une maladie inflammatoire chronique de l'intestin. Les traitements actuels ne sont que symptomatiques, ciblant principalement les manifestations inflammatoires. L'étiologie de la SpA est multifactorielle avec une composante génétique dominée par l'association forte et bien connue avec l'allèle HLA-B27. Cependant, ce facteur génétique n'est clairement pas suffisant pour induire le développement de la maladie. L'objectif de ce projet de thèse était donc d'identifier d'autres facteurs génétiques à l'origine du développement de la SpA.

Mon travail a porté sur l'analyse de deux jeux de données complémentaires, dans une perspective de biologie des systèmes. Dans une première partie, j'ai conduit une analyse de liaison dans 210 familles atteintes de la maladie représentant 1310 personnes génotypées avec des puces Affymetrix 250k. Une nouvelle région significativement liée à la SpA a été détectée en 13q13, avec un intervalle de 1,3 Mb défini par des haplotypes recombinants chez les patients. Cette région est en cours de séquençage pour identifier les variants causaux.

Ensuite, une analyse transcriptomique des cellules dendritiques dérivées des monocytes de 23 patients HLA-B27+, 23 témoins sains HLA-B27+ et 21 témoins sains HLA-B27-, et stimulées ou non par du LPS, a tenté de distinguer les gènes dont l'expression est modifiée par la maladie de ceux influencés par l'allèle HLA-B27 seul. L'annotation fonctionnelle et une analyse par réseau de gènes ont mis en évidence l'inhibition chez les patients des étapes précoces de la biosynthèse du cholestérol. La validation expérimentale par qPCR et par analyse de profils lipidiques est en cours dans le laboratoire. Pour conduire ces analyses de réseaux, j'ai développé un package R, stringgaussnet, permettant de combiner simplement des réseaux de gènes sémantiques et gaussiens et de les visualiser dans Cytoscape.

Abstract

Spondyloarthritis is a frequent chronic inflammatory rheumatism, with a prevalence of 0.43 % in France. This disease presents axial skeleton injuries, but also on peripheral joints, and can results in a total spinal and sacro-iliac motility loss. Extra-articular features including uveitis, psoriasis and inflammatory bowel disease are frequent. Current SpA treatments are only symptomatic, relieving inflammatory symptoms. SpA etiology is largely multifactorial with a genetic component dominated by the long-known strong association with the HLA-B27 allele. This allele, however, is not sufficient for the disease to occur. This thesis project objective was then to identify other genetic factors in the origin of SpA.

My work was mainly divided in two complementary data analyses, in a way to get a systems biology approach. The first one consisted in proceed linking analyses on data from Affymetrix genotyping chips gathered from DNA of 1310 people grouped in 210 families. This study allowed notably to detect a new significantly linked region to SpA : 13q13, with an interval of 1.3 Mb. This part of genome is currently being sequenced to allow a better causal SNP identification.

Secondly, an Affymetrix HumanGene 1.0 st transcriptomic chips analysis was performed on MD-DCs extracted from 68 people, stimulated or not by LPS during 6 or 24 hours. This cohort was grouped between 23 patients HLA-B27+, 23 healthy controls HLA-B27+ and 21 healthy controls HLA-B27-. I could notice that HLA-B27 allele is farly enough to considerably affect cell transcriptomic profiles, which encourages to include HLA-B27+ healthy controls. Otherwise, a gene network analysis allowed me to highlight on an inhibition of early steps of cholesterol biosyntthesis. Validations by qPCR and and lipid profiles are currently done in the laboratory. Finally, I published an R package, stringaussnetnn allowing to simply create semantic and gaussien gene networks, and then to import those into Cytoscape.

REMERCIEMENTS

Je tiens à remercier les membres de mon jury de thèse, qui m'ont permis de valider et de valoriser mon travail effectué pendant trois ans dans ce laboratoire. Je remercie particulièrement le docteur Valentina Boeva et le professeur Philippe Broet pour avoir généreusement accepté de rapporter mon travail. Je suis également reconnaissant au professeur Christophe Ambroise pour avoir accepté d'être examinateur au sein de mon jury.

J'adresse tous mes remerciements au professeur Henri-Jean Garchon pour m'avoir permis de rejoindre ce laboratoire en tant qu'Ingénieur d'Etudes pendant un an, puis d'avoir cru en moi pour me diriger dans un projet de thèse aussi ambitieux que passionnant. Tout en ayant une certaine liberté et autonomie sur mon projet, j'ai été guidé exactement quand il le fallait, comme il est attendu pour ce type de projet. Cette collaboration était également pour moi très enrichissante, et j'ai pu apprendre à avoir toujours un recul important sur les résultats obtenus, et à ne jamais trop être optimiste tant que tout n'a pas été validé. C'est pour moi l'esprit même d'un chercheur, et notamment dans les domaines de la bioinformatique et de la biologie des systèmes.

Je remercie également le professeur Maxime Breban et le directeur Gilles Chiocchia, co-directeurs de l'équipe « Inflammatory Response and Immune System » (IRIS), pour m'avoir accueilli dans cette équipe.

J'apporte mes remerciements à l'Ecole Doctorale des Génomes Aux Organismes de l'UVSQ, ainsi qu'à l'ANR, pour avoir financé mes quatre années à travailler dans ce laboratoire.

Je souhaite de nouveau remercier Christophe Ambroise, statisticien au laboratoire LaMME à Evry, pour ses conseils avisés sur les approches de permutation et de bootstrap, sur l'utilisation du package SIMoNe pour l'inférence de nos réseaux gaussiens, ainsi que pour avoir participé à mon comité de thèse de mi-parcours.

Je remercie Sébastien Gaumer et Mathieu Sourdeval, qui m'ont encadré pendant mon monitorat à l'UVSQ et m'ont permis de vivre cette grande expérience que d'encadrer des TPs d'analyses expérimentales avec des étudiants de niveau L3.

Je remercie très humblement Félicie Costantino et Alice Talpin, qui m'ont permis de prendre la suite des analyses des puces de génotypage et transcriptomiques. Vous avez su m'apporter votre expertise acquise au cours de vos thèses, afin de mieux comprendre l'enjeu des analyses de ces données.

Je remercie toutes les personnes qui étaient présentes à l'Institut Cochin dans le même bâtiment que notre laboratoire, et avec qui j'ai pu beaucoup échanger. Merci à Nathalie pour avoir partagé ce bureau et m'avoir apporté ta joie de vivre pendant la fin de ta thèse. Merci à Barbara, pour ces

moments d'amitié et de voyage à Amsterdam, sans compter tout le groupe d'amis que j'ai pu connaître grâce aux personnes que tu m'as présentées. Merci à Maeva et Aurélie pour avoir amené une énergie débordante dans la vie du laboratoire. Merci également à la plate-forme PIPA, et notamment à Didier Fradelizi, qui en outre d'apporter une humeur débordante de joie de vivre autour de lui, m'a permis de développer un projet aussi intéressant que le développement du site web de l'association APEMM.

Je remercie toutes les personnes de l'équipe IRIS, qui ont contribué à la vie du laboratoire. Merci à Ingrid, qui a partagé mon bureau et dont le départ a créé un vide certain. Merci à Luiza, toujours fidèle à l'équipe et pour son humeur à la brésilienne. Merci à Isabelle pour avoir partagé mon bureau : je te souhaite bonne continuation dans tes projets. Merci aux doctorants, Quentin et Ketia, et également à Nadège, pour apporter un jeune dynamisme à ce laboratoire. Merci à Christophe et Clémence, ce duo de choc aussi efficace dans la vie du laboratoire que dans la mise au points des manips. Merci à Olivier pour gérer tout ce planning des réunions de laboratoire. Merci à Claudine, qui est encore plus fidèle au laboratoire !

Je remercie une partie des anciens de Sup'Biotech avec qui j'ai gardé contact et qui m'ont soutenu après l'obtention du diplôme : Yann, Maxime, Guillaume, Maximilien. Merci à l'administration de Sup'Biotech pour m'avoir soutenu dans la présidence de l'association des anciens. J'exprime également mes remerciements aux personnes plus ou moins extérieures à ma thèse et avec qui j'ai pu partager de très bons moments. Toutes ces rencontres très diverses m'ont permis de m'ouvrir encore plus l'esprit, et de me rendre compte de ce que la vie parisienne pouvait m'apporter.

Enfin, je te remercie, Priscillia, pour m'avoir soutenu et supporté pendant ces trois années de thèse. Tu as su me diriger et me donner des conseils très précieux dans la gestion de mon planning et des tâches, pour la finalisation de tous mes projets. En plus d'être une excellente partenaire de vie, tu me fournis la structure et la joie de vivre dont j'ai besoin. Je crois fort en toi et en tous tes projets.

Merci à tous !

TABLE DES MATIÈRES

LISTE DES FIGURES ET TABLEAUX

Figures

Tableaux

LISTE DES ABRÉVIATIONS

ACP : analyse en composantes principales

ADNc : ADN complémentaire

ADN : acide désoxyribonucléique

AIC : critère d'information d'Akaike

AINS : anti-inflammatoires non stéroïdiens

API : interface de programmation

AQCg/p : accuracy quality control for genes/pathways

ARMM : anti-rhumatismal modificateur de la maladie

ARN : acide ribonucléique

ASAS : assessment of spondyloarthritis international society

BASDAI : bath ankylosing spondylitis disease activity index

BASFI : bath ankylosing spondylitis functional index

BIC : critère d'information bayésienne

CMH : complexe majeur d'histocompatibilité

CNG : Centre National de Génotypage

CPA : cellule présentatrice d'antigène

CQCg/p : consistency quality control for genes/pathways

CRP : protéine C-réactive

DC : cellule dendritique

DE : différentiellement exprimé

DL : déséquilibre de liaison

eQTL : expression quantitative trait locus

EFS : Etablissement Français du Sang

EQC : external quality control

ESR : taux de sédimentation des érythrocytes

ESSG : european spondyloarthropathy study group

GEO : gene expression omnibus

GFEGS : groupe français d'étude génétique de la spondyloarthrite

GO : gene ontology

GRR : graphical relationship representation

GSEA : analyse d'enrichissement des jeux de gènes

GWAS : étude d'association sur génome entier

HLA : human leukocyte antigen

HLOD score : LOD score d'hétérogénéité

HW : Hardy-Weinberg

IBD : identité par descendance

IGAS : consortium international de la génétique de la spondylarthrite ankylosante

IQC : internal quality control

IRM : imagerie par résonance magnétique

KIR : killer-cell immunoglobulin-like receptor

LIMMA : linear models for microarray analysis

LOD : logarithm of odds

MAF : fréquence de l'allèle mineur

MD-DC : cellule dendritique dérivée de monocyte

MICI : maladie inflammatoire chronique de l'intestin

miRNA : micro-ARN

MLR : maximum de vraisemblance

MMP : métalloprotéinase

MNDA : antigène de différentiation nucléaire des myéloïdes

NASC : consortium nord-américain de la spondyloarthrite

NK : natural killer

NPL : analyse de liaison non paramétrique

OA : ostéoarthrite

ODE : équation différentielle ordinaire

PBMC : cellule mononucléée du sang périphérique

PPA : protéine de phase aigüe

PPI : interaction protéine-protéine

PR : polyarthrite rhumatoïde

quSAGE : quantitative set analysis for gene expression

RE : réticulum endoplasmique

RGS1 : régulateur 1 de la signalisation de la protéine G

RIC : rhumatisme inflammatoire chronique

RIN : score d'intégrité de l'ARN

RMA : robust multi-array average

SA : spondylarthrite ankylosante

SIMoNe : statistical inference for modular networks

SNP : polymorphisme sur nucléotide simple

SpA : spondyloarthrite

STRING : search tool for retrieval of interacting genes

TDT : test de déséquilibre de transmission

TNF : facteur de nécrose tumorale

UPR : unfolded protein response

USpA : spondyloarthrite non différenciée

VAR1 : vecteur auto-régressif de premier ordre

VIF : facteur d'inflation de la variance

WGCNA : weighted gene co-expression network analysis

β2m : β2-microglobuline humaine

Chapitre 1

INTRODUCTION GÉNÉRALE

Les maladies inflammatoires et multifactorielles

I/ Les maladies multifactorielles

Afin de développer des traitements contre une maladie chez l'Homme, il est primordial de comprendre comment celle-ci se développe. Il existe un nombre très important de types de pathologies qui sont classées en fonction de leurs origines. Nous pouvons citer par exemple les maladies infectieuses, métaboliques, du vieillissement, ou génétiques. Lorsqu'une seule source est à l'origine du développement d'une pathologie, il est plus facile avec les technologies actuelles de savoir ce qui peut empêcher de provoquer ou d'arrêter complètement les symptômes, et de partir sur une voie de guérison. Lorsque plusieurs facteurs peuvent influencer le développement d'une maladie, on dit qu'elle est multifactorielle. Ce groupe de maladies est en général plus complexe, avec des symptômes très diverses et un diagnostic difficile. De plus, il est particulièrement compliqué de déterminer ce qui est à l'origine du développement de la pathologie, et encore plus son traitement, tant le bruit d'information dû à des facteurs indirects est difficile à discriminer. Des maladies très connues et étudiées comme les cancers, l'obésité, le diabète de type 1 ou le spectre autistique en sont des exemples représentatifs.

II/ L'exemple des maladies inflammatoires

Des maladies inflammatoires font également partie de cette classe de pathologies, et se caractérisent par une réaction immunitaire de type inflammatoire trop importante qui entraîne un endommagement de tissus du cors humain. Ces pathologies s'accompagnent généralement de crises aiguës de douleurs et de situations de handicap importants chez les patients. Les origines d'une telle réaction inflammatoire sont particulièrement difficiles à déterminer, celles-ci pouvant être génétiques ou environnementales. Notamment, des études très importantes portent à croire que la flore bactérienne montrerait une signature spécifique dans son écosystème entre les patients et les personnes saines, bien qu'il est encore difficile de déterminer s'il s'agit de l'origine du développement pathologique ou une conséquence.

III/ Les rhumatismes inflammatoires chroniques

Parmi ces maladies inflammatoires, les rhumatismes inflammatoires chroniques se caractérisent par une atteinte principale au niveau des articulations et structures péri-articulaires, ce qui a pour conséquence d'impacter fortement la mobilité de la personne malade. Le RIC le plus fréquent est

la polyarthrite rhumatoïde (PR), qui se caractérise par une inflammation de la synoviale, suivie d'une hypertrophie de celle-ci avec multiplication des franges, un amincissement du cartilage et un épanchement de liquide synovial. Il s'en suit une poursuite de l'amincissement du cartilage et un développement d'ulcérations osseuses. La spondyloarthrite est ensuite le deuxième RIC le plus fréquent chez les adultes en France, avec une atteinte inflammatoire articulaire prédominante au niveau de la colonne vertébrale. Cette pathologie se divise en plusieurs sous-types, avec la spondylarthrite ankylosante comme forme principale. Il s'agit d'une maladie non mortelle dans l'immédiat, mais avec un très fort impact économique et social dû à l'immobilisation importante dans laquelle se trouve le patient et la difficulté à diminuer complètement celle-ci. L'étude de ce RIC est à la fois intéressant et un défi, notamment par sa complexité à définir les multiples facteurs à l'origine de son développement.

Qu'est-ce-que la spondylarthrite ankylosante et la SpA ?

La spondylarthrite ankylosante (SA) est un rhumatisme inflammatoire chronique (RIC) présentant une atteinte principale du squelette axial. Plus généralement, nous parlons de spondyloarthrite (SpA), qui inclut la SA et d'autres sous-types, pouvant eux-mêmes se développer chez la même personne et avec des caractéristiques communes.

I/ Nosologie et sémiologie

I.1) Les spondyloarthropathies séronégatives

Les spondyloarthropathies désignent un groupe de maladies articulaires qui peuvent se succéder à un degré divers chez une personne ou dans sa famille, et qui présentent des facteurs communs. Dans un sens plus large, ce terme inclut les atteintes articulaires de la colonne vertébrale à partir de n'importe quelle maladie articulaire, y compris certaines formes de PR et d'ostéoarthrite (OA). Nous les distinguons donc des spondyloarthropathies séronégatives, un groupe de maladies impliquant le squelette axial et possédant un statut sérique négatif, c'est-à-dire qui ne présentent pas le facteur rhumatoïde.[1] Ce dernier est l'auto-anticorps qui a été découvert en premier dans la PR, défini comme un anticorps dirigé contre la portion Fc des IgG. La SpA est une spondyloarthropathie séronégative accompagnée d'une inflammation.

I.2) Atteinte principale du squelette axial et du sacro-iliaque

La SpA est la seconde cause de maladies inflammatoires articulaires chez l'adulte après la PR. La SA est la forme la plus classique de cette maladie, avec une atteinte prédominante de la colonne vertébrale (ou squelette axial) et une sacro-iliite radiologique avancée (Figure 1). Des caractéristiques communes, en plus de l'absence de facteur rhumatoïde, ont regroupé d'autres pathologies dans la SpA. Ces caractéristiques sont entre autres une liaison avec l'allèle HLA-B27 (gène du complexe majeur d'histocompatibilité ou CMH), une arthrite inflammatoire de l'axe (généralement sacro-iliite ou spondylite), des oligoarthrites avec une présentation généralement asymétrique, et une agrégation familiale significative. Les oligoarthrites représentent une atteinte d'une à quatre articulations pendant les six premiers mois de la maladie.[2]

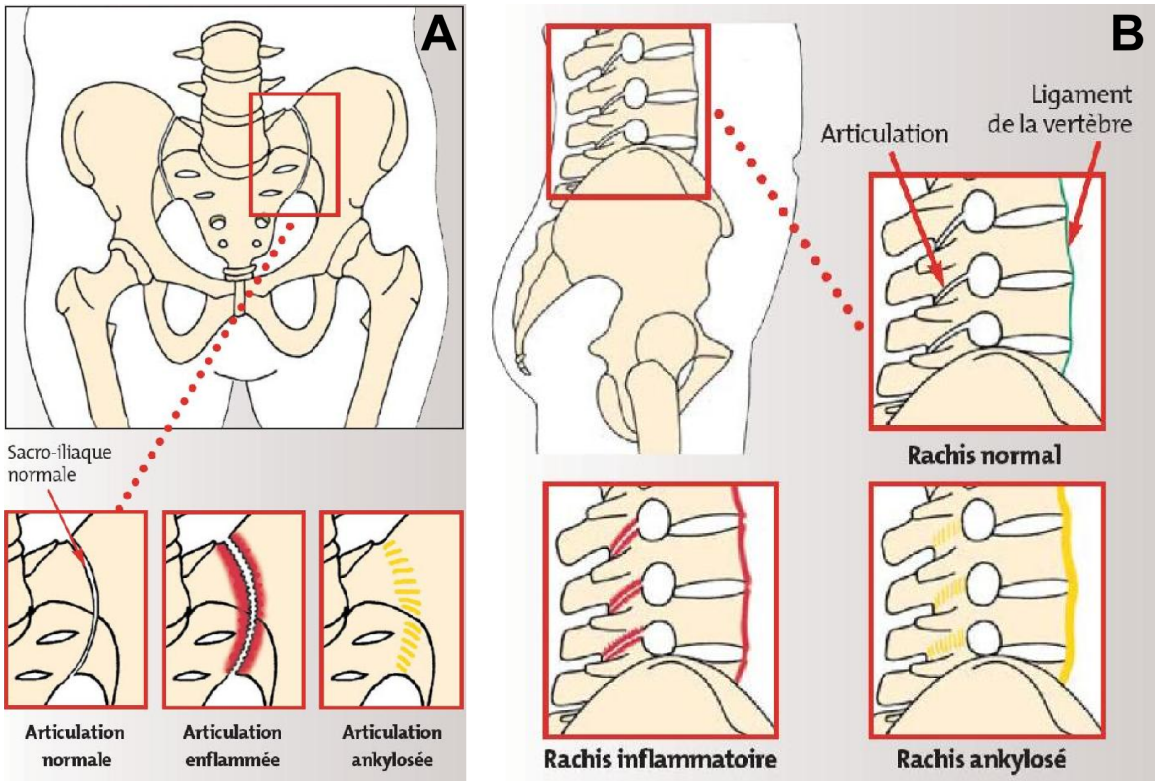


Figure 1: Représentations des atteintes de la sacro-iliaque (A) et du rachis (B) dans la SpA

D'après « La Spondylarthrite Ankylosante en 100 questions », édition 2005, Dougados M. et al.[3]

I.3) Atteintes périphériques

En plus de la manifestation axiale typique, la SpA présente souvent des symptômes musculo-squelettiques, sous forme d'arthrite, d'enthésite (notamment au niveau du tendon d'Achille) ou de dactylite (inflammation du doigt). L'enthésite est une inflammation de l'enthèse, c'est-à-dire le tissu connectif entre le tendon ou le ligament et l'os (Figure 2). Les symptômes extra-articulaires de la SpA comprennent une uvéite antérieure aiguë, un psoriasis (inflammation de la peau), et des maladies inflammatoires chroniques de l'intestin (MICI). Ces dernières sont principalement similaires à la maladie de Crohn et la colite ulcéreuse. L'ankylose, c'est-à-dire une fixation et une immobilité d'une articulation, était une manifestation clinique courante de la SA avant qu'elle ne soit mieux traitée comme c'est le cas actuellement. Elle se manifeste généralement au niveau du rachis, mais parfois également aux hanches ou à la cage thoracique, entraînant dans ce dernier cas une diminution de l'amplitude respiratoire. L'ankylose de la SA est un processus douloureux pour le patient et entraîne un handicap important et non réversible avec une position très inconfortable.[2]

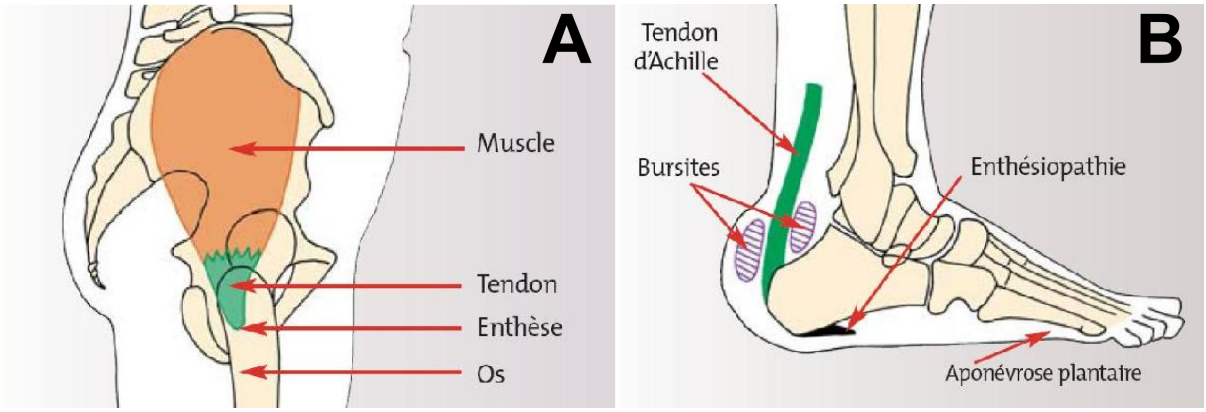


Figure 2: Représentation d'une enthèse (A) et de son atteinte au talon (B) lors de la SpA

D'après « La Spondylarthrite Ankylosante en 100 questions », édition 2005, Dougados M. et al.[3]

I.4) Les sous-types de la SpA, un groupement justifié

Des motifs d'implication spécifiques reconnus ont entraîné l'identification de plusieurs sous-types de la SpA en plus de la SA, incluant l'arthrite psoriasique, l'arthrite associée avec une MICI, et l'arthrite réactionnelle en réponse à une infection. Cependant, une personne atteinte de SA peut développer par la suite un ou plusieurs sous-types précédemment décrits, supportant l'idée de facteurs de susceptibilités génétiques communs. De plus, plusieurs membres d'une même famille peuvent présenter plusieurs sous-types différents de la SpA. Lorsqu'une forme de SpA ne peut être classée, on parlera de spondyloarhtrithe non différenciée (USpA). Bien qu'il y ait une prévalence plus forte de la SpA chez les personnes adultes, elle peut aussi apparaître chez l'enfant sous forme de spondyloarthrite juvénile.[4] Les différents sous-types de SpA sont résumés dans la Figure 3. De nombreuses études sur la SpA résument cette pathologie par sa forme la plus fréquente, la SA. De nombreux efforts ont été effectués pour bien définir les différents sous-types associés, et les termes peuvent encore évoluer à force que la connaissance de la maladie gagne du terrain.

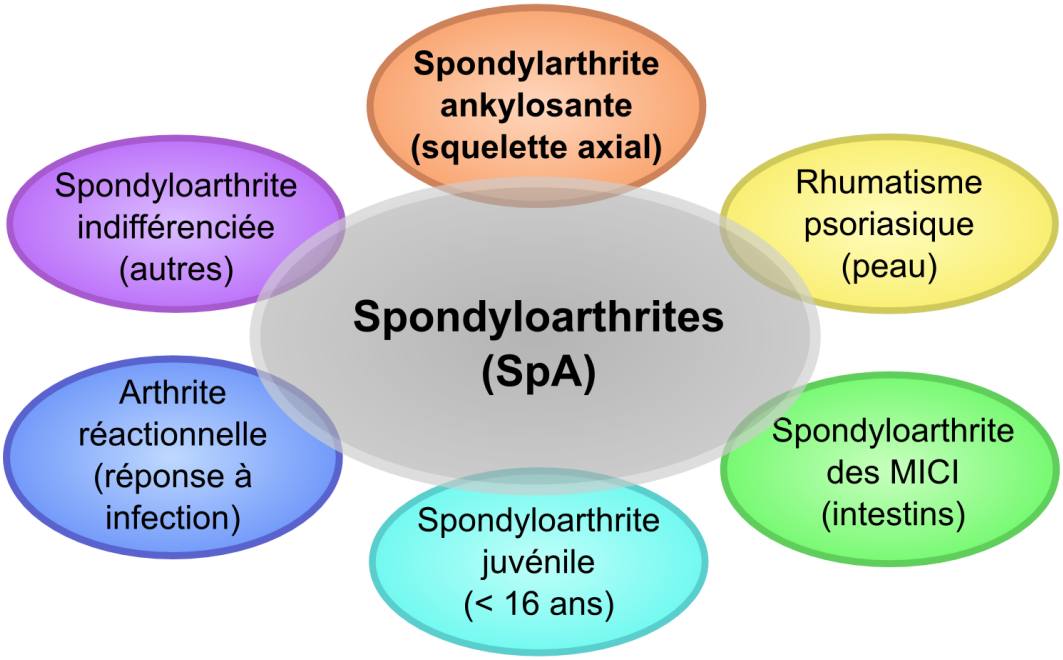


Figure 3: Les différents sous-types de SpA

II/ Epidémiologie, diagnostic et traitements actuels

II.1) Les critères de classification de la SpA

L'épidémiologie représente l'étude de la distribution et des facteurs déterminants des maladies dans les populations humaines. Pour cela, au vu des différents sous-types de la SpA, il est important de bien les classer. C'est pour cela que les méthodes de choix des critères doivent être bien définis, qui se présentent en partie pour certains sous forme de questionnaire. Nous pouvons citer par exemple les critères d'Amor ou European Spondyloarthropathy Study Group (ESSG). Chaque protocole a ses avantages, et étudie la question sous un angle légèrement différent, ce qui peut les rendre complémentaire et explique pourquoi plusieurs méthodes sont parfois utilisées en même temps pour une même cohorte. Les critères ESSG, proposés en 1991, inclusent les formes non différenciées, ce qui n'est pas le cas pour les critères d'Amor. Ces derniers, proposés un an plus tôt, représentent un système de score attribué au patient en fonction de signes basés sur des caractéristiques radiologiques, cliniques et de laboratoire, sans critère d'entrée. Les critères Assessment of SpondyloArthritis international Society (ASAS), proposé en 2009, ont pour avantage majeur d'inclure des techniques d'imagerie par résonance magnétique (IRM) afin de

détecter une sacro-iliite.[4] Les trois critères cités ont pour point commun de se baser sur les symptômes axiaux, périphériques et extra-articulaires, et sur la détection d'une sacro-iliite. De plus, ce sont actuellement les critères les plus utilisés pour des études cliniques. Ils présentent l'avantage d'inclure plusieurs formes de SpA, et non la SA de façon isolée. Les critères d'Amor et ESSG utilisent la présence de l'allèle HLA-B27. Les critères les plus anciens proposés sont ceux de Rome pour la SA en 1963, qui étaient principalement ciblés pour les études de populations. Les critères modifiés de New York (NY), proposés en 1984, sont également utilisés comme moyen de diagnostic, et se basent sur les symptômes axiaux, les limitations dans la mobilité et une sacro-iliite radiographique.[5] Les principaux critères de classification sont résumés dans le Tableau 1.

II.2) Epidémiologie dans le Monde

La SpA est une maladie qui apparaît de façon graduelle, plus couramment entre 15 et 45 ans, avec une atteinte équivalente entre les hommes et les femmes. Cependant, des formes plus sévères sont en général observées chez les hommes. La SA est fortement associée à l'allèle HLA-B27, puisque ce dernier est présent dans environ 80 % des patients. Cependant, la distribution de cet allèle varie beaucoup entre chaque population. Cela explique certainement pourquoi les taux de prévalence et d'incidence de la maladie varient énormément en fonction de la zone géographique. Par exemple, cette pathologie semble beaucoup plus rare en Afrique noire, dû à la faible prévalence de l'allèle HLA-B27 dans la population liée (moins de 5%). La prévalence s'appuie sur le nombre total de patients sur une période donnée, tandis que l'incidence ne tient compte que des nouveaux cas. Globalement, le taux de prévalence de la SpA dans les pays d'Europe de l'Ouest oscille entre 0.3 et 2.5 %. Lorsque les sous-types de SpA sont pris en compte, les formes indifférenciées semblent constituer environ 40 % de ces cas. Les taux de prévalence de la SA et de l'arthrite psoriasique semblent similaires entre les pays occidentaux, pouvant atteindre 0.53 %.[5] Une étude publiée en 2015 a estimé que la prévalence de la SpA dans la population française est de 0.43 %. L'allèle HLA-B27 est estimé comme étant porté par 75 % des patients contre 6.9 % des témoins, ce qui attribue à cet allèle un risque de développement de la maladie augmenté de 39 fois par rapport aux témoins HLA-B27 négatifs.[6]

Critères de classification, année	Cibles	Méthode de développement	Critères d'entrée	Critères cliniques	Critères de laboratoire ou radiographiques
Critères de Rome pour la SA, 1963[7]	Etudes de populations	Expériences cliniques des rhumatologues	Aucun	• Symptômes axiaux • Limitations de la mobilité • Atteintes extra-articulaires (uvéite)	• Sacro-iliite (radiographique)
Critères de NY pour la SA, 1966[8]	Etudes de populations	Expériences cliniques et résultats d'une application du critère de Rome	Sacro-iliite radiographique	• Symptômes axiaux • Limitations de la mobilité	• Sacro-iliite (radiographique)
Critères modifiés de NY pour la SA, 1984[9]	Etudes de diagnostique et cliniques	Modification du critère de NY par des experts basée sur les résultats d'une étude des deux critères précédents	Sacro-iliite radiographique	• Symptômes axiaux • Limitations de la mobilité	• Sacro-iliite (radiographique)
Critères d'Amor pour la SpA, 1990[10]	Etudes cliniques	Test rétrospectif et prospectif des composantes candidates (sélectionnées par des experts)	Aucun	• Symptômes axiaux • Symptômes périphériques • Atteintes extra-articulaires	• Sacro-iliite (radiographique) • HLA-B27
Critères ESSG pour la SpA, 1991[11]	Etudes cliniques	Tests statistiques des variables candidats chez les patients SpA (opinion d'experts) et de témoins, en plus de raisonnements cliniques	Mal de dos inflammatoire ou synovite	• Symptômes axiaux • Symptômes périphériques • Atteintes extra-articulaires	• Sacro-iliite (radiographique)
Critères ASAS SpA axial, 2009[12]	Etudes cliniques	Critères candidats (par des experts basés sur le raisonnement clinique) testés prospectivement (option d'experts) ; critères finals basés sur l'analyse statistique et un vote des experts pour les critères candidats	Plus de 3 mois de mal de dos et âge de suivi inférieur à 45 ans	• Symptômes axiaux • Symptômes périphériques • Manifestations extra-articulaires	• Sacro-iliite (radiographique ou IRM) • HLA-B27 • CRP
Critères ASAS SpA périphérique, 2011[13]	Etudes cliniques	Idem que ASAS SpA axial	Arthrite, enthèsite ou dactylite	Idem que ASAS SpA axial	• Sacro-iliite (radiographique ou IRM) • HLA-B27

Tableau 1: Liste des principaux critères de classification de la SpA.

D'après « Epidemiology of Spondyloarthritis », Stolwijk et al.[5]

II.3) Les moyens de diagnostic

Les méthodes de critères pour classer les SpA peuvent également être utilisées comme moyens de diagnostic. Bien qu'il n'y ait pas de diagnostic direct pour la SpA, d'autres tests non spécifiques peuvent être utiles pour détecter si le patient développe ce rhumatisme inflammatoire. Le test de Schober consiste à mesurer la capacité du patient à plier son dos, et représente déjà une mesure clinique utile pour déterminer la flexibilité du rachis lombaire. Le rhumatologue peut également déterminer les caractéristiques radiographiques, comme au niveau des articulations sacro-iliaques ou du rachis, grâce à des technologies de rayons X ou IRM. Des tests sanguins peuvent être effectués, afin de détecter une augmentation du taux protéines C-réactives (CRPs) et du taux de sédimentation des erythrocytes (ESR), mais ce ne sont pas des paramètres observables chez tous les patients atteints de SpA. Les CRPs font partie des protéines de phase aigüe (PPAs), tout comme des cytokines telles que l'IL-6 ou le TNF-α. Il s'agit de très bons biomarqueurs pour mesurer l'activité inflammatoire. Un test génétique permettant de tester l'allèle HLA-B27 peut être utilisé, mais la fréquence de ce variant et le risque attribué varie en fonction des zones géographiques. Le Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) a pour but de détecter la charge inflammatoire d'une maladie active. Cette méthode peut aider à établir un diagnostic de la SpA en présence d'autres facteurs comme l'allèle HLA-B27, la fessalgie, ou l'évidence d'une atteinte aux articulations sacro-iliaques par imagerie. Il s'agit d'un calcul de score compris entre 1 et 10 et qui mesure l'inconfort du patient. Il se base entre autres sur l'épuisement, la douleur rachidienne, l'arthralgie (douleur articulaire), l'atteinte des enthèses et la raideur des articulations.[2] Le BASFI (pour Functional Index) est une mesure précise de l'infirmité fonctionnelle induite par la maladie. Elle est plus utilisée pour étudier une réponse à un traitement que comme moyen de diagnostic.

II.4) Les traitements médicamentaux

Il n'y a actuellement aucun traitement spécifique permettant la guérison de la SpA. L'un des obstacles majeurs à un traitement efficace du patient est le diagnostic tardif, dû à la difficulté pour le patient de détecter qu'il est bien atteint d'une SpA. Cependant, beaucoup de recherches ont été effectuées au cours des dernières années pour développer des traitements et médicaments pouvant réduire les symptômes et la douleur. Notamment, l'ankylose se développe maintenant très peu si la

maladie est bien prise en charge chez le patient. Les médicaments utilisés en majorité ont pour fonction de diminuer la douleur ou de ralentir la progression de la maladie. Les médicaments anti-inflammatoires sont utilisés dans ce but, qui incluent les anti-inflammatoires non stéroïdiens (AINS) comme l'ibuprofène, le phénylbutazone, le diclofénac, l'indométhacine, le naproxène et d'autres inhibiteurs de COX-2. Il a été montré en 2012 que ce type de traitement est très bénéfique pour les patients atteints de SA avec un taux élevé de PPAs.[14] Des opioïdes antalgiques peuvent également être utilisés pour soulager la douleur, comme la morphine, le tramadol ou la mépéridine.[15] Des anti-rhumatismaux modificateurs de la maladie (ARMM) peuvent aussi être utilisés comme la sulfasalazine pour les patients atteints d'arthrite périphérique. Des bloqueurs anti-TNF-α, comme l'infliximab, le golimumab ou l'adalimumab, sont pour la plupart des antagonistes de TNF-α et sont souvent utilisés pour la PR. Cependant, il a été démontré qu'ils pouvaient également améliorer les symptômes cliniques dans le traitement de la SA.[16] Des inhibiteurs anti-IL-6 comme la tocilizumab et le rituximab (anticorps monoclonal anti-CD20), bien qu'ils soient efficaces contre la PR, se sont révélés décevants contre la SpA.[17] Une phase clinique publiée en 2013 teste l'efficacité du secukinumab, un anticorps monoclonal contre l'IL-17A, et donne des résultats prometteurs dans le traitement de la SA.[18]

II.5) Les traitements chirurgicaux et physiques

Dans les cas sévères de SpA, les médecins ont parfois recours à la chirurgie pour remplacer les articulations, notamment dans les genoux et les hanches. Bien que l'opération soit très risquée, il est possible d'effectuer une correction chirurgicale pour les patients atteints de déformations graves de flexion du rachis, notamment en cas de courbure sévère vers le bas. La rigidité des nervures thoraciques, l’insuffisance aortique de certains patients et la calcification des ligaments rendent les anesthésies particulièrement difficiles à mettre en place. Enfin, des exercices thérapeutiques peuvent être utilisés, comme c'est le cas pour beaucoup d'arthrites. Elles peuvent soulager les douleurs au niveau des lombes, du cou, du genou et de l'épaule. Celles-ci incluent des exercices aérobiques de faible intensité, de la neurostimulation électrique transcutanée, de la thermothérapie, de la facilitation proprioceptive neuromusculaire, et des exercices supervisés ou à faire chez soi.[19] Cependant, des exercices modérés ou à hauts efforts comme le jogging sont proscrits car ils entraînent trop d'endommagements au niveau des articulations.

III/ La SpA, une maladie multifactorielle

III.1) Implication d'autres facteurs suspectée

Outre la forte association de l'allèle HLA-B27 à la SpA, une forte ségrégation familiale de la maladie est observée chez les patients. Il a été démontré de nombreuses fois que la susceptibilité à la SpA a une forte composante génétique. Les facteurs génétiques actuellement connus comme étant associés à la SpA sont décrits un peu plus loin dans le manuscrit (Études d'association sur la SpA). Cependant, bien que l'allèle HLA-B27 soit présent chez une majorité de patients, seulement 4 % des sujets HLA-B27 positifs développent la SpA. Il est démontré que cet allèle ne comprend qu'une minorité de la susceptibilité génétique à cette pathologie.[20] De plus, les locus identifiés par des études d'association sur génome entier en dehors du complexe majeur d'histocompatibilité (CMH) n'ont que peu d'effet sur le risque de développer la SpA (Figure 4). Il est donc évident que d'autres facteurs sont à l'origine du développement de la SpA et doivent encore être expliqués.[21]

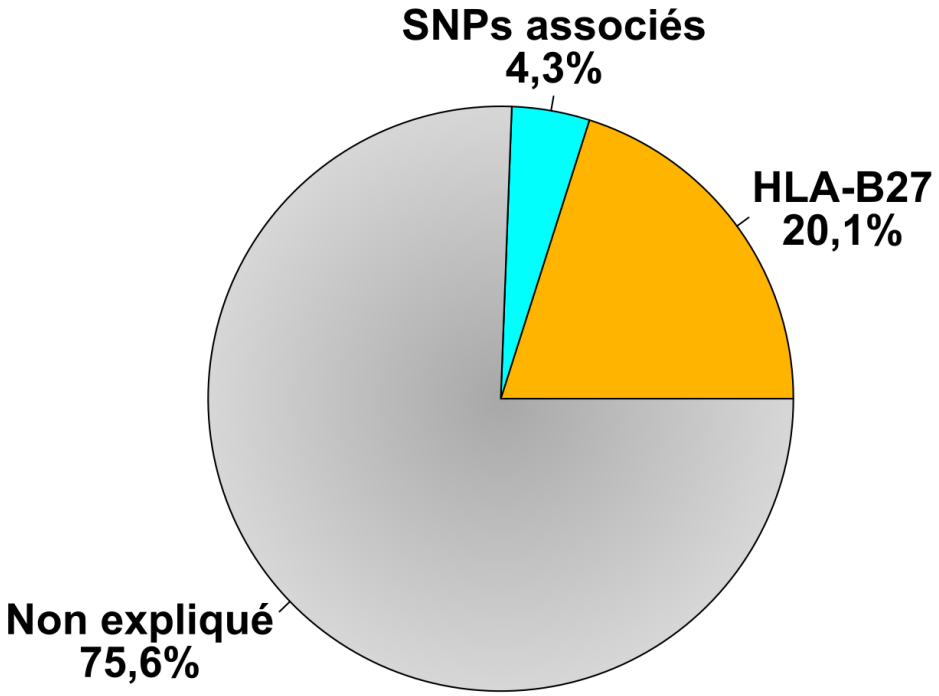


Figure 4: Part de l'héritabilité expliquée actuellement par les études d'associations de la SpA

D'après « Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci », IGAS et al.[22]

III.2) Le microbiote, principal candidat comme facteur environnemental

Parmi les facteurs environnementaux, un intérêt s'est largement développé pour les changements de l'éco-système du microbiote pouvant être associés aux maladies inflammatoires. Le microbiote est l'ensemble des micro-organismes vivant dans un environnement spécifique (le microbiome), incluant les bactéries, levures, champignons et virus. Le microbiote intestinal a été démontré comme contrôlant ou régulant chez la souris la masse osseuse, le stockage des graisses corporelles, mais aussi le bon développement de la réponse immunitaire.[23] De plus, il a été démontré une association entre la barrière bactérienne intestinale et les MICI, notamment la maladie de Crohn, pour laquelle un traitement anti-TNF est capable de restaurer un microbiote sain.[24] Par ailleurs, la souche *clostridium difficile* peut provoquer une colite.[25] En sachant que les patients atteints de SpA développent beaucoup plus une MICI que dans la population générale, il y a une forte supposition que des mécanismes liés à l'inflammation et des facteurs prédisposant peuvent être communs entre ces deux pathologies. Il a également été démontré que des rats transgéniques pour HLA-B27, un très bon modèle animal de la SpA, ne développent pas la maladie quand ils sont maintenus en absence de micro-organismes.[26] Un autre lien direct entre la SpA et le microbiote sont des formes d'arthrites réactionnelles qui sont déclenchées après des infections bactériennes de l'appareil intestinal et urinaire.[27] Concernant les autres facteurs environnementaux, une étude publiée en 2013 a démontré pour la première fois un lien entre les événements stressants et l'activité de la maladie.[28]

Les facteurs génétiques et fonctionnels liés à la SpA

Ce projet de thèse se base sur l'analyse de données issues de deux technologies pour mieux étudier l'étiologie de la SA : les puces de génotypage et les puces transcriptomiques. Nous verrons ici les approches similaires majeures déjà effectuées à ce jour afin d'identifier les deux différents types de facteurs associés à la maladie, dont certaines incluent des études menées par notre laboratoire.

I/ Associations génétiques

I.1) Les puces de génotypage pour analyser les SNPs

Des études d'association sur génome entier (GWAS) ont été effectuées sur de nombreuses cohortes afin d'explorer d'autres facteurs génétiques que l'allèle HLA-B27 manquant à l'explication de l'héritabilité de la SpA. Ce type d'étude se fait principalement grâce à des techniques utilisant des puces de génotypage, ou puces à ADN, sur génome entier. Ces dernières vont permettre de tester l'association de très nombreux polymorphismes sur nucléotide simple (SNPs), pouvant dépasser le million, et répartis sur tout le génome. Il s'agit de l'approche inverse de celle que l'on appelle gène-candidat, cette dernière consistant à tester des gènes ou locus sélectionnés a priori. Un SNP dénomme un marqueur génétique qui témoigne d'une mutation ancestrale et conservée dans les populations, et qui permet par association de déterminer quels peuvent être les facteurs génétiques à l'origine du développement d'un phénotype observé. Pour toutes les analyses incluant des SNPs dont nous parlerons, nous nous restreindrons sur les polymorphismes bi-alléliques, c'est-à-dire ceux incluant seulement un allèle majeur et un allèle mineur. Le type d'approche sur génome entier permet de découvrir d'autres facteurs auparavant non connus ou soupçonnés, ce qui est très avantageux pour mieux expliquer le mécanisme déclenchant les maladies multifactorielles. Des efforts considérables ont été effectués pour rassembler tous les SNPs connus dans une base de données appelée dbSNP, qui est mise à jour régulièrement et se connecte sur d'autres bases de données issues de consortiums, comme la très reconnue HapMap.[29] dbSNP intègre les polymorphismes de plusieurs espèces et des informations plus complètes telles que la cartographie physique, la génétique des populations ou l'investigation des relations évolutives. Par ailleurs, le consortium Immunochip a développé une puce à ADN dont le choix des polymorphismes détectés a été choisi dans le but précis de répertorier les marqueurs génétiques détectés lors de GWAS ou de séquençages en profondeur de plusieurs

maladies autoimmunes et inflammatoires.[30] Bien que les puces de génotypage représentent une technique avec de nombreux avantages qui s'est démocratisée au fil des dernières années, elles commencent à être remplacées peu à peu par des technologies de séquençage à haut débit (NGS), qui permettent notamment d'avoir des séquences complètes au lieu de la mesure d'un seul marqueur, et donc de découvrir des nouveaux SNPs pendant l'analyse avec un meilleur recouvrement.[31]

I.1.A) Précautions sur les analyses

Il appartient à ceux qui analysent ce type de données d'être prudents sur les résultats statistiques. En effet, le nombre important de marqueurs testés simultanément impose d'utiliser une correction sur tests multiples. Par exemple, le seuil de p-value accepté pour les GWAS est de 5×10^{-8}, ce qui implique d'avoir une puissance statistique importante et donc un effort substantiel sur le recrutement des patients dans une cohorte.[32] Il est de plus préférable d'analyser des variants communs, même si nous verrons par la suite que l'analyse de liaison familiale permet de détecter de façon plus fine des SNPs dont la fréquence de l'allèle mineure (MAF) est basse. Par ailleurs, il est nécessaire de prendre en compte le déséquilibre de liaison (DL) lorsqu'on veut définir les SNPs causaux d'un phénotype. Le DL est la conséquence des recombinaisons homologues qui ont lieu pendant la méiose et qui sont en grande partie à l'origine de la diversité génétique que nous observons entre chaque individu. Il dépendra également de l'ancienneté de la mutation associée dans l'évolution. Ce phénomène s'observe lorsque la probabilité d'observer un couple d'allèles sur un locus n'est pas le produit des probabilités des allèles isolés.[33] Il est possible d'éviter le biais par le DL lors de découvertes de SNPs d'intérêt, en estimant des blocs de marqueurs dont le coefficient de corrélation r2 est très proche. Nous pouvons ainsi restreindre le nombre de SNPs à analyser et économiser considérablement les besoins en ressources de calculs, en estimant qu'un seul SNP servira à prévoir les allèles présents sur les autres marqueurs du bloc de DL.

I.1.B) Contrôle qualité des échantillons

L'analyse des SNPs nécessite un contrôle qualité conséquent afin d'éviter au maximum des biais sur les résultats. La première vérification qui peut être faite, lorsqu'on détient les données de génotypage du chromosome X, est de vérifier le sexe des individus. En effet, on peut facilement détecter les gènes hétérozygotes qui ne devraient pas exister dans les régions non pseudo-

autosomales chez les hommes. Cette région définit la partie tronquée dans le chromosome Y par rapport au X. Ensuite, il est également possible de vérifier les liens de parenté sans nettoyage au préalable des SNPs.[34] Cela permet de détecter des erreurs de prélèvement, ou tout simplement des liens de parentés entrés dans les informations de la cohorte qui s'avéraient être faux. Il s'agit pour l'instant de nettoyages d'erreurs qui se trouvent au niveau des échantillons, mais nous verrons qu'il est également impératif de vérifier des erreurs au niveau de chaque SNP.

I.1.C) Contrôle qualité des SNPs

L'erreur mendélienne est celle qui est la plus évidente au niveau du SNP lors d'études familiales, puisqu'elle se traduit par l'apparition d'un allèle chez le descendant sans qu'aucun parent ne soit porteur de celui-ci. Compte-tenu de la rareté extrême d'une mutation pouvant apparaître spontanément au niveau d'un SNP d'une génération à l'autre, il convient de retirer ce marqueur pour les trois individus testés. Etant donné que les puces de génotypage donnent les allèles observés sur chaque chromosome homologue sans savoir de quel parent vient celui-ci, il est parfois difficile de détecter des erreurs mendéliennes dans le cas où le descendant porte les mêmes allèles que les parents, mais sur des chromosomes homologues discordants. Il existe pour cela des outils permettant de reconstruire les haplotypes avec déduction et de détecter d'une manière plus fine ces erreurs.[35] A cause du fait que la reconstruction des haplotypes dépend fortement des SNPs analysés, il convient de renouveler cette détection d'erreurs plusieurs fois, ce qui peut être très coûteux en calculs pour les données sur chromosome entier. En revanche, l'avantage est que ce type de nettoyage peut être effectué sur des chromosomes isolés, étant donné que la reconstruction des haplotypes ne souffre d'aucune dépendance inter-chromosome, et donc de lancer ce type de nettoyage en parallèle sur des ordinateurs suffisamment puissants.

Il convient également de retirer les déviations trop importantes de l'équilibre de Hardy-Weinberg (HW), qui peuvent témoigner d'une erreur de génotypage ou d'une stratification de la population.[36,37] Cette dernière est un biais de l'association des SNPs liés à l'origine ethnique plutôt qu'à la présence ou non de la maladie. Cette mesure se base sur la loi de HW, où les fréquences alléliques et génotypiques dans une population de très grande taille doivent être équilibrées de façon conservée d'une génération à l'autre. Les fréquence génotypiques doivent donc être déduites des fréquences alléliques.[38] La fréquence génotypique représente les différentes combinaisons

possibles de plusieurs allèles sur des chromosomes homologues d'un même individu. Cependant, un fort déséquilibre de HW est parfois la conséquence d'une réelle association avec la maladie, c'est pourquoi il est conseillé d'effectuer ce nettoyage en ne se basant que sur les sujets sains.

Une fois ces nettoyages sur les SNPs effectués, il est recommandé de vérifier le taux de génotypage de chaque marqueur, c'est-à-dire le ratio du nombre d'individus avec des données sur les allèles de ce marqueur, comparé au nombre d'individus total. En général, le seuil toléré pour le taux de génotypage est au minimum 95 ou 99 %, en fonction de la taille de la cohorte et du type d'analyse effectuée.[39] Une fois toutes ces erreurs de génotypage nettoyées, il est possible de lancer des analyses permettant d'identifier les marqueurs causaux d'une maladie. Nous verrons qu'il en existe principalement deux catégories, l'analyse d'association et l'analyse de liaison.

I.1.D) Les analyses d'association

Les analyses d'association sont celles qui sont les plus utilisées sur génome entier, et se basent sur le principe de DL, puisque celui-ci peut également être la conséquence d'un marqueur spécifiquement déterminant pour la maladie étudiée. Deux types d'approches sont principalement utilisées : les études cas-témoins et les études intra-familiales. L'étude cas-témoins, qui est la plus courante, consiste à déterminer les SNPs pour lesquels la fréquence allélique est plus élevée ou plus basse entre les patients et les témoins. Il s'agit donc d'une méthode très proche du principe des odds ratio, ou rapports de chance. Le test du χ^2 ou le test exact de Fisher peuvent être utilisés pour vérifier la significativité de l'association de chaque marqueur séparément. Afin de faire la différence entre les SNPs réellement causaux et ceux qui ont une association indirecte, il est primordial de prendre en compte le DL pour tenter d'enlever ce bruit d'information.[40] Les études d'associations intra-familiales sont principalement effectuées sur des trios, c'est-à-dire des structures familiales constituées de deux parents et d'un descendant. On peut ainsi effectuer un test de déséquilibre de transmission (TDT), qui consiste à comparer la fréquence de transmission allélique observée et celle attendue selon les lois de Mendel. Si les fréquences sont significativement différentes (déterminé par un test du χ^2) pour un SNP donné, l'hypothèse nulle est rejetée et nous pouvons donc conclure que le polymorphisme est à la fois lié et associé à la maladie.[41]

Les analyses d'association présentent néanmoins comme inconvénient de se limiter aux SNPs

communs. En effet, les variants rares ne seront pas détectés à cause de leur faible probabilité de les trouver dans la population générale. Or, les SNPs avec une faible MAF peuvent être très intéressants pour détecter les facteurs génétiques associés à une maladie.

I.1.E) Les analyses de liaison non paramétriques

Les analyses de liaison familiales permettent d'étudier aussi les SNPs rares, car elles se basent sur le principe des recombinaisons homologues et des haplotypes transmis au sein de plusieurs générations pour identifier les variants qui seraient la cause du développement de la maladie. Cette méthode permet donc, par une approche de co-ségrégation de la maladie avec un haplotype particulier, de définir une région liée au phénotype observé. On utilise ici la notion de partage d'allèles identiques par descendance (IBD) chez les personnes malades. D'abord basées sur des analyses de paires de germains, il est maintenant possible d'effectuer des analyses de liaison sur des structures de familles très larges, ce qui peut s'avérer très informatif pour les maladies complexes. La région liée est définie en comparant les haplotypes qui sont co-transmis en commun dans chaque famille avec la maladie. Pour chacune de ces transmissions, des recombinaisons homologues peuvent se produire sur les haplotypes, qui permettront de restreindre la région à étudier. Ces recombinaisons sont dites informatives, et elles sont cruciales pour les analyses de liaison. Ensuite, cette région peut de nouveau être séquencée ou annotée pour déterminer quel est le facteur génétique de susceptibilité de façon plus précise.

La mesure de la liaison d'un marqueur se fait grâce au logarithm of odds (LOD) score. Son calcul se fait en plusieurs étapes successives :[42,43]

A. **Fraction de recombinaison** (entre deux locus) $\boldsymbol{\theta} = \dfrac{\text{Nombre de gamètes recombinants}}{\text{Nombre total de méioses}}$

B. **Rapport de vraisemblance** $= \dfrac{\text{Vraisemblance des résultats sous l'hypothèse de liaison } (\theta < 1/2)}{\text{Vraisemblance des résultats sous l'hypothèse d'indépendance } (\theta = 1/2)}$

C. **Maximum de vraissemblance** = MLR = $\max_{0<\theta<1/2}$ (Rapport de vraissemblance)

D. **Logarithm of odds** = LOD = log(MLR)

Le logarithme décimal du MLR permet de manipuler plus facilement les résultats, puisqu'il suffit d'additionner les scores respectifs de chaque famille pour obtenir un score global La méthode du

LOD score pour une analyse de liaison dans le cas d'une maladie utilise la fraction θ entre un marqueur et le locus morbide, c'est-à-dire qu'on effectue une cartographie afin de trouver le marqueur qui a le plus de chances d'être proche du locus responsable de la pathologie. Lorsque le LOD score d'un marqueur est supérieur ou égal à 3, nous pouvons dire qu'il y a une liaison entre le marqueur testé et le locus morbide. Un score inférieur à -2 témoigne d'une absence de liaison. Entre ces deux seuils, il est impossible de conclure.[44] Quoi qu'il en soit, l'analyse de liaison nécessite de connaître les haplotypes, c'est-à-dire l'enchaînement des allèles sur un même chromosome homologue, et donc de les reconstruire dans le cas de l'analyse des puces de génotypage. Nous différencions les analyses de liaison bi-point, qui se contentent de calculer les fractions θ pour chaque marqueur de façon isolée, et les analyses multi-point qui sont plus puissants et permettent de localiser le locus morbide par rapport à une carte fixe de marqueurs génétiques dont la position est connue.

I.1.F) Les analyses de liaison paramétriques

Pour étudier des maladies complexes et multifactorielles, des analyses de liaison paramétriques sont couramment utilisées. Cette approche se base sur un modèle de transmission, qui nous permettra de tester l'hypothèse nulle en tenant compte de ce modèle. Par exemple, des modèles simples comme une transmission par allèle dominant ou récessif permettront de détecter par une analyse paramétrique les variants qui sont les plus liés en nous basant sur cette hypothèse. Il est possible d'utiliser des modèles plus complexes, comme la codominance avec un trait connu. Dans le cas de la SpA, nous pouvons prendre par exemple le modèle de codominance avec HLA-B27, puisque nous savons que cet allèle est fortement associé à la maladie. Il est recommandé, dans le cas de maladies multifactorielles, de tester en premier les deux modèles simples donnés en exemple (dominant et récessif), et de se baser sur le modèle avec le LOD score le plus élevé. Ensuite, un seuil de significativité corrigé de 3,3 est conseillé, compte tenu de l'augmentation de la probabilité de trouver des faux positifs dans le cas d'une analyse paramétrique.[45,46] Lorsqu'on étudie une maladie multifactorielle, il est important d'intégrer et de prendre en compte l'hétérogénéité de locus, par la mesure du LOD score d'hétérogénéité (HLOD score).[47] Cette hétérogénéité correspond à plusieurs locus indépendants pouvant être à l'origine du même phénotype, c'est-à-dire que plusieurs facteurs génétiques peuvent être en cause. Le calcul du

HLOD score se base sur la fraction α, qui correspond à la proportion de familles liées au marqueur d'intérêt, et est obtenu par la formule suivante :[48]

$$\mathbf{HLOD}(\alpha,\theta) = \log_{10}\left\langle \frac{\text{Vraissemblance des résultats sous l'hypothèse } \alpha < 1 \text{ et } \theta < 1/2}{\text{Vraissemblance des résultats sous l'hypothèse } \alpha = 1 \text{ et } \theta = 1/2} \right\rangle$$

Ce score permet donc de tester la liaison en cas d'hétérogénéité de locus, ce qui ne serait pas détectable par une analyse classique de plusieurs familles liées à différents facteurs.

La Figure 5 résume les différences entre les analyses d'association sur génome entier et les analyses de liaison.

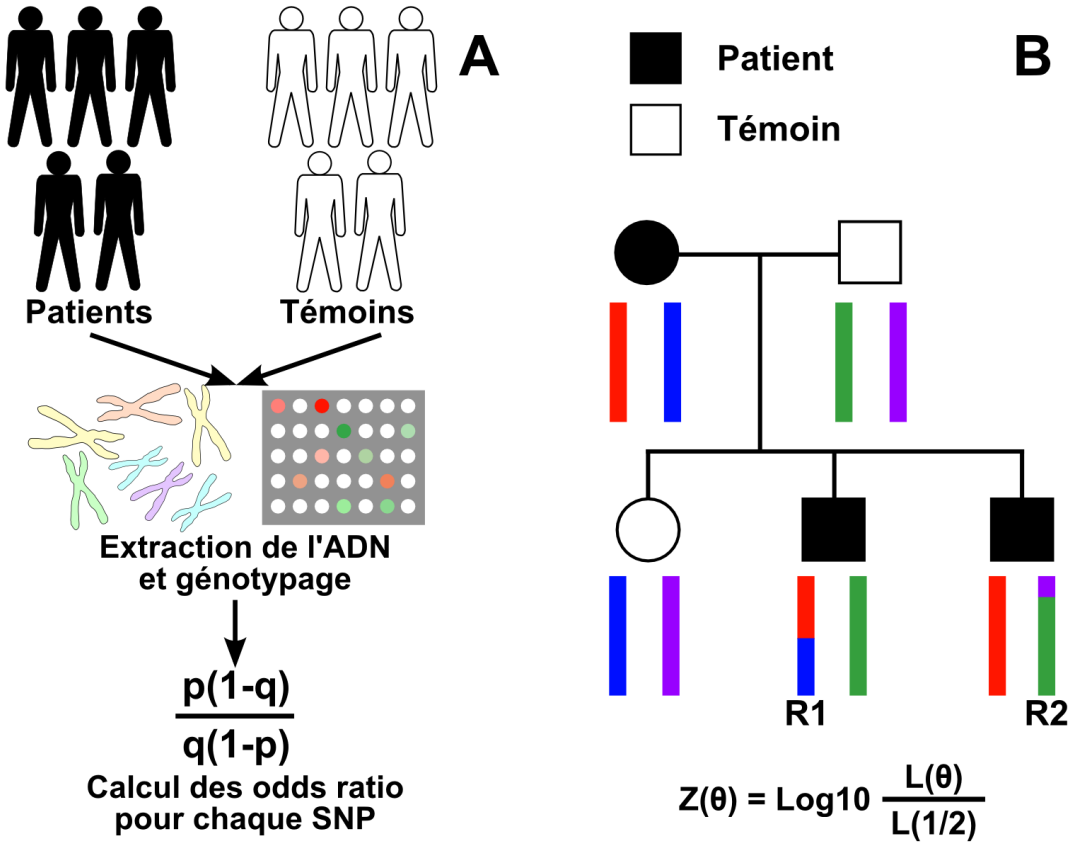


Figure 5: Différences entre les GWAS (A) et les analyses de liaison familiales (B)

L'analyse d'association compare les fréquences des allèles entre patients et témoins, tandis que l'analyse de liaison prend en compte la co-ségrégation de la maladie avec le locus à risque. Dans notre exemple, R1 est une recombinaison informative, contrairement à R2. Le locus à risque est situé dans la zone rouge. Z correspond au LOD score, et L est le rapport de vraisemblance, pour une fraction de recombinaison θ.

I.2) Les agrégations familiales liées à la SpA

Bien que des facteurs environnements soient impliqués dans la SpA, la ségrégation familiale qui apparaît dans les familles de patients est une caractéristique qui démontre clairement l'existence de facteurs de susceptibilité génétique. La SpA présente un ratio de risque de récurrence chez les apparentés au premier degré de 40 dans son ensemble. Pour le sous-type SA, ce risque est estimé à 80.[49,50] Ce ratio, appelé aussi lambda s, est défini comme le rapport de la manifestation de la maladie, compte tenu du fait qu'un frère ou une sœur est affecté, comparée avec la prévalence de la maladie dans la population.[51] La concordance entre les jumeaux monozygotes est supérieure à 50 %, ce qui renforce l'implication de facteurs génétiques.[52] L'héritabilité au sens large correspond à la proportion de la variance de la population due aux effets génétiques, et est déterminée par la somme des composantes génétiques additives et dominantes divisée par la variance de la population pour n'importe quelle maladie. L'héritabilité au sens strict ne prendra en compte que la composante génétique additive, et déterminera la composante génétique des déterminants d'héritage d'une génération à la prochaine. Celle-ci est estimée, pour la SA, à plus de 90 %.[20]

I.3) La SpA et HLA-B27

L'allèle HLA-B27 (pour Human Leukocyte Antigen B27) est très fortement associé à la SA et de façon plus modérée aux autres sous-types de la SpA (Tableau 2). En effet, les études épidémiologiques ont montré une présence de cet allèle chez au moins 80 % des patients atteints par cette pathologie, ce qui est largement supérieur à la prévalence générale de ce marqueur (inférieur à 10 % dans la plupart des pays). HLA-B27 est un antigène de surface de classe I codé par le locus B dans le complexe majeur d'histocompatibilité (CMH) sur le chromosome 6 et présente des peptides antigéniques aux lymphocytes T.

