Combinaison d'approches expérimentales et bioinformatiques pour caractériser les intéractions entre Plasmodium falciparum et son hôte humain

Yvon Mbouamboua

AIX-MARSEILLE UNIVERSITÉ
UNIVERSITE MARIEN NGOUABI
ECOLE DOCTORALE SCIENCES DE LA VIE ET DE LA SANTE
FONDATION CONGOLAISE POUR LA RECHERCHE MEDICALE
THEORIES AND APPROACHES OF GENOMIC COMPLEXITY/INSERM

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Biologie santé
Spécialité : Génomique et Bioinformatique

Yvon MBOUAMBOUA

Titre de la thèse : Combinaison d'approches expérimentales et bioinformatiques pour caractériser les intéractions entre *Plasmodium falciparum* et son hôte humain

Thesis: Combinaison of experimental and bioinformatics approaches to characterize the interactions between *Plasmodium falciparum* and its human host

Soutenue le 23/09/2019 devant le jury composé de :

Prof. Florence ROBERT-GANGNEUX	Université de Rennes 1	Rapporteur
Prof. François SABOT	Université de Montpellier	Rapporteur
Dr. Laurence FLORI	Université de Montpellier	Examinateur
Prof. Pedro MALDONADO COUTINHO	Aix-Marseille Université	Président
Prof. associée Francine NTOUMI	Université Marien Ngouab	Co-directrice
Prof. Jacques van HELDEN	Université de Tübingen Aix-Marseille Université	Directeur de thèse

Numéro national de thèse/suffixe local : 2019AIXM0001/001ED62

Résumé

La première partie de ma thèse porte sur la caractérisation expérimentale des infections submicroscopiques à *Plasmodium falciparum* chez les femmes Congolaises asymptomatiques lors de l'accouchement.

Le paludisme demeure un problème majeur de santé publique dans le monde avec environ 219 millions de cas et 435 000 décès par an, majoritairement en Afrique sub-saharienne (90%). Dans la zone de forte transmission, l'infection palustre se caractérise principalement par le déclenchement d'une anémie maternelle et par la présence de parasites dans le placenta. Une forte proportion d'individus vivant dans des zones d'endémie palustre héberge des parasites indétectables par microscopie, mais qui peuvent néanmoins être détectées à l'aide d'outils moléculaires tels que la PCR. La morbi-mortalité liée au paludisme gestationnel a diminué depuis la mise en place du traitement préventif intermittent à base de la sulfadoxine-pyriméthamine (TPI-SP). Dans la zone endémique, on observe une baisse de la prévalence du paludisme placentaire et l'augmentation du poids moyen du nouveau-né. Malgré tout, l'infection plasmodiale chez la femme enceinte persiste et n'est pas maîtrisée.

Dans le but de comprendre les causes de cette persistance, j'ai caractérisé les populations parasitaires de *P. falciparum* chez les femmes enceintes Congolaises du sud de Brazzaville sous TPI-SP et j'ai analysé leur profil génétique dans le sang périphérique, placentaire et du cordon ombilical. L'évaluation de la fréquence de l'infection plasmodiale a montré qu'il y'a une baisse de la fréquence des infections microscopiques chez les femmes sous traitement préventif intermittent et une augmentation de la fréquence des infections submicroscopiques avec une diversité génétique modérée de *P. falciparum*. L'âge, la gravidité et les doses de SP n'interfèrent pas avec la multiplicité des infections qui est similaire dans les trois types de sang. Ces résultats contribuent à la compréhension de la dynamique des parasites dans le sang périphérique, placentaire et du cordon ombilical et de la population parasitaire en circulation au Congo.

La seconde partie de ma thèse a consisté à utiliser des approches bioinformatiques pour détecter les variants régulateurs associés à la susceptibilité au paludisme sévère.

Les progrès récents des technologies de séquençage ont permis d'identifier un spectre de plus en plus large de variants dans le génome humain. Cependant, l'identification des variants régulateurs associés à des maladies complexes reste un défi, en particulier pour identifier des variants pertinents du point de vue fonctionnel dans des régions non-codantes. Nous avons développé une méthode bioinformatique de prédiction des variants régulateurs des régions non-codantes associés à une maladie et agissant sur la régulation transcriptionnelle. L'approche repose sur l'intégration d'éléments d'informations collectées automatiquement à partir des bases de données Ensembl, dbSNP et GWAS catalog, et sur la sélection des variations susceptibles d'affecter la régulation, en combinant des outils bioinformatiques spécialisés : analyse de motifs (Regulatory Sequence Analysis

Tools) et de données ChIP-seq (ReMap). Pour ce faire, nous avons développé un workflow d'analyse dans le langage de statistiques R pour invoquer des ressources à distance (Web services). L'outil est conçu de façon générique, et peut s'adapter pour l'étude des variants régulateurs de n'importe quelle maladie documentée dans le GWAS catalog. Afin de faciliter son utilisation par un biologiste, l'outil génère automatiquement (en R markdown) un rapport d'analyse illustré par des figures et tableaux. J'ai testé l'outil avec un cas d'exemple du paludisme sévère qui a montré que sur un ensemble de 375 variants candidats, l'outil prédit 11 potentiels variants régulateurs qui altèrent les sites de liaisons des facteurs de transcription. Trois de ces variants candidats (rs1541253, rs1541254, rs1541255) associés au gène ATP2B4 (*ATPase 4 transportant le calcium sur la membrane plasmique*, code pour la pompe de calcium principale des érythrocytes) sont en cours de validation expérimentale pour leur impact sur l'activité du promoteur.

Mots clés : paludisme, *Plasmodium falciparum*, femme enceinte, vanriant régulateur,

Abstract

The first part of my thesis deals with the experimental characterization of *Plasmodium falciparum* submicroscopic infections in asymptomatic Congolese women during childbirth.

Malaria remains a major public health problem in the world with about 219 million cases and 435,000 deaths a year, mostly in sub-Saharan Africa (90%). In the area of high transmission, malaria infection is mainly characterized by the onset of maternal anemia and the presence of parasites in the placenta. A high proportion of individuals living in malaria-endemic areas harbor parasites undetectable by microscopy, but which can nevertheless be detected using molecular tools such as PCR. Morbidity and mortality related to gestational malaria has decreased since the introduction of intermittent preventive treatment with sulfadoxine-pyrimethamine (IPT-SP). In the endemic area, there is a decrease in the prevalence of placental malaria and an increase in the average weight of the newborn. Nevertheless, the plasmodial infection in the pregnant woman persists and is not controlled. In order to understand the causes of this persistence, I characterized the parasitic populations of *P. falciparum* in Congolese pregnant women from southern Brazzaville on IPT-SP and I analyzed their genetic profile in the peripheral blood, placental blood and umbilical cord. The evaluation of the frequency of the plasmodial infection has shown that the treatment correlates with a decrease in the frequency of microscopic infections and an increase in the frequency of submicroscopic infections with a moderate genetic diversity of *P. falciparum*. Age, pregnancy, and doses of SP do not interfere with the multiplicity of infections that are similar in all three types of blood. These results contribute to the understanding of parasite dynamics in peripheral, placental and umbilical cord blood and circulating parasite population in Congo.

The second part of my thesis involved the use of bioinformatic approaches to detect regulatory variants associated with susceptibility to severe malaria.

Recent advances in sequencing technologies enable to identify an increasingly broad spectrum of variants in the human genome. However, the identification of regulatory variants associated with complex diseases remains a challenge, particularly for identifying functionally relevant variants in non-coding regions. We have developed a bioinformatic method to predict regulatory variants of noncoding regions associated with a disease and acting on transcriptional regulation. The approach is based on the integration of various elements of information collected automatically from Ensembl database, dbSNP and GWAS catalog, and on the selection of variations that may affect regulation, by combining specialized bioinformatics tools: analysis Regulatory Sequence Analysis Tools and ChIP-seq (ReMap) data. To do this, we developed an analysis workflow in the R statistics language, which invokes remote resources (Web services). The tool is designed generically and can be adapted for the study of regulatory variants of any disease

documented in the GWAS catalog. In order to facilitate its use by a biologist, the tool automatically generates (in R markdown) an analysis report illustrated by figures and tables. I tested the tool with an example case of severe malaria that showed that on a set of 375 candidate variants, the tool predicts eleven potential regulatory variants that alter the binding sites of transcription factors. Three of these candidate variants (rs1541253, rs1541254, rs1541255) associated with the ATP2B4 gene (*plasma membrane calcium-transporting ATPase 4*, encodes the main calcium pump of erythrocytes) are being experimentally validated for their impact on promoter activity.

Keywords: malaria, *Plasmodium falciparum*, pregnant women, regulatory variant.

Remerciements

Ce travail est le fruit d'un investissement de plusieurs personnalités qui, ayant donné de leur temps, stratégie, énergie et patience, méritent ma reconnaissance.

Je remercie ma famille. Ma mère Florence EKOUNDA m'a emmené à l'école pour la première. C'est une femme dynamique et visionnaire qui espérait ce que je suis devenu aujourd'hui ; merci maman ! Mon père Ignace MBOUAMBOUA mérite ma reconnaissance. Il m'a montré l'exemple de la persévérance et m'a donné le goût de la biologie ; je lui dois ma reconnaissance.

Je remercie sincèrement Jacques van HELDEN et Francine NTOUMI , mes co-directeurs de thèse pour leur investissement avec beaucoup d'amour. Chaque instant auprès d'eux a été un moment très enrichissant d'apprentissage. Je ne saurai leur dire avec mes faibles mots combien ils sont généreux et très consacrés à la transmission du savoir. Je puis dire qu'ils sont de très bons formateurs.

Je remercie Catherine NGUYEN en sa qualité d'ancienne directrice lors de mon admission comme doctorant au laboratoire des *Theories and Approaches of Genomic Complexity* (TAGC).

Je remercie Pascal RIHET, l'actuel Directeur du laboratoire des Technologies Avancées pour la Génomique et la Complexité (TAGC), pour son soutien et sa participation active dans mon projet de thèse.

Je remercie Benoît BALLESTER , Emmanuel TALLA et Laurence RODER d'avoir joué leur rôle du comité de suivi de ma thèse.

Je ne saurai oublier Jean-Maurille OUAMBA qui a été un canal par lequel ma thèse a été financé en partie par Campus France. Aussi, l'aboutissement de ce travail est l'œuvre de la bonne volonté d'un complément financier de la Fondation Congolaise pour la Recherche Médicale (FCRM) et de L'Institut national de la santé et de la recherche médicale (INSERM).

Grâce à Pascal HINGAMP qui m'a enseigné le ABC de la bioinformatique, je puis dire que *je suis tombé amoureux de la bioinformatique*. Je lui dois ma reconnaissance.

Je remercie Félix KOUKUIKILA-KOUSSOUINDA pour son implication active de mes travaux expérimentales.

Je tiens à remercier mes collègues de la FCRM et du TAGC pour l'ambiance et les collaborations que nous avons partagé.

Je ne saurai terminer sans dire un grand merci à ma famille pour le soutien de tout genre. Je remercie mes parents pour leur

Encore merci !

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

ACT	Artemisinin-based Combination Therapy (association à base d'artémisinine)
AL	Artemether–Lumefantrine
LAMP	Loop-mediated isothermal amplification (amplification isothermique à boucle)
API	Application Programming Interface (interface de programmation d'application)
ASAQ	
BD	Base de données
BED	Browser Extensive Data
ChiP-seq	Chromatin ImmunoPrecipitation Sequencing
ECODE	Encyclopedia of DNA Elements (encyclopédie des éléments de l'ADN)
EIR	Entomological Inoculation Rate (taux d'inoculation entomologique)
HTTP	Hypertext Transfer Protocol (Protocole de transfert hypertexte)
ICAM-1	Inter cellular cell Adhésion molécule-1
ITPp-SP	intermittent preventive treatment with sulfadoxine-pyrimethamine for pregnant women (traitement préventif intermittent à base de la sulfadoxine-pyriméthamine chez la femme enceinte)
MHP	Ministry of Health and Population (Ministère de la Santé et de la Population)
LLINs	Long-Lasting Insecticide-treated mosquito Nets (moustiquaires imprégnées d'insecticide de longue durée)
ReMap	Regulatory Mapping (site web du catalogue des régions régions régulatrices chez l'homme)
PSSM	Position-Specific Scoring Matrices (Matrices de notation par position)
POST	Power-On Self-Test (l'auto-test au démarrage)
PAG	paludisme associé la grossesse
PNLP	Programme National de Lutte contre le Paludisme
PCR	Polymerase Chain Reaction
PfEMP1	Plasmodium falciparum-erythrocyte membrane protein-1
qPCR	Real-time quantitative (PCR quantitative en temps réel)
REST	Representational State Transfert (style d'architecture de construction d'applications)
RDC	République Démocratique du Congo

RSAT	Regulatory Sequence Analysis Tools (Outils d'analyse des séquences régulatrices)
SOI	SNP Of Interest (SNP d'interêt)
TDR	Test de Diagnostic Rapide
TF	Transcription factor (acteur de Transcription)
TFBS	Transcription Factor Binding Site (Site de fixation des facteurs de transcription)
TPIp	Traitement Préventif Intermittent pendant la grossesse
NGS	Next Generation Sequencing
URL	Uniform Resource Locator (localisateur uniforme de ressource)
CSA	Chondroïtine sulfate A
GE	Goutte Epaisse
FS	Frottis Sanguin
CQ	Chloroquine
CPN	Consultation Prénatale
DP	Dihydroartémisinine Pipéraquine
SP	Sulfadoxine Pyriméthamine
MII	Moustiquanire Imprégnée d'Insecticide
GWAS	Genome Wide Association Study (étude d'association du génome entier)
CRM	Cis-regulatory Module (module cis-régulateur)
DA-SNPs	Disease-Associated SNPs (SNPs associés à la maladie)
LD-SNPs	Linkage disequilibrium SNPs (SNPs en déséquilibre de liaison)

Introduction

Le paludisme est une maladie parasitaire, potentiellement mortelle, causée par un protozoaire du genre Plasmodium. Le parasite est transmis à l'homme par piqûre de moustiques femelles infestés du genre Anopheles. Cinq espèces sont responsables du paludisme chez l'homme : *Plasmodium falciparum, P. vivax, P. malariae, P. ovale et P. knowlesi. P. falciparum* est l'espèce responsable de la plupart des cas graves et des décès dus au paludisme surtout en Afrique sub-saharienne (White, Pukrittayakamee, Hien et al., 2014).

Malgré la baisse signalée de la mortalité par paludisme, le fardeau de l'infection par *P. falciparum* reste élevé, en particulier en Afrique subsaharienne, où survient environ 90% de la morbi-mortalité globale (*World Health Organization , 2015. Malaria Treatment Guidelines. Geneva, Switzerland : World Health Organization*). L'infection à *P. falciparum* entraîne l'un des trois résultats possibles : (i) la parasitémie asymptomatique, définie comme la présence de parasites asexués dans le sang sans symptômes ; (ii) le paludisme simple, qui implique une maladie fébrile non associée à des signes de gravité ; et (iii) le paludisme grave, caractérisé par divers syndromes de dysfonctionnement organique qui, s'ils ne sont pas traités rapidement, peuvent entraîner la mort (*World Health Organization , 2015. Malaria Treatment Guidelines. Geneva, Switzerland : World Health Organization*). Les nourrissons, les enfants de moins de 5 ans, les femmes enceintes, les patients infectés par le VIH/sida ainsi que les migrants non immunisés, les populations mobiles et les voyageurs sont des groupes de populations les plus vulnérables.

Le paludisme gestationnel dû à *P. falciparum* est une cause majeure de morbidité et de mortalité pour les femmes et leur progéniture (Steketee, Nahlen, Parise et al., 2001, Duffy et Fried, 1999). Pendant la grossesse, les parasites se fixent généralement dans le tissu placentaire (Fried et Duffy, 1996 ; Dahlbäck, Nielsen et Salanti, 2010), alors que les frottis du sang périphérique semblent souvent être négatifs. En outre, de nombreuses femmes infectées restent asymptomatiques, en particulier dans les zones de forte transmission où l'immunité systémique est élevée, bien que des séquelles, notamment l'anémie maternelle et le retard de croissance intra-utérine, se développent insidieusement et augmentent le risque de décès.

Le paludisme est une maladie multifactorielle (divers facteurs génétiques et environnementaux). La sévérité et le devenir de l'infection plasmodiale sont influencés par les interactions complexes entre de nombreux facteurs, notamment la génétique de l'hôte. Les différences génétiques entre individus peuvent modifier leur sensibilité au paludisme. La contribution des facteurs génétiques de l'hôte n'a montré que 2% des formes de résistance par la mutation de l'hémoglobine S. Les bases moléculaire de la susceptibilité/résistance au paludisme sont encore loin d'être élucidées. Il a donc été nécessaire d'étendre le champ des recherches sur la susceptibilité/résistance au paludisme en réalisant des études d'association à l'échelle du génome entier (GWAS).

Une des questions fondamentales de la génétique médicale est d'identifier les variants d'ADN fonctionnels responsables d'une maladie ou d'un phénotype d'intérêt. Les résultats d'études génétiques à grande échelle, telles que les GWAS, et la disponibilité de technologies de séquençage à haut débit offrent des possibilités sans précédent pour l'identification des variants associés à des maladies. Pour tirer bénéfice des progrès techniques résultant des technologies à haut débit, des méthodologies informatiques doivent être développées afin de hiérarchiser des millions de variants en fonction de leurs effets causatifs potentiels.

Mon travail combine une approche in vivo qui traite de la caractérisation des infections sub-microscopiques de *Plasmodium falciparum* dans le sang périphérique, placentaire et du cordon ombilical chez les femmes à l'accouchement en République du Congo, et une approche in silico visant à identifier des variants génétiques humains pouvant expliquer la survenue d'accès graves, en développant un workflow bioinformatique qui intègre des informations provenant de différentes bases de données et outils logiciels.

1. Généralités

Sommaire

1.1. Paludisme chez la femme enceinte

1.1.1. Généralité sur le paludisme

1.1.1.1. Historique du paludisme

Hippocrate (460 à 370 av. J-C), a été le premier à décrire les signes cliniques du paludisme, notamment la fièvre intermittente qu'il avait associé à certaines conditions climatiques et environnementales. Depuis l'antiquité, le paludisme a été décrit comme une fièvre survenant dans les zones marécageuses, sans qu'on en connaisse l'agent responsable. Au XVIIe siècle, les Indiens d'Amérique du Sud

remarquaient la disparition de ces fièvres quand ils absorbaient des décoctions d'écorce d'un arbuste, le quinquina. Rapportée en Europe, cette plante a soulagé de nombreuses personnalités de la cour du roi Louis XIV, qui étaient atteintes de fièvres intermittentes à Versailles, les bassins formant d'excellents gîtes larvaires. Mais il a fallu attendre 1820 pour que Pelletier et Caventou en isolent le principe actif : la quinine. C'est seulement en 1880 que Laveran va identifier le protozoaire responsable de la maladie qui sera appelé en 1885, *Plasmodium* (ce qui lui vaudra le prix Nobel en 1907). Treize ans après (1898), Ronald Ross, prouvait que le paludisme est transmis par piqûre d'un moustique ; l'Anophèle. Les antimalariques de synthèse (chloroquine) sont découverts en 1940 et sont très largement utilisés. Mais la résistance de *Plasmodium* à la chloroquine apparaît dès 1960, en Amérique du Sud et en Asie du Sud-Est, et se propage vite, jusqu'à la constatation de la diffusion de la chloroquino-résistance dans toutes les zones d'endémie, à partir des années 2000.

1.1.1.2. Situation du paludisme dans le monde

Le paludisme est un fléau qui sévit dans plus de 100 pays de la zone tropicale et subtropicale, nomment en Afrique sub-saharienne, en Asie, dans le Pacifique, en Amérique latine (Organisation mondiale de la santé, 2018).

Au niveau mondial, le nombre de cas de paludisme est estimé à 219 millions en 2017, contre 239 millions en 2010 et 217 millions en 2016 (Organisation mondiale de la santé, 2018)

Même si les estimations du nombre de cas de paludisme pour 2017 sont en baisse de 20 millions par rapport à 2010, les données pour la période 2015-2017 mettent en évidence l'absence de progrès significatifs par rapport à cet indicateur durant ce laps de temps.