Sous-types SpA	**Prévalence de l'allèle HLA-B27**
Spondylarthrite ankylosante	≥ 90 %
Arthrite réactionnelle	70 %
Rhumatisme psoriasique	60-70 %
Spondylarthrites des entérocolopathies inflammatoires (MICI)	50-60 %

Tableau 2: Prévalence de l'allèle HLA-B27 dans différents sous-types de la SpA

D'après « The genetic basis of spondyloarthritis », Reveille JD[53]

I.3.A) Rappels sur le CMH et le système HLA

Le gène HLA-B appartient au CMH, lui-même étant un ensemble de molécules de surface codés par une large famille de gènes qui contrôlent une majeure partie du système immunitaire chez les vertébrés. Chez l'humain ce complexe est appelé human leukocyte antigen (HLA). Il a été initialement identifié par ses effets majeurs dans le rejet des greffes, d'où la notion d'histocompatibilité. Il est situé sur le chromosome 6, dans la région 6p21.3, s'étend sur 3600 kb et comprend 226 gènes.[54] L'étude approfondie des séquences situées de part et d'autre du CMH a, par la suite, permis d'identifier des gènes synténiques chez la souris, c'est-à-dire une conservation de l'ordre des gènes entre l'humain et cette espèce. Ainsi, la région du CMH a étendu ses limites sur une distance de 7600 kb, donnant la définition du « CMH étendu ».[54] Ce dernier comprend 421 loci, constituant 60 % de gènes exprimés et 7 % de gènes transcrits, et est associé à plus de cent maladies, majoritairement liées avec le système immunitaire.[55] Une autre particularité de la région HLA est qu'elle est la plus dense en gènes et la plus polymorphe.

La région HLA est organisée en trois sous-régions, ou plus communément appelées classes (Figure 6) :

- La région du HLA de classe II, qui est la plus centromérique, code pour des gènes permettant la présentation de l'antigène situé à l'extérieur de la cellule aux lymphocytes T. Ces antigènes présentés vont ensuite stimuler la multiplication des lymphocytes T helper, qui stimuleront eux-mêmes la production des anticorps par les lymphocytes B contre l'antigène spécifique. Les antigènes du soi n'induisent pas de réaction humorale grâce aux lymphocytes T régulatrices. Les molécules correspondant à ce locus sont entre autres HLA-DP, -DQ, -DR, -DM, DOA, DOB, et les transporteurs associés au peptide TAP1 et TAP2.
- La région HLA de classe III, qui représente la zone intermédiaire, encode pour les composants du système du complément. Ce dernier représente une partie du système immunitaire qui aide ou qui complète la capacité des anticorps et des cellules phagocytaires à retirer les pathogènes d'un organisme. Les gènes de cette région codent notamment pour des molécules de l'inflammation (TNFα et β) et des protéines de choc thermique (HSP70).

- La région HLA de classe I, la plus télomérique, code pour les molécules de type HLA-A, B ou C. Les molécules issus de cette région, comme HLA-B27, sont des dimères formés d'une chaîne lourde α de 44 kDa liée de manière non covalente à la β2-microglobuline humaine (β_2m), chaîne invariante légère de 11,5 kDa non codée par le CMH. La chaîne α se compose d'une partie intracytoplasmique, une partie transmembranaire et une partie extracellulaire qui compte trois domaines : α1, α2 et α3. Ces molécules présentent à la surface de la cellule des peptides situés originellement à l'intérieur, comme des fragments d'antigènes issus de virus ou des antigènes non tolérés marqueurs de cellules cancéreuses. Contrairement aux peptides présentés par les molécules du HLA de classe II, ceux présentés par le HLA de classe I ont une longueur petite et fixe, d'environ 9 acides aminés. Les antigènes ainsi présentés attirent les lymphotyce T CD8 (ou killer) pour détruire la cellule cible. Les gènes codant pour le CMH sont principalement exprimés par les cellules présentatrices d'antigène (CPA). Celles-ci sont de deux catégories, les non-professionnelles, qui n'expriment pas de façon constitutive HLA de classe I ou II, et les professionnelles, comme les cellules dendritiques.

Les régions de classe I et II sont riches en pseudogènes, suggérant une duplication ancestrale multiple de certains gènes, générant de nouveaux membres de ces familles de gènes pour de nouvelles fonctions.[56] Les pseudogènes sont des gènes inactifs mais qui ont été démontrés récemment comme jouant un rôle.

Les gènes du CMH sont considérés comme co-dominants et fortement polyalléliques, atteignant 2128 allèles pour les gènes de classe I et 954 pour ceux de classe II. L'ensemble de ces gènes se transmet généralement en bloc, laissant les recombinaisons arriver rarement ($\leq$ 1 % en moyenne). Les pressions de sélection sur ces régions sont particulièrement importantes, résultant ainsi d'un fort DL entre les allèles de certains gènes de ce complexe.

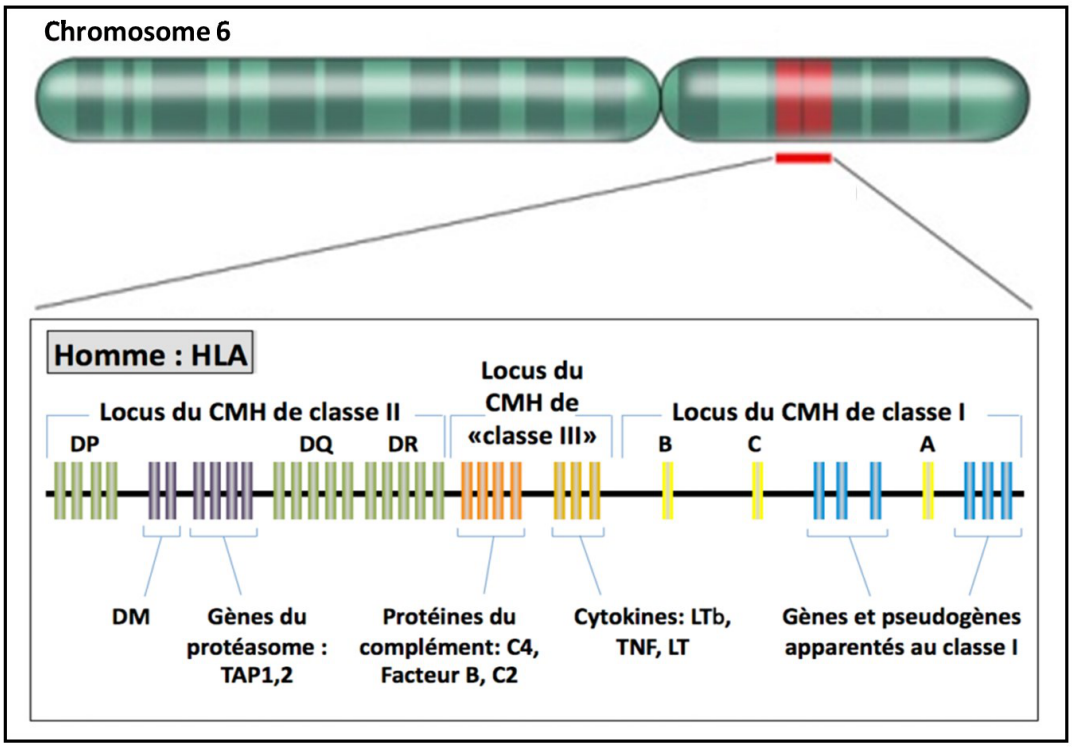


Figure 6: Organisation de la région HLA et de ses 3 sous-loci, le CMH de classe I, II et III.

D'après « Les bases de l'immunologie fondamentale et clinique », édition 2005, Abbas A.K. et Lichtman A.H.[57]

I.3.B) Hypothèses sur la physiopathologie

L'association entre la SA et HLA-B27 est connue depuis 1973.[58] Cependant, la physiopathologie, c'est-à-dire les mécanismes à l'origine de cette association, n'est pas encore bien définie. Deux types de théorie ont été développées sans confirmation définitive, qui se basent spécifiquement sur l'antigène ou non. L'hypothèse du mimétisme moléculaire lie directement le mécanisme d'auto-immunité induite par réponse immunitaire croisée contre des bactéries, et pose comme postulat que les anticorps dirigés contre ces pathogènes auraient une réponse croisée contre HLA-B27 au niveau des tissus. Cette théorie peut notamment s'appuyer sur le facteur de déclenchement des arthrites réactionnelles.[59] L'autre théorie spécifique de l'antigène impliquerait des peptides arhtritogènes, c'est-à-dire des fragments d'antigènes bactériens possédant une homologie de séquence avec des peptides à tropisme articulaire qui généreraient des clones auto-réactifs.[60] Les théories indépendantes de l'antigène se basent plus sur la structure de la molécule HLA-B27. La

première théorie part du principe que la chaîne lourde de HLA-B27 se replie plus lentement que d'autres molécules du HLA. Cela entraînerait une accumulation des chaînes lourdes mal repliées durant l'induction de l'expression de HLA par les cytokines, et ainsi un stress du réticulum endoplasmique (RE) qui résulterait d'une activation de l'unfolded protein response (UPR).[61] Cette réponse induit l'activation du facteur de transcription NFκB et de la production de cytokines pro-inflammatoires. L'autre théorie se base aussi sur la chaîne lourde de la molécule HLA-B27, mais plutôt sur la formation d'homodimères. Ceux-ci, en étant exportés à la surface cellulaire, pourraient être reconnus par des récepteurs KIR sur les cellules natural killer (NK) et les lymphocytes T. Cette reconnaissance peut protéger les cellules NK de l'apoptose (mort cellulaire programmée) et favoriser la production d'IL-17 par les lymphocytes T, ce qui induit une réponse inflammatoire.[62,63]

I.3.C) Les sous-types de HLA-B27 et la SpA

Il existe plusieurs sous-types de la molécule HLA-B27 par l'importante variabilité génétique associée. Ceux-ci dériveraient tous par mutations de l'allèle ancestral et majoritaire dans les populations caucasiennes, HLA-B2705.[64] La plupart des autres sous-types communs sont associés à la maladie, à savoir B2702, B2704 et B2707.[65] Les sous-types B2706 (fréquent en Asie du Sud-Est) et B2709 sont peu ou pas associés à la maladie.[66,67] D'autres allèles HLA de classe I ont été proposés comme associés à la SpA, mais présentent une forte discordance au niveau des études. Seuls HLA-B60 et HLA-B1403 montrent une association convaincante, le premier augmentant la susceptibilité chez les sujets HLA-B27 positifs, et le deuxième dans les populations sub-sahariennes. Ces dernières populations ont une prévalence très faible de l'allèle HLA-B27.[64,68]

I.4) Autres facteurs génétiques que le CMH

Seulement une minorité de la population HLA-B27 positif générale développe la SpA, c'est-à-dire entre 1 et 5 %. D'autres facteurs génétiques supplémentaires ont été depuis identifiés, mais qui ont pourtant un effet faible sur la prédisposition de la maladie, suggérant un mécanisme multifactoriel avec une hétérogénéité importante au niveau des locus.

I.4.A) Études d'association sur la SpA

De multiples études ont été conduites ces dernières années pour déterminer quels sont les facteurs

de susceptibilités génétiques prédisposant à la SpA en dehors du CMH. En effet, comme décrit précédemment, l'allèle HLA-B27 explique très peu l'héritabilité de la SpA.

Plusieurs études gène-candidat dans la SA (ou la SpA en général) ont été menées, mais la majorité des associations démontrées n'ont pu être répliquées de façon indépendante, ce qui laisse particulièrement perplexe sur la véracité de l'association de ces locus. Cette absence de réplication s'observe dans la plupart des études de maladies complexes. Ces gènes sont pour la plupart impliqués dans la régulation de la prolifération des lymphocytes T ou B (IL-10, TGFB1, CTLA4, TLR4)[69–72], la réponse médiée cellulaire aux interleukines (STAT3)[73], des gènes associés au lupus erythémateux systémique (une autre maladie inflammatoire, PDCD1 et FCGR2B)[74,75] ou bien au diabète de type 2 et à l'angiotensine (ACE)[76].

Des GWAS menées depuis 2007 se sont montrées plus convaincantes pour associer des SNPs ou des gènes à la SpA. Quatre GWAS ont été publiées, dont trois dans des populations caucasiennes et une dans la population chinoise Han. Elle a permis notamment de découvrir 11 locus associés, qui sont IL23R, ERAP1, les régions 2p15 et 21q22, PTGER4, GPR25-KIF21B, RUNX3, IL12B, LTBR-TNFRSF1A, NPEPPS-TBKBP1-TBX21 et CARD9, et qui se recoupent sur plusieurs études.[77–80] Une étude publiée en 2013 a utilisé des puces dédiées Immunochip pour une approche genes-candidats à grande échelle. Cette analyse imposante, menée par le consortium international de la génétique de la SA (IGAS), regroupe 10 619 patients SA et 15 145 témoins. Elle a permis de confirmer l'association des 11 locus listés précédemment, et de découvrir 13 nouveaux locus associés à la SA. Ceux-ci comprennent IL6R, FCGR2A, UBE2E3, GPR35, BACH2, ZMIZ1, NXX2-3, SH2B3, GPR65, IL27-SULT1A1, NOS2, TYKE2, ICOSLG. Bien que de nombreux locus ont été découverts récemment, l'ensemble des facteurs identifiés expliquent moins d'un quart de l'héritabilité de la SA. En effet, HLA-B27 explique à lui seul 20,1%, tandis que les autres facteurs en expliquent seulement 4,3 % (Figure 4).[22] Cependant, ces locus permettent de mieux préciser la compréhension de la SpA et de développer de nouvelles hypothèses. Il est tout de même crucial de découvrir encore d'autres facteurs génétiques, et notamment les variants plus rares grâce à l'analyse de liaison familiale, afin de pouvoir encore creuser d'autres facteurs non découverts par les analyses d'association.

I.4.B) Études de liaison sur la SpA

Plusieurs analyses de liaison sur génome entier ont été publiées afin d'identifier des régions liées à la maladie qui ne seraient pas détectées par les GWAS. Depuis 1998, trois études ont été menées pour identifier les régions liées à la SA, dont une qui incluait également d'autres sous-types de SpA.

La première étude ne se portait pas sur des SNPs, mais plutôt des microsatellites. Ces derniers étaient utilisés pour les études de marqueurs avant les nouvelles technologies de puces à ADN, et correspondent à des répétitions de motifs dans la séquence dont le nombre diffère entre chaque individu. Cette étude comptait 185 familles comprenant 255 paires de germains atteints de SA. Leur analyse de liaison non paramétrique multi-point a permis de découvrir deux régions significativement liées à la SA, qui sont la région du CMH avec un LOD score de 15,6 et la région 16q23.3 avec un LOD score maximal de 4,7. Cinq locus ayant atteint un seuil de significativité suggestif (entre 2,2 et 3,6) ont été également été découverts, qui sont 1p, 2q, 9q, 10q et 19q.[81] Le consortium Nord-Américain de la SpA (NASC) a publié en 2004 une autre étude des micro-satellites, qui compte 180 familles, ne mettant en évidence que la région du CMH avec une liaison significative à la SA.[82] Le Groupe Français d'Etude Génétique de la SpA (GFEGS) a mené une autre étude avec 120 familles et a trouvé une liaison significative avec la région 9q31-34, en plus du CMH, qui chevauchait la région avec une liaison suggestive démontrée dans la première étude. C'est ainsi que cette région a été nommée SPA2. L'autre intérêt de cette étude est qu'elle incluait également les autres sous-types de SpA.[83] Deux méta-analyses ont été ensuite effectuées, se basant sur les trois études décrites précédemment ; une utilisant une comparaison des résultats d'analyse statistiques directement ;[84] l'autre se basant sur les données brutes de génotypage.[85] Les deux ont permis de mettre en évidence une liaison suggestive avec les locus 6q et 16q, tandis que la première méta-analyse a identifié les régions 9q, 17p et 19q, et la deuxième la région 10q.

I.4.C) Étude approfondie du locus SPA2

Parmi toutes les régions précédemment décrites comme liées de manière significative à la SpA, seul le locus SPA2 (qui correspond à la région 9q31-34) a été étudiée de manière plus approfondie. Une analyse de liaison sur des données de génotypage plus denses (149 familles), en 2009, a permis de réduire la région d'intérêt à un intervalle de 13 Mb. Cette région a ensuite été

soumise à une cartographie fine de DL (136 familles) afin d'identifier un SNP significativement associé à la SpA, rs4979459. Grâce au génotypage de ce SNP et d'une trentaine d'autres autour dans 287 familles, une association significative a été démontrée et répliquée dans une population indépendante entre un haplotype de six SNPs et la maladie.[86] La ténascine C, un gène-candidat positionné dans cette région, a été re-séquencé et les polymorphismes les plus intéressants ont été par la suite génotypés dans une collection de trios. Aucune association n'a pu être détectée.[87] D'autre part, les régions codantes et régulatrices de neuf gènes-candidats supplémentaires de cette région ont été re-séquencées, qui comprennent ZNF616, A1L4R1_HUMAN, AMBP, KIF12, ORM1, ORM2, C9ORF91, ENSESTG000000230601 et TNFSF8. Plusieurs études d'extension (dans une population de cas-témoins français indépendants) et de réplication (cohortes portugaises et belges) ont démontré une association significative du SNP intronique rare du gène TNFSF8. Nous avons ici l'exemple qu'une analyse beaucoup plus fine de la liaison permet de mettre en cause des SNPs avec des MAF très faibles, dont le rôle dans la SpA n'aurait jamais été soupçonné auparavant.[88] Cependant, ce SNP n'explique pas à lui seul toute la liaison qu'on peut accorder au locus SPA2.

Le Tableau 3 représente une synthèse de toutes les études d'association et de liaison sur la SpA, en dehors de l'étude approfondie du locus SPA2.

Etude	Cohorte étudiée	Gènes ou locus identifiés
	Etudes d'association	
Burton et al.[77]	Découverte : 922 cas/1466 témoins Réplication : 471 cas/625 témoins	**IL23R, ERAP1**
Reveille et al.[78]	Découverte : 2053 cas/5140 témoins Réplication : 998 cas/1518 témoins	**IL23R, ERAP1** 2p15, 21q22 ***ANTXR2, IL1R2***
Evans et al.[79]	Découverte : 3023 cas/8779 témoins Réplication : 2111 cas/4483 témoins	**IL23R, ERAP1, 2p15, 21q22,** KIF21B, RUNX3, IL12B, LTBR-TNFRSF1A, ***ANTXR2, PTGER4, CARD9, TBKBP1***
Lin et al.[80]	Découverte : 1837 cas/4231 témoins Réplication : 2100 cas/3496 témoins	EDIL3-HAPLN1, ANO6, **2p15**
IGAS et al.[22]	10 619 cas/15 145 témoins	**RUNX3, IL23R, GPR25-KIF21B, 2p15,** **PTGER4, ERAP1, IL12B, CARD9, LTBR-TNFRSF1A, NPEPPS, 21q22** IL6R, FCGR2A, UBE2E3, GPR35, BACH2, ZMIZ1, NKX2-3, SH2B3, GPR65, IL27-SULT1A1, NOS2, TYK2, ICOSLG
	Etudes de liaison	
Oxford[81,89]	185 familles	Significatifs : CMH, 16q23.3 Suggestifs : 1p, 2q, **9q**, 10q et 19q
NASC[82]	180 familles	Significatif : CMH
GFEGS[83]	120 familles	Significatifs : CMH, **9q31-34**
Lee et al.[84]	Méta-analyse	Significatif : CMH Suggestifs : 6q, **9q**, 16q, 17p et 19p
Carter et al.[85]	Méta-analyse	Significatif : CMH Suggestifs : 6q, 10q et 16q

Tableau 3: Synthèse des études d'association et de liaison sur la SpA

I.4.D) L'intérêt des approches fonctionnelles

De manière générale, nous pouvons remarquer qu'il est nécessaire de découvrir d'autres facteurs génétiques de susceptibilité à l'origine de la SpA. C'est pourquoi des travaux de grande envergure de séquençage sont lancés, avec des outils statistiques plus spécialisés et des moyens informatiques conséquents, afin de pouvoir encore plus investiguer le génome humain à l'origine de cette pathologie. De plus, avec l'ère du séquençage haut débit qui vient s'ajouter aux connaissances plus approfondies des mécanismes épigénétiques, il est évident qu'un nombre

conséquent de facteurs restent encore à découvrir.

Cependant, les études d'association ou de liaison génétiques ne suffisent pas à poser des hypothèses sur tous les mécanismes pouvant expliquer l'étiologie de la SpA. Une approche de biologie des systèmes est nécessaire, d'autant plus pour une maladie multifactorielle. L'analyse du transcriptome est par exemple un excellent moyen complémentaire d'investiguer sur les mécanismes fonctionnels à l'origine de la SpA.

II/ Associations du transcriptome

II.1) Utilité de l'étude de l'expression des gènes

Le transcriptome constitue l'ensemble de tous les ARN issus de la transcription du génome. Ils comprennent les ARN messagers, ribosomiques, de transfert et autres espèces d'ARN comme les micro-ARN (miRNA). L'analyse du transcriptome est une approche qui peut être complémentaire aux études génétiques pour expliquer la physiopathologie de maladies multifactorielles, et se base sur l'identification de gènes qui sont différentiellement exprimés (DE) entre les patients et les témoins, en ajoutant des facteurs supplémentaires de contrôle afin de retirer des facteurs de susceptibilité déjà mis en cause.

II.1.A) Les biomarqueurs

Cette approche permet par exemple de définir des biomarqueurs qui peuvent être utilisés par la suite de manière standardisée comme moyen de diagnostic. En effet, la mesure de l'expression de gènes identifiés comme biomarqueurs chez un patient est un moyen rapide, peu coûteux, et plus précis pour définir si la personne est bien atteint de la maladie ou non. De plus, de telles techniques peuvent identifier des phases précoces de la maladie, sans qu'aucun symptôme ne soit visible à l'œil nu expert du médecin. Ce type de mesure sur les biomarqueurs peut se faire de manière plus fonctionnelle et indirecte, par des réactions enzymatiques ou des tests ELISA.[90]

II.1.B) Les nouvelles voies de signalisation

Une autre utilité de l'analyse d'expression des gènes par rapport à une maladie est de découvrir des nouvelles voies de signalisation comprenant les gènes DE. Ces voies permettent de mieux comprendre les mécanismes à l'origine d'une maladie multifactorielle, et ne seraient pas forcément découverts par des études génétiques. En effet, l'expression d'un gène dépend très faiblement de la

séquence ADN qui la concerne, mais plus d'autres facteurs de transcription ou des mécanismes épigénétiques, comme la méthylation de l'histone, qui sont très difficilement détectables par l'étude du génome.[91]

II.1.C) Les eQTLs

L'analyse du transcriptome peut être complémentaire aux études génétiques de manière encore plus étroite, par l'approche des expression quantitatite trait loci (eQTLs). Les eQTLs représentent les traits (principalement des SNPs) dont l'allèle change l'expression d'un gène cible. Nous distinguons deux types d'eQTLs : les cis et les trans. Les cis-eQTLs représentent les marqueurs situés de manière assez proche des gènes affectés et qui modifieraient l'habilité aux facteurs de transcription à se fixer sur leurs promoteurs. Les trans-eQTLs sont des marqueurs plus éloignés de leurs gènes cibles et influencent de manière indirecte l'expression de ceux-ci, par des éléments régulateurs positionnés près du variant tels que les miRNA ou des facteurs de transcription (Figure 7). Un eQTL peut être détecté en calculant le facteur de corrélation entre un allèle (trait qualitatif) et l'expression d'un autre gène (trait quantitatif). C'est donc une approche qui utilise exactement les deux types de données présentées dans cette thèse.[92]

La cartographie des eQTLs sur génome entier fait preuve d'un effort considérable qui se développe depuis seulement quelques années, et qui donne une approche novatrice sur la compréhension des mécanismes moléculaires régulant le fonctionnement d'une cellule. Les cis-eQTLs sont plus simples à déterminer, car nous avons déjà un tri effectué a priori sur les corrélations à tester par les localisations génomiques des marqueurs et des gènes. En revanche, les trans-eQTLs ajoutent une difficulté supplémentaire à l'analyse, car plusieurs faux positifs peuvent survenir par des éléments régulateurs indirects qui ne sont pas forcément liés au SNP testé. Des outils statistiques, bioinformatiques et de réseaux peuvent être utilisés afin de permettre une évaluation des gènes candidats et des systèmes entiers d'interactions.[88] Notamment, il est possible de retirer par des régressions linéaires les faux positifs par des effets cis-eQTL indirects. Il est primordial de faire une sélection fine des marqueurs des gènes à corréler afin d'éviter des temps de calculs et un bruit de fond considérables.

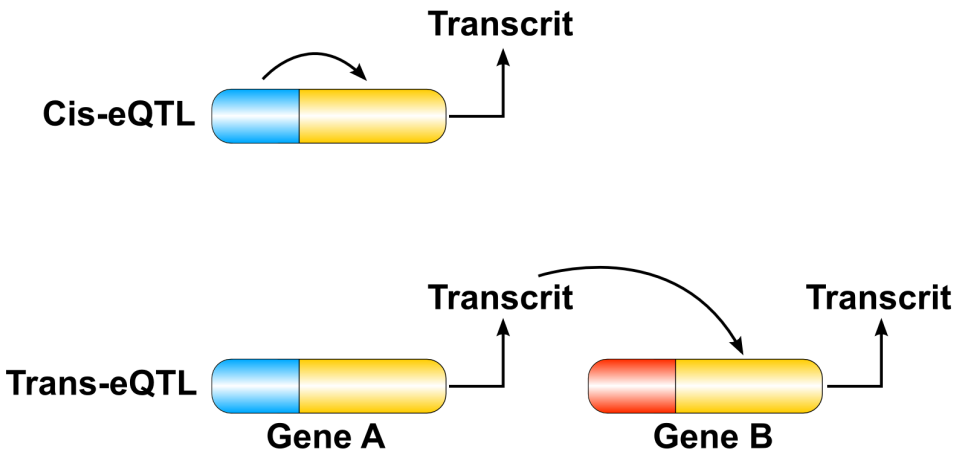


Figure 7: Principes du Cis- et Trans-eQTL

Dans le cas du Cis-eQTL, le SNP modifie directement l'expression du gène. Le Trans-eQTL modifiera sa cible indirectement via la régulation par un autre gène ou micro ARN.

II.2) L'expression des gènes et le type cellulaire

L'étude de l'expression des gènes, contrairement à l'analyse des polymorphismes, dépend fortement des cellules étudiées et du contexte expérimental dans lequel elles sont placées. Tandis qu'une étude génétique traduira une condition pérenne pendant toute la vie du patient, nous pouvons voir l'étude d'expression des gènes comme une photographie de l'état de fonctionnement des cellules du patient dans un contexte donné. Il est, pour cette raison, primordial de bien définir le design expérimental, afin d'être le plus représentatif du contexte environnemental dans lequel se trouve le patient pendant la phase réactive de la maladie. L'avantage d'une telle approche est que nous passons à une étape supérieure dans le lien entre le fonctionnel et l'étiologie de la maladie. Il existe par ailleurs des bases de données recensant des profils d'expressions typiques qu'on peut trouver dans les différents tissus et différentes cellules, et même entre les différentes espèces, afin de s'assurer de la qualité de nos expériences.[93–96]

II.3) Les puces transcriptomiques

De la même manière que pour les GWAS, des puces ont été développées pour évaluer l'expression des gènes sur le génome entier, plus communément appelées des puces transcriptomiques. Ces puces sont constituées de molécules d'ARN fixées immobilisées qui représentent des séquences spécifiques de gènes localisés le long du génome. La plupart des puces utilisent une série de

sondes pour reconnaître un seul gène, afin d'augmenter la spécificité, et certains incluent même les différents épissages alternatifs du transcrit. L'ARN total d'une cellule est tout d'abord rétro-transcrit en ADN complémentaire (ADNc), puis utilisé pour la synthèse d'un ARN complémentaire marqué par de la biotine ou un autre marqueur fluorescent. Ces séquences étiquetées sont ensuite fragmentées et déposées sur les sondes pour s'hybrider, entraînant ensuite la production d'un signal (fluorescence émise après excitation par un laser pour la grande majorité). Les sondes sont organisées par spots sur la puce, et un spot permet de quantifier le nombre de sondes hybridées par la quantité de fluorescence, permettant ainsi de déterminer le nombre de transcrits et donc le niveau d'expression d'un gène (Figure 8). Bien que les données issues de puces transcriptomiques sont d'une précision remarquable, il convient de valider les résultats des gènes DE les plus intéressants par d'autres technologies, le plus souvent par qPCR. Afin de s'affranchir du biais expérimental dû à l'importante variabilité inter-individuelle de l'expression des gènes (qu'il est très difficile de contrôler), il est recommandé d'effectuer ces validations sur une cohorte indépendante.[97]

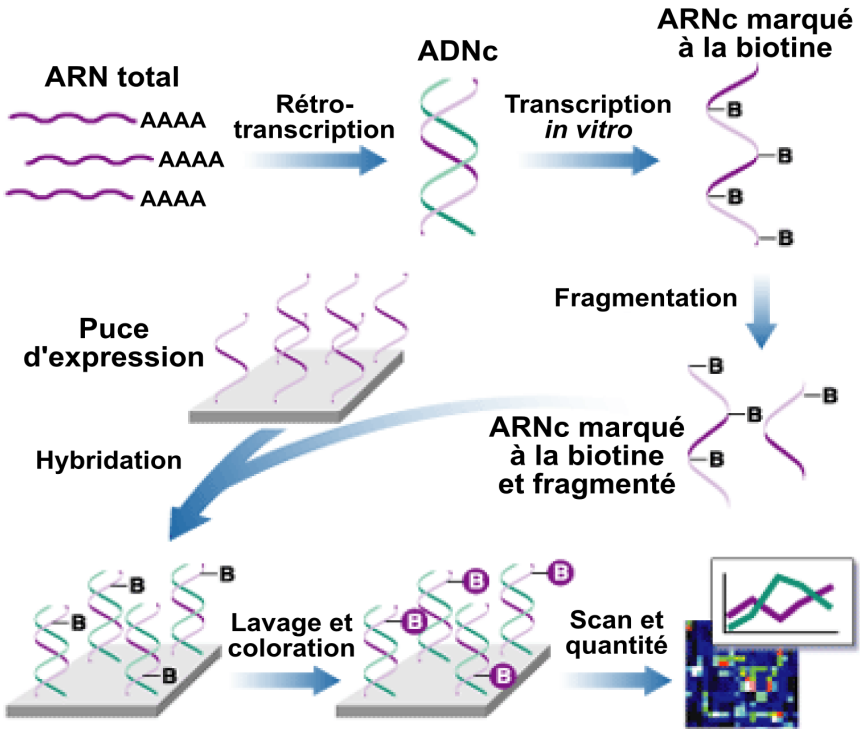


Figure 8: Etapes d'une étude d'expression globale des gènes par puce transcriptomique

D'après les protocoles donnés sur affymetrix.com

II.4) Les technologies RNA-seq

Les puces d'analyses du transcrit sur génome entier se sont démocratisées ces dernières années, tout comme les puces à ADN de génotypage. De la même manière, des technologies de séquençage à haut débit ont été adaptées pour l'ARN, qui sont appelées les technologies RNA-seq. Elles permettent de séquencer directement l'ARN total et de compter le nombre de copies de séquences détectées. Les données quantitatives obtenues pour mesurer l'expression de chaque gène sont donc sous forme de variables discrètes, contrairement à l'intensité de fluorescence. Cela permet donc d'obtenir un résultat d'une extrême précision, puisque la mesure se fait plus directement sur l'extrait d'ARN. De plus, les puces transcriptomiques se limitent souvent à quelques milliers de transcrits et n'incluent que les gènes déjà connus. Or, la technologie RNA-seq permet aussi d'analyser toutes les séquences extraites dans l'ARN total (en prenant en compte aussi les épissages alternatifs).[98] Bien que ces technologies génèrent des données beaucoup plus importantes et requièrent un effort supplémentaire au niveau des analyses statistiques et des ressources bioinformatiques, il est évident que ces technologies commencent à remplacer les puces transcriptomiques souvent utilisées ces dernières années.

II.5) Analyse des données de puces transcriptomiques

Les puces transcriptomiques génèrent un nombre important de données. Tout comme pour les GWAS, une qualité supérieure est requise pour les analyses statistiques. Des compétences en bioinformatique sont nécessaires, même si des outils intuitifs et très efficaces existent, qui intègrent toutes les étapes d'analyse des puces, de la normalisation jusqu'à l'étude des réseaux de gènes.[99,100] Cependant, ces logiciels sont pour la plupart payants, avec des licences pouvant atteindre des sommes très élevées, ce qui restreint leur utilisation à des plate-formes d'analyse et qui n'est pas forcément avantageux pour les laboratoires qui souhaitent effectuer leurs propres analyses. De plus, par leur souci de protection des propriétés intellectuelles, ces outils ne laissent pas forcément un contrôle total sur la qualité, l'explication ou les sources des résultats. De nombreux outils open-source sont développées par la communauté scientifique, notamment bioconductor qui se base sur le langage statistique R, mais qui demandent tout de même des connaissances plus élevées en informatique.[101] Il est important d'avoir une certaine expertise lorsqu'on analyse des données de technologies à haut débit, surtout pour des designs

expérimentaux complexes à plusieurs facteurs, ce qui est maintenant le cas pour la plupart des études sur puces transcriptomiques.

II.5.A) Normalisation des données

En plus de la variabilité très importante des expressions des gènes entre chaque individu, les puces transcriptomiques peuvent présenter une variabilité indépendante entre elles par les différentes qualités expérimentales, mis en cause par des facteurs que l'expérimentateur ne contrôle pas forcément. C'est pourquoi il est important, comme pour tout type d'expériences permettant d'analyser les expressions de gènes, d'effectuer une normalisation entre tous les échantillons, afin d'obtenir une homogénéité dans les résultats et de s'affranchir de tout biais d'échantillonnage. Actuellement, il est largement accepté par la communauté scientifique d'utiliser l'approche robust multi-array average (RMA), qui est maintenant la méthode standard pour cette étape.[102] Les données sont normalisées de telle sorte que les valeurs soient toutes positives, et que la moyenne globale des échantillons soit homogène.

II.5.B) Méta-analyses

Afin d'augmenter la puissance de ses résultats ou de contrôler la qualité de ses expériences de puces d'expression, il est possible d'effectuer des méta-analyses, c'est-à-dire de regrouper des données issues de différentes expériences. En effet, les résultats d'études d'expressions sur génome entier donnent très rarement un recouvrement satisfaisant, et il est parfois intéressant de regrouper toutes ces études afin de trouver un compromis entre les données. Il est possible de les récupérer par des bases de données regroupant toutes les données d'expression pan-génomiques publiées, comme Gene Expression Omnibus (GEO).[103] Il existe principalement deux types de méta-analyses sur les données génomiques, en fonction de la méthode de rassemblement des données. Les méta-analyses horizontales combinent différentes cohortes pour la même classe de données (par exemple puces transcriptomiques) et donc le même événement moléculaire. Les méta-analyses verticales rassemblent différentes mesures génomiques, comme du séquençage ADN ou des études protéiques, sur une même cohorte afin d'effectuer une analyse à différents niveaux (Figure 9). Par exemple, l'approche eQTL est une méta-analyse verticale, tandis que le rassemblement de plusieurs expériences de puces transcriptomiques en est une horizontale.[104]

De même, il existe deux niveaux de méta-analyse. Le premier consiste à comparer les statistiques issues des analyses, plus précisément les p-values, afin de s'affranchir de la grande variabilité des niveaux d'expressions des gènes que nous pouvons trouver entre les études (ce qui est d'autant plus le cas lorsqu'elles viennent de technologies de puces différentes). Cette approche, utilisée de façon très commune, consiste à comparer les résultats de plusieurs études et à recouper les gènes qui sont DE dans chacune d'elle, pour représenter ensuite ces croisements par un diagramme de Venn. Nous pouvons citer les méthodes de Fisher et Stouffer, qui utilisent des p-values transformées en logarithmes ou normales inverses, et minP et maxP, qui utilisent les p-values aux extrémités comme statistiques de test. MetaDE, un package R développé à l'Université de Pittsburg et publié en 2012, permet d'inclure toutes les méthodes de comparaison des statistiques majeures décrites dans la littérature.[105]

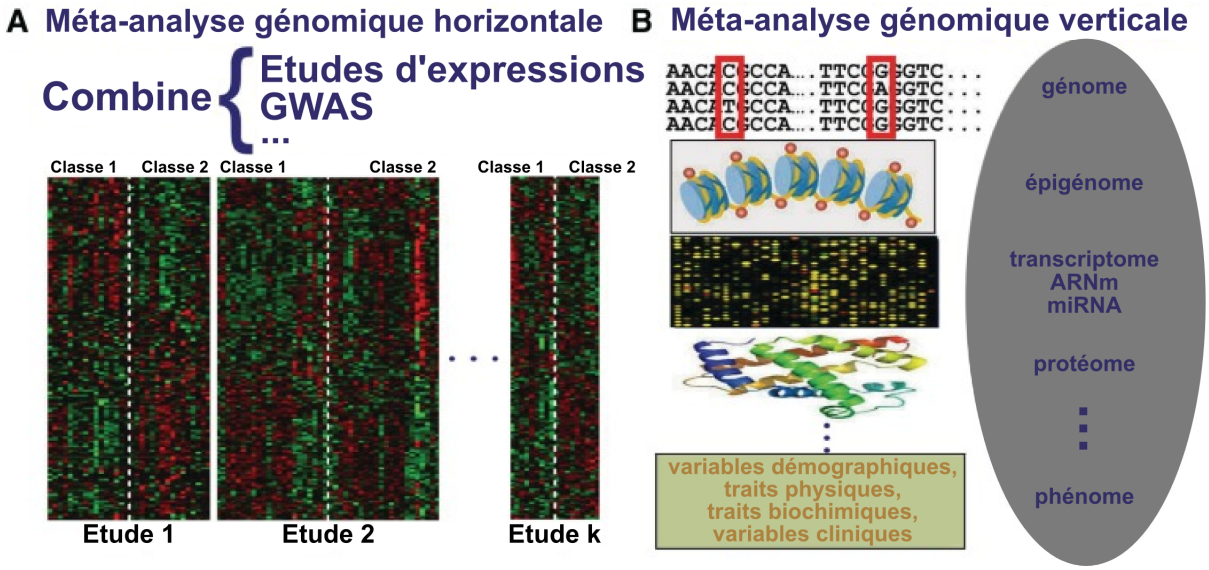


Figure 9: Types d'informations intégrées dans les méta-analyses

D'après « Comprehensive literature review and statistical considerations for microarray meta-analysis », Tseng et al.[104]

Le deuxième niveau de méta-analyse rassemble les niveaux d'expressions brutes et effectue des analyses statistiques directement sur les résultats. Cette approche doit être effectuée avec beaucoup de précautions, et une homogénéité extrême est indispensable entre toutes les études. Cependant, il existe des contrôles qualités objectifs qui permettent de déterminer au préalable s'il est possible d'effectuer ce type d'approche. Le même laboratoire que ceux qui ont développé

MetaDE proposent MetaQC, un package R développé dans cette optique et publié la même année. Cet outil inclut principalement six critères d'évaluation. Les indices IQC (internal quality control) et EQC (external quality control) permettront d'évaluer respectivement : l'homogénéité des co-expressions entre les études ; et la cohérence de celles-ci avec des voies de signalisation connues et stockées dans des bases de données publiques (MSigDB, KEGG, GO et Biocarta). Les indices de critères AQCg (accuracy quality control for genes) et AQCp (accuracy quality control for pathways) mesurent la précision de la détection des biomarqueurs (ou gènes DE) et des voies de signalisation enrichies. Le même type de critères est utilisé pour CQCg et CQCp (consistency quality control for genes/pathways), qui se basent plus les rangs des gènes que les niveaux d'expressions. Ce package, en plus de mesurer des scores de qualité en fonction de chacun de ces critères, effectue une analyse en composantes principales (ACP) et cartographie chaque étude sur un graphe expliquant les variances respectives. Ainsi, il est possible de savoir au préalable si des études peuvent être rassemblées ou non pour une méta-analyse plus poussée.[106]

II.5.C) Identification des gènes DE

L'identification des gènes DE consiste à effectuer des tests statistiques pour déterminer si le niveau d'expression d'un gène est significativement modifié en fonction d'un phénotype observé. Plus particulièrement, dans le cas d'une étude d'une maladie, il s'agit de lister les gènes dont le niveau du transcrit est affecté par la pathologie. Depuis 2004, la méthode linear models for microarrays analysis (LIMMA), développée par le laboratoire de Smyth Gordon K., est considérée comme la méthode standard, à la fois souple et robuste pour cette application.[107] Elle se base sur l'utilisation de modèles linéaires pour décrire les données de chaque gène en fonction de chaque facteur, de telle sorte que la somme des coefficients est égale à la moyenne d'expression des échantillons. L'utilisation des modèles linéaires a pour avantage de s'adapter pour tous les designs, et correspond parfaitement à des études multifactorielles complexes. Il est même possible d'ajouter une correction par un facteur d'erreur connu et de prendre en compte la variabilité intra-individuelle en informant sur les échantillons dupliqués. Une étape très importante dans l'utilisation de LIMMA est de bien définir le design, c'est-à-dire la description des échantillons et des facteurs associés, afin que le package construise tous les modèles linéaires nécessaires. C'est à partir de cet ensemble de régressions qu'il est possible d'identifier les gènes DE en définissant tous

les contrastes nécessaires. Ces derniers doivent également être entrés avec précaution pour les analyses multifactorielles. Lorsque toutes les étapes sont bien définies, LIMMA effectue un test de Student, avec une inférence bayésienne empirique en empruntant les informations entre les gènes, afin de déterminer les gènes DE affectés par les contrastes définis. L'inférence bayésienne est une approche statistique qui utilise des lois de probabilité choisies a priori pour déterminer une fonction de vraisemblance, contrairement à une probabilité d'hypothèse nulle pour la statistique classique. Une distribution des paramètres est ensuite calculée a posteriori. La statistique bayésienne possède l'avantage de se baser sur des expériences passées afin de déterminer des paramètres plus proches de ce qu'on attendait en fonction des connaissances acquises.[108]

En plus des analyses statistiques poussées comme LIMMA, il est possible d'utiliser des approches de ré-échantillonnage, comme la permutation ou le bootstrap, afin de mieux contrôler le signal observé. Le bootstrap a pour objectif principal de vérifier que la statistique observée n'est pas due à un biais d'échantillonnage. Le principe consiste simplement à faire un nombre important de tirages avec remise des échantillons, en conservant les facteurs associés à ceux-ci. Ainsi, on calcule une distribution de statistiques qui doit être très proche de la statistique calculée initialement si le signal observé est assez fort. La permutation est une approche similaire, à la différence que les facteurs associés à chaque puce sont tirés sans remise, tout en gardant le même ordre et les mêmes valeurs pour les échantillons. Par cette approche, on calcule une distribution empirique de notre statistique pour l'hypothèse nulle que les résultats observés ne sont pas associés aux facteurs mesurés. Si la statistique calculée initialement est suffisamment éloignée de cette distribution, on peut rejeter cette hypothèse nulle. Cette étape de permutation n'est pas négligeable, même pour une méthode aussi sophistiquée que LIMMA. Cette dernière se base sur un test de Student, qui est un test paramétrique et se base sur une distribution supposée a priori.[109]

II.5.D) Annotations des gènes DE et approche GSEA

Lorsque nous obtenons des listes de gènes DE, ceux-ci sont souvent très nombreux compte tenu des fluctuations importantes des expressions des gènes par de nombreux mécanismes indirects régulant les fonctions cellulaires. Même dans le cas de l'étude d'une maladie non multifactorielle, l'expression des gènes ne se modifie pas de façon isolée. Il est clairement défini que le taux de transcrit global dépend lui-même de nombreux facteurs de régulations, qui peuvent également

s'opérer entre les gènes régulés eux-mêmes.[110] C'est donc une problématique non négligeable pour des études pan-génomiques d'expression, même avec des connaissances très poussées sur les fonctions des gènes, où il est difficile de déterminer les voies de signalisation spécifiques affectées par la maladie étudiée. De plus, les gènes ont souvent des rôles très diverses en fonction du contexte physiologique dans lequel se trouve la cellule. C'est particulièrement le cas pour le système immunitaire qui montre une plasticité impressionnante en fonction du milieu dans lequel il se trouve.[111]

C'est dans cette optique qu'un effort considérable a été effectué pour classer les gènes en fonction d'annotations fonctionnelles. C'est notamment le cas de Gene Ontology (GO), qui est une initiative bioinformatique majeure dans la compréhension globale des fonctions des gènes. Il s'agit d'une base de données considérable regroupant différents termes techniques et bien définis pour associer un rôle différent à chaque gène. Ces termes sont organisés sous forme d'arbres, c'est-à-dire qu'un terme plus global peut regrouper plusieurs termes d'un niveau plus précis. De manière générale, cette base de données se divise en trois parties : les composants cellulaires, les fonctions moléculaires et les processus biologiques. Ces informations sont dynamiques et mis à jour très régulièrement en fonction des découvertes majeures effectuées sur les fonctions. Un terme GO est associé à chaque fois à un groupe de gènes plus ou moins grand en fonction de la spécificité de celui-ci.[112] De nombreuses bases de données regroupent des annotations plus ou moins spécifiques, comme KEGG pour les réactions enzymatiques du système biologique, ou bien OMIM pour les associations mendéliennes des gènes à des maladies définies.[113,114]

Plusieurs outils ont été développés dans le but d'effectuer des tests statistiques pour déterminer des annotations spécifiquement enrichies dans une liste de gènes donnée. En sachant qu'une annotation est associée à plusieurs gènes, nous parlerons de jeux de gènes. L'analyse d'enrichissement des jeux de gènes (GSEA) est la méthode la plus utilisée et la plus inspirée pour d'autres outils plus spécifiques. Elle se base sur MSigDB, une base de données conséquente regroupant plusieurs types d'annotations.[115] Elles comprennent notamment les termes GO, mais également des listes de gènes détectées par d'autres expériences d'expression sur génome entier ou rassemblées en fonction de leur localisation dans le génome. La méthode GSEA teste l'enrichissement d'un jeu de gènes en se basant sur l'hypothèse nulle que celui-ci peut être trouvé

au hasard dans la liste de gènes initiale, et donc n'est pas associé au phénotype. De nombreux autres outils sont inspirés de cette approche et permettent d'inclure des statistiques issues de l'identification des gènes DE et une directionnalité (fold change). Nous pouvons notamment citer Panther, WebGestalt, DAVID ou Enrichr.[116–119] Chaque méthode apporte une approche originale et il est avantageux d'utiliser chacune d'elles pour comparer les résultats. Panther permettra d'obtenir une visualisation intuitive de chaque annotation sous forme de camembert. Enrichr possède une interface très intuitive pour regrouper une liste quasi-exhaustive des bases de données connues d'annotations. WebGestalt donne une interface simplifiée des jeux de gènes enrichis sous forme d'arbre et permet par simple clic d'entrer plus en détail dans les résultats.

Les approches GSEA sont clairement biaisées quand la directionnalité des gènes n'est pas prise en compte, et encore plus quand la corrélation n'est pas mesurée.[120] C'est pourquoi d'autres outils existent, qui permettent d'inclure les données d'expressions brutes des gènes afin de calculer leurs corrélations et de tester la probabilité que des jeux de gènes ne sont pas corrélés, bien que présents dans le jeu de données. Les mêmes développeurs que LIMMA proposent CAMERA, une méthode développée dans cette optique.[121] Cet outil est même déjà intégré de base avec le package R limma. D'après une revue publiée en 2013, il est parmi les méthodes les plus sensibles et les plus spécifiques.[122] Son avantage principal est qu'il se base également sur les modèles linéaires et inclut la même souplesse au niveau du design du modèle statistique. Plus récemment, l'approche quantitative set analysis for gene expression (quSAGE) possède l'avantage de calculer le facteur d'inflation de la variance (VIF) pour chaque jeu de gènes avec un intervalle de confiance.[123] Ainsi, il est possible d'obtenir un fold change par jeu de gènes plutôt que par gène isolé, ce qui est une approche beaucoup plus intuitive pour la mesure d'activité de chacune des annotations.

Cependant, l'analyse des jeux de gènes simplifie les interactions possibles qui existent entre ceux-ci. Elle ne se base finalement que sur des groupes de gènes, qui peuvent eux-mêmes présenter une dynamique bien plus complexe que d'affirmer simplement qu'ils agissent ensemble. De plus, les annotations se regroupent, ce qui a pour conséquence que lors des approches GSEA, nous pouvons nous retrouver avec de nombreuses fonctions redondantes et qui sont parfois trop générales, donc pas forcément pertinentes pour pouvoir conclure sur les mécanismes biologiques à l'origine du phénomène observé. Cette approche est donc maintenant souvent utilisée comme

première synthèse des données de gènes DE identifiés, avant d'entrer plus en détail dans l'exploration des fonctions cellulaires mises en évidence.

II.5.E) La théorie des graphes comme outil d'analyse

La théorie des graphes regroupe l'étude des graphes, ces derniers étant des objets issus des mathématiques discrètes représentant des réseaux avec des nœuds connectés par des arcs. Cette notion était introduite en 1736 par Leonhard Euler, avec le problème des sept ponts de Königsberg.[124] Il existe différents types de graphes en fonction des informations représentées. Ils peuvent être dirigés, quand l'arc représente une connexion à sens unique entre une source et une cible. Il peut également être pondéré quand un score est attribué à chaque arc, ce qui donne une notion de distance. Il s'agit d'un outil très intéressant pour étudier des réseaux complexes, car il est possible d'effectuer des manipulations plus poussées pour identifier des mécanismes clés. La structure d'un réseau, appelée aussi topologie, permet d'obtenir différentes propriétés propres du graphe étudié. Il est également possible de calculer des chemins les plus courts pour voyager d'un nœud à un autre, ce qui permet de développer des théories sur les voies les plus empruntées à moindre coût. Nous parlons de réseau dynamique quand un ensemble de graphes est utilisé à plusieurs temps, regroupant des graphes d'états permanents et des graphes de transition. La représentation visuelle d'un graphe est simple et trivial. On dessine un nœud par un polygone, dont la taille, la forme et la couleur peuvent varier en fonction d'attributs associés à celui-ci. Les arcs sont représentés par des traits dessinés entre chaque nœud, avec toujours une représentation graphique en fonction de différentes propriétés. Un arc dirigé est représenté par une flèche.

Cet outil est grandement apprécié dans l'étude de phénomènes issus du monde réel, comme dans le domaine de l'économie, des réseaux sociaux ou de l'aéronautique.[125–127] La plupart des réseaux du monde réel montrent des propriétés particulières, parmi lesquelles se trouvent les réseaux du monde petit. Ceux-ci représentent un type de graphe où la plupart des nœuds ne sont pas voisins directement entre eux, mais sont connectés par un faible nombre d'étapes.[128] Ces réseaux peuvent être générés aléatoirement en respectant que la distance entre deux nœuds est proportionnelle au logarithme du nombre de nœuds total. Ainsi, des techniques de permutations et de statistiques poussées peuvent être effectués pour tester la robustesse d'un réseau, comme la simulation de Monte Carlo.[129] Il est également possible de détecter des sous-graphes, ou des modules, qui

représenteront une connectivité particulière par rapport au reste du réseau. Dans le cas des réseaux dynamiques, il est intéressant de mettre en évidence des cycles attracteurs, c'est-à-dire un ensemble de nœuds et d'arcs pour lesquels un état stable se met en place grâce à une boucle d'interactions.[130]

II.5.F) Les réseaux biologiques et propriétés

La théorie des graphes apporte son intérêt dans le cas d'études de phénomènes biologiques depuis un certain nombre d'années. En effet, cette approche trouve de nombreuses applications, que ce soit au niveau des espèces ou des interactions moléculaires. Par exemple, des réseaux écologiques peuvent être développés pour représenter les interactions entre espèces ou des chaînes alimentaires.[131–133] Des graphes intégrant le système nerveux permettent de construire des réseaux neuronaux et de mieux comprendre comment interagissent ces cellules pour permettre au cerveau de réagir au monde extérieur.[134–136] Des graphes peuvent aussi être utilisés en bioinformatique structurale, afin de représenter la structure d'une protéine comme un réseau d'acides aminés reliés par des liaisons covalentes ou secondaires.[137–139] Enfin, nous pouvons aussi citer l'utilisation des réseaux biologiques dans l'étude d'interactions cellulaires, notamment dans le système immunitaire.[140,141]

Là où se trouvent le plus grand nombre de réseaux biologiques est dans les interactions que nous pouvons étudier au niveau des gènes, c'est-à-dire les réseaux de gènes. La notion de gènes regroupe elle-même l'ADN, les transcrits et les protéines, ou bien même les enzymes et les substrats. De nombreuses interactions variées peuvent être listées, comme les complexes protéiques, les régulations inter-géniques, ou bien les récepteurs et leurs ligands. C'est pourquoi les réseaux de gènes sont eux-mêmes divisés en quelques ensembles de types de réseaux, en fonction des informations représentées et de leur utilité (Figure 10). Les réseaux d'interactions protéine-protéine (PPI) regroupent de manière générale toutes les interactions que nous pourrons trouver au niveau fonctionnel, comme par exemple des complexes protéiques.[142] Les réseaux de régulation de gènes regroupent les interactions entre l'ADN et la protéine, ce qui est notamment le cas avec les phénomènes épigénétiques et les facteurs de transcription.[143] Les réseaux de co-expression de gènes impliquent les interactions transcrits-transcrits, c'est-à-dire qu'un niveau d'ARN transcrit augmentera ou diminuera le niveau d'un autre transcrit. Les réseaux PPI peuvent

regrouper tous ces types d'interaction, car il est souvent accepté de simplifier le problème en résumant qu'un gène peut être aux trois niveaux à la fois (ADN, ARN et protéine, Figure 11). Il existe également des réseaux métaboliques, qui regroupent des interactions enzymes-substrats, et des réseaux de voies de signalisation, comme c'est le cas de la base de données KEGG.[113]

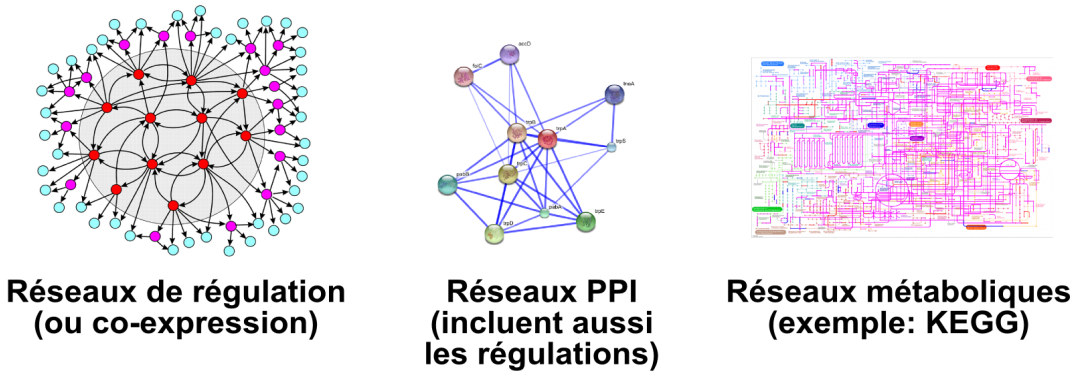


Figure 10: Les différents réseaux de gènes pouvant être construits

Les réseaux de régulation représentent les régulations des gènes au niveau du transcrit principalement. Un réseau de co-expression est un réseau de régulation non dirigé. Les réseaux PPI incluent toutes les interactions protéine-proteine possibles, mais comprennent aussi les régulations. Les réseaux métaboliques ne regroupent que les enzymes et comprennent aussi les substrats.

Les réseaux de gènes dynamiques, notamment ceux de régulation, peuvent être analysés par des outils mathématiques permettant de modéliser des niveaux d'état des gènes (ou nœuds) en fonction du temps. Des modèles très simples, comme la théorie de René Thomas, modélisent l'état d'un nœud avec une variable booléenne, c'est-à-dire qu'un gène est dit allumé ou éteint (exprimé ou réprimé).[144] Les équations différentielles ordinaires (ODE) sont d'une complexité supplémentaire mais sont utiles pour des réseaux issus de mesures prises à de nombreux temps, ce qui est notamment le cas pour les études cinétiques.[145]

Il a souvent été observé qu'un réseau de gènes montre une topologie de type invariant d'échelle, c'est-à-dire que la connectivité suit une loi de puissance indépendante de la taille du réseau.[146,147] Ainsi, la fraction de nœuds avec k voisins directs, appelée P(k), est égale à $k^{-\gamma}$. Le coefficient γ est appelé exposant d'invariance d'échelle. Plus concrètement, cela implique que les gènes se regroupent sous forme de modules fonctionnels, avec souvent des gènes présentant une forte connectivité et jouant le rôle de centres parmi ces modules.[148] Un réseau invariant d'échelle est un

cas particulier du réseau du monde petit, ce dernier étant appelé aussi le paradoxe de Milgram.[149] Ce découpage de réseaux de gènes a souvent fait ses preuves dans la bibliographie, et il est démontré que des gènes présentant une forte connectivité jouent un rôle important dans les mécanismes mettant en place tout le réseau observé, et notamment dans le phénotype étudié.[150–154] Cette topologie particulière n'est pas un fruit du hasard et des théories expliquent qu'elle serait le produit de la sélection naturelle, car grâce à cette structure, la perturbation d'un gène n'aura pas forcément un impact majeur sur l'ensemble du réseau.[155] Cependant, elle présente des points faibles par ces gènes centraux trouvés dans les modules. Les réseaux métaboliques montrent des réseaux invariants d'échelle sous forme de nœud papillon, c'est-à-dire que plusieurs facteurs peuvent être reliés à un seul autre, qui sera lui-même relié à de nombreuses autres cibles.[156]

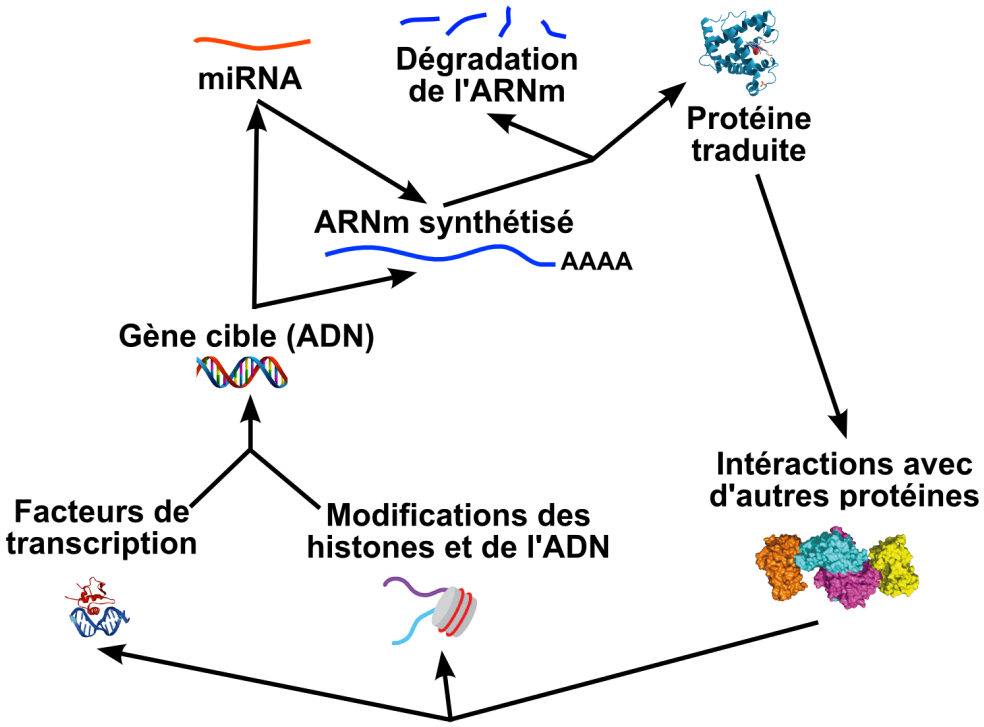


Figure 11: Schéma des principaux mécanismes de régulation d'un gène

Cette figure justifie le fait de simplifier un gène dans un réseau PPI par ses trois principales formes, à savoir l'ADN, l'ARN et la protéine.

Les réseaux biologiques, et plus particulièrement les réseaux de gènes, permettent donc d'entrer plus en détail dans l'étude d'un phénotype et montrent des propriétés très intéressantes. Tout cela représente un outil très utile pour donner une importance particulière à certains gènes, sans

forcément connaître la fonction de tous ceux qui constituent le réseau. Une fois que ces réseaux sont obtenus, il est intéressant de pouvoir les visualiser et les manipuler avec des logiciels intuitifs. Cytoscape est un programme fonctionnant sous Java et qui est très puissant pour la visualisation des réseaux complexes.[157,158] Il possède également l'avantage de proposer de nombreux plugins à installer et qui sont spécialisés dans l'analyse de réseaux de gènes. Notamment, ils permettent d'inclure des annotations fonctionnelles ou de pouvoir importer des réseaux automatiquement par ligne de commande.[159] Cette dernière fonction est particulièrement utile lorsque nous voulons étudier de nombreux réseaux dans une seule session, et cyREST est justement un plugin développé dans cette optique.[160] D'autres outils de visualisation existent permettant de représenter par exemple les réseaux en trois dimensions.[161]

II.5.G) Construction des réseaux de gènes sémantiques

Les réseaux de gènes sémantiques sont construits en se basant sur les interactions connues et listées dans différentes bases de données et dans la littérature. L'objectif de cette approche est de développer des hypothèses sur les voies de signalisation à l'origine du développement de la maladie, et quels gènes peuvent être les plus causaux. Ce sont souvent des outils d'enrichissement de gènes qui permettent aussi de représenter les gènes sous forme de réseau. C'est notamment le cas de g:profiler et la suite de logiciels toppgene, qui regroupent les gènes par différents jeux de gènes et montrent ceux qui se recoupent.[162,163] Il s'agit donc là d'une méthode de visualisation, et non d'un niveau supérieur d'information. InnateDB est un outil très intéressant, car il intègre de nombreux outils d'enrichissement, mais permet également de créer des réseaux regroupant un grand nombre d'interactions inter-géniques mises en évidence dans des publications scientifiques.[164] Ce logiciel web a été développé principalement dans l'optique d'étudier des gènes liés au système immunitaire inné, et il faut donc l'utiliser avec la conscience qu'il ne regroupe pas forcément toutes les interactions connues. D'autres outils plus spécifiques existent, comme GeneMania, qui inclut entre autres des données de structure des protéines et prédit des interactions en se basant sur des domaines partagés.[165] EnrichNet est un outil qui est développé dans le but inverse, c'est-à-dire qu'il permettra de tester l'enrichissement de jeux de gènes, mais en se basant sur des réseaux de gènes connus dans différents tissus et cellules.[166] Nous pouvons donc remarquer que les réseaux de gènes sémantiques et les GSEA ont quelques points communs,

notamment l'objectif de savoir quelles sont les voies de signalisation les plus représentées par une liste de gènes.

Search Tool for Retrieval of INteracting Genes (STRING) est une base de données développée dans le seul but de créer des réseaux PPI sémantiques, et possède donc l'avantage de bénéficier d'une certaine expertise dans le regroupement des interactions connues et de la hiérarchisation de celles-ci.[167,168] Cet outil pourra lister, à partir d'un seul ou une liste de gènes, les interactions connues à partir de diverses sources. Celles-ci peuvent être issues de la paralogie entre espèces, de données expérimentales, de co-expressions observées dans d'autres études de transcrits, de bases de données internes (assez proches des jeux de gènes), ou de text-mining. Ce dernier se basera sur la vraisemblance d'une interaction entre deux gènes par le fait qu'ils soient co-cités dans l'abstract d'une publication. Cette source a encore ses limites, notamment par le fait qu'il existe encore des ambiguïtés sur l'utilisation des abréviations des gènes dans les articles. STRING permet donc d'obtenir différentes interactions, avec un score donné qui représente la probabilité qu'une interaction existe entre ces deux gènes. Un score combiné est aussi calculé pour rassembler toutes les sources d'interactions entre deux nœuds. L'autre avantage particulier de cet outil est qu'il intègre une interface de programmation (API) permettant d'interroger automatiquement la base de données et de créer des réseaux PPI sémantiques, sans avoir à passer par l'interface web. Enfin, STRING permet également d'ajouter un nombre défini d'interacteurs afin de compléter le réseau par des interactions indirectes entre les gènes étudiés.

II.5.H) Inférence des réseaux de gènes gaussiens

Les réseaux gaussiens se basent cette fois-ci sur les corrélations entre les expressions des gènes, et sur l'hypothèse que celles-ci témoignent d'une dépendance directe ou indirecte entre les différents gènes. Cette méthode donnera un réseau biologique de type co-expression, puisque nous inférons l'effet d'un transcrit sur un autre. Le type de mécanisme derrière cet effet est multiple. Il peut s'agir d'un effet direct de la protéine traduite sur le gène cible, dans le cas où la source serait un facteur de transcription. Des effets de silencing par miRNA peuvent être détectés, même si les puces les plus utilisées en intègrent un très faible nombre. Les gènes corrélés peuvent être également issus d'une même voie de signalisation et un même gène ou jeu de gènes serait dans ce cas à l'origine de leur dérégulation au niveau de l'ARN messager. Ces différents mécanismes ne peuvent pas être

déterminés par la méthode d'inférence du réseau en elle-même, mais plutôt par les connaissances des fonctions et des localisations des gènes mis en cause.

La méthode la plus courante et la plus utilisée pour d'autres outils est Weighted Gene Co-expression Network Analysis (WGCNA), qui existe sous forme de package R et qui part d'un calcul des facteurs de corrélations entre les expressions des gènes.[169] Ensuite, ces facteurs de corrélation (principalement des rhos de Spearman) sont convertis en scores de similarité, en utilisant les valeurs absolues ou en centrant autour de 0,5 entre 0 et 1 (pour garder le signe de corrélation). Les scores de similarités peuvent être filtrés directement par un seuil défini, ce qui représente un seuillage dur. Dans un autre cas, il est possible de transformer les scores de similarités en scores de proximité grâce à une fonction de puissance ou sigmoïde, afin d'effectuer un seuillage plus doux. Ensuite, ce package calcule les dissimilarités des nœuds, en se basant sur les connectivités de ceux-ci. Ce calcul permet d'identifier des modules dans le réseau, par une approche de clustering hiérarchique sur une matrice de chevauchement topologique. Enfin, des coefficients de clustering et de connectivités sont attribués aux gènes en fonction de différents facteurs extérieurs définis par le modèle, ce qui permet de définir des gènes qui sont différentiellement connectés entre les différents niveaux du facteur étudié. Ces gènes sont eux-mêmes décrits comme propres à leurs modules, c'est-à-dire qu'ils ont une importance topologique particulière au niveau des clusters. La description des réseaux de régulations par modules fonctionnels est utile pour les réseaux de large taille, mais il faut tenir compte du fait qu'un gène peut appartenir à plusieurs modules fonctionnels, de la même manière qu'un gène peut être lié à plusieurs termes GO.

Cette approche d'inférence de réseaux de co-expression apporte une limite considérable par l'important bruit de fond généré, dû au nombre d'interactions pouvant être inférés qui est largement supérieur aux mesures indépendantes. Il est donc important de pouvoir différentier les interactions entre gènes qui seraient dues à un facteur commun qui est déjà présent dans le réseau, ce qui représente le défi principal de l'inférence des réseaux gaussiens. D'autres méthodes, comme SIRENE ou Aracne, incluent des données extérieures sur les régulations possibles entre les gènes, ce qui apporte une approche d'inférence supervisée des réseaux de régulation.[170,171] Ainsi, les régulations indirectes peuvent être retirées par déduction, si des interactions directes sont déjà

connues. Cependant, cette approche peut retirer des faux négatifs lorsque nous cherchons à identifier de nouvelles voies qui seraient perturbées ou induites spécifiquement par la maladie étudiée. En effet, tout comme pour les réseaux sémantiques, cette approche combinée tient ses limites du fait que les données extérieures servant à la supervision se basent seulement sur ce qui est déjà connu. De plus, ce type de modèle s'applique principalement pour des cellules issues d'espèces simples, comme les bactéries ou les levures. Lorsque nous nous intéressons aux réseaux de régulation issus des cellules de mammifères, les régulations indirectes sont beaucoup plus nombreuses et complexes.[172]

Un autre moyen de s'affranchir des interactions indirectes entre gènes, en ne se basant que sur les expressions des gènes, est d'utiliser les corrélations partielles, c'est-à-dire la corrélation qui existe entre une variable A et B, si la variable C était constante. Cette approche prend donc en compte toutes les variables existant dans le système, c'est-à-dire les gènes du réseau, afin de calculer les dépendances qui existent entre chaque expression. Ainsi, on ne se contente pas de regarder la corrélation entre deux gènes, mais bien l'influence de tous les autres sur celle-ci. Cela aura pour effet de déterminer si la corrélation observée est bien indépendante de toutes les autres mesures.[173] C'est sur cette approche que se base le package R Statistical Inference for MOdular NEtworks (SIMoNe, Figure 12).[174] En plus de ce calcul de corrélations partielles, l'outil utilisera une version revisitée du LASSO graphique, afin d'attribuer des pénalités sur les dépendances entre les gènes en se basant sur une structure modulaire latente du réseau. En plus de cela, ce package propose une sélection parmi plusieurs modèles de réseaux en calculant les scores de qualité à partir du critère d'information bayésienne (BIC) et le critère d'information d'Akaike (AIC). Ces deux méthodes permettent de tester la vraisemblance d'un modèle statistique et de le pénaliser en fonction du nombre de paramètres, afin de satisfaire le critère de parcimonie.[175] Le score BIC dépendra aussi du nombre d'échantillons, contrairement au score AIC. Il s'agit donc d'un outil très intéressant pour inférer des réseaux de co-expression à des états permanents, tout en retirant un certain nombre d'interactions indirectes considérées comme du bruit de fond sans a priori sur les fonctions des gènes. Ce package inclut des inférences de graphes de transition en se basant sur le modèle des vecteurs auto-régressifs de premier ordre (VAR1), à condition d'avoir des données d'expressions sur plusieurs temps rapprochés (de l'ordre de quelques minutes).[176]

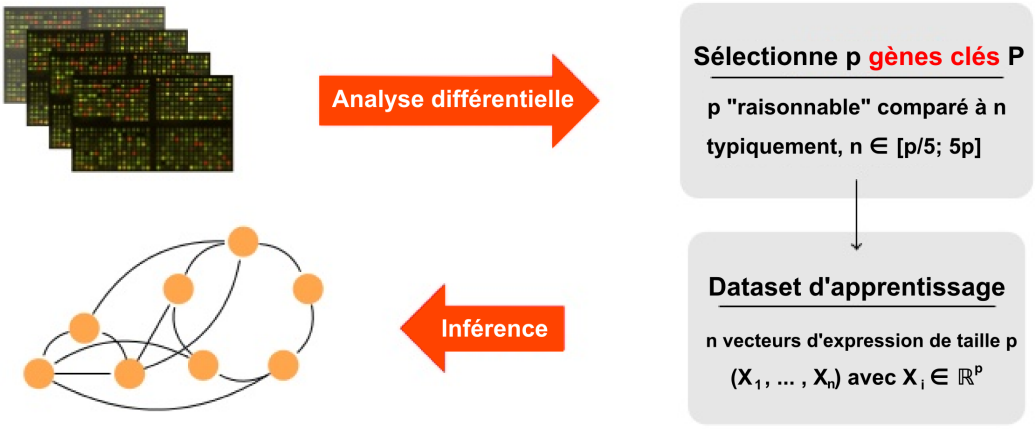


Figure 12: Principe de l'inférence des réseaux de gènes à partir des données d'expression

D'après la présentation du package R SIMoNe (http://www.math-evry.cnrs.fr/logiciels/simone)

II.6) Études de puces transcriptomiques sur la SpA

Grâce à la démocratisation des technologies de puces transcriptomiques sur génome entier, de nombreuses études ont été menées afin d'identifier les changements qui ont lieu au niveau des transcrits chez les patients atteints de SA ou autres sous-types de SpA, dans différents tissus ou cultures cellulaires.

II.6.A) Études des tissus inflammatoires

Comme indiqué précédemment, le choix des organes, tissus ou types cellulaires étudiés est important lors d'études d'expression pour se rapprocher au mieux des mécanismes clés à l'origine de la maladie. Dans le cas de la SA ou de la SpA en général, les articulations inflammatoires ou ankylosées sont les premiers organes touchés. Bien que l'accès et l'extraction d'échantillons à ces endroits est extrêmement difficile, des études incluent des expressions sur génome entier de biopsies de tissus inflammatoires. Cependant, celles-ci n'étaient pas concluantes et non confirmées, compte-tenu principalement de la très faible taille d'échantillons. Une première étude publiée en 2004 a comparé les transcrits de biopsies de tissus synoviaux issus d'articulations périphériques de trois patients atteints de SpA et de trois autres sujets témoins atteints d'arthrose, et a mis en évidence des gènes DE associés aux cytokines et chimiokines.[177] Une autre étude publiée en 2008 s'est intéressée à l'articulation sacro-iliaque, qui joue un rôle plus central et plus

caractéristique de la SA, permettant d'identifier une surexpression de l'IL-7.[178] Une étude de 2006 s'est intéressée aux mécanismes qui peuvent avoir lieu dans l'intestin, plus précisément dans le colon, afin de découvrir ce qui peut être à l'origine du développement des MICI associés à la SpA.[179] La cohorte étudiée regroupait des patients de SpA associés ou non avec une MICI, des patients atteints de la maladie de Crohn et des individus ne présentant aucune de ces pathologies. En plus des 95 gènes DE identifiés, cette étude a trouvé des profils regroupant les MICI avec la maladie de Crohn et les patients SpA sans MICI avec les individus sains, supportant l'idée de mécanismes similaires entre ces différentes pathologies. En revanche, aucune voie de signalisation n'a pu être mise en cause dans les gènes DE déterminés. Une étude plus récente datant de 2013 a comparé le transcriptome de biopsies du tissu synovial du genou chez six patients SpA (dont deux atteints de SA), trois atteints d'ostéoarthrites et quatre individus sains.[180] Un des gènes validés par qPCR et identifié comme étant sous-exprimé est DKK3, un inhibiteur de la voie de signalisation Wnt. Les autres gènes DE confirmés par qPCR comprennent MMP3 et PTGER4 étant sur-exprimés chez les patients.

II.6.B) Études sur le sang total

D'autres études transcriptomiques ont été menées sur le sang total, dont l'extraction est beaucoup plus aisée et qui peut aussi représenter les mécanismes inflammatoires mis en jeu par le système immunitaire. Une première étude publiée en 2009 a regroupé 18 patients atteints de SpA et 25 témoins sains.[181] Elle a ensuite fait l'objet d'une étude de réplication dans une deuxième cohorte, ce qui conforte beaucoup les résultats d'analyses. Une vingtaine de gènes ont été identifiés comme DE, comprenant deux récepteurs à l'IL-1, MAPK14, le récepteur des cellules NK NCR3, ainsi que 4 gènes impliqués dans le remodelage osseux (dont BMP6 et ostéonectine). Cette étude a permis de mettre en évidence une signature immune de la SA. Une autre étude parue en 2011 regroupait 16 patients atteints de SA et 14 individus sains, ainsi que des patients atteints d'autres maladies inflammatoires tels que le lupus érythémateux et la sclérodermie systémique.[182] 83 gènes DE ont pu être ainsi identifiés, dont certains étant liés aux voies de signalisation de la maturation des cellules dendritiques, NF-κB et TREM1. Par ailleurs, TLR4 et TLR5 étaient surexprimés chez ces patients et ces résultats ont pu être confirmés dans un second échantillon. Une autre étude de la même année comparant 18 patients SA à 18 témoins ont identifié 221 gènes DE, parmi lesquels

14 ont pu être confirmés dans une étude de réplication.[183] Parmi ceux-ci nous pouvons citer l'ostéonectine (également identifié dans une autre étude) et EP300, impliqués tous les deux dans le métabolisme du cartilage et des os. En 2013, une autre équipe a repris ces données pour construire un réseau PPI et a priorisé 11 gènes reliés à la SA par analyse topologique, comprenant XRCC6, SMAD3, HDAC1, SHC1, TP53, FYN, ZAP70, MEN1, RUNX3, TUBB et POLR2A.[184] Une autre étude publiée en 2010 était conçue dans le but de déterminer les gènes à l'origine de la réponse à un traitement, et notamment aux anti-TNF.[185] Dans ce type d'étude, il n'est pas rare d'identifier un nombre important de transcrits, car des traitements ont tendance à perturber considérablement le motif d'expression de la cellule.[186–189] Ainsi, ce ne sont pas moins de 1428 gènes qui ont été identifiés comme DE en réponse à ce traitement, ce qui confirme qu'il est important de prendre en compte la prise d'un traitement lors d'études du transcriptome. 4 gènes parmi ceux-là ont été confirmés par qPCR, dont LIGHT, pour lequel la quantité de protéine est également corrélée au taux de CRP.

II.6.C) Études sur des PBMCs

Cependant, les ARNs issus du sang total représentent un nombre très important de types cellulaires, et c'est donc très compliqué de déterminer quelle est l'origine de la voie de signalisation observée. C'est pourquoi d'autres études ont été menées sur des cellules mononucléées du sang périphérique (PBMCs), ce qui permet de discriminer les granulocytes et de pouvoir cibler l'ensemble des lymphocytes, les cellules NK et les monocytes, ces cellules reflétant plus le système immunitaire que le sang total. Une étude publiée en 2002 a étudié le transcriptome de PBMCs extraits de patients atteints de SpA, en comparant avec des individus sains et des personnes atteintes d'arthrite psoriasique et de PR.[190] L'analyse des données a mis en évidence des gènes DE impliqués dans les trois types de RIC, mais un seul a pu se différencier des autres : l'antigène de différentiation nucléaire des myéloïdes (MNDA), impliqué entre autre dans la réponse à l'interféron par la lignée des monocytes.[191] Une autre étude de 2009 comparant des patients atteints de SA ou de SpA indifférenciée avec des témoins, a identifié le gène codant pour le régulateur 1 de la signalisation de la protéine G (RGS1) comme biomarqueur potentiel par sa surexpression chez les patients.[192] Ce gène est activé par le TNF-α et l'IL-17, tous les deux étant des cibles thérapeutiques de la SA. Une étude publiée en 2010 a comparé 18 patients SA et 18

témoins appariés sur l'âge et le sexe, et a mis en évidence la sous-expression de trois gènes dans les PBMCs confirmés dans une étude de réplication : NR4A2, TNFAIP3 et CD69.[193] Ces gènes présentent une pertinence particulière par leur rôle dans l'inflammation ou l'immunité.

II.6.D) Études sur des cellules isolées de PBMCs

Les PBMCs représentent de nouveau un groupe de types cellulaires et peuvent être discutables quant à la spécificité des régulations du transcriptome identifiées. C'est pourquoi d'autres études se sont intéressées à isoler des cellules spécifiques sur les PBMCs, grâce à des techniques de tri cellulaire. Une première étude parue en 2008 relate l'analyse du transcriptome de macrophages dérivés à partir des monocytes isolés des PBMCs de 8 patients SA et de 9 individus sains.[194] Ces cellules ont ensuite été stimulées avec de l'IFN-γ et du LPS, afin de reproduire un contexte inflammatoire in vitro. L'analyse du transcriptome de ces cellules a permis d'identifier une signature IFN-γ inversée, qui est de nouveau supprimée par le traitement de ces macrophages avec de l'IFN-γ exogène. Cette signature est proposée comme étant à l'origine de l'augmentation des Th17 observée dans la SA, puisque cet interféron favorise une différenciation inverse vers les Th1. Une autre étude a trouvé des résultats très similaires lors de l'étude du transcriptome des CPAs du modèle animal rat transgénique pour HLA-B27.[195] Une méta-analyse de ces résultats par une autre équipe en 2013 a utilisé une approche par inférence d'un réseau de régulation des gènes gaussien en calculant les coefficients de corrélation de Pearson entre les facteurs de transcriptions et les gènes DE, soulignant l'implication de NFKB1, STAT1, STAT4, TNFSF10, IL2RA et IL2RB.[196]

Par ailleurs, comme décrit plus loin dans l'introduction(Implication des cellules dendritiques dans la SpA), les cellules dendritiques (DCs) sont fortement suspectées comme jouant un rôle central dans la SA. C'est pourquoi une autre étude du transcriptome s'est intéressée aux cellules dendritiques dérivées de monocytes (MD-DCs) de 9 patients atteints de SpA axial et de 10 individus sains.[197] Cette analyse s'est portée sur des puces Affymetrix HumanGene 1.0 st, qui regroupent environ 23 000 transcrits. Les MD-DCs étaient stimulées ou non par du LPS pendant 6h ou 24h. Le LPS a comme propriété d'imiter la paroi bactérienne et d'induire une réaction immunitaire non spécifique proche de l'inflammation. Cette étude a permis d'identifier des différences d'expression au niveau de la voie Wnt co-régulée par CITED2 sous-exprimé. Par

ailleurs, l'expression de CITED2 est inversement corrélée avec celle d'ADAMTS15, une métalloprotéinase (MMP). Les MMPs ont été démontrées comme associées à la susceptibilité et la sévérité de la SpA.[198] En plus de ces deux gènes confirmés par qPCR s'ajoutent F13A1 et SELL, respectivement un facteur de coagulation et une protéine d'adhésion de surface des cellules. Une autre publication de 2015 met en évidence un eQTL entre l'haplotype de trois SNPs (rs17482078, rs10050860 et rs30187) et l'expression d'ERAP1 dans des MD-DCs d'une cohorte de 23 patients SpA et 44 témoins.[199]

II.6.E) Une méta-analyse avec réseau PPI

Une étude très intéressante publiée en 2012 a effectué une méta-analyse considérable en regroupant des gènes causaux identifiés comme reliés à la SpA dans OMIM (HLA-B, TNFA, IL23R, CYP2D6, TNFSF13, TNFSF13B, B2M et COL2A1), une étude du protéome de monocytes de patients atteints de SA[200], et deux études du transcriptome (dont une sur la forme de SpA juvénile)[181,201], afin de construire un réseau PPI avec la base de données STRING et d'identifier des modules fonctionnels liés à la SpA.[202] Cette étude a permis de refléter le lien entre la SpA et l'inflammation médiée par le système immunitaire, ainsi qu'avec le remodelage perturbé de l'os causant une néo-formation ou une perte osseuse. Ils ont notamment priorisé 7 gènes, qui sont PTGS2, MMP2, MAPK3, HRAS, EGF, NFKB1 et TNF.

II.6.F) Bilan des études du transcriptome

Nous pouvons remarquer que les études ciblées sur le transcriptome de patients atteints de SA ou de SpA sont très variées, et un faible recouvrement des résultats est observé (Tableau 4). Il semble cependant que les voies du système immunitaire et de signalisation Wnt ont un rôle essentiel dans le développement de la maladie, ainsi que celles du remodelage de l'os. Par ailleurs, un nombre très faible d'études s'intéresse aux gènes DE spécifiquement dans les DCs, qui sont tout de même suspectées comme étant centrales et qui expriment les gènes du CMH (y compris HLA-B27). A l'heure actuelle, aucune étude transcriptomique n'a développé un design expérimental permettant de discriminer les effets spécifiques de HLA-B27, par exemple en incluant des individus sains porteurs de cet allèle. De même, très peu d'études se sont intéressées à l'étude des gènes sous forme de réseau, ce qui peut pourtant être un outil très intéressant pour étudier des gènes à l'origine de maladies multifactorielles comme la SpA.

Auteurs, date	Type de puce	Nombre de sondes / EST / transcrits	Nombre de patients/témoins	Tissus ou cellules étudiées	Nombre de gènes DE
Gu et al., 2002[190]	Atlas Human Array 7740–1 Nylon	588 transcrits	7 SpA, 6 PsA et 6 PR/7	PBMCs	4 SpA vs T 4 PsA vs T 3 SpA vs PsA
Rihl et al., 2004[177]	Atlas Human 1.2 array Nylon Array	1185 EST	3 SpA/4	Biopsies tissus synoviaux	18
Laukens et al., 2006[179]	Type VII silane-coated slides Array	74 828 EST	15 SpA/10	Biopsies colon	464
Rihl et al., 2008[178]	Atlas Human 1.2 array Nylon Array	1185 EST	3 SpA/4	Fluide de l'articulation sacro-iliaque	47
Smith et al., 2008[194]	Affymetrix U133 Plus 2.0 GeneChip	47 000 transcrits	8 SA/9	Macrophages dérivés de monocytes	141
Gu et al., 2009[192]	Sentrix Human Ref-8_v2 Beadchips	20 000 sondes	46 USpA et 44 SA/46	PBMCs	4 SA vs T 38 USpA vs T
Sharma et al., 2009[181]	Affymetrix U133 Plus 2.0 GeneChip	47 000 transcrits	18 SpA/25	Sang total	107
Duan et al., 2010[193]	Illumina HT-12 BeadChips	48 000 transcrits	18 SA/18	PBMCs	452
Haroon et al., 2010[185]	Affymetrix U133 Plus 2.0 GeneChip	47 000 transcrits	16 SA (effet traitement)	Sang total	1428
Assassi et al., 2011[182]	Illumina Human BeadChips	24 000 sondes	16 SA/14	Sang total	83
Pimentel-Santos et al., 2011[183]	Illumina HT-12 BeadChips	48 000 transcrits	18 SA/18	Sang total	221
Gethin et al., 2013[180]	Illumina Whole-Genome DASL	24 000 transcrits	6 SpA, 2 SA et 3 OA/4	Biopsies tissu synovial du genou	416
Zhu et al., 2013[196]	Méta-analyse de l'étude de Smith et al. De 2008	47 000 transcrits	8 SA/9	Macrophages dérivés de monocytes	16 (réseau de régulation)
Talpin et al., 2014[197]	Affymetrix HumanGene 1.0 st	23 000 transcrits	9 SA/10	MD-DCs	81
Zhao et al., 2012[202]	Méta-analyse de 3 études (ARN + protéine)	Ajout des gènes dans la base OMIM	GSE1402 : 18 SpA/25 GSE18781 : 11 SpA/12	PBMCs	380

Tableau 4: Synthèse des études incluant des données de puces transcriptomiques

III/ Implication des cellules dendritiques dans la SpA

III.1) Rappels sur les cellules dendritiques

Les cellules dendritiques (DCs) sont des CPAs professionnelles, c'est-à-dire qu'elles expriment de manière constitutive le CMH de classe II. Elles portent ce nom par leur structure de la membrane cellulaire particulière, qui se présente sous forme de dendrites, de manière assez semblable aux neurones. Leur rôle est de capter les antigènes présents dans les tissus, afin de les traiter pour les présenter aux lymphocytes T dans les organes lymphoïdes. Une synapse immunologique se met en place, ayant pour effet d'induire une réponse globale dans tout l'organisme contre l'antigène présenté, avec une activation et une amplification clonale des lymphocytes T effecteurs. Avant l'activation des lymphocytes T, les DCs immatures deviennent matures, et expriment donc fortement les molécules de CMH de classe I et II et des molécules de co-stimulation.

Il est possible d'obtenir des cellules dendritiques dérivées de monocytes en cultivant ces dernières pendant 7 jours avec du GM-CSF et de l'IL-4.[203] Il est donc facile de les étudier à partir des PBMCs, les DCs naturelles étant plus difficiles à prélever et isoler chez les patients. De plus, une analyse transcriptomique parue en 2013 dans Immunity a démontré que les DCs inflammatoires partageaient des signatures génétiques avec les MD-DCs obtenues in vitro.[204]

III.2) Modèle du rat HLA-B27 et les DCs

En plus des études sur des individus patients et témoins, notre laboratoire étudie les mécanismes biologiques à l'origine de la SpA grâce à un modèle animal, le rat transgénique HLA-B27. Ce modèle a été développé en 1990 en introduisant dans des rats le transgène du HLA-B2705 et de la β_2m.[205] A partir de 40 copies du transgène, ce modèle animal développe spontanément au bout de 6 semaines des symptômes proches de la SpA, comme les arthrites et le psoriasis, avec une prévalence prononcée de MICI proche de la maladie de Crohn. Ce développement de la maladie chez le rat est bien spécifique de l'allèle HLA-B27 et n'est pas le fruit d'une surexpression d'un autre allèle du CMH de classe I.[206] Ce modèle rat a permis de développer de nombreuses hypothèses sur l'étiologie de la SpA, notamment sur l'implication du microbiote. En effet, des rats élevés en condition sans germe pathogène, en environnement stérile, ne développeront plus la maladie. Les inflammations se manifestent à nouveau lorsque la flore bactérienne intestinale est reconstituée.[207] Par ailleurs, les rats transgéniques HLA-B27 ont permis de démontrer un rôle joué

par des cellules d'origine hématopoïétiques exprimant fortement ce transgène et par les lymphocytes T CD4+.[208,209]

Parmi les cellules hématopoïétiques mis en cause qui sont les meilleurs candidats, se trouvent les CPA professionnelles (notamment les DCs), les lymphocytes B activés et les monocytes/macrophages. Le modèle du rat a mis en évidence un rôle important joué par les DCs dans le développement de la SpA. Notamment, les DCs extraits de ces rats présentent un défaut d'activation des lymphocyte T CD4+ allogéniques ou syngéniques, qui est néanmoins plus un effet spécifique de HLA-B27 que de la maladie.[210] Une diminution de la formation d'une synapse immunologique stable est observée entre les DCs et les lymphocytes T de rats HLA-B27, mettant en cause des molécules de co-stimulation comme CD86/CD28, indépendamment de l'immaturité des DCs ou de facteurs inhibant l'activation des lymphocytes T.[211,212] Trois groupes fonctionnels de gènes ont été identifiés comme DE au niveau des protéines dans les DCs spléniques de rats HLA-B27, qui se divisent ainsi : les protéines impliquées dans le traitement de l'antigène et la réponse UPR, celles dans la mobilité du cytosquelette, et enfin dans la synthèse du CMH de classe II. De plus, ces cellules présentent un défaut de mobilité, une augmentation de l'apoptose, ainsi qu'une diminution de l'expression des molécules du CMH de classe II.[213]

Ces études laissent donc supposer fortement que les DCs jouent un rôle important dans le développement de la SpA par leur forte expression de HLA-B27, induisant une rupture de la tolérance immunitaire médiée par les lymphocytes T CD4+, et ainsi une réponse immunitaire croisée contre les bactéries du tractus intestinal.

Objectifs et présentation du projet de thèse

Notre laboratoire a comme axe de recherche principal l'exploration de l'étiologie de la SpA. C'est pourquoi plusieurs approches sont utilisées en parallèle, combinant à la fois des études sur le modèle animal (rat transgénique HLA-B27) et les facteurs associés chez l'Homme. De plus, par un rattachement avec l'Hôpital Ambroise-Paré de Boulogne-Billancourt, une cohorte considérable de plus de 3600 personnes est à disposition pour effectuer des analyses sur génome entier ou d'autres études nécessitant des prélèvements. En outre de l'analyse du microbiote des patients par des partenaires, des études in vitro sont également développées pour étudier les mécanismes cellulaires à l'origine de la SpA. C'est cette richesse de moyens développés pour explorer tous les facteurs pouvant être à l'origine de cette pathologie, qui donne une approche méritant sa place dans le domaine de la biologie intégrative.

Ce projet se divise principalement en deux volets, afin d'aborder une approche complémentaire et de pouvoir développer de nouvelles hypothèses sur la physiophatologie des facteurs génétiques associés à la maladie. Le premier comprend une analyse de liaison de puces de génotypage sur une cohorte de 1310 personnes réparties dans 210 familles afin d'identifier de nouveaux facteurs pouvant expliquer l'hétérogénéité de la maladie. Le deuxième volet implique une étude multivariée de données de puces transcriptomiques de MD-DCs stimulées ou non par le LPS, et issues de patients atteints de SpA axial porteurs de l'allèle HLA-B27 et d'individus sains porteurs ou non de cet allèle. L'analyse des gènes DE s'est prolongée par une approche en réseaux, afin d'identifier d'autres voies de signalisation affectées par la maladie et de prioriser les gènes pour des études fonctionnelles. Ce projet s'est développé dans un effort de réflexion inter-disciplinaire, et dans l'esprit de rester le plus objectif et sans a priori sur les résultats des analyses statistiques. L'aspect le plus innovant de cette étude est d'intégrer une inférence des réseaux de gènes à partir de données d'expression propres au laboratoire pour cette maladie, et d'analyser une cohorte permettant de discriminer l'effet de l'allèle HLA-B27. Nous verrons que cette approche ouvre de nouvelles voies sur la compréhension du développement de cette pathologie et sur les futures validations et explorations fonctionnelles des cellules de l'immunité. La SpA est une maladie particulièrement intéressante à étudier, que ce soit par sa forte prévalence, ou sa nature multifactorielle qui est tout de même simplifiée par son association avec l'allèle HLA-B27.

Chapitre 2

MATÉRIELS ET MÉTHODES

Analyse de liaison des SNPs associés à la SpA

I/ Cohorte, puces de génotypages

I.1) Cohortes étudiées

Les études de génotypage présentées dans ce manuscrit se sont basées principalement sur trois cohortes, qui ont été par la suite fusionnées pour plus de puissance statistique (Tableau 5). Les individus, majoritairement d'origine caucasienne, étaient recrutés par le GFEGS, et les études qui les incluent étaient approuvées par les comités locaux d'éthique de l'Institut Cochin de Paris et de l'hôpital Ambroise-Paré de Boulogne-Billancourt. Tous les individus des trois cohortes ont donné leur consentement écrit. Les critères Amor, ESSG et ASAS ont servi de diagnostic pour classer les patients des témoins. Le BASDAI moyen des patients dans les trois cohortes était d'environ 34,38, avec un écart-type d'environ 21,51. Les âges moyens correspondaient à l'âge des individus lors de leur dernier suivi.

Cohorte	Individus (familles)	Nombre de patients/témoins	Nombre d'hommes/femmes	Age moyen (± SD)	Nombre de HLA-B27+/-	Nombre de SA/USpA
Cohorte 1	956 (154)	480/465	469/487	50,42 (±16,09)	646/308	272/135
Cohorte 2	319 (56)	169/141	154/165	51,30 (±17,92)	200/116	68/68
Cohorte 3	35 (12)	19/14	20/15	49,51 (±17,00)	15/16	7/8
Total	1310 (210)	668/620	643/667	50,61 (±16,57)	861/440	347/211

Tableau 5: Caractéristiques des trois cohortes étudiées pour le génotypage sur génome entier

I.1.A) Cohorte 1

La cohorte 1 regroupait 956 individus génotypés, comprenant 480 patients et 465 témoins. Parmi les patients, 242 étaient des hommes, 238 des femmes, 35 ne portaient pas l'allèle HLA-B27 et 445 sont HLA-B27 positifs. Les témoins étaient répartis avec 221 hommes, 244 femmes, 267 HLA-B27 négatifs et 197 HLA-B27 positifs. Les témoins étaient principalement des individus sains, sauf 1 atteint de PR. Parmi les patients, 272 étaient atteints de SA, 105 de psoriasis, 62 de rhumatisme, 17 de colite ulcérative, 16 de la maladie de Crohn, 3 d'arthrite réactionnelle, et 135 avaient une USpA. 128 patients présentaient plusieurs sous-types de SpA.

I.1.B) Cohorte 2

La cohorte 2 représentait 319 individus issus de familles différentes de la première cohorte,

comptant 169 patients et 141 témoins. Les patients se recoupaient entre 91 hommes et 78 femmes. 29 patients étaient HLA-B27 négatifs, contre 140 HLA-B27 positifs. 60 témoins étaient des hommes pour 81 témoins femmes. 83 portaient l'allèle HLA-B27, contre 57 non porteurs. Les patients comptaient autant d'atteintes de SA que d'USpA, c'est-à-dire 68. 32 avaient développé un psoriasis, 23 un rhumatisme, 9 une maladie de Crohn, et 2 une colite ulcérative. 38 patients regroupaient plusieurs sous-types de SpA. Tous les témoins dans cette cohorte étaient des individus sains.

I.1.C) Cohorte 3

La cohorte 3 comptait 35 individus issus des mêmes familles que les cohortes 1 et 2, et qui ont été ajoutés car ils présentaient un intérêt particulier dans l'analyse familiale. Cette cohorte regroupait 19 patients et 14 témoins. 14 patients étaient des hommes pour 5 femmes. Du côté des témoins, 5 étaient des hommes pour 9 femmes. Parmi les patients, 8 étaient HLA-B27 négatifs et 11 portaient cet allèle. 7 témoins étaient HLA-B27 négatifs pour 4 HLA-B27 positifs. Tous les témoins étaient des individus sains. 7 patients étaient atteints d'une SA, 4 d'un psoriasis, 3 d'un rhumatisme, 2 d'une MICI (1 atteint de maladie de Crohn et 1 de colite ulcérative), et 8 d'USpA. 5 patients étaient atteints d'au moins deux sous-types de SpA.

I.2) Puces de génotypage

Les individus des cohortes citées ci-dessus ont été génotypés en utilisant des puces Affymetrix 250K qui regroupent environ 262 000 SNPs répartis le long du génome. Le génotypage était effectué au Centre National de Génotypage (CNG) d'Evry. L'ADN total était digéré par l'enzyme NspI, ligaturé à l'adaptateur et amplifié par PCR. Les amplicons issus des produits PCR purifiés ont ensuite été quantifiés, fragmentés, marqués et hybridés aux puces 250K. L'interprétation du signal des puces en matrices d'allèles attribués pour chaque individu et chaque SNP a été déterminée par l'algorithme BRLMM.[214]

II/ Nettoyage et annotations des données

II.1) Vérification des échantillons

Une fois les données de génotypage rassemblées pour tous les individus et par chromosome, une vérification des erreurs d'échantillonnage était effectuée grâce à une inférence des liens de parenté

par graphical relationship representation (GRR), un logiciel avec une interface graphique développé pour Windows. Cette étape de vérification était effectuée après un échantillonnage de 20 000 SNPs tout le long du génome.

II.2) Mise à jour des annotations des SNPs

Après que les pedigree aient été corrigés, une mise à jour était effectuée vers la version 132 de dbSNP quand c'était nécessaire, notamment pour une majorité des génotypages issus de la première cohorte qui étaient sous la version 129. Cette mise à jour a entraîné une modification des localisations des SNPs, puisque le séquençage du génome humain était passé de la version hg18 à hg19. Ensuite, les variants représentant plusieurs allèles ou qui n'étaient pas assez spécifiques ont été retirés. 5 587 SNPs ont ainsi été retirés entre les deux versions : 5 188 par leur nature multi-allélique ; 203 ayant été supprimés directement par dbSNP ; 96 n'étant pas assez spécifiques.

II.3) Vérification des SNPs

Une série de nettoyages d'erreurs était ensuite effectuée au niveau des SNPs par le logiciel PLINK.[215] Tout d'abord, les SNPs avec une MAF inférieure à 1 % ont été retirés. Ensuite, les variants étaient soumis à un test exact d'équilibre de HW.[216] Les SNPs avec une p-value de ce test supérieure à 10^{-3} ont été retirés. Les erreurs mendéliennes basiques ont été retirées. Enfin, les SNPs avec un taux de génotypage inférieur à 95 % à la suite de ces nettoyages ont été supprimés des études de génotypage.

Les données nettoyées étaient de nouveau soumis à un test d'erreurs mendéliennes plus subtiles après reconstruction des haplotypes grâce au logiciel Merlin.[35] Cette étape de nettoyage était répétée plusieurs fois, jusqu'à ce qu'aucune erreur ne soit détectée. Une fois tous ces nettoyages effectués, un nouveau filtrage sur les taux de génotypage des SNPs supérieurs à 95 % était initié.

III/ Analyses de liaison

III.1) Analyses de liaison non paramétriques

Les analyses de liaison non paramétriques ont été effectuées sous le logiciel Merlin, avec un lissage de 0,5 cM pour prendre en compte le DL. Systématiquement, un format de sortie additionnel affichant les LOD scores par famille était demandé. Lorsque le chromosome X était analysé, la commande minx était entrée, une version de Merlin spécialisée dans la prise en compte

des régions non pseudo-autosomales. Pour quelques analyses, des blocs de DL ont été calculés en plus par le logiciel en indiquant un seuil de coefficient de corrélation r^2 de 0,1. Ces blocs étaient réutilisés pour d'autres analyses ultérieures sur les mêmes chromosomes.

III.2) Analyses de liaison paramètriques

Les analyses de liaison paramétriques étaient effectuées également sous Merlin. Les modèles utilisés se basaient tous sur la prévalence de la maladie en France estimée à 0,43 %.[6] Les modèles principalement utilisés se basaient sur l'hypothèse d'une transmission dominante, récessive, co-dominante, ou en prenant en compte que la maladie était fortement associée à l'allèle HLA-B27. Pour chaque SNP, un HLOD score était calculé avec la proportion α de familles liées à celui-ci. Le ratio α pour le SNP avec un HLOD score maximal était utilisé pour sélectionner les familles les plus liées au locus à risque, afin de visualiser la transmission des haplotypes dans les pedigree correspondants.

III.3) Analyses de liaison de la protection

Afin de tester la liaison des locus protecteurs, nous avons inversé les statuts, en transformant tous les patients en témoins. Ensuite, tous les témoins ou seulement les témoins HLA-B27 positifs étaient transformés en patients, en fonction de l'approche utilisée. Par la suite, seules les familles avec au moins deux témoins inversés étaient sélectionnées pour les analyses de liaison, car considérées comme informatives.

IV/ Ressources informatiques

Les étapes de nettoyage d'erreurs mendéliennes subtiles et d'analyses de liaison sous Merlin demandaient une reconstruction des haplotypes, et donc des temps de calculs et des ressources informatiques considérables. C'est pourquoi, pour ces étapes, nous disposions de deux serveurs DELL connectés en réseau local dans l'UVSQ, les deux disposant de 397 Go de RAM, respectivement de 16 et 25 To d'espace disque, et de processeurs Intel Xeon 32 cœurs 2,9 GHz. Afin de pouvoir y accéder de l'extérieur du réseau de l'Université, un système de tunnel ssh par un autre routeur était disponible.

Analyses du transcriptome de patients

I/ Approche fonctionnelle

I.1) Cohortes étudiées

Tout comme pour les études de génotypage, les cohortes étudiées pour les analyses du transcriptome se découpaient en plusieurs groupes (Tableau 6). Celles-ci représentaient deux études, appelées ici étude 1 et étude 2. Au total, ces deux groupes rassemblaient 68 personnes, avec 23 patients HLA-B27 positifs, 24 témoins HLA-B27 positifs et 21 témoins HLA-B27 négatifs. Le BASDAI moyen des patients était de 39,15 avec un écart-type de 20,94. Ils répondaient tous aux critères ASAS de la SpA axial. Tous les participants ont donné leur consentement écrit et les études étaient approuvées par le comité local d'étique de l'Ile-de-France XI (Saint-Germain-en-Laye).

Cohorte	Nombre d'individus	Nombre de patients	Nombre de témoins HLA-B27+/-	Nombre d'hommes/femmes	Age moyen (± SD)	BASDAI moyen (± SD)
Etude 1	19	9	0/10	7/2	42,89 (± 9,94)	53,15 (± 22,14)
Etude 2	49	14	24/11	8/6	48,64 (± 7,69)	34,49 (± 19,21)
Total	68	23	24/21	15/8	46,39 (± 8,90)	39,16 (± 20,94)

Tableau 6: Caractéristiques des deux cohortes étudiées pour les analyses du transcriptome

Les nombres d'hommes/femmes, les âges et le BASDAI moyens ne se basent ici que sur les patients.

I.1.A) Etude 1

La cohorte de l'étude 1 reprennait les données d'expressions publiées en 2014 par notre laboratoire lors d'une étude de découverte sur des MD-DCs.[197] Cette cohorte rassemblait 9 patients atteints de SA (dont deux avec un psoriasis) issus de familles indépendantes. Le ratio homme/femme des patients était de 77,78 %. Leur âge moyen lors du dernier suivi était de 42,89, avec un écart-type de 9,94. Ils étaient tous porteurs de l'allèle HLA-B27. Cette étude regroupait aussi 10 individus sains comme témoins, qui étaient des donneurs de sang prélevés par l'Etablissement Français du Sang (EFS).

I.1.B) Etude 2

La cohorte de l'étude 2 était une réplication de l'étude 1, avec un recrutement plus poussé permettant de faire des analyses plus fines. Globalement, celle-ci regroupait 14 patients HLA-B27

positifs, 24 témoins HLA-B27 positifs et 11 témoins HLA-B27 négatifs. Parmi les patients, 13 étaient appariés au 1er degré avec des témoins HLA-B27 positifs, ce qui donnait 13 paires de germains. Parmi les patients, 9 étaient atteints d'une SA, 4 d'une USpA, et 1 présentait à la fois les symptômes d'un psoriasis et d'un rhumatisme. Le ratio hommes/femmes des patients était de 57,14 %, avec un âge moyen lors du dernier suivi de 48,64 (écart-type de 7,69). Les 24 témoins étaient tous des individus sains, une partie était prélevée par l'EFS.

I.2) Isolations, Cultures et stimulations des cellules

Le même protocole de manipulations cellulaires était utilisé dans les deux cohortes, afin de garantir une homogénéité des résultats. Les PBMCs étaient isolées à partir de 50 mL de sang par centrifugation de gradient de densités Ficoll (technologies STEMCELL). Ensuite, les monocytes étaient isolés par tri magnétique des cellules avec des billes recouvertes d'anticorps monoclonaux anti-CD14 (BD Imag). La cytométrie de flux a permis de déterminer que les monocytes triés présentaient une homogénéité morphologique avec 99 % de cellules CD14 positives.

Les monocytes étaient ensuite cultivés pendant 6 jours dans des plaques 24 puits (400 000 cellules/500 µL) dans du milieu RPMI 1640 supplémenté avec 10 % de sérum de veau inactivé, 100 U/mL de pénicilline, 100 µg/mL de streptomycine, 500 U/mL de GM-CSF recombinant humain et 500 U/mL d'IL-4 recombinant humain (AbCys S.A., France). A la suite de cette différenciation, les MD-DCs étaient stimulées ou non par du LPS d'E. Coli (Sigma-Aldrich, St Louis, MO) à une concentration de 100 ng/mL pour les dernières 6 ou 24 heures de culture. Les temps de stimulation LPS sont référés dans ce manuscrit comme H0, H6 et H24. Seuls les MD-DCs ont servi pour les études transcriptomiques décrites par la suite.

Bien que les patients prenaient un traitement pendant leurs prélèvements, nous estimions que les 6 jours de culture cellulaire, additionnés à la différentiation, annulaient les effets du traitement sur l'expression globale des gènes.

I.3) Isolation des ARNs et puces transcriptomiques

Les MD-DCs étaient rompues et homogénéisées en utilisant du tampon RLT (Qiagen, Valence, CA). L'ARN total était ensuite isolé en utilisant le Mini Kit RNeasy Plus (Qiagen, Valence, CA). La quantité et la qualité de l'ARN étaient évaluées par le Bioanalyseur Agilent 2100 (Agilent,

Santa Clara, CA). Seuls les échantillons d'ARN avec un score d'intégrité de l'ARN (RIN) supérieur à 8 étaient utilisés pour la suite.

L'ARN était ensuite rétro-transcrit, converti en ARNc biotinylé par le protocole standard Affymetrix (Affymetrix, Santa Clara, CA), et hybridé sur des puces à ADN Affymetrix Human Gene 1.0 st, regroupant environ 23 000 transcrits tout le long du génome. Un protocole de routine était ensuite établi par la plateforme de génomique de l'Institut Cochin, pour convertir les signaux des spots en données d'expressions pour chaque sonde.

II/ Analyses des gènes DE

II.1) Normalisation et remplissage des données manquantes

Les données étaient normalisées par la méthode RMA fournie par le package affy de la librairie R Bioconductor. Les sondes étaient convertis en identifiants Ensembl grâce aux fichiers de définition des puces personnalisés fournis par le site internet de Brain Array.[217] Après normalisation, des matrices d'expressions étaient obtenues avec chaque colonne correspondant à un échantillon et chaque ligne correspondant à un gène.[218] Chacun des identifiants Ensembl était ensuite annoté en se connectant à la base de données biomaRt par le package R correspondant, afin de récupérer les symboles HGNC et les données de localisation dans le génome.[219]

Par la suite, les données de quelques temps de stimulation LPS manquaient pour certains individus : 1 patient pour H24 ; 1 patient et 1 témoins HLA-B27 négatif pour H6 ; et 1 patient, 1 témoin HLA-B27 positif et 1 témoin HLA-B27 négatif pour H0. Afin de remplir les données manquantes et d'assurer une analyse homogène, nous avons ajouté la différence moyenne de tous les échantillons entre le temps initial et celui manquant, aux données d'expressions du temps non manquant. Par exemple, lorsque le temps H0 manquait, nous ajoutions la différence moyenne entre H6 et H0 au temps H6.

II.2) Méta-analyses des deux études du transcriptome

Afin d'étudier la réplication des résultats entre nos deux cohortes, nous avons comparé les expressions des gènes entre patients HLA-B27 positifs et HLA-B27 négatifs dans les deux cohortes. Nous avons ensuite représenté les croisements des deux listes par un diagramme de Venn pour mesurer le taux de discordance entre les deux études. Afin d'être plus sensible, nous

avons également utilisé la méthode maxP du package R MetaDE, en filtrant sur la p-value de maxP inférieure à 1 %.[105]

Afin de vérifier l'homogénéité de la qualité de nos deux études, nous avons utilisé le package R MetaQC.[106] Les contrastes utilisés pour nos deux études se basaient sur les temps de stimulation LPS H0, H6 et H24. Afin de mieux cartographier la qualité des données, nous avons ajouté 8 études externes publiques trouvées dans la base de données GEO (Tableau 7). Toutes les études étaient effectuées sur des cellules humaines, avec l'ARN total hybridé sur des puces Affymetrix HumanGene 1.0 ou 1.1 st. Seules les sondes communes entre ces deux technologies étaient prises en compte. La normalisation des données de toutes ces études s'est faite avec le même protocole que pour nos données.

Titre de l'étude (identifiant GSE)	Taille	Cellules principales impliquées	Bases pour le contraste	Technologie de puces
Transcriptional and functional profiling of human intestinal dendritic cells (GSE50380)	11	DCs intestinales	Marqueurs CD103 et Sirpa	Affymetrix HumanGene 1.0 st
Interleukin-27 is a potent inhibitor of cis HIV-1 replication in Monocyte-derived Dendritic Cells via a Type I Interferon-independent pathway (GSE44732)	6	MD-DCs	Stimulation ou non avec de l'IL-27	Affymetrix HumanGene 1.0 st
Gene expression analysis of dendritic cells from normal or tumor sections of human prostates (GSE26747)	10	DCs de biopsies de tumeurs de la prostate	Échantillons issues de tumeurs ou de tissus normaux	Affymetrix HumanGene 1.0 st
Expression data from laser capture microdissected human atherosclerotic plaque, rich or void of plasmacytoid dendritic cells (GSE49670)	11	Micro-dissections de plaques d'athérosclérose	Plaques enrichies ou non avec des DCs plasmacytoïdes	Affymetrix HumanGene 1.0 st
Transcriptome analysis of five population of Antigen Presenting Cells: inflammatory macrophages, Inflammatory dendritic cells, CD14+CD16- monocytes, CD14 dim Cd16+ monocytes and BDCA1+ Dendritic cells (GSE40484)	22	DCs inflammatoires	DCs inflammatoires et autres types cellulaires (monocytes, macrophages)	Affymetrix HumanGene 1.1 st
Gene expression profiling of CD4 T-Cells (CD4+CD62L+) from human peripheral blood mononuclear cells (PBMCs). PBMCs were isolated from healthy individuals from the Boston area (GSE56033)	499	Lymphocytes T CD4+ isolées de PBMCs	Création de deux groupes aléatoires (tous correspondent à des lymphocytes T CD4+)	Affymetrix HumanGene 1.0 st
Expression data measured by microarray of monocyte-derived dendritic cells from healthy individuals stimulated with LPS, influenza, or left unstimulated (GSE53166)	113	MD-DCs	Stimulation ou non par du LPS ou de l'influenza	Affymetrix HumanGene 1.0 st
Human intestinal T cell and paired blood transcriptomes (GSE49877)	36	Lymphocytes T intestinaux	Marqueurs CD4 ou CD8	Affymetrix HumanGene 1.0 st

Tableau 7: Études externes publiques ajoutées pour la méta-analyse par MetaQC.

II.3) Analyses statistiques des gènes DE

Les analyses statistiques pour détecter les gènes DE étaient effectuées avec LIMMA.[107] Les matrices de design étaient créées de manière à prendre en compte les interactions entre les facteurs temps de stimulation LPS (H0, H6 ou H24), le statut HLA-B27 (POS ou NEG) et le statut de la maladie (PAT ou TEM). Pour les calculs des modèles linéaires, la corrélation intra-individuelle était prise en compte. Lors de l'analyse de nos études fusionnées, une correction sur le facteur étude (etude1 ou etude2) était ajoutée.

Les comparaisons effectuées se sont divisées en trois listes : les patients HLA-B27 positifs contre les témoins HLA-B27 négatifs (liste A), les témoins HLA-B27 positifs contre les témoins HLA-B27 négatifs (liste B), et les patients HLA-B27 positifs contre les témoins HLA-B27 négatifs (liste C). Pour chacune de ces listes, des comparaisons différentes étaient effectuées en fonction des temps de stimulation LPS : comparaisons aux trois temps fixes ; comparaisons différentielles entre H0 et les temps H6 et H24 ; et comparaison globale à tous les temps de stimulations.

II.4) Bootstrap des échantillons

Afin de vérifier la fiabilité des modèles linéaires entrés dans LIMMA et que nos résultats présentaient un signal stable, nous avons utilisé une approche de bootstrap. Pour cela, nous avons mélangé les échantillons par un tirage avec remise en conservant les facteurs associés, puis effectué les mêmes analyses LIMMA que pour l'identification des gènes DE, en répétant cela 1000 fois. Pour chaque analyse LIMMA, la corrélation intra-individuelle était de nouveau inférée. Nous avions ainsi une distribution, pour chaque gène, de p-values issues des statistiques t bayésiennes inférées par LIMMA sur ces données mélangés. Nous avons comparé, par un calcul de facteur de corrélation r^2, les moyennes de ces distributions à l'échelle logarithmique avec les p-values initiales LIMMA qui étaient inférieure à 10 %.

II.5) Permutation des facteurs

Nous voulions nous affranchir du biais qui pouvait apparaître sur nos résultats statistiques par le fait que LIMMA se basait sur l'hypothèse d'une distribution de Student des expressions des gènes. Pour cela, nous avons utilisé une approche de permutation pour calculer une p-value empirique. Nous avons donc mélangé les facteurs des échantillons avec un tirage sans remise, de telle sorte

qu'il y ait au final le même nombre d'échantillons attribués à chaque groupe de facteurs. Ces données étaient ensuite analysées par LIMMA de la même façon que pour le bootstrap, à l'exception que les corrélations intra-individuelles n'étaient pas calculées de nouveau (même structure des duplicats pour chaque tirage). Ce tirage était effectué 10 000 fois, et permettait d'avoir pour chaque gène une distribution empirique des statistiques t bayésiennes inférées par LIMMA. La p-value empirique était calculée à partir d'une fonction de distribution cumulative déduite de la distribution des valeurs absolues des statistiques permutées. Concrètement, elle correspondait à la proportion de valeurs absolues des statistiques permutées supérieures ou égales à la statistique initiale. Un gène était considéré alors comme DE lorsque la p-value LIMMA initiale et la p-value empirique issue des permutations étaient tous les deux inférieurs à un certain seuil (5%).

III/ Analyses des jeux de gènes DE

III.1) Jeux de gènes testés

Pour chacune des analyses d'enrichissement des jeux de gènes effectuées sur les données normalisées de nos études, nous avons utilisé la base de données de jeux de gènes MsigDB, la même qui est utilisée par la GSEA initiale.[115] Cette base de données se divisait en 8 collections majeures. La collection C1 se basait sur les positions proches des gènes le long du génome. La C2 était une organisation des gènes en fonction d'autres bases de données de voies de signalisation, de publications dans PubMed et de la connaissance des experts dans certains domaines biologiques. La C3 était une collection se basant sur les motifs de régulations conservés entre espèces homologues. La C4 était définie par l'analyse de large collections de données de puces transcriptomiques utilisées dans le cadre d'études de cancers. La collection C5 se basait sur les termes GO. Les collections C6 et C7 rassemblaient respectivement des signatures oncogéniques et immunologiques. La collection H était un rassemblement cohérent des 7 autres collections pour représenter des états ou processus biologiques bien définis. En tout, ces collections rassemblaient 10 294 jeux de gènes sur 17 300 gènes.

III.2) Méthode d'analyse des jeux de gènes

La méthode quSAGE était utilisée avec le package R correspondant pour effectuer une GSEA sur les données normalisées de nos études.[123] Les mêmes matrices de design et de contraste que pour

les analyses LIMMA étaient utilisées, afin d'obtenir des statistiques par jeu de gène pour chaque comparaison effectuée.

IV/ Construction et analyse des réseaux

IV.1) Construction des réseaux de gènes sémantiques

Les réseaux sémantiques entre les gènes DE étaient construits à partir des interactions connues en interrogeant STRING directement par son API, sous forme d'URI.[167] Afin d'avoir un recouvrement maximal, le double du nombre de gènes initial était demandé comme nombre de nœuds additionnels. Pour ne pas être bloqué par la limite de la longueur de l'URI pour l'API, nous avons défini une limite de nombre de gènes initial à un maximum de 400. Seules les interactions issues de co-expressions, de données expérimentales ou de bases de données internes étaient sélectionnées. Ensuite, les scores combinés étaient de nouveau calculés en ajoutant une correction antérieure qui considérait que tous les scores des PPIs étaient sur un taux de base de 6 %. Ce calcul de score combiné était utilisé dans la version 8 de STRING.[220] Enfin, un filtrage sur les scores avec un seuil minimum de 0,6 était appliqué.

Compte tenu du nombre important d'interacteurs ajoutés par nos requêtes dans l'API de STRING, nous avons résumé les interactions indirectes entre deux gènes initiaux par une seule interaction reprenant le chemin le plus court entre les deux. Pour cela, nous avons converti les scores combinés S en distances D, ainsi nous calculions pour une interaction i : $D_i=max(S)+1-S_i$. Plus le score était grand, plus la distance était petite. D représentait un vecteur de distances de même longueur que le nombre d'interactions, avec des valeurs comprises entre 1 et 2. Les chemins les plus courts étaient calculés entre tous les gènes initiaux par l'algorithme de Dijkstra, implémenté dans le package R igraph. Chaque interaction indirecte possédait comme attribut le nombre d'interacteurs, les noms des gènes intermédiaires et la distance calculée.

IV.2) Construction des réseaux de gènes gaussiens

Les réseaux gaussiens entre les gènes DE étaient inférés à l'aide du package R SIMoNe.[174] Tous les réseaux étaient inférés sur les deux études fusionnées. Afin d'éviter le biais dû aux études différentes, nous mettions à la même échelle et centrions les deux séries d'expressions pour chaque gène. La sélection du modèle SIMoNe était effectuée en prenant la moyenne des nombres

d'arcs avec un score BIC et un score AIC maximal. Aucune contrainte de clustering n'était requise pendant l'inférence. Ensuite, des tests de Spearman étaient effectués entre les expressions des gènes pour lesquels SIMoNe avait inféré une interaction. La méthode d'approximation de la distribution nulle pour ces tests était celle de AS89. Enfin, un dernier filtrage sur les valeurs absolues de rho était effectuée en ne gardant que celles supérieures à 0,4.

Les réseaux inférés par SIMoNe étaient comparés avec une approche WGCNA, sans le calcul de matrice de dissimilarités et des modules.[169] Pour cela, nous avons calculé les rhos de Spearman entre tous les gènes du réseau, puis nous les avons converti en scores de similarités σ par σ=(1+rho)/2. Enfin, nous avons calculé les scores de proximités A avec $A=1/(1+e^{-\alpha*(\sigma-0,5)})$. α était fixé à 8, et un filtrage sur A était effectué comme devant être supérieur à 0,85 ou inférieur à 0,15.

IV.3) Comparaison des connectivités

Pour un même jeu de gènes, nous avons parfois inféré deux réseaux gaussiens par SIMoNe, en calculant les corrélations des expressions d'une part chez les patients, et d'autre part chez les témoins. Ainsi, nous obtenions deux réseaux superposables. Nous avons ensuite calculé les connectivités de chaque gène spécifiquement chez les patients ou témoins, en additionnant les valeurs absolues de rho. La différence de connectivité était ainsi calculée entre les deux groupes. Les gènes différentiellement connectés chez les patients avaient une différence positive, tandis que ceux chez les témoins avaient une différence négative.

IV.4) Visualisation des réseaux

Tous les réseaux construits par les méthodes décrites précédemment étaient visualisés et manipulés à l'aide du logiciel Cytoscape.[158] Le plugin cyREST permettait de les importer automatiquement.[160]

Différents styles de visualisation étaient appliqués en fonction de chaque type de réseau. En dehors des réseaux de comparaison des connectivités, la taille des nœuds était inversement proportionnelle à la p-value donnée par LIMMA pour la liste C. La couleur changeait en fonction des fold changes de cette même liste, avec le bleu traduisant une sous-expression et le rouge une sur-expression. Les arcs étaient représentés en fonction des attributs spécifiques de chaque type de réseau (score, distance, rho de Spearman). Les interactions indirectes dans les réseaux STRING

résumés étaient représentées par des pointillés bleus, contrairement à un trait noir solide pour les interactions directes. Pour les réseaux de comparaison entre patients et témoins, les différences de connectivités étaient utilisés pour représenter les nœuds.

Chapitre 3

PRÉSENTATION DES RÉSULTATS

Cohorte et base de données relationnelle

Dans le cadre d'une uniformisation des données, nous avions rassemblé les informations d'une cohorte de quelques milliers de personnes suivies atteintes ou non de SpA. Toutes les cohortes décrites pour nos études de génotypage et transcriptomiques sont comprises dans cet ensemble. Cette partie des résultats décrit dans un premier temps la cohorte totale, puis comment les données recueillies étaient uniformisées, stockées et rendues facilement accessibles dans une base de données relationnelle.

I/ Présentation de la cohorte totale

La cohorte totale représentait une base de données considérable regroupant 3923 personnes. Ils étaient recrutés par notre laboratoire pour diverses raisons qui peuvent se recouper. Par exemple, 2345 étaient recrutés pour des analyses de familles multiplex, 978 pour des études de cas-témoins, 383 car ils présentaient des récurrences systématiques, 221 pour des études simplex, ou 150 pour des analyses du microbiote intestinal. Cette cohorte est donc riche d'informations et permet d'effectuer de nombreux types d'études avec une approche de biologie intégrative.

Le ratio hommes/femmes était au total de 48,01 %. Chez les 1818 patients, ce ratio était de 50,94 %. Les 1773 témoins présentaient 46,98 % d'hommes, avec 0,28 % pour lesquels le sexe était inconnu. Parmi les 1806 patients testés, 78,96 % étaient HLA-B27 positifs. Les 1616 témoins séquencés pour cette région présentaient une tendance inverse dans cette cohorte, avec seulement 34,10 % d'individus HLA-B27 positifs. Leur de leur dernier suivi, 99 des patients avaient moins de 25 ans, 1021 avaient entre 25 et 50 ans, et 695 avaient 50 ans ou plus. Chez les témoins, ces tranches d'âges étaient réparties respectivement chez 143, 667 et 891 individus. Les individus étaient répartis dans 1048 familles, parmi lesquelles 420 ne rassemblaient qu'une seule personne enregistrée.

Parmi les individus témoins, 96,79 % étaient sains, et 2,99 % avaient un phénotype non enregistré. Parmi les 1612 patients classifiés, 794 avaient une AS, 543 une UspA, 389 un psoriasis, 271 un rhumatisme, 74 une maladie de Crohn, 47 une colite ulcérative et 5 une arthrite réactionnelle. 470 présentaient plusieurs de ces phénotypes. 874 patients avaient développé leurs symptômes avant 25 ans, 486 entre 25 et 35 ans, 266 entre 35 et 50 ans, et 50 à 50 ans ou plus. Le BASDAI moyen

des patients était de 37,72, avec un écart-type de 22,43. Le BASFI moyen était de 26,33, pour un écart-type de 25,12. Les ADNs de 99,72 % des patients et 99,89 % des témoins étaient prélevés et étaient associés à des numéros de tubes ADN bien définis.

II/ Base de données relationnelle

Compte tenu du nombre important d'informations recueillies dans cette cohorte, il était primordial d'installer un système permettant d'uniformiser les données et de les rendre accessibles au laboratoire. De plus, les données devaient être anonymisées, avec comme identifiant unique le numéro CEPH. Pour cela, nous avions pris l'initiative de créer une base de données relationnelle sous le logiciel MySQL, permettant ainsi de créer des contraintes et des routines d'importation et de vérification des données.

Ainsi, nous avions créé une base de données dont la structure était divisée en deux tableaux : un rassemblait les données cliniques de chaque individu ; l'autre décrivait les échantillons ADNs prélevés dans la cohorte (Figure 13). Les deux tableaux étaient séparés pour la raison qu'un individu avait pu être prélevé plusieurs fois. Ce lien entre les deux tableaux était très important, car un clinicien se référerait plus au numéro CEPH, tandis qu'une personne manipulant les échantillons ADN serait plus à l'aise avec les numéros de tube. Cette base de données a été créée en Novembre 2011 et était mise à jour régulièrement, tous les deux ou trois mois, par un envoi sous format Excel de la part des gestionnaires de la base de données (à la base enregistrée sous FileMaker Pro) puis par importation automatique grâce à des scripts SQL. Après chaque importation, des vérifications étaient effectuées pour vérifier la cohérence des numéros de tube ADN et des numéros CEPH grâce à une double entrée manuelle, et que les liens de parenté définis ne présentaient pas d'aberrance.