La grande majorité des cas de paludisme (200 millions ou 92 %) a été enregistrés dans la région Afrique de l'OMS, loin devant la région Asie du Sud-Est (5 %) et la région Méditerranée orientale (2 %).

Quinze pays d'Afrique subsaharienne et l'Inde ont concentré quasiment 80 % du nombre total de cas de paludisme dans le monde, parmi lesquels 5, à eux seuls, ont enregistré près de la moitié des cas : le Nigéria (25 %), la République démocratique du Congo (11 %), le Mozambique (5 %), l'Inde (4 %) et l'Ouganda (4 %).

Les 10 pays où le paludisme sévit le plus en Afrique ont rapporté une hausse du nombre de cas en 2017 par rapport à 2016. Parmi ces pays, le Nigéria, Madagascar et la République démocratique du Congo auraient enregistré les plus fortes augmentations, toutes estimées à plus d'un demi-million de cas. À l'inverse, l'Inde a déclaré 3 millions de cas en moins durant la même période, soit une baisse de 24 % par rapport à 2016.

Au niveau mondial, l'incidence du paludisme a reculé entre 2010 et 2017, passant de 72 cas pour 1 000 habitants exposés au risque de paludisme à 59 pour 1 000. Même si ces chiffres représentent une baisse de 18 % sur la période, le nombre de cas pour 1 000 habitants exposés au risque de paludisme a stagné à 59 ces 3 dernières années

À l'exception de la région Asie du Sud-Est de l'OMS, où l'incidence du paludisme continue à baisser (17 cas pour 1 000 habitants exposés au risque de paludisme en

2010 contre 7 pour 1 000 en 2017, soit une baisse de 59 %), toutes les régions de l'OMS ont enregistré des progrès très modestes, voire une hausse de l'incidence. Dans la région Amériques de l'OMS, l'incidence du paludisme a augmenté, principalement à cause d'une transmission accrue au Brésil, au Nicaragua et au Venezuela. Dans la région Afrique de l'OMS, elle est restée à 219 cas pour 1 000 habitants exposés au risque de paludisme pour la deuxième année consécutive.

P. falciparum est le parasite du paludisme le plus prévalent dans la région Afrique de l'OMS ; il est en effet à l'origine de 99,7 % des cas de paludisme estimés en 2017, tout comme dans les régions Asie du Sud-Est (62,8 %), Méditerranée orientale (69 %) et Pacifique occidental (71,9 %). *P. vivax* prédomine dans la région Amériques de l'OMS, représentant 74,1 % des cas de paludisme.

Au niveau mondial, on estime que le paludisme est responsable de 435.000 décès chaque année contre 451 000 en 2016 et 607 000 en 2010. Ces décès sont majoritairement en Afrique sub-saharienne (Figure 1.1)[1]. Les femmes enceintes et les enfants de moins de 5 ans sont les plus vulnérables.

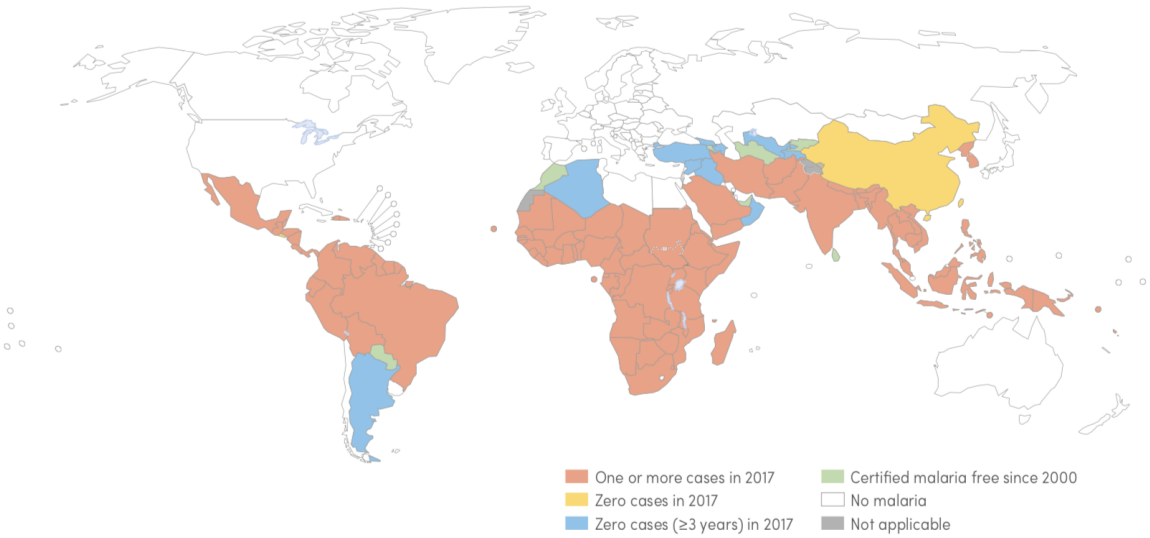


Figure 1.1. – Situation du paludisme dans le monde (source : Organisation mondiale de la santé, 2018).

Selon l'estimation mondiale de *Roll Back Malaria Partnership*, chaque année, environ 50 millions de femmes vivant dans des pays d'endémie palustre à travers le monde tombent enceintes[2]. Environ 30 millions de femmes africaines vivant dans des zones d'endémie palustre tombent enceintes et risquaient le paludisme chaque année (Dellicour, Tatem, Guerra et al., 2010).

Les enfants de moins de 5 ans sont les plus vulnérables face au paludisme. En 2017, ils ont représenté 61 % (266 000) des décès associés au paludisme dans le monde.

À elle seule, la région Afrique de l'OMS a enregistré 93 % des décès liés au paludisme au niveau mondial en 2017 ; elle a cependant représenté 88 % des 172 000 décès en moins dus à la maladie par rapport à 2010.

Près de 80 % des décès dus au paludisme dans le monde en 2017 ont été concentrés dans 17 pays de la région Afrique de l'OMS et en Inde. Sept de ces pays représentent 53 % des décès associés : le Nigéria (19 %), la République démocratique du Congo (11 %), le Burkina Faso (6 %), la République-Unie de Tanzanie (5 %), la Sierra Leone (4 %), le Niger (4 %) et l'Inde (4 %).

Par rapport à 2010, la mortalité liée au paludisme a diminué dans toutes les régions de l'OMS en 2017 à l'exception des Amériques. Les baisses les plus prononcées ont été observées en Asie du Sud-Est (54 %), Afrique (40 %) et Méditerranée orientale (10 %). Malgré ces progrès, la baisse de la mortalité liée au paludisme a ralenti depuis 2015, ce qui reflète les tendances estimées en matière d'incidence de la maladie.

1.1.1.3. Situation du paludisme en République du Congo

La République du Congo est située dans la partie centre-ouest de l'Afrique subsaharienne et traversé par l'équateur. Les deux grandes villes du pays (Brazzaville et Pointe-Noire) comptent 61% de la population congolaise. Le pays est alterné par deux saisons sèches (une courte de janvier à février et une longue de juin à septembre) et deux saisons de pluies ou saison humide (une courte de mars à mai et une longue d'octobre à décembre). Les températures varient entre 18 et 20°C pendant la saison sèche et entre 28 et 32°C pendant la saison des pluies (Figure 1.2) (Kimbatsa, Mahoungou et Ofoueme, 2018).

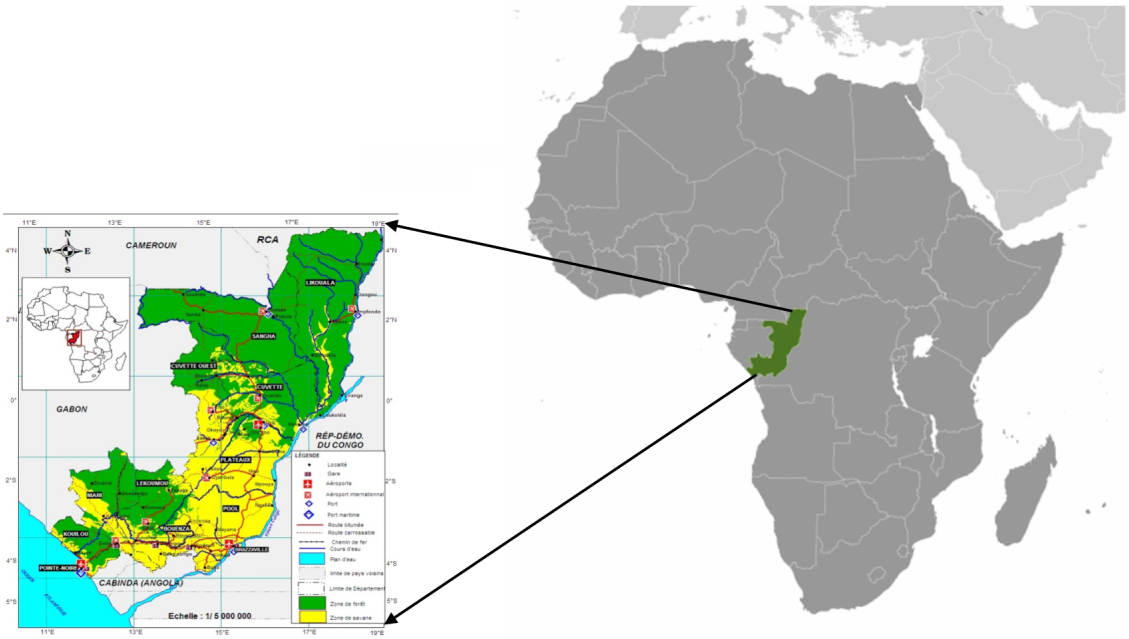


Figure 1.2. – Carte de la République du Congo (source : adapté de Kimbatsa, Mahoungou et Ofoueme, 2018).

Brazzaville, la capitale politique du Congo, est située le long du fleuve Congo, en face de Kinshasa (capitale de la République Démocratique du Congo, RDC). La transmission du paludisme demeure élevée [3]. Une récente enquête entomologique

menée conjointement par l'OMS et le Ministère de la Santé et de la Population (*ministry of health and population*, MHP) a montré que la transmission du paludisme dans le pays est pérenne toute l'année dans les zones forestières, avec un taux d'inoculation entomologique (*entomological inoculation rate*, EIR) de 200–1000 piqûres infectantes/personne/an[4].

Malgré des efforts et des progrès considérables en matière de contrôle du paludisme au cours des dernières décennies par l'adoption des combinaisons thérapeutiques à base d'artémicinine (*artemisinin-based combination therapy* , ACT) pour le traitement du paludisme simple en 2006, l'utilisation d'Artemether–Lumefantrine (AL) ou d'artesunate–amodiaquine (ASAQ), du traitement préventif intermittent (TPI) à base de la sulfadoxine-pyrimethamine (SP) pendant la grossesse (TPIp-SP), de distribution massive de moustiquaires imprégnées d'insecticide de 2008 à 2012 et traitement antipaludique gratuit pour les enfants âgés de moins de 15 ans depuis 2008 ; il reste l'un des problèmes de santé publique importants[5]. *P. falciparum* est le principal parasite du paludisme et *Anopheles gambiae ss* (Trape et Zoulani, 1987) qui transmet le paludisme généralement entre 18 heures et 6 heures, est le principal moustique-vecteur. Les dernières estimations du programme national de lutte contre le paludisme (PNLP) indiquent que le paludisme clinique représente 47,9% de toutes les consultations ambulatoires dans les hôpitaux publics, 64,8% des hospitalisations et 18,4% des décès[6].

1.1.1.4. L'agent vecteur et son écologie

Les *Anophèles* femelles sont les moustiques vecteurs et hôtes définitifs du parasite *Plasmodium* responsable du paludisme. Le genre *Anophèles* appartient au règne animal, embranchement des *arthropodes*, classe des insectes, ordre des *diptères*, sous ordre des *Nematocera*, famille des *Culicidae* et sous famille des *Anophelinae* (Carnevale, Robert, Manguin et al., 2009).

Le paludisme est transmis à l'homme par des moustiques femelles du genre Anopheles pendant leur repas sanguin pour la production d'œufs, et ces repas de sang constituent le lien entre l'homme et les hôtes moustiques dans le cycle de vie du parasite. Le développement réussi du parasite du paludisme chez le moustique (du stade « gamétocyte » au stade « sporozoïte ») (**??**) dépend de plusieurs facteurs. Le plus important est la température et l'humidité ambiantes (des températures plus élevées accélèrent la croissance du parasite chez le moustique) et la survie des anophèles assez longtemps pour permettre au parasite de terminer son cycle chez le moustique hôte (9 à 18 jours).

Actuellement 484 espèces d'anophèles sont identifiées, mais seulement une soixantaine d'entre elles sont des vecteurs du *Plasmodium* pour l'être humain, le principal vecteur en cause étant *Anopheles gambiae* sur le continent africain, où il cohabite avec *A.funestus* et *A. arabiensis*. Les anophèles ont une répartition quasiment mondiale (Figure 1.3).

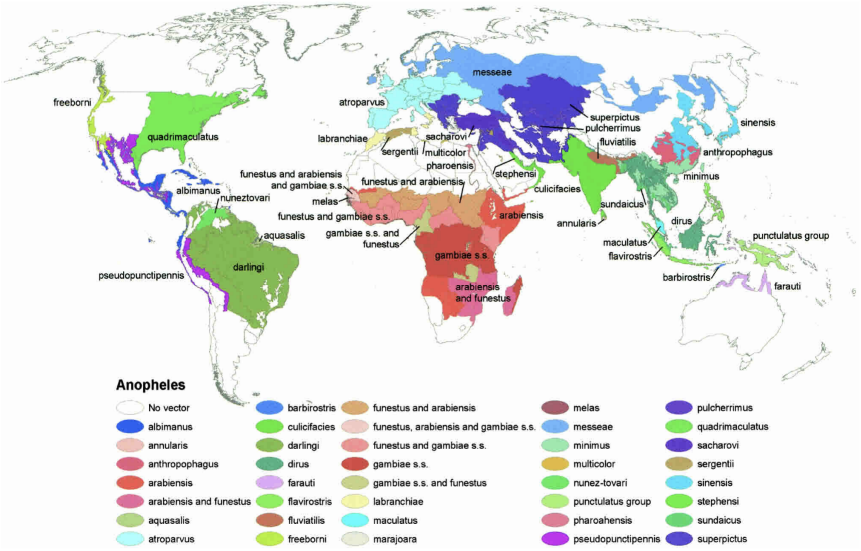


FIGURE 1.3. – Distribution mondiale (projection de Robinson) des vecteurs du paludisme dominants ou potentiellement importants (source : KISZEWSKI, MELLINGER, SPIELMAN et al., 2004).

1.1.1.5. Le parasite

Les parasites du paludisme sont des micro-organismes appartenant au genre *Plasmodium*. Il existe plus de 100 espèces de *Plasmodium* qui peuvent infecter de nombreuses espèces animales telles que les reptiles, les oiseaux et divers mammifères. On sait depuis longtemps que quatre espèces de *Plasmodium* infectent l'homme dans la nature (*P. falciparum, P. vivax, P. malariae, P. ovale*) (WHITE, PUKRITTAYAKAMEE, HIEN et al., 2014). En outre, *P. knowlesi* est une espèce infectant naturellement les à longue queue (*Macaca fascicularis*) a récemment été reconnue comme une cause du paludisme zoonotique chez l'homme (SINGH, KIM SUNG, MATUSOP et al., 2004).

— *P. falciparum* : est présent dans le monde entier dans les zones tropicales et subtropicales, et particulièrement en Afrique, où cette espèce prédomine. *P. falciparum* peut provoquer un paludisme grave car il se multiplie rapidement dans le sang et peut donc entraîner une perte de sang importante (anémie). De plus, les EI infectés peuvent obstruer les petits vaisseaux sanguins. Lorsque cela se produit dans le cerveau, il s'ensuit un paludisme cérébral, complication qui peut être fatal.

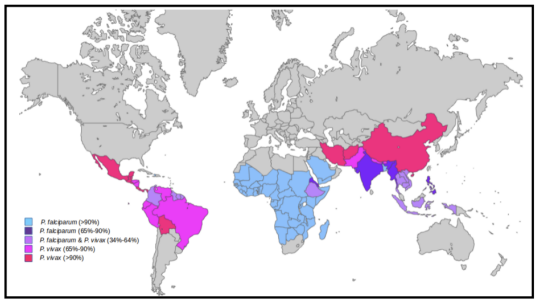
— *P. vivax* : on le trouve principalement en Asie, en Amérique latine et dans certaines régions d'Afrique. En raison de la densité de sa population, particulièrement en Asie, il s'agit probablement du parasite du paludisme humain le plus répandu. *P. vivax* (ainsi que *P. ovale*) présente des stades hépatiques en sommeil (« hypnozoïtes ») qui peuvent activer et envahir le sang (« rechute ») plusieurs mois ou années après la piqûre de moustique infectante.

— *P. ovale* se rencontre principalement en Afrique (en particulier l'Afrique de l'Ouest) et dans les îles du Pacifique occidental. Il est très similaire sur le plan biologique et morphologique à *P. vivax*. Cependant, contrairement à *P. vivax*, il peut infecter des personnes négatives pour le groupe sanguin Duffy, ce qui

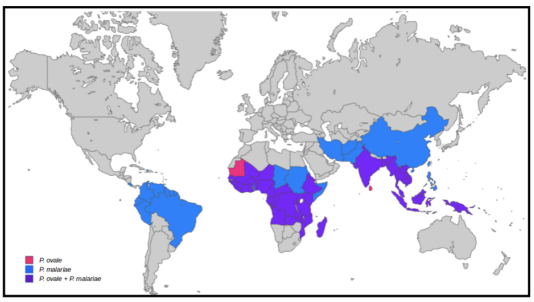
est le cas de nombreux résidents de l'Afrique subsaharienne. Ceci explique la plus grande prévalence de *P. ovale* (plutôt que *P. vivax*) dans la plupart des pays d'Afrique.

— *P. malariae*, présent dans le monde entier, est la seule espèce parasite humaine du paludisme à avoir un cycle quartan (cycle de trois jours). (Les trois autres espèces ont un cycle tertiaire de deux jours.) En l'absence de traitement, *P. malariae* provoque une infection chronique de longue durée qui, dans certains cas, peut durer toute la vie. Chez certains patients atteints d'une infection chronique, *P. malariae* peut entraîner des complications graves telles que le syndrome néphrotique.

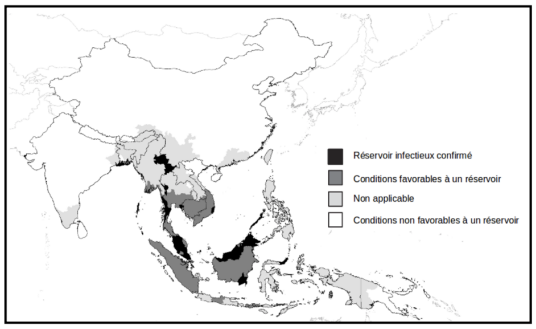
— *P. knowlesi* est présent dans toute l'Asie du Sud-Est en tant que pathogène naturel des macaques à longue queue et à queue de cochon. Il a récemment été démontré qu'il était une cause importante de paludisme zoonotique dans cette région, en particulier en Malaisie. *P. knowlesi* a un cycle de réplication de 24 heures et peut donc rapidement évoluer d'une infection simple à une infection grave ; des cas mortels ont été signalés.



Répartition géographique de *P. falciparum et P. vivax* en 2016 (OMS 2016b).



Répartition géographique de *P. ovale et P. malariae* en 2016 (Rutledge et al., 2017).



Répartition géographique de *P. knowlesi* en 2014 (Moyes et al., 2014).

FIGURE 1.4. – Repartition géographique des 5 espèces de Plasmodium infectant l'homme.

1.1.1.6. Cycle de vie du *Plasmodium*

La Figure 1.5 (ARGY et HOUZÉ, 2018) schématise le cycle de vie de *Plasmodium* chez l'hôte humain. Pour initier l'infection chez l'homme, des moustiques anophèles infectés injectent dans la peau des sporozoïtes du parasite *Plasmodium* qui transitent dans la circulation générale et, en quelques minutes (30 à 45 minutes) migrent vers le foie et infectent les hépatocytes. En l'espace de 7 à 10 jours, les parasites se développent et se divisent en mérozoïtes qui sont libérés dans le sang.