Le logiciel MySQL et son langage SQL sont de très bons outils pour stocker et vérifier l'intégrité des données, mais son interface n'est pas forcément très intuitive. Les requêtes par SQL demandent une formation initiale du langage et peuvent être très complexes lors de demandes précises. C'est pourquoi, en plus de la maintenance de la base de données, nous avions créé une interface web plus intuitive et spécialisée que PhpMyAdmin, grâce à un site web en local. Pour cela, nous avions utilisé un système LAMP sur un serveur dédié, c'est-à-dire que l'ordinateur où étaient stockées les données fonctionne sous Linux avec les logiciels Apache, MySQL et PHP. Ces

trois permettaient de créer un site web dynamique, et dont l'affichage dépendait d'une base de données connectée. Ainsi, le client (c'est-à-dire l'utilisateur qui visitait le site web) pouvait accéder de manière intuitive à la base de données sans pouvoir l'éditer et de manière sécurisée. Nous avions donc créé une application web sous Joomla !, un CMS permettant d'intégrer facilement les trois logiciels pour éditer des sites web. Ainsi, toute personne connectée au réseau local pouvait visiter ce site internet et s'identifier pour pouvoir entrer des paramètres de filtrage et d'exportation des deux tableaux rassemblant les données sur la cohorte. Cela permettait ainsi facilement de décrire de façon uniforme les cohortes étudiées dans chaque expérience. La connexion était cryptée par ssh, et un ordinateur extérieur au réseau local ne pouvait pas se connecter, à part par un système de tunnel ssh avec un autre routeur.

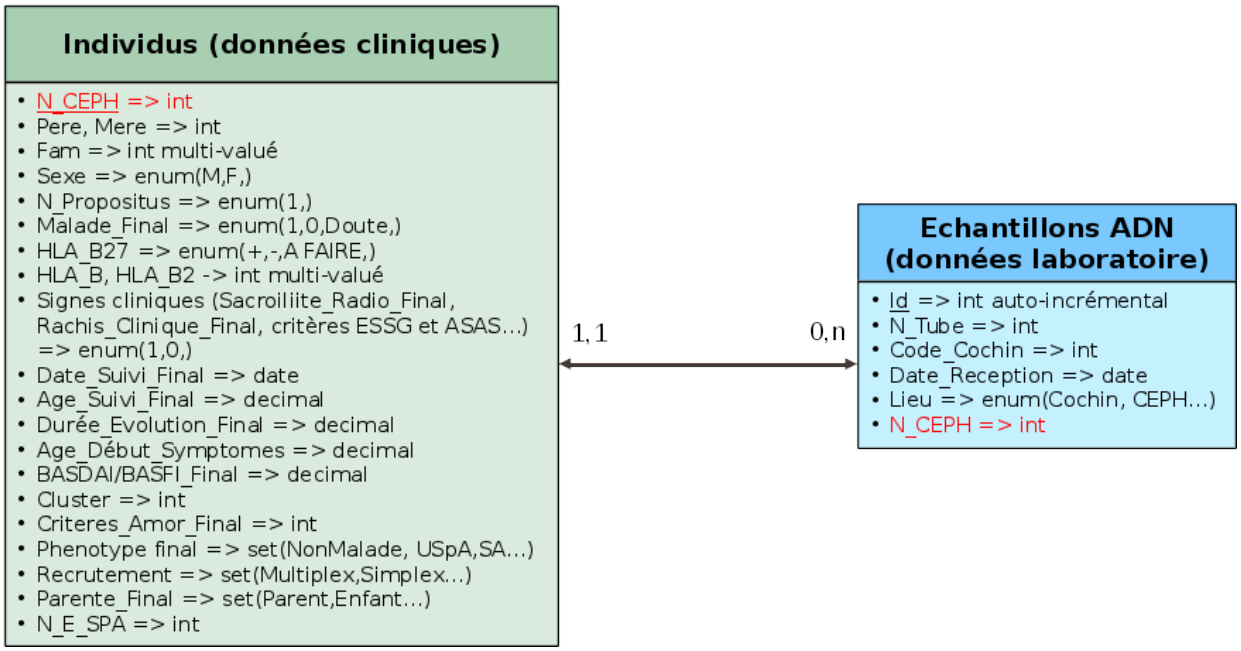


Figure 13: Diagramme UML de la base de données cliniques en lien avec les données d'échantillons d'ADN.

Les couples de chiffres au-dessus de la flèche représente les liens de cardinalité entre les tableaux. Dans ce cas, cela veut dire qu'un échantillon n'est associé qu'à un individu, tandis qu'un individu peut être associé à aucun ou plusieurs échantillons ADNs. Les attributs en rouge représentent la clé étrangère reliant les deux tableaux.

Régions génétiques liées à la SpA

I/ Nettoyage des puces

Les étapes de vérification des échantillons et des SNPs décrites dans les matériels et méthodes étaient effectuées sur les trois cohortes de manière indépendante, car elles rassemblaient des individus issus de familles différentes. Ainsi, les erreurs mendéliennes ou les vérifications de liens de parenté n'auraient pas été affectées par une quelconque fusion des données. Les descriptions des trois cohortes correspondaient déjà aux échantillons après leur vérification et la correction par GRR.

Les données de génotypage des cohortes 1 et 2 étaient soumises aux étapes de vérification d'erreurs par PLINK séparément. Concernant la première cohorte, l'analyse des erreurs partait sur 256 546 marqueurs répartis sur le long du génome. Après filtrage sur les MAF, les déviations de HW, les erreurs mendéliennes et les taux de génotypage des SNPs, 199 117 marqueurs étaient gardés, ce qui représentait une perte de 22,39 %. Concernant la deuxième cohorte, ces étapes de nettoyage avaient retiré 48 213 des 256 517 SNPs initiaux, soit une perte de 18,80 %.

Les erreurs mendéliennes subtiles étaient retirées avec le logiciel Merlin pour ces deux cohortes, toujours de manière indépendante, et séparément sur chaque chromosome. Les temps de calculs pour ces vérifications étaient considérables, d'autant plus qu'elles étaient répétées jusqu'à huit fois pour le chromosome 6 dans la deuxième cohorte. Chaque étape de nettoyage pouvait durer jusqu'à 1 mois, entraînant des durées totales parfois de 5 ou 6 mois. De plus, ces analyses demandaient une quantité de mémoire considérable, ce qui ne laissait la possibilité de lancer en parallèle ces procédures que sur moins d'une dizaine de chromosomes. Ceci explique donc pourquoi, même si nous avions les données à disposition depuis le début de la thèse, nous ne pouvions lancer les analyses de liaison qu'en milieu de projet. Après ces étapes de nettoyage effectuées dans la cohorte 1, 5509 SNPs étaient retirés au profit de 193 608 restant, soit un retrait de 2,77 %. Pour la deuxième cohorte, 7033 marqueurs étaient supprimés pour laisser 201 271 SNPs, soit une perte de 3,38 % (Tableau 8).

Chromosome	Cohorte 1		Cohorte 2	
	Nombre d'étapes	Nombre de SNPs restants	Nombre d'étapes	Nombre de SNPs restants
1	4	14 273	5	14 940
2	4	16 249	5	16 910
3	5	13 600	6	14 184
4	5	13 862	4	14 620
5	4	12 890	5	13 448
6	4	12 886	8	13 415
7	5	10 444	5	10 884
8	6	11 092	6	11 499
9	4	8 981	5	9 329
10	5	10 662	5	11 012
11	4	9 671	5	10 056
12	4	9 523	6	9 978
13	4	8 249	6	8 641
14	4	6 032	5	6 260
15	4	5 064	5	5 212
16	3	5 181	4	5 255
17	2	3 567	4	3 642
18	5	5 904	4	6 187
19	3	2 009	3	2 000
20	4	4 444	5	4 513
21	3	2 890	5	3 037
22	4	1 895	4	1 898
X	4	4 240	4	4 351

Tableau 8: Bilan des nettoyages d'erreurs sous Merlin dans les cohortes 1 et 2.

Le nombre de SNPs restant constituent ceux qui avaient un taux de génotypage supérieur à 95 % après les étapes de nettoyage par Merlin.

Compte tenu du temps de calcul important des nettoyages d'erreurs mendéliennes subtiles et de la très faible taille de la cohorte 3 comparée aux deux autres, nous avions décidé de n'effectuer que les étapes de nettoyage par PLINK sur cette cohorte. Nous nous étions basés sur les SNPs en commun entre les deux premières cohortes, ce qui représentait 179 907 marqueurs. Après les nettoyages effectués par PLINK, 178 881 SNPs étaient gardés, soit un retrait de 0,57 % des marqueurs (Tableau 9).

	Cohorte 1	Cohorte 2	Cohorte 3
Nombre de SNPs initial	256 546	256 517	179 907
Nombre de SNPs avec MAF < 1 %	34 552	32 919	32
Nombre de SNPs avec un déséquilibre de HW	4 488	2 077	161
Nombre d'erreurs mendéliennes	226 191	69 202	3 563
Nombre de SNPs avec un taux de génotypage < 95 %	18 389	13 217	833
Nombre de SNPs restants	199 117	208 304	178 881

Tableau 9: Bilan des nettoyages d'erreurs sous PLINK dans les trois cohortes.

II/ Région liée à la région 13q13 - partie de la cohorte 1

Avant que n'arrivaient les données de génotypage de la cohorte 2, la cohorte 1 avait fait l'objet d'une analyse de liaison non paramétrique tout le long du génome, afin de tester la présence de nouveaux locus associés à la SpA. En vérité, les résultats présentés ici ne représentaient qu'une partie de la cohorte 1, même si elle était très majoritaire, l'autre partie des données étant arrivée ultérieurement. Cette cohorte faisait par la suite l'objet d'une présentation de résultats d'une étude d'extension et une autre de réplication décrites plus loin (Etude d'extension - cohortes 1 et 2 fusionnées).

Les analyses de liaison non paramétriques (NPL) effectuées sur ces données ont permis d'associer pour la première fois une région significativement liée à la SpA en dehors du CMH en 13q13. De plus, 8 autre locus avaient obtenu un score liaison suggestif, qui étaient sur les chromosome 1, 8, 9, 10 et 17, et qui comprenaient le locus SPA2. Une analyse approfondie de la région 13q13 avait permis d'estimer la fraction α de familles liées à ce locus à environ 30 %, d'après une analyse paramétrique avec un modèle de transmission co-dominant avec des pénétrances de 30 et 60 % en fonction de l'hétérozygotie de l'allèle. De plus, l'analyse de recombinaisons homologues informatives avait permis de restreindre la région d'intérêt sur un intervalle d'environ 1,3 Mb contenant seulement 6 gènes. Une analyse d'association au niveau de cette région n'avait détecté aucun SNP avec un odds ratio significatif. Cela souligne l'importance des analyses de liaison pour détecter d'autres variants permettant de mieux expliquer l'hétérogénéité de la SpA.

Ces résultats ont fait l'objet d'un article publié dans la revue Annals of the Rheumatic Diseases. La région 13q13 est actuellement séquencée par notre laboratoire pour une analyse approfondie,

similaire à celle du locus SPA2.

Mon implication principale pour cette étude était de mettre en place tout le flux de travail de nettoyage des données de puces et d'analyses non paramétriques et paramétriques. Cela représentait des temps de calculs considérables, comme indiqué dans les Ressources informatiques des Matériels et méthodes, ce qui est habituellement observé pour une cohorte qui dépasse largement les 1000 personnes. Un soin particulier, ainsi qu'une rigueur indiscutable, étaient nécessaires pour éviter des erreurs lors des transferts de fichiers et de connaissances. Ensuite, les analyses plus approfondies, comme l'analyse des recombinaisons homologues informatives ou la détermination de la fraction de familles liées, étaient effectuées par d'autres personnes du laboratoire.

Whole-Genome Single Nucleotide Polymorphism-based Linkage Analysis in Multiplex Families of Spondyloarthritis Reveals a New Susceptibility Locus in 13q13.

Félicie Costantino, M D, Ph D,[1,2,3] Emmanuel Chaplais, M S,[1,3] Tifenn Leturcq, M D,[1,3] Roula Said-Nahal, M D,[2] Ariane Leboime, M D,[2] Elena Zinovieva, Ph D,[1,3] Diana Zelenika, Ph D,[4] Ivo Gut, Ph D,[4] Céline Charon, Ph D,[4] Gilles Chiocchia, Ph D,[1,3] Maxime Breban, M D, Ph D,[1,2,3*] , Henri-Jean Garchon,M D, Ph D,[1,3,5*]

1 – INSERM U1173, UFR Simone Veil, Versailles-Saint Quentin University, France

2 – Rheumatology Division, Ambroise Paré Hospital (AP-HP), Boulogne-Billancourt, France

3 – Université Paris Diderot, Sorbonne Paris Cité, Laboratoire d'Excellence, Paris, France

4 – National Genotyping Center (CNG), Evry, France

5 – Genetics Division, Ambroise Paré Hospital (AP-HP), Boulogne-Billancourt, France

* The two last authors shared equal contribution

This work was supported by a grant from Agence Nationale de la Recherche (grant ANR 2010 GEMISA). Félicie Costantino and Tifenn Leturcq were supported by a grant from the Société Française de Rhumatologie (SFR).

Address reprint requests and correspondence to:
Henri-Jean Garchon
UFR Simone Veil, 2 avenue de la source de la Bièvre, 78180 Montigny le Bretonneux, France
Ph: 33-(0)170 429 470
E-mail : henri-jean.garchon@inserm.fr

Keywords: ankylosing spondylitis; spondyloarthritis; linkage analysis; genetics.

Word Count: 2,715

ABSTRACT

Objective: Spondyloarthritis (SpA) is a chronic inflammatory disorder with high heritability but with complex genetics. Apart from HLA-B27, most of the underlying genetic component remains to be identified. We conducted a whole-genome high density non-parametric linkage analysis to identify new genetic factors of susceptibility to SpA.

Methods: 914 subjects including 462 with SpA from 143 multiplex families were genotyped using Affymetrix 250K microarrays. After quality control, 189,368 single-nucleotide polymorphism (SNPs) were kept for further analyses. Both non-parametric and parametric linkage analyses were performed using Merlin software. Association was tested with Unphased.

Results: Non-parametric linkage analysis identified two regions significantly linked to SpA: the MHC (LOD_{max} = 24.77) and a new 13q13 locus (LOD_{max} = 5.03). Additionally, 8 loci achieved suggestive LOD scores, including the previously identified SPA2 locus at 9q33 (LOD_{max} = 3.51). Parametric analysis supported a co-dominant model in 13q13 with a maximum heterogeneity LOD, "HLOD" score of 3.084 (α = 0.28). Identification of meiotic recombination events around the 13q13 linkage peak in affected subjects from the 43 best-linked families allowed us to map the disease interval between 38.753 and 40.040 Mb. Family-based association analysis of the SNPs inside this interval in the best-linked families identified a SNP near *FREM2* (rs1945502) which reached a p-value close to statistical significance ($p = 5.8 \times 10^{-4}$).

Conclusions: We report here for the first time a significant linkage between 13q13 and SpA. Identification of susceptibility factor inside this chromosomal region through targeted sequencing in linked families is underway.

INTRODUCTION

Spondyloarthritis (SpA) is one of the most common forms of chronic inflammatory rheumatisms with an estimated prevalence of 0.43% in the Western adult population (1). It is characterized by axial and/or peripheral joint inflammation, often in association with extra-articular inflammatory features such as psoriasis, uveitis or inflammatory bowel disease (IBD). Depending on its clinico-radiological presentation, several subsets have been described: ankylosing spondylitis (AS), psoriatic arthritis, arthritis associated with IBD, reactive arthritis and undifferentiated SpA. Familial aggregation among these subsets has been established long ago suggesting shared genetic factors (2). This concept was reinforced by extensive analysis of a large panel of families with multiple cases of SpA, leading to the conclusion that all of the subtypes should be considered together in genetic studies (3–5).

Familial aggregation of SpA is high with a sibling recurrence risk ratio of 40 (6). Disease heritability, as estimated by twin studies in AS, exceeds 90% (7). The most important part of heritability comes from the HLA-B27 allele in the major histocompatibility complex (MHC), present in approximately 75% of the patients (1). However, there is strong epidemiologic evidence to suggest that other genes are involved. First, only 1 to 5% of HLA-B27 positive individuals develop SpA (1,8). Second, the disease risk is 5.6 to 16 times greater in HLA-B27 positive relatives of AS patients than in the general population (8). Third, in HLA-B27 twin pairs, the concordance rate was higher in monozygotic than in dizygotic ones (7,9,10). Recurrence-risk modeling based on familial data supports the hypothesis of an oligogenic model with between 3 and 9 genes operating in addition to HLA-B27 (11).

Several groups have identified genetic polymorphisms outside the MHC involved in AS susceptibility. The most recent findings resulted from genome-wide association studies (GWAS).

Since 2007, three GWAS conducted in Caucasian population have led to the identification of 12 susceptibility loci outside the MHC (12–14). More recently, an international study performed in 10,619 AS cases and 15,145 controls, using the Immunochip array especially designed for immunogenetics studies, has uncovered 13 additional loci (15). However, these newly discovered loci explain only a small additional fraction of AS heritability. In total, less than 25% of the heritability of AS was considered as explained (including 20% contributed by HLA-B27 and 5% by other loci) (15).

Most of the associations with AS discovered through GWAS involve common variants. It has been suggested that a significant part of the "missing heritability" in complex diseases might be due to effects of rare variants that are poorly detected in GWAS (16). Linkage analyses could be more powerful than GWAS for the detection of such variants (17). In SpA, three genome-wide linkage studies have been previously published, two in AS and one in SpA as a whole (18–21). Only two loci besides the MHC reached significance threshold: one on 16q and the other on 9q31-34 (19,21). Here we report on a new genome-wide linkage analysis using a high-density panel of single nucleotide polymorphisms (SNP) in an extended set of multiplex SpA families.

SUBJECTS & METHODS

Multiplex SpA families

Caucasian multiplex SpA families, *i.e.* including more than one SpA case per family, were recruited through the Groupe Français d'Etude Génétique des Spondylarthropathies (GFEGS), as previously described (3). This study was approved by the local ethical committees of Cochin Hospital (Paris, France) and Ambroise Paré Hospital (Boulogne-Billancourt, France). Written informed consent was obtained from each participant.

The diagnosis of SpA was originally ascertained according to the internationally validated classification criteria of Amor and/or European Spondyloarthropathy Study Group (22,23), as previously described (3). All anteroposterior radiographs of the pelvis were examined blindly and independently by 2 qualified examiners (RSN, MB), using an established grading system (24). Even if they were not developed at the time of recruitment, the Assessment of SpondyloArthritis international Society classification criteria for axial and peripheral SpA were applied *a posteriori* (25,26).

The family set consisted of 143 families including 914 genotyped subjects of whom 462 had SpA. Detailed clinical characteristics of the dataset are provided in **Table 1**. Pedigrees' structures are summarized in **Table 2**.

Genotyping

A SNP-based genome scan was conducted using the Affymetrix 250K SNPs array that covers the entire genome with around 262,000 SNPs. Genotyping was performed according to the manufacturer's instructions by the Centre National de Génotypage (Evry, France).

Briefly, total genomic DNA was digested with the NspI enzyme, ligated to the adaptor and amplified by polymerase chain reaction (PCR). After purification of the PCR products, amplicons

were quantified, fragmented, labeled and hybridized to the 250K SNPs mapping array. Genotyping calling was determined by the BRLMM algorithm (27).

Genotyping quality control

Gender corresponding to each DNA sample was checked by analysis of X chromosome heterozygosity using PLINK software (28). The relatedness of individuals was verified using the GRR program (29).

Quality control was also performed for each SNP. First, information data provided by Affymetrix was updated using dbSNP 135. A total of 5,587 SNPs were removed from analysis for various reasons (5,188 had more than 2 alleles, 303 had been deleted from dbSNP and 96 were double-hits). Physical position on the genome and allele phasing was also updated if necessary. Pedstats was used to detect Mendelian errors (30). SNPs were then selected according to the following parameters: genotyping rate >95%, minor allelic frequency >1% and a significant deviation from Hardy-Weinberg proportions ($P>1x10^{-3}$). Subtle genotyping inconsistencies were also detected and removed with Merlin (31). Finally, 189,368 SNPs were kept for further analysis.

Linkage analysis

Multipoint non-parametric linkage was tested for all autosomal chromosomes using Merlin software (31) to calculate the Kong and Cox LOD score (32). Two genome-wide thresholds of LOD score were used according to Lander & Kruglyak recommendations: 3.65 for significance, 2.2 for suggestiveness (33).

Linkage disequilibrium (LD) between SNPs could inflate LOD score in linkage studies (34). To address this issue, we performed linkage analysis handling LD as implemented in Merlin (with r^2 threshold of 0.1) in the best-linked regions.

Following non-parametric analysis, we additionally performed parametric linkage analysis

of the best-linked loci under three generic models, dominant and recessive (each with an arbitrary 50% penetrance of the disease allele) and co-dominant (with 30% and 60% penetrance in the presence of one or two doses of the disease allele).

Haplotype reconstruction

Regional haplotypes were reconstructed with Merlin, using a subset of SNPs selected on a minor allelic frequency > 0.2 and pairwise $r^2 > 0.8$. LD for reconstructed haplotypes was determined using the Haploview program V4.2 (35). Haplotypes were visualized on family tree using HaploPainter (36).

Association analyses

Family-based association study was performed using Unphased version 3.1.6 (37). Bonferroni correction was applied to determine significance threshold.

RESULTS

Non-parametric linkage analysis

Results of the whole-genome non-parametric linkage analysis in the whole dataset are summarized in **Figure 1** and **Table 3**. Two regions achieved significant LOD scores: the MHC (LOD_{max} = 24.77) and the 13q13 region with a linkage peak at 39.7 Mb from the p-telomere (LOD_{max} = 5.03) which represents a new hit. Additionally, 8 loci achieved suggestive LOD scores, notably including the previously identified SPA2 locus at 9q33 (LOD_{max} = 3.51 at 130 Mb).

Detailed analysis of the new locus in 13q13

Non-parametric linkage analysis of the chromosome 13 revealed a significantly linked region around 40 Mb from the p-telomere as shown in **Figure 2**. Linkage was still significant after taking in account LD (**Figure 2**).

A parametric analysis was then conducted to test the likelihood of generic models of the disease locus and to assess the proportion α of linked families under these models. Using the same subset of SNPs as for haplotype reconstruction, the co-dominant model (with 30% and 60% penetrance in the presence of one or two disease alleles, respectively) was best supported, with a maximum heterogeneity LOD (HLOD) score of 3.084 (α = 0.28)..

To further circumscribe the disease locus, we reconstructed haplotypes around the linkage peak in the 43 best-linked families having a ponderated LOD score $\geq$ 0.16 (i.e. a threshold consistent with the α coefficient of 0.28). In 23 of these families, we identified one haplotype segregating with the disease, whereas in the remaining ones, all the affected members shared both of their haplotypes. Moreover analysis of the 5 meiotic recombination events occurring in patients (an example of such recombination event is shown in **Supplementary figure 1)** allowed us to

map the disease interval between 38.753 Mb and 40.040 Mb, consistent with the results of non-parametric and parametric linkage analysis (**Supplementary figure 2**). The SNPs of this interval were tested for association in the 43 best-linked families. We observed no significant association after Bonferroni correction. However, the SNP rs1945502, an intergenic SNP located in the upstream region of *FREM2*, yielded a p-value close to significance (nominal $p = 5.8 \times 10^{-4}$, corrected $p = 0.08$) with an odds ratio of 2.3 (95% confidence interval: 1.3 to 4.0) (**Figure 3**). All the haplotypes segregating with disease in the best-linked families, but one, harboured the risk allele of this SNP.

The MHC

As expected, the MHC region on 6p21, carrying the HLA-B27 alleles, yielded the highest LOD scores ($LOD_{max} = 24.77$). More than 92% of the patients included in the study were HLA-B27 positive (**Table 1**) and a vast majority of the families were linked to this locus. Four families had no HLA-B27 positive affected member (one consisted of 4 SpA patients, another one of 3 cases and the 2 remaining contained 2 cases each). Only one of those families had no MHC haplotype cosegregating with the disease (**Supplementary figure 3**). In the other three ones, all the affected members shared at least one MHC haplotype, still consistent with MHC linkage and suggesting allelic heterogeneity. Considering the 14 families in the lower decile of the ponderated LOD score distribution in the MHC, we observed in 7 of them two HLA-B27 haplotypes coming from distinct unrelated members. In the other ones, at least one affected subject was HLA-B27 negative but there was no obviously recurring HLA-B allele.

Genetic heterogeneity

Except for the MHC, we observed high linkage heterogeneity with an estimated fraction of 28% of the families linked to 13q13 and of 16% linked to the previously identified SPA2 locus in

9q33. We compared the LOD scores obtained by each family at the two non-MHC best-linked loci (13q13 and 9q33) and observed no correlation (pairwise Spearman's test: $p = 0.73$; rho = 0.03). (**Supplementary figure 4**).

DISCUSSION

This high density genome-wide linkage analysis performed in a large set of multiplex families provides evidence for the existence of a new non-MHC susceptibility locus for SpA on 13q13. A suggestive linkage signal was also observed in this same region in the previous genome scan performed by our group, using microsatellite markers (NPL_{max} = 2.55; LOD_{max} = 1.41) (21).

The present study however cannot be considered as a replication because of the overlap between both family sets. Hence, 47 of the 65 families included in the initial microsatellite genome scan were also part of the present study. In the pooled analysis of the 3 previously published linkage analyses in SpA, conducted by Carter *et al* (38), the microsatellite marker D13S218 in 13q13 achieved moderate evidence of linkage ($P < 0.01$) with suggestive linkage ($P < 0.05$) in the GFEGS and North American Spondylitis Consortium (NASC) datasets (20,21). Besides, the 13q13 locus has also been shown as significantly linked to Crohn's disease (CD), a chronic IBD which is frequently associated with SpA (39). In our dataset, however, comparison of the patients belonging to 13q13-linked families with the others showed no significant increase in CD frequency (**Table 1**).

Further analyses of the 13q13 allowed us to delimit a 1.3 Mb disease interval mapped between 38.7 and 40.0 Mb. This interval contains 6 genes (UFM1, FREM2, STOML3, PROSER1, NHLRC3 and LHFP), 4 pseudogenes and 4 non-coding RNA genes. Among them, LHFP, UFM1 and FREM2 are potential candidates. Indeed, polymorphisms of LHFP (lipoma HMGIC fusion partner) were previously associated with psoriasis, a major extra-articular component of the SpA spectrum (40). However, we observed no increased psoriasis frequency in patients of 13q13-linked families (**Table 1**). Some of the functions of UFM1 may also been related to SpA pathogenesis. Indeed components of the Ufm1 cascade have been demonstrated to

be induced specifically under endoplasmic reticulum stress (41,42), a process that has been implicated in SpA pathogenesis (43). Finally, FREM2 is known to encode a large integral membrane protein of 3,169 aa with numerous repeats of a chondroitin sulfate proteoglycan domain. It is expressed at the basement membrane of several epithelia. Mutations of FREM2 were described in patients affected with type 2 Fraser syndrome, a rare malformative recessive disorder. Concerning the other genes located in the susceptibility interval, there is either little information on their function in public databases or their known functions could not be easily related to SpA pathogenesis.

Except for the MHC, there was a poor overlap between the findings of our scan and those of the studies previoulsy published by the Oxford group and NASC (19,20). Among the 8 suggestive loci, only two in 9q and 17q overlapped with or were close to those identified by the pooled analysis of Carter *et al* (38). The 17q region also overlaps with the NPEPPS-TBKBP1-TBX21 gene cluster recently shown as associated with AS (15). Indeed, the suggestiveness threshold corresponds to statistical evidence for a hit expected to occur less than once at random in a genome-wide scan. Thus some of the suggestive loci we detected may be true positives.

Several explanations could be proposed to explain discrepancies between studies. First, there is probably a high level of genetic heterogeneity in SpA, as reported for most complex diseases (44). Such genetic heterogeneity may be increased by differences in the geographic origin of patients (for instance, French versus British or North American). However, genetic heterogeneity may be seen in an ethnically homogenous population. For instance, we estimated that only 28% of the families studied herein were linked to 13q13 while a smaller set (16%) was linked to 9q33.

Genetic heterogeneity might also depend on variations in the studied phenotype. Here, we

included all subtypes of SpA and not only AS patients (requiring advanced radiographic sacroiliitis) as it was the case in the Oxford and NASC studies (19,20). Our choice was based on segregation analyses indicating that major shared genetic factors were expected to account for all SpA subtypes coexisting in families (3,4). On the other hand, some of the loci previously identified in studies restricted to AS could be related to structural damage, i.e. more to disease severity than causality. Hence, several reports suggested that structural damage in SpA is partly genetically determined (45,46) and we recently identified that the SNP rs11209026 in the *IL23R* gene was specifically associated with AS but not with SpA lacking radiographic sacroiliitis (47).

Another difference between this study and those previously published in the early 2000's concerns the type of genetic marker used (a dense array of roughly 190,000 effective SNPs here, and a limited number of around 370 microsatellites from the Applied Biosystems Prism Linkage Mapping Set Version 2.0, in the previous ones). It is likely that the chance of detecting linkage was maximized in the present study by using a dense map of markers that extracted the maximum inheritance information content (48). Although they are bi-allelic and thus less informative, SNPs are present at a far greater density than microsatellites throughout the genome. Several studies have demonstrated that a dense map of SNPs has greater power to detect linkage than low density microsatellite maps (49–52). In some instances, the same dataset was genotyped using both microsatellites and SNPs and some significantly linked loci were identified only in the SNP-based analysis (50–52). Moreover, a high SNP density allows a more precise localization of disease locus (51,52).

To conclude, we report here for the first time a significant linkage between 13q13 and SpA, highlighting the interest of high density SNP-based genome scan in large pedigrees dataset. Sequencing of the region of interest delimited between 38.7 and 40.0 Mb from the p-telomere in

patients from the best linked families could allow us to identify variants associated with the disease. Additionally, 8 loci (including the previously identified SPA2 locus) achieved genome-wide suggestiveness threshold. Replication study in another familial cohort would be expected to discriminate true linkage from false positivity.

REFERENCES

1. Costantino F, Talpin A, Said-Nahal R, Goldberg M, Henny J, Chiocchia G, et al. Prevalence of spondyloarthritis in reference to HLA-B27 in the French population: results of the GAZEL cohort. *Ann Rheum Dis* 2013. Available at: Dec 18. doi:10.1136/annrheumdis-2013-204436. [Epub ahead of print].

2. Moll JM, Haslock I, Macrae IF, Wright V. Associations between ankylosing spondylitis, psoriatic arthritis, Reiter's disease, the intestinal arthropathies, and Behcet's syndrome. *Medicine (Baltimore)* 1974;53:343–364.

3. Said-Nahal R, Miceli-Richard C, Berthelot JM, Duché A, Dernis-Labous E, Blévec G Le, et al. The familial form of spondylarthropathy: a clinical study of 115 multiplex families. Groupe Français d'Etude Génétique des Spondylarthropathies. *Arthritis Rheum* 2000;43:1356–1365.

4. Said-Nahal R, Miceli-Richard C, D'Agostino MA, Dernis-Labous E, Berthelot JM, Duché A, et al. Phenotypic diversity is not determined by independent genetic factors in familial spondylarthropathy. *Arthritis Rheum* 2001;45:478–484.

5. Porcher R, Said-Nahal R, D'Agostino M-A, Miceli-Richard C, Dougados M, Breban M. Two major spondylarthropathy phenotypes are distinguished by pattern analysis in multiplex families. *Arthritis Rheum* 2005;53:263–271.

6. Dernis E, Said-Nahal R, D'Agostino M-A, Aegerter P, Dougados M, Breban M. Recurrence of spondylarthropathy among first-degree relatives of patients: a systematic cross-sectional study. *Ann Rheum Dis* 2009;68:502–507.

7. Brown MA, Kennedy LG, MacGregor AJ, Darke C, Duncan E, Shatford JL, et al. Susceptibility to ankylosing spondylitis in twins: the role of genes, HLA, and the environment. *Arthritis Rheum* 1997;40:1823–1828.

8. Linden SM van der, Valkenburg HA, Jongh BM de, Cats A. The risk of developing ankylosing spondylitis in HLA-B27 positive individuals. A comparison of relatives of spondylitis patients with the general population. *Arthritis Rheum* 1984;27:241–249.

9. Järvinen P. Occurrence of ankylosing spondylitis in a nationwide series of twins. *Arthritis Rheum* 1995;38:381–383.

10. Pedersen OB, Svendsen AJ, Ejstrup L, Skytthe A, Harris JR, Junker P. Ankylosing spondylitis in Danish and Norwegian twins: occurrence and the relative importance of genetic vs. environmental effectors in disease causation. *Scand J Rheumatol* 2008;37:120–126.

11. Brown MA, Laval SH, Brophy S, Calin A. Recurrence risk modelling of the genetic susceptibility to ankylosing spondylitis. *Ann Rheum Dis* 2000;59:883–886.

12. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet* 2007;39:1329–1337.

13. Reveille JD, Sims A-M, Danoy P, Evans DM, Leo P, Pointon JJ, et al. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nat Genet* 2010;42:123–127.

14. Evans DM, Spencer CCA, Pointon JJ, Su Z, Harvey D, Kochan G, et al. Interaction between

ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet* 2011;43:761–767.

15. International Genetics of Ankylosing Spondylitis Consortium (IGAS), Cortes A, Hadler J, Pointon JP, Robinson PC, Karaderi T, et al. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat Genet* 2013;45:730–738.

16. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;40:695–701.

17. Bailey-Wilson JE, Wilson AF. Linkage analysis in the next-generation sequencing era. *Hum Hered* 2011;72:228–236.

18. Brown MA, Pile KD, Kennedy LG, Campbell D, Andrew L, March R, et al. A genome-wide screen for susceptibility loci in ankylosing spondylitis. *Arthritis Rheum* 1998;41:588–595.

19. Laval SH, Timms A, Edwards S, Bradbury L, Brophy S, Milicic A, et al. Whole-genome screening in ankylosing spondylitis: evidence of non-MHC genetic-susceptibility loci. *Am J Hum Genet* 2001;68:918–926.

20. Zhang G, Luo J, Bruckel J, Weisman MA, Schumacher HR, Khan MA, et al. Genetic studies in familial ankylosing spondylitis susceptibility. *Arthritis Rheum* 2004;50:2246–2254.

21. Miceli-Richard C, Zouali H, Said-Nahal R, Lesage S, Merlin F, Toma C De, et al. Significant linkage to spondyloarthropathy on 9q31-34. *Hum Mol Genet* 2004;13:1641–1648.

22. Amor B, Dougados M, Mijiyawa M. [Criteria of the classification of spondylarthropathies].

Rev Rhum Mal Osteoartic 1990;57:85–89.

23. Dougados M, Linden S van der, Juhlin R, Huitfeldt B, Amor B, Calin A, et al. The European Spondylarthropathy Study Group preliminary criteria for the classification of spondylarthropathy. *Arthritis Rheum* 1991;34:1218–1227.

24. Linden S van der, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis Rheum* 1984;27:361–368.

25. Rudwaleit M, Heijde D van der, Landewé R, Listing J, Akkoc N, Brandt J, et al. The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part II): validation and final selection. *Ann Rheum Dis* 2009;68:777–783.

26. Rudwaleit M, Heijde D van der, Landewé R, Akkoc N, Brandt J, Chou CT, et al. The Assessment of SpondyloArthritis International Society classification criteria for peripheral spondyloarthritis and for spondyloarthritis in general. *Ann Rheum Dis* 2011;70:25–31.

27. Anon. BRLMM: an Improved Genotype Calling Method for the GeneChip Human Mappking 500K Array Set. Available at: [http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf].

28. 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;81:559–575.

29. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. GRR: graphical representation of

relationship errors. *Bioinformatics* 2001;17:742–743.

30. Wigginton JE, Abecasis GR. PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* 2005;21:3445–3447.

31. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;30:97–101.

32. Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* 1997;61:1179–1188.

33. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995;11:241–247.

34. Huang Q, Shete S, Amos CI. Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am J Hum Genet* 2004;75:1106–1112.

35. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;21:263–265.

36. Thiele H, Nürnberg P. HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics* 2005;21:1730–1732.

37. Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered* 2008;66:87–98.

38. Carter KW, Pluzhnikov A, Timms AE, Miceli-Richard C, Bourgain C, Wordsworth BP, et al.

Combined analysis of three whole genome linkage scans for Ankylosing Spondylitis. *Rheumatology (Oxford)* 2007;46:763–771.

39. Shugart YY, Silverberg MS, Duerr RH, Taylor KD, Wang M-H, Zarfas K, et al. An SNP linkage scan identifies significant Crohn's disease loci on chromosomes 13q13.3 and, in Jewish families, on 1p35.2 and 3q29. *Genes Immun* 2008;9:161–167.

40. Liu Y, Helms C, Liao W, Zaba LC, Duan S, Gardner J, et al. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genet* 2008;4:e1000041.

41. Azfer A, Niu J, Rogers LM, Adamski FM, Kolattukudy PE. Activation of endoplasmic reticulum stress response during the development of ischemic heart disease. *Am J Physiol Heart Circ Physiol* 2006;291:H1411–1420.

42. Zhang Y, Zhang M, Wu J, Lei G, Li H. Transcriptional regulation of the Ufm1 conjugation system in response to disturbance of the endoplasmic reticulum homeostasis and inhibition of vesicle trafficking. *PLoS ONE* 2012;7:e48587.

43. Colbert RA, Tran TM, Layh-Schmitt G. HLA-B27 misfolding and ankylosing spondylitis. *Mol Immunol* 2014;57:44–51.

44. McClellan J, King M-C. Genetic heterogeneity in human disease. *Cell* 2010;141:210–217.

45. Brophy S, Hickey S, Menon A, Taylor G, Bradbury L, Hamersma J, et al. Concordance of disease severity among family members with ankylosing spondylitis? *J Rheumatol* 2004;31:1775–1778.

46. Ward MM, Hendrey MR, Malley JD, Learch TJ, Davis JC Jr, Reveille JD, et al. Clinical and immunogenetic prognostic factors for radiographic severity in ankylosing spondylitis. *Arthritis Rheum* 2009;61:859–866.

47. Kadi A, Costantino F, Izac B, Leboime A, Said-Nahal R, Garchon H-J, et al. Brief report: the IL23R nonsynonymous polymorphism rs11209026 is associated with radiographic sacroiliitis in spondyloarthritis. *Arthritis Rheum* 2013;65:2655–2660.

48. Kruglyak L, Lander ES. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* 1995;57:439–454.

49. Kruglyak L. The use of a genetic map of biallelic markers in linkage studies. *Nat Genet* 1997;17:21–24.

50. Middleton FA, Pato MT, Gentile KL, Morley CP, Zhao X, Eisener AF, et al. Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Am J Hum Genet* 2004;74:886–897.

51. Vieland VJ, Walters KA, Azaro M, Brzustowicz LM, Lehner T. The value of regenotyping older linkage data sets with denser marker panels. *Hum Hered* 2014;78:9–16.

52. John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W, et al. Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am J Hum Genet* 2004;75:54–64.

TABLES

Table 1. Clinical characteristics of SpA patients included in the genome-wide linkage study

Characteristic	**All**	**Linked to 13q13***	**Not linked to 13q13****
	(n = 462)	**(n = 134)**	**(n = 328)**
Age in years, mean ± SD	47.1 ± 14.5	45.2 ± 14.3	47.9 ± 14.6
Disease duration in years, mean ± SD	23.1 ± 13.4	26.5 ± 18.0	23.5 ± 16.0
Sex ratio, % of men	51.3	54.5	50.0
HLA-B27 positivity, %	92.4	93.3	92.1
Axial manifestations			
Inflammatory back pain, %	98.1	98.5	97.9
Radiographic sacroiliitis***, %	58.8	59.7	58.4
Peripheral manifestations			
Peripheral arthritis, %	44.2	42.5	44.8
Peripheral enthesitis, %	69.3	70.1	68.9
Extra-articular manifestations			
Uveitis, %	27.9	25.4	29.0
Psoriasis, %	25.5	25.4	25.6
Inflammatory bowel disease, %	6.7	6.7	6.7
Crohn's disease	3.5	3.7	3.35
Ulcerative colitis	3.2	3	3.35
Classification criteria fulfillment			
Amor, %	99.6	99.3	99.7
ESSG, %	94.8	94.0	95.1
ASAS	99.3	99.2	99.4
axial, %	94.3	94.7	94.2
peripheral, %	5.0	4.5	5.2

* refers to patients belonging to families reaching a ponderated LOD score ≥ 0.16 in 13q13 locus;
** refers to patients belonging to families having a ponderated LOD score < 0.16 in 13q13 locus;
*** refers to radiographic sacroiliitis ≥ grade II bilateral or grade III unilateral

Table 2. Characteristics of the SpA multiplex families included in the study

Characteristic	**Family dataset (n =143)**
Individuals per family	
all, average no. (range)	7.57 (4 – 27)
affected, average no. (range)	3.28 (2 – 11)
Generations per family, average no. (range)	2.47 (2 – 6)
Affected relative pairs no.	282
Sib-pairs	2
Half-sibs	62
Cousins	163
Parent-child	10
Grandparent-grandchild	118
Avuncular	

Table 3. Significant (bolded) and suggestive loci resulting from the non-parametric linkage analysis.

Chromosome	Position (Mb)	LOD score
1	73	2.87
1	189	2.55
6	**28.7**	**24.77**
8	48.2	3.06
8	114.2	2.4
9	15.7	3.34
9	130.2	3.51
10	58.7	2.44
13	**39.7**	**5.03**
17	44.0	2.37

FIGURES

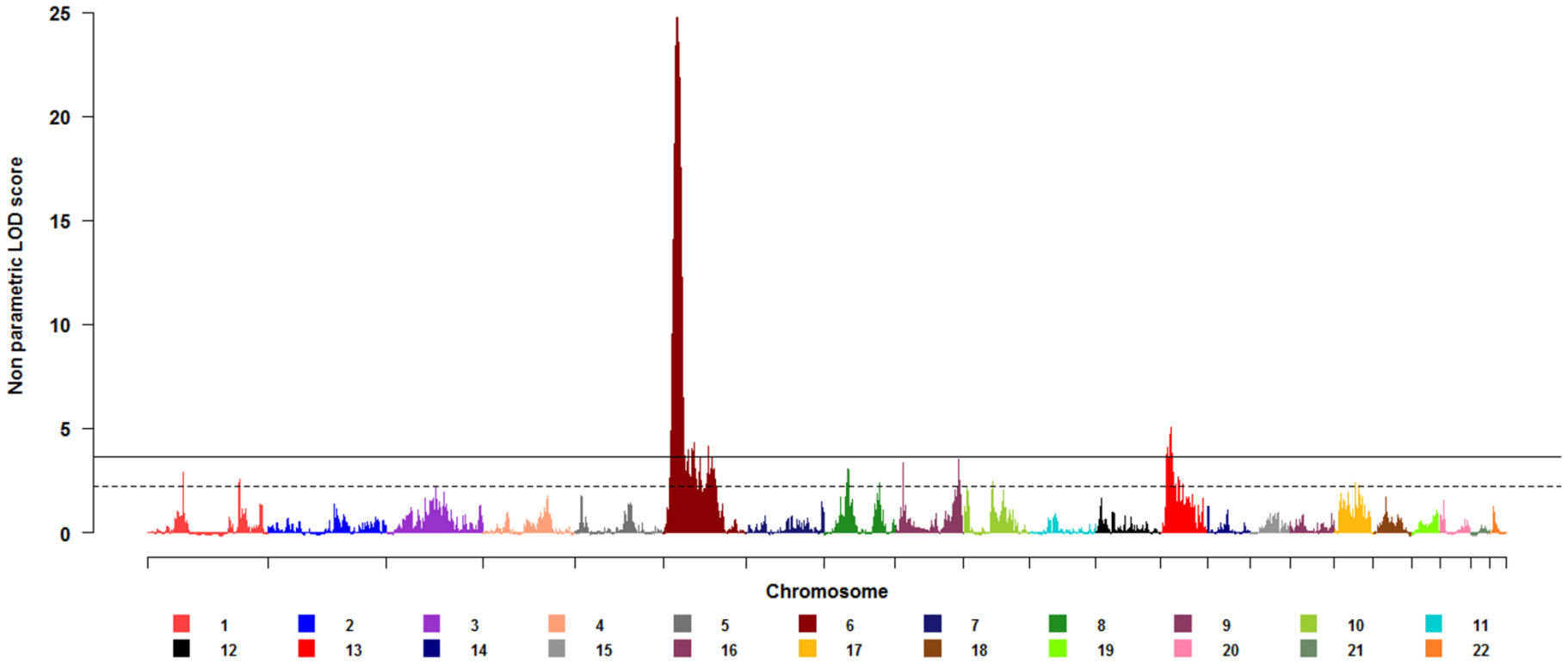


Figure 1. Genome-wide non-parametric linkage analysis results. The x axis indicates the marker positions on the genome with each chromosome represented with a distinct color. The y axis shows the Kong & Cox LOD score. Significant and suggestive LOD score thresholds are represented by solid and dashed line, respectively.

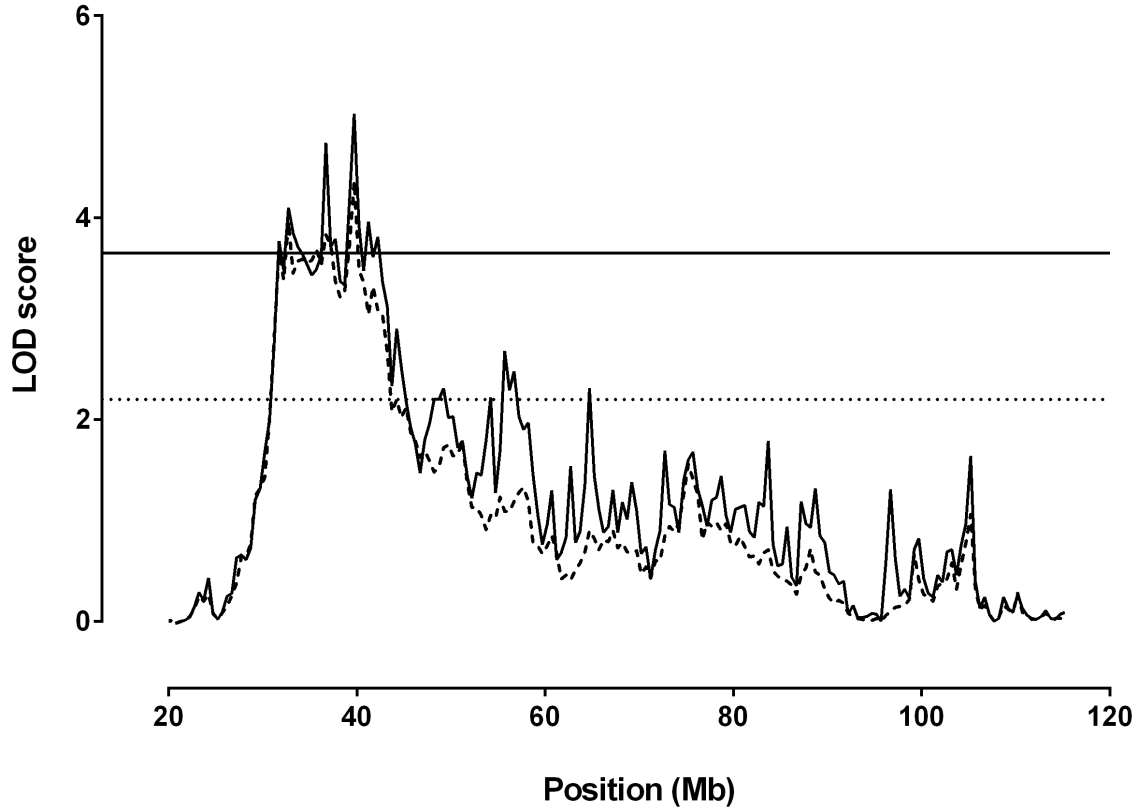


Figure 2. Non-parametric linkage analysis of chromosome 13, taking in account (dashed line) or not (solid line) the LD. The x axis represents the physical distance from the p-telomere and the y axis represents the Kong & Cox LOD score. Significant and suggestive LOD score thresholds are represented by solid and dotted horizontal lines, respectively.

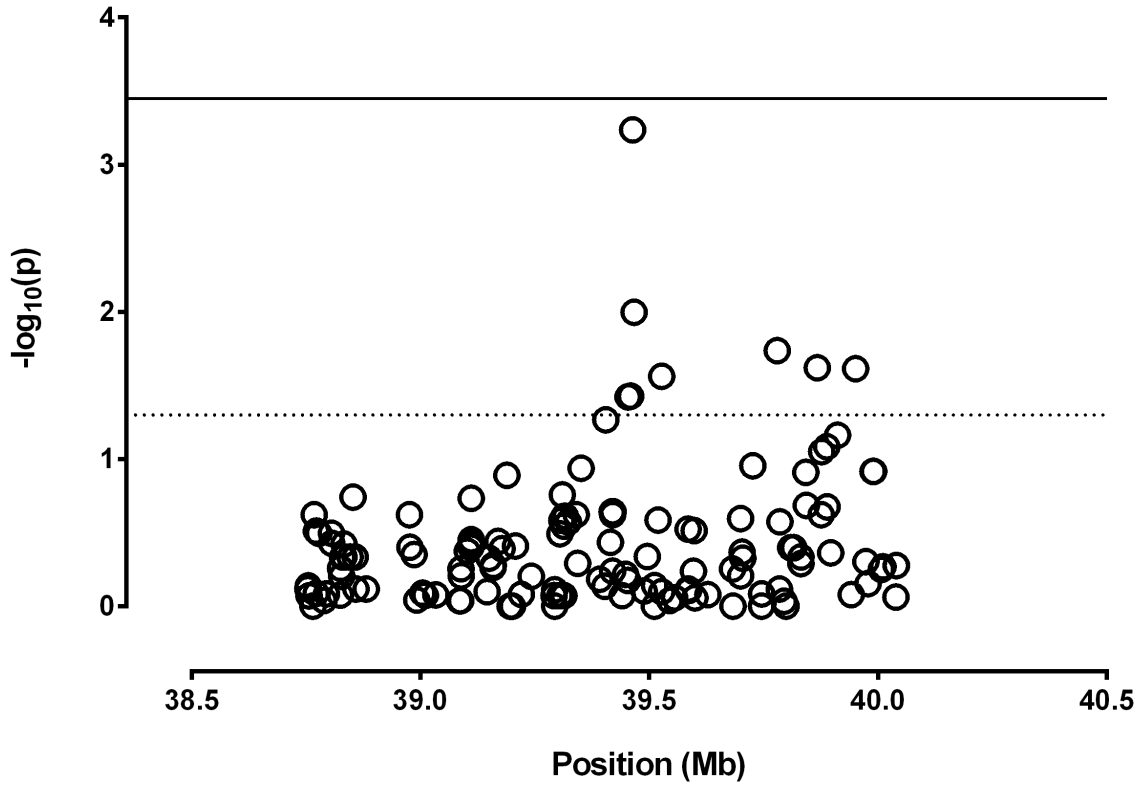
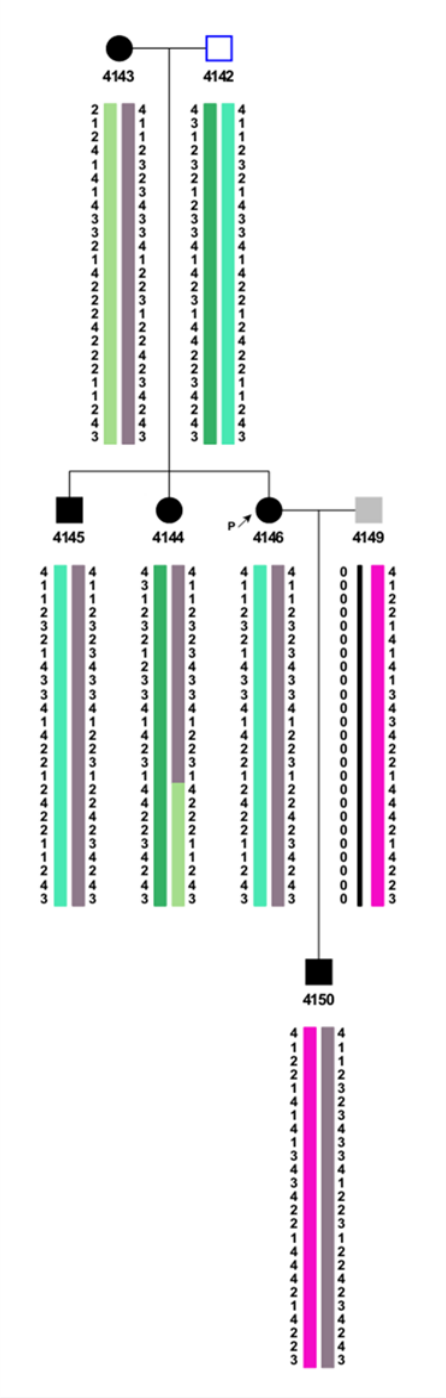
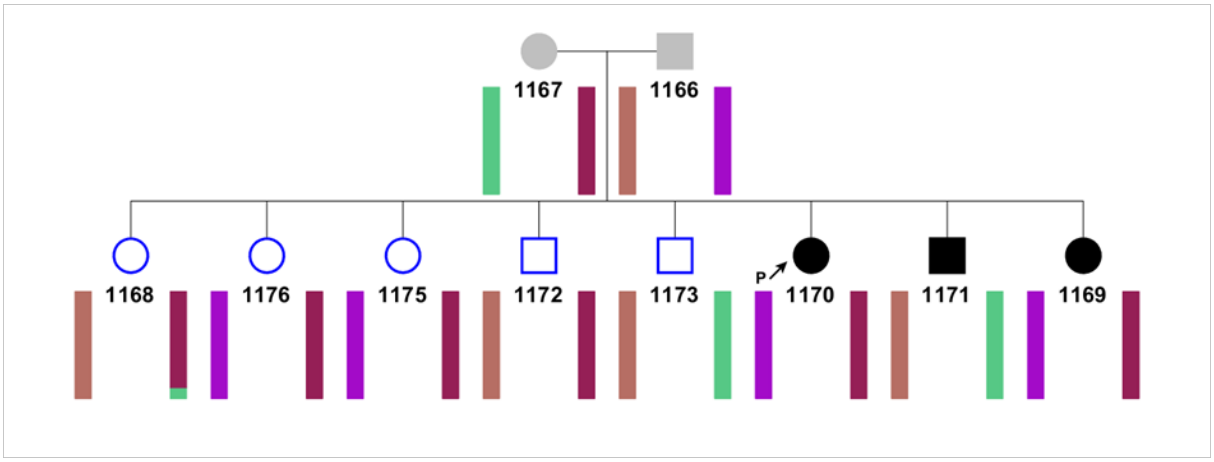


Figure 3. Family-based association study results. All the SNPs of the 13q13 disease interval have been tested for association in the 43 best-linked families (ponderated pLOD ≥ 0.16). The x axis represents position and the y axis represents the negative decimal logarithm of the p-value. Significant p-value threshold according to Bonferroni correction is represented by solid line and the dotted line corresponds to a nominal p-value of 0.05.

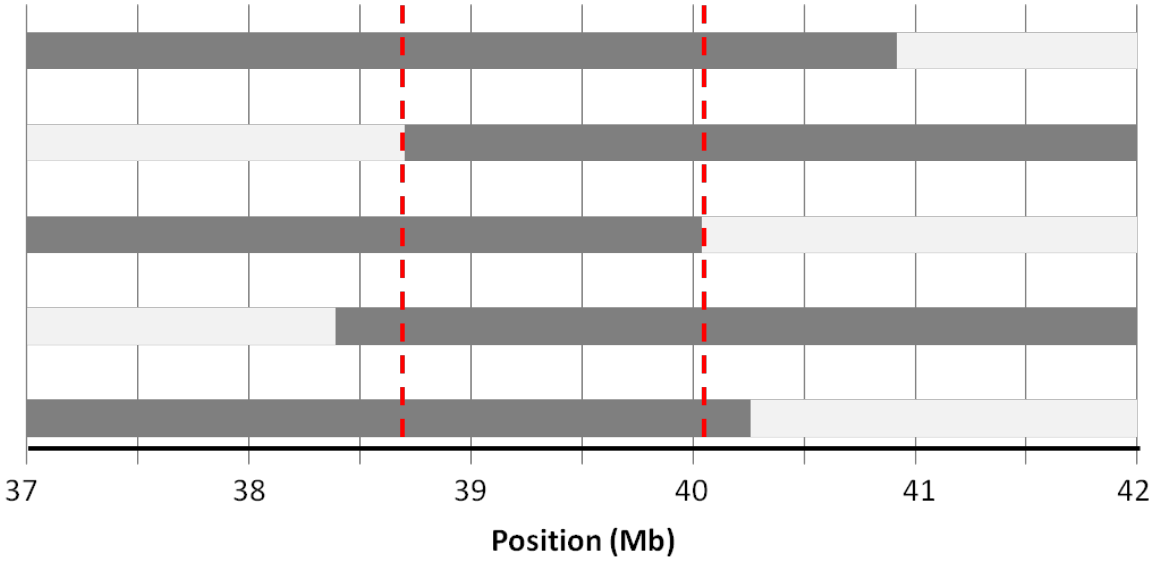


Supplementary figure 1. Example of an informative meiotic recombination event taking place at the 13q13 locus, in one linked pedigree. Haplotypes were reconstructed, using tag-SNP between 37 and 42 Mb from the chromosome 13 p-telomere. All the affected members of the pedigree harbor the grey-coloured haplotype with a recombination event in subject no. 4144. Closed

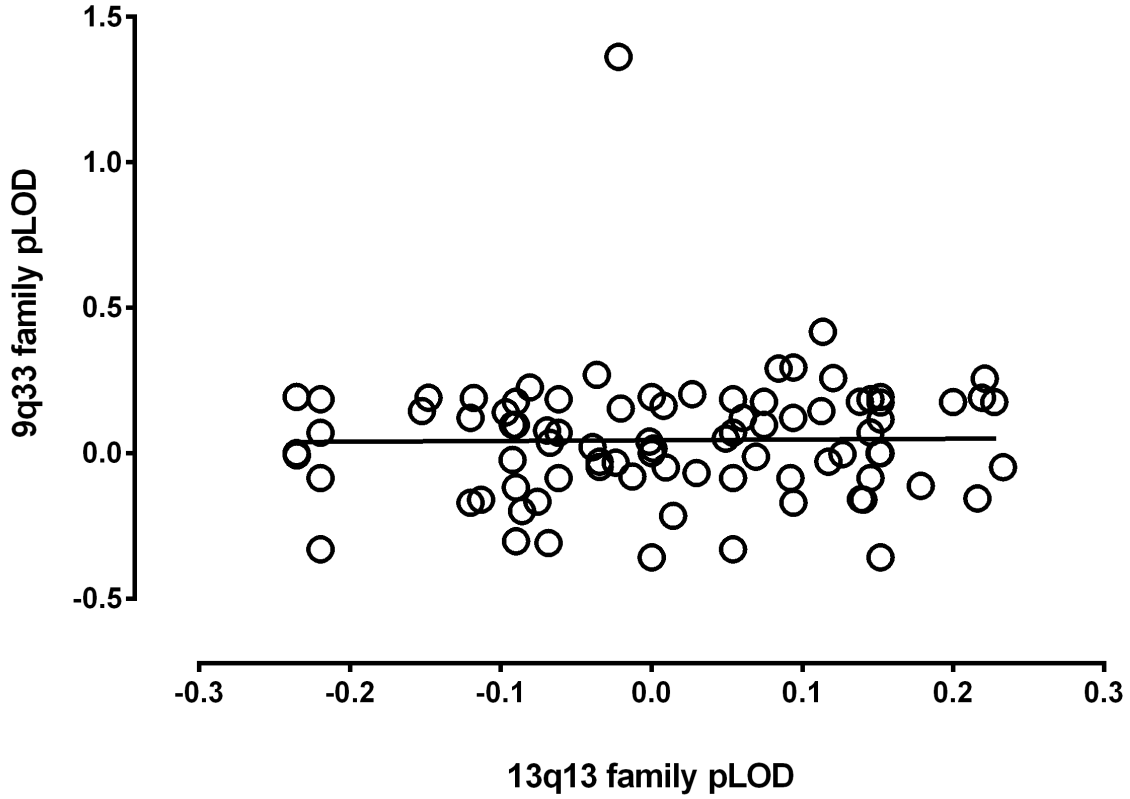
symbols correspond to patients and open ones to nonaffected subjects. Shaded symbols indicate subjects who were not examined and genotyped. Their haplotypes were therefore deduced. SNP alleles are coded numerically.



Supplementary figure 2. MHC haplotype reconstruction of the family no. 92. All the subjects (affected or not) are HLA-B27 negative. Subject no. 1171 does not share any haplotype with his two affected sisters (P: proband).



Supplementary figure 3. Disease interval (framed in red) delimited by informative crossover events in patients. For each informative crossover, the grey bar represents the portion of the recombinant haplotype which segregates with the disease in the other members of the family.



Supplementary figure 4. No correlation between family pLODs at the new 13q13 locus and at the SPA2 locus in 9q33. For each family represented as a dot, the x axis represents the pLOD at 39.7 Mb from the p-telomere of chromosome 13 and the y axis represents the pLOD at 130.2 Mb from the p-telomere of chromosome 9. The black line represents the regression line.

III/ Etude d'extension - cohortes 1 et 2 fusionnées

Afin d'obtenir plus de puissance statistique, nous avions fusionné les données des deux cohortes 1 et 2, ce qui représentait 1 275 individus génotypés dans 210 familles, comprenant 649 patients et 606 témoins. La fusion s'était faite sur les SNPs en commun, ce qui représentait au total 179 907 marqueurs.

Tout comme pour l'analyse NPL de la cohorte 1, les LOD scores n'avaient donné qu'une seule région liée significativement. Cependant, il ne s'agit pas de la région 13q13, qui passait cette fois-ci dans l'intervalle de suggestivité. De manière inattendue, les résultats montraient un pic significatif de liaison au niveau du chromosome 1. Cependant, ce pic était tellement fin que lorsqu'on effectuait une analyse de liaison en prenant en compte le DL (r^2>0,1), il disparaissait, ce qui suggérait plus un effet de DL qu'une réelle liaison à la SpA (Figure 15). 8 régions avaient atteint le seuil de suggestivité, dont deux sur le même chromosome 8 (Figure 14). De même que pour le chromosome 1, le pic le plus fort de suggestivité sur le chromosome 8, situé entre 40 et 60 Mb, disparaissait complètement après la construction des blocs de DL. Il en était de même pour les régions suggestives sur les chromosomes 3 et 9. Seule la région 13q13 conservait sa suggestivité après la prise en compte du DL (Figure 16). Les autres régions suggestives n'étaient pas testées, compte tenu des pics ayant un trop faible recouvrement et de leurs faibles puissances, ainsi que de leur faible recouvrement avec l'étude sur la cohorte 1 seule.

Des analyses paramétriques sur le chromosome 13 étaient effectuées sous Merlin afin de tester si nous détections une liaison plus forte en prenant en compte des modèles de transmission et un facteur d'hétérogénéité. Nous avions donc testé plusieurs modèles, avec un lissage sur 2 cM. Le premier était un modèle de transmission dominant, avec une prévalence de 50 % quelle que soit l'hétérozygotie. Le deuxième modèle était récessif, avec une prévalence de 10^{-3} pour un allèle hétérozygote, comme pour la présence d'un allèle homozygote pour le variant non à risque. Le troisième modèle prenait en compte HLA-B27, en supposant que la transmission était dominante lorsque la personne portait l'allèle (80 %, comme la fréquence des personnes malades HLA-B27 positifs). Le quatrième modèle était récessif en présence de HLA-B27, et le cinquième proposait une co-dominance avec HLA-B27, c'est-à-dire une prévalence de 40 % pour l'allèle hétérozygote. Quoi qu'il en soit, aucun de ces modèles ne donnait un score HLOD satisfaisant, ni même avec un

seuil de suggestivité.

Tous ces résultats confirment l'hétérogénéité génétique très présente dans la SpA. En effet, il suffit d'ajouter la moitié d'une première cohorte dans une étude d'extension pour ne plus répliquer les résultats obtenus auparavant. C'est pourquoi les résultats de cette analyse n'ont pas fait l'objet d'analyses paramétriques plus poussés sur les autres régions détectées. Cependant, le séquençage de la région 13q13 initié dans notre laboratoire promet de découvrir d'autres facteurs génétiques liés à la SpA, permettant d'expliquer mieux l'héritabilité manquante. De plus, une série de séquençage à haut débit de l'exome est en cours, afin d'obtenir un meilleur recouvrement et une précision améliorée dans les résultats. L'exploration des facteurs génétiques liés ou associés à la SpA ne s'arrêtent pas là, et il reste en plus à découvrir des haplotypes pouvant être protecteurs, ce qui représente une approche innovante dans le cadre de cette maladie.

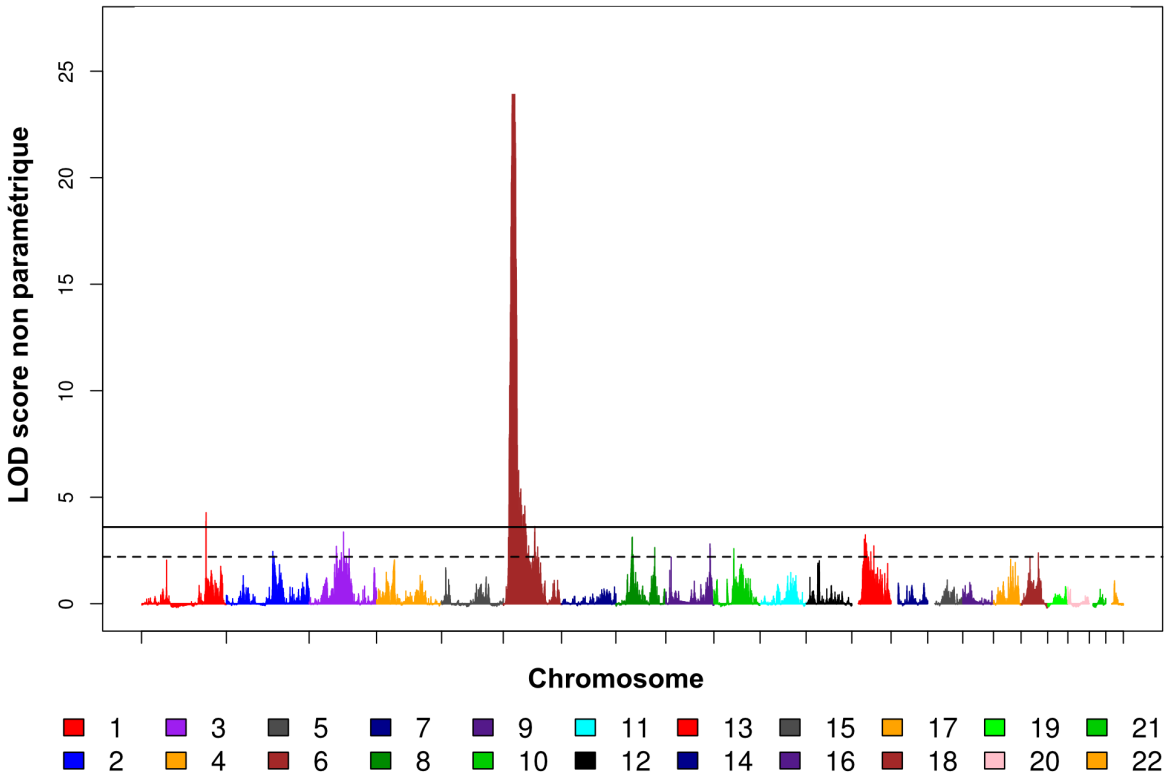


Figure 14: Bilan des analyses de liaison non paramétriques (NPL) sur tout le génome dans les cohortes 1 et 2 fusionnées.

L'axe x représente les coordonnées des SNPs sur les chromosomes, chacune représentée par une couleur

spécifique définie par la légende du dessous. L'axe y représente le LOD score attribué par Merlin après une analyse NPL. Le trait horizontal plein représente le seuil de significativité (3,6), et le trait en pointillés représente le seuil de suggestivité (2,2), tels qu'ils sont définis pour les analyses NPL sur génome entier.

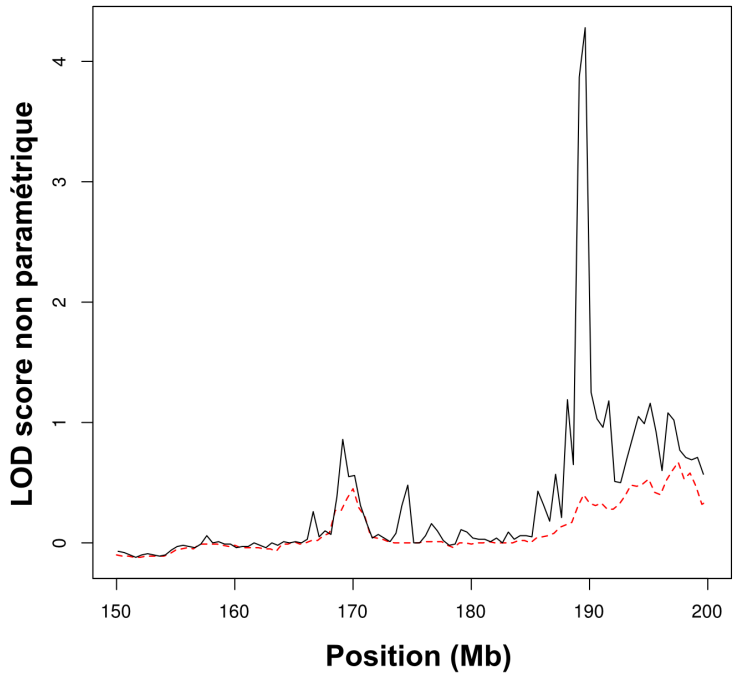


Figure 15: Analyse NPL du chromosome 1 sur les deux cohortes fusionnées.

Le trait noir solide représente l'analyse NPL sans prise en compte du DL (seulement un lissage sur 0,5 cM). La ligne rouge en pointillés représente la même analyse après construction des blocs de DL (r^2>0,1).

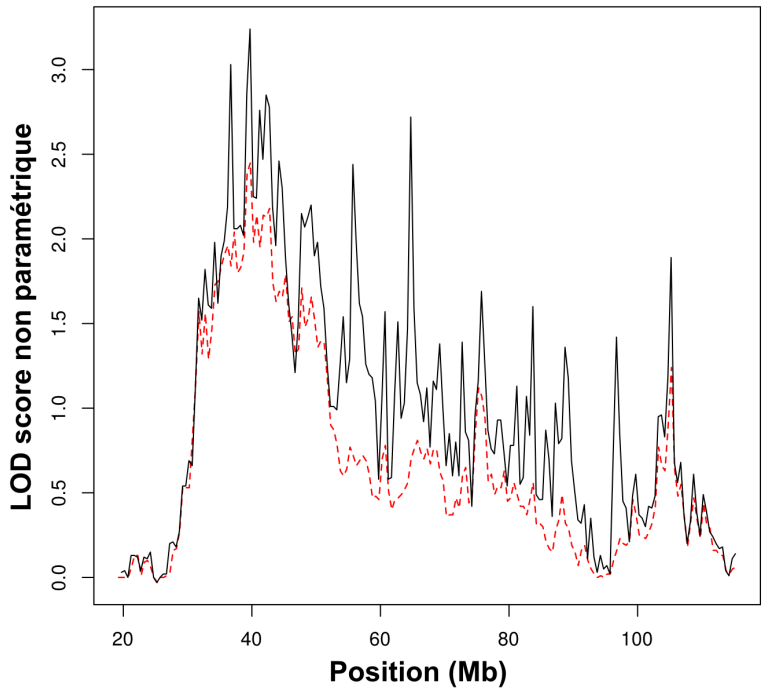


Figure 16: Analyse NPL du chromosome 13 sur les deux cohortes fusionnées.

IV/ Analyse de la protection - cohortes 1, 2 et 3 fusionnées

Nous avions par la suite ajouté les données de la cohorte 3 à celles des deux premières cohortes fusionnées, pour un total de 1 310 individus dans 210 familles, avec 668 patients et 620 témoins. Après les filtrages effectués par PLINK, 178 881 SNPs pouvaient être testés le long du génome pour des analyses de liaison.

Étant donné la faible taille de la cohorte 3, nous n'avions pas de nouveau entrepris des analyses NPL classiques sur toutes ces données, estimant que les résultats seraient très similaires. Cependant, nous avions construit les blocs de DL sur chacund des chromosomes (r^2>0,1), afin d'accélérer les calculs qui suivaient. Nous avions ensuite entrepris d'effectuer une analyse de liaison des locus qui confèreraient une protection aux témoins HLA-B27 positifs. Pour cela, nous avions inversé les statuts des individus en considérant tous les patients comme des témoins sains et les témoins HLA-B27 positifs comme des patients. Seuls les témoins HLA-B27 positifs dont l'âge de dernier suivi était supérieur ou égal à 25 ans étaient considérés transformés en patients.

Pour ceux de moins de 25 ans, un statut inconnu était alors entré. Nous n'avions sélectionné ensuite que les familles présentant au moins deux témoins HLA-B27 positifs originaux génotypés, afin de ne garder que les familles informatives. Cela donnait une sous-cohorte de 64 familles d'une taille moyenne de 10,7 personnes, avec entre 4 et 29 individus dans chaque famille. Parmi tous ces individus, 551 étaient génotypés, les 134 autres étant ajoutés pour mieux expliquer les structures familiales. Les individus génotypés se recoupaient entre 362 nouveaux témoins transformés (188 hommes et 174 femmes), et 175 nouveaux patients (79 hommes et 96 femmes). Nous estimions donc que cette cohorte était suffisamment grande pour pouvoir effectuer des analyses NPL sur tout le génome par Merlin, afin de tester la présence d'haplotypes protecteurs.

Après ces analyses, le chromosome 6 conservait sa significativité, avec un LOD score légèrement supérieur à 5. Cependant, le seul autre locus ayant dépassé le seuil de suggestivité était le chromosome 22, avec un LOD score maximal de 2,25 situé entre 31,5 et 33,1 Mb (Figure 17), à la jonction entre les régions 22q12.2 et 22q12.3. Parmi les 15 gènes codants dans cette région, 8 avaient un rôle bien défini dans la base de données GeneCards, qui étaient SCL5A1, DEPDC5, SYN3, TIMP3, YWHAH, FBXO7, RTCB et PISD. Parmi ceux-ci, TIMP3 est un inhibiteur des matrices de métalloprotéinases, comme la collagénase, et peut prendre part à une réponse aïgue spécifique du tissu à des stimuli de remodelage. Les autres gènes ont des rôles éloignés du système immunitaire et ne donnent pas de lien direct avec la SpA ou le système immunitaire.

Bien que ces résultats soient décevants au niveau de la non-découverte de régions significatives, il fallait prendre en compte que la taille de la cohorte représentée était restreinte par rapport aux autres étudiées plus haut. Cela était dû au fait que les individus n'étaient pas recrutés dans cette optique, et les familles n'étaient donc pas toutes informatives. Cette approche est novatrice, et mérite de concevoir une expérience de génotypage spécialement dédiée à l'analyse de ces haplotypes protecteurs.

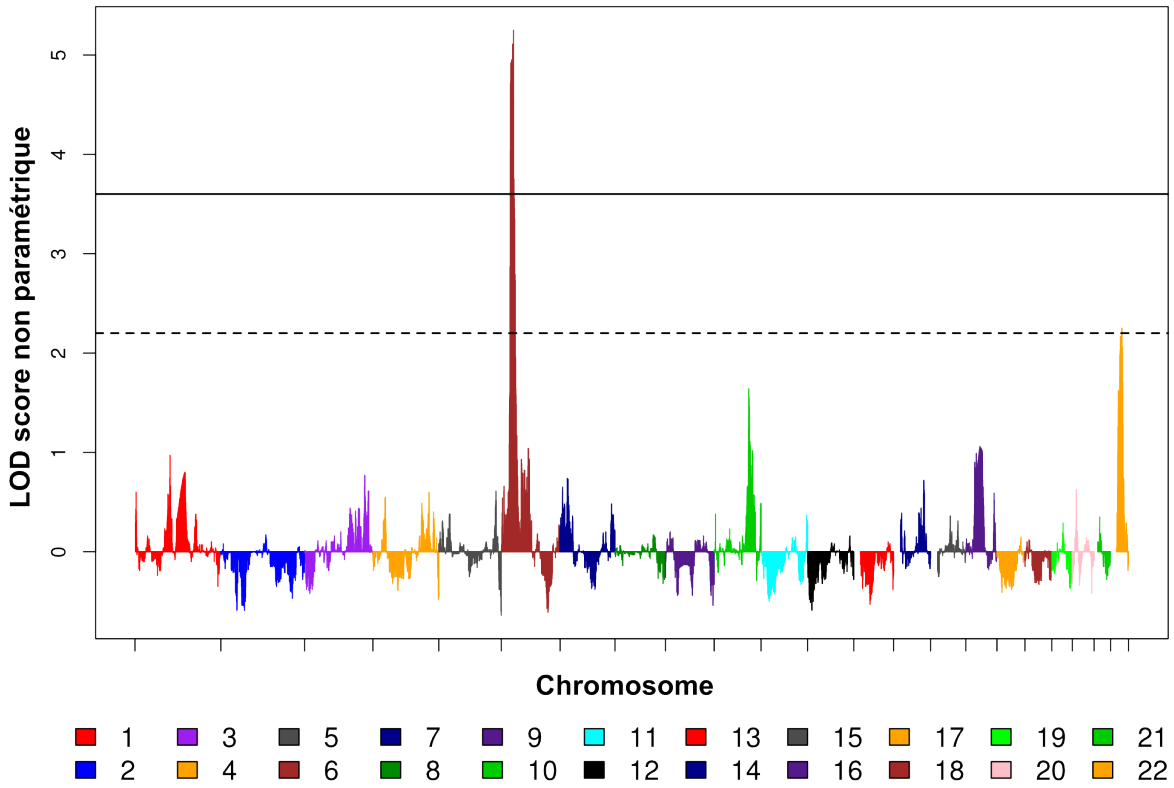


Figure 17: Bilan des analyses de liaison non paramétriques (NPL) sur tout le génome dans les cohortes 1, 2 et 3 fusionnées, après inversion des statuts chez les témoins HLA-B27 positifs.

L'axe x représente les coordonnées des SNPs sur les chromosomes, chacune représentée par une couleur spécifique définie par la légende du dessous. L'axe y représente le LOD score attribué par Merlin après une analyse NPL. Le trait horizontal plein représente le seuil de significativité (3,6), et le trait en pointillés représente le seuil de suggestivité (2,2), tels qu'ils sont définis pour les analyses NPL sur génome entier.

Analyses des gènes DE à la SpA

I/ Méta-analyse

Lorsque nous avions obtenu les données de l'étude 2 pour les données transcriptomiques, nous voulions dans un premier temps savoir si les données se répliquaient avec l'étude 1. Pour cela, nous avions effectué les comparaisons sous LIMMA entre les patients HLA-B27 positifs et les témoins HLA-B27 négatifs dans les deux études séparément, en ajustant pour tous les temps de stimulation LPS. Nous considérions un gène comme DE lorsque la p-value LIMMA obtenue était en dessous de 5 %. L'étude 1 donnait 1383 gènes DE, tandis que ceux de l'étude 2 en détectaient 808. Malheureusement, les deux études se recoupaient très faiblement, avec 7,55 % des gènes de l'étude 2 en commun avec l'étude 1. De plus, parmi les 61 gènes communs, 26 étaient discordants dans leurs fold changes, ce qui représentait 42,62 % de ces gènes (Figure 18). Nous avions voulu tester une méthode plus sensible, maxP, comme décrit dans les matériels et méthodes. De nouveau, les résultats ne donnaient que 241 gènes communs, soit 29,83 %, qui comptaient eux-mêmes un taux de discordance élevé (34,85%). Ces résultats, bien que décevants, n'étaient pas surprenants, car il arrive souvent que des études transcriptomiques sur génome entier se recoupent peu par la forte variabilité intra-individuelle des expressions des gènes et de la faible taille d'échantillon des études.

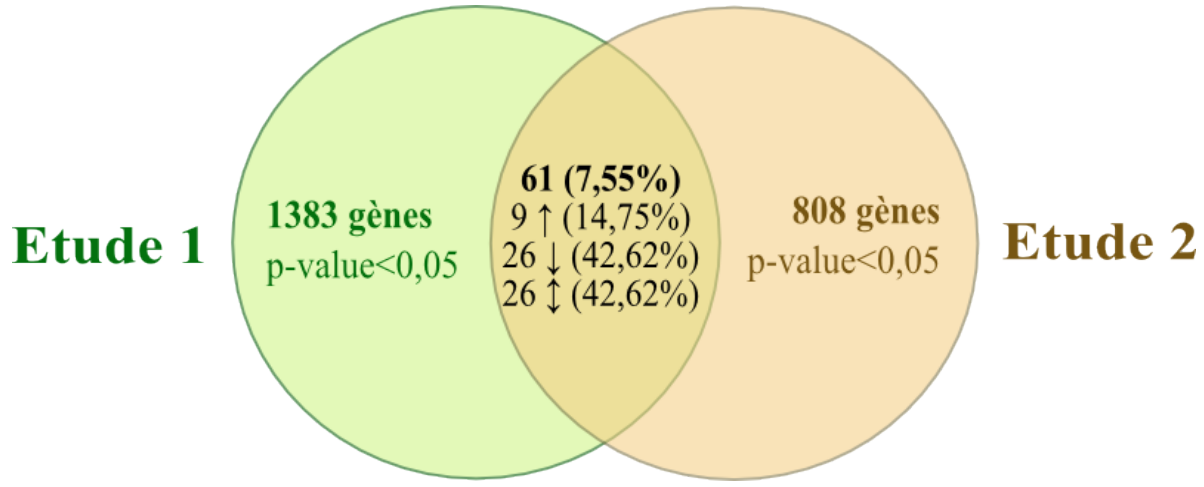


Figure 18: Diagramme de venn des résultats d'analyses LIMMA entre les deux études transcriptomiques.

Les nombres dans chacun des cercles séparés comprennent tous les gènes DE de l'étude correspondante, y compris les gènes communs.

C'est pourquoi nous nous étions demandés s'il était possible de fusionner les deux études afin d'augmenter la puissance statistique de nos résultats. Pour cela, nous nous étions d'abord posés la question de savoir si les qualités des données des deux études étaient homogènes, malgré les discordances que nous trouvions dans les analyses séparées. Nous avions donc utilisé MetaQC sur celles-ci, avec 8 études publiques supplémentaires. L'ACP sur les scores de qualité calculés donnait une proximité très forte entre les données de nos deux études, comparées à toutes les autres. Par ailleurs, nos deux études se trouvaient très proches des deux études impliquant des MD-DCs et de celle impliquant des DCs inflammatoires, ce qui nous confortait dans l'intégrité des données et du type cellulaire étudié. De plus, ces résultats nous permettaient de fusionner les données de nos deux études en toute confiance, avec toutefois une correction sur l'étude dans les modèles linéaires utilisés par LIMMA ou une normalisation entre les deux études pour les calculs de corrélation. En effet, un biais pouvait subsister lorsque nous ne le prenions pas en compte.

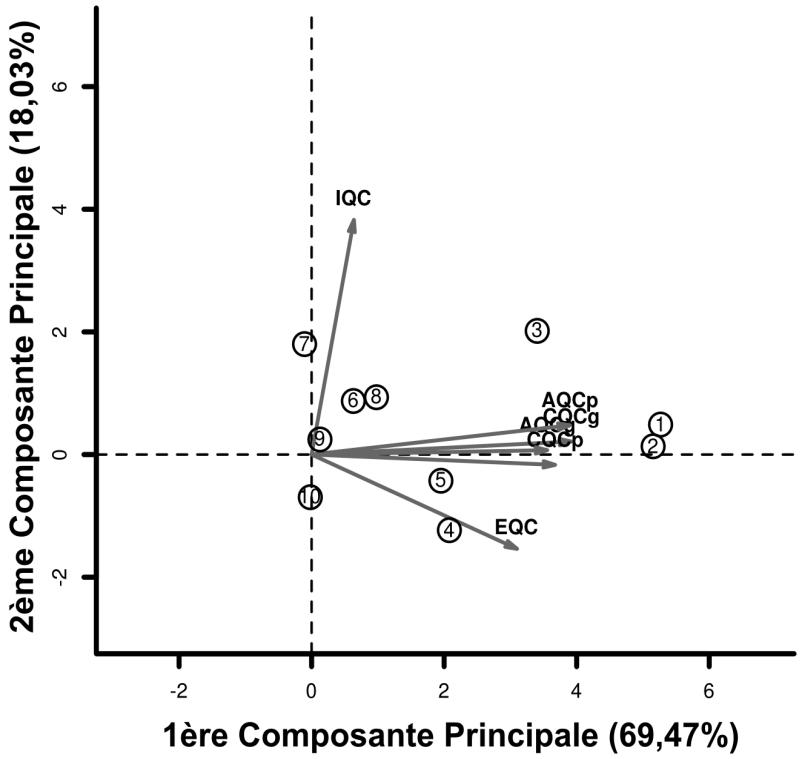


Figure 19: ACP sur les résultats MetaQC incluant nos deux études transcriptomiques.

Les pourcentages entre parenthèses des labels des axes représentent la variance expliquée par les composantes principales. Les numéros 1 et 2 représentent la cartographie au niveau des scores de qualité des études 1 et 2. Les études 3 et 4 représentent les études impliquant les MD-DCs, et l'étude 5 est celle qui implique les DCs inflammatoires.

II/ Analyses des gènes DE

Nous avions par la suite voulu déterminer quels étaient les gènes DE en fonction de différentes comparaisons effectuées sur nos deux études fusionnées. Dans un premier temps, nous avions défini les modèles linéaires par LIMMA en décrivant les données par une matrice de design prenant en compte les interactions de nos trois facteurs principaux, à savoir le temps de stimulation LPS, le statut HLA-B27 et le statut maladie. C'était donc à partir des coefficients calculés par LIMMA que nous avions effectué trois comparaisons principales, générant ainsi trois listes : la liste A comparant les patients HLA-B27 positifs et les témoins HLA-B27 négatifs ; la liste B comparant les témoins HLA-B27 positifs et les témoins HLA-B27 négatifs ; et la liste C comparant les patients HLA-B27 positifs et les témoins HLA-B27 positifs. Pour chacune de ces listes, les analyses temporelles suivantes étaient prises en compte en fonction de la stimulation LPS : H0, H6-H0, H6, H24-H0, H24 et tous temps. H6-H0 et H24-H0 testaient des réponses différentielles des gènes après 6 heures ou 24 heures de stimulation LPS, par rapport à l'état basal sans stimulation. Tous temps correspondait à une analyse ajustée pour tous les temps de stimulation.

Les listes A et C ne donnaient presque aucun gène avec une p-value corrigée de Benjamini et Hochberg. Afin d'avoir tout de même un contrôle qualité sur nos résultats, nous avions utilisé des approches de bootstrap pour vérifier la stabilité de notre signal. Les corrélations obtenues pour chaque comparaison entre les moyennes des logarithmes des p-values fournies par LIMMA issues du bootstrap et celles des données originales étaient très convaincantes, avec un r^2 en moyenne de 0,92 ($\pm$ 0,04). A chaque fois, les p-values étaient inférieures à 2.10^{-16}. Cela nous confortait ainsi dans la force du signal observé. Par ailleurs, nous avions permuté les méta-données attribuées à chaque échantillon, afin d'obtenir une distribution empirique et de s'affranchir de la limite de LIMMA qui se basait sur une distribution de Student pour les expressions des gènes. Ainsi, la p-value de permutation servait de filtre supplémentaire sur les gènes DE, avec un retrait d'une moyenne de 20,56% ($\pm$ 9,71%) des gènes pour un seuil à 5 %. Ce filtrage supplémentaire retirait une part beaucoup plus importante de gènes DE pour les analyses ajustées à tous les temps LPS,

avec une moyenne de 38,36 % (Tableau 10).

Liste	Analyse temporelle	r^2 obtenu par bootstrap	Gènes retirés par permutation	Nombre de gènes DE final
A	**H0**	0,7698	18,62 %	647
	H6-H0	0,9152	14,66 %	1980
	H6	0,9139	15,39 %	2518
	H24-H0	0,9038	23,59 %	991
	H24	0,9115	23,80 %	1265
	Tous temps	0,9314	34,32 %	798
B	**H0**	0,9143	8,55 %	2545
	H6-H0	0,9573	6,60 %	3695
	H6	0,9155	15,87 %	2088
	H24-H0	0,9537	11,66 %	3970
	H24	0,9210	25,40 %	1974
	Tous temps	0,9520	40,05 %	681
C	**H0**	0,8880	15,32 %	1343
	H6-H0	0,9545	14,69 %	2793
	H6	0,9177	18,64 %	1829
	H24-H0	0,9547	18,69 %	2641
	H24	0,9251	23,53 %	1901
	Tous temps	0,9429	40,71 %	466
Moyenne		**0,9190**	**20,56 %**	**1895,83**
Ecart-type		**0,0425**	**9,71 %**	**1016,62**

Tableau 10: Résultats des approches de bootstrap et de permutations sur les analyses LIMMA.

La liste correspond aux comparaisons décrites dans les résultats. Les gènes retirés par permutation sont représentés sous forme de pourcentage des gènes avec une p-value de permutation inférieure à 5 % parmi ceux avec une p-value LIMMA inférieure à 5 %. Le nombre de gènes DE final correspond aux gènes avec les p-values LIMMA et de permutation inférieurs à 5 %.

Nous avions donc effectué les trois comparaisons, suivies d'un calcul de p-values empiriques par permutation, et filtré les gènes à la fois avec la p-value LIMMA et la p-value de permutation inférieures à 5 %. Ainsi, la liste A représentait les gènes DE pouvant être affectés par la maladie ou l'allèle HLA-B27. La liste B représentait les gènes affectés spécifiquement par HLA-B27, tandis que la liste C regroupait ceux affectés par la SpA seule. Afin d'avoir une liste plus robuste

de gènes spécifiquement affectés par la SpA, nous avions effectué un croisement de liste, (A-B)∩C, avec les fold change entre A et C devant être concordants. Ainsi, nous avions effectué ce croisement de listes à toutes les comparaisons prenant en compte le temps de stimulation LPS. Nous avions déjà pu remarquer que les gènes DE issus de la liste B, c'est-à-dire affectés par l'allèle HLA-B27, représentaient un nombre non négligeable et présentaient même des p-values ajustées de Benjamini et Hochberg très significatives, ce qui confirmait l'importance de prendre en compte l'effet de ce facteur génétique dans l'identification des gènes dont l'expression était affectée spécifiquement chez les patients atteints de SpA (Tableau 10). Par ailleurs, comme pour l'étude 1, les gènes étaient majoritairement DE dans les MD-DCs des patients atteints de SpA après 6h de stimulation LPS, que ce soit au temps fixe ou en comparaison avec H0. Par ailleurs, les gènes étaient majoritairement sur-exprimés, avec 85,91 % à H6-H0 et 68,03 % à H6. Les gènes DE sans stimulation LPS étaient peu nombreux, tout comme dans l'analyse ajustée à tous les temps (Tableau 11). Cela confirme bien qu'il est important dans notre étude d'effectuer des analyses qui prennent en compte chaque temps de stimulation LPS séparément.

Analyse temporelle	Nombre de gènes DE			P-value LIMMA minimale	
	Total (HGNC)	Sur-exprimés	Sous-exprimés	Liste A	Liste C
H0	50 (49)	24	26	$9{,}13.10^{-4}$	$7{,}68.10^{-5}$
H6-H0	369 (353)	317	52	$3{,}64.10^{-6}$	$7{,}18.10^{-6}$
H6	466 (443)	317	149	$4{,}71.10^{-6}$	$1.82.10^{-5}$
H24-H0	98 (96)	39	59	$4{,}75.10^{-5}$	$9{,}22.10^{-5}$
H24	200 (194)	107	93	$2{,}26.10^{-6}$	$2{,}37.10^{-5}$
Tous temps	64 (59)	22	42	$7{,}45.10^{-4}$	$2{,}77.10^{-6}$

Tableau 11: Synthèse des croisements de listes (A-B)∩C à chaque analyse LIMMA temporelle de stimulation LPS.

Les nombres totaux des gènes entre parenthèses correspondent aux gènes avec un identifiant Ensembl et un symbole HGNC.

Parmi les gènes DE pour l'analyse ajustée à tous les temps dans le croisement de listes (A-B)∩C, celui avec la p-value LIMMA la plus significative dans la liste C était PCOLCE2. La p-value ajustée de Benjamini et Hochberg était relativement faible comparée aux autres gènes identifiés comme DE, avec une valeur de 6,11 %. Il était sur-exprimé chez les patients avec un fold change de 2,35 dans la liste A et de 3,57 dans la liste C (Figure 20). Ce gène est connu pour interagir avec

BMP1, une enzyme impliquée dans l'ossification et la formation du cartilage.[221] Des mutations causant la perte de fonction de BMP1 peuvent induire une ostéogenèse imparfaite, et donc une fragilité osseuse importante.[222] Il est donc possible qu'un gain de fonction de PCOLCE2 et de la voie de signalisation BMP1 soit impliqué dans la pathogenèse de la SpA, et notamment dans l'ankylose.

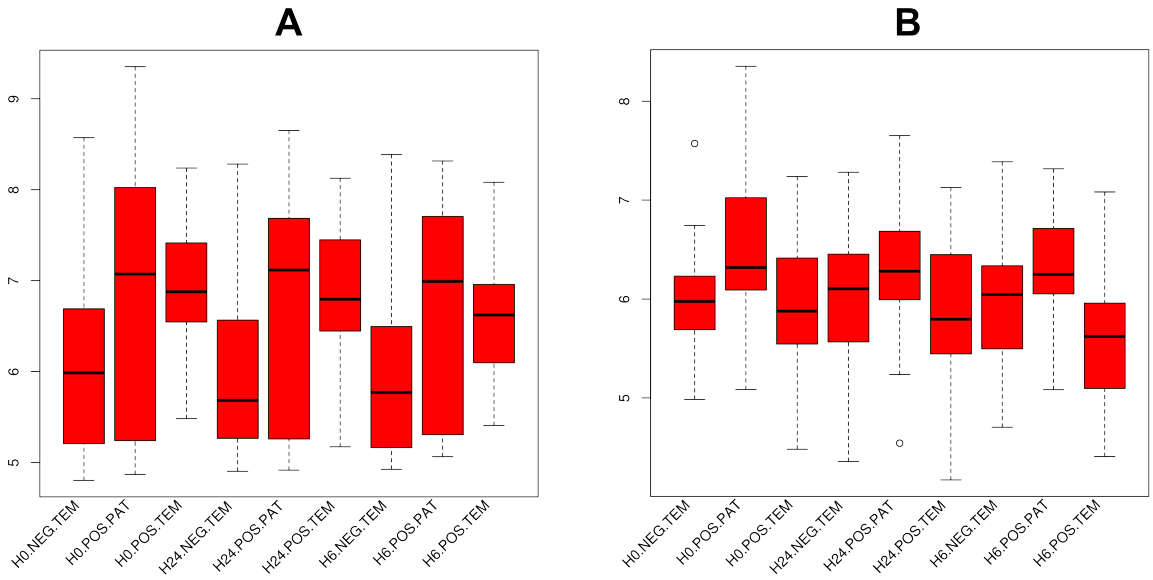


Figure 20: Données d'expression de PCOLCE2 par Affymetrix dans les différents groupes d'individus et aux différents temps de stimulation LPS.

Les données d'expressions de PCOLCE2 sont représentées sans la correction sur le facteur étude (A), et avec correction (B). Nous pouvons remarquer que la correction est primordiale pour éviter un biais statistique. Enfin, nous pouvons voir que globalement, PCOLCE2 est plus exprimé à tous les temps chez les patients (POS.PAT) par rapport aux témoins HLA-B27 positifs (POS.TEM) et HLA-B27 négatifs (NEG.TEM).

Nous avions voulu savoir quels gènes étaient également DE dans tous nos croisements de listes aux analyses temporelles décrites parmi les 73 gènes publiés comme affectés par la SpA dans l'étude 1.[197] Parmi ceux-là, 18 étaient DE à différentes analyses temporelles, soit 24,66 %, avec des sens de variation qui concordaient (Tableau 12). 2 gènes, ENY2 et RPL10A, se retrouvaient également sous-exprimés dans notre étude à tous les temps de stimulation LPS. Par ailleurs, RPL10A était aussi DE à H6, et c'était le seul qui l'était à différentes analyses temporelles. LRRC4 était sous-exprimé à H24-H0, et il représentait le seul gène DE à une analyse temporelle différentielle entre H0 et un temps suivant. 5 gènes DE à H0 se recoupaient dans notre analyse

avec l'étude 1, avec notamment MNDA sous-exprimé et ADAMTS15 sur-exprimé. FBXL4 et FBXO18 étaient sur-exprimés à H24. Enfin, c'était au temps H6 que nous retrouvions le plus de gènes DE en commun, avec 9 gènes sous-exprimés.

Nom du gène	Analyse temporelle	Liste A		Liste C	
		Fold change	P-value LIMMA	Fold change	P-value LIMMA
ADAMTS15	H0	1,163	$6{,}21.10^{-3}$	1,160	$7{,}61.10^{-3}$
KIAA0907	H0	1,192	$1{,}40.10^{-3}$	1,145	$1{,}37.10^{-2}$
SIGLEC15	H0	1,169	$1{,}28.10^{-2}$	1,210	$2{,}53.10^{-3}$
TBCK	H0	1,157	$2{,}94.10^{-2}$	1,173	$1{,}79.10^{-2}$
MNDA	H0	0,764	$3{,}27.10^{-3}$	0,812	$2{,}31.10^{-2}$
ANKRD36BP1	H6	0,766	$5{,}13.10^{-4}$	0,822	$1{,}04.10^{-2}$
CKAP2	H6	0,814	$1{,}24.10^{-2}$	0,747	$4{,}71.10^{-4}$
CRTAP	H6	0,904	$4{,}28.10^{-2}$	0,892	$2{,}33.10^{-2}$
HAUS1	H6	0,831	$5{,}09.10^{-3}$	0,841	$9{,}14.10^{-3}$
MUT	H6	0,840	$1{,}89.10^{-3}$	0,849	$3{,}55.10^{-3}$
RBBP9	H6	0,833	$1{,}06.10^{-2}$	0,867	$4{,}60.10^{-2}$
RPL10A	H6	0,854	$8{,}17.10^{-3}$	0,876	$2{,}64.10^{-2}$
SAP130	H6	0,829	$4{,}71.10^{-6}$	0,874	$8{,}83.10^{-4}$
SLU7	H6	0,843	$1{,}36.10^{-4}$	0,896	$1{,}37.10^{-2}$
LRRC4	H24-H0	0,819	$4{,}18.10^{-2}$	0,787	$1{,}15.10^{-2}$
FBXL4	H24	1,221	$5{,}66.10^{-3}$	1,242	$2{,}81.10^{-3}$
FBXO18	H24	1,109	$1{,}85.10^{-2}$	1,099	$3{,}14.10^{-2}$
ENY2	Tous temps	0,753	$5{,}12.10^{-3}$	0,811	$4{,}58.10^{-2}$
RPL10A	Tous temps	0,724	$9{,}60.10^{-3}$	0,719	$1{,}10.10^{-2}$

Tableau 12: Gènes DE dans (A-B)∩C étant aussi DE dans l'étude 1.

Les gènes en rouge sont sur-exprimés, alors que les bleus sont sous-exprimés chez les patients SpA.

Nous avions également voulu utiliser une approche de biologie intégrative en recherchant les gènes DE identifiés par les croisements de listes situés dans des régions génomiques associées ou liées à la SpA. Pour cela, nous avions regardé : les gènes situés sur les régions liées 13q13, 9q31-4 et 16q23.3 ; les régions associées 2p15 et 21q22 ; les 26 gènes identifiés comme associés par l'IGAS en 2013 ; les principaux gènes localisés dans la région du CMH en prenant les noms qui commencent par « HLA- ». Cela représentait donc 713 gènes avec symboles HGNC disponibles dans les puces Affymetrix utilisées pour nos profils transcriptomiques. Parmi ceux-ci, 40 étaient

identifiés comme DE dans (A-B)∩C et dans nos différentes analyses prenant en compte le temps de stimulation LPS (Tableau 13). 8 gènes étaient identifiés à plusieurs analyses temporelles. BRCA2, situé dans la région 13q13, était sous-exprimé à tous les temps, H6 et H0. FCN2 et LCN12, situés dans la région 9q34.3, étaient sur-exprimés aux temps H6-H0 et H6. Dans la même région, ARRDC1 était sur-exprimé à tous les temps et à H24, et LCN9 était sur-exprimé à H6-H0 mais sous-exprimé à H0. KRTAP10-10 et MRAP, situés dans la région 21q22, étaient sur-exprimés aux temps H6-H0 et H6. PCP4, situé à la même région, était sous-exprimé à tous les temps et à H0. Un autre gène intéressant sur-exprimé à H6 était AIRE, qui est situé dans la région 21q22.3 et qui est un régulateur de l'auto-immunité en présentant aléatoirement des antigènes du soi aux cellules immunitaires. COL27A1 est un gène lié aussi à la synthèse du collagène, situé dans la région 9q32, et était sur-exprimé à H6. La région la plus représentée par nos listes de gènes était 9q34.3, avec 7 gènes DE. Ensuite, les régions 21q22.3, q22.11, q22.13 et q22.2 comptaient respectivement 6, 4, 3 et 3 gènes DE. Parmi les gènes associés à la SpA, seul BACH2 était identifié comme DE et était sur-exprimé au temps H6.

Tous ces résultats laissent à penser que les gènes que nous retrouvons comme DE dans nos croisements de listes ont une certaine pertinence biologique avec les facteurs génétiques identifiés comme associés à la SpA dans des études précédentes. Cependant, il est difficile de dire quels sont les gènes réellement clés dans la physiopathologie parmi les 986 gènes uniques avec symboles HGNC identifiés dans toutes nos listes. De plus, les régulations des gènes se font rarement de façon isolée, et il est important de savoir quelles peuvent être les voies de signalisation à l'origine des changements d'expressions observés. C'est pourquoi d'autres approches complémentaires, comme l'enrichissement des jeux de gènes, sont importants, même si l'analyse des gènes DE reste primordial pour la suite des analyses.

Nom du gène	Région	Analyse(s) temporelle(s)	Fold change(s)	P-value(s)
XPO1	2p15	H24	1,097	$2{,}63.10^{-2}$
HLA-DOA	6p21.32	H24	0,748	$4{,}04.10^{-3}$
HLA-DPA1	6p21.32	H6	0,897	$4{,}53.10^{-3}$
HLA-DRA	6p21.32	H6	0,944	$3{,}16.10^{-2}$
BACH2	6q15	H6	1,077	$1{,}24.10^{-2}$
ABCA1	9q31.1	H24	0,711	$5{,}87.10^{-3}$
IKBKAP	9q31.3	H24	1,127	$5{,}70.10^{-3}$
LRRC37A5P	9q31.3	H6-H0	1,098	$1{,}38.10^{-2}$
COL27A1	9q32	H6	1,077	$2{,}05.10^{-2}$
KIF12	9q32	H6-H0	1,104	$3{,}54.10^{-3}$
DEC1	9q33.1	H24	0,949	$4{,}9.10^{-2}$
CNTRL	9q33.2	H6	0,784	$2{,}38.10^{-3}$
OR1K1	9q33.2	H6	1,137	$9{,}39.10^{-4}$
PTGES	9q34.11	H24	0,914	$3{,}63.10^{-2}$
FIBCD1	9q34.12	H6	1,083	$3{,}10.10^{-2}$
ARRDC1	9q34.3	H24 | Tous temps	1,148 | 1,273	$5{,}04.10^{-3}$ | $2{,}39.10^{-2}$
CYSRT1	9q34.3	H6-H0	1,102	$2{,}95.10^{-3}$
ENTPD2	9q34.3	H6-H0	1,112	$1{,}42.10^{-2}$
FCN2	9q34.3	H6-H0 | H6	1,175 | 1,110	$1{,}10.10^{-3}$ | $1{,}15.10^{-2}$
LCN12	9q34.3	H6-H0 | H6	1,108 | 1,112	$6{,}49.10^{-3}$ | $8{,}28.10^{-4}$
LCN9	9q34.3	H0 | H6-H0	0,936 | 1,108	$2{,}69.10^{-2}$ | $4{,}48.10^{-3}$
MAMDC4	9q34.3	H6	1,070	$4{,}73.10^{-2}$
BRCA2	13q13.1	H0 | H6 | Tous temps	0,836 | 0,798 | 0,628	$2{,}69.10^{-2}$ | $5{,}33.10^{-3}$ | $8{,}41.10^{-3}$
PROSER1	13q13.3	H6	0,891	$1{,}75.10^{-3}$
DNAJC28	21q22.11	H6-H0	0,805	$2{,}36.10^{-2}$
KRTAP21-2	21q22.11	H24-H0	0,892	$1{,}01.10^{-3}$
MRAP	21q22.11	H6-H0 | H6	1,141 | 1,086	$1{,}84.10^{-4}$ | $4{,}65.10^{-3}$
OLIG1	21q22.11	H6	1,066	$1{,}21.10^{-2}$
DYRK1A	21q22.13	H6	0,863	$7{,}52.10^{-4}$
RNU6-696P	21q22.13	H24-H0	0,922	$3{,}69.10^{-3}$
SIM2	21q22.13	H6-H0	1,091	$1{,}98.10^{-2}$
PCP4	21q22.2	H0 | Tous temps	0,950 | 0,881	$3{,}63.10^{-2}$ | $1{,}74.10^{-2}$
PSMG1	21q22.2	H24	1,117	$2{,}79.10^{-2}$
RPL23AP12	22q22.2	H24	0,928	$1{,}11.10^{-2}$
AIRE	21q22.3	H6	1,103	$3{,}14.10^{-3}$

Nom du gène	Région	Analyse(s) temporelle(s)	Fold change(s)	P-value(s)
CRYAA	21q22.3	H6	1,105	$5{,}79.10^{-3}$
FTCD	21q22.3	H6-H0	1,065	$4{,}64.10^{-2}$
KRTAP10-10	21q22.3	H6-H0 | H6	1,189 | 1,112	$9{,}88.10^{-4}$ | $1{,}57.10^{-2}$
TFF1	21q22.3	H6-H0	1,127	$6{,}39.10^{-3}$
TMPRSS3	21q22.3	H6-H0	1,127	$1{,}22.10^{-2}$

Tableau 13: Gènes DE dans (A-B)∩C localisés dans des locus associés ou liés à la SpA.

Les fold changes et p-values correspondent à la liste C.

III/ Analyses des jeux de gènes enrichis

Afin d'identifier les voies de signalisation pouvant être impliqués dans la SpA et plus particulièrement dans les changements observés dans nos puces transcriptomiques, nous avions effectué une GSEA avec quSAGE. Cette méthode permet de mesurer en premier lieu une activité de chaque jeu de gènes en fonction du modèle linéaire utilisé. Elle effectue ensuite une analyse différentielle directement sur ces jeux de gènes et donne un fold change pour chacun de ceux-ci. Ainsi, nous incluions tous les gènes, y compris ceux qui pourraient avoir une forte corrélation différentielle mais qui ne seraient pas détectés par une analyse LIMMA isolée sur chaque gène. L'approche quSAGE était donc utilisée pour exactement toutes les comparaisons effectuées pour les analyses de gènes DE par LIMMA. Ensuite, nous avions effectué les croisements de listes (A-B)∩C pour toutes les analyses temporelles au niveau des jeux de gènes, et nous n'avions gardé que les jeux de gènes qui comprenaient des gènes DE par LIMMA dans les mêmes croisements de listes.

Au total sur toutes les analyses temporelles, 250 jeux de gènes étaient identifiés comme DE par la SpA dans les croisements de listes respectifs, avec 170 sur-exprimés et 80 sous-exprimés. Parmi ceux-ci, 17 se retrouvaient DE à plusieurs analyses temporelles (Tableau 15). Nous pouvons citer parmi ceux-ci des gènes liés aux synapses des neurones et la voie de biosynthèse des lipides et du cholestérol. Le nombre de jeux de gènes DE à H0, H6-H0, H6, H24-H0, H24 et tous temps étaient respectivement 13, 11, 97, 1, 133 et 14 (Tableau 14).

Analyse temporelle	Nombre de jeux de gènes DE			P-value minimale	
	Total	Sur-exprimés	Sous-exprimés	Liste A	Liste C
H0	13	0	13	$2{,}23.10^{-4}$	$6{,}30.10^{-5}$
H6-H0	11	8	3	$6{,}79.10^{-3}$	$8{,}85.10^{-3}$
H6	97	29	68	$6{,}61.10^{-4}$	$7{,}36.10^{-4}$
H24-H0	1	1	0	$3{,}15.10^{-2}$	$3{,}43.10^{-2}$
H24	133	132	1	$6{,}68.10^{-4}$	$3{,}50.10^{-5}$
Tous temps	14	7	7	$1{,}08.10^{-5}$	$7{,}81.10^{-6}$

Tableau 14: Synthèse des croisements de listes (A-B)∩C à chaque analyse quSAGE temporelle de stimulation LPS.

Identifiant MSigDB	Taille	Analyses temporelles	Fold changes	P-values	Gènes DE
chr6q	4	H6-H0 | H6	0,926 | 0,924	$4,17.10^{-2}$ | $4,01.10^{-3}$	1 | 1
NEUROPEPTIDE BINDING	19	H6-H0 | H6	1,052 | 1,036	$1,77.10^{-2}$ | $2,22.10^{-2}$	4 | 2
REGULATION OF HORMONE SECRETION	13	H6-H0 | H6	1,078 | 1,054	$8,85.10^{-3}$ | $1,13.10^{-2}$	3 | 1
GENTILE UV RESPONSE CLUSTER D8	36	H0 | H6	0,956 | 0,967	$2,16.10^{-3}$ | $2,31.10^{-2}$	3 | 1
GNF2 SPRR1B	23	H6-H0 | H6	1,053 | 1,038	$3,18.10^{-2}$ | $3,54.10^{-2}$	2 | 2
HORMONE SECRETION	16	H6-H0 | H6	1,073 | 1,046	$1,61.10^{-2}$ | $3,49.10^{-2}$	4 | 2
NEUROPEPTIDE RECEPTOR ACTIVITY	18	H6-H0 | H6	1,053 | 1,036	$1,67.10^{-2}$ | $2,28.10^{-2}$	4 | 2
NEUROTRANSMITTER BINDING	49	H6-H0 | H6	1,046 | 1,035	$3,10.10^{-2}$ | $2,36.10^{-2}$	6 | 3
NEUROTRANSMITTER RECEPTOR ACTIVITY	46	H6-H0 | H6	1,053 | 1,036	$1,67.10^{-2}$ | $2,28.10^{-2}$	4 | 2
GUO TARGETS OF IRS1 AND IRS2	91	H0 | H6	0,976 | 0,973	$2,70.10^{-2}$ | $1,50.10^{-2}$	2 | 3
REACTOME CHOLESTEROL BIOSYNTHESIS	20	H0 | H6	0,874 | 0,836	$3,68.10^{-2}$ | $5,65.10^{-3}$	2 | 2
GENTILE UV RESPONSE CLUSTER D9	24	H0 | Tous temps	0,976 | 0,942	$4,22.10^{-2}$ | $6,60.10^{-3}$	1 | 1
KREPPEL CD99 TARGETS DN	6	H0 | Tous temps	0,863 | 0,859	$6,30.10^{-5}$ | $2,30.10^{-2}$	1 | 1
SCHMIDT POR TARGETS IN LIMB BUD UP	24	H0 | Tous temps	0,855 | 0,823	$3,06.10^{-3}$ | $4,22.10^{-2}$	5 | 3
MODULE 509	14	H6 | Tous temps	0,922 | 0,894	$7,36.10^{-4}$ | $1,07.10^{-2}$	1 | 1
KEGG STEROID BIOSYNTHESIS	15	H0 | H6 | Tous temps	0,882 | 0,844 | 0,819	$1,72.10^{-2}$ | $1,28.10^{-3}$ | $3,72.10^{-2}$	2 | 1 | 1
WENG POR TARGETS GLOBAL UP	19	H0 | H6 | Tous temps	0,917 | 0,929 | 0,867	$4,02.10^{-3}$ | $1,48.10^{-2}$ | $8,89.10^{-3}$	2 | 1 | 1

Tableau 15: Jeux de gènes dans (A-B)∩C DE dans plusieurs analyses temporelles.

La colonne « Gènes DE » correspond au nombre de gènes DE qu'on retrouve dans le même croisement de listes par LIMMA. Les fold changes et p-values correspondent à la liste C.

Cependant, les fonctions ressorties par quSAGE étaient très diverses et souvent très vagues, sans mise en évidence de relations pertinentes avec la SpA ou la nature des cellules. Nous détections des jeux de gènes liés au métabolisme, à la régulation des gènes, au système immunitaire ou à la mort cellulaire programmée. Cela ne dépendait pas de la méthode quSAGE, puisque nous

obtenions des résultats similaires avec d'autres outils ne se basant que sur des listes de gènes DE (Enrichr, g:profiler) ou sur les données brutes d'expression (CAMERA). De plus, beaucoup de jeux de gènes DE ne comptaient aucun gène DE identifiés par LIMMA et après permutation et croisement de listes effectués. Parmi ceux qui en comptaient, seulement 4,84 % en moyenne étaient des gènes DE parmi tous les gènes qui constituaient les jeux de gènes. Les tailles des jeux de gènes pouvaient varier entre 4 et 1834 gènes. Par ailleurs, nous avions voulu observer les recoupements entre les 250 jeux de gènes DE dans toutes nos listes. Pour cela, nous avions regardé tous les gènes (DE ou non) qui étaient en commun entre chaque paire de jeux de gènes. Puis nous avions divisé ce nombre de gènes communs par la taille du jeu de gènes le plus petit, afin d'obtenir le ratio de gènes partagés. Cela permettait d'obtenir un réseau de jeux de gènes avec les nœuds représentant chaque jeu de gènes et les arcs représentant le ratio calculé. Nous avions limité ce dernier à devant être supérieur à 15 %. Nous pouvions observer que très peu de jeux de gènes étaient spécifiques dans leur contenu, et la grande majorité se recoupaient entre eux (Figure 21).

L'utilisation de GSEA se trouvait donc assez décevante, tant par la non-spécificité des voies reconnues que par le manque de recouvrement avec les gènes identifiés par LIMMA. Nous étions donc toujours au même point, c'est-à-dire que nous ne savions toujours pas quels gènes prioriser pour les études fonctionnelles. C'est pourquoi nous avions utilisé la théorie des graphes, qui se trouvait être un outil très intéressant dans notre cas.

Figure 21: Réseau représentant la redondance des 250 jeux de gènes DE identifiés par quSAGE dans (A-B)∩C.

La transparence des arcs est proportionnelle au ratio du nombre de gènes partagés par rapport à la taille du jeu de gènes le plus petit. Seuls les ratios de plus de 15 % sont représentés.

IV/ Analyses par réseau

Bien que nous avions des résultats intéressants par l'analyse des gènes DE, l'approche GSEA s'était trouvée décevante sur l'exploration des voies de signalisation à l'origine des changements d'expression observés, et sur la priorisation des gènes pour les études fonctionnelles. C'est pourquoi nous avions poursuivi notre analyse plus en profondeur pour observer quelles étaient les interactions pouvant subsister entre les gènes DE grâce à la théorie des graphes, et pour accorder une centralité à certains gènes en fonction de la topologie du réseau observé. L'objectif principal de notre inférence de réseaux était de comparer les expressions des gènes, car nous avions

suffisamment de données pour observer des corrélations très significatives. De plus, nous voulions découvrir de nouvelles interactions qui pouvaient apparaître spécifiquement chez les patients ou chez les témoins. C'est pourquoi nos réseaux sémantiques, construits avec STRING, avaient pour utilité principale de déterminer si les interactions observées dans nos réseaux gaussiens étaient connues dans la littérature. Nos réseaux gaussiens inférés n'utilisaient que les gènes avec des symboles HGNC existants, pour une interprétation plus facile.

IV.1) Comparaison des résultats SIMoNe avec WGCNA

Nous avions inféré tous nos réseaux gaussiens avec l'approche SIMoNe, ce qui avait pour avantage principal d'utiliser une méthode permettant de prendre en compte les corrélations partielles, et donc de retirer le bruit de fond dû à des interactions indirectes de façon non supervisée. Cependant, cette approche n'était pas forcément très intuitive et n'était pas beaucoup utilisée dans la littérature sur des données biologiques réelles. C'est pourquoi, dans un premier temps, nous avions voulu comparer les résultats inférés par SIMoNe avec une approche WGCNA, qui était plus triviale et plus ancienne, ce qui faisait qu'elle était déjà énormément citée pour les analyses de réseaux de gènes dans la littérature. Afin d'avoir une comparaison globale sur nos résultats, nous avions inféré nos réseaux avec les deux approches sur les croisements de listes (A-B)∩C à toutes les analyses temporelles : H0, H6-H0, H6, H24-H0, H24 et tous temps. De plus, pour chacune de ces analyses, nous avions inféré les réseaux spécifiquement chez les patients, ou les témoins HLA-B27 positifs et négatifs réunis. Afin d'avoir des modèles avec des scores BIC et AIC corrects sous SIMoNe, nous avions pris à chaque fois le nombre gènes les plus DE dans la liste C de telle sorte à ce qu'il soit égal au nombre d'échantillons moins deux. Aucun filtrage sur les rhos de Spearman n'était effectué après les inférences par SIMoNe, afin de comparer les deux méthodes sans a priori.

A première vue nous avions observé une connectivité bien plus importante dans les réseaux inférés par SIMoNe que par WGCNA. En regardant de plus près, nous avions remarqué que parmi les arcs inférés par WGCNA, une bonne partie était aussi détectée par SIMoNe (61,62 %, Figure 22). Dans la grande majorité des cas, les arcs spécifiques de WGCNA étaient en vérité des corrélations fortes dues à des interactions indirectes avec un même gène présent dans le réseau. L'autre propriété intéressante de nos réseaux était que nous bénéficions de corrélations très

significatives, avec des rhos de Spearman pouvant dépasser 0,80 et des p-values de l'ordre de 10^{-12}. Par conséquent, nous avions décidé d'utiliser une approche mixte entre SIMoNe et WGCNA : après avoir inféré nos réseaux par SIMoNe, nous effectuions un filtrage a posteriori sur la valeur absolue de rho dépassant 0,4 pour chaque arc, ce qui correspondait à une p-value inférieure à 5 % (Figure 23). En effet, après ce filtrage, nous obtenions des réseaux beaucoup plus interprétables, car SIMoNe pouvait inférer des arcs avec une faible corrélation entre les expressions des gènes.

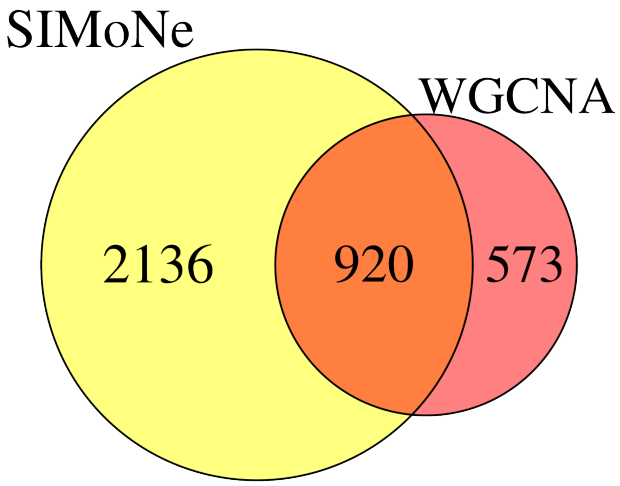


Figure 22: Diagramme de venn du nombre d'arcs total inféré soit par SIMoNe, soit par WGCNA.

Les nombres d'arcs représentent la somme des connectivités des réseaux inférés sur tous les croisements de listes (A-B)∩C à toutes les analyses temporelles, sur la totalité des échantillons, ceux extraits des patients, ou ceux extraits des témoins.

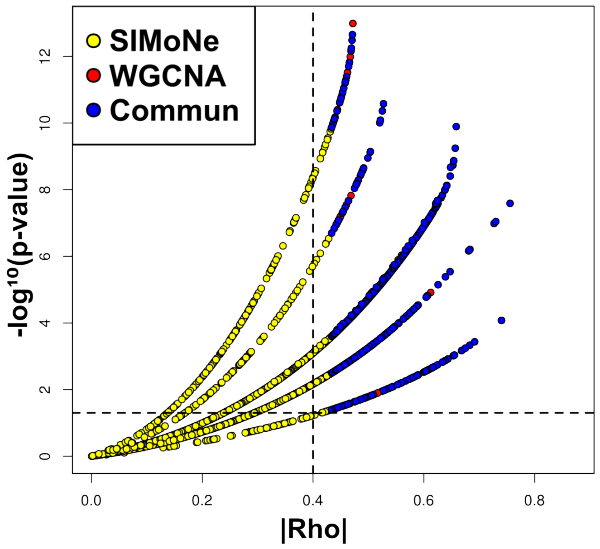


Figure 23: P-values en fonction des rhos de Spearman des corrélations entre les expressions des gènes pour les arcs inférrés par SIMoNe ou WGCNA.

Le pointillé horizontal représente un |rho| égal à 0,4, et le vertical représente une p-value égale à 5 %. Les corrélations avec des p-values exactement égales à zéro ne sont pas montrées, la transformation en logarithme ne donnant pas des valeurs finies.

IV.2) Implication de la voie de biosynthèse du cholestérol

En premier lieu, nous avions observé comment se connectaient les 49 gènes DE avec symboles HGNC détectés par LIMMA sans stimulation LPS dans (A-B)∩C. Pour cela, nous avions inféré un réseau sous SIMoNe avec tous les échantillons, pour détecter les interactions qui pouvaient apparaître globalement. De plus, les 68 échantillons à H0 permettaient d'obtenir des modèles avec des critères BIC et AIC convenables. Après filtrage sur les rhos de Spearman, nous avions inféré un réseau de gènes gaussien de 33 nœuds et 39 arcs (Figure 24). Nous avions détecté 30 corrélations positives pour 9 corrélations négatives. Parmi les gènes connectés, nous retrouvions KIAA0907 et TBCK qui étaient également DE dans l'étude 1. Nous avions également PCP4, un gène sous-exprimé chez les patients dans nos données et localisé dans la région associée 21q22.2. Mais de façon plus intéressante et inattendue, nous avions également 5 gènes sous-exprimés chez les patients et liés à la voie de biosynthèse du cholestérol : SREBF2, SQLE, LDLR, INSIG1 et MSMO1. De plus, SQLE et MSMO1 font partie de la même voie de signalisation dans la base de

données KEGG : Steroid biosynthesis. C'était également dans ce sous-graphe que nous avions les corrélations les plus fortes, dont celle entre SQLE et MSMO1 donnant un rho de Spearman de 0,87 et une p-value égale à zéro (Figure 25). SQLE catalyse la première étape d'oxygénation dans la biosynthèse du cholestérol. SREBF2 est un activateur transcriptionnel requis pour l'homéostasie lipidique. MSMO1 est impliqué dans la biosynthèse du cholestérol à la membrane du RE. LDLR est un récepteur permettant l'endocytose des vésicules LDL. Enfin, INSIG1 est impliqué dans le rétrocontrôle de la synthèse du cholestérol et retient le complexe SCAP-SREBF2 dans le RE. A ces gènes-ci nous pouvions ajouter PLIN2, sur-exprimé chez les patients, qui est impliqué dans le stockage des gouttelettes de lipide intracellulaires, et dont l'absence chez la souris protège contre l'inflammation des tissus adipeux.[223] Sa corrélation négative avec SREBF2 était la deuxième plus forte dans le réseau, avec un rho de 0,57 et une p-value de $7{,}88.10^{-7}$ (Figure 24 et Tableau 16).

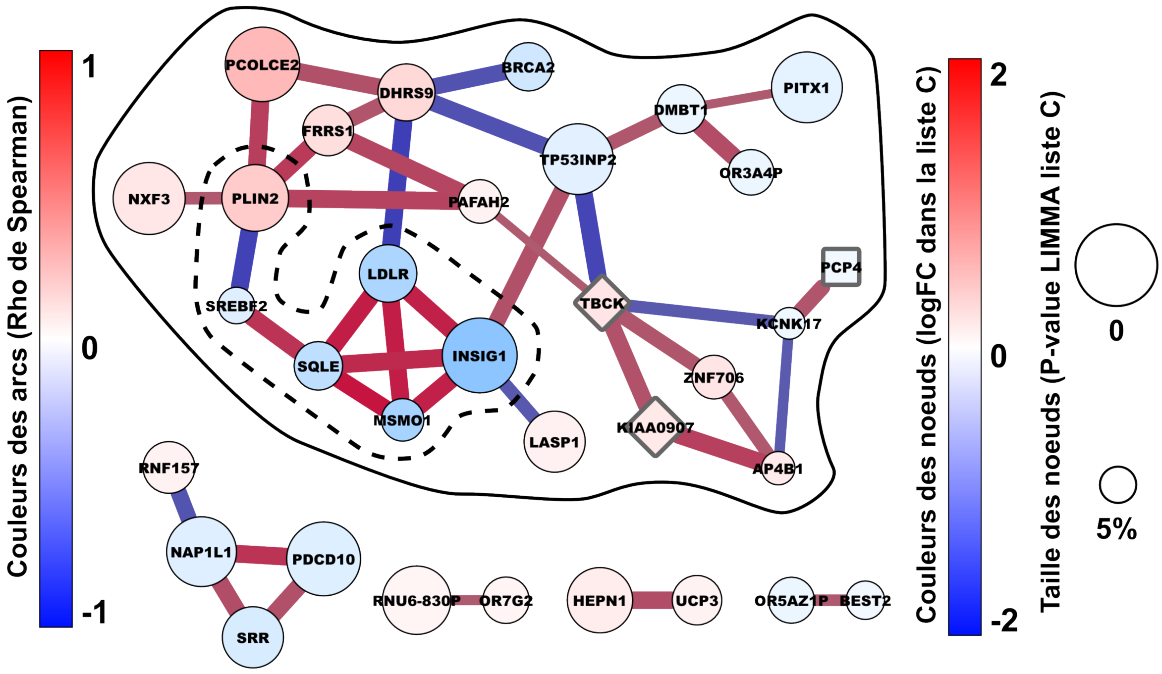


Figure 24: Réseau inféré avec SIMoNe sur les 49 gènes DE à H0 dans (A-B)∩C, avec tous les échantillons de H0.