Au cours du stade sanguin d'infection par *Plasmodium*, la forme mérozoïte du parasite envahit les globules rouges (érythrocytes matures) et se réplique à l'intérieur de ceux-ci (Figure 1.5). Les cycles de réplication au stade sanguin prennent environ 48 heures pour *P. falciparum* et *P. vivax*, mais seulement 24 heures pour *P. knowlesi*. Le paludisme survient au cours de la phase sanguine de l'infection. L'apparition des gamétocytes a lieu en général la deuxième semaine qui suit l'infection et ces formes peuvent persister plusieurs semaines après la guérison.

Il est important de noter que l'érythrocyte, ne possédant pas de système de synthèse et de transport des protéines et n'exprimant pas de molécules du complexe majeur d'histocompatibilité (CMH) de classe I ou II à sa surface, est un refuge idéal pour un parasite qui doit perdurer de longues périodes chez son hôte, afin d'être transmis au moustique.

A la suite d'une nouvelle piqûre par un Anophèle, les gamétocytes mâles et femelles (au dimorphisme sexuel marqué) sont ingérés avec le repas sanguin. L'anophèle femelle ingère des gamétocytes lors d'un repas sanguin sur un individu infecté. Ceux-ci migrent vers l'estomac du moustique et se transforment en gamètes. Par un processus d'exflagellation du gamète mâle, les gamètes femelles sont fécondés et il en résulte un zygote appelé ookinète. Celui-ci s'implante sous la paroi stomacale en formant l'oocyste. Après une division méiotique suivie par plusieurs mitoses, les sporozoïtes sont générés. Ils se libèrent après éclatement de l'oocyste pour se concentrer au niveau des glandes salivaires en attendant la prochaine piqûre infectante (Argy et Houzé, 2018).

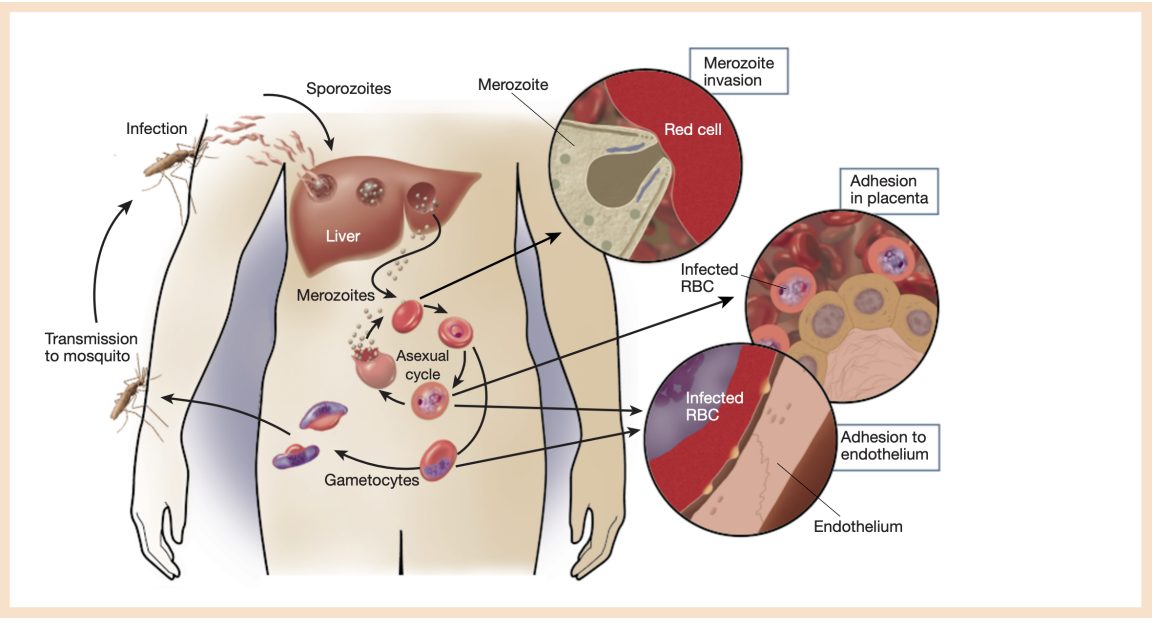


Figure 1.5. – Cycle de vie des parasites et pathogenèse du paludisme à *falciparum*. Les globules rouges infectés par *P. falciparum* peuvent se lier à l'endothélium ou au placenta (Source : Miller, Baruch, Marsh et al., 2002).

Le cycle de vie de *Plasmodium* se déroule entre 10 et 40 jours, suivant la température extérieure et les espèces en cause (Gentilini, Danis et Mouchet,

1990). La compréhension des événements moléculaires et cellulaires impliqués dans l'invasion par les mérozoïtes et la réponse immunitaire de l'hôte aux antigènes des mérozoïtes sont essentielles au développement de vaccins et de nouveaux traitements.

1.1.1.7. Pathogenèse du paludisme

Chez l'homme le parasite a besoin de proliférer et de survivre en échappant au système immunitaire de l'hôte. Pour ce faire, *P. falciparum* utilise différentes techniques afin de contourner le système immunitaire de l'hôte. Les techniques les mieux connues incluent notamment la cytoadhérence et la séquestration, la formation de rosettes et la variation antigénique.

— **La cytoadhérence**

P. falciparum est responsable de la séquestration des EI. La surface des EI est recouverte de protubérances appelées *knobs* qui sont le point de contact avec les cellules de l'hôte. L'adhésion protège EI de la destruction car les EI circulants sont éliminés dans la rate. Plusieurs protéines parasitaires sont localisées dans ces protubérances et participent directement ou indirectement à la cytoadhérence.

La pathogenèse du paludisme dépend de la capacité d'adhésion du *Plasmodium* à la surface endothéliale de l'hôte. La cytoadhérence est médiée par la protéine parasitaire nommée *P. falciparum-erythrocyte membrane protein-1 (PfEMP-1)* qui reconnait les récepteurs à la surface des cellules endothéliales. (Figure 1.6A) (Fried et Duffy, 1996, Beeson, Rogerson, Cooke et al., 2000, Sherman, Crandall, Guthrie et al., 1995) La protéine *PfEMP-1* appartient à la famille des gènes variables *Var* (Rowe et Kyes, 2004) dont on distingue 60 copies par génome haploïde. *PfEMP-1* est composée de trois parties dont un segments extra cellulaire N-terminale variable, un segment transmembranaire, un segment intracellulaire C-terminal conservé, de plusieurs domaines *Duffy Binding like (DBL)* et des régions *inter-domaines riches en cystéine* (CIDR) (Smith, Subramanian, Gamain et al., 2000).

La cytoadhérence des EI implique la liaison de *PfEMP1* aux molécules d'adhésion exprimées à la surface des cellules endothéliales, telles qu'ICAM-1, VCAM-1, CD36 parmi les six récepteurs qui ont été mis en évidence : la thrombospondine, CD36, ICAM-1 , VCAM-1 (vascular cell adhesion molecule 1), E-sélectine et la chondroïtine sulfate A (CSA) (Fried et Duffy, 1996, Warrell, 1997). Seuls ICAM-1 et E-sélectine sont exprimés plus massivement dans l'accès pernicieux, probablement sous l'effet des cytokines, et seraient ainsi partiellement responsables de la sévérité du paludisme (Muanza, Traoré, Gay et al., 1999, Newbold, Warn, Black et al., 1997).

Différents phénotypes du paludisme dépendent d'adhésions de *PfEMP-1*. Le paludisme cérébral est définie par l'adhésion de *PfEMP-1* sur le récepteur *inter cellular adhesion molecule-1 (ICAM-1)* (Figure 1.6A). Les accès palustres graves sont déterminés par l'adhésion des *PfEMP-1* au récepteur *cluster of differentiation 36 (CD36)* (Figure 1.6A) dans les capillaires viscéraux et cérébraux. Le paludisme placentaire à une particularité définie par l'adhésion du *PfEMP-1* à la CSA (Fried et Duffy, 1996) (Figure 1.6B).

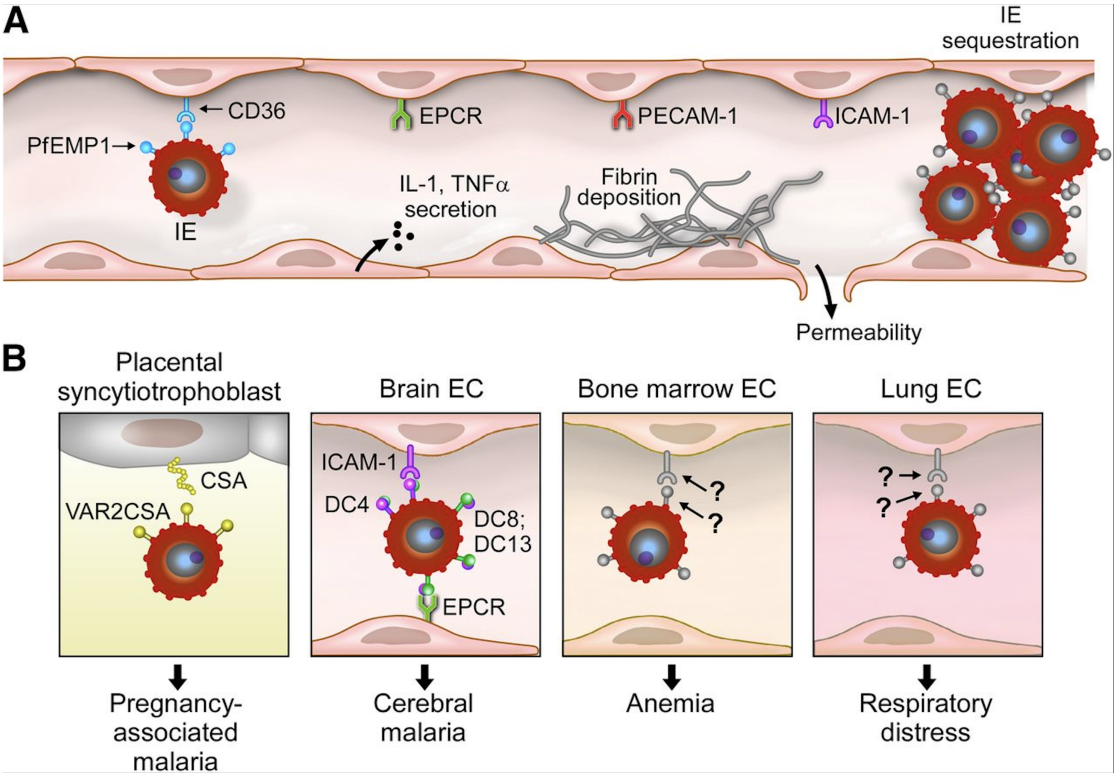


FIGURE 1.6. – Interactions entre antigène *PfEMP1* et récepteurs endothéliaux intervenant dans la séquestration spécifique de la paroi microvasculaire des EI. (A) Micro-vaisseau typique trouvé dans divers organes et tissus chez des patients atteints de paludisme. (B) Séquestration des érythrocytes infectés (EI) dans quatre organes différents. Le paludisme associé à la grossesse (PAG) est un syndrome spécifique à un organe initié par l'expression du variant de *PfEMP1*, VAR2CSA, qui assure la liaison de l'EI au CSA placentaire exprimé par les syncytiotrophoblastes (William C. et al, 2013).

Le paludisme placentaire à *P. falciparum*, caractérisé par la séquestration tissulaire spécifique des EI dans l'espace intervilleux, est reconnu depuis longtemps (CLARK, 1915). Cependant, la concentration marquée de paludisme placentaire chez les primigravidés est restée inexpliquée jusqu'à ce que l'on se rende compte que les EI dans le placenta ont un phénotype d'adhésion très distinct qui n'est pas compatible avec la survie du parasite chez un hôte non enceinte (HVIID, 2011). Il est maintenant largement admis que l'accumulation sélective d'EI dans le placenta est médiée par *PfEMP1* de type VAR2CSA avec affinité pour la CSA (FRIED et DUFFY, 1996) (Figure 1.7) et non au CD36 et ICAM-1 dans des organes autres que le placenta (FRIED et DUFFY, 1996, BEESON, ROGERSON, COOKE et al., 2000) (Figure 1.6). Suite à l'identification de la CSA en tant que récepteur unique qui favorise l'adhésion des parasites au placenta (Figure 1.7), des études supplémentaires menées dans différents sites ont confirmé ce phénotype de liaison (BEESON, ROGERSON, COOKE et al., 2000, DAHLBÄCK, NIELSEN et SALANTI, 2010, FRIED, NOSTEN, BROCKMAN et al., 1998 ;VAINBERG, TAP, LEWIS et al., 1995).

L'infection placentaire contribue à perturber les échanges materno-fœtal et, de ce

fait, à diminuer le poids de l'enfant à la naissance, source indirecte d'une morbidité et d'une mortalité néonatale élevée dans les pays en développement.

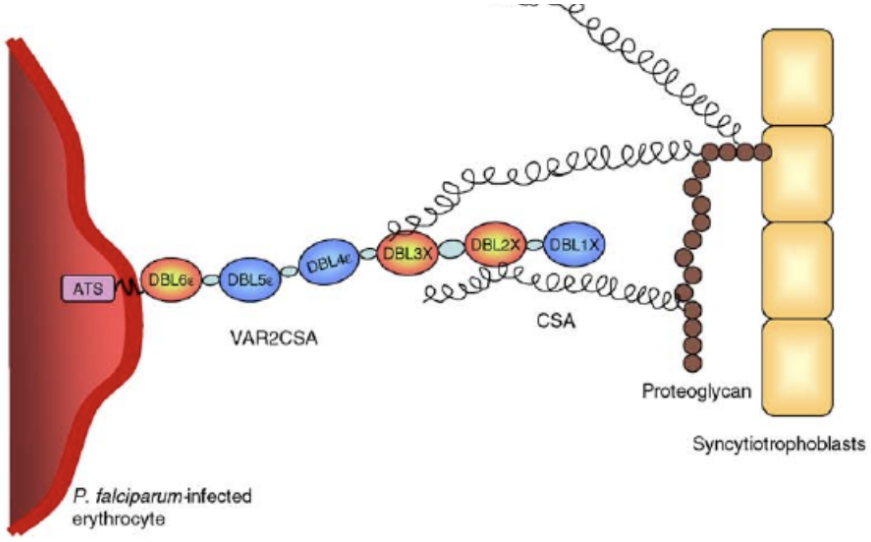


FIGURE 1.7. – Interaction chondroïtine Sulfate A et VAR2SCA.

— **La rosetting**

La formation de rosettes entraînerait une accumulation de globules rouges dans la microvasculature. Il s'agit de la capacité de certaines érythrocytes saines à s'allier avec 1-2 EIs via les héparanes sulfates, le récepteur CR1 (récepteur du complément), mais aussi à la liaison des protéines parasitaires STEVOR et RIFIN (à la surface des EI) à la glycophorine C et l'antigène du groupe sanguin A respectivement (à la surfaces des érythrocytes sains). La formation de rosette créée ainsi un obstacle au flux sanguin (Figure 1.6 A). Même si ce phénomène a été observé avec certains isolats provenant de cas de neuropaludisme (CARLSON, HELMBY, HILL et al., 1990) il ne représente pas une condition nécessaire puisque de nombreux malades graves peuvent être infectés avec des souches qui ne possèdent pas ces capacités (YAMAN, GENTON, MOKELA et al., 1995). A l'inverse, des parasites prélevés sur des personnes asymptomatiques peuvent former de nombreuses rosettes in vitro sans entraîner de problèmes chez les porteurs.

1.1.1.8. Variabilité antigénique du *Plasmodium*

La protéine *PfEMP1* est impliquée dans la variation antigénique du paludisme à *P. falciparum*. Avec environ 60 gènes var codant pour *PfEMP1* et un seul gène var dominant exprimé au stade adulte du parasite, *PfEMP1* a atteint une forme de variabilité qui permet au parasite de contourner le système immunitaire de l'hôte. De plus, les fréquentes recombinaisons et remaniements génétiques au cours des processus de fusion et de division dans le moustique et les érythrocytes humains peuvent entraîner une grande diversité génétique et antigénique du parasite (FLICK, AHUJA, CHENE et al., 2004).

— **Le génome du *Plasmodium***

Le génome du *Plasmodium* (clone 3D7) a été entièrement séquencé en 2002 (GARDNER, HALL, FUNG et al., 2002). Il compte 14 chromosomes de tailles variantes

entre 0,7 et 3,3 Mb totalisant 22,8 Mb. On compte 5314 gènes avec une densité moyenne de 1 gène par 4338 pb. En plus de l'ADN nucléaire, le *Plasmodium* contient un fragment mitochondrial de 6 kb et un fragment circulaire de 35 kb dans l'apicoplaste (Gardner, Hall, Fung et al., 2002). Le séquençage du génome de *Plasmodium* a permis des avancées dans la recherche des nouveaux médicaments contre le paludisme, sur l'analyse protéomique, sur le développement des nouveaux outils de suivi d'évaluation des activités de lutte contre le paludisme et dans la recherche vaccinale.

On distingue plusieurs antigènes du *P. falciparum* codés par des gènes en copie unique comme les *merozoïte surface proteine* (MSP1, MSP2, MSP3), *Glutamate-rich protein* (GLURP) (Gosi, Lanteri, Tyner et al., 2013, Amoah, Nuvor, Obboh et al., 2017) et bien d'autres. Bon nombre de ces protéines de surface des mérozoïtes sont en cours d'évaluation en tant que composants potentiels d'un vaccin contre *P. falciparum*.

Nous avons utilisé le marqueur MSP2 pour la caractérisation par PCR des infections de *P. falciparum* aux niveaux du sang périphérique maternel, du sang placentaire et du sang du cordon ombilical.

— **La protéine de surface du mérozoïte 2**

Plusieurs protéines de surface des mérozoïtes sont en cours d'évaluation en tant que composants potentiels d'un vaccin contre *P. falciparum*. Une de ces protéines, la protéine de surface mérozoïte 2 (MSP2), est particulièrement hydrophile et contient des répétitions de séquences en tandem, caractéristiques des protéines intrinsèquement non-structurées. Une série d'études physico-chimiques ont confirmé que les formes recombinantes de MSP2 sont en grande partie non structurées. MSP2 est une protéine de 45 Kd codée par un gène localisé sur le chromosome 2. Le gène est divisé en cinq blocks. Les allèles du gène MSP2 sont groupés en 2 familles principales 3D7 et FC27 (Figure 1.8), correspondant à la structure dimorphique de la région centrale variable (Block 3) (Kanunfre, Leoratti, Hoffmann et al., 2003).

La protéine MSP2 de *P. falciparum* est considérée comme un bon candidat pour l'inclusion dans un vaccin antipaludique. Plusieurs études ont rapporté la relation entre la protection et les réponses immunitaires humorales aux antigènes MSP2 (Taylor, Smith, Robinson et al., 1995). De plus, le gène de la protéine MSP2 est également couramment utilisé comme marqueur unique pour la caractérisation moléculaire des parasites du paludisme de terrain (Ntoumi, Contamin, Rogier et al., 1995). Le gène MSP2 est représenté en un seul exemplaire sur le génome de *P. falciparum* (Smythe, Peterson, Coppel et al., 1990) et un degré élevé de polymorphisme a été rapporté dans sa région variable centrale (Smythe, Peterson, Coppel et al., 1990). Le typage de ces différentes régions polymorphes du génome de *P. falciparum* a permis de déterminer des indicateurs d'infection par le paludisme, par exemple la diversité des souches de *P. falciparum* et la multiplicité d'infection (*Multiplicity of Infections,* **MOI**), qui peuvent contribuer à la description de la situation du paludisme dans un lieu donné.

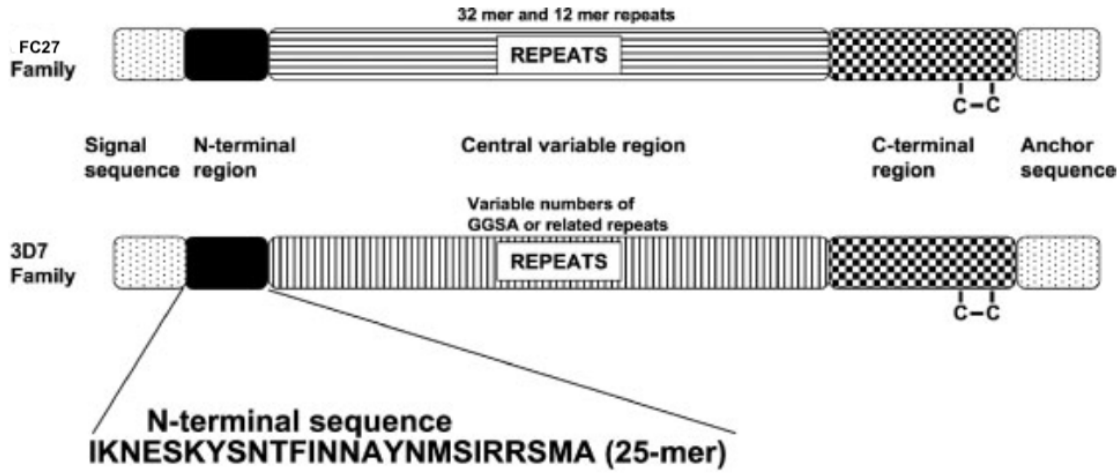


FIGURE 1.8. – Schéma de principe des deux familles alléliques du gène *meroite surface protein 2* (MSP2), FC27 et 3D7. Les boîtes avec des points fins représentent les régions avant la région N-terminale (séquence signal) et après la région C-terminale (séquence ancre) ; la région conservée N-terminal est représentée par des cases noires. Les zones de lignes horizontales et verticales désignent les régions à variation centrale pour les familles FC27 et 3D7, respectivement. La principale différence entre les deux familles se situe dans cette région où le FC27 a des répétitions de 12 et 32 mères, alors que 3D7 a des répétitions de séquence GGSA ; la région conservée C-terminale est représentée par des carrés noirs et blancs. Les deux C sont la liaison disulfure dans la région C-terminale (adapté de Andrew Low et al., 2007).

1.1.1.9. Signes cliniques et symptômes

Le paludisme peut se manifester sous des formes variées, allant de l'infection asymptomatique (porteur de parasites sans signe clinique) à des formes graves pouvant conduire au décès de l'individu. Au sein de cette variabilité, l'accès palustre simple constitue la majorité des cas diagnostiqués. Celui-ci est caractérisé par des épisodes de fièvres aigus accompagnés de symptômes plus ou moins marqués (vomissement, maux de tête, frissons,...). Les accès palustres graves affectent principalement les enfants de moins de 5 ans n'ayant pas encore développé une immunité protectrice contre le parasite. Ceux-ci entraînent des manifestations sévères souvent mortelles (anémie sévère, détresse respiratoire ou paludisme cérébral) nécessitant une prise en charge hospitalière d'urgence.

Dans les zones où le paludisme est fortement endémique, les groupes les plus à risque sont les jeunes enfants (susceptibles de développer un paludisme sévère), ainsi que les femmes enceintes, susceptibles de souffrir d'anémie et d'accoucher d'un nouveau-né de faible poids à la naissance. Dans les zones où le paludisme est transmis tout au long de l'année, les enfants plus âgés et les adultes développent une immunité partielle après des infections répétées et présentent un risque relativement faible de contracter un paludisme sévère.

Les voyageurs se rendant dans des zones touchées par la malaria n'ont généralement jamais été exposés aux parasites du paludisme ou ont perdu leur immunité s'ils quittaient la zone d'endémie ; ils présentent un risque très élevé de contracter un paludisme sévère s'ils sont infectés par *P. falciparum* (Wilson, Weld, Boggild et al., 2007).

Dans la plupart des cas, la période d'incubation de l'infection à *P. falciparum* est d'environ 12 à 14 jours (extrêmes : 7 à 30 jours). La plupart des infections à *P. falciparum* deviennent cliniquement évidentes un mois après l'exposition[7]. Des périodes d'incubation plus longues sont plus susceptibles chez les personnes semi-immunes et les personnes prenant une prophylaxie antipaludique inefficace.

— **Paludisme simple**

Les symptômes cliniques les plus fréquents du paludisme non compliqué à *P. falciparum* sont : une fièvre élevée accompagnée d'une fatigue intense, d'une sensation de malaise général avec frissons, des douleurs musculaires et squelettiques, des céphalées, des troubles digestifs (nausées, vomissements, diarrhées, douleurs abdominales) suivis de sueurs (Grobusch et Kremsner, 2005). La survenue de la fièvre est caractérisée par la présence du parasite dans la circulation sanguine suite à la lyses des EI. Cette lyse libère, non seulement des mérozoïtes, mais aussi des exoantigènes parasitaires tels que l'hémozoïne et le glycosylphosphatidylinositol (GPI). Classiquement (mais rarement observées), les attaques se produisent tous les deux jours avec *P. falciparum, P. vivax et P. ovale* et tous les trois jours avec *P. malariae*.

— **Paludisme sévère**

L'infection plasmodiale peut évoluer vers un paludisme sévère qui peut entraîner la mort. La majorité des formes graves du paludisme sont dues essentiellement à *P. falciparum* et dans dans les cas rares à *P. vivax* (Barber, William, Grigg et al., 2013).

Le paludisme sévère résulte d'une combinaison de facteurs spécifiques du parasite : adhérence, séquestration dans le système vasculaire et libération de molécules bioactives ; mais aussi de la réponse inflammatoire de l'hôte avec notamment la production de cytokines, chémokines et d'infiltrats cellulaires (Mackintosh, Beeson et Marsh, 2004, Dunst, Kamena et Matuschewski, 2017).

Deux syndromes majeurs caractérisent le paludisme sévère : l'anémie sévère et le paludisme cérébral. La pathogénie du paludisme cérébral est attribuée à la cytoadhérence des EIs aux cellules endothéliales dans les micros vaisseaux du cerveau entraînant l'obstruction puis la rupture des vaisseaux sanguins. Un des mécanismes de la pathogenèse du paludisme grave est l'acidose métabolique qui entraîne la détresse respiratoire. L'acidose métabolique est un bon prédicteur de la mortalité due au paludisme sévère (Taylor, Borgstein et Molyneux, 1993, Marsh, Forster, Waruiru et al., 1995). La pathogénie de l'acidose métabolique est mal connue.

Les cytokines pro-inflammatoires jouent un rôle central dans les manifestations cliniques du paludisme cérébral. L'interleukine 10 (IL-10) aurait un rôle protecteur dans le paludisme cérébral (Clark et Cowden, 2003). Par contre, une cytokine comme le facteur de nécrose tumorale alpha (TNF-alpha) est plutôt associée à la sensibilité au paludisme cérébral (Grau, Taylor, Molyneux et al., 1989, Brown, Turner, Rogerson et al., 1999). Le rôle de l'interféron gamma (IFN-gamma)

est ambigu, elle est associée à la fois à la pathogenèse et à la protection contre le paludisme chez l'homme (Day, Hien, Schollaardt et al., 1999). Un autre mécanisme impliqué dans la pathogenèse du paludisme cérébral est l'induction du monoxyde d'azote (NO), il serait responsable non seulement d'une hypotension intracrânienne due à la dilatation excessive des vaisseaux sanguins (vasodilatation) mais aussi à une neurotransmission aberrante (Clark et Rockett, 1994).

1.1.2. Paludisme et Grossesse

Plus de 50 millions de femmes sont exposées au risque de paludisme pendant la grossesse chaque année. Le paludisme associé à la grossesse (PAG) entraîne une augmentation substantielle de la mortalité maternelle et surtout fœtale et la morbidité infantile, causant entre 75 000 et 200 000 décès infantiles chaque année (Steketee, Nahlen, Parise et al., 2001, Desai, Kuile, Nosten et al., 2007). On estime que chaque année, des centaines de milliers de femmes meurent pendant la grossesse du paludisme sévère et d'anémie liée à l'infection palustre.

Dans les zones de forte transmission, les femmes enceintes, en particulier les primigestes (Schantz-Dunn et Nour, 2009), sont plus susceptibles à l'infection palustre que les multigestes (Mayengue, Rieth, Khattab et al., 2004, Brabin, 1983) et leurs homologues non enceintes (Brabin, 1983, McGregor, 1984). Les femmes multigestes acquièrent naturellement une résistance à *P. falciparum* qui réduit la densité parasitaire et sont souvent asymptomatiques. Toutefois, les plasmodies peuvent être présentes dans le placenta et contribuer à une anémie maternelle même en l'absence d'une parasitémie périphérique avérée. Une anémie maternelle, comme une parasitémie placentaire, peut être responsable d'un faible poids de naissance (Figure 1.9), un facteur important dans la mortalité infantile (Davison, Cogswell, Baskin et al., 2000).

En zone de faible transmission, les femmes de toutes les parités (nombre de grossesses ayant atteint l'âge de viabilité de 24 semaines) ont une immunité acquise relativement faible contre le paludisme (Nosten, Kuile, Maelankirri et al., 1991). Le manquent d'une forte immunité systémique entraîne une susceptibilité de développer des syndromes graves comme la détresse respiratoire et le paludisme cérébral (Nosten, Kuile, Maelankirri et al., 1991). Les infections plasmodiales entraînent les accès palustre graves, une anémie sévère, des fausses couches spontanées et un risque élevé de faible poids de naissance (Desai, Kuile, Nosten et al., 2007, Nosten, McGready, Simpson et al., 1999, Bourée et Bisaro, 2007).

Les femmes enceintes manquent d'une forte immunité systémique à cause de la production des cytokines dites régulatrices ou anti-inflammatoires et la mise en place de l'état de tolérance materno-foetale. En effet, l'immuno-modulation pendant la grossesse, en particulier pendant les deuxième et troisième trimestres est favorisé par la production de certaines hormones ou substance comme le cortisol et la prolactine qui sont favorisent la migration des lymphocytes T régulatrices (LTreg). Ces LTreg vont produire les cytokines anti-inflammatoires à savoir interlettrage 10 (IL10) et le facteur de croissance transformant (TGF-beta) qui vont après inhiber les lymphocytes T activés (Hanssen, Collinet, Leblanc et al., 2013). La tolérance foeto-maternelle représente l'unique exemple non pathologique de tolérance naturelle induite par une greffe semi-allogénique. Cette tolérance est rendu

possible par l'expression des molécules HLA-G par les cellules cytotrophoblastiques, empêchant ainsi leur lyse par les cellules *natural killer* (NK) maternelles qui infiltrent l'endomètre. Aussi, la molécule HLA-G est loisible à l'embryon de na pas être rejeté (Carosella, Paul, Moreau et al., 2000).

Les perturbations placentaires peuvent se traduire par des lésions histologiques et une obstruction des microvillosités placentaires empêchant ainsi les échanges entre la mère et le fœtus.

La baisse du risque du paludisme gestationnel avec le nombre de grossesses s'explique par le fait qu'au cours de leur première grossesse et en présence de l'infection plasmodiale, les femmes ont un déficit en anticorps anti-CSA (ou anti-VAR2CSA). Et lors des prochaines grossesses, avec la mémoire immunitaire, la production des anticorps anti-CSA se fait de manière exponentielle, empêchant ainsi l'adhérence des parasites au niveau du placenta (Fried, Nosten, Brockman et al., 1998).

Le paludisme est la seule maladie infectieuse qui à un tropisme essentiellement placentaire et les conséquences fœtales seront plus liées au degré d'altération du placenta qu'à l'atteinte du fœtus lui-même (paludisme congénital). En Afrique sub-saharienne, le paludisme est presque exclusivement dû à *P. falciparum*, mais dans le reste du monde (beaucoup plus peuplé) *P. vivax* coexiste avec *P. falciparum*. Les effets de l'infection à *P. vivax* pendant la grossesse sont moins sévères, mais associent l'anémie maternelle et la réduction du poids de naissance (Nosten, McGready, Simpson et al., 1999). Les accès fébriles dus à *P. falciparum* et *P. vivax* sont responsables d'accouchements prématurés (Figure 1.9) et d'interruptions de grossesses. Le paludisme congénital est peu fréquent mais pas exceptionnel.

Le paludisme est une cause importante de morbidité et de mortalité chez les femmes enceintes infectées par le VIH dans les régions endémiques de l'Afrique sub-saharienne, où au moins 1 million de co-infections du paludisme surviennent chaque année chez les femmes enceintes (Uneke et Ogbonna, 2009). Les femmes enceintes infectées par le VIH et atteintes de paludisme clinique courent un risque accru d'accouchement prématuré, d'anémie sévère, de faible poids de naissance et de décès[8]. La co-infection VIH-paludisme chez les femmes enceintes favorise l'infestation placentaire, le passage trans-placentaire des parasites et cause près de 20% de tous les décès par an en Afrique sub-saharienne.

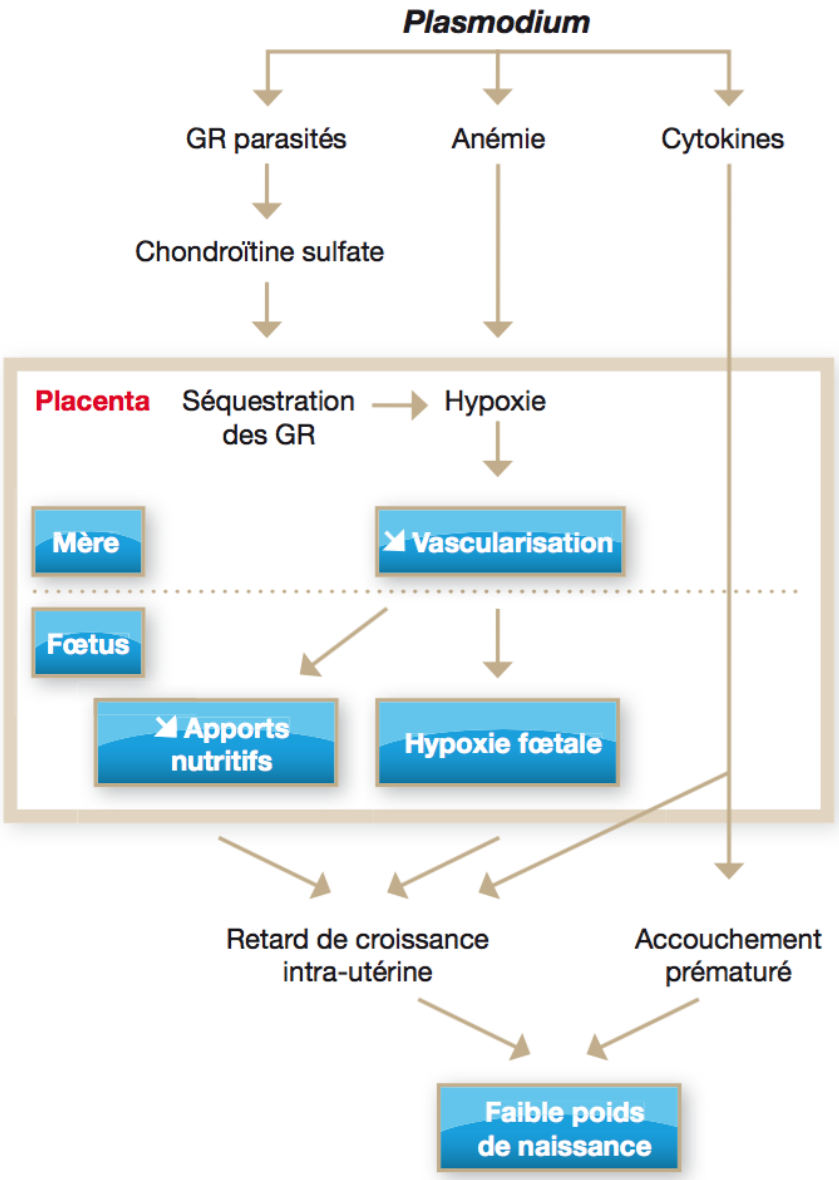


FIGURE 1.9. – Paludisme placentaire lié au faible poids de naissance (source : adapté de BOURÉE, 2008).

1.1.3. Paludisme asymptomatique chez la femme enceinte

Le paludisme asymptomatique est actuellement reconnu comme potentielle menace et frein au contrôle du paludisme. L'infection asymptomatique à *P. falciparum* est très répandue dans les régions d'endémie palustre (Figure 1.10) (LANGHORNE, NDUNGU, SPONAAS et al., 2008), ce qui suscite de vives inquiétudes quant aux efforts déployés pour éliminer le parasite (RILEY, WAGNER, AKANMORI et al., 2001). La plupart des études ont suggéré que les personnes exposées étaient probablement porteuses d'infections asymptomatiques chroniques à *P. falciparum* (STURROCK, HSIANG, COHEN et al., 2013 ;MBOUAMBOUA, KOUKOUIKILA-KOUSSOUNDA, NTOUMI et al., 2019), une situation dans laquelle la majorité de cette population

peut servir de réservoir de transmission au paludisme (Francine, Damien, Anna et al., 2016 ; Mbouamboua, Koukouikila-Koussounda, Ntoumi et al., 2019).

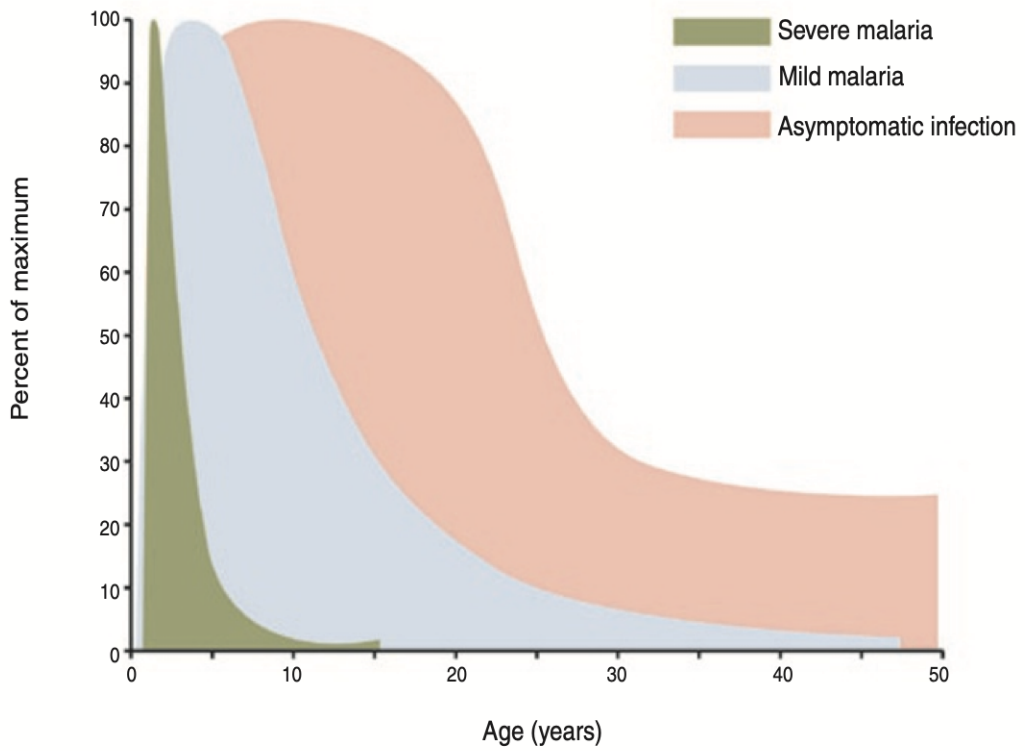


Figure 1.10. – Évolution dans le temps des indices de population en fonction de l'immunité dans une zone d'endémie palustre. Infection asymptomatique (rose), paludisme simple (bleu) et paludisme sévère (vert). (Source : Langhorne, Ndungu, Sponaas et al., 2008).

Les symptômes et les complications du paludisme au cours de la grossesse diffèrent selon l'intensité de la transmission et le taux d'immunité acquise par la femme enceinte. Dans des zones de faible transmission ou instable du paludisme, les femmes enceintes n'ont pas acquis un taux d'immunité élevé et tombent généralement malades lorsqu'elles sont infectées par *P. falciparum*. Dans des zones de transmission stable du paludisme, la plupart des femmes adultes ont développé une immunité suffisante pour que, même pendant la grossesse, l'infection à *P. falciparum* n'entraîne généralement ni fièvre ni autre symptôme clinique (Figure 1.11). Dans ces zones, l'infection palustre se caractérise principalement par le déclenchement d'une anémie secondaire et par la présence de parasites dans le placenta. Les carences nutritives qui en résultent pour le fœtus contribuent à un faible poids à la naissance (OMS, 2005).