Le contour en pointillés représente les gènes impliqués dans la voie de biosynthèse du cholestérol. Celui en trait solide constitue le cluster de 23 gènes utilisé pour comparer la connectivité entre patients et témoins. Les deux nœuds en forme de losanges représentent les gènes répliqués par rapport à l'étude 1. Le gène représenté par un carré (PCP4) est un gène situé dans une région associée à la SpA.

Nom du gène	Liste A		Liste C	
	Fold change	P-value LIMMA	Fold change	P-value LIMMA
INSIG1	0,737	$4{,}39.10^{-3}$	0,650	$7{,}68.10^{-5}$
LDLR	0,704	$6{,}88.10^{-3}$	0,734	$1{,}76.10^{-2}$
SQLE	0,776	$2{,}03.10^{-2}$	0,785	$2{,}67.10^{-2}$
MSMO1	0,700	$2{,}19.10^{-2}$	0,717	$3{,}28.10^{-2}$
SREBF2	0,855	$5{,}56.10^{-3}$	0,890	$3{,}88.10^{-2}$
PLIN2	1,354	$1{,}26.10^{-2}$	1,380	$8{,}34.10^{-3}$

Tableau 16: Synthèse des résultats LIMMA des gènes impliqués dans la voie de biosynthèse du cholestérol détectés à H0.

Les gènes bleus sont sous-exprimés chez les patients, alors que le rouge est sur-exprimé.

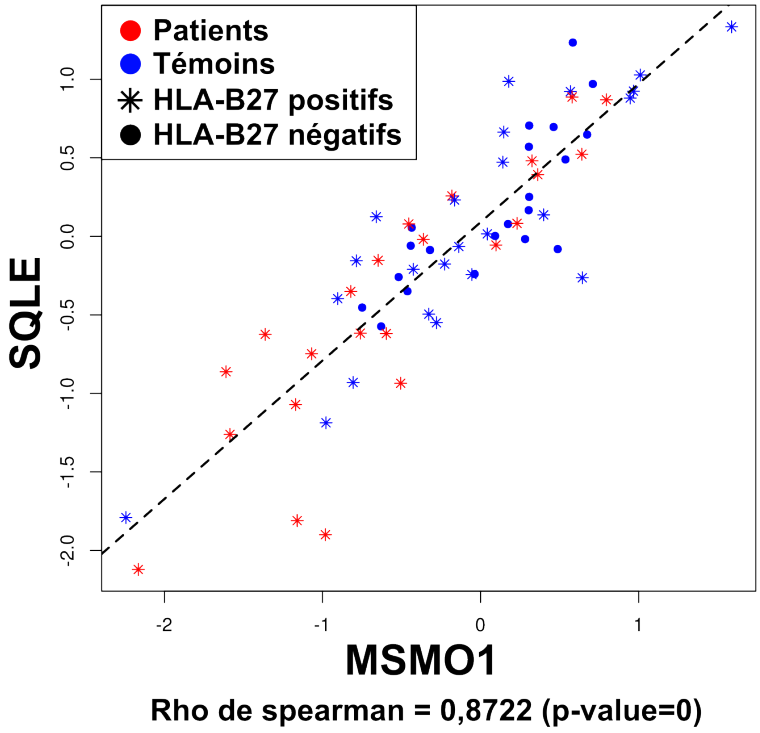


Figure 25: Données d'expressions centralisées et mises à l'échelle de SQLE en fonction de MSMO1.

Le trait en pointillés représente la régression linéaire des deux variables.

Nous avions donc inféré un réseau de gènes fortement corrélés et qui avaient un sens au niveau des voies de signalisation dans leurs interactions. Cependant, nous voulions vérifier si ces interactions étaient réellement mises en évidence dans la littérature, et nous avions pour cela

comparé nos résultats avec celles du réseau STRING construit. Comme attendu compte tenu des rôles de chaque gène, ce réseau est retrouvé dans ce réseau sémantique (Figure 26). Il était constitué de 4 interactions issues de bases de données internes, 4 de données expérimentales, et 2 de données de co-expression. Les scores de confiance étaient compris entre 0,621 et 0,937, et un interacteur était ajouté entre PLIN2 et SREBF2, ainsi que LDLR et SREBF2 : CREBBP, un activateur transcriptionnel ubiquitaire. Bien que les connections n'étaient pas exactement les mêmes entre SIMoNe et STRING, nous retrouvions et confirmions un réseau très fortement corrélé et qui existait aussi dans la littérature. Cela nous confortait donc dans les résultats des réseaux inférés par SIMoNe et dans nos données d'expression. Nous estimions que les quelques différences de connectivités étaient dues aux très fortes corrélations entre les expressions de chaque gène, rendant difficile la priorisation non supervisée des arcs dans notre réseau.

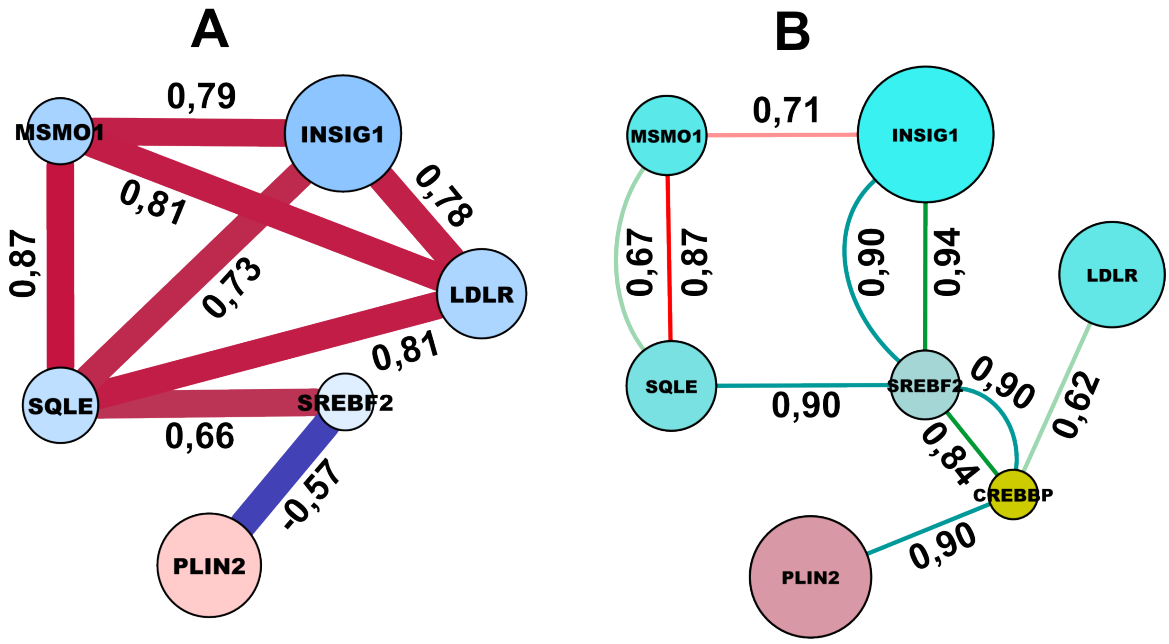


Figure 26: Réseau de gènes impliqués dans la voie de biosynthèse du cholestérol identifiés à H0 inféré par SIMoNe (A) et construit par STRING (B).

Les numéros associés à chaque arc du réseau A correspondent aux rhos de Spearman après test de corrélation entre les expressions des gènes correspondants. Ceux du réseau B correspondent aux scores de confiance donnés par STRING. Toujours dans ce réseau, les arcs rouges correspondent aux co-expressions, les verts clairs aux données expérimentales et les turquoises aux bases de données internes de STRING. CREBBP est un gène ajouté par STRING pour déterminer les interactions indirectes. Les tailles et les couleurs des nœuds dans les deux réseaux dépendent respectivement du fold change et de la p-value donnée par LIMMA pour la liste C.

IV.3) Centralité de MSMO1 spécifique des patients

L'autre particularité du réseau inféré avec les 49 gènes DE à H0 dans notre croisement de listes était la formation d'un cluster constituant 23 gènes, ce qui le rendait évaluable sous SIMoNe seulement avec les échantillons des patients ou des témoins (Figure 24). De plus, ce cluster comprenait les gènes impliqués dans la voie de biosynthèse du cholestérol, ainsi que PCOLCE2. La question que nous nous étions alors posés était : comment ce cluster de gènes se connecte-t-il chez les patients ou les témoins ? C'est ainsi que nous avions inféré sous SIMoNe des réseaux gaussiens spécifiques de ces deux groupes avec ce cluster, puis nous les avions superposé et calculé les différences de connectivités entre patients et témoins pour chaque nœud. Cela nous permettait d'avoir une approche complémentaire à l'expression différentielle, qui était de détecter les gènes différentiellement connectés.

De manière générale, nous avions pu remarquer que ce réseau était plus connecté chez les témoins que chez les patients, avec respectivement 36 et 19 arcs. Nous pouvions d'ailleurs identifier principalement deux clusters spécifiques des témoins, constitués de 8 et 5 gènes. Par ailleurs, les gènes impliqués dans la voie de biosynthèse des lipides étaient connectés à la fois chez les patients et les témoins, ce qui confirmait un lien très stable entre ces gènes. Le plus intéressant était la connectivité particulière de MSMO1 entre les deux groupes. En effet, ce gène était connecté directement avec trois autres gènes supplémentaires chez les patients par rapport aux témoins : INSIG1, OR3A4P et AP4B1. De plus, il s'agissait du gène avec la plus grande différence de connectivité spécifique des patients, qui était égale à environ 2,23 (Figure 27). Cela confèrait donc une centralité particulière à MSMO1 à l'état basal des MD-DCs issus des patients, c'est-à-dire sans stimulation LPS. Nous avions donc considéré que ce gène était représentatif de l'état d'expression des gènes connectés à H0, par sa topologie particulière dans le réseau et par le fait qu'il soit exactement en amont de la biosynthèse du cholestérol.

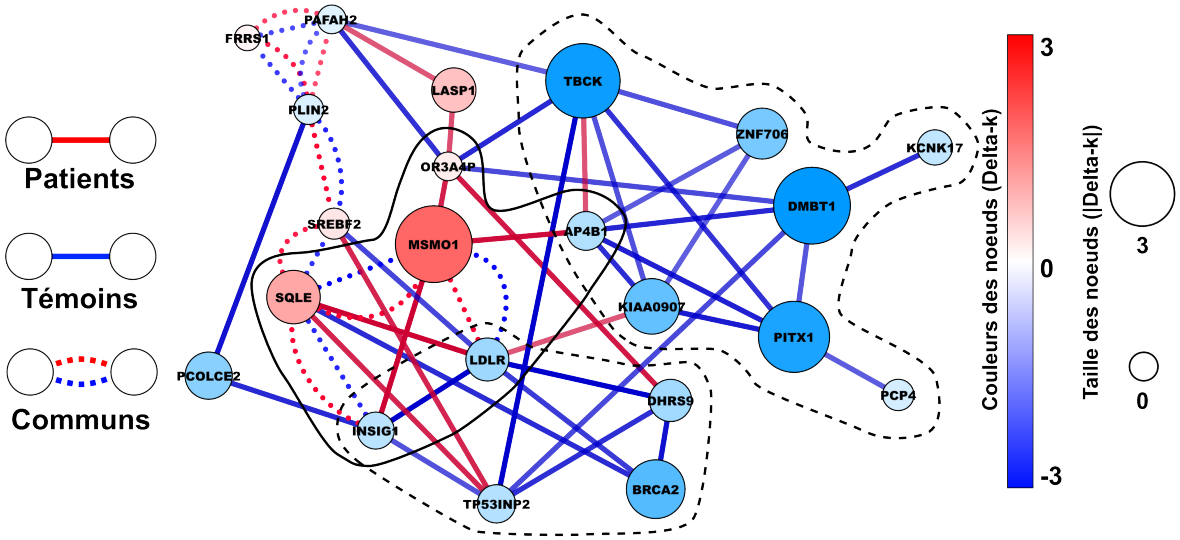


Figure 27: Superposition du cluster de 23 gènes DE à H0 inférés par SIMoNe chez les patients et témoins.

Les arcs rouges représentent les interactions inférées chez les patients, et les bleus celles chez les témoins. Les contours en pointillés représentent les deux clusters spécifiquement connectés chez les témoins. Celui en trait plein regroupe MSMO1 et ses 5 voisins directs. Delta-k représente la différence de connectivité entre les patients et les témoins.

IV.4) Centralité de la réponse de APOA4 à MSMO1

Bénéficiant des données d'expressions des MD-DCs après 6h et 24h de stimulation LPS, nous voulions savoir quel était l'impact de l'expression différentielle de MSMO1 sur la réaction des cellules au LPS. Plus concrètement dans nos données, cela revenait à observer comment ce gène se connectait avec les autres gènes DE à H6-H0, H6, H24-H0, et H24. Nous voulions savoir dans un premier temps quels étaient les gènes les plus corrélés à ces différents temps par rapport à MSMO1, chez les patients et les témoins. Pour cela, nous avions effectué notre inférence WGCNA sur les données de MSMO1 à H0 et celles des autres gènes aux temps correspondant. Pour les gènes DE à H6-H0 et H24-H0, nous avions pris les données d'expressions respectivement à H6 et H24. Nous obtenions ainsi en tout 8 réseaux, répartis en deux groupes chez les patients et témoins, regroupant chacun quatre réseaux correspondant aux analyses temporelles citées ci-dessus. Enfin, nous sélectionnions les voisins directs de MSMO1. Afin d'avoir un nombre optimal de gènes pouvant être inférés dans les deux groupes d'individus, soit nous ne prenions que les gènes les plus DE parmi les voisins directs, soit nous ajoutions quelques voisins secondaires.

Nous utilisions WGCNA au lieu de SIMoNe car nous voulions inclure tous les gènes DE, et nous étions moins gênés par le bruit de fond des interactions indirectes, puisque nous ne nous intéressions qu'aux voisins les plus proches de MSMO1.

Nous avions pu obtenir des réseaux d'une taille moyenne de 20 gènes chez les patients, contre 17,25 chez les témoins (Figure 28). Cette baisse de la moyenne chez les témoins était principalement due aux gènes DE à H24, où nous étions contraints de diminuer considérablement le nombre de gènes pour pouvoir obtenir des modèles SIMoNe corrects à la fois chez les patients et les témoins par la suite. Quoi qu'il en soit, nous avions pu remarquer que nous avions beaucoup plus de facilités à trouver des voisins directs de MSMO1 à H6-H0 et H6 qu'à H24-H0 et H24. La première explication à celle-ci pourrait être que nous avions moins de gènes DE après 24h de stimulation LPS qu'après 6h. Cette différence se percevait surtout entre H24-H0 et H6-H0, avec respectivement 369 et 98 gènes DE. Cependant, bien qu'il y avait 466 gènes DE à H6 et que cela représentait beaucoup plus que les 200 gènes DE à H24, ce dernier nombre n'était pas très inférieur par rapport à H6-H0. De plus, si les réseaux étaient invariants d'échelle comme c'est le cas pour la plupart des réseaux de gènes, la connectivité suivrait une loi de puissance indépendante de la taille totale du réseau. L'autre explication, qui semblait plus probable et en rapport avec les expériences, serait qu'il est plus difficile de détecter un effet de l'expression de MSMO1 après 24h de stimulation que 6h de stimulation. En effet, il est connu que les mécanismes de régulations des expressions des gènes sont très dynamiques, et il n'est pas surprenant de ne plus pouvoir observer d'effet direct de MSMO1 sur les autres gènes sur une période aussi prolongée. Les réseaux inférés à H6-H0 par WGCNA chez les patients et témoins avaient détecté au total respectivement 25 et 18 voisins directs. La sélection des voisins pour l'inférence par SIMoNe s'était donc faite sur la quasi-totalité de ces gènes. Il en était de même pour H6 chez les témoins, avec 21 gènes sélectionnés sur les 27 voisins directs de MSMO1 détectés. En revanche, pour les patients, nous n'avions sélectionné que 21 voisins directs sur 45, soit environ 46,67 %, ce qui n'était pas très représentatif de la totalité des effets possibles de MSMO1 sur les gènes DE à cette analyse temporelle.

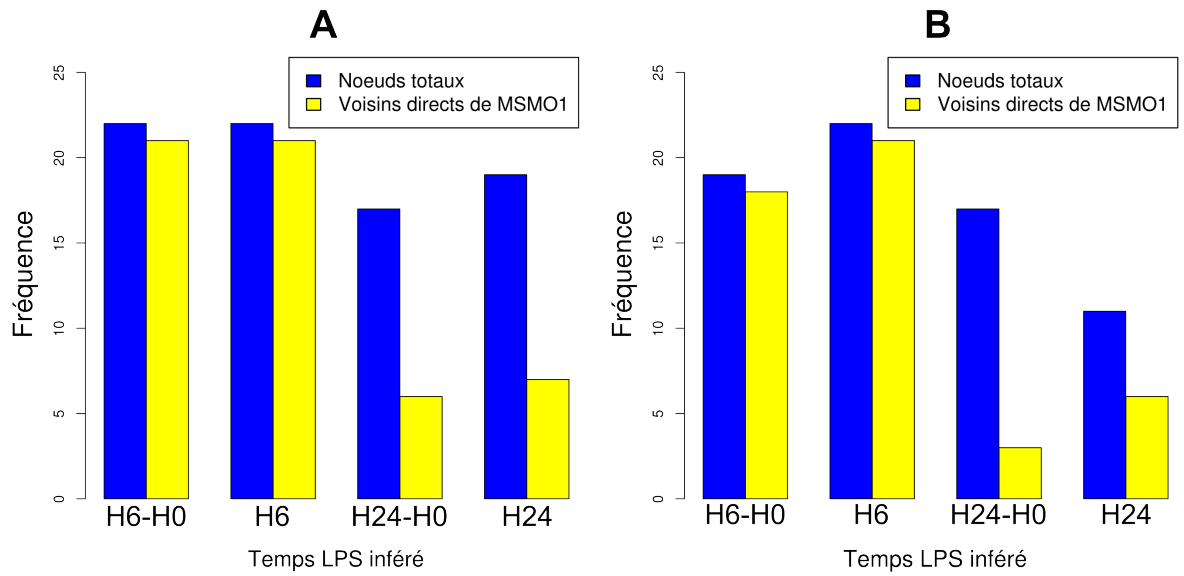


Figure 28: Résumé des voisins sélectionnés de MSMO1 aux temps LPS suivants inférés par WGCNA chez les patients (A) et chez les témoins (B).

Nous pouvons remarquer que nous avions plus de difficultés à trouver des voisins directs de MSMO1 après 24h de stimulation LPS qu'après 6h.

Par la suite, nous avions voulu déterminer si les clusters définis par WGCNA entre MSMO1 et ses voisins aux différentes analyses temporelles étaient bien spécifiques des groupes d'individus chez qui nous avions utilisé WGCNA. Pour cela, nous avions utilisé la même approche que pour le cluster de 23 gènes DE à H0, mais pour chacun des 8 clusters définis plus haut : nous avions superposé les réseaux inférés par SIMoNe sur ces clusters chez les patients et témoins, et déterminé les gènes différentiellement connectés. Dans un premier temps, en comparant les connectivités globales de chaque réseau par leurs nombres d'arcs totaux respectifs, nous avions remarqué que globalement et comme attendu, tous les clusters étaient plus connectés dans les groupes d'individus qui leur correspondaient (Figure 29). Cependant, nous avions observé que les réseaux étaient très peu spécifiquement connectés à H24-H0 chez les patients et à H24 chez les témoins. Inversement, le réseau le plus spécifique que nous obtenions était à H6-H0 chez les patients.

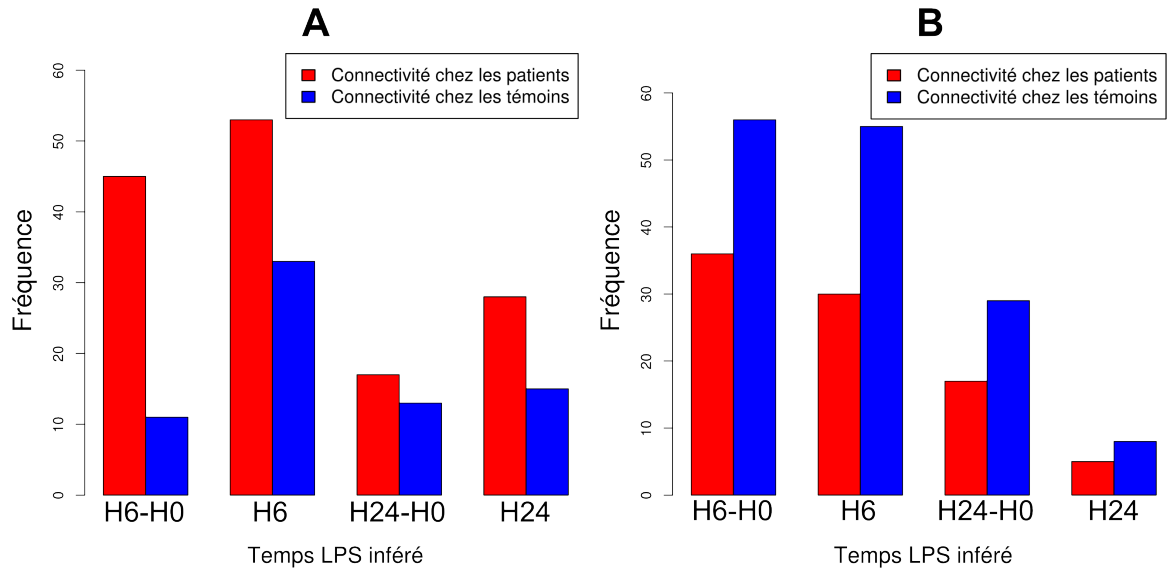


Figure 29: Comparaison de chaque cluster sélectionné dans la Figure 28 et inféré par SIMoNe chez les patients et témoins.

Nous avions observé de plus près ce réseau en superposant ceux inférés par SIMoNe chez les patients et les témoins. Comme attendu d'après les résultats décrits ci-dessus, le réseau inféré était très spécifiquement connecté chez les patients, avec des différences de connectivité positives entre patients et témoins pour exclusivement tous les gènes (Figure 30). Parmi les gènes connectés, nous retrouvions DNAJC28 et ENTPD2, qui sont localisés dans deux régions liées à la SpA, respectivement 21q22.11 et 9q34.3. Cependant, le résultat le plus intéressant se trouvait être le gène le plus différentiellement connecté, APOA4. Il possèdait 7 voisins spécifiques directs, qui étaient ADORA2A-AS1, MAGEB3, MYADML, SLC5A11, LINGO1 et FOXI2, et une différence de connectivité de 5,72. De plus, le gène APOA4 est un composant majeur des vésicules HDL et des chylomicrons, et est donc impliqué de nouveau dans les voies de signalisations associées au cholestérol. Ce gène était sur-exprimé chez les patients, avec un fold change de 1,08 et une p-value de $8{,}63.10^{-3}$ dans la liste C. Même s'il n'était pas connecté directement à MSMO1 dans notre réseau inféré par SIMoNe, les deux possèdaient une corrélation négative significative entre H0 et H6, avec un rho de -0,51 et une p-value de 1,49 % chez les 23 patients.

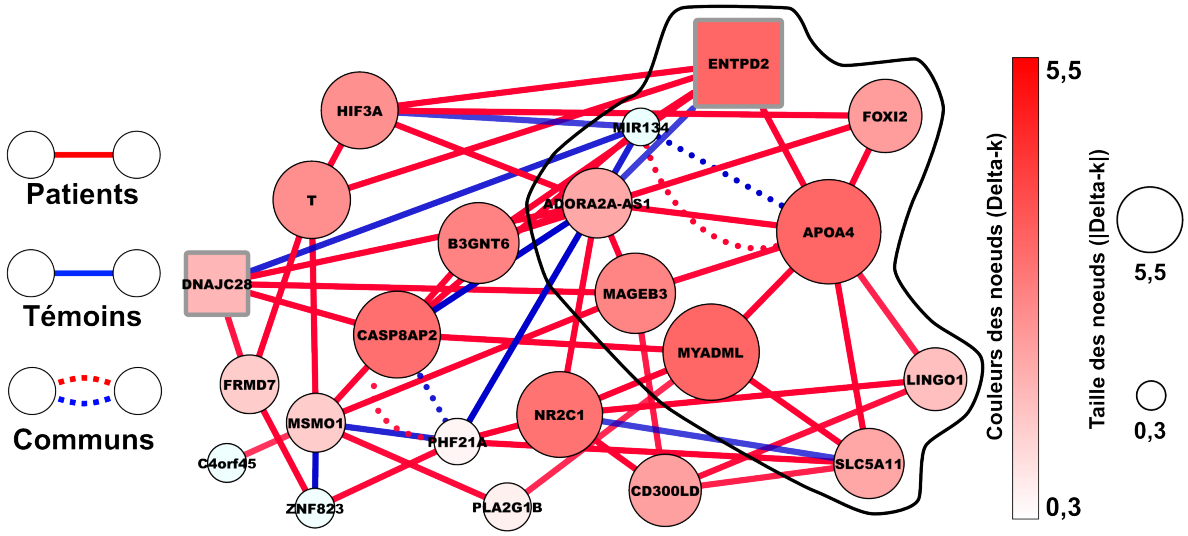


Figure 30: Superposition des réseaux inférés sous SIMoNe chez les patients et témoins, à partir des voisins de MSMO1 inférés par WGCNA chez les patients.

Le contour représente les 8 voisins directs de MSMO1 inférés chez les patients par SIMoNe. Les gènes sous forme carré représentent ceux qui sont localisés dans des régions liées à la SpA.

Nous avons donc pu remarquer que la théorie des graphes appliquée à nos données transcriptomiques se révèle être un outil très intéressant, et qu'il nous a permis de mettre en évidence des voies de signalisation, et notamment celles incluant la voie de biosynthèse du cholestérol, qui étaient complètement noyées dans les informations fournies par l'approche GSEA et le nombre important de gènes DE identifiés. De plus, il nous a permis de donner un sens biologique à nos résultats et plus de confiance dans les statistiques issues des analyses LIMMA, notamment par les fortes corrélations observées entre les expressions de nos gènes.

V/ Méthode de construction des réseaux

Toutes les méthodes de construction des réseaux décrites dans les matériels et méthodes, en dehors de la superposition des réseaux entre plusieurs facteurs, ont été rassemblées et publiées sous forme de package R nommé stringgaussnet. Il a été déposé dans le CRAN et un article sous forme d'application note a été publié dans Bioinformatics. L'objectif principal de cet outil était de partir de résultats d'analyses de gènes DE et de construire des réseaux sémantiques avec STRING ou des réseaux gaussiens avec SIMoNe et WGCNA, le tout en seulement quelques lignes de commande. Il inclut également le calcul des voies les plus courtes pour les réseaux STRING, ainsi que beaucoup de paramètres par défaut qui faciliteront la prise en main au début. Par ailleurs, ce

package inclut des fonctions génériques permettant de construire automatiquement différents réseaux gaussiens en fonction d'un facteur attribué aux échantillons, ou bien de générer tous les types de réseaux sur plusieurs listes de gènes. De plus, la partie la plus intéressante de ce package est de pouvoir importer automatiquement tous les réseaux générés dans une session Cytoscape active, sans aucune connaissance de la structure des objets R et sans installation de langage de programmation tiers.

Nous avons créé ce package car nous avons remarqué qu'aucun outil dans cette optique n'a encore été développé, c'est-à-dire la création facile de réseaux de gènes sans connaître précisément tous les paramètres potentiellement compliqués à prendre en compte. Pourtant, beaucoup de biologistes souhaitent aller plus loin dans leurs analyses de gènes DE en intégrant la théorie des graphes, et ce manque d'outil peut les bloquer ou ajouter des difficultés non nécessaires. Le but de ce package est de guider l'utilisateur à travers la construction des réseaux de gènes ; il appartiendra ensuite au biologiste de conclure sur les gènes les plus intéressants en manipulant son réseau et en y effectuant des analyses complémentaires. Ce package a donc été développé dans un souci d'être à la fois souple et ergonomique. Seules quelques notions du langage de programmation R sont nécessaires. Ces travaux ont fait l'object d'une Application Note publiée dans la revue Bioinformatics.

Application note

stringgaussnet: from differentially expressed genes to semantic and gaussian networks generation

Emmanuel Chaplais[1], Henri-Jean Garchon[1,2]

[1]Inserm U1173 and University of Versailles Saint-Quentin, 78180 Montigny-le-Bretonneux, France

[2]Ambroise Paré Hospital, Division of Genetics, 92100, Boulogne-Billancourt, France

Keywords : Biological networks, Gene regulatory networks, Network analysis

ABSTRACT

Motivation: Knowledge-based and co-expression networks are two kinds of gene networks that can be currently implemented by sophisticated but distinct tools. We developed stringgaussnet, an R package that integrates both approaches, starting from a list of differentially expressed genes.

Availability and implementation: Freely available on the web at http://cran.r-project.org/web/packages/stringgaussnet

1 INTRODUCTION

Analysis of genes differentially expressed (DE) depending on a condition has become a standard procedure in current biology. However, identification of biologically relevant DE genes is far from being trivial. Yet efficient prioritization of DE genes is an essential step before undertaking rate-limiting wet lab experiments.(Smyth, 2004) In this regard, the network theory appears as a powerful framework. The aim is to connect genes (the nodes) by means of their interactions (the edges).(Dong and Horvath, 2007) These interactions may be based on prior knowledge, curated and stored in databases, or extracted from the experimental dataset, e.g. using coexpression information.(Xue et al., 2014; Verfaillie et al., 2015; Lin et al., 2015; Cotney et al., 2015) Sophisticated but distinct tools are available to implement either one of these approaches separately. We introduce stringgaussnet, an R package that allows to infer gene networks starting from a list of DE genes by integrating both of these approaches with ease and flexibility.

2 INITIAL OBJECTS AND EXAMPLE DATA

Stringgaussnet requires two data frames that store, one the expression measurements, the other DE gene statistics. They are combined to create an object of class DEGeneExpr. Expression data are usually normalized for efficient correlation computation. In the example data we provided transcriptomic profiles of monocyte-derived dendritic cells from 9 patients affected with ankylosing spondylitis and 10 healthy controls. Transcriptomic data were obtained using microarrays. Gene expression levels in patients and in controls were then compared with LIMMA.(Talpin et al., 2014) We limited the number of DE genes to 75.

3 STRING NETWORK CONSTRUCTION

Stringgaussnet allows to construct a protein-protein interaction network using the DE gene names (Ensembl IDs or HGNC symbols) with the use of the STRING API with specific URIs.(Franceschini et al., 2013; Szklarczyk et al., 2015) The number of additional nodes can be set by the user, and these are useful to detect indirect relationships between input genes. The default value is twice the number of initial DE genes. Different species can be curated, the default being Homo sapiens. This process generates an object of class STRINGNet, a network with multiple edges depending on sources and combined scores. One can select specific sources of interactions and filter on the scores given by STRING. Stringgaussnet can calculate a new combined score, based on the algorithm provided for STRING version 8 (http://string-stitch.blogspot.fr/2010/03/combining-scores-right-way.html).

4 SHORT PATHS FROM STRING NETWORKS

The generated network can be large and dense. As a STRINGNet object, it can be reduced by computing shortest paths between genes of a user's list. To this aim, combined scores S are converted to distances D for each node pair i with Di=max(Si)+1-Si, where max(Si) is the maximum of S over all interactions. Shortest paths between each pair of nodes are computed with the Dijkstra's algorithm. This method creates an object of class ShortPathSTRINGNet, with unique edges giving distances and intermediates as attributes. It is also possible to filter edges on D.

5 SIMONE NETWORK INFERENCE

Alternatively, stringgaussnet can help infer Gaussian networks from expression data, using the R package SIMoNe.(Chiquet et al., 2009) Default parameters are set for easy use. The number of edges that is selected is the mean of the number of edges corresponding to maximal AIC and BIC scores. By default, the algorithm computes a network under two models, with or without clustering constraints, and picks edges common to both models. Stringgaussnet performs a Spearman's test for each inferred edge. This generates an object of class SIMoNeNet, a network of unique edges including theta, Spearman's rho and p-value as attributes. One can filter on edge attributes, notably on Spearman's rho.

6 WGCNA APPROACH TO COMPARE RESULTS

SIMoNe is a powerful tool to infer unsupervised Gaussian networks. For comparison, we also propose the use of the popular WGCNA package. (Langfelder and Horvath, 2008) Stringgaussnet performs a Spearman's test between all pairs of genes. The respective rho coefficients are converted to similarity scores $\sigma=(1+\rho)/2$ that are then converted to adjacency scores $A=1/(1+e-\alpha*(\sigma-0.5))$ where α is the soft power threshold; its default is set

to 8. A method helps adjust this parameter by plotting relationships between A and rho. Then a filtering step is done using a threshold t, A being superior to t or inferior to 1-t. Dissimilarity and module computation are not implemented, because the main purpose is to compare with SIMoNe results. The network is saved in a WGCNANet class object. A function is provided to compare networks inferred by SIMoNe and WGCNA, with a Venn diagram displaying the numbers of shared and specific edges, and a series of plots showing correlation coefficients of selected interactions.

7 ADDING ANNOTATIONS TO GENES

While networks are being generated, it is possible to add gene annotations as node attributes. They are of two kinds, including genomic localization and brief gene description, using the biomaRt R package, and cellular component terms, using the GO.db package. In addition, genes are ranked depending on the localization of their protein product, from nuclear, the most relevant, to extracellular, to plasma membrane and last to cytoplasm.

8 AUTOMATIC NETWORK CREATIONS IN ONE STEP

An overlay of functions allows the user to create multiple networks in only one step, with all options configurable in the same method. One can create multiple Gaussian networks from the same DEGeneExpr object, depending on a grouping factor and for a given list of genes. The package then allows to compare networks inferred for multiple levels of the factor, and for the same DE genes list. One can create an object of class MultiDEGeneExpr, a list of DEGeneExpr objects. Then, both kinds of networks, semantic or Gaussian, can be generated for each data set and stored in a MultiNetworks object.

9 AUTOMATIC EXPORT TO CYTOSCAPE

In addition to be exported in standard file formats, generated networks can be imported automatically from R objects into Cytoscape, without requiring an intermediate language.(Cline et al., 2007) This is performed in an operating-system independent manner through the plugin cyREST. (http://apps.cytoscape.org/apps/cyrest) Stringgaussnet proposes predefined styles for displaying the exported networks; they can be easily customized in Cytoscape or directly in the package.

ACKNOWLEDGEMENTS

We thank Maxime Breban and Gilles Chiocchia for discussion and support and Christophe Ambroise for advice.

Funding: EC was funded by a PhD scholarship from the doctoral school “Du génome aux organismes”. This work was supported by an ANR grant (2010 GEMISA).

REFERENCES

Chiquet,J. et al. (2009) SIMoNe: Statistical Inference for MOdular NEtworks. Bioinforma. Oxf. Engl., 25, 417–418.

Cline,M.S. et al. (2007) Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc., 2, 2366–2382.

Cotney,J. et al. (2015) The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. Nat. Commun., 6.

Dong,J. and Horvath,S. (2007) Understanding network concepts in modules. BMC Syst. Biol., 1, 24.

Franceschini,A. et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res., 41, D808–815.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559.

Lin,Y. et al. (2015) MiRNA and TF co-regulatory network analysis for the pathology and recurrence of myocardial infarction. Sci. Rep., 5.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol., 3, Article3.

Szklarczyk,D. et al. (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res., 43, D447–452.

Talpin,A. et al. (2014) Monocyte-derived dendritic cells from HLA-B27+ axial spondyloarthritis (SpA) patients display altered functional capacity and deregulated gene expression. Arthritis Res. Ther., 16, 417.

Verfaillie,A. et al. (2015) Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. Nat. Commun., 6.

Xue,J. et al. (2014) Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. Immunity, 40, 274–288.

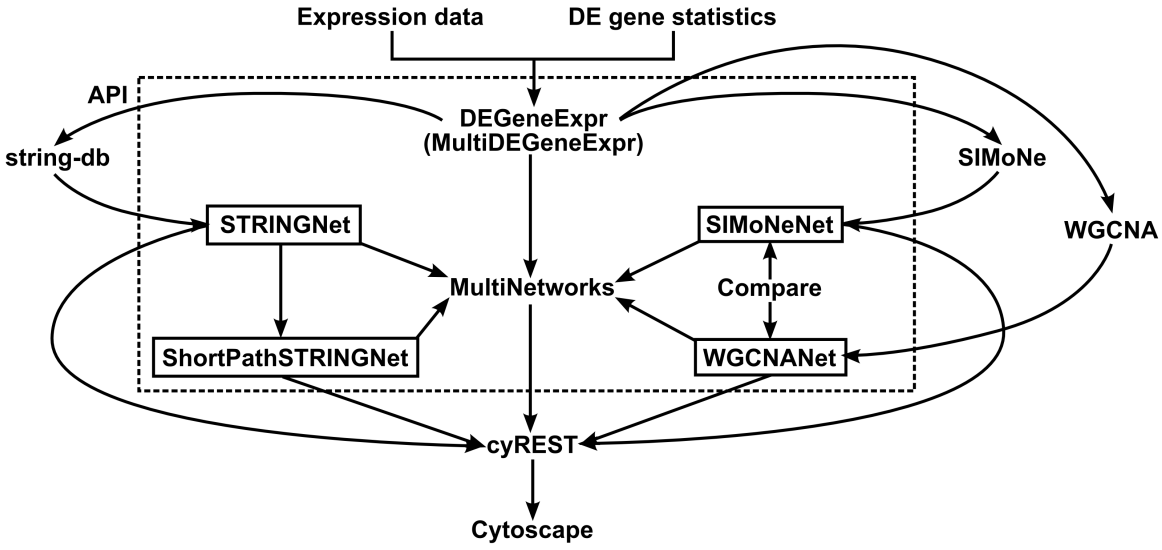


Fig. 1. Stringgaussnet operating principle. Starting from expression data and DE gene statistics, both semantic and Gaussian networks can be inferred and then exported into Cytoscape. The package environment is circumscribed by the dashed rectangle.

Stringgaussnet: user's guide

Emmanuel Chaplais
2015-06-30

INTRODUCTION

Analysis of genes differentially expressed (DE) depending on a condition has become a standard procedure in current biology. However, identification of biologically relevant DE genes is far from being trivial. Yet efficient prioritization of DE genes is an essential step before undertaking rate-limiting wet lab experiments.(Smyth 2004) In this regard, the network theory appears as a powerful framework. The aim is to connect genes (the nodes) by means of their interactions (the edges).(Dong and Horvath 2007) These interactions may be based on prior knowledge, found in databases, or extracted from the experimental dataset, e.g. using coexpression information.(Xue et al. 2014; Verfaillie et al. 2015; Lin et al. 2015; Cotney et al. 2015) Sophisticated but distinct tools are available to implement either one of these approaches separately. We introduce stringgaussnet, an R package that allows to infer gene networks starting from a list of DE genes by integrating both of these approaches with ease and flexibility.

The main objective of this tool is to be much flexible in function of your needs and to proceed automatically all necessary steps. Semantic networks are constructed by extracting all wished interaction types and possible additional nodes by requesting the STRING Application Programming Interface (API). This tool can also reduce the network by calculating shortest paths between a given number of targeted genes. Gaussian networks are inferred by SIMoNe, with a possible ad-hoc filtering of edges on Spearman's rho coefficients. This tool can also use a WGCNA-like approach, a simple correlation calculation with a soft thresholding, to compare results with SIMoNe. This package integrates R commands allowing to export automatically all created networks in Cytoscape, through the cyREST plugin.(Cline et al. 2007)

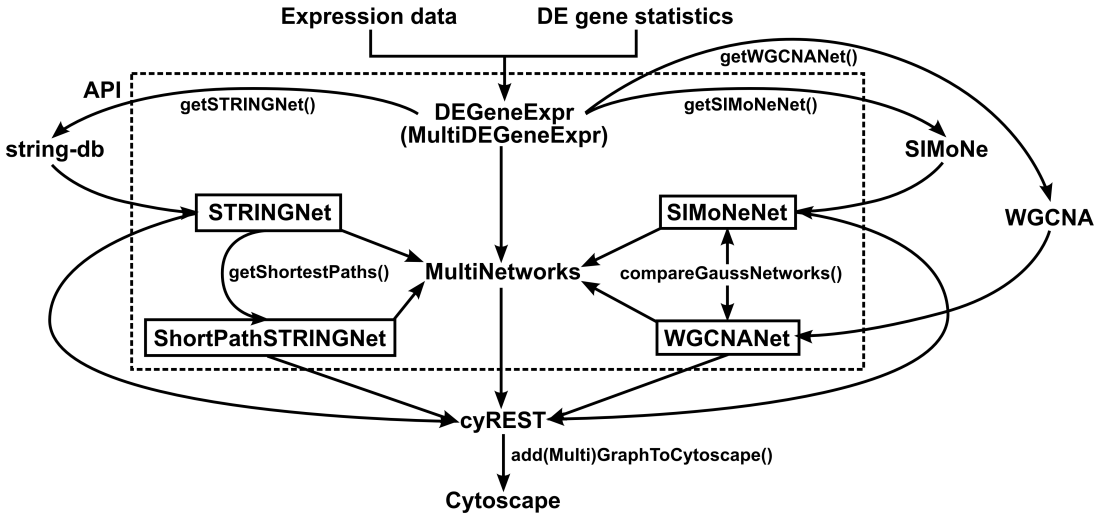


Figure 1: Stringgaussnet operating principle. Starting from expression data and DE gene statistics, it is possible to create all kinds of semantic and gaussian networks, and then to export graphs into Cytoscape. Dashed square represents original methods of stringgaussnet. Main functions are displayed by ending with brackets.

HARDWARE AND SOFTWARE REQUIREMENTS

This R package is operating system independent. However, some precautions should be taken before using stringgaussnet.
Firstly, one considers that the user (you) already knows how to use basic functions from R and to install necessary packages.
The hardware limitations depends on the network sizes you wish to compute. You mainly have to consider that whatever was your differential analysis tool, stringgaussnet will surely require a lower memory usage, because you will analyze a subset of gene expression data. A computer with at least 2 Go of RAM, a sufficient free disk space (> 1 Go) and a reasonably recent CPU (< 5 years old) is recommended.
Stringgaussnet has only been tested with R version of at least 3.2. We can not guarantee stable computation with previous versions. Otherwise, some R packages must be installed to use all functions from stringgaussnet, which are:

- AnnotationDbi
- GO.db
- VennDiagram
- simone
- biomaRt
- limma
- pspearman
- igraph
- httr
- RJSONIO
- Rcurl
- org.Hs.eg.db

Regarding packages from CRAN, you must install on his own all necessary secondary packages. Packages from bioconductor are advised to be installed with the function biocLite().
In order to be able to export networks into Cytoscape from R, this software must be independently installed on the same machine (http://www.cytoscape.org/download.php). Please use at least the version 3.2.1 and make sure to have installed java runtime environnment with version > 8 (https://www.java.com/download/). The communication between R and Cytoscape can not be performed without the plugin cyREST (version > 0.9.17, http://apps.cytoscape.org/apps/cyrest). In order to test if the plugin works fine, please turn on Cytoscape and launch the command checkCytoscapeRunning() in R. If this does not work, please try to create a new variable called port.number with a value of 1234. Then restart your computer. If this still does not work, check your Cytoscape, cyREST and java versions.

STARTING WITH DIFFERENTIAL ANALYSIS RESULTS

Stringgaussnet is not a tool of differential analysis for gene expressions. Other powerful tools exist, like limma, and it is considered that you have already identified key DE genes to analyze in a network before using this R package.
The differential analysis results constitute the basis of the package use, and two data frames are required, which are combined into an object of class DEGeneExpr. Those are expression data and DE genes statistics.
Expression data must count samples as row names and genes as column names. Values are usually normalized for efficient correlation computation.
DE genes statistics can be obtained by analyzing expression data to get genes that are affected by a given phenotype. Those results can be obtained for example by LIMMA. This variable corresponds to a data frame with genes as rows and statistics as columns. The minimum suggested columns are fold changes and p-values for the visualization in Cytoscape, but you are free to set other values as properties for node color and size. Genes in expression data must be exactly the same as those in DE genes statistics.
Those data frames must be firstly combined in an object of class DEGeneExpr, which is then a list with its own print function, giving the number of samples and genes and a preview of the both data frames.
In the example data we provided transcriptomic profiles of monocyte-derived dendritic cells (MD-DCs) from 9 patients affected with ankylosing spondylitis and 10 healthy controls. Transcriptomic data were obtained using microarrays. Gene expression levels in patients and in controls were then compared with LIMMA.(Talpin et al. 2014) We limited the number of DE genes to 75.
Let's see how it looks like in the example data inside the package:

```
library(stringgaussnet)

data(SpADataExpression) # Import example expression data
data(SpADEGenes) # Import example DE genes analysis results
data(SpASamples) # Import example sample description
# We firstly import all example data from stringgaussnet package.
# SpASamples is not compulsory for using stringgaussnet, but is useful for creating a
# factor for subsetting gaussian networks generation.

head(SpASamples,5)
##   chipnum  status LPStime subject
## 1      21 Patient      H0     I18
## 2      22 Patient      H0     I17
## 3      23 Patient      H0      I4
## 4      24 Patient      H0      I1
## 5      25 Patient      H0      I2
# We can see here what sample descriptions look like. LPStime is the LPS
stimulation
# duration for MD-DCs.
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)

print(SpAData,5)
```

```
## Object of class DEGeneExpr (package stringgaussnet)
##
## Number of samples: 57
## Number of genes: 75
##
## DataExpression preview:
##      NUDT3    P2RX1    SGMS2    WDR25    F13A1  FAM204A    LRRC4
## 21 10.25609 7.779726 7.478363 7.395941 13.53042 8.865439 6.196102
## 22 10.17532 7.713649 7.426126 7.482414 13.39109 8.743199 6.715109
## 23 10.09892 7.736853 7.514560 7.598745 13.37703 8.716703 7.489576
## 24 10.24614 7.995944 7.962326 7.658044 13.65390 8.732653 6.611568
## 25 10.02167 7.884883 6.473152 7.645395 13.67628 8.581200 6.637558
##       EIF4H   SAP130     SLU7  POU5F1B   POLR1D   TTC39C ADAMTS15
## 21 10.355450 10.17557 8.517671 6.425804 9.553580 7.429002 7.779027
## 22 10.104375 10.18127 7.944881 6.761638 9.350876 7.147367 8.855387
## 23 11.089198 10.28844 7.796919 6.634245 9.377427 7.614758 9.434300
## 24 10.188845 10.28227 8.153847 6.697428 9.390345 7.488865 7.944880
## 25  9.787701 10.14357 7.824845 6.638128 9.036362 7.173620 8.397508
##    TNFSF13B   TSPYL5    CSF3R      FAU     TFAM     FAIM2   CITED2
## 21 10.17795 8.561254 7.383586 10.299998 6.797904  6.753533 11.23097
## 22 10.84027 8.214381 7.135964 10.300351 6.786641 10.433228 11.00969
## 23 10.90312 8.316282 7.008943 10.109672 6.656924  9.121431 10.86765
## 24 10.87046 8.395607 7.803192  9.546860 6.669851 10.279958 11.39621
## 25 10.64746 8.070187 7.843946  9.631554 6.621111  9.579760 11.83044
##    SIGLEC15     MBIP    HAUS1     RFC3     TBCK   TRIM24  ANAPC15     PIGB
## 21 7.730661 8.265060 9.547622 6.717628 9.507269 9.164891 10.42642
8.663999
## 22 8.486743 8.373733 9.286334 6.572705 9.561226 9.012223 10.77170
8.803228
## 23 8.774401 8.302195 9.210819 6.721516 9.318380 9.086190 10.71807
9.102257
## 24 8.724044 8.059957 8.926340 6.859887 9.486730 9.269940 10.73594
8.996247
## 25 8.460145 7.942978 8.907234 6.503483 9.719901 8.715967 10.48000
8.208808
##     DNAJA4    RBBP9 KIAA0907      MUT   FBXO18    USP30      NDP     OLR1
## 21 8.272130 8.143958 9.352399 8.188985 8.265647 8.903849 7.663387
6.220560
## 22 8.370383 7.665083 9.118642 8.275170 8.363612 8.475835 7.667593
6.641498
## 23 8.160428 7.922728 8.891830 8.369798 8.099168 8.736526 6.961104
7.722695
## 24 8.407753 8.101788 9.011711 8.337796 8.353551 8.483161 6.080360
5.939017
## 25 8.047116 7.251024 9.234585 7.995534 7.991599 8.523176 5.715268
6.317065
##      PLP2     MNDA    IFT52    COX20   BAK1P1    CRTAP    RPS15   ALG10B
## 21 11.36747 8.173222 7.287404 7.968966 6.412227 11.28916 10.43043
```

```
7.672037
## 22 11.09409 9.675167 6.828304 8.073868 6.570384 11.00393 10.43679
6.686629
## 23 11.20175 8.463822 6.903684 8.310869 6.447920 10.97360 10.63966
7.217247
## 24 11.78391 9.009064 7.119040 8.297897 6.767743 11.39861 10.35509
7.628294
## 25 10.61192 8.824307 6.888093 8.136609 7.137114 11.16194 10.16442
6.972385
##       ENY2    PIAS2    FBXL4    HSPH1    PTPLA    COX7B    EDEM3
## 21 9.043001 9.677796 8.894753 10.96024 7.228479 7.967758 9.340578
## 22 8.295627 9.842157 8.790365 11.17939 7.226040 7.412543 9.267813
## 23 8.808935 9.400758 8.702894 11.01773 6.944539 8.053695 8.921224
## 24 8.651800 9.368522 8.476081 11.29588 6.667556 7.071694 9.336514
## 25 8.298026 9.675684 8.861115 11.21505 7.316875 7.129328 9.400450
##    CTBP1-AS1   HSPA1A     SELL    P4HA1    CKAP2    ELMO1    GTSF1
## 21  8.373863 10.60785 5.866700 9.232794 9.014485 10.88822 5.578275
## 22  8.126615 10.85421 5.726393 9.200150 8.656653 10.68401 5.890657
## 23  8.484404 10.42632 6.278107 9.226551 8.244108 10.68071 5.845840
## 24  8.157645 10.87067 5.989361 9.794671 8.899038 10.76450 5.937787
## 25  8.084694 10.64348 6.251385 9.481582 8.180849 10.71228 4.922962
##       RPS4X     FHL3   SEMA3C    RPL10A  SPATA20   ITPRIP  ZNF804A
## 21 10.20354 9.475687 9.297923  9.780683 8.969004 8.030501 8.230648
## 22 10.22314 9.537080 9.813916  9.642061 9.218515 8.032228 7.943240
## 23 10.44483 9.741039 9.820876 10.185917 9.321994 8.382751 8.378532
## 24 10.22493 9.920751 9.454757  9.691150 9.098727 8.347813 7.909494
## 25 10.72840 8.996963 9.572225  9.261363 8.838107 7.775624 7.602159
##    ANKRD36BP1    BACE2    PORCN    USP40     RND3    ACADM   GPR180
## 21   7.906530 7.728812 8.595057 7.496214 5.611142 9.417765 7.893242
## 22   7.161434 7.261641 8.539727 7.473354 6.057174 9.251663 8.089585
## 23   6.989984 7.505255 8.345492 7.083213 5.914650 9.560359 7.692392
## 24   7.254221 7.185580 8.189664 7.286869 5.472070 9.639629 8.128288
## 25   6.451797 7.973019 8.424378 7.440147 5.745261 8.978150 7.554904
##       TRBC2    PARVG
## 21 7.062023 9.102626
## 22 7.054005 9.251641
## 23 8.127213 9.117042
## 24 6.748430 9.739595
## 25 7.617592 9.866001
##
## DEGenesResults preview:
##       GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## P2RX1      P2RX1 ENSG00000108405 8.45e-06        0.52 -0.9434165
## SGMS2      SGMS2 ENSG00000164023 4.13e-05        0.36 -1.4739312
## WDR25      WDR25 ENSG00000176473 4.45e-05        0.56 -0.8365013
## F13A1      F13A1 ENSG00000124491 6.12e-05        2.63  1.3950628
# Here we see that we have 57 samples and 75 DE genes.
```

```
# There is a preview of the expression data. Row names correspond to the column chipnum
# in SpAsamples.

# For each gene in row, we have attributes that will be added further as node attributes
# in our networks. Notably, we have here the both gene identifiers HGNC symbols and
# Ensembl IDs, and p-values and fold changes computed by LIMMA.
```

SEMANTIC NETWORK CREATION WITH STRING

Stringgaussnet allows to construct a protein-protein interaction (PPI) network using the DE gene names. Then all known PPIs between those genes are explored. To this aim, the package uses the STRING API, which is a query application that works through the construction of a specific URL, which is also called an uniform resource identifier (URI) in this case. The entered gene names can be HGNC symbols or Ensembl IDs, the latter being more specific but less intuitive. Due to STRING server limitations, the number of characters with all gene IDs is limited to 8198 (which represents around 400 genes). The number of additional nodes can be set, but the API gives at least 10 additional genes. Then this package proposes to remove all additional nodes a posteriori by your request. Those added nodes are useful to see indirect interactions between initial nodes. By default, two times the number of initial DE genes identifiers are requested to STRING, in order to get a maximal covering. Different species can be curated, the default being homo sapiens. Species are entered with taxon identifiers. To see correspondence, please have a look here: http://www.uniprot.org/taxonomy
This request through STRING API constructs an object of class STRINGNet, with a network with multiple edges (depending on sources and combined scores). Its print function gives the number of initial and added nodes, and respective numbers of interactions. The summary function displays minimum, maximum, mean and median scores from different sources of interactions. From this object, it is possible to select specific sources of interactions, and to filter on the scores given by STRING. After this step, stringgaussnet calculates a new combined score, based on the calculation given for STRING version 8.1. Computation of more recent methods is not implemented, due to the lack of information concerning their precise algorithm.
Let's see an example in our package:

```
library(stringgaussnet)
data(SpADataExpression)
data(SpADEGenes)
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)
SpASTRINGNet<-getSTRINGNet(SpAData)
# Here we get the STRING network with default parameters.
# You can type help("getSTRINGNet") for more details.

print(SpASTRINGNet,5)
## Object of class STRINGNet (package stringgaussnet)
##
```

```
## Total number of nodes: 235
## Number of initial nodes: 53
## Number of added nodes: 182
##
## Total number of interactions: 5099
## Number of interactions between initial nodes: 17
## Number of interactions with added nodes: 5082
##
## Edges preview:
##              node1 node2    Interaction Score
## coexpression   RPLP1 RPL36   coexpression 0.931
## experimental   RPLP1 RPL36   experimental 0.929
## knowledge      RPLP1 RPL36      knowledge 0.900
## textmining     RPLP1 RPL36     textmining 0.233
## combined_score RPLP1 RPL36 combined_score 0.999
##
## DEGenes preview:
##       GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## SGMS2      SGMS2 ENSG00000164023 4.13e-05        0.36 -1.4739312
## F13A1      F13A1 ENSG00000124491 6.12e-05        2.63  1.3950628
## LRRC4      LRRC4 ENSG00000128594 8.03e-05        0.42 -1.2515388
## EIF4H      EIF4H ENSG00000106682 1.11e-04        0.34 -1.5563933
# We can see that STRING gave a network with 53 from the 75 initial genes
# entered in the API. 183 additional nodes were used to construct the network.

# We can see also that 17 interactions between initial nodes were found, on contrary
to
# 5168 including added nodes. Multiple edges are not taken into account, which
means
# that the print function displays unique pairs of genes as interactions.

# We have here, for each edge, an interaction source attribute and the
corresponding
# score given by STRING. Combined scores are also entered, with the label
“combined”.

# As we can see, differential analysis results are used as node attributes.

summary(SpASTRINGNet)
## All interactions:
##              coexpression cooccurence experimental fusion    knowledge
## Count       4533.0000000 116.0000000 4106.0000000  1.000 3278.0000000
## Min score      0.0640000   0.0057970    0.0430000  0.485    0.3600000
## Max score      0.9750000   0.5250000    0.9990000  0.485    0.9000000
## Mean score     0.7713536   0.2541574    0.7979408  0.485    0.8955888
## Median score   0.9360000   0.2405000    0.9300000  0.485    0.9000000
##              neighborhood  textmining
```

```
## Count       1124.0000000 4793.000000
## Min score      0.0650000    0.002376
## Max score      0.6080000    0.999000
## Mean score     0.3569448    0.419683
## Median score   0.4620000    0.401000
##
## Interactions between initial nodes:
##              coexpression cooccurence experimental knowledge neighborhood
## Count          11.0000000   2.0000000   12.0000000       7.0      4.00000
## Min score       0.1570000   0.1009470    0.1090000       0.9      0.27300
## Max score       0.9750000   0.3750000    0.9990000       0.9      0.46200
## Mean score      0.6133636   0.2379735    0.5840833       0.9      0.41475
## Median score    0.7390000   0.2379735    0.4700000       0.9      0.46200
##              textmining
## Count        15.0000000
## Min score     0.1791130
## Max score     0.7420000
## Mean score    0.3596075
## Median score  0.3150000
##
## Interactions with added nodes:
##              coexpression cooccurence experimental fusion    knowledge
## Count        4522.0000000 114.0000000 4094.0000000  1.000 3271.0000000
## Min score       0.0640000   0.0057970    0.0430000  0.485    0.3600000
## Max score       0.9750000   0.5250000    0.9990000  0.485    0.9000000
## Mean score      0.7717379   0.2544413    0.7985677  0.485    0.8955793
## Median score    0.9360000   0.2405000    0.9300000  0.485    0.9000000
##              neighborhood   textmining
## Count        1120.0000000 4778.0000000
## Min score       0.0650000    0.0023760
## Max score       0.6080000    0.9990000
## Mean score      0.3567384    0.4198716
## Median score    0.4620000    0.4010000
# Here we have score summaries for each interaction source and by making a
distinction
# between initial and added nodes.

PPISpASTRINGNet <- selectInteractionTypes(SpASTRINGNet,
  c("coexpression","experimental","knowledge"), 0.9)
# Here we select only interactions of kind “coexpression”, “experimental” and
# “knowledge”, with a score filtering threshold of 0.9.

print(PPISpASTRINGNet,5)
## Object of class STRINGNet (package stringgaussnet)
##
## Total number of nodes: 197
## Number of initial nodes: 25
## Number of added nodes: 172
```

```
##
## Total number of interactions: 4027
## Number of interactions between initial nodes: 7
## Number of interactions with added nodes: 4020
##
## Edges preview:
##              node1 node2  Interaction Score
## coexpression  RPL39 RPL36 coexpression 0.966
## experimental  RPL39 RPL36 experimental 0.986
## knowledge     RPL39 RPL36    knowledge 0.900
## coexpression1 RPL11 RPL13 coexpression 0.975
## experimental1 RPL11 RPL13 experimental 0.999
##
## DEGenes preview:
##         GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## LRRC4         LRRC4 ENSG00000128594 8.03e-05        0.42 -1.2515388
## SAP130       SAP130 ENSG00000136715 1.15e-04        0.65 -0.6214884
## SLU7           SLU7 ENSG00000164609 1.50e-04        0.63 -0.6665763
## POLR1D       POLR1D ENSG00000186184 2.24e-04        0.65 -0.6214884
## TNFSF13B   TNFSF13B ENSG00000102524 3.15e-04        0.61 -0.7131189
summary(PPISpASTRINGNet)
## All interactions:
##              coexpression experimental knowledge
## Count       2597.0000000 2352.0000000    3232.0
## Min score      0.9000000    0.9000000       0.9
## Max score      0.9750000    0.9990000       0.9
## Mean score     0.9589634    0.9751926       0.9
## Median score   0.9670000    0.9890000       0.9
##
## Interactions between initial nodes:
##              coexpression experimental knowledge
## Count             5.0000      4.00000       7.0
## Min score         0.9180      0.93000       0.9
## Max score         0.9750      0.99900       0.9
## Mean score        0.9518      0.98025       0.9
## Median score      0.9580      0.99600       0.9
##
## Interactions with added nodes:
##              coexpression experimental knowledge
## Count       2592.0000000  2348.000000    3225.0
## Min score      0.9000000     0.900000       0.9
## Max score      0.9750000     0.999000       0.9
## Mean score     0.9589772     0.975184       0.9
## Median score   0.9670000     0.989000       0.9
```

SHORT PATHS FROM STRINGNET

The generated network can be large and dense. As a STRINGNet object, it can be reduced by

computing shortest paths between genes of a user's list (Figure 2). To this aim, combined scores S are converted to distances D for each node pair i with Di=max(Si)+1-Si, where max(Si) is the maximum of S over all interactions. The distance represents a value comprised between 1 and 2, and higher is the score, lower is the distance. The shortest paths between each pair of given nodes are computed with the Dijkstra's algorithm, provided in the R package igraph. This method creates an object of class ShortPathSTRINGNet, with unique edges giving distances and intermediates as attributes. The print and summary functions are quite similar to STRINGNet, but they focus more on the distance attributes than on the scores. It is then possible to filter on the distance, if you wish to look only for closest interactions.

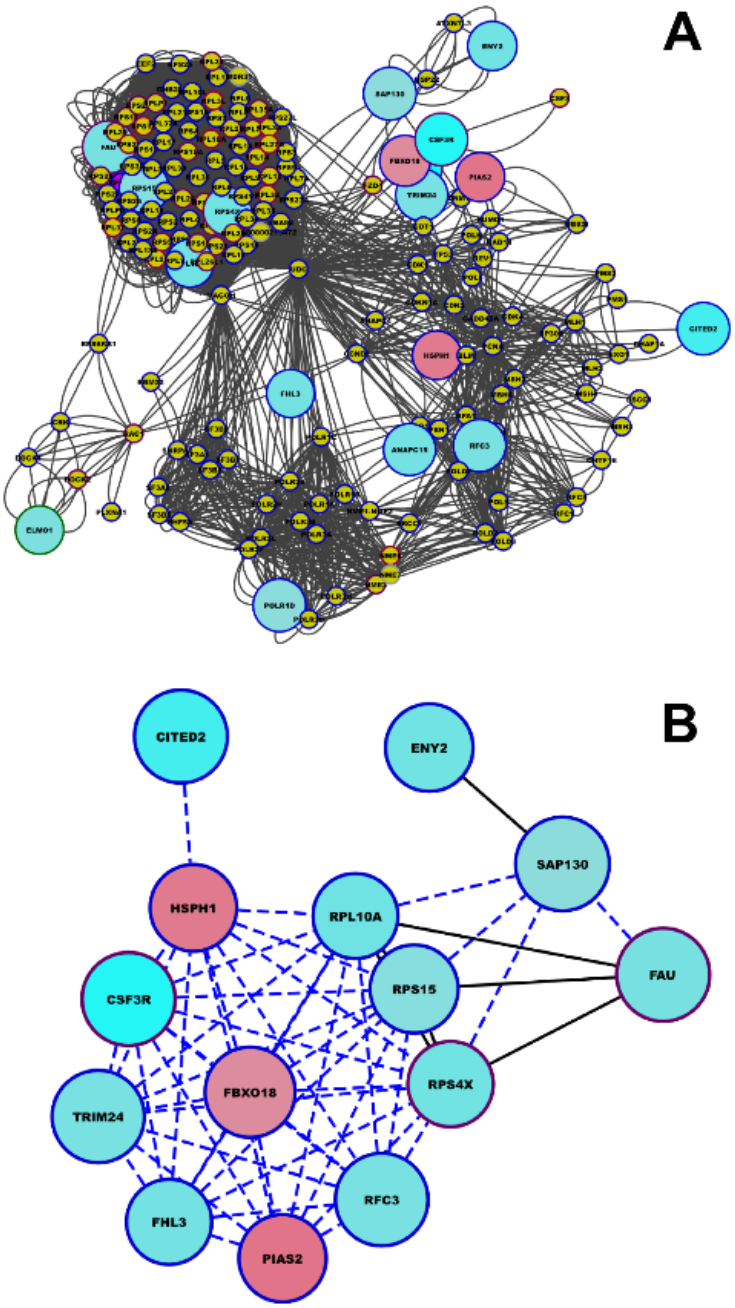


Figure 2: From STRINGnet (A) to ShortPathSTRINGNet (B) object. We can see that we considerably reduce displayed information noise for large networks. Dashed blue lines in B represent indirect interactions.

Let's take a look in the example:

```
library(stringgaussnet)
data(SpADataExpression)
data(SpADEGenes)
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)
```

```
SpASTRINGNet<-getSTRINGNet(SpAData)
PPISpASTRINGNet <- selectInteractionTypes(SpASTRINGNet,
  c("coexpression","experimental","knowledge"), 0.9)
shortPathSpANet<-getShortestPaths(PPISpASTRINGNet)
# Here we get the short paths STRING network with default parameters.
# You can type help("getShortestPaths") for more details.
shortPathSpANet<-FilterEdges(shortPathSpANet,5)
# Here we can filter on the distance between two nodes.
print(shortPathSpANet,5)
## Object of class ShortPathSTRINGNet (package stringgaussnet)
##
## Total number of nodes: 18
## Number of initial nodes: 18
## Number of added nodes: 0
## Number of intermediate nodes: 20
## Total number of interactions: 151
## Number of interactions between initial nodes: 151
## Number of interactions with added nodes: 0
##
## Edges preview:
##    node1  node2     Interaction Distance NIntermediates    Intermediates
## 1 SAP130 POLR1D shortestpathway 3.109328             2     MAGOH,POLR2E
## 2 SAP130  CSF3R shortestpathway 3.120991             2       MAGOH,UBC
## 3 SAP130    FAU shortestpathway 2.103994             1           MAGOH
## 4 SAP130 CITED2 shortestpathway 4.139095             3 DNMT1,PCNA,EP300
## 5 SAP130   RFC3 shortestpathway 3.101098             2      DNMT1,PCNA
##
## DEGenes preview:
##        GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## SAP130     SAP130 ENSG00000136715 0.000115        0.65 -0.6214884
## POLR1D     POLR1D ENSG00000186184 0.000224        0.65 -0.6214884
## CSF3R       CSF3R ENSG00000119535 0.000564        0.32 -1.6438562
## FAU           FAU ENSG00000149806 0.000603        0.56 -0.8365013
## CITED2     CITED2 ENSG00000164442 0.000729        0.40 -1.3219281
# Here we don't have any added node, because we summarized the network only
between
# initial nodes. We can see that 20 genes were used as intermediates.

# We have unique edges with distance, number of intermediates and intermediate
names
# as attributes.
```

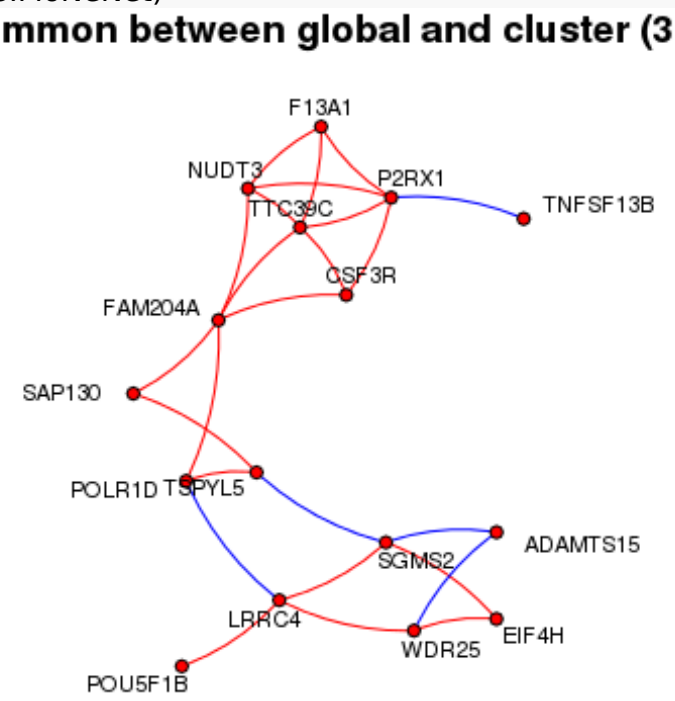
SIMONE NETWORK INFERENCE

Another use of this package is to infer a gaussian network from expression data. We implemented the use of the R package simone, in order to strongly reduce noise from indirect interactions without supervision.(Chiquet et al. 2009) All options from SIMoNe are changeable in stringgaussnet, and default values are given for users who want to discover this tool. The default

method to select the inferred model by SIMoNe in our package helps to make a choice with a better compromise. The number of edges is selected by computing the mean between those with maximal AIC and BIC scores. You can choose otherwise a fixed edges number, or to base only on AIC or BIC score. Without choice from you, the algorithm computes a network with or without clustering constraints, and selects only common edges between the both models. A function is provided to help in selecting the best model inferred by SIMoNe with a series of graphs, which are already implemented in the simone package. Notably, you can see all BIC and AIC scores as a function of the penalty level given in the graphical LASSO method. In addition to the theta score given by SIMoNe, stringgaussnet computes Spearman's test for each inferred edge, with an AS89 approximation of null distribution.
This inference creates an object of class SIMoNeNet, with a network of unique edges including theta, Spearman's rho and p-value as attributes. The associated print function displays the number of nodes and edges, with a preview of node and edge attributes. The summary function summarizes theta scores, Spearman's rho and their absolute values, and Spearman's test p-values. This is possible to filter on edge attributes, notably on Spearman's rho. This is useful for large networks, considering that SIMoNe can infer interactions without strong correlations. A function is provided to plot a network preview, based on the original simone package.
Let's see how it works with the example data:

```
library(stringgaussnet)
data(SpADataExpression)
data(SpADEGenes)
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)
NodesForSIMoNe<-rownames(SpADEGenes)[1:17]
GaussianSpAData<-DEGeneExpr(
  t(SpADataExpression[NodesForSIMoNe,]),
  SpADEGenes[NodesForSIMoNe,])
# We select a reasonable number of genes for SIMoNe network inference.
# We advice to take a number of genes being inferior to the sample size.
#pickSIMoNeParam(GaussianSpAData)
# We use a series of plot provided with the simone package to see which penalty
level
# we can use for the graphical LASSO regression.
GlobalSIMoNeNet<-getSIMoNeNet(GaussianSpAData)
##
## Found a network with 36 edges.
##
## Found a network with 36 edges.
##
## Found a network with 36 edges.
# Here we get the SIMoNe network with default parameters.
# You can type help("getSIMoNeNet") for more details.
GlobalSIMoNeNet<-FilterEdges(GlobalSIMoNeNet,0.4)
# Here we can filter on the absolute values of rho being superior to 0.4.
print(GlobalSIMoNeNet,5)
## Object of class SIMoNeNet (package stringgaussnet)
##
## Number of nodes: 16
```

```
## Number of interactions: 25
##
## Edges preview:
##   node1   node2     Interaction       Theta       Rho P.Value
## 1 NUDT3   P2RX1 SIMoNeInference -0.38204142 0.9008297       0
## 3 NUDT3   F13A1 SIMoNeInference -0.08337518 0.7824086       0
## 4 P2RX1   F13A1 SIMoNeInference -0.04243323 0.7665932       0
## 5 NUDT3 FAM204A SIMoNeInference -0.14221514 0.7543550       0
## 6 SGMS2   LRRC4 SIMoNeInference -0.44771356 0.8002982       0
##
## DEGenes preview:
##       GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## P2RX1      P2RX1 ENSG00000108405 8.45e-06        0.52 -0.9434165
## SGMS2      SGMS2 ENSG00000164023 4.13e-05        0.36 -1.4739312
## WDR25      WDR25 ENSG00000176473 4.45e-05        0.56 -0.8365013
## F13A1      F13A1 ENSG00000124491 6.12e-05        2.63  1.3950628
# We have unique edges with Theta scores, and Spearman's rhos and p-values.
plot(GlobalSIMoNeNet)
```




```
# Here we have the network displayed with the plot function provided with the
# simone package.
```

WGCNA NETWORK INFERENCE AND COMPARISON WITH SIMONE

SIMoNe is a powerful statistical approach to infer non-supervised gaussian networks. However, the algorithm principle is not intuitive for any beginner in the graph theory domain and statistical inference. This is why we propose to compare with a more trivial approach: WGCNA. Stringgaussnet allows to compute Spearman's test between all pairs of genes, and respective rhos are converted to similarity scores σ with $\sigma=(1+rho)/2$, in order to keep the correlation signs. Then those scores are converted to proximity scores A, with $A=1/(1+exp(-\alpha*(\sigma-0,5)))$. α is the soft power threshold, and is by default 8 in the package. A function is provided to help in choosing this parameter, by giving a series of plots representing relations between A and rho (Figure 3). Then a filtering step is proceeded with a threshold t, A being superior to t or inferior to 1-t. By default, t is 0,85, as suggested by the WGCNA tutorial. Dissimilarities and modules computations are not implemented, because the main purpose is to compare with SIMoNe and to empower its use.

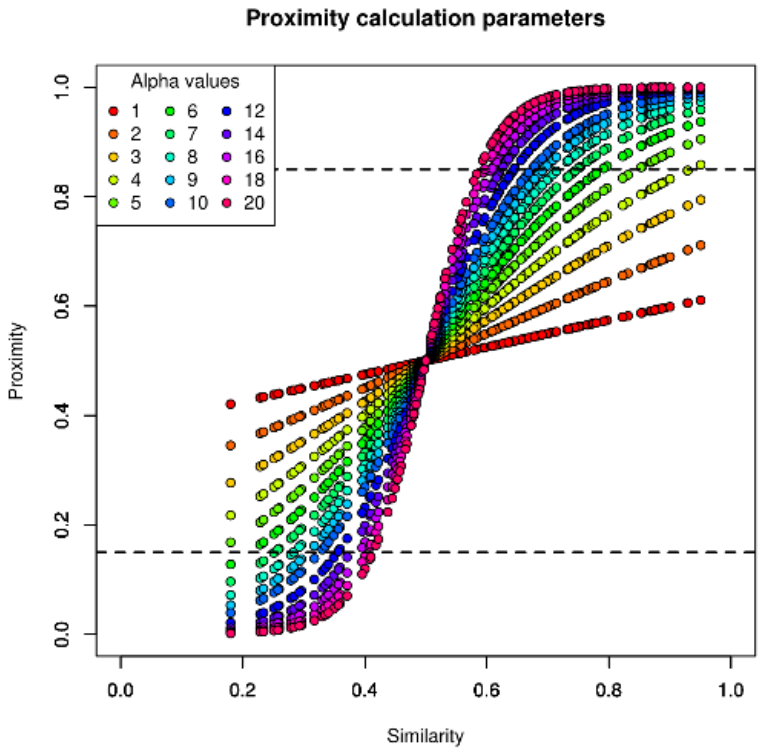


Figure 3: A plot to help in choosing the soft thresholding parameter. Different alpha values are used to visualize different possibilities.

The obtained network is saved in a WGCNANet class object, with print and summary functions much similar to SIMoNeNet. This is also possible to draw the inferred network in the same way as for SIMoNeNet. A function is provided to compare inferred networks from both SIMoNe and WGCNA, with venn diagram displaying respective connectivities, and a series of plots showing correlations of picked interactions (Figure 4). Please see help("compareGaussNetworks") for more informations.

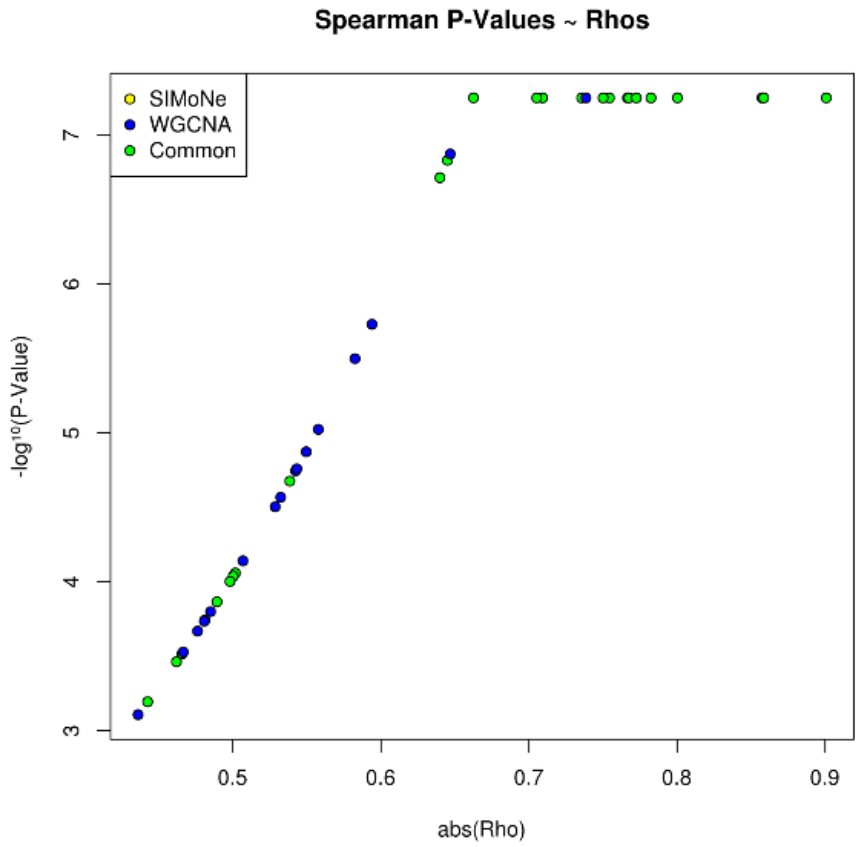
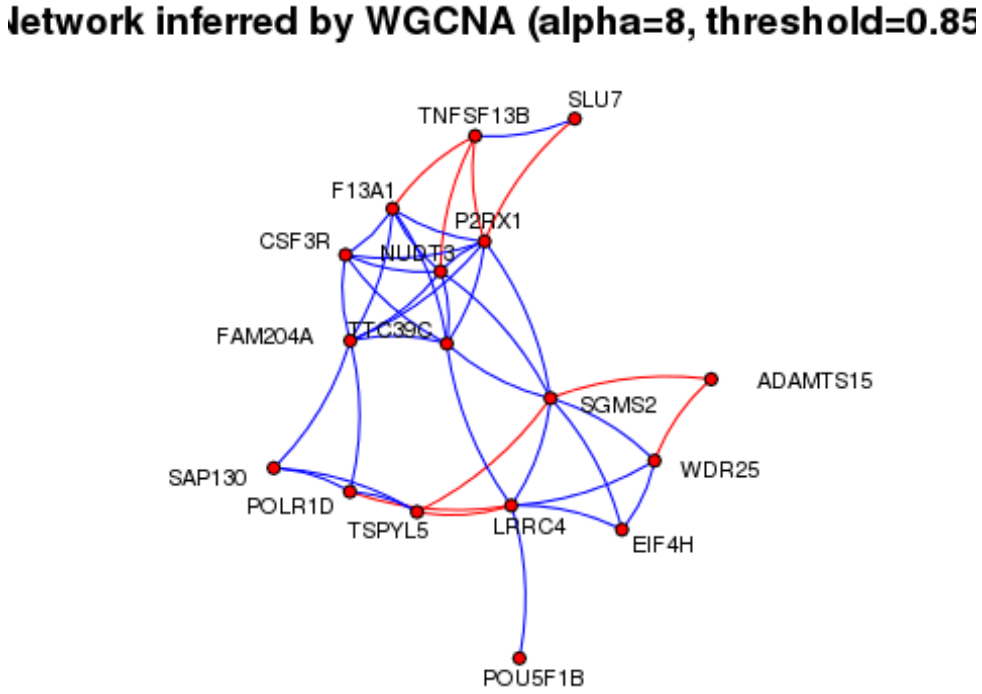


Figure 4: Comparison plot of Spearman's rho and p-value between SIMoNe WGCNA infered networks. We can notice that SIMoNe removes a lot of interactions thanks to the partial correlation computation.

Let's have a look in the example:

```
library(stringgaussnet)
data(SpADataExpression)
data(SpADEGenes)
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)
NodesForSIMoNe<-rownames(SpADEGenes)[1:17]
GaussianSpAData<-DEGeneExpr(
  t(SpADataExpression[NodesForSIMoNe,]),
  SpADEGenes[NodesForSIMoNe,])
#pickWGCNAParam(GaussianSpAData)
# Here we use a list of plots to help in choosing the right parameter for
```

```
# WGCNA computing. You can see help("pickWGCNAParam") for more details.
GlobalWGCNANet<-getWGCNANet(GaussianSpAData)
# Here we get the WGCNA network with default parameters.
# You can type help("getWGCNANet") for more details.
print(GlobalWGCNANet,5)
## Object of class WGCNANet (package stringgaussnet)
##
## Number of nodes: 17
## Number of interactions: 41
## Edges preview:
##   node1 node2     Interaction Adjacency       Rho      P.Value
## 1 NUDT3 P2RX1 SIMoNeInference 0.9734888 0.9008297 0.000000e+00
## 2 NUDT3 SGMS2 SIMoNeInference 0.8836258 0.5068058 7.250921e-05
## 3 P2RX1 SGMS2 SIMoNeInference 0.8724723 0.4807493 1.840595e-04
## 4 SGMS2 WDR25 SIMoNeInference 0.8923138 0.5286492 3.143367e-05
## 5 NUDT3 F13A1 SIMoNeInference 0.9580987 0.7824086 0.000000e+00
##
## DEGenes preview:
##       GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## P2RX1      P2RX1 ENSG00000108405 8.45e-06        0.52 -0.9434165
## SGMS2      SGMS2 ENSG00000164023 4.13e-05        0.36 -1.4739312
## WDR25      WDR25 ENSG00000176473 4.45e-05        0.56 -0.8365013
## F13A1      F13A1 ENSG00000124491 6.12e-05        2.63  1.3950628
# We have adjacency scores, and Spearman's rhos and p-values for each edge.
plot(GlobalWGCNANet)
```




```
# Here we have the network displayed with the plot function provided in the
# simone package.
```

ADDING ANNOTATIONS TO GENES

For each of those different described kinds of networks, this is possible to add gene annotations as node attributes. This option is usable at the same step as for network generation. This enrichment adds two kinds of informations. First, stringgaussnet uses the R package biomaRt to get mainly genomic localization and gene description. Secondly, it adds cellular component terms with the package GO.db. Because several components can be linked to one gene, a prioritization is performed to rank genes products localizations from nuclear, the most important, and then extracellular, plasma membrane and cytoplasm. Indeed, we suppose that cytoplasmic localization is the lowest important information, because the most part of genes go through this component to arrive in others. To use this feature, you can use the parameter AddAnnotations=TRUE for each network creation function.

Let's see an example for SIMoNeNet:

```
library(stringgaussnet)
data(SpADataExpression)
data(SpADEGenes)
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)
NodesForSIMoNe<-rownames(SpADEGenes)[1:17]
GaussianSpAData<-DEGeneExpr(
  t(SpADataExpression[NodesForSIMoNe,]),
  SpADEGenes[NodesForSIMoNe,])
GlobalSIMoNeNet<-getSIMoNeNet(GaussianSpAData,
```

```
  AddAnnotations=TRUE)
# Here we use the parameter "AddAnnotations=TRUE" to add annotations to genes
from
# the network.
print(GlobalSIMoNeNet,5)
## Object of class SIMoNeNet (package stringgaussnet)
##
## Number of nodes: 17
## Number of interactions: 36
##
## Edges preview:
##   node1   node2     Interaction       Theta       Rho    P.Value
## 1 NUDT3   P2RX1 SIMoNeInference -0.38204142 0.9008297 0.00000000
## 2 NUDT3   WDR25 SIMoNeInference -0.04485639 0.3279103 0.01311308
## 3 NUDT3   F13A1 SIMoNeInference -0.08337518 0.7824086 0.00000000
## 4 P2RX1   F13A1 SIMoNeInference -0.04243323 0.7665932 0.00000000
## 5 NUDT3 FAM204A SIMoNeInference -0.14221514 0.7543550 0.00000000
##
## DEGenes preview:
##       GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## P2RX1      P2RX1 ENSG00000108405 8.45e-06        0.52 -0.9434165
## SGMS2      SGMS2 ENSG00000164023 4.13e-05        0.36 -1.4739312
## WDR25      WDR25 ENSG00000176473 4.45e-05        0.56 -0.8365013
## F13A1      F13A1 ENSG00000124491 6.12e-05        2.63  1.3950628
##
## Annotations preview:
##         ensembl_gene_id  localization hgnc_symbol chromosome_name   band
## ADAMTS15 ENSG00000166106 extracellular    ADAMTS15              11  q24.3
## CSF3R    ENSG00000119535 extracellular       CSF3R               1  p34.3
## EIF4H    ENSG00000106682     cytoplasm       EIF4H               7 q11.23
## F13A1    ENSG00000124491 extracellular       F13A1               6  p25.1
## FAM204A  ENSG00000165669         <NA>     FAM204A              10 q26.11
##         strand start_position end_position
## ADAMTS15      1     130448974    130476641
## CSF3R        -1      36466043     36483278
## EIF4H         1      74174245     74197101
## F13A1        -1       6144085      6321013
## FAM204A      -1     118297930    118342328
##                                                                  description
## ADAMTS15 ADAM metallopeptidase with thrombospondin type 1 motif, 15
[Source:HGNC Symbol;Acc:HGNC:16305]
## CSF3R            colony stimulating factor 3 receptor (granulocyte) [Source:HGNC
Symbol;Acc:HGNC:2439]
## EIF4H                eukaryotic translation initiation factor 4H [Source:HGNC
Symbol;Acc:HGNC:12741]
## F13A1                    coagulation factor XIII, A1 polypeptide [Source:HGNC
Symbol;Acc:HGNC:3531]
```

```
## FAM204A          family with sequence similarity 204, member A [Source:HGNC
Symbol;Acc:HGNC:25794]
# We can see that we have gene annotations added by biomaRt and gene product
# localizations provided by Gene Ontology.
```

MULTIPLE GAUSSIAN NETWORKS INFERENCE AS A FUNCTION OF A FACTOR

An overlay of functions allows you to create multiple networks in only one step, with all options configurable in the same method. One can create multiple Gaussian networks from the same DEGeneExpr object, depending on a grouping factor and for a given list of genes. The package then allows to compare networks inferred for multiple levels of the factor, and for the same DE genes list (Figure 5).

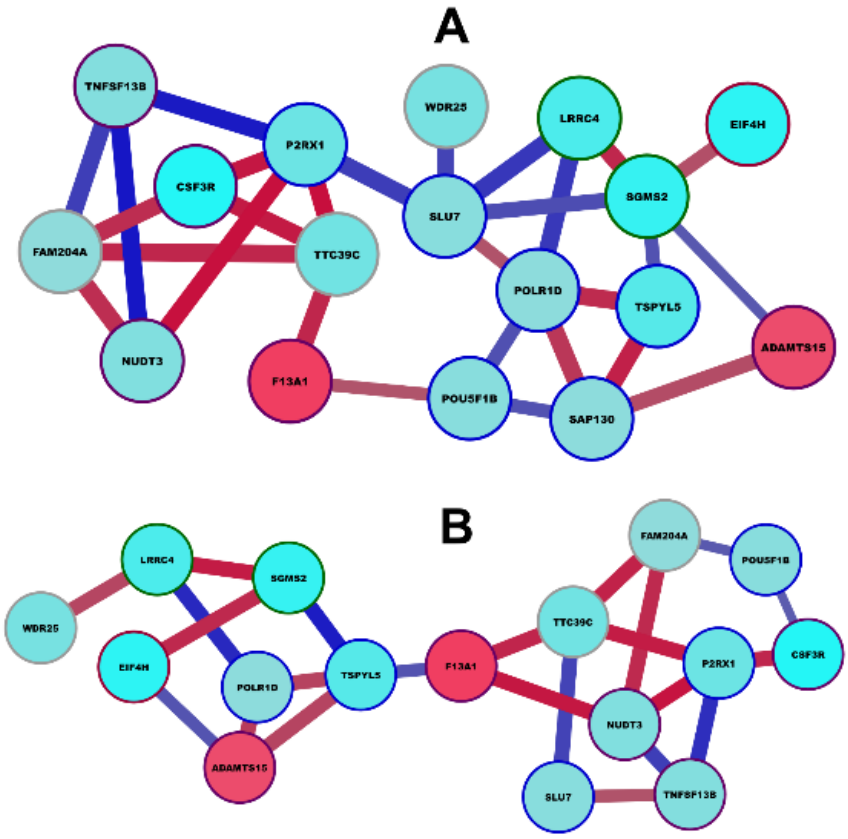


Figure 5: Example of infered SIMoNe networks between two groups (A and B) of a given factor (disease status in our example). Stringgaussnet allows to generate automatically the both networks in one step.

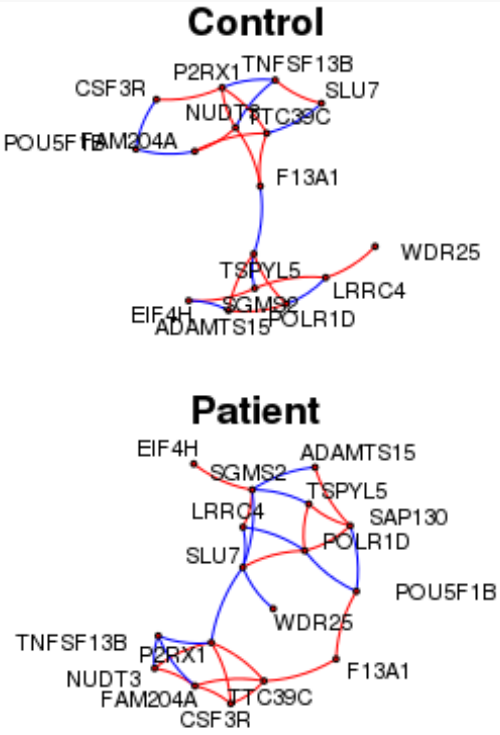
Here is an example for SIMoNeNet and patient status:

```
library(stringgaussnet)
data(SpADataExpression)
data(SpADEGenes)
data(SpASamples)
```

```
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)
StatusFactor<-SpASamples$status
names(StatusFactor)<-SpASamples$chipnum
# We create a factor vector based on the status.
NodesForSIMoNe<-rownames(SpADEGenes)[1:17]
GaussianSpAData<-DEGeneExpr(
  t(SpADataExpression[NodesForSIMoNe,]),
  SpADEGenes[NodesForSIMoNe,])
StatusFactorSIMoNeNet<-FactorNetworks(GaussianSpAData,
  StatusFactor,"SIMoNe")
## Level: Control
##
## Found a network with 34 edges.
##
## Found a network with 34 edges.
##
## Found a network with 34 edges.
## Level: Patient
##
## Found a network with 38 edges.
##
## Found a network with 38 edges.
##
## Found a network with 38 edges.
# We infer different SIMoNe networks on different groups of samples
# (patients and controls).
StatusFactorSIMoNeNet<-FilterEdges(StatusFactorSIMoNeNet,0.4)
# We can filter on edges, like for SIMoNeNet.
print(StatusFactorSIMoNeNet)
## Object of class FactorNetworks (package stringgaussnet)
##
## Levels distribution:
## Control Patient
##      30      27
##
## Control:
##
## Object of class DEGeneExpr (package stringgaussnet)
##
## Number of samples: 30
## Number of genes: 17
##
## DataExpression preview:
##       NUDT3    P2RX1    SGMS2    WDR25    F13A1  FAM204A    LRRC4    EIF4H
## 31 10.53476 8.356974 7.519848 7.661825 13.35763 8.663832 6.544332
10.73350
## 32 10.47816 8.205153 7.915057 7.674115 13.61540 8.873416 6.676184
11.10315
```

```
##     SAP130    SLU7  POU5F1B   POLR1D   TTC39C ADAMTS15 TNFSF13B   TSPYL5
## 31 10.21238 8.166741 6.673612 9.473133 7.468633 7.515297 10.46620
8.811882
## 32 10.39199 8.212908 6.523527 9.760605 7.596354 8.080091 11.00189
8.496030
##      CSF3R
## 31 7.508418
## 32 8.067584
##
## DEGenesResults preview:
##      GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## P2RX1      P2RX1 ENSG00000108405 8.45e-06        0.52 -0.9434165
##
## Object of class SIMoNeNet (package stringgaussnet)
##
## Number of nodes: 16
## Number of interactions: 23
##
## Edges preview:
##   node1 node2     Interaction      Theta       Rho      P.Value
## 1 NUDT3 P2RX1 SIMoNeInference -0.2182752 0.8758621 4.668259e-07
## 2 NUDT3 F13A1 SIMoNeInference -0.3869448 0.8491657 7.174537e-07
##
## DEGenes preview:
##      GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## P2RX1      P2RX1 ENSG00000108405 8.45e-06        0.52 -0.9434165
##
## Patient:
##
## Object of class DEGeneExpr (package stringgaussnet)
##
## Number of samples: 27
## Number of genes: 17
##
## DataExpression preview:
##      NUDT3    P2RX1    SGMS2    WDR25    F13A1  FAM204A    LRRC4    EIF4H
## 21 10.25609 7.779726 7.478363 7.395941 13.53042 8.865439 6.196102
10.35545
## 22 10.17532 7.713649 7.426126 7.482414 13.39109 8.743199 6.715109
10.10438
##     SAP130    SLU7  POU5F1B   POLR1D   TTC39C ADAMTS15 TNFSF13B   TSPYL5
## 21 10.17557 8.517671 6.425804 9.553580 7.429002 7.779027 10.17795
8.561254
## 22 10.18127 7.944881 6.761638 9.350876 7.147367 8.855387 10.84027
8.214381
##      CSF3R
```

```
## 21 7.383586
## 22 7.135964
##
## DEGenesResults preview:
##       GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## P2RX1      P2RX1 ENSG00000108405 8.45e-06        0.52 -0.9434165
##
## Object of class SIMoNeNet (package stringgaussnet)
##
## Number of nodes: 17
## Number of interactions: 28
##
## Edges preview:
##   node1   node2     Interaction      Theta       Rho      P.Value
## 1 NUDT3   P2RX1 SIMoNeInference -0.3228745 0.8925519 1.164776e-06
## 3 NUDT3 FAM204A SIMoNeInference -0.1920048 0.6990232 7.705345e-05
##
## DEGenes preview:
##       GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## P2RX1      P2RX1 ENSG00000108405 8.45e-06        0.52 -0.9434165
# We can see that we have a preview of inferred networks for each level.
par(mfrow=c(2,1))
plot(StatusFactorSIMoNeNet,interactiveMode=F)
```



```
# Here we can display networks inferred in patients and controls specifically, like in
# the Figure 5, with the provided function in the simone package.

#compareFactorNetworks(StatusFactorSIMoNeNet)
# Here we can have a series of plots to compare results of inferred networks for
# each level. You can see help ("compareFactorNetworks") for more details.

#StatusFactorSIMoNeNet<-FactorNetworks(GaussianSpAData,
#  StatusFactor,"SIMoNe",list(AddAnnotations=TRUE))
# This is how you should do if you wanted to add gene annotations.
```

MULTIPLE NETWORKS GENERATION FROM A LIST OF DIFFERENTIAL ANALYSIS RESULTS

One can create an object of class MultiDEGeneExpr, a list of DEGeneExpr objects. Then, both kinds of networks, semantic or Gaussian, can be generated for each data set and stored in a MultiNetworks object.. This wrapper is useful to explore all possible interactions between a set of DE genes lists. All options for each network generation are accessible through an unique function. Let's have a look in an example:

```
library(stringgaussnet)
data(SpADataExpression)
data(SpADEGenes)
data(SpASamples)
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)
StatusFactor<-SpASamples$status
names(StatusFactor)<-SpASamples$chipnum
NodesForSIMoNe<-rownames(SpADEGenes)[1:17]
GaussianSpAData<-DEGeneExpr(
  t(SpADataExpression[NodesForSIMoNe,]),
  SpADEGenes[NodesForSIMoNe,])
MultiSpAData<-MultiDEGeneExpr(GaussianSpAData,
  DEGeneExpr(t(SpADataExpression[18:34,]),
  SpADEGenes[18:34,]),
  DEGeneExpr(t(SpADataExpression[35:51,]),
  SpADEGenes[35:51,]))
# We create multiple lists of DE genes results, and then a list of DEGeneExpr
objects,
# by subsetting the original data.
print(MultiSpAData)
## Object of class MultiDEGeneExpr (package stringgaussnet)
##
## 3 objects of class DEGeneExpr (List1, List2, List3)
##
## List1:
## Object of class DEGeneExpr (package stringgaussnet)
##
## Number of samples: 57
```

```
## Number of genes: 17
##
## DataExpression preview:
##      NUDT3    P2RX1    SGMS2    WDR25    F13A1  FAM204A    LRRC4    EIF4H
## 21 10.25609 7.779726 7.478363 7.395941 13.53042 8.865439 6.196102
10.35545
## 22 10.17532 7.713649 7.426126 7.482414 13.39109 8.743199 6.715109
10.10438
##     SAP130     SLU7  POU5F1B   POLR1D   TTC39C ADAMTS15 TNFSF13B   TSPYL5
## 21 10.17557 8.517671 6.425804 9.553580 7.429002 7.779027 10.17795
8.561254
## 22 10.18127 7.944881 6.761638 9.350876 7.147367 8.855387 10.84027
8.214381
##      CSF3R
## 21 7.383586
## 22 7.135964
##
## DEGenesResults preview:
##      GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## NUDT3      NUDT3 ENSG00000112664 4.60e-06        0.60 -0.7369656
## P2RX1      P2RX1 ENSG00000108405 8.45e-06        0.52 -0.9434165
##
## List2:
## Object of class DEGeneExpr (package stringgaussnet)
##
## Number of samples: 57
## Number of genes: 17
##
## DataExpression preview:
##       FAU     TFAM     FAIM2    CITED2 SIGLEC15     MBIP    HAUS1
## 21 10.30000 6.797904  6.753533 11.23097 7.730661 8.265060 9.547622
## 22 10.30035 6.786641 10.433228 11.00969 8.486743 8.373733 9.286334
##      RFC3     TBCK   TRIM24  ANAPC15     PIGB    DNAJA4    RBBP9 KIAA0907
## 21 6.717628 9.507269 9.164891 10.42642 8.663999 8.272130 8.143958
9.352399
## 22 6.572705 9.561226 9.012223 10.77170 8.803228 8.370383 7.665083
9.118642
##       MUT   FBXO18
## 21 8.188985 8.265647
## 22 8.275170 8.363612
##
## DEGenesResults preview:
##     GeneSymbol       EnsemblId  P.Value Fold.Change      logFC
## FAU         FAU ENSG00000149806 0.000603        0.56 -0.8365013
## TFAM       TFAM ENSG00000108064 0.000640        0.63 -0.6665763
##
## List3:
## Object of class DEGeneExpr (package stringgaussnet)
```

```
##
## Number of samples: 57
## Number of genes: 17
##
## DataExpression preview:
##      USP30     NDP     OLR1     PLP2     MNDA    IFT52    COX20   BAK1P1
## 21 8.903849 7.663387 6.220560 11.36747 8.173222 7.287404 7.968966
6.412227
## 22 8.475835 7.667593 6.641498 11.09409 9.675167 6.828304 8.073868
6.570384
##      CRTAP    RPS15   ALG10B     ENY2    PIAS2    FBXL4    HSPH1    PTPLA
## 21 11.28916 10.43043 7.672037 9.043001 9.677796 8.894753 10.96024
7.228479
## 22 11.00393 10.43679 6.686629 8.295627 9.842157 8.790365 11.17939
7.226040
##      COX7B
## 21 7.967758
## 22 7.412543
##
## DEGenesResults preview:
##       GeneSymbol       EnsemblId P.Value Fold.Change      logFC
## USP30      USP30 ENSG00000135093 0.00320        0.60 -0.7369656
## NDP         NDP ENSG00000124479 0.00321        0.13 -2.9434165
# We have, by the specific print function, a preview of each DEGeneExpr object
# in the list.
MultiSpANetworks<-MultiNetworks(MultiSpAData,
  SelectInteractionsSTRING=c("coexpression",
  "experimental","knowledge"),STRINGThreshold=0.9,
  FilterSIMoNeOptions=list(Threshold=0.4),
  Factors=StatusFactor)
# We create an object of class MultiNetworks, which allows to generate all kinds of
# networks in multiple lists of DEGeneExpr objects in one line.
print(MultiSpANetworks)
## Object of class MultiNetworks (package stringgaussnet)
##
## 3 object(s) of class DEGeneExpr used (List1, List2, List3)
##
## 3 method(s) of network creation used (STRING, SIMoNe, WGCNA)
##
## A factor with 2 levels has been entered by the user (Control, Patient)
# We simply have a summary of the used method to generate the MultiNetworks
object.

#MultiSpANetworks<-MultiNetworks(MultiSpAData,
#  SelectInteractionsSTRING=c("coexpression",
#  "experimental","knowledge"),STRINGThreshold=0.9,
#  FilterSIMoNeOptions=list(Threshold=0.4),
#  Factors=StatusFactor,
```

```
#  STRINGOptions=list(AddAnnotations=TRUE),
#  SIMoNeOptions=list(AddAnnotations=TRUE),
#  WGCNAOptions=list(AddAnnotations=TRUE))
# This is how you should do if you wanted to add gene annotations for all network
generations.
```

NETWORK EXPORTATION INTO FILES AND CYTOSCAPE

An important feature of stringgaussnet is to allow to visualize and manipulate all generated networks only in few steps, without any knowledge in network files manipulations and package object structures. Firstly, a generic export function is available to save all generated networks in standard edge and node attributes format.
Let's see here an example for SIMoNeNet:

```
library(stringgaussnet)
data(SpADataExpression)
data(SpADEGenes)
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)
NodesForSIMoNe<-rownames(SpADEGenes)[1:17]
GaussianSpAData<-DEGeneExpr(
  t(SpADataExpression[NodesForSIMoNe,]),
  SpADEGenes[NodesForSIMoNe,])
GlobalSIMoNeNet<-getSIMoNeNet(GaussianSpAData,
  AddAnnotations=TRUE)
GlobalSIMoNeNet<-FilterEdges(GlobalSIMoNeNet,0.4)
export(GlobalSIMoNeNet,YourDirPath)
# Replace YourDirPath by the directory where will be saved edge and node
attributes.
# If the directory exists, this will not be overwritten,
# excepted if you use the parameter “overwrite=TRUE”.
```

But the most interesting part is to be able to import automatically all generated networks from R objects to Cytoscape, without any requirement of secondary language. Cytoscape is a powerful software to compute biological networks. Automatic importation through stringgaussnet uses the plugin cyREST, which works like a local API. Thus, for this feature, Cytoscape must be running and this plugin installed. At least java version 8 must be installed on the computer, which is the case for the most windows running computers. Moreover, this importation is not operating system dependent, such as the R programming language. For more information about software requirements, please see Hardware and software requirements.
Cytoscape uses styles to display imported networks. Our package generates predefined styles, which can be then easily modified in Cytoscape. But you can choose to handle custom styles available in the running Cytoscape session. Node sizes and colors are dependent by default on fold change and p-value from DE genes analysis, but you can choose what attributes can be at the basis of those properties. Edge views depend on attributes from each kind of generated network, e.g. Spearman's rho for gaussian networks. Indirect interactions from ShortPathSTRINGNet are displayed with dashed blue lines. The layout can also be set, which is by default force-directed.
Let's see an example for MultiNetworks:

```
library(stringgaussnet)
data(SpADataExpression)
data(SpADEGenes)
data(SpASamples)
SpAData<-DEGeneExpr(t(SpADataExpression),SpADEGenes)
StatusFactor<-SpASamples$status
names(StatusFactor)<-SpASamples$chipnum
NodesForSIMoNe<-rownames(SpADEGenes)[1:17]
GaussianSpAData<-DEGeneExpr(
  t(SpADataExpression[NodesForSIMoNe,]),
  SpADEGenes[NodesForSIMoNe,])
MultiSpAData<-MultiDEGeneExpr(GaussianSpAData,
  DEGeneExpr(t(SpADataExpression[18:34,]),
  SpADEGenes[18:34,]),
  DEGeneExpr(t(SpADataExpression[35:51,]),
  SpADEGenes[35:51,]))
MultiSpANetworks<-MultiNetworks(MultiSpAData,
  SelectInteractionsSTRING=c("coexpression",
  "experimental","knowledge"),STRINGThreshold=0.9,
  FilterSIMoNeOptions=list(Threshold=0.4),
  Factors=StatusFactor,
  STRINGOptions=list(AddAnnotations=TRUE),
  SIMoNeOptions=list(AddAnnotations=TRUE),
  WGCNAOptions=list(AddAnnotations=TRUE))
resetCytoscapeSession()
# We reset and create an empty session in Cytoscape.
# Please be sure that Cytoscape is running with the cyREST plugin installed.
addMultiGraphToCytoscape(MultiSpANetworks,
  points.size.map="P.Value",points.fill.map="logFC")
# We add automatically all generated networks in the Cytoscape session,
# with predefined styles.
saveCytoscapeSession(YourFilePath)
# We can save the current Cytoscape session in a .cys file.
# Replace YourFilePath with the path where you would like to save.
# The .cys extension is automatically added if necessary.
```

REFERENCES

Chiquet, Julien, Alexander Smith, Gilles Grasseau, Catherine Matias, and Christophe Ambroise. 2009. “SIMoNe: Statistical Inference for MOdular NEtworks.” *Bioinformatics (Oxford, England)* 25 (3): 417–18. doi:10.1093/bioinformatics/btn637.

Cline, Melissa S., Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, et al. 2007. “Integration of Biological Networks and Gene Expression Data Using Cytoscape.” *Nature Protocols* 2 (10): 2366–82. doi:10.1038/nprot.2007.324.

Cotney, Justin, Rebecca A. Muhle, Stephan J. Sanders, Li Liu, A. Jeremy Willsey, Wei Niu,

Wenzhong Liu, et al. 2015. “The Autism-Associated Chromatin Modifier CHD8 Regulates Other Autism Risk Genes During Human Neurodevelopment.” *Nature Communications* 6 (March). doi:10.1038/ncomms7404.

Dong, Jun, and Steve Horvath. 2007. “Understanding Network Concepts in Modules.” *BMC Systems Biology* 1: 24. doi:10.1186/1752-0509-1-24.

Lin, Ying, Vusumuzi Leroy Sibanda, Hong-Mei Zhang, Hui Hu, Hui Liu, and An-Yuan Guo. 2015. “MiRNA and TF Co-Regulatory Network Analysis for the Pathology and Recurrence of Myocardial Infarction.” *Scientific Reports* 5 (April). doi:10.1038/srep09653.

Smyth, Gordon K. 2004. “Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.” *Statistical Applications in Genetics and Molecular Biology* 3: Article3. doi:10.2202/1544-6115.1027.

Talpin, Alice, Félicie Costantino, Nelly Bonilla, Ariane Leboime, Franck Letourneur, Sébastien Jacques, Florent Dumont, et al. 2014. “Monocyte-Derived Dendritic Cells from HLA-B27+ Axial Spondyloarthritis (SpA) Patients Display Altered Functional Capacity and Deregulated Gene Expression.” *Arthritis Research & Therapy* 16 (4): 417. doi:10.1186/s13075-014-0417-0.

Verfaillie, Annelien, Hana Imrichova, Zeynep Kalender Atak, Michael Dewaele, Florian Rambow, Gert Hulselmans, Valerie Christiaens, et al. 2015. “Decoding the Regulatory Landscape of Melanoma Reveals TEADS as Regulators of the Invasive Cell State.” *Nature Communications* 6 (April). doi:10.1038/ncomms7683.

Xue, Jia, Susanne V. Schmidt, Jil Sander, Astrid Draffehn, Wolfgang Krebs, Inga Quester, Dominic De Nardo, et al. 2014. “Transcriptome-Based Network Analysis Reveals a Spectrum Model of Human Macrophage Activation.” *Immunity* 40 (2): 274–88. doi:10.1016/j.immuni.2014.01.006.

Chapitre 4

DISCUSSION ET PERSPECTIVES

Bases de données relationnelles

Notre laboratoire combine plusieurs approches complémentaires afin de mieux déceler les mécanismes à l'origine du développement de la SpA. A cela s'ajoute une cohorte conséquente de près de 4000 personnes avec des informations importantes recueillies par des médecins de l'Hôpital Ambroise-Paré (Boulogne-Billancourt) ou des personnes de notre laboratoire. Il apparaît donc évident qu'une uniformisation des données est cruciale pour éviter des erreurs de manipulation et d'entrée/sortie. De plus, il est important que tout le laboratoire soit d'accord sur les cohortes que chacun étudie. C'est pourquoi nous avions créé une base de données relationnelles sous MySQL pour rassembler toutes ces informations. En outre d'effectuer plus simplement et de façon plus fiable une synthèse de la cohorte totale, elle nous avait permis de rendre plus accessible de façon sécurisée et anonymisée les informations à toutes les personnes du laboratoire. De plus, l'interface web avait permis de rendre plus ergonomique cet accès. Un point crucial pour faire communiquer ces données entre cliniciens et biologistes était de faciliter la correspondance entre les numéros CEPH et ceux de tubes ADNs, ces derniers étant la base pour beaucoup d'études effectuées par notre laboratoire.

D'autre part, nous avions développé une base de données relationnelles regroupant les résultats de puces transcriptomiques. Cette base de données était développée également sous MySQL et regroupait la description des modèles linéaires utilisés par LIMMA et quSAGE (matrices de designs et de contrastes), les annotations des gènes et des jeux de gènes testés, la description des deux cohortes étudiées, ainsi que les résultats des analyses LIMMA et quSAGE. Elle a nous avait permis notamment d'effectuer rapidement et facilement les croisements de listes $(A-B)\cap C$ à toutes les analyses temporelles de stimulation LPS, et de vérifier par exemple quels gènes associés ou liés étaient DE. Au vu de son utilité et de sa simplicité d'utilisation, et compte tenu du fait que les données servaient de base pour des validations fonctionnelles en cours, cette base de données était rendue accessible sur un serveur connecté en local pour le laboratoire. Il pourrait également être intéressant de rendre accessible plus facilement ces données par une application web proche de celle utilisée pour les données cliniques.

Les analyses de liaison

I/ La région 13q13

L'analyse de liaison non paramétrique sur une partie de la cohorte 1 avait permis d'identifier 13q13 comme nouvelle région liée significativement à la SpA. Après une analyse plus approfondie de cette région, nous avions pu définir plus précisément la région liée comme étant comprise entre 38,753 et 40,040 Mb, ce qui représente un intervalle d'environ 1 Mb. Parmi les quelques gènes présentant des informations connues sur leurs fonctions, UFM1 se révèle un bon candidat par son rôle dans le stress du RE, ou bien LHFP qui présente des polymorphismes associés au psoriasis.[224–226]

Cette région était déjà décrite auparavant comme liée avec un signal suggestif dans l'étude du GFEGS comprenant des microsatellites, ainsi que dans la méta-analyse de 2007 reprenant les trois études de liaison précédemment publiées sur la SpA.[83,84] Cette région était publiée en 2008 comme étant liée significativement à la maladie de Crohn, cette dernière présentant une fréquence plus élevée d'apparition chez les patients atteints de SpA.[227] Par ailleurs, une étude publiée en 2014 basée sur des variants de nombres de copies avait identifié la région 13q13.1 comme étant associée significativement à la SA dans une cohorte de sujets coréens.[228] Cependant, cette région est un peu éloignée, en amont de celle découverte par notre laboratoire. De plus, il faut noter que nous n'avions pas pu répliquer ces résultats avec la fusion de nos deux cohortes 1 et 2, et nous n'avions pas obtenu non plus de résultats intéressants lors des analyses paramétriques. Cependant, cela ne signifie pas forcément qu'il s'agit là d'un faux positif, mais permet surtout de se rendre compte de l'importance de prendre en compte le fait que nous étudions une maladie avec une très forte hétérogénéité. C'est pourquoi notre laboratoire continue tout de même l'étude de cette région par un séquençage plus précis. Tout comme pour la région SPA2, nous espérons grandement découvrir de nouveaux facteurs génétiques pouvant s'ajouter à ceux expliquant le développement de la SpA. Bien que les puces de génotypage permettent d'analyser à moindre coût et rapidement des polymorphismes tout le long du génome, ils possèdent l'inconvénient de ne pas recouvrir la totalité de la séquence génomique et peuvent donc passer à côté du SNP causal. C'est pourquoi le séquençage précis d'une région identifiée comme étant liée est la suite logique des analyses de liaison par puces de génotypage.

II/ L'analyse des haplotypes protecteurs

Lorsque nous effectuons une analyse de liaison chez des patients, nous recherchons les facteurs génétiques prédisposant au développement de la maladie. Dans le cas de la SpA, nous partons en général d'une première hypothèse, qui est la suivante : en dehors de la forte association connue avec HLA-B27, d'autres locus s'additionnent à ce facteur et rendent possible le développement de la SpA à différents degrés d'hétérogénéité. Dans ce cas, nous estimons que l'allèle HLA-B27 est nécessaire mais ne suffit pas au développement de la maladie. Cependant, il est intéressant aussi d'explorer une autre possibilité, qui est que l'allèle HLA-B27 suffit au développement de la SpA, mais que d'autres facteurs génétiques présents chez les témoins HLA-B27 positifs protègent de la maladie. C'est dans cette optique que nous avions entrepris une analyse de liaison différente sur nos trois cohortes fusionnées, en effectuant une inversion des statuts seulement chez les témoins HLA-B27 positifs. Ainsi, nous espérions trouver d'autres facteurs génétiques qui seraient présents spécifiquement chez les témoins HLA-B27 positifs et qui pourraient leur conférer une protection. Malheureusement, nous n'obtenions aucune région significativement liée. Cependant, une région à la jonction entre 22q12.2 et 22q12.3 semble avoir une liaison suggestive, et porterait donc des facteurs génétiques qui empêcheraient les individus HLA-B27 positifs de développer la SpA. Une seule étude cite cette région comme étant déterminante dans l'uvéite et partagée avec d'autres maladies autoimmunes.[229]

Dans cette région, peu de gènes possèdent des informations claires sur leurs fonctions, et seul un semble avoir un lien avec le système immunitaire : TIMP3, qui est un inhibiteur des matrices de métalloprotéinases, et notamment les collagénases. Il a été démontré que cette protéine régule les populations lymphocytaires du foie, et confère une protection contre l'hépatite auto-immune dirigée par les lymphocytes Th1.[230] Par ailleurs, dans des modèles murins, la perte de TIMP3 augmente l'inflammation et la polarisation pro-inflammatoire des macrophages dans l'athérosclérose, ainsi que la dégradation de l'agrécane et du collagène avec le vieillissement dans les articulations.[231,232] Cela conforte dans la possibilité d'un rôle protecteur de ce gène chez les individus sains porteurs de l'allèle HLA-B27. Bien que le signal ne soit que suggestif après une analyse de liaison dans cette région, il faut prendre en compte que la cohorte était fortement diminuée après la sélection des familles comptant au moins deux témoins HLA-B27 positifs génotypés. En effet, nous passions de 1310 à 551 individus génotypés, ce qui représentait encore

moins de personnes que lorsque nous avions identifié une liaison significative avec la région 13q13. Il serait donc intéressant d'étudier plus précisément cette région avec une cohorte sélectionnée spécifiquement dans l'optique d'étudier les haplotypes protecteurs.

III/ Comment explorer encore plus l'hétérogénéité de la SpA

Nous avions pu remarquer qu'en analysant les cohortes 1 et 2 fusionnés, nous n'arrivions pas à reproduire nos résultats obtenus avec une partie de la cohorte 1 seule, et notamment une liaison significative avec la région 13q13. Nous ne nous attendions pas forcément à cela, compte tenu du fait que nous conservions les familles liées à cette région pour cette deuxième analyse. Ce manque de recouvrement entre différentes études est une chose assez courante pour les analyses sur génome entier dans le cadre de la SpA, malgré un soin apporté sur la qualité statistique des analyses. Cela confère une part importante à l'hétérogénéité de la maladie, et donne une explication aux différentes formes de SpA qui peuvent apparaître en fonction de chaque individu, même au sein d'une même famille. Nous avions pu observer qu'une analyse de liaison intra-familiale permettait tout de même de détecter des polymorphismes causaux plus rares que lors d'une GWAS, ce qui prouve que l'efficacité de l'approche utilisée pour détecter les facteurs génétiques à l'origine de la SpA dépend fortement de la méthode d'analyse statistique.

Il serait d'ailleurs intéressant, lors des futures analyses, de prendre en compte l'effet de synergie pouvant apparaître entre plusieurs SNPs. En effet, cette interaction nommée également l'épistasie, est un phénomène génétique qui part du principe que l'effet d'un gène dépend d'un ou plusieurs autres gènes, constituant le fond génétique global. Ainsi, de la même manière que lors de l'approche par réseaux de gènes, on ne considère pas un SNP comme une entité seule, mais plutôt comme un élément d'un système constitué de plusieurs polymorphismes agissant indirectement entre eux. Concrètement, cela se manifeste par un effet de la présence de plusieurs allèles en même temps qui ont un effet différent de ce qui serait attendu s'ils étaient indépendants. L'épistasie peut apparaître entre des mutations situées sur plusieurs gènes, ou bien sur un seul même gène. Dans ce dernier cas, on parle de complémentation intragénique. Les effets observés sur la maladie pourraient être synergistiques ou antagonistiques. Une telle analyse serait particulièrement adaptée dans le cas de l'étude de la SpA, puisque nous observons une forte hétérogénéité au niveau des facteurs liés ou associés. Il est donc probable que les SNPs identifiés

n'ont pas un effet seul sur la maladie, mais que ce soit plutôt la combinaison de plusieurs polymorphismes qui entraînent le développement de ce RIC. Des outils permettent d'analyser l'épistasie entre les polymorphismes issus de GWAS avec des techniques de permutation, ou bien des modèles linéaires mixtes (comme FaST-LMM-Select).[233]

Maintenant que nous avons à disposition un certain nombre de locus en dehors du CMH qui montrent une association ou une liaison significative à la SpA, il serait intéressant d'utiliser des modèles statistiques plus fins qui prendraient en compte qu'une part de l'héritabilité est expliquée par les locus connus. Il est aussi important, dans ce cas, de prendre en compte qu'une majorité des facteurs héréditaires est écrasée par l'association avec HLA-B27. C'est pourquoi les analyses paramétriques sont importantes une fois que l'analyse non paramétrique détecte une région significativement liée. D'autre part, les nouvelles technologies de séquençage, ou autrement appelées les next-generation sequencing (NGS), permettent maintenant d'avoir un meilleur recouvrement et une meilleure qualité des données concernant les SNPs. C'est pourquoi dans notre laboratoire, nous entreprenons une stratégie de séquençage à haut débit sur l'exome, les données étant actuellement en cours de génération. Il est aussi important de pouvoir augmenter davantage la cohorte d'individus génotypés, de la même manière que pour l'étude d'association de l'IGAS.[22] Bien entendu, cela demande un effort supplémentaire par rapport à une GWAS, car il est primordial de choisir seulement des formes familiales pour les analyses de liaison. Enfin, il est intéressant de pouvoir directement lier les informations de génotypage des invididus avec d'autres informations les concernant directement, comme les données transcriptomiques pour une stratégie eQTL, ou bien avec le microbiote pour pouvoir associer facteurs génétiques et environnementaux. C'est dans cette optique que notre laboratoire a entrepris une stratégie de séquençage RNAseq chez une cohorte également génotypée, et une collaboration avec la plate-forme MICALIS localisée à l'INRA de Jouy en Josas pour étudier le microbiote intestinal.

Les analyses du transcriptome

I/ Utilité de la méta-analyse

En plus de nous conforter dans le fait de pouvoir fusionner nos deux cohortes pour les analyses transcriptomiques, notre méta-analyse avec l'outil MetaQC avait permis de tester la cohérence de nos données avec d'autres lignées cellulaires. Nous pouvions clairement définir que nos données

transcriptomiques étaient suffisamment proches d'autres études incluant des MD-DCs ou des DCs inflammatoires. Ces dernières cellules correspondent à des populations de cellules dendritiques extraites dans des fluides inflammatoires humains, comme définies dans une publication de l'équipe d'Amigorena S.[204] Nous nous étions rendus compte du potentiel de ce type d'analyse, qui pouvait aller bien au-delà d'un simple contrôle qualité de nos données. En effet, il pourrait être intéressant de déterminer si nos études sont proches d'autres incluant des stimulations par du LPS ou d'autres substances induisant une réponse inflammatoire non spécifiques, comme la flagelline en extra-cellulaire ou TLR-3 en intra-cellulaire.[234–236] En effet, le LPS, bien qu'il imite une paroi bactérienne, est aussi discutée dans le fait qu'il représente toute situation d'une cellule de l'immunité en cas d'inflammation. De plus, des hypothèses portent à croire qu'une infection virale pourrait être aussi à l'origine de l'inflammation dans la SpA.[237,238] Par ailleurs, il serait intéressant de comparer nos données transcriptomiques avec d'autres études concernant la SpA ou des maladies inflammatoires plus ou moins proches, afin d'effectuer une cartographie de chacune de ces maladies. Nous pourrions par exemple comparer l'expression des gènes dans la SpA avec la maladie de Crohn, le psoriasis ou la PR.[201,239–242] Cependant, bien que cette approche se révèle simple dans la mesure où il suffit de normaliser séparément les données et de déterminer les contrastes les plus discriminants dans les échantillons, il faut faire attention à prendre en compte que toutes ces maladies ne portent pas forcément autant d'attention sur les cellules dendritiques. Ce projet mérite donc une attention particulière sur les études à inclure et sur l'homogénéité des données étudiées. Une telle méta-analyse permettrait de définir de manière encore plus précise et cohérente le design expérimental des futures études transcriptomiques.

II/ Gènes DE associés à la SpA et croisements de liste

II.1) Bootstrap et permutation comme qualités statistiques supplémentaires

Lors de la génération de nos listes A et C de gènes DE par LIMMA, qui représentaient respectivement les gènes affectés par la maladie ou HLA-B27 et par la maladie spécifiquement, nous ne pouvions pas utiliser les p-values ajustées de Benjamini et Hochberg afin de filtrer les gènes significativement DE, car aucune ou très peu obtenaient une valeur inférieure à 5 %. Or, une correction sur test multiple aurait été l'idéal pour pouvoir affirmer avec plus de sûreté que nous n'avions pas affaire à des faux positifs. C'est pourquoi nous avions préféré nous conforter

plus sur nos résultats statistiques grâce à une approche de bootstrap sur nos données. Au vu de la forte corrélation entre les moyennes des p-values bootstrappées et celles issues des données réelles, nous pouvions supposer que nos modèles linéaires utilisés pour LIMMA représentaient bien la variation observée au niveau des expressions des gènes. La méthode du bootstrap a été introduite par Elfron B. en 1979, et est utilisée de plus en plus dans la sélection de modèles dans le cadre d'analyses statistiques de données biologiques.[243–247] Cette méthode permet de simuler une distribution nulle, à l'instar des méthodes de Monte-Carlo, et possède l'avantage de ne pas nécessiter d'autres informations supplémentaires que celle fournie par l'échantillon initial. Cependant, bien qu'encore très peu étudié, nous pourrions discuter sur un biais possible par le fait que nous ne nous basons que sur l'échantillon initial, sans connaissance a priori de la distribution. Récemment, il a été proposé d'utiliser une méthode proche, qui consiste à sous-échantillonner l'échantillon initial, à la place d'une autre méthode de machine learning nommée la forêt d'arbres décisionnels.[248] Ce sous-échantillonnage consiste à tirer au sort sans remise un nombre d'observations inférieur à la taille totale de l'échantillon, et d'utiliser la même approche par la suite que le bootstrap pour comparer les résultats.[249] Il pourrait être intéressant d'utiliser cette approche sur nos données et de comparer ces résultats avec le bootstrap présenté dans notre analyse.

Par ailleurs, nous voulions effectuer un filtre supplémentaire sur nos gènes DE qui permettraient de retirer des faux positifs, mais de façon moins sévère que la correction sur tests multiples. Pour cela, nous avions inféré des distributions empiriques de nos statistiques par une technique de permutation, de manière très proche de notre technique de bootstrap. La principale différence ici est que nous calculions une p-value empirique au lieu d'un facteur de corrélation, qui se basait sur l'hypothèse nulle que nous obtiendrions la même statistique t par LIMMA si nous avions mélangé au hasard les facteurs attribués à chaque échantillon. Cela nous avait permis tout de même de retirer en moyenne 1/5 des gènes initialement identifiés comme DE par LIMMA. Cependant, il faut tout de même être conscient que nous ne sommes pas aussi reproductibles sur le retrait des faux positifs qu'avec la méthode de Benjamini et Hochberg. C'est pourquoi des validations par qPCR sont en cours pour vérifier l'expression différentielle dans notre même cohorte de gènes identifiés comme DE par notre analyse de puces transcriptomiques. Ces gènes sont FADS2, INSIG1, LDLR, PCOLCE2, RPL30, SCAP, SIRT6, SQLE, SREBF2, MSMO1, PLIN2, CLEC4D, ABCA1 et MIR33. Les gènes de ménages utilisés sont GAPDH, ACTB et HPRT1. De plus, la

même cohorte est actuellement en train d'être séquencée par RNAseq et nous pourrons bientôt confirmer par cette technologie si nos résultats sont valables. Enfin, la même approche de séquençage à haut débit des transcrits par RNAseq est sur le point d'être effectuée sur une cohorte d'une centaine de personnes, indépendante des deux présentées dans ce manuscrit. Nous espérons que toutes ces expériences permettront non seulement de confirmer nos résultats malgré la limite statistique, mais en plus de découvrir éventuellement d'autres voies de signalisation ou gènes clés permettant d'expliquer l'étiologie de la SpA.