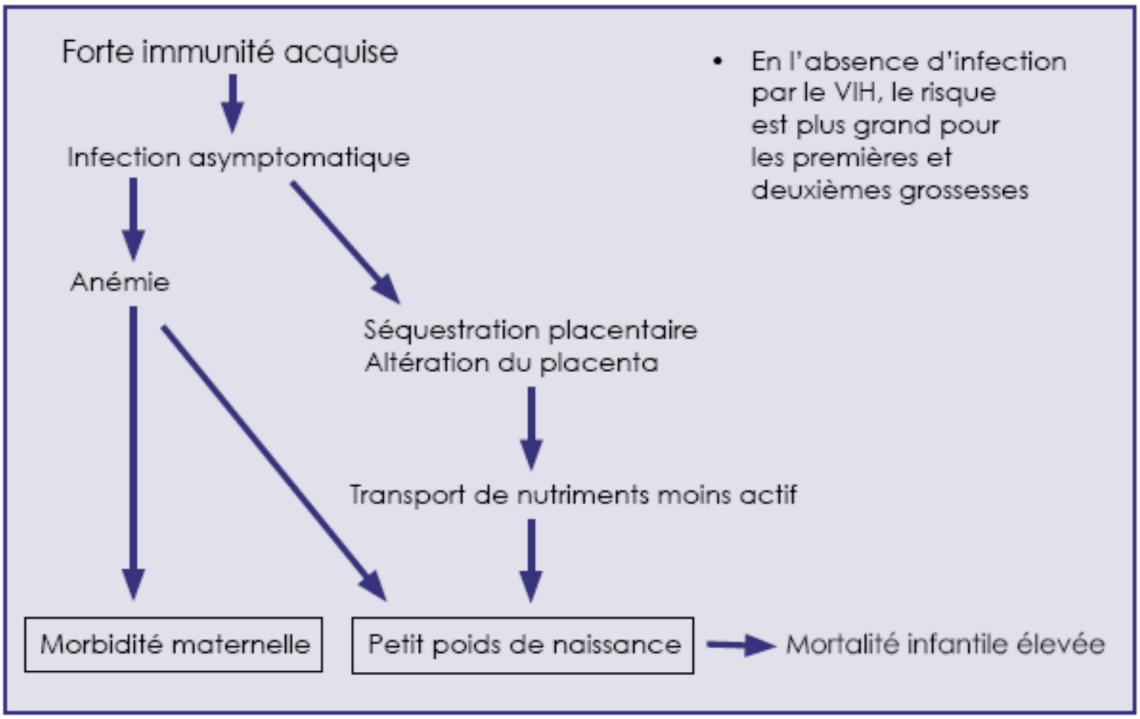


FIGURE 1.11. – Conséquences du paludisme pendant la grossesse : zone de transmission forte ou modérée (stable) (vert). (Source : OMS 2005).

1.1.4. Diagnostic du paludisme

Une prise en charge efficace de la maladie requiert un diagnostic posé sans délai. Le diagnostic repose sur la suspicion clinique d'un paludisme et, la recherche des hématozoaires par l'examen microscopique certifie ce diagnostic en mettant en évidence le parasite dans le sang circulant.

L'infection à *P. falciparum* peut-être difficile à diagnostiquer pendant la grossesse, en particulier chez les femmes semi-immunes qui sont souvent asymptomatiques pendant l'infection. Bien que les EIs s'accumulent dans le placenta, la densité parasitaire dans le sang périphérique peut être trop faible pour être détectée par la goutte épaisse (GE). Le frottis sanguin (FS) est la référence en matière de diagnostic du paludisme et est idéale pour distinguer les différentes espèces plasmodiales. Cependant, la qualité varie considérablement et les exigences en matière de microscope et de microscopiste qualifié limitent la disponibilité ou la qualité des FS dans de nombreux endroits.

1.1.4.1. Diagnostic microscopique direct par goutte épaisse(GE) et frottis sanguin (FS)

La microscopie est la technique de référence en matière de recherche sur le paludisme et reste un diagnostic sur le lieu de traitement en milieu clinique et épidémiologique. La méthode consiste à examiner au microscope optique des gouttes de sang après hémolyse des EI tapissés sur une lame en verre et leur coloration avec du Giemsa. Cependant, la sensibilité et de fiabilité de la GE dépendent d'une

bonne expérience du microscopiste du le niveau de parasitémie du sujet infecté. L'identification morphologique des parasites et des hématies est faite par le FS dont le seuil de détection est de parasites 100/µL pour 200 champs de lecture. Le seuil de détection de la GE pour une lecture de 100 champs est de 10 parasites/µL de sang[9] . Le diagnostic microscopic ne permet pas de détecter une faible parasitémie et des *Plasmodiums* altérés par un traitement présomptif (Moody, 2002). La parasitémie est souvent indétectable chez les femmes enceintes du fait de la séquestration placentaire.

1.1.4.2. Les tests de diagnostic rapide (TDR)

Les progrès technologiques ont permis le développement des TDR du paludisme. Le TDR est la détection d'antigènes plasmodiaux par immunochromatographie. Plusieurs techniques sont disponibles actuellement : Parasight® (Becton Dickinson), ICT Malaria P.f. Test (ICT Diagnostics), OptiMal®. Elles sont basées sur la mise en évidence soit de *Histidine Rich Protein 2* (HRP2), spécifique de *P. falciparum* (Parasight® et ICT Malaria®), soit de la Lactate Déshydrogénase Parasitaire, LDH (OptiMal®) ou aldolase (LDHp) communes aux 5 espèces de *Plasmodium* et la *LDHpv* spécifique de *P. vivax* (Pieroni, Mills, Ohrt et al., 1998). Les TDR sont faciles à utiliser et peuvent rendre un résultat positif ou négatif en 15 minutes.

L'évaluation de ces test par l'OMS pour des parasitémies de l'ordre de 200 parasites/ul et de 2000 parasites/ul. Une vingtaine de ces tests ont une sensibilité de l'ordre de 100% pour les parasitémies de 200 parasites/µl. En général, les tests basés sur la HRP2 ont une sensibilité plus élevée (ils identifient davantage de cas réels de paludisme) et moins spécifiques (ils rendent davantage de résultats faussement positifs) que les tests basés sur la LDHp (Abba, Deeks, Olliaro et al., 2011).

Les TDR peuvent être très précis par rapport à la microscopie et à la méthode d'amplification en chaîne par polymérase (*polymerase chain reaction*, PCR) (Abba, Deeks, Olliaro et al., 2011). Cependant, même les TDR d'excellente sensibilité peuvent s'avérer moins sensibles dans le cas du paludisme gestationnel de faible parasitémie indétectable par goutte épaisse ou frottis sanguin. Plusieurs études ont comparé les performances des TDR détectant l'enzyme HRP2 ou la LDHp à d'autres méthodes telles que la GE périphérique, la GE placentaire, l'histologie placentaire ou la PCR de sang périphérique et placentaire. Les résultats de ces études montrent que la sensibilité des TDR basés sur la HRP2 à l'aide d'échantillons de sang périphérique est beaucoup plus faible que la détection par PCR des acides nucléiques du parasite dans le sang périphérique ou placentaire (Fried, Muehlenbachs et Duffy, 2012) (voir annexe 1).

1.1.4.3. Détection de parasites par amplification d'ADN et infection sous-microscopique

— **Les méthodes de *polymerase chain reaction* (PCR)**

Les techniques de PCR de détection de l'infection plasmodiale ont été décrites il y a plus de deux décennies (Barker, Banchongaksorn, Courval et al., 1992, Snounou, Viriyakosol, Zhu et al., 1993). Elles reposent sur la détection de l'ADN parasitaire, et sont généralement plus sensibles que la GE (Barker,

Banchongaksorn, Courval et al., 1992, Snounou, Viriyakosol, Zhu et al., 1993).

La PCR est une réaction enzymatique qui permet de sélectionner puis d'amplifier en une très grande quantité, un fragment d'ADN particulier, présent en très faible quantité au départ, parmi des millions d'autres fragments. La PCR est basée sur le mécanisme de réplication de l'ADN *in vivo* : l'ADN bicaténaire est déroulé en ADN monocaténaire, puis dupliqué et ré-enroulé, selon des cycles répétitifs comprenant les trois étapes suivantes :

— Dénaturation de l'ADN par fusion à haute température pour convertir l'ADN bicaténaire en ADN monocaténaire. Cette étape est réalisée à une température comprise entre 93 et 96°C.
— Hybridation à l'ADN cible de deux oligonucléotides utilisés comme amorces. Cette hybridation a lieu à une température comprise entre 55 et 65°C.
— Extension de la chaine d'ADN par addition de nucléotides à partir des amorces en utilisant l'ADN polymérase comme catalyseur en présence d'ions Mg2+. La température optimale de travail de l'ADN polymérase est de 72°C.

Lorsque la quantité de produits d'amplification est suffisante, ceux-ci sont soumis à une électrophorèse en gel d'agarose. Cette électrophorèse permet de faire migrer les acides nucléiques au travers du gel (d'agarose ou de polyacrilamide) additionné de BEt (Bromure d'éthydium : produit intercalant qui se glisse entre les bases des acides nucléiques faisant apparaitre à la molécule d'ADN une fluorescence orange sous illumination par des UV courts (environ 30 nm)). La vitesse de migration étant dépendant de la masse de la molécule, donc du nombre de bases de l'ADN testé, la présence et la taille des amplicons peuvent être facilement vérifiable sur le gel.

Une technique a été développée pour lutter contre les amplifications parasites et s'assurer que la bande observée correspond bien à la séquence recherchée. Cette technique appelée *nested PCR* (PCR nichée), consiste à réaliser deux PCR successives en utilisant des couples d'amorces différents, le deuxième couple d'amorces encadrant une séquence incluse dans celle qui est amplifiée par le premier couple d'amorces. Ainsi si le premier couple d'amorces est artéfactuelle, lors de la deuxième PCR les amorces du deuxième couple ne pourront pas s'hybrider et il n'y aura pas d'amplification. Cette technique permet d'augmenter dans le même temps la spécificité et le taux d'amplification. Elle est particulièrement souhaitée lorsque la quantité de départ est faible.

La PCR quantitative (qPCR), ou PCR en temps réel est souvent utilisée comme méthode particulière de PCR classique permettant de mesurer la quantité initiale d'ADN. En réalité, la PCR quantitative mesure le nombre d'amplicons (portion d'ADN définie par un couple d'amorces).

En clinique, on continue de pratiquer essentiellement la GE et le TDR, car les PCR sont très sensibles mais restent trop complexes pour une mise en œuvre sur le terrain en raison de la nécessité de conditions de laboratoire sophistiquées, d'une formation avancée et d'un délai relativement long avant l'obtention des résultats.

— **La technique *loop-mediated isothermal amplification* (LAMP)**

LAMP est une nouvelle alternative à la PCR et à la qPCR. C'est une méthode d'amplification isothermique à médiation par boucle. Elle présente un très haut niveau de spécificité, lié à l'utilisation de 6 amorces. Cette méthode d'amplification

est en général très rapide. Elle se fait à une température comprise entre 60-65°C et ne comporte pas d'étape de purification et de dénaturation de la séquence nucléique cible. Elle utilise une instrumentation simple (bain-marie ou bloc chauffant) et peut être réalisée en moins d'une heure (Poon, Wong, Ma et al., 2006).

La méthode LAMP se fait en trois étapes suivantes :

1. Hybridation avec 3 amorces sens suivie d'une élongation
2. Hybridation du nouveau brin avec 3 amorces anti-sens suivie d'une élongation
3. Révélation. Cette méthode génère des sous-produits qui forment un précipité blanc et augmentent la turbidité de la solution. La détection de l'amplification peut donc être réalisée par un grand nombre de méthodes telles que l'électrophorèse, la détermination turbidimétrique, des méthodes électrochimiques ou simplement colorimétriques (par exemple via l'utilisation de SybrGreen).

Comme pour la PCR, il existe plusieurs déclinaisons de cette méthode : la LAMP classique, la LAMP en temps réel, la RT-LAMP ou encore la RT-LAMP en temps réel.

L'étude de Vasquez et al (Vásquez, Zuluaga, Tobón et al., 2018) a rapporté une sensibilité (100%) plus élevée de la technique LAMP comparé à la microscopie (79,5%) et le TDR (76,9%) pour la détection de l'infection plasmodiale dans le sang périphérique des femmes à l'accouchement. Dans cette étude, la spécificité de ces trois techniques était similaire (100%). De même, la sensibilité pour la détection des infections placentaires était quatre fois plus élevée avec LAMP (88%) comparée à la microscopie et le TDR (22% pour les deux techniques). Une autre étude réalisée par Serra-Casas en 2017 (Serra-Casas, Manrique, Ding et al., 2017) a rapporté une sensibilité de 91,8% pour LAMP et 91,9% pour la qPCR qui était largement supérieur à celle de la microscopie (30%).

Plusieurs études comparant la méthode LAMP à la microscopie, à la PCR et au TDR, ont montré une sensibilité et une spécificité élevées par rapport à la microscopie ou à la PCR (Polley, Mori, Watson et al., 2010 ;Pöschl, Waneesorn, Thekisoe et al., 2010) (voir annexe 2).

1.1.5. Prévention et traitement du paludisme chez la femme enceinte

Dans les zones de transmission modérée à élevée de *P.falciparum*, l'OMS recommande un ensemble d'interventions pour contrôler le paludisme et ses effets pendant la grossesse, ce qui inclut la promotion et l'utilisation des moustiquaires imprégnées d'insecticide (MII), le TPIp-SP, et prise en charge appropriée des cas par le biais de traitement rapide et efficace chez la femme enceinte[10].

1.1.5.1. Traitement préventif intermittent

Le premier médicament utilisé pour prévenir le paludisme placentaire était la chloroquine (CQ) hebdomadaire à une dose prophylactique de 300 mg (Dose totale : 25 mg/kg, répartis sur 3 jours. 1er jour : 10 mg/kg, puis 5 mg/kg, 6 à 8 heures plus tard. 2e et 3e jours : 5 mg/kg, en une seule prise)[11].

Cependant, l'apparition de parasites résistants à la choloroquine (CQ) en Afrique sub-saharienne au cours des années 1980 a incité à rechercher de nouvelles stratégies.

Au début des années 2000, l'OMS a recommandé un TPI chez les femmes enceintes des régions d'endémie palustre, avec au moins deux doses curatives du médicament antipaludique à base de la sulfadoxine pyriméthamine (SP), une première dose au second trimestre et une deuxième au troisième trimestre de la grossesse (Publication Data, 2007). La première dose était administrée aux environs de la 16ème semaine d'aménorrhée à raison de 3 comprimés (500 mg de Sulfadoxine et 25 mg de Pyriméthamine par comprimé). La deuxième dose était prise au deuxième trimestre entre la 28ème et la 34ème semaine d'aménorrhée. La SP est une combinaison de molécules ayant pour cibles : la dihydrofolate réductase (DHFR) pour la sulfadoxine et la synthase dihydroptéroate (DHPS) pour la pyriméthamine (Gregson et Plowe, 2005). Une étude réalisée en 1992 au Malawi a montré que deux doses de SP administrées au deuxième et au début du troisième trimestre réduisaient considérablement la prévalence du paludisme gestationnel par rapport à la CQ (Schultz, Steketee, Macheso et al., 1994). Un essai ultérieur au Kenya a confirmé que deux doses de traitement à la SP réduisaient la prévalence du paludisme placentaire chez les femmes infectées par le VIH (Parise, Ayisi, Nahlen et al., 1998).

En 2012, l'OMS a mis à jour la recommandation, en augmentant le nombre à trois doses de SP ou plus. En pratique, les femmes vivant dans des zones de transmission modérée à élevée du paludisme devraient recevoir une dose de SP à chaque visite de soins prénatals au cours des deuxième et troisième trimestres (quatre visites étant recommandées), avec un intervalle d'un mois entre les doses[12].

Dès le deuxième trimestre de la grossesse ou dès la perception des premiers mouvements fœtaux par la femme enceinte, il est recommandé à chaque femme enceinte de faire 4 consultations prénatales (CPN). A partir de la 16ème semaine de grossesse (2ème trimestre) ou dès la perception des premiers mouvements actifs fœtaux par la femme enceinte, le TPI doit être administré lors de chaque CPN. Dans tous les cas, l'intervalle entre deux doses doit être au moins de 1 mois.

En raison de la propagation de la résistance à la SP en Afrique sub-saharienne, les traitements thérapeutiques à base d'artémicinine (*Artemicine combinaison therapeutic*, ACT) ont été adoptées comme traitement de première intention du paludisme simple dans les années 2000 (Eastman et Fidock, 2009). Les ACT ont été adoptés en République du Congo en 2006. Des études d'efficacité thérapeutiques de Ndounga et al (2007) réalisées en 2003-2004 et de Nsimba et al (2004) réalisées en 2004 avaient rapporté des niveaux d'échec thérapeutiques, respectivement de 31,2% et de près de 25% pour la SP d'une part et respectivement de 95,5% et plus de 25% pour la chloroquine d'autre part. Alors même que la population générale adoptait les ACT comme politique de traitement, la stratégie TPIp-SP était en train d'être largement adoptée par les femmes enceintes. Au cours des années 1992-2002, le TPIp-SP a réduit de manière significative le paludisme placentaire dans les études menées en Afrique. Cependant, la plupart des données recueillies après 2001-2002 en Afrique de l'Est et du Sud-Est indiquaient que le TPIp-SP a perdu de son efficacité pour réduire la prévalence du paludisme placentaire et / ou la densité parasitaire. Cette tendance a progressé jusqu'en Afrique de l'Ouest, où un site au Ghana a signalé que le TPIp-SP n'a pas réduit la prévalence du

paludisme placentaire (Spronsen, Schneider et Atasige, 2012). Cependant, à l'heure actuelle, l'OMS continue de recommander le TPIp-SP, même dans les zones présentant des niveaux élevés de résistance à la SP et d'échec thérapeutique.

La résistance à la SP résulte de l'accumulation de mutations dans les gènes dhfr et dhps. En effet, les mutations quintuples de *P. falciparum* (trois en *Pfdhfr* et deux en *Pfdhps*) ont été associées à un échec du traitement (Kublin, Dzinjalamala, Kamwendo et al., 2002, Naidoo et Roper, 2013) et à une densité accrue de parasites placentaires corrélée avec un nombre croissant de mutations du Pfdhfr (Mockenhaupt, Bedu-Addo, Junge et al., 2007).

Un document de l'OMS publié en novembre 2015[13] indiquait : Une association entre des haplotypes mutants sextuple de *P. falciparum* et une diminution du poids à la naissance a ont été rapportés dans des études d'observation dans quelques sites d'Afrique orientale. D'autres études sont nécessaires pour évaluer cela et concevoir les stratégies de prévention les meilleures et les plus rentables dans les zones de très forte résistance à la SP. La politique de poursuite de TPIp-SP dans les zones de haute résistance est déroutante et contraire aux directives de l'OMS pour le traitement du paludisme (Nosten et McGready, 2015), ainsi que des études associant fortement les mutations dhfr / dhps à l'échec du traitement.

Actuellement, le TPIp-SP reste efficace pour réduire le taux du paludisme placentaire et / ou de charge parasitaire sur certains sites en Afrique de l'Ouest. Cependant, même dans les zones où la résistance à la SP est faible ou modérée, la stratégie TPIp-SP ne prévient pas complètement le paludisme placentaire et les effets protecteurs dépendent du moment de la première dose et de l'intervalle entre les traitements (Nosten et McGready, 2015).