II.2) Les croisements de listes (A-B)∩C et autres possibilités

En outre le filtrage supplémentaire de nos gènes DE par les calculs de p-values empiriques après permutation, nous avions utilisé les différents contrastes étudiés générant les listes A (gènes affectés par la SpA et/ou HLA-B27), B (affectés par HLA-B27 seulement) et C (affectés par la SpA seulement), afin d'effectuer un croisement de liste comme moyen de réplication directement à partir de nos données. Nous partions du principe que retirer les gènes de la liste B à la liste A permettrait de ne sélectionner que les gènes affectés par la SpA, puis que de croiser cette nouvelle liste avec la liste C servirait de confirmation. Nous avions ainsi pu garder des nombres de gènes suffisants pour développer des hypothèses biologiques, tout en étant suffisamment petits pour estimer que nous retirions un nombre suffisant de faux positifs.

En dehors de la sur-expression de PCOLCE2 qui présente un intérêt par son implication dans les voies de signalisation liées à l'ossification, il est rassurant de retrouver des gènes DE qui étaient soit identifiés comme DE par la cohorte de l'étude 1, soit localisés dans des régions associées ou liées à la SpA.[22,82–84,89,197] Même si notre étude était une extension de la première, et que les gènes en commun étaient retrouvés comme DE à des analyses temporelles plus spécifiques, il n'était pas si évident de retrouver presque 1/5 des gènes publiés comme DE dans notre croisement de listes (A-B)∩C, qui de plus variaient dans le même sens chez les patients. Nous retrouvions particulièrement ADAMTS15, qui avait fait l'objet d'une étude plus poussée de corrélation d'expression avec CITED2 par qPCR. De plus, il s'agit d'une MMP qui a été démontrée comme sous-exprimée chez les patients atteints d'OA.[250] Il est donc probable que cette enzyme soit impliquée dans le renouvellement du cartilage et/ou de l'os au cours de l'inflammation induite par la SpA. Aucune implication d'ADAMTS15 n'a cependant été mise en évidence directement avec

la SpA, mise à part l'expression différentielle publiée par l'analyse de l'étude 1.[197] Les 40 gènes DE localisés dans des régions associées ou liées à la SpA donnent un premier lien direct entre la génomique et l'aspect fonctionnel des gènes. Cependant, nous sommes conscients que nous avions ratissé large sans a priori au niveau des gènes testés, puisqu'ils représentaient au total 713 gènes. Cela veut dire que nous n'avions finalement trouvé que 5,61 % de ces gènes comme étant DE. D'autre part, ces 40 gènes représentaient seulement 4,06 % de tous les gènes identifiés comme DE dans nos croisements de listes aux différentes analyses temporelles. Il faut donc noter que la probabilité que ces gènes ne sont pas forcément affectés par les liaisons ou associations génétiques n'est pas négligeable (p-value égale à 37,56 % lors d'un test exact de Fisher). Cependant, ces résultats démontrent tout de même que nos résultats d'analyses différentielles font sens avec la pathologie étudiée, et qu'il serait intéressant d'effectuer une réelle stratégie eQTL. D'autre part, il pourrait être intéressant de tester si des SNPs associés ou liés avec la SpA peuvent présenter des effets trans-eQTLs sur des gènes DE grâce à des bases de données publiques regroupant ces informations.[251–253] Quoi qu'il en soit, l'approche eQTL directe grâce aux résultats NGS permettront une analyse plus fine du lien entre les polymorphismes et l'expression des gènes.

Par ailleurs, de la même manière que pour l'analyse des haplotypes protecteurs, nous pourrions aussi étudier les gènes qui confèreraient une protection aux individus sains porteurs de l'allèle HLA-B27. Ce projet était le sujet d'un stage en Master 2 dans notre laboratoire, et nous avions ainsi effectué un autre croisement de listes : (B-A)∩C, avec les variations des gènes entre B et C devant être opposées. De plus, le fait que la liste B donnait des gènes avec des p-values ajustées très significatives a permis de filtrer davantage les gènes DE dans ce contexte. Les résultats, bien que non présentés ici, donnaient des choses très intéressantes, comme l'implication de la voie de signalisation PI3K/AKT et la détection de plusieurs gènes associés à la SpA comme IL12B et TNFSF15. Enfin, nous avions également comparé les 13 paires de germains entre patients et témoins HLA-B27 positifs à toutes les analyses temporelles de stimulation LPS. Cette liste de gènes DE, encore non étudiée en détail, pourrait également servir à identifier des gènes DE affectés spécifiquement par des formes familiales de SpA, car les personnes sont issues de familles multiplex. Par ailleurs, il permettrait de retirer les effets génétiques et d'analyser plus en détail des facteurs environnementaux.

III/ Jeux de gènes enrichis associés à la SpA

Même si l'analyse d'enrichissement des jeux de gènes associés à la SpA n'avait rien donné de spécifique au niveau des voies de signalisation, elle avait permis de guider sur les directions à suivre et donné quelques informations supplémentaires. Notamment, nous pouvions déjà avoir un indice sur le fait d'explorer les voies de biosynthèse du cholestérol, puisque plusieurs jeux de gènes liés à ce mécanisme se retrouvaient enrichis, et ce de manière conservée à plusieurs temps de stimulation LPS. Nous avions pu remarquer également que les MD-DCs réagissaient différemment après 6 heures de stimulation chez les patients par rapport aux témoins au niveau des voies de signalisation liées à l'activité des récepteurs des neurones. Cela n'est pas si surprenant, car nous comparons souvent ces deux cellules au niveau de leurs activités synaptiques. Ces résultats pourraient traduire donc une dérégulation au niveau de l'activité des DCs dans leurs interactions avec d'autres lymphocytes, et donc induire une réponse inflammatoire non contrôlée.

Cependant, le faible recouvrement des gènes DE dans les jeux de gènes enrichis et la forte redondance de ceux-ci au niveau de leurs contenus en gènes ne nous permettait pas d'aller plus loin dans les hypothèses pouvant expliquer la physiopathologie des gènes DE identifiés. En effet, il était difficile de déterminer précisément quelle voie de signalisation était la plus affectée chez les patients. Il semblerait plutôt qu'il fallait approcher le problème autrement, en considérant que ce n'était pas une seule voie de signalisation entière qui était affectée, mais plutôt un ensemble de gènes liés à différentes voies qui, en interagissant entre eux de manière directe ou indirecte, influerait sur le fonctionnement global des DCs à réguler une réponse auto-immune. Nous pouvions donc constater la limite de l'approche GSEA dans l'étude de maladies complexes et multifactorielles comme la SpA. C'est pour cela que nous préférions nous orienter vers une approche avec moins de supervision sur les fonctions des gènes et avec plus de précisions dans l'étude des interactions gène-gène : la théorie des graphes et les réseaux de gènes. En ce qui concerne la redondance des jeux de gènes, il pourrait être intéressant d'utiliser une approche différente qui permettrait de regrouper les jeux de gènes redondants et de construire par hiérarchisation un jeu de gène spécifiquement lié à la SpA. Cette approche pourrait être intéressante, bien qu'il mériterait un travail considérable sur le fait de prendre en compte le croisement de listes et les différentes analyses temporelles pour pouvoir résumer toutes ces voies enrichies.

IV/ Approche par réseaux de gènes

IV.1) La théorie des graphes

La théorie des graphes appliquée dans les données biologiques est une méthode de plus en plus utilisée en application à l'analyse de données d'expérience à haut débit du transcriptome, et cette approche commence à être indispensable et à remplacer la GSEA pour pouvoir définir des fonctions biologiques à l'origine du phénotype observé. Il existe de multiples façons de construire des réseaux de gènes, mais nous avons voulu nous concentrer pour ce projet sur les réseaux de régulation, et particulièrement sur l'inférence des réseaux gaussiens à l'aide des corrélations entre les expressions des gènes. Nous avons également voulu comparer nos réseaux expérimentaux avec les interactions connues dans la littérature à l'aide de la base de données STRING et son API. Cette approche mixte permettait d'utiliser deux approches complémentaires, qui consistaient à observer des réseaux de gènes sémantiques et gaussiens, et d'observer laquelle nous donnait le plus d'informations.

Nous avons pu remarquer que l'inférence des réseaux gaussiens nous donnait des informations particulièrement pertinentes, notamment sur l'implication des voies de biosynthèse du cholestérol, et ce grâce à la comparaison des réseaux inférés chez les patients ou les témoins. Cela nous permettait, au-delà de l'analyse différentielle, de vérifier la présence de gènes différentiellement connectés, et plus précisément des gènes qui seraient plus impliqués dans des régulations inter-géniques chez les patients que chez les témoins. Cette approche d'étude des perturbations de réseaux induites spécifiquement par une maladie était déjà intégrée par WGCNA, ainsi que d'autres packages R.[169,254–257] Notre approche de comparaison de connectivités utilisée dans notre projet était encore simple, compte tenu du fait que nous appliquions cela de façon exploratoire. De plus, nous ne pouvions pas utiliser des outils classiques de calculs de différences de connectivités sur tous nos gènes, à cause de la complexité du design de nos analyses multivariées et de l'application de nos croisements de listes. Cependant, il serait intéressant d'effectuer des tests statistiques pour comparer si la différence de connectivité est bien significative. De même que pour LIMMA, il serait intéressant d'effectuer un bootstrap et/ou une permutation sur des résultats aléatoires de ces tests statistiques.

Par ailleurs, nous avions des réseaux inférés après différents temps de stimulations, et donc des

graphes d'états stables spécifiques des patients et témoins. Nous avions également inféré des graphes de transition en étudiant l'effet de MSMO1 à H0 sur les temps de stimulation LPS suivants. Il est donc faisable d'utiliser ces réseaux sous forme de variables discrètes ou logiques pour les interactions, et d'utiliser des algorithmes permettant d'effectuer des simulations sur le réseau pour détecter quels gènes pourraient être ciblés pour une analyse fonctionnelle. Nous pourrions par exemple utiliser la théorie de René Thomas, ou bien le logiciel GinSIM, qui sont spécialement conçus dans cette optique.[144,258] Il n'est pas nécessaire d'utiliser des variables continues pour les interactions et des équations différentielles ordinaires, car les temps de stimulation étudiés sont trop éloignés entre eux. En effet, cette méthode est surtout conseillée pour des études portées plus sur la cinétique des expressions des gènes. Tout comme pour le calcul de différences de connectivités, l'utilisation de réseaux logiques pour simuler l'effet d'une induction ou non de l'expression d'un gène sur l'ensemble des autres gènes doit se faire avec précaution, d'autant plus que les gènes que nous étudions interagissent très certainement de manière indirecte.

IV.2) Utilité de STRING

Dans le cadre de notre projet, nous avions utilisé STRING principalement pour comparer nos réseaux inférés par SIMoNe avec les interactions déjà connues, ainsi que dans un objectif d'exploration des mécanismes pouvant lier les différents gènes DE identifiés. Il est possible d'attribuer par exemple une importance particulière à un gène plus connecté. En effet, les gènes centraux ont une plus grande probabilité de jouer un rôle central dans le phénotype observé, et inversement. Cependant, dans la pratique, comme dans notre projet, il est rare d'identifier tous les gènes causaux à l'origine du réseau parmi tous ceux qui sont DE. Nous pouvions ajouter des gènes intermédiaires, mais sans l'exacte certitude que c'étaient bien ceux-ci qui interagissaient dans notre réseau. De plus, l'ajout de ces gènes, bien que potentiellement informatifs, rendaient les réseaux très complexes à analyser lorsque nous dépassions les 100 gènes DE. C'est pourquoi nous utilisions des réseaux condensés en transformant les scores en distances et en calculant les chemins les plus courts entre chaque gène DE. Ceci réduisait considérablement les informations visibles sur notre réseau, tout en ne perdant pas les informations sur les gènes intermédiaires grâce aux attributs des arcs. Quoi qu'il en soit, nous avions tout de même des difficultés à distinguer des voies de signalisations spécifiquement affectés par la SpA avec cette approche, comme nous

pouvions nous y attendre avec les résultats de l'approche GSEA. Il serait intéressant de l'utiliser comme moyen de supervision automatisée sur nos réseaux inférés par SIMoNe, de la même manière que le logiciel Aracne[171], afin de prioriser des voies différentiellement connectées déjà connues dans la littérature. Contrairement à Aracne, nous aurions déjà retiré des interactions indirectes par une approche non supervisée : SIMoNe.

IV.3) L'approche SIMoNe (comparaison avec WGCNA)

Nous avions également voulu, à travers ce projet, essayer l'outil SIMoNe dans l'inférence de réseaux gaussiens à partir de nos données d'expressions. En effet, nous avions profité de notre collaboration avec le Laboratoire de Mathématiques et Modélisation d'Evry pour tester leur package R sur des données biologiques à des fins d'exploration d'interactions dans un réseau de gènes, ce qui n'avait pour l'instant jamais été effectué. C'est pour cela que nous avions comparé nos résultats avec une inférence plus classique proche de WGCNA, et nous étions confortés par le fait que SIMoNe pouvait détecter avec une certaine sensibilité presque tous les arcs inférés par WGCNA. Les quelques arcs qui étaient spécifiques de WGCNA étaient justement principalement des corrélations fortes dues à des interactions indirectes au sein du réseau même. Cependant, l'approche SIMoNe pouvait également donner des arcs avec des faibles coefficients de corrélations de Spearman, ce qui était difficile à interpréter biologiquement. Nous avions donc utilisé une méthode mixte, et adopté SIMoNe dans une optique de guide des interactions indirectes à retirer, afin de filtrer a posteriori sur les valeurs absolues des rhos de Spearman devant être supérieures à 0,4. Nous étions satisfaits de cette approche, qui avait pour principal avantage de ne pas se limiter sur des interactions déjà connues pour retirer le bruit d'informations des corrélations indirectes, comme le ferait une approche supervisée plus classique.

IV.4) L'implication des voies de biosynthèse du cholestérol

L'analyse des gènes DE, des jeux de gènes enrichis, et surtout des réseaux de gènes, convergent toutes vers une conclusion commune : la voie de biosynthèse du cholestérol est affectée dans les DCs issus de patients atteints de SpA. Cette découverte est novatrice, car aucun lien fonctionnel au niveau des gènes n'avait été élucidé avec cette pathologie. La littérature montre quelques études portant sur le profil lipidique affecté et les risques cardio-vasculaires présents chez les patients atteints de SpA.[259–262] Cependant, aucune ne portait sur le taux de cholestérol

intracellulaire des DCs, ce qui ne donne pas forcément un lien direct avec nos résultats. L'hypothèse que nous pourrions développer est que la dérégulation des voies de biosynthèse du cholestérol induirait un stress du RE, et donc une activation de la voie UPR. Parmi les régulateurs suspectés de nos gènes cibles se trouve MIR33, c'est pourquoi ce miRNA fait partie des gènes actuellement en cours de validation par qPCR. De plus, MIR33 ne fait pas partie des gènes inclus dans les puces Affymetrix que nous avions étudié. Il est donc possible que le transcrit soit altéré sans que nous ayons pu le détecter. Nos gènes DE impliqués dans cette voie de signalisation, qui sont PLIN2, INSIG1, LDLR, MSMO1, SQLE, SREBF2 et APOA4, ont tous un lien très fort, que ce soit par une implication dans la même voie enzymatique ou des mécanismes de régulation communs. MSMO1 semble jouer un rôle central dans nos réseaux de gènes. Des mutations dans ce locus ont été détectées comme pouvant causer une forme de psoriasis.[263] Quoi qu'il en soit, ces résultats méritent d'être explorés plus en profondeur, notamment par une analyse du profil lipidique de ces cellules actuellement en cours par une technique de spectrométrie de masse.

IV.5) Le package stringgaussnet

Lorsque nous voulions inférer nos réseaux gaussiens et les comparer avec des réseaux sémantiques à partir de nos résultats de gènes DE, nous étions confrontés au fait qu'il n'existait pas encore d'outil spécialisé dans la construction de ce type de réseau. Par conséquent, nous avions créé nos propres scripts R permettant d'inférer automatiquement un nombre important de réseaux à partir de plusieurs listes de gènes. De plus, au vu de la quantité importante de réseaux à visualiser dans Cytoscape, nous avions créé un système d'importation automatique à travers des commandes envoyées à l'API local généré par le plugin cyREST. Nous avions ainsi un système stable et ergonomique permettant d'étudier rapidement les interactions entre nos gènes listés.

Comme nous avions pu le remarquer, les approches d'analyses différentielles des gènes et GSEA ne suffisent pas forcément à définir une voie de signalisation précise à l'origine d'un phénomène biologique observé. L'application de la théorie des graphes est maintenant devenu une étape indispensable avant les validations et études fonctionnelles. Cependant, cette notion n'est pas forcément très connue de la plupart des biologistes ayant effectué des expériences transcriptomiques à haut débit. C'est pourquoi nous avions voulu assembler nos codes R permettant la génération et la visualisation de nos réseaux en un seul package : stringgaussnet. La

construction des réseaux sémantiques et gaussiens sont deux approches complémentaires et apportent des informations différentes. A ce jour, aucun outil ne permettait de construire de façon simple et didactique ces deux types de réseaux à partir de listes de gènes pré-définis. C'est dans cette optique que nous avions développé ce package R. Nous espérons que cet outil permettra aux biologistes d'explorer plus en profondeur et plus simplement les interactions possibles entre les gènes DE grâce à la théorie des graphes.

Des mises à jour pourront être effectuées par la suite pour améliorer l'expérience utilisateur et la fiabilité des calculs. Nous pourrons par exemple faire en sorte que le calcul de score combiné soit le dernier utilisé par STRING, c'est-à-dire dans la version 10. En effet, nous n'incluions que la méthode de la version 8.1 par souci de non divulgation de l'algorithme précis permettant ce calcul. D'autre part, nous pourrions ajouter la possibilité de retirer des gènes ou d'en inclure d'autres, sans avoir à refaire toutes les étapes de construction des réseaux. Enfin, nous pourrions inclure un système de tests statistiques pour définir de façon précise et objective les différences de connectivité entre deux groupes d'échantillons.

Chapitre 5

CONCLUSIONS

La SpA est une maladie multifactorielle, hétérogène, avec une forte composante génétique, et qui présente plusieurs sous-types pouvant se regrouper dans une même famille. Il est actuellement difficile d'expliquer toutes les voies de signalisation affectées chez les patients et quelle est l'étiologie de la maladie. En dehors de sa forte association avec HLA-B27, nous ne savons finalement que peu de choses sur sa physiopathologie. Les projets menés durant ces trois années de thèse m'ont permis d'apporter de nouvelles pistes sur la compréhension de la SpA et les voies de signalisation à explorer, en combinant à la fois une approche d'analyse génétique et transcriptomique.

Les analyses des facteurs génétiques sur la cohorte 1, et le soin apporté sur les flux de travail concernant les nettoyages et les analyses de liaison des données de génotypage, ont permis d'identifier une nouvelle région, 13q13, comme étant significativement liée à la SpA. Cette liaison a fait l'objet de la publication d'un article, et d'un séquençage plus précis dans les familles les plus liées. Par ailleurs, un séquençage par technologie NGS de l'exome est en cours sur une cohorte conséquente, et les données pourront servir à répliquer nos résultats et/ou à poursuivre l'exploration d'autres facteurs génétiques. Aucun lien direct n'a pu être mis en évidence entre la région 13q13 et les gènes DE détectés par l'analyse de notre laboratoire. Cependant, nous pensons que des mécanismes plus indirects peuvent être à l'origine d'une interaction entre celle-ci et d'autres facteurs de transcription.

Afin d'utiliser une approche complémentaire de l'analyse du génotype des individus, nous avions entrepris une analyse du transcriptome. Elle représente la première étude des transcrits incluant des cellules dendritiques et des témoins porteurs de l'allèle HLA-B27. En plus de la découverte de nouveaux gènes différentiellement exprimés chez les patients, j'ai poursuivi par une analyse en réseaux de gènes, en construisant des réseaux sémantiques et gaussiens. Cette méthode, de plus en plus utilisée dans la biologie des systèmes, m'a permis de mettre la lumière sur l'implication des phases précoces de la voie de biosynthèse du cholestérol dans la physiopathologie de la SpA. Plusieurs gènes sont en cours de validation par qPCR chez les mêmes patients, et le profil lipidique est en cours d'analyse. Par ailleurs, des données RNA-seq issues d'une autre cohorte plus importante sont sur le point d'être générées, et serviront à la fois de réplication de mes résultats, et d'outils d'explorations plus avancées des changements d'expressions des gènes pouvant être à

l'origine du développement de la SpA.

Mes travaux ont donc permis d'aller encore plus loin dans les hypothèses sur les mécanismes physiopathologies de la SpA, et font l'objet de travaux supplémentaires au sein même du laboratoire. Ce RIC est difficile à diagnostiquer, et aucun traitement actuel ne permet de soigner directement les patients. Même si elle n'est pas directement mortelle, cette maladie est particulièrement handicapante et douloureuse pour les patients. Nous espérons que ces résultats, en plus des travaux qui suivront, permettront de proposer des nouvelles méthodes de diagnostique et de traitement pour améliorer définitivement la vie des patients.

Enfin, j'ai pris l'initiative de développer et de rendre accessible un package R regroupant les méthodes m'ayant permis de créer les deux types de réseaux de gènes utilisés pour ce projet de thèse. Nous espérons également que cet outil permettra aux biologistes d'utiliser plus facilement la théorie des graphes dans le cas d'analyse de réseaux de gènes.

Sixième partie

BIBLIOGRAPHIE

1. Wright, V. Seronegative polyarthritis: a unified concept. *Arthritis Rheum.* **21,** 619–633 (1978).
2. Garrett, S. *et al.* A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *J. Rheumatol.* **21,** 2286–2291 (1994).
3. Dougados, M. *et al.* La spondylarthrite en 100 questions. (2005).
4. Sieper, J. *et al.* New criteria for inflammatory back pain in patients with chronic back pain: a real patient exercise by experts from the Assessment of SpondyloArthritis international Society (ASAS). *Ann. Rheum. Dis.* **68,** 784–788 (2009).
5. Stolwijk, C., Boonen, A., van Tubergen, A. & Reveille, J. D. Epidemiology of spondyloarthritis. *Rheum. Dis. Clin. North Am.* **38,** 441–476 (2012).
6. Costantino, F. *et al.* Prevalence of spondyloarthritis in reference to HLA-B27 in the French population: results of the GAZEL cohort. *Ann. Rheum. Dis.* **74,** 689–693 (2015).
7. Kellgren, J. H. Diagnostic criteria for population studies. *Bull. Rheum. Dis.* **13,** 291–292 (1962).
8. Kellgren, J. H. Population Studies of the Rheumatic Diseases. *Ann. Rheum. Dis.* **28,** 63 (1969).
9. Van der Linden, S., Valkenburg, H. A. & Cats, A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis Rheum.* **27,** 361–368 (1984).
10. Amor, B., Dougados, M. & Mijiyawa, M. [Criteria of the classification of spondylarthropathies]. *Rev. Rhum. Mal. Ostéo-Articul.* **57,** 85–89 (1990).
11. Dougados, M. *et al.* The European Spondylarthropathy Study Group preliminary criteria for the classification of spondylarthropathy. *Arthritis Rheum.* **34,** 1218–1227 (1991).
12. Rudwaleit, M. *et al.* The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part II): validation and final selection. *Ann. Rheum. Dis.* **68,** 777–783 (2009).

13. Rudwaleit, M. *et al.* The Assessment of SpondyloArthritis international Society classification criteria for peripheral spondyloarthritis and for spondyloarthritis in general. *Ann. Rheum. Dis.* **70,** 25–31 (2011).
14. Kroon, F., Landewé, R., Dougados, M. & van der Heijde, D. Continuous NSAID use reverts the effects of inflammation on radiographic progression in patients with ankylosing spondylitis. *Ann. Rheum. Dis.* **71,** 1623–1629 (2012).
15. Balfour, J. A., Fitton, A. & Barradell, L. B. Lornoxicam. A review of its pharmacology and therapeutic potential in the management of painful and inflammatory conditions. *Drugs* **51,** 639–657 (1996).
16. Maxwell, L. J. *et al.* TNF-alpha inhibitors for ankylosing spondylitis. *Cochrane Database Syst. Rev.* **4,** CD005468 (2015).
17. Van den Bosch, F. & Deodhar, A. Treatment of spondyloarthritis beyond TNF-alpha blockade. *Best Pract. Res. Clin. Rheumatol.* **28,** 819–827 (2014).
18. Baeten, D. *et al.* Anti-interleukin-17A monoclonal antibody secukinumab in treatment of ankylosing spondylitis: a randomised, double-blind, placebo-controlled trial. *Lancet* **382,** 1705–1713 (2013).
19. Dagfinrud, H., Kvien, T. K. & Hagen, K. B. Physiotherapy interventions for ankylosing spondylitis. *Cochrane Database Syst. Rev.* CD002822 (2008). doi:10.1002/14651858.CD002822.pub3
20. Brown, M. A. *et al.* Susceptibility to ankylosing spondylitis in twins: the role of genes, HLA, and the environment. *Arthritis Rheum.* **40,** 1823–1828 (1997).
21. O'Rielly, D. D. & Rahman, P. Advances in the genetics of spondyloarthritis and clinical implications. *Curr. Rheumatol. Rep.* **15,** 347 (2013).
22. International Genetics of Ankylosing Spondylitis Consortium (IGAS) *et al.* Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-

related loci. *Nat. Genet.* **45,** 730–738 (2013).

23. Gaboriau-Routhiau, V. *et al.* The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses. *Immunity* **31,** 677–689 (2009).
24. Suenaert, P. *et al.* Anti-tumor necrosis factor treatment restores the gut barrier in Crohn's disease. *Am. J. Gastroenterol.* **97,** 2000–2004 (2002).
25. Guarner, F. & Malagelada, J.-R. Role of bacteria in experimental colitis. *Best Pract. Res. Clin. Gastroenterol.* **17,** 793–804 (2003).
26. Jacques, P. & Elewaut, D. Joint expedition: linking gut inflammation to arthritis. *Mucosal Immunol.* **1,** 364–371 (2008).
27. Hill Gaston, J. S. & Lillicrap, M. S. Arthritis associated with enteric infection. *Best Pract. Res. Clin. Rheumatol.* **17,** 219–239 (2003).
28. Zeboulon-Ktorza, N. *et al.* Influence of environmental factors on disease activity in spondyloarthritis: a prospective cohort study. *J. Rheumatol.* **40,** 469–475 (2013).
29. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35,** D5–D12 (2007).
30. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13,** 101 (2011).
31. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5,** 16–18 (2008).
32. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).
33. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **4,** 587–597 (2003).
34. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).

35. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30,** 97–101 (2002).
36. Hosking, L. *et al.* Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur. J. Hum. Genet. EJHG* **12,** 395–399 (2004).
37. Tiret, L. & Cambien, F. Departure from Hardy-Weinberg equilibrium should be systematically tested in studies of association between genetic markers and disease. *Circulation* **92,** 3364–3365 (1995).
38. Hardy, G. H. Mendelian proportions in a mixed population. 1908. *Yale J. Biol. Med.* **76,** 79–80 (2003).
39. Schutte, B. C. & Murray, J. C. Current Protocols in Human Genetics. *Am. J. Hum. Genet.* **57,** 735–736 (1995).
40. Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70,** 425–434 (2002).
41. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52,** 506–516 (1993).
42. Hutchinson, J. B. The Application of the ‘Method of Maximum Likelihood’ to the Estimation of Linkage. *Genetics* **14,** 519–537 (1929).
43. Morton, N. E. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7,** 277–318 (1955).
44. Dawn Teare, M. & Barrett, J. H. Genetic linkage studies. *Lancet* **366,** 1036–1044 (2005).
45. Hodge, S. E., Abreu, P. C. & Greenberg, D. A. Magnitude of type I error when single-locus linkage analysis is maximized over models: a simulation study. *Am. J. Hum. Genet.* **60,** 217–227 (1997).

46. Greenberg, D. A., Abreu, P. & Hodge, S. E. The power to detect linkage in complex disease by means of simple LOD-score analyses. *Am. J. Hum. Genet.* **63,** 870–879 (1998).
47. Ott, J. Linkage analysis and family classification under heterogeneity. *Ann. Hum. Genet.* **47,** 311–320 (1983).
48. Abreu, P. C., Hodge, S. E. & Greenberg, D. A. Quantification of type I error probabilities for heterogeneity LOD scores. *Genet. Epidemiol.* **22,** 156–169 (2002).
49. Brown, M. A., Laval, S. H., Brophy, S. & Calin, A. Recurrence risk modelling of the genetic susceptibility to ankylosing spondylitis. *Ann. Rheum. Dis.* **59,** 883–886 (2000).
50. Dernis, E. *et al.* Recurrence of spondylarthropathy among first-degree relatives of patients: a systematic cross-sectional study. *Ann. Rheum. Dis.* **68,** 502–507 (2009).
51. Wickramaratne, P. J. & Hodge, S. E. Estimation of Sibling Recurrence-Risk Ratio under Single Ascertainment in Two-Child Families. *Am. J. Hum. Genet.* **68,** 807–812 (2001).
52. Breban, M., Miceli-Richard, C., Zinovieva, E., Monnet, D. & Said-Nahal, R. The genetics of spondyloarthropathies. *Jt. Bone Spine Rev. Rhum.* **73,** 355–362 (2006).
53. Reveille, J. D. The genetic basis of spondyloarthritis. *Ann. Rheum. Dis.* **70 Suppl 1,** i44–50 (2011).
54. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425,** 805–811 (2003).
55. Horton, R. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5,** 889–899 (2004).
56. The MHC sequencing Consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401,** 921–923 (1999).
57. Abbas, A. K. & Lichtman, A. H. *Les bases de l'immunologie fondamentale et clinique*. (Elsevier Masson, 2008).
58. Brewerton, D. A. *et al.* Ankylosing spondylitis and HL-A 27. *Lancet* **1,** 904–907 (1973).
59. Schwimmbeck, P. L., Yu, D. T. & Oldstone, M. B. Autoantibodies to HLA B27 in the sera of

HLA B27 patients with ankylosing spondylitis and Reiter's syndrome. Molecular mimicry with Klebsiella pneumoniae as potential mechanism of autoimmune disease. *J. Exp. Med.* **166,** 173–181 (1987).

60. Benjamin, R. & Parham, P. Guilt by association: HLA-B27 and ankylosing spondylitis. *Immunol. Today* **11,** 137–142 (1990).
61. Colbert, R. A., DeLay, M. L., Layh-Schmitt, G. & Sowders, D. P. HLA-B27 misfolding and spondyloarthropathies. *Prion* **3,** 15–26 (2009).
62. Chan, A. T., Kollnberger, S. D., Wedderburn, L. R. & Bowness, P. Expansion and enhanced survival of natural killer cells expressing the killer immunoglobulin-like receptor KIR3DL2 in spondylarthritis. *Arthritis Rheum.* **52,** 3586–3595 (2005).
63. Bowness, P. *et al.* Th17 cells expressing KIR3DL2+ and responsive to HLA-B27 homodimers are increased in ankylosing spondylitis. *J. Immunol. Baltim. Md 1950* **186,** 2672–2680 (2011).
64. Díaz-Peña, R., López-Vázquez, A. & López-Larrea, C. Old and new HLA associations with ankylosing spondylitis. *Tissue Antigens* **80,** 205–213 (2012).
65. Reveille, J. D. Major histocompatibility genes and ankylosing spondylitis. *Best Pract. Res. Clin. Rheumatol.* **20,** 601–609 (2006).
66. López-Larrea, C. *et al.* HLA-B27 subtypes in Asian patients with ankylosing spondylitis. Evidence for new associations. *Tissue Antigens* **45,** 169–176 (1995).
67. Paladini, F. *et al.* Distribution of HLA-B27 subtypes in Sardinia and continental Italy and their association with spondylarthropathies. *Arthritis Rheum.* **52,** 3319–3321 (2005).
68. Van Gaalen, F. A. *et al.* Epistasis between two HLA antigens defines a subset of individuals at a very high risk for ankylosing spondylitis. *Ann. Rheum. Dis.* **72,** 974–978 (2013).
69. Lv, C. *et al.* Association of Interleukin-10 gene polymorphisms with ankylosing spondylitis. *Clin. Investig. Med. Médecine Clin. Exp.* **34,** E370 (2011).
70. Wu, W. *et al.* Susceptibility to ankylosing spondylitis: evidence for the role of ERAP1, TGFb1

and TLR9 gene polymorphisms. *Rheumatol. Int.* **32,** 2517–2521 (2012).

71. Lee, W.-Y. *et al.* Polymorphisms of cytotoxic T lymphocyte-associated antigen-4 and cytokine genes in Taiwanese patients with ankylosing spondylitis. *Tissue Antigens* **75,** 119–126 (2010).
72. Xu, W.-D., Liu, S.-S., Pan, H.-F. & Ye, D.-Q. Lack of association of TLR4 polymorphisms with susceptibility to rheumatoid arthritis and ankylosing spondylitis: a meta-analysis. *Jt. Bone Spine Rev. Rhum.* **79,** 566–569 (2012).
73. Danoy, P. *et al.* Association of variants at 1q32 and STAT3 with ankylosing spondylitis suggests genetic overlap with Crohn's disease. *PLoS Genet.* **6,** e1001195 (2010).
74. Liu, X. *et al.* Programmed cell death 1 gene polymorphisms is associated with ankylosing spondylitis in Chinese Han population. *Rheumatol. Int.* **31,** 209–213 (2011).
75. Duan, Z.-H. *et al.* The FCGR2B rs10917661 polymorphism may confer susceptibility to ankylosing spondylitis in Han Chinese: a case-control study. *Scand. J. Rheumatol.* **41,** 219–222 (2012).
76. Inanır, A. *et al.* Significant association between insertion/deletion polymorphism of the angiotensin-convertig enzyme gene and ankylosing spondylitis. *Mol. Vis.* **18,** 2107–2113 (2012).
77. Wellcome Trust Case Control Consortium *et al.* Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* **39,** 1329–1337 (2007).
78. Australo-Anglo-American Spondyloarthritis Consortium (TASC) *et al.* Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nat. Genet.* **42,** 123–127 (2010).
79. Evans, D. M. *et al.* Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat. Genet.* **43,** 761–767 (2011).
80. Lin, Z. *et al.* A genome-wide association study in Han Chinese identifies new susceptibility

loci for ankylosing spondylitis. *Nat. Genet.* **44,** 73–77 (2012).

81. Laval, S. H. *et al.* Whole-genome screening in ankylosing spondylitis: evidence of non-MHC genetic-susceptibility loci. *Am. J. Hum. Genet.* **68,** 918–926 (2001).
82. Zhang, G. *et al.* Genetic studies in familial ankylosing spondylitis susceptibility. *Arthritis Rheum.* **50,** 2246–2254 (2004).
83. Miceli-Richard, C. *et al.* Significant linkage to spondyloarthropathy on 9q31-34. *Hum. Mol. Genet.* **13,** 1641–1648 (2004).
84. Lee, Y. H., Rho, Y. H., Choi, S. J., Ji, J. D. & Song, G. G. Ankylosing spondylitis susceptibility loci defined by genome-search meta-analysis. *J. Hum. Genet.* **50,** 453–459 (2005).
85. Carter, K. W. *et al.* Combined analysis of three whole genome linkage scans for Ankylosing Spondylitis. *Rheumatol. Oxf. Engl.* **46,** 763–771 (2007).
86. Zinovieva, E. *et al.* Comprehensive linkage and association analyses identify haplotype, near to the TNFSF15 gene, significantly associated with spondyloarthritis. *PLoS Genet.* **5,** e1000528 (2009).
87. Zinovieva, E. *et al.* Lack of association between Tenascin-C gene and spondyloarthritis. *Rheumatol. Oxf. Engl.* **47,** 1655–1658 (2008).
88. Zinovieva, E. *et al.* Systematic candidate gene investigations in the SPA2 locus (9q32) show an association between TNFSF8 and susceptibility to spondylarthritis. *Arthritis Rheum.* **63,** 1853–1859 (2011).
89. Brown, M. A. *et al.* A genome-wide screen for susceptibility loci in ankylosing spondylitis. *Arthritis Rheum.* **41,** 588–595 (1998).
90. Strimbu, K. & Tavel, J. A. What are Biomarkers? *Curr. Opin. HIV AIDS* **5,** 463–466 (2010).
91. Bird, A. Perceptions of epigenetics. *Nature* **447,** 396–398 (2007).
92. Michaelson, J. J., Loguercio, S. & Beyer, A. Detection and interpretation of expression

quantitative trait loci (eQTL). *Methods San Diego Calif* **48,** 265–276 (2009).

93. Rosikiewicz, M. & Robinson-Rechavi, M. IQRray, a new method for Affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics. *Bioinforma. Oxf. Engl.* **30,** 1392–1399 (2014).
94. Cahoy, J. D. *et al.* A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *J. Neurosci.* **28,** 264–278 (2008).
95. Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C. & Hume, D. A. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics* **14,** 632 (2013).
96. Wan, Q. *et al.* BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database J. Biol. Databases Curation* **2015,** (2015).
97. Rajeevan, M. S., Ranamukhaarachchi, D. G., Vernon, S. D. & Unger, E. R. Use of real-time quantitative PCR to validate the results of cDNA array and differential display PCR technologies. *Methods San Diego Calif* **25,** 443–451 (2001).
98. Chu, Y. & Corey, D. R. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Ther.* **22,** 271–274 (2012).
99. www.genomatix.de.
100. www.ingenuity.com.
101. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12,** 115–121 (2015).
102. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostat. Oxf. Engl.* **4,** 249–264 (2003).
103. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30,** 207–210 (2002).

104. Tseng, G. C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* **40,** 3785–3799 (2012).

105. Wang, X. *et al.* An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinforma. Oxf. Engl.* **28,** 2534–2536 (2012).

106. Kang, D. D., Sibille, E., Kaminski, N. & Tseng, G. C. MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Res.* **40,** e15 (2012).

107. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3,** Article3 (2004).

108. Konishi, S. & Kitagawa, G. *Information Criteria and Statistical Modeling*. (Springer New York, 2008). at <http://link.springer.com/10.1007/978-0-387-71887-3>

109. Li, D., Le Pape, M. A., Parikh, N. I., Chen, W. X. & Dye, T. D. Assessing differential expression in two-color microarrays: a resampling-based empirical Bayes approach. *PloS One* **8,** e80099 (2013).

110. Veitia, R. A. One thousand and one ways of making functionally similar transcriptional enhancers. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **30,** 1052–1057 (2008).

111. Hirahara, K. *et al.* Mechanisms underlying helper T-cell plasticity: implications for immune-mediated disease. *J. Allergy Clin. Immunol.* **131,** 1276–1287 (2013).

112. The Gene Ontology project in 2008. *Nucleic Acids Res.* **36,** D440–D444 (2008).

113. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

114. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33,** D514–D517 (2005).

115. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for

interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).

116. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41,** D377–386 (2013).

117. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33,** W741–748 (2005).

118. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).

119. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14,** 128 (2013).

120. Tamayo, P., Steinhardt, G., Liberzon, A. & Mesirov, J. P. The limitations of simple gene set enrichment analysis assuming gene independence. *Stat. Methods Med. Res.* (2012). doi:10.1177/0962280212460441

121. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40,** e133 (2012).

122. Tarca, A. L., Bhatti, G. & Romero, R. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLoS ONE* **8,** (2013).

123. Yaari, G., Bolen, C. R., Thakar, J. & Kleinstein, S. H. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Res.* **41,** e170 (2013).

124. Shields, R. Cultural Topology: The Seven Bridges of Königsburg, 1736. *Theory Cult. Soc.* **29,** 43–57 (2012).

125. Michael, D. K. & Battiston, S. in *Networks, Topology and Dynamics* (eds. Naimzada, A. K., Stefani, S. & Torriero, A.) 23–63 (Springer Berlin Heidelberg, 2009). at

<http://link.springer.com/chapter/10.1007/978-3-540-68409-1_2>

126. 6, N. & 2013. The graph theory of friendship, community, and leadership. *UBC News* at <http://news.ubc.ca/2013/11/06/the-graph-theory-of-friendship-community-and-leadership/>

127. Arney, D. C. & Wilhite, A. W. Modeling Space System Architectures with Graph Theory. *J. Spacecr. Rockets* **51,** 1413–1429 (2014).

128. Barthelemy, M. & Amaral, L. A. N. Small-world networks: Evidence for a crossover picture. *Phys. Rev. Lett.* **82,** 3180–3183 (1999).

129. Guizani, M., Rayes, A., Khan, B. & Al-Fuqaha, A. in *Network Modeling and Simulation* 69–96 (John Wiley & Sons, Ltd, 2010). at <http://onlinelibrary.wiley.com/doi/10.1002/9780470515211.ch4/summary>

130. Zemel, R. S. & Mozer, M. C. Localist attractor networks. *Neural Comput.* **13,** 1045–1064 (2001).

131. MacArthur, R. Fluctuations of Animal Populations and a Measure of Community Stability. *Ecology* **36,** 533–536 (1955).

132. Network structure and biodiversity loss in food webs: robustness increases with connectance. at <http://www.academia.edu/6357545/Network_structure_and_biodiversity_loss_in_food_webs_robustness_increases_with_connectance>

133. Bascompte, J. Disentangling the Web of Life. *Science* **325,** 416–419 (2009).

134. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10,** 186–198 (2009).

135. Stephan, K. E. *et al.* Computational analysis of functional connectivity between areas of primate cerebral cortex. *Philos. Trans. R. Soc. B Biol. Sci.* **355,** 111–126 (2000).

136. Gopinath, K., Krishnamurthy, V., Cabanban, R. & Crosson, B. A. Hubs of Anticorrelation in High-Resolution Resting-State Functional Connectivity Network Architecture. *Brain*

Connect. (2015). doi:10.1089/brain.2014.0323

137. Ryslik, G. A., Cheng, Y., Cheung, K.-H., Modis, Y. & Zhao, H. A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations. *BMC Bioinformatics* **15,** 86 (2014).

138. Xue, W., Jiao, P., Liu, H. & Yao, X. Molecular modeling and residue interaction network studies on the mechanism of binding and resistance of the HCV NS5B polymerase mutants to VX-222 and ANA598. *Antiviral Res.* **104,** 40–51 (2014).

139. Sborgi, L. *et al.* Interaction Networks in Protein Folding via Atomic-Resolution Experiments and Long-Timescale Molecular Dynamics Simulations. *J. Am. Chem. Soc.* (2015). doi:10.1021/jacs.5b02324

140. Naldi, A., Carneiro, J., Chaouiya, C. & Thieffry, D. Diversity and plasticity of Th cell types predicted from regulatory network modelling. *PLoS Comput. Biol.* **6,** e1000912 (2010).

141. Csikász-Nagy, A., Cavaliere, M. & Sedwards, S. in *New Challenges for Cancer Systems Biomedicine* (eds. d' Onofrio, A., Cerrai, P. & Gandolfi, A.) 3–18 (Springer Milan, 2012). at <http://link.springer.com/chapter/10.1007/978-88-470-2571-4_1>

142. Mashaghi, A. R., Ramezanpour, A. & Karimipour, V. Investigation of a protein complex network. *Eur. Phys. J. B - Condens. Matter Complex Syst.* **41,** 113–121 (2004).

143. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10,** 252–263 (2009).

144. Thomas, R. & Kaufman, M. Conceptual tools for the integration of data. *C. R. Biol.* **325,** 505–514 (2002).

145. Cao, J., Qi, X. & Zhao, H. Modeling gene regulation networks using ordinary differential equations. *Methods Mol. Biol. Clifton NJ* **802,** 185–197 (2012).

146. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data.

SIAM Rev. **51,** 661–703 (2009).

147. Shaw, S. Evidence of Scale-Free Topology and Dynamics in Gene Regulatory Networks. *ArXivcond-Mat0301041* (2003). at <http://arxiv.org/abs/cond-mat/0301041>

148. Dong, J. & Horvath, S. Understanding network concepts in modules. *BMC Syst. Biol.* **1,** 24 (2007).

149. Travers, J., Milgram, S., Travers, J. & Milgram, S. An Experimental Study of the Small World Problem. *Sociometry* **32,** 425–443 (1969).

150. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).

151. Xue, J. *et al.* Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity* **40,** 274–288 (2014).

152. Verfaillie, A. *et al.* Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat. Commun.* **6,** (2015).

153. Lin, Y. *et al.* MiRNA and TF co-regulatory network analysis for the pathology and recurrence of myocardial infarction. *Sci. Rep.* **5,** (2015).

154. Cotney, J. *et al.* The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat. Commun.* **6,** (2015).

155. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 1974–1979 (2005).

156. Zhao, J., Yu, H., Luo, J.-H., Cao, Z.-W. & Li, Y.-X. Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinformatics* **7,** 386 (2006).

157. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).

158. Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2,** 2366–2382 (2007).

159. Saito, R. *et al*. A travel guide to Cytoscape plugins. *Nat. Methods* **9,** 1069–1076 (2012).

160. cyREST: Cytoscape as a Service. (12:50:16 UTC). at <http://fr.slideshare.net/keiono/cy-rest-recombcytoscapeworkshop2014>

161. Theocharidis, A., van Dongen, S., Enright, A. J. & Freeman, T. C. Network visualization and analysis of gene expression data using BioLayout Express3D. *Nat. Protoc.* **4,** 1535–1550 (2009).

162. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35,** W193–W200 (2007).

163. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37,** W305–311 (2009).

164. Lynn, D. J. *et al.* InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.* **4,** 218 (2008).

165. Montojo, J. *et al.* GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinforma. Oxf. Engl.* **26,** 2927–2928 (2010).

166. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: network-based gene set enrichment analysis. *Bioinforma. Oxf. Engl.* **28,** i451–i457 (2012).

167. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41,** D808–815 (2013).

168. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–452 (2015).

169. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

170. Mordelet, F. & Vert, J.-P. SIRENE: supervised inference of regulatory networks.

Bioinformatics **24,** i76–i82 (2008).

171. Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1,** S7 (2006).

172. De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* **8,** 717–729 (2010).

173. Baba, K., Shibata, R. & Sibuya, M. Partial Correlation and Conditional Correlation as Measures of Conditional Independence. *Aust. N. Z. J. Stat.* **46,** 657–664 (2004).

174. Chiquet, J., Smith, A., Grasseau, G., Matias, C. & Ambroise, C. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinforma. Oxf. Engl.* **25,** 417–418 (2009).

175. Banerjee, O., Ghaoui, L. E. & d' Aspremont, A. Model Selection Through Sparse Maximum Likelihood Estimation. *ArXiv07070704 Cs* (2007). at <http://arxiv.org/abs/0707.0704>

176. Weighted-LASSO for Structured Network Inference from Time Course Data : Statistical Applications in Genetics and Molecular Biology. at <http://www.degruyter.com/view/j/sagmb.2010.9.1/sagmb.2010.9.1.1519/sagmb.2010.9.1.1519.xml>

177. Rihl, M. *et al.* Technical validation of cDNA based microarray as screening technique to identify candidate genes in synovial tissue biopsy specimens from patients with spondyloarthropathy. *Ann. Rheum. Dis.* **63,** 498–507 (2004).

178. Rihl, M. *et al.* Identification of interleukin-7 as a candidate disease mediator in spondylarthritis. *Arthritis Rheum.* **58,** 3430–3435 (2008).

179. Laukens, D. *et al.* Altered gut transcriptome in spondyloarthropathy. *Ann. Rheum. Dis.* **65,** 1293–1300 (2006).

180. Thomas, G. P. *et al.* Expression profiling in spondyloarthropathy synovial biopsies highlights changes in expression of inflammatory genes in conjunction with tissue

remodelling genes. *BMC Musculoskelet. Disord.* **14,** 354 (2013).

181. Sharma, S. M. *et al.* Insights in to the pathogenesis of axial spondyloarthropathy based on gene expression profiles. *Arthritis Res. Ther.* **11,** R168 (2009).

182. Assassi, S. *et al.* Whole-blood gene expression profiling in ankylosing spondylitis shows upregulation of toll-like receptor 4 and 5. *J. Rheumatol.* **38,** 87–98 (2011).

183. Pimentel-Santos, F. M. *et al.* Whole blood transcriptional profiling in ankylosing spondylitis identifies novel candidate genes that might contribute to the inflammatory and tissue-destructive disease aspects. *Arthritis Res. Ther.* **13,** R57 (2011).

184. Chen, K. *et al.* Whole-blood gene expression profiling in ankylosing spondylitis identifies novel candidate genes that may contribute to the inflammatory and tissue-destructive disease aspects. *Cell. Immunol.* **286,** 59–64 (2013).

185. Haroon, N. *et al.* From gene expression to serum proteins: biomarker discovery in ankylosing spondylitis. *Ann. Rheum. Dis.* **69,** 297–300 (2010).

186. Cheok, M. H. *et al.* Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat. Genet.* **34,** 85–90 (2003).

187. Chuang, Y.-Y. E. *et al.* Gene expression after treatment with hydrogen peroxide, menadione, or t-butyl hydroperoxide in breast cancer cells. *Cancer Res.* **62,** 6246–6254 (2002).

188. Folgueira, M. A. A. K. *et al.* Gene expression profile associated with response to doxorubicin-based therapy in breast cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **11,** 7434–7443 (2005).

189. Kauraniemi, P. *et al.* Effects of Herceptin treatment on global gene expression patterns in HER2-amplified and nonamplified breast cancer cell lines. *Oncogene* **23,** 1010–1013 (2004).

190. Gu, J. *et al.* A 588-gene microarray analysis of the peripheral blood mononuclear cells of spondyloarthropathy patients. *Rheumatol. Oxf. Engl.* **41,** 759–766 (2002).

191. Briggs, R. *et al.* Interferon alpha selectively affects expression of the human myeloid cell nuclear differentiation antigen in late stage cells in the monocytic but not the granulocytic lineage. *J. Cell. Biochem.* **54,** 198–206 (1994).

192. Gu, J. *et al.* Identification of RGS1 as a candidate biomarker for undifferentiated spondylarthritis by genome-wide expression profiling and real-time polymerase chain reaction. *Arthritis Rheum.* **60,** 3269–3279 (2009).

193. Duan, R., Leo, P., Bradbury, L., Brown, M. A. & Thomas, G. Gene expression profiling reveals a downregulation in immune-associated genes in patients with AS. *Ann. Rheum. Dis.* **69,** 1724–1729 (2010).

194. Smith, J. A. *et al.* Gene expression analysis of macrophages derived from ankylosing spondylitis patients reveals interferon-gamma dysregulation. *Arthritis Rheum.* **58,** 1640–1649 (2008).

195. Fert, I. *et al.* Reverse interferon signature is characteristic of antigen-presenting cells in human and rat spondyloarthritis. *Arthritis Rheumatol. Hoboken NJ* **66,** 841–851 (2014).

196. Zhu, Z.-Q., Tang, J.-S. & Cao, X.-J. Transcriptome network analysis reveals potential candidate genes for ankylosing spondylitis. *Eur. Rev. Med. Pharmacol. Sci.* **17,** 3178–3185 (2013).

197. Talpin, A. *et al.* Monocyte-derived dendritic cells from HLA-B27+ axial spondyloarthritis (SpA) patients display altered functional capacity and deregulated gene expression. *Arthritis Res. Ther.* **16,** 417 (2014).

198. Mattey, D. L. *et al.* Association of cytokine and matrix metalloproteinase profiles with disease activity and function in ankylosing spondylitis. *Arthritis Res. Ther.* **14,** R127 (2012).

199. Costantino, F. *et al.* ERAP1 gene expression is influenced by non-synonymous polymorphisms associated with predisposition to spondyloarthritis. *Arthritis Rheumatol. Hoboken NJ* (2015). doi:10.1002/art.39072

200. Wright, C. *et al.* Ankylosing spondylitis monocytes show upregulation of proteins involved in inflammation and the ubiquitin proteasome pathway. *Ann. Rheum. Dis.* **68,** 1626–1632 (2009).

201. Barnes, M. G. *et al.* Gene expression in juvenile arthritis and spondyloarthropathy: pro-angiogenic ELR+ chemokine genes relate to course of arthritis. *Rheumatol. Oxf. Engl.* **43,** 973–979 (2004).

202. Zhao, J., Chen, J., Yang, T.-H. & Holme, P. Insights into the pathogenesis of axial spondyloarthropathy from network and pathway analysis. *BMC Syst. Biol.* **6 Suppl 1,** S4 (2012).

203. Ohgimoto, K. *et al.* Difference in production of infectious wild-type measles and vaccine viruses in monocyte-derived dendritic cells. *Virus Res.* **123,** 1–8 (2007).

204. Segura, E. *et al.* Human Inflammatory Dendritic Cells Induce Th17 Cell Differentiation. *Immunity* **38,** 336–348 (2013).

205. Hammer, R. E., Maika, S. D., Richardson, J. A., Tang, J. P. & Taurog, J. D. Spontaneous inflammatory disease in transgenic rats expressing HLA-B27 and human beta 2m: an animal model of HLA-B27-associated human disorders. *Cell* **63,** 1099–1112 (1990).

206. Taurog, J. D., Maika, S. D., Simmons, W. A., Breban, M. & Hammer, R. E. Susceptibility to inflammatory disease in HLA-B27 transgenic rat lines correlates with the level of B27 expression. *J. Immunol. Baltim. Md 1950* **150,** 4168–4178 (1993).

207. Taurog, J. D. *et al.* The germfree state prevents development of gut and joint inflammatory disease in HLA-B27 transgenic rats. *J. Exp. Med.* **180,** 2359–2364 (1994).

208. Transfer of the inflammatory disease of HLA-B27 transgenic rats by bone marrow engraftment. *J. Exp. Med.* **178,** 1607–1616 (1993).

209. Breban, M. *et al.* T cells, but not thymic exposure to HLA-B27, are required for the inflammatory disease of HLA-B27 transgenic rats. *J. Immunol. Baltim. Md 1950* **156,** 794–

803 (1996).

210. Hacquard-Bouder, C. *et al.* Defective costimulatory function is a striking feature of antigen-presenting cells in an HLA–B27–transgenic rat model of spondylarthropathy. *Arthritis Rheum.* **50,** 1624–1635 (2004).

211. Hacquard-Bouder, C. *et al.* Alteration of antigen-independent immunologic synapse formation between dendritic cells from HLA-B27-transgenic rats and CD4+ T cells: selective impairment of costimulatory molecule engagement by mature HLA-B27. *Arthritis Rheum.* **56,** 1478–1489 (2007).

212. Fert, I. *et al.* Correlation between dendritic cell functional defect and spondylarthritis phenotypes in HLA-B27/HUMAN beta2-microglobulin-transgenic rat lines. *Arthritis Rheum.* **58,** 3425–3429 (2008).

213. Dhaenens, M. *et al.* Dendritic cells from spondylarthritis-prone HLA-B27-transgenic rats display altered cytoskeletal dynamics, class II major histocompatibility complex expression, and viability. *Arthritis Rheum.* **60,** 2622–2632 (2009).

214. Affymetrix. BRLMM: an improved genotype calling method for the genechip human mapping 500k array set. at <http://www.biostat.jhsph.edu/~iruczins/teaching/misc/gwas/papers/affymetrix2006.pdf>

215. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

216. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A Note on Exact Tests of Hardy-Weinberg Equilibrium. *Am. J. Hum. Genet.* **76,** 887–893 (2005).

217. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33,** e175–e175 (2005).

218. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–D755 (2014).

219. Baker, M. Quantitative data: learning to share. *Nat. Methods* **9,** 39–41 (2012).

220. http://string-stitch.blogspot.fr/2010/03/combining-scores-right-way.html.

221. Zhu, J. *et al.* Regulation of apoAI processing by procollagen C-proteinase enhancer-2 and bone morphogenetic protein-1. *J. Lipid Res.* **50,** 1330–1339 (2009).

222. Martínez-Glez, V. *et al.* Identification of a mutation causing deficient BMP1/mTLD proteolytic activity in autosomal recessive osteogenesis imperfecta. *Hum. Mutat.* **33,** 343–350 (2012).

223. McManaman, J. L. *et al.* Perilipin-2-null mice are protected against diet-induced obesity, adipose inflammation, and fatty liver disease. *J. Lipid Res.* **54,** 1346–1359 (2013).

224. Azfer, A., Niu, J., Rogers, L. M., Adamski, F. M. & Kolattukudy, P. E. Activation of endoplasmic reticulum stress response during the development of ischemic heart disease. *Am. J. Physiol. Heart Circ. Physiol.* **291,** H1411–H1420 (2006).

225. Zhang, Y., Zhang, M., Wu, J., Lei, G. & Li, H. Transcriptional regulation of the Ufm1 conjugation system in response to disturbance of the endoplasmic reticulum homeostasis and inhibition of vesicle trafficking. *PloS One* **7,** e48587 (2012).

226. Petit, M. M. *et al.* LHFP, a novel translocation partner gene of HMGIC in a lipoma, is a member of a new family of LHFP-like genes. *Genomics* **57,** 438–441 (1999).

227. Shugart, Y. Y. *et al.* An SNP linkage scan identifies significant Crohn's disease loci on chromosomes 13q13.3 and, in Jewish families, on 1p35.2 and 3q29. *Genes Immun.* **9,** 161–167 (2008).

228. Jung, S.-H. *et al.* Genome-wide copy number variation analysis identifies deletion variants associated with ankylosing spondylitis. *Arthritis Rheumatol. Hoboken NJ* **66,** 2103–2112 (2014).

229. Mattapallil, M. J. *et al.* Common genetic determinants of uveitis shared with other autoimmune disorders. *J. Immunol. Baltim. Md 1950* **180,** 6751–6759 (2008).

230. Murthy, A. *et al.* Stromal TIMP3 regulates liver lymphocyte populations and provides

protection against Th1 T cell-driven autoimmune hepatitis. *J. Immunol. Baltim. Md 1950* **188,** 2876–2883 (2012).

231. Stöhr, R. *et al.* Loss of TIMP3 exacerbates atherosclerosis in ApoE null mice. *Atherosclerosis* **235,** 438–443 (2014).

232. Sahebjam, S., Khokha, R. & Mort, J. S. Increased collagen and aggrecan degradation with age in the joints of Timp3(-/-) mice. *Arthritis Rheum.* **56,** 905–909 (2007).

233. Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat. Genet.* **45,** 470–471 (2013).

234. Roach, J. C. *et al.* Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. *Proc. Natl. Acad. Sci.* **104,** 16245–16250 (2007).

235. Clark, J. G., Kim, K.-H., Basom, R. S. & Gharib, S. A. Plasticity of airway epithelial cell transcriptome in response to flagellin. *PloS One* **10,** e0115486 (2015).

236. Yamazaki, K. *et al.* Suppression of iodide uptake and thyroid hormone synthesis with stimulation of the type I interferon system by double-stranded ribonucleic acid in cultured human thyroid follicles. *Endocrinology* **148,** 3226–3235 (2007).

237. Neumann-Haefelin, C. HLA-B27-mediated protection in HIV and hepatitis C virus infection and pathogenesis in spondyloarthritis: two sides of the same coin? *Curr. Opin. Rheumatol.* **25,** 426–433 (2013).

238. McMichael, A. & Bowness, P. HLA-B27: natural function and pathogenic role in spondyloarthritis. *Arthritis Res. Ther.* **4,** S153 (2002).

239. Rosenberg, A. *et al.* Divergent gene activation in peripheral blood and tissues of patients with rheumatoid arthritis, psoriatic arthritis and psoriasis following infliximab therapy. *PloS One* **9,** e110657 (2014).

240. Bernstein, P. *et al.* Expression pattern differences between osteoarthritic chondrocytes and mesenchymal stem cells during chondrogenic differentiation. *Osteoarthr. Cartil. OARS*

Osteoarthr. Res. Soc. **18,** 1596–1607 (2010).

241. Costello, C. M. *et al.* Dissection of the Inflammatory Bowel Disease Transcriptome Using Genome-Wide cDNA Microarrays. *PLoS Med.* **2,** (2005).

242. Johnson-Huang, L. M. *et al.* A single intradermal injection of IFN-γ induces an inflammatory state in both non-lesional psoriatic and healthy skin. *J. Invest. Dermatol.* **132,** 1177–1187 (2012).

243. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **7,** 1–26 (1979).

244. Wiley: Bootstrap Methods: A Guide for Practitioners and Researchers, 2nd Edition - Michael R. Chernick. at <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471756210,subjectCd-STZ0.html>

245. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and Their Application*. (Cambridge University Press, 1997).

246. Manly, B. F. J. *Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition*. (Chapman and Hall/CRC, 2006).

247. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. (Springer-Verlag, 2005). at <http://link.springer.com/10.1007/b138696>

248. Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8,** 25 (2007).

249. Wu, C. F. J. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *Ann. Stat.* **14,** 1261–1295 (1986).

250. Kevorkian, L. *et al.* Expression profiling of metalloproteinases and their inhibitors in cartilage. *Arthritis Rheum.* **50,** 131–141 (2004).

251. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6,** (2015).

252. Yang, T.-P. *et al.* Genevar: a database and Java application for the analysis and

visualization of SNP-gene associations in eQTL studies. *Bioinformatics* **26,** 2474–2476 (2010).

253. Xia, K. *et al.* seeQTL: a searchable database for human eQTLs. *Bioinformatics* **28,** 451–452 (2012).

254. Jayaswal, V., Schramm, S.-J., Mann, G. J., Wilkins, M. R. & Yang, Y. H. VAN: an R package for identifying biologically perturbed networks via differential variability analysis. *BMC Res. Notes* **6,** 430 (2013).

255. Dimont, E., Shi, J., Kirchner, R. & Hide, W. edgeRun: an R package for sensitive, functionally relevant differential expression discovery using an unconditional exact test. *Bioinformatics* btv209 (2015). doi:10.1093/bioinformatics/btv209

256. Fukushima, A. DiffCorr: An R package to analyze and visualize differential correlations in biological networks. *Gene* **518,** 209–214 (2013).

257. Bao-Hong Liu, H. Y. DCGL: an R package for identifying differentially coexpressed genes and links from gene expression microarray data. *Bioinforma. Oxf. Engl.* **26,** 2637–8 (2010).

258. in (eds. Helden, J. van, Toussaint, A. & Thieffry, D.) (Springer New York, 2012). at <http://link.springer.com/protocol/10.1007%2F978-1-61779-361-5_23>

259. Semb, A. G. *et al.* Prediction of cardiovascular events in patients with ankylosing spondylitis and psoriatic arthritis: role of lipoproteins in a high-risk population. *J. Rheumatol.* **39,** 1433–1440 (2012).

260. Papagoras, C. *et al.* Cardiovascular risk profile in patients with spondyloarthritis. *Jt. Bone Spine Rev. Rhum.* **81,** 57–63 (2014).

261. Sundström, B., Johansson, G., Johansson, I. & Wållberg-Jonsson, S. Modifiable cardiovascular risk factors in patients with ankylosing spondylitis. *Clin. Rheumatol.* **33,** 111–117 (2014).

262. Genre, F. *et al.* Osteoprotegerin correlates with disease activity and endothelial activation

in non-diabetic ankylosing spondylitis patients undergoing TNF-α antagonist therapy. *Clin. Exp. Rheumatol.* **32,** 640–646 (2014).

263. He, M. *et al.* Mutations in the human SC4MOL gene encoding a methyl sterol oxidase cause psoriasiform dermatitis, microcephaly, and developmental delay. *J. Clin. Invest.* **121,** 976–984 (2011).