Chaque dose de sulfadoxine-pyriméthamine (comprimés de 500 mg de sulfadoxine et 25 mg de pyriméthamine [Kamsidar, Kampala Pharmaceutique Industries]) consistait en trois comprimés pris ensemble ; des doses ont été administrées à trois reprises pendant la grossesse (groupe sulfadoxine-pyriméthamine). Chaque dose de dihydroartémisinine-pipéraquine (comprimés de 40 mg de dihydroartémisinine et 320 mg de pipéraquine [Duo-Cotecxin, Holley-Cotec]) consistait en trois comprimés administrés une fois par jour pendant 3 jours consécutifs ; des doses ont été administrées soit trois fois au cours de la grossesse (groupe des trois doses dihydroartémisinine-piperaquine) ou une fois par mois (groupe mensuel dihydroartémisinine-pipéraquine). Les participants assignés au groupe sulfadoxine-pyriméthamine ou au groupe trihydroxy dihydroartémisinine-pipéraquine ont reçu les agents actifs de l'étude à 20, 28 et 36 semaines de gestation. Les participants assignés au groupe mensuel dihydroartémisinine-pipéraquine ont reçu des agents actifs de l'étude toutes les 4 semaines à partir de 16 ou 20 semaines de gestation. Des placebos de sulfadoxine-pyriméthamine et de dihydroartémisinine – pipéraquine ont été utilisés de telle sorte que toutes les 4 semaines, les participants recevaient le même nombre de comprimés de même apparence. L'administration des premières doses quotidiennes de principe actif ou de placebo a été directement observée à la clinique et les deuxième et troisième doses quotidiennes ont été administrées à domicile. Des détails supplémentaires concernant l'administration des agents de l'étude sont fournis dans le protocole de l'essai.

1.1.5.2. Alternatives au TPIp-SP : la Dihydroartémisinine-Pipéraquine (DP)

Une comparaison entre trois doses de TPIp-SP et trois doses de DP (comprimés de 40 mg de dihydroartémisinine et 320 mg de pipéraquine) a récemment été réalisée en Ouganda (Kakuru, Jagannathan, Muhindo et al., 2016). Chaque dose de SP (comprimés de 500 mg de sulfadoxine et 25 mg de pyriméthamine) consistait en trois comprimés pris ensemble ; des doses ont été administrées à trois reprises pendant la grossesse (groupe SP). Chaque dose de DP (comprimés de 40 mg de dihydroartémisinine et 320 mg de pipéraquine) consistait en trois comprimés administrés une fois par jour pendant 3 jours consécutifs ; des doses ont été administrées soit trois fois au cours de la grossesse (groupe des trois doses DP) ou une fois par mois (groupe mensuel DP).

La parasitémie sanguine périphérique détectée par LAMP était significativement plus élevée dans le groupe TPIp-SP de trois doses que de DP mensuelle. De même, les infection combinées actives et passées étaient significativement plus élevées chez les femmes recevant du TPIp-SP que chez celles recevant trois doses ou un traitement mensuel avec DP (Figure 1.12).

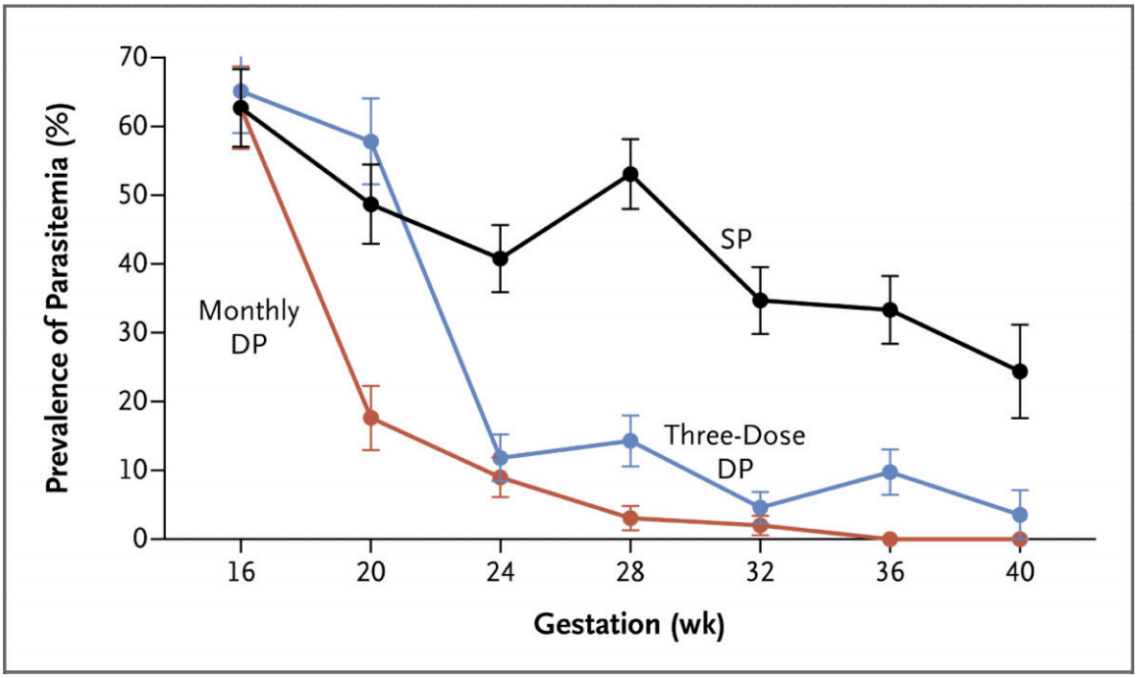


Figure 1.12. – Prévalence de la parasitémie pendant la grossesse, selon la semaine de gestation (Source : Kakuru, Jagannathan, Muhindo et al., 2016).

Bien que, chez les femmes primigravidaes, le taux de paludisme placentaire était similaire dans les trois groupes, la quantité d'hémozoïne était significativement plus élevée dans les groupes TPIp-SP, ce qui pourrait indiquer une densité parasitaire plus élevée lors d'infections passées. Le risque de mauvaise issue de la grossesse (accouchement prématuré, faible poids de naissance , anomalie congénitale, mortinatalité, avortement spontané) était significativement plus faible chez

les femmes recevant une DP mensuelle que chez celles recevant trois doses de DP ou de TPIp-SP.

1.1.5.3. Moustiquaires imprégnées d'insecticide (MII)

Les moustiquaires traitées à l'insecticide devraient être fournies aux femmes enceintes dès le début de la grossesse possible. Les femmes devraient être encouragées à utiliser des MII tout au long de la grossesse, ainsi que pendant la période post-partum, lorsque le risque de paludisme est également accru. Le TPI-SP ne remplace pas l'utilisation des MII ; les deux interventions offrent des avantages importants.

1.2. Approches génomiques de la résistance et la susceptibilité aux maladies

1.2.1. Les variations génétiques des populations humaines

Le premier séquençage du génome humain (entre 1990 et 2003) a constitué un modèle en génétique humaine. L'ère du big data combinée à une réduction du coût du séquençage a ouvert une nouvelle recherche, dont l'un des volets est l'étude systématique de la variation génétique humaine. Le séquençage du génome a été très utile dans la compréhension de la localisation et la structure des gènes ; mais n'a pas renseigné sur la diversité du génome humain en fonction des individus et à l'échelle des populations. Pour affronter cette question, différents consortium internationaux se sont constitués, qui avaient pour but principal de caractériser le taux de polymorphismes qu'on trouve entre les différents individus et populations.

1.2.1.1. Classification des variants

On distingue deux types de variations dans le génome humain : les variations de séquence et les variation structurelles. Les variations de séquence incluent les SNP (Single Nucleotide Polymorphism), les insertions, les délétions, les indel et les substitutions. Les variations structurelles sont catégorisées en CNV (Copy Number Variant), inversion et translocation (Figure 1.13)[14].

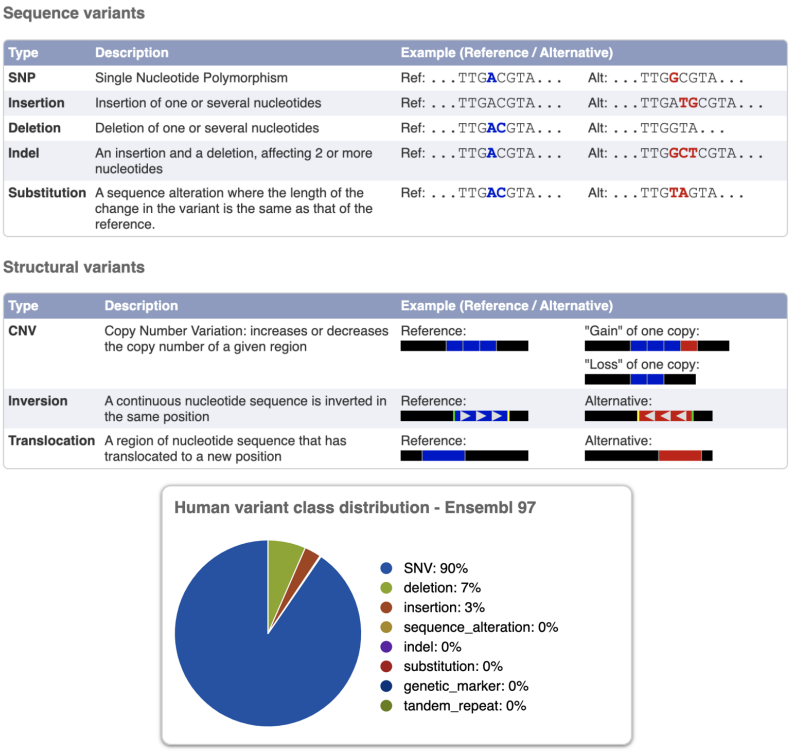


FIGURE 1.13. – Classification des variants génétiques. Les SNV (Single Nucleotide Variant) sont des positions de nucléotide simple dans l'ADN génomique pour lesquelles différentes alternatives de séquence existent. Ils représentent 90 pourcent des variations génétiques (adapté de Ensemb).

Le génome humain est composé de deux copies de chromosomes totalisant 3 milliards de paires de bases (pb). Les hommes héritent tous de deux copies (une copie de la mère et une copie du père) de ces 3 milliards de pb. Si l'on considère deux copies du génome humain, on observe dans la plupart des position le même nucléotide (donc la même information génétique dans les deux génomes). Mais certaines positions diffèrent et sont appelées polymorphismes (THE 1000 GENOMES PROJECT CONSORTIUM, 2015) qui sont repartis sur les 23 pairs de chromosomes et mitochondrie. En moyenne 2 copies du génome humain pris au hasard diffèrent l'une de l'autre par environ 0,1% des sites de nucléotides (c'est-à-dire 1 variant par 1 000 bases en moyenne) (WANG, 1998). Les données de *Ensembl 2018* font état de 150 millions de variations dans le génome humain (Tableau 1.1).

TABLE 1.1. – Nombre de variations polymorphiques humains par chromosome (source : *Ensembl variation 2018*).

Chromosome	Variation	Chromosome	Variation
1	11539842	14	4673783
2	12403763	15	4323507
3	10202364	16	4868520
4	9819229	17	4338697

Chromosome	Variation	Chromosome	Variation
5	9139217	18	3888598
6	8771545	19	3542478
7	8254393	20	3291351
8	7925640	21	1969843
9	6319783	22	2031211
10	6924480	X	5415253
11	7202023	Y	342525
12	6896404	MT	1760
13	4903156		NA
Total	NA		148989365

1.2.2. Single Nucleotide Polymorphism (SNP)

Un polymorphisme nucléotidique unique, ou SNP (pour *Single Nucleotide Polymorphism*), est une variation en une position unique dans une séquence d'ADN parmi des individus. Ils représentent plus de 90% de toutes les différences entre individus. Les SNPs sont stables et répartis à une fréquence d'environ une paire de base (pb) pour mille dans le génome humain (*L'étude des Snips*). Pour deux génomes humains tirés au hasard, 99,9% de la séquence d'ADN est identique, alors que 0,1% de la séquence contient des variations. On estime qu'il y a en moyenne un SNP pour 1 000 nucléotides (Wang, 1998). Chaque SNP représente une différence dans un seul bloc constitutif de l'ADN, appelé nucléotide (Figure 1.14). Un gène est décrit comme ayant plus d'un allèle s'il contient un SNP. Dans ce cas les SNPs peuvent entraîner des variations dans la séquence d'acide aminés. Cependant, les SNPs peuvent aussi se produire dans les régions non-codantes de l'ADN et peuvent avoir un impact considérable sur la régulation des gènes (Gusev, Lee, Trynka et al., 2014), contribuant ainsi à la diversité phénotypique et à un large éventail de troubles humains (*The role of regulatory variation in complex traits and disease. - PubMed - NCBI* ;Deplancke, Alpern et Gardeux, 2016). Certains SNPs sont associés à des maladies, bien qu'une majorité des SNPs ne causent pas spécialement de troubles. Ces associations permettent la recherche des SNPs afin d'évaluer la prédisposition génétique d'un individu à développer une maladie. De plus, si l'on sait que certains SNPs sont associés à un trait, on peut examiner des segments d'ADN proches de ces SNPs afin d'identifier le ou les gènes responsables de ce trait.

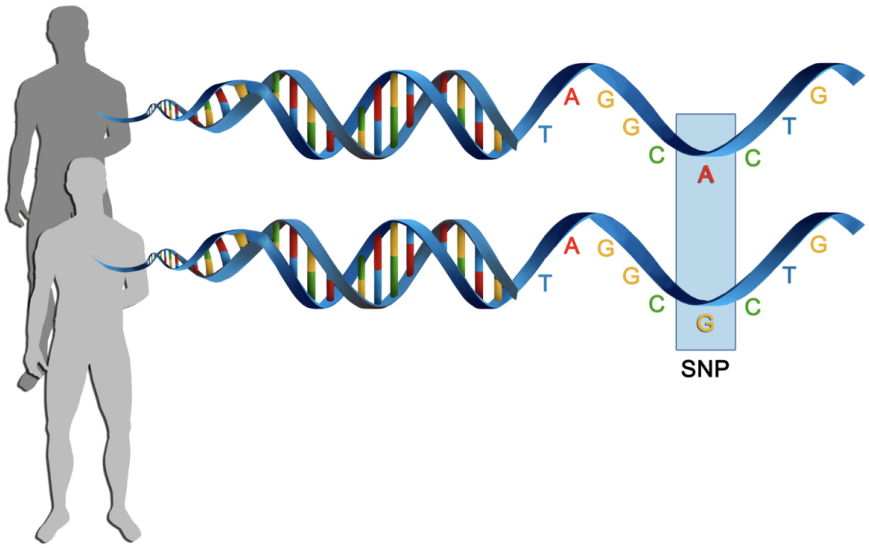


FIGURE 1.14. – Polymorphisme mononucléotidique (Source : https ://atlasofscience.org/single-nucleotide-polymorphisms-as-genomic-markers-for-high-throughput-pharmacogenomic-studies/).

1.2.3. Les variations structurelles

Les variations structurelles (*structural variation*, SV) sont des réarrangements génomiques affectant plus de 50 pb. La taille moyenne des SVs détectées dans les données du projet 1000 Génomes est de 8 kb (THE 1000 GENOMES PROJECT CONSORTIUM, 2015). Les SVs peuvent être identifiés par la méthode d'hybridation comparée du génome (*comparative genome hybridization*, CGH). C'est une technique permettant de détecter les modifications du nombre de copies chromosomiques sans recourir à la culture cellulaire (CONRAD, PINTO, REDON et al., 2010). Cette méthode rapporte une valeur quatre fois plus grande (CONRAD, PINTO, REDON et al., 2010). Les SVs comprennent des événements équilibrés et non équilibrés, à savoir des variantes modifiant le nombre total de paires de bases dans un génome. Ainsi, les SVs incluent les délétions, les insertions, les inversions, les transpositions d'éléments mobiles, les translocations, les répétitions en tandem et les variants de nombre de copies (*copy number variant*, CNV) (voir Figure 1.15). Les SVs expliquent 1,2% de la variation entre les génomes humains et les SNPs, 0,1% (PANG, MACDONALD, PINTO et al., 2010).

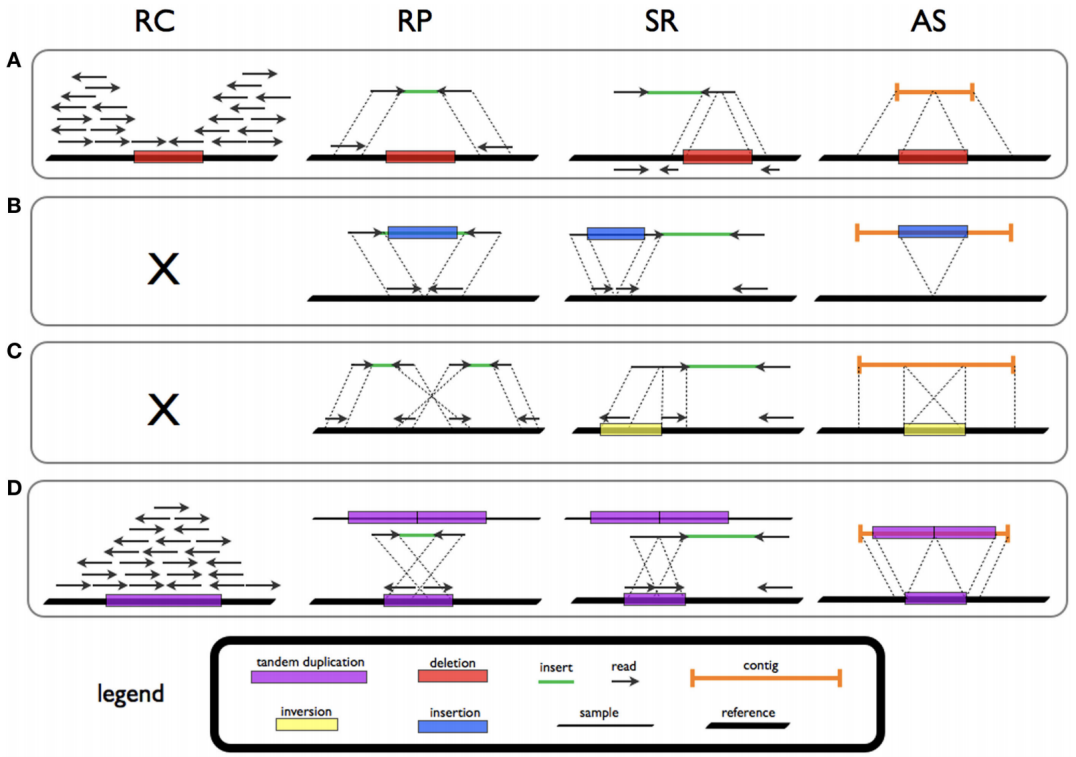


FIGURE 1.15. – Signatures et modèles de SV pour suppression (A), insertion de nouvelle séquence (B), inversion (C) et duplication en tandem (D) en nombre de lectures (RC), paire de lecture (RP), lecture partagée (SR), et les méthodes d'assemblage de novo (AS). (Source : adapté de LORENZO T. et al., 2015).

Comme pour les variations mononucléotidiques, une grande partie des SVs évoluera probablement de manière neutre, mais dans certains cas, il existe des preuves d'un équilibre ou d'une évolution adaptative (SCHRIDER, HAHN et BEGUN, 2016 ;LEFFLER, BAND, BUSBY et al., 2017). Il a également été démontré que le SVs générait de nouveaux gènes ayant des conséquences fonctionnelles. Un exemple de génération du gène SRGAP2 spécifique à l'homme augmentent la densité des épines dendritiques dans le cerveau (CHARRIER, JOSHI, COUTINHO-BUDD et al., 2012). Une étude pangénomique récente a montré que l'assemblage des génomes de 910 humains d'ascendance africaine contient environ 10% plus d'ADN que le génome de référence humain actuel (GRCh38) (SHERMAN, FORMAN, ANTONESCU et al., 2019). En plus des SVs dans la lignée germinale, de grandes SVs somatiques ont été observées dans le cerveau humain (MCCONNELL, MORAN, ABYZOV et al., 2017), dans les fibroblastes de la peau (ABYZOV, MARIANI, PALEJEV et al., 2012) et dans le sang de jumeaux identiques (BRUDER, PIOTROWSKI, GIJSBERS et al., 2008).

1.2.4. Le projet 1000 Génomes

Entre 2008 et 2015, le projet 1000 Génomes a été réalisé avec comme objectif : l'identification de la plupart des variations génétiques ayant une fréquence d'au

moins 1% dans la population étudiée. Dans la phase finale de ce projet, on a reconstitué 2504 génomes d'individus de 26 populations (Afrique, Asie de l'Est, Asie du sud, Europe et les Amériques) en utilisant une combinaison de séquençage du génome entier à faible couverture, de séquençage d'exome profond et de génotypage dense de micro-réseaux (Figure 1.16) (The 1000 Genomes Project Consortium, 2015). Le projet a mis en évidence un large spectre de variations totalisant 88 millions dont 84,7 millions de SNPs, 3,6 millions d'insertions / délétions courtes (indels) et 60 000 variants structuraux associés à des haplotypes de haute qualité (The 1000 Genomes Project Consortium, 2015). Cette ressource inclut de 99% des variants de type SNPs avec une fréquence de 1% d'allèles mineurs.

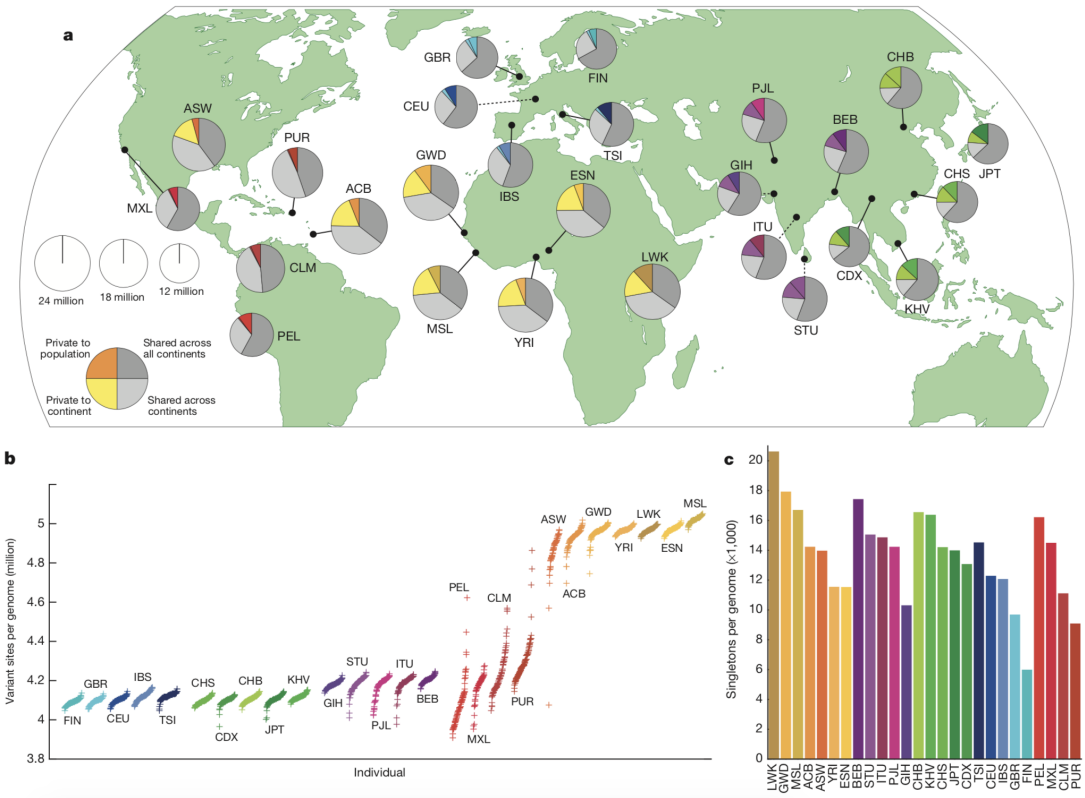


Figure 1.16. – Échantillonnage de populations pour le projet 1000 genomes. (a), variants polymorphes dans l'échantillon des populations. L'aire de chaque diagramme en camembert est proportionnelle au nombre de polymorphismes au sein d'une population. Les camemberts sont divisées en quatre tranches, représentant respectivement les variants propres à une population (couleur plus foncée unique à la population), propres à une zone continentale (couleur plus claire partagée par un groupe continental), partagées entre des zones continentales (gris clair) et répartis sur tous les continents (gris foncé). Les lignes pointillées indiquent les populations échantillonnées en dehors de leur région continentale ancestrale. (b), le nombre de sites variants par génome. (c), le nombre moyen de singletons par génome (Source : The 1000 Genomes Project Consortium, 2015).

L'une des utilisations les plus immédiates des données du projet 1000 Génomes concerne les études d'association du génome entié (*genome-wide association study*, GWAS) qui analyse de nombreuses variations génétiques chez de nombreux individus, afin d'étudier leurs corrélations avec des traits phénotypiques. La notion de GWAS est traité dans le paragraphe 3.9.

1.2.5. Projet population islandaise

Un rapport d'une étude de l'ensemble des génomes d'un grand groupe d'Islandais par un séquençage à grande échelle du génome entier a mise en évidence un ensemble de 20 millions de SNPs et 1,5 million d'indels pour un séquençage de 2 636 génomes d'individus Islandais (Gudbjartsson, Helgason, Gudjonsson et al., 2015). Les auteurs ont constaté un excès d'homozygotie et de variants rares codant pour les protéines dans cette population. Ils ont par la suite imputé ces variants à 104 220 personnes (avec une fréquence d'allèles mineurs d'au moins 0,1%) et ont montré l'intérêt de ces données pour détecter des variations associées à des traits particulier. Par exemple, ils ont identifié deux mutations récessifs de type indel dans le gène MYL4 (*Myosin Light Chain 4*) qui provoque une fibrillation auriculaire précoce ; plusieurs mutations dans le gène ABCB4 (*ATP Binding Cassette subfamily B member 4*) qui augmentent le risque de maladies du foie et un variant dans la région intronique du gène GNAS (*Guanine Nucleotide binding protein, Alpha Stimulating*) associée à une augmentation du niveau d'hormones stimulant la thyroïde (allèles agissent via un effet spécifiquement maternel).

1.2.6. Les stratifications dans les populations humaines

La stratification des populations humaines est caractérisée par la présence de différences de fréquences alléliques entre les sous-populations ayant une origine ancestrale différente. Des groupes humains ont quitté le site de leurs origines géographiques en Afrique (McDougall, Brown et Fleagle, 2005, White, Asfaw, DeGusta et al., 2003) et ont migré vers l'Europe, l'Asie et les Amériques (Gravel, Henn, Gutenkunst et al., 2011, Mallick, Li, Lipson et al., 2016). Ces migrations ont produit le métissage et l'adaptation qui différencient les populations humaines (Schlebusch, Skoglund, Sjödin et al., 2012). L'isolement physique de ces groupes d'individus sur plusieurs générations a produit trois grands groupes ancestraux (Africain, Asiatique et Européen) et conduisant à la dérive génétique qui offrent aux Africains le plus haut niveau de diversité génétique parmi les populations continentales humaines actuelles (Rosenberg, Pritchard, Weber et al., 2002). Une conséquence de la stratification de la population est le mélange génétique qui découle d'un métissage de groupes ancestraux tel que le mélange génétique de la population afro-américaine, qui a des origines Africaines et Européennes.

La stratification et le mélange de population, est un facteur de confusion important dans les études de GWAS, car elle peut produire des associations erronées. La recherche d'un locus causal d'un trait à partir des études de GWAS est une tâche qui nécessite l'élimination des associations fausses qui découlent de la structure de population et des facteurs de confusion environnementaux. La corrélation induite par l'ascendance Des associations parasites peuvent survenir en raison de corréla-

tions induites lorsque l'ascendance est corrélée à la fois au génotype et au trait (Campbell, Ogburn, Lunetta et al., 2005). L'analyse en composantes principales (PCA) (Price, Patterson, Plenge et al., 2006) et les modèles linéaires mixtes (LMM) (Kang, Sul, Service et al., 2010) sont deux classes populaires de méthodes de correction de la confusion entre ascendances dans les études de génétique humaine. Ces deux méthodes testent les associations génétiques dans des modèles prospectifs décrivant la distribution du caractère en fonction du génotype.

1.2.7. Implications fonctionnelles des variants génétiques non-codants

Les implications fonctionnelles de la variation génétique des régions non-codantes pour la susceptibilité aux maladies et régulation des gènes fait l'objet de nombreuses études. Les facteurs génétiques qui déterminent notre susceptibilité individuelle à des traits complexes (notamment des maladies) se sont révélés difficiles à localiser malgré des efforts de recherche importants. Les effets génétiques impliquent probablement plusieurs loci de susceptibilité d'ampleur individuelle, ce qui limite l'application des études de liaison. L'approche du "gène-candidat" a été largement utilisée pour analyser les associations possibles entre variants génétiques et évolution de la maladie, avec la sélection de gènes basés sur une connaissance a priori de la pathogenèse de la maladie. Cependant, l'analyse de liaisons et les études d'association ne permettent généralement pas de déterminer si un variant génétique associé à une maladie est important du point de vue fonctionnel ou s'il ne s'agit que d'un marqueur génétique qui a été détecté parce qu'il est génétiquement lié au locus fonctionnel hérité de l'allèle polymorphe.

Une variation qui se produit dans l'ADN codant peut induire un effet de modification de la séquence peptidique pouvant entraîner un changement de conformation tridimensionnelle qui rend la protéine inactive ou lui confère de nouvelles propriétés (Ramensky, 2002). Même lorsque le polymorphisme reste silencieux au niveau de la protéine, comme dans le cas de mutations synonymes, l'effet du polymorphisme peut être analysé au niveau de l'ARNm. Ceci est basé sur le fait que lorsqu'un polymorphisme apparaît dans l'ADN codant, il sera présent dans l'ARNm transcrit.

En revanche, l'évaluation de l'effet fonctionnel des polymorphismes apparaissant dans l'ADN non-codant est plus problématique. L'identification de variants fonctionnels des régions génomiques non-codantes reste difficile en raison de leur abondance, de leur spécificité de type cellulaire / tissulaire et de leur mode d'action complexe (Khurana, Fu, Chakravarty et al., 2016). Dans cette étude, l'attention s'est surtout portée sur les SNPs apparaissant dans régions régulatrices dans lesquelles un changement de base dans la séquence d'ADN peut affecter le niveau d'expression des gènes.

1.2.8. Les variants régulateurs

Les variants dans les régions génomiques non-codantes peuvent avoir un effet important sur la régulation transcriptionnelle, et certains d'entre eux sont impliqués dans l'étiologie de maladies. Ces dernières années, l'accent a été mis de plus en plus sur ces variants et sur la manière dont ils peuvent modifier des éléments régulateurs,

tels que les amplificateurs, les sites de liaison aux facteurs de transcription (TFs) (Figure 1.17) et les régions de méthylation de l'ADN. De telles variants peuvent être considérés comme des variants régulateurs. Parallèlement, de nombreux efforts ont été déployés pour créer des consortiums internationaux afin d'entreprendre de grands projets visant à découvrir les éléments régulateurs dans différents tissus, lignées cellulaires et organismes, et à sonder les effets des variants génétiques sur la régulation en mesurant l'expression des gènes.

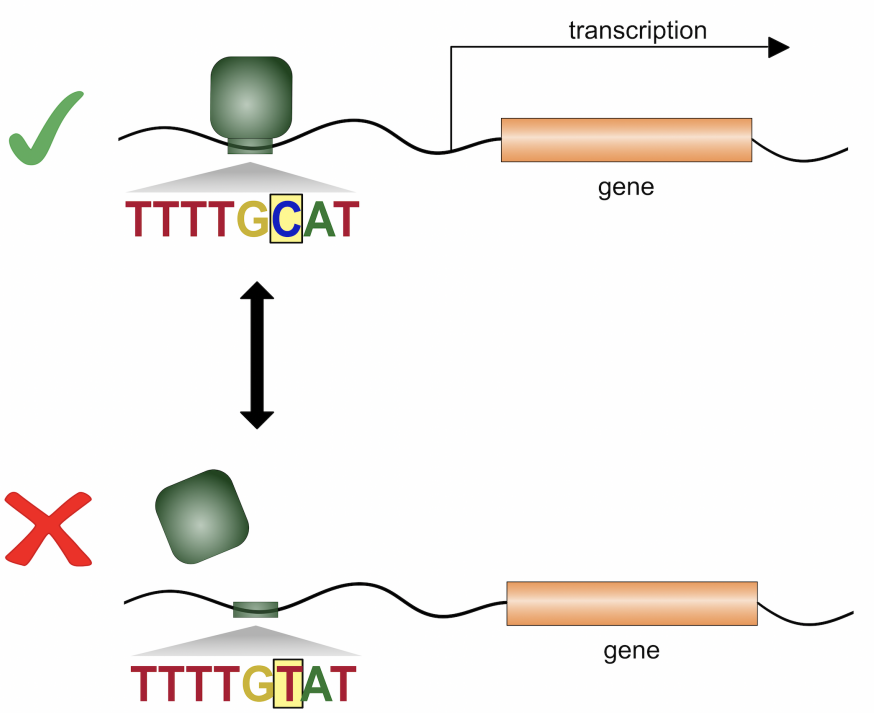


FIGURE 1.17. – Schéma simplifié de l'impact d'un variant régulateur sur la transcription (en vert). En présence l'allèle C (en bleu), le facteur de transcription se fixe à son site activateur au promoteur et induit l'a transcription de l'ADN'expression du gène (en orange). L'activité du promoteur en présence du variant T empêche la fixation du facteur transcriptionnel et inactive l'expression du gène (en orange).

Il est désormais plus facile et économique de séquencer, grâce aux avancées technologiques du séquençage, le génome d'un patient dans le but d'identifier les variants associés à une maladie. Plusieurs études ont été concentré sur les variants recouvrant des régions du génome codant pour des protéines.

Il existe plusieurs moyens par lesquels les variants non-codants peuvent affecter la régulations transcriptionnelle. La plupart de ces variants sont impliqués dans la perturbation d'éléments régulateurs génomiques régulant l'expression des gènes : on les appels éléments régulateurs (NARLIKAR et OVCHARENKO, 2009). On distingue des éléments régulateurs :

— les éléments cis-régulateurs (agissant en cis) : c'est une séquence d'ADN ayant un effet régulateur sur la transcription d'un gène.

— Les éléments trans-régulateurs (agisant en trans) : c'est une protéine (facteur de transcription) qui se fixe sur la séquence régulatrice et module le taux de transcription.

La prédiction des variants régulateurs associés à une maladie relève d'une grande complexité qui implique les procédures expérimentales et des ressources informatiques. Certaines étapes sont les mêmes que pour la détection des variants codants, telles que le séquençage du génome, l'analyse informatique des données et l'identification des variants associés, par exemple en comparant les fréquences alléliques entre individus. D'autres étapes sont spécifiques à la détection de variants régulateurs telles que : la méthode de gène rapporteur (*Reporter Gene Assays - FR*), la méthode *Survey of Regulatory Elements (SuRE)*(Arensbergen, FitzPatrick, Haas et al., 2017). Ces deux méthodes sont décrites dans le chapitre 5.

Les variants régulateurs sont une source potentielle d'influence de la régulation de l'expression des gènes et sont le plus souvent non annotés. Afin de comprendre l'impact fonctionnel de ces variants sur un trait phénotypique, il est nécessaire :

1) de mettre en place des méthodes d'annotation des régions qui chevauchent ces variants ;
2) de prioriser ces variants à base ces annotations (prévoir leurs effets fonctionnels) à l'aide des outils bioinformatiques.

Au cours des ces dernières années, de nombreuses études ont adopté une approche intégrative consistant à combiner des annotations non-codantes disponibles afin de prédire l'impact fonctionnel probable des variants sur la régulation. La plupart de ces méthodes, telles que : CADD (Kircher, Witten, Jain et al., 2014), GWAVA (Ritchie, Dunham, Zeggini et al., 2014), FATHMM-MKL (Shihab, Rogers, Gough et al., 2015) et Genomiser (Smedley, Schubach, Jacobsen et al., 2016), IW-Scoring (Wang, Dayem Ullah et Chelala, 2018), TAGOOS (« TAGOOS ») ont utilisé des algorithmes d'apprentissage automatique pour développer des classificateurs intégrant une gamme d'annotations entre autres, les caractéristiques régulatrices des variants non-codants.

Dans le cadre de ma thèse, j'ai développé une approche bioinformatique qui intègre principalement les outils d'analyse de variations régulatrices présents dans la suite logicielle *Regulatory Sequence Analysis Tools (RSAT variation-tools)*(*RSAT Var-tools : an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding | bioRxiv*) pour la détection des variants régulateurs et *ReMap* (Chèneby, Gheorghe, Artufel et al., 2018) pour identifier les facteurs transcriptionnels dont les régions de pics de ChIP-seq chevauchent les variants régulateurs.

1.2.9. Genome-Wide Association Studies (GWAS)

Les études d'association du génome entier (*Genome-Wide Association Study*, GWAS) ont pour objectif d'identifier des facteurs de susceptibilité génétiques des traits phénotypiques ou maladies multifactorielles. Elles consistent à comparer la fréquence de centaines de milliers de variants génétiques distribués sur l'ensemble des chromosomes entre un groupe de cas atteints de la maladie et un groupe de témoins (Figure 1.18), en utilisant des technologies de génotypage à haut débit.

L'approche GWAS (Manolio, 2010) a été proposée pour identifier les loci associés à des traits complexes (Hindorff, Sethupathy, Junkins et al., 2009 ;McCarthy et Hirschhorn, 2008) sans aucune dépendance vis-à-vis d'hypothèses préalable sur les gènes d'intérêt contrairement aux études d'association génétique de type gène candidat. Le nombre très important de tests statistiques effectués nécessite une puissance de calcul importante et une correction sur tests multiples, une valeur $5x10^{-8}$ étant généralement considérée comme significative.

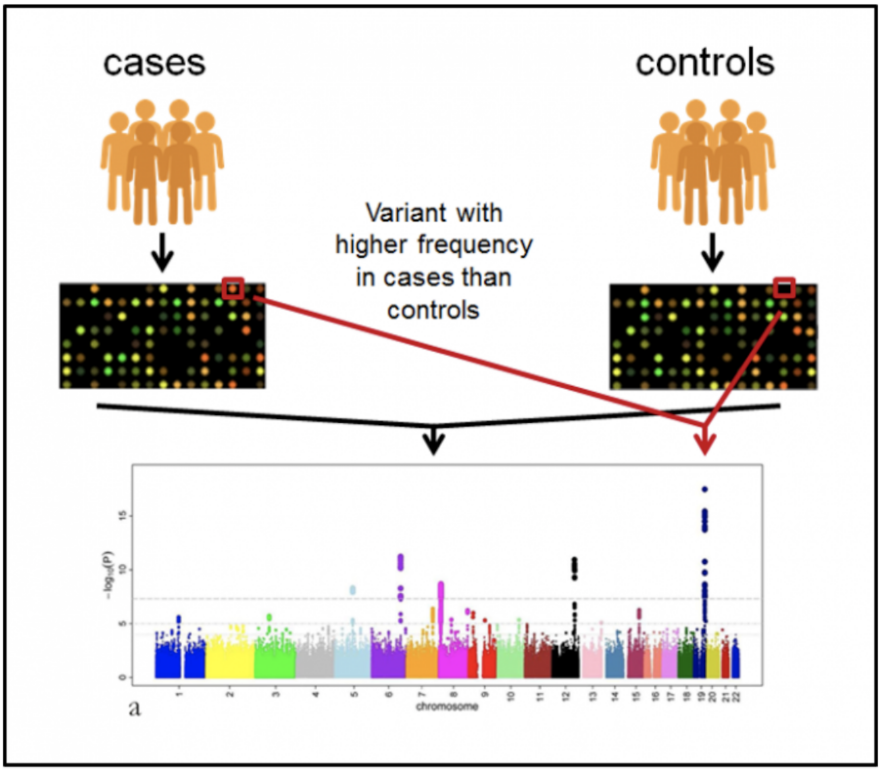


Figure 1.18. – Illustration de l'identification de variants alternatifs dans les cas et les contrôles à l'aide d'une méthode de typage basée sur des tableaux. Les résultats sont soumis à des analyses statistiques pour attribuer une p-valeur à chaque variant (Source : https ://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations/why-do-we-need-gwas-catalog/what-are-genome).

Les études de GWAS ont permis de cartographier les variations génétiques dans le génome humain. La dernière version du diagramme de GWAS indique toutes les positions pour lesquelles les variations génétiques ont été associés à un trait (par exemple risque d'une maladie). (Figure 1.19).

FIGURE 1.19. – Diagramme GWAS. Chaque position sur laquelle la variation est associée à un trait est représentée en cercle et la couleur représente le type de trait que chaque variation affecte. Les variations sont réparties sur l'ensemble du génome et que pour chaque phénotype, on distingue un grand nombre de milliers de variations à risque (Source : https ://www.ebi.ac.uk/gwas/docs/diagram-downloads).

Une grande majorité des variants (généralement les SNPs) associés à une susceptibilité à des maladies complexes identifées par le biais des GWAS sont situés dans des régions non codantes du génome (EDWARDS, BEESLEY, FRENCH et al., 2013). On a émis l'hypothèse que ces variants affecteraient la régulation des gènes, notamment via la variation de la liaison aux facteurs de transcription (REDDY, GERTZ, PAULI et al., 2012, KILPINEN, WASZAK, GSCHWIND et al., 2013).

Une des difficultés majeure rencontrée est l'établissement de la causalité entre variants associés et maladie / traits complexes, en particulier pour identifier les variants pertinents du point de vue fonctionnel dans les régions non codantes (TODOROVIC, 2016, SAINT PIERRE et GÉNIN, 2014). Pour le petit nombre de cas où les variants causaux ont été validés expérimentalement, il a été démontré qu'ils : (a) perturbaient les sites de liaison des facteurs de transcription, (b) la structure de la chromatine locale ou (b) le recrutement de co-facteurs, entraînant finalement des modifications du résultat de la transcription du ou des gènes cibles.

Malgré l'intérêt croissant et les ressources disponibles pour étudier ces variants, comprendre leur effet fonctionnel reste difficile pour plusieurs raisons : (i) la plupart des SNPs associés sont encore identifiés par réseaux de génotypage à l'échelle du génome, ce qui ne permet pas d'étudier tous les variants mais seulement les

SNPs marqueurs ; (ii) il est nécessaire d'étudier le bon type de cellule dans le bon environnement pour découvrir le mécanisme d'action d'un variant car l'expression des gènes et la liaison aux facteurs de transcription varient à travers les tissus et les conditions (Albert et Kruglyak, 2015).

Les premières GWAS fonctionnaient dans l'espoir que les variantes sous-jacentes aux traits complexes auraient des effets de taille moyenne à grande et seraient détectables dans des échantillons relativement petits. Bien que cela soit vrai pour certains phénotypes, tels que la dégénérescence maculaire liée à l'âge et le glaucome exfoliant (Thorleifsson, Magnusson, Sulem et al., 2007, Haines, Hauser, Schmidt et al., 2005), il est devenu évident que, pour la majorité des traits complexes, des échantillons beaucoup plus grands sont nécessaires pour obtenir la puissance statistique nécessaire pour détecter les faibles effets des locus à risque. Il est maintenant reconnu que, dans la plupart des cas, les locus de risque identifiés pourraient n'expliquer qu'une petite partie de la variance des traits (c'est-à-dire le problème d'héritabilité manquant) (Manolio, Collins, Cox et al., 2009). Ainsi, on a élargi le champ d'application de l'utilisation d'un seuil de signification standard à l'échelle du génome corrigé par Bonferroni de $5x10^{-8}$ à l'utilisation d'autres indicateurs d'association (Yang, Lee, Goddard et al., 2011).

Le variant génétique associé à la maladie n'est souvent pas le variant causal. L'allèle du variant est associé à la maladie si sa fréquence est élevée dans la population de cas que de témoins, et ceci n'étant pas le simple hasard. Les variants étudiés n'étant souvent pas directement fonctionnels, mais uniquement en déséquilibre de liaison avec le variant causal qui n'est pas observé (Figure 1.20) (Debette, 2012).

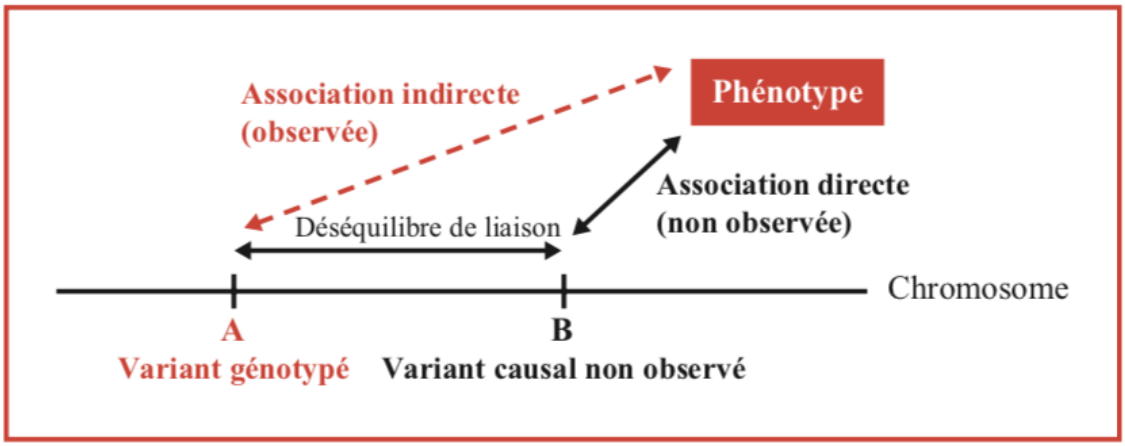


Figure 1.20. – Distinction entre association génétique et relation causale. Le variant associé indirectement au phénotype est en déséquilibre de liaison génétique avec le variant causal (association directe) sur le même haplotype (Source : Debette, 2012).

On dit qu'il y a déséquilibre de liaison si la fréquence de co-occurrence des allèles de deux loci différents A et B est différente de ce que donnerait une association aléatoire de ces allèles. Autrement dit, c'est un signe qu'il y a association préférentielle entre deux allèles.

La fréquence jointe de la liaison est :

$$f_{AB} = f_A \cdot f_B$$

où f_{AB} indique la fréquence jointe de deux allèles, et f_{AB} l'espérance de cette fréquence jointe sous hypothèse d'indépendance des allèles A et B.

La différence entre fréquence observée et fréquence attendue est :

$$D = f_{AB} - f_A \cdot f_B$$

Le déséquilibre de liaison (LD) peut-être calculé par le coefficient de corrélation r^2. En considérant $f_A, f_a, f_B et f_b$ comme fréquences alléliques des allèles AB, Ab, aB et ab les fréquences des haplotypes, le r^2 est défini de la manière suivante :

$$r^2 = \frac{D^2}{f_A.f_a.f_B.f_b}$$

1.3. Facteurs génétiques liés à la susceptibilité/résistance de l'hôte au paludisme

Les manifestations cliniques du paludisme dépendent des paramètres du parasite, de l'hôte, de l'environnement et des facteurs sociaux. Ceux-ci convergent pour aboutir à des conséquences variables, pouvant aller d'une infection asymptomatique à la maladie grave et la mort. La forte pression de sélection exercée par la diversité génétique de l'hôte a contraint le parasite à une meilleure adaptabilité et a sélectionné la mise en place de la variation antigénique (Deroost, Pham, Opdenakker et al., 2016).

Plusieurs études démontrent une association entre les variants génétiques et divers phénotypes liés susceptibilité/résistance au paludisme (López, Saravia, Gomez et al., 2010 ;Hill, 1999). Les premiers facteurs génétiques ayant été identifiés sont impliqués dans la physiologie des érythocytes (HbS, HbC, l'alpha-globuline, G6PD) et dans la réponse immunitaire (certains gènes du système HLA, TNF-alpha, IFN-gamma, NCR3, ...). Un grand nombre d'études d'épidémiologie génétique ont permis par la suite d'identifier des gènes associés aux différents phénotypes du paludisme (niveau de densité parasitaire, accès palustre simple, accès palustre grave) (Hernandez-Valladares, Rihet et Iraqi, 2014). La plupart de ces études se sont basées sur des approches gènes candidats sélectionnés a priori pour leur rôle potentiel dans la maladie (majoritairement impliqués dans le contrôle de la réponse immunitaire) et montrent des résultats souvent discordants. Des études de liaison, étudiant la co-transmission des marqueurs de la maladie au sein des familles, ont également mis en évidence l'implication de la région 5q31-5q33 dans la réduction de la parasitémie asymptomatique (Rihet, Traoré, Abel et al., 1998).

Les bases moléculaires de la susceptibilité/résistance au paludisme restant encore très mal connues à ce jour. L'hémoglobine S, facteur génétique le plus important connu jusqu'à présent représente seulement 2% de forme de résistence aux formes sévères du paludisme. Il est nécessaire d'étendre l'étude des facteurs génétiques de l'hôte dans la susceptibilité/résistance au paludisme à l'ensemble du génome par des études GWAS.

1.3.1. Gènes impliqués dans la résistance aux accès palustre simple

Les accès palustres simples sont déterminés par une parasitémie de *P. falciparum* dans le sang, par la fièvre et les symptômes cliniques tels que les céphalés, les douleurs, les vomissements ou la diarrhée. La susceptibilité aux épisodes de fièvres palustre est liée en partie par des facteurs génétiques (Jepson, Banya, Sisay-Joof et al., 1995). Ainsi, certaines études pangénomiques montrent par exemple que la parasitémie asymptomatique et les accès palustres simples sont liés au locus 6p21-p23 (Flori, Sawadogo, Esnault et al., 2003, Brisebarre, Kumulungui, Sawadogo et al., 2014). Parmi les gènes identifiés dans ces études, on peut noter les gènes TNF (*Tumor Necrosis Factor*) et NCR3 (*Natural Cytotoxicity Triggering Receptor 3*).

1.3.1.1. *Tumor Necrosis Factor* (TNF)

TNF ou facteur de nécrose tumorale code pour une cytokine TNF-a produite par les monocytes / macrophages et lymphocytes T. TNF-a joue un rôle important dans la réponse de l'hôte à l'inflammation et infection. Plusieurs études ont mis en évidence une association du TNF avec les infections de *P. falciparum*. Une étude menée au Burkina Faso sur contribution des variations génétiques du TNF aux phénotypes du paludisme a montré que les polymorphismes du TNF (TNF-1031, TNF-308, TNF851 et TNF1304) étaient liés à des épisodes non compliqués de paludisme (Flori, Delahaye, Iraqi et al., 2005). Une autre étude faite chez les enfants Congolais à montré une association de TNF1304, TNF-308, TNF-238 et TNF-244 avec paludisme simple et la parasitémie symptomatique maximale.

1.3.1.2. *Natural Cytotoxicity Triggering Receptor 3* (NCR3)

C'est un gène essentiel dans la biologie des cellules NK. En effet, NCR3 code pour un récepteur activateur exprimé à la surface des cellules NK. Les cellules NK sont capables de reconnaître directement les globule rouges infectés par *P. falciparum* via les récepteurs NKp30 (Mavoungou, Held, Mewono et al., 2007). Un polymorphisme nucléotidique unique dans la région non traduite 5 'de ce gène a été associé à une légère susceptibilité au paludisme (Delahaye, Barbier, Fumoux et al., 2007, Baaklini, Afridi, Nguyen et al., 2017). Les différentes études de Mavoungou et al. (Mavoungou, Held, Mewono et al., 2007) et Pende (Pende, Parolini, Pessino et al., 1999), d'une part sur la reconnaissance des antigènes plasmodiaux par les récepteurs NKp30 et d'autre part sur l'effet de leur sous expression sur la cytotoxicité des cellules NK indiquent que les récepteurs NKp30 jouent un rôle important dans la réponse immunitaire de l'hôte infecté par Plasmodium. Parallèlement à ces études fonctionnelles, une étude génétique menée au Burkina Faso a montré qu'il existait une association entre le polymorphisme rs2736191 (NCR3-412C) et la forme simple de paludisme (Delahaye, Barbier, Fumoux et al., 2007).

1.3.2. Gènes impliqués dans le contrôle de la parasitémie

Le contrôle génétique de la parasitémie est complexe et implique plusieur facteurs génétiques (Rihet, Traoré, Abel et al., 1998, Garcia, Cot, Chippaux et al., 1998). Une étude de liaison pangénomique portant sur des phénotypes relatifs à la parasitémie, a été conduite à Dielmo au Sénégal sur 190 familles nucléaires et a mis en évidence une liaison significative entre la région chromosomique 5q31-q33 et la densité parasitaire moyenne asymptomatique, confirmant ainsi le résultat de 3 précédentes études de liaison génétique indépendantes (d'approche région-candidate) (Rihet, Traoré, Abel et al., 1998, Flori, Sawadogo, Esnault et al., 2003). Une étude menée au Cameroun avait déjà montré une liaison suggestive entre la région 5q31 et la parasitémie dans 9 familles nucléaires (Marquet, Garcia, Fievet et al., 1998). An niveau de la région du complexe majeur d'histocompatibilité (CMH), on a découvert que le locus 6p21-p2 est associé à une liaison suggestive avec la parasitémie asymptomatique dans une étude pangénomique menée auprès de 63 familles du district urbain Logoforousso et de 55 autres familles vivant dans la zone

rurale de Bobo Dioulasso au Burkina Faso (Brisebarre, Kumulungui, Sawadogo et al., 2014). Le gène du TNF a été associé à la parasitémie symptomatique et asymptomatique, et plus précisément les polymorphismes TNF-1031, TNF-308, TNF-238, TNF1304 et TNF851 (Flori, Delahaye, Iraqi et al., 2005). Un lien a été identifié entre l'intensité de l'infection plasmodiale et un polymorphisme du gène ARHGAP26 (*Rho GTPase activating protein 26*) dans la région 5q31-q33 (Milet, Nuel, Watier et al., 2010). Ce gène code pour une protéine impliquée dans la voie de transduction du signal d'intégrine (Ohta, Goto, Matsumoto et al., 2016) et pourrait ainsi jouer un rôle dans la stabilité des interactions ICAM-1/LFA-1 entre les globules rouges infectés et les cellules NK. Différents polymorphismes du gène IL-12 (IL-12pro, rs2288831, rs3213099, rs3181221, rs3213105, rs11574790, IL-12B 3'UTR) ont été trouvés lié à la parasitémie dans la population urbaine de Bobo-Dioulaso au Burkina Faso (Barbier, Atkinson, Fumoux, & Rihet, 2008). Il existe bien d'autres études sur les gènes impliqués dans le contrôle de la parasitémie qui ne sont pas mentionnés dans ce chapitre.

1.3.3. Gène impliqués dans la résistance aux formes sévère du paludisme

On distingue plusieurs études génétiques sur les phénotypes du paludisme sévère qui inclu le paludisme cérébral, l'anémie sévère et la détresse respiratoire. Cependant, peu d'études d'association à l'échelle du génome (GWAS) portant sur l'évidence entre des régions et polymorphismes d'intérêt dans le cas de l'infection palustre grave ; et celles qui ont été conduites n'ont pas été en mesure de répliquer des variants admis comme influençant l'évolution de la maladie.

Des études de GWAS ont récemment été mis en œuvre dans les zones d'endémie palustre et ont répliqué les variants bien connues, notamment le système de groupe sanguin ABO et l'hémoglobine S (Jallow, Teo, Small et al., 2009 ;Ravenhall, Campino, Sepúlveda et al., 2018), ce qui renforce l'importance des variants érythrocytaires pour la protection contre le paludisme sévère. Malgré ce succès, seules quelques variants originaux ont été identifiées suggérant que certains des vrais signaux d'association ont peut-être été atténués. Cela peut être dû à plusieurs facteurs de confusion tels que : la grande diversité génétique de la population dans les zones d'endémie palustre, les différences de fréquence et la taille des effets des allèles protecteurs d'une population à l'autre et les limites inhérentes aux approches GWAS, entre autres. Une observation intéressante dans ces études est l'identification de plusieurs signaux d'association répartis dans le génome qui n'ont pas dépassé le seuil de significativité des GWAS ($p < 5x10^{-8}$) ; suggérant l'existence d'effets polygéniques (Jallow, Teo, Small et al., 2009 ;Ravenhall, Campino, Sepúlveda et al., 2018).

Quelques nouveaux variants associés au paludisme sévère ont été découvert dans les gènes suivants :

1.3.3.1. *ATPase plasma membrane Ca2+ transporting 4* (ATP2B4)

L'association du gène ATP2B4 au paludisme grave a été identifiée pour la première fois par Timmann et al. (Timmann, Thye, Vens et al., 2012) dans

les populations ghanéennes. Dans cette étude, plusieurs SNP du gène ATP2B4, situés sur le chromosome 1q32, ont montré des associations significatives. Le rôle protecteur des variants du gène ATP2B4 a ensuite été signalé dans d'autres études12,23. Le gène ATP2B4 code pour une ATPase4 (PMCA4b) (Guerini, Pan et Carafoli, 2003), transportant le calcium, sur une membrane plasmique omniprésente. PMCA4b est largement exprimé dans la majorité des tissus humains et constitue le principal transporteur de Ca2+ dans la membrane érythrocytaire (Guerini, Pan et Carafoli, 2003). Une étude récente a montré que les individus possédant un haplotype précédemment non reconnu appelé « haplotype1 » dans leur gène ATP2B4 présentaient un niveau d'expression réduit de PMCA4b (Zámbó, Várady, Padányi et al., 2017). Plusieurs SNPs impliqués dans la protection contre le paludisme sont localisés dans cet haplotype. Timmann et al. (Timmann, Thye, Vens et al., 2012) ont proposé que la perturbation de l'homéostasie du Ca2+ intracellulaire provoquée par une régulation altérée de PMCA4b puisse jouer un rôle important dans l'interférence avec les effets pathologiques du paludisme dans différents tissus. La perturbation de l'homéostasie intracellulaire de Ca2+ dans les érythorocytes pourrait entraver le développement et la reproduction du parasite dans une membrane vacuole parasitophore (PVM) (Gazarini, Thomas, Pozzan et al., 2003). La dérégulation du Ca2+ intracellulaire dans d'autres tissus tels que le placenta, les plaquettes et le cerveau pourrait modifier les phénotypes d'adhésion et par conséquent minimiser les pathologies associées à l'adhérence (Adams, Kuhnrae, Higgins et al., 2014).

Les mécanismes de protection des variants du gène ATP2B4 pourraient également être associés à une régulation altérée de l'oxyde nitrique (NO), l'une des molécules importantes de la pathogenèse du paludisme et de ses protections (Sobolewski, Gramaglia, Frangos et al., 2005). L'augmentation du niveau de NO était impliquée dans la protection contre le paludisme cérébral alors qu'elle augmentait le risque de paludisme non compliqué (Turner, Morrison, Jones et al., 1994). Dans les cellules neuronales, PMCA4 module l'activité de la synthèse d'oxyde nitrique neuronale (nNOS), laquelle régule à son tour la biodisponibilité du NO de manière dépendante du calcium (Trovoada, Martins, Ben Mansour et al., 2014). Ainsi, la production réduite de PMCA4 pourrait affecter les activités de l'enzyme de synthèse de l'oxyde nitrique dans le cerveau, ce qui pourrait affecter la concentration de NO et donc la protection contre le paludisme grave.

1.3.3.2. Groupe des 3 gènes de glycophorine (GYPA/B/E)

Band et al. (Band, Le, Jostins et al., 2013) ont mené à bien la plus grande GWAS de prédilection pour le paludisme multicentrique comprenant onze populations. Dans cette étude, 34 régions génomiques contenant des loci de susceptibilité potentiels au paludisme grave ont été identifiées. Parmi ceux-ci, un signal fort a été observé au locus entre le gène FREM3 et le groupe de 3 gènes de glycophorine (GYPA/B/E) sur le chromosome 4. Un haplotype (au niveau du SNP rs184895969) réduit le risque de développer un paludisme grave d'environ 40% et est fréquent dans les populations kényanes avec une fréquence d'allèle d'environ 10% (Band, Le, Jostins et al., 2013).

Outre les signaux proches des gènes de glycophorine, six autres loci présumés

jouant un rôle clé en biologie membranaire ont été identifiés dans cette étude (Band, Le, Jostins et al., 2013). Une étude ultérieure dans les mêmes populations a identifié un grand nombre de variants du nombre de copies caractérisés par la suppression, la duplication et les structures hybrides des gènes GYPA et GYPB (Leffler, Band, Busby et al., 2017) dont un variant distincte appelée DUP4 réduirait le risque de paludisme grave 40% dans les populations d'Afrique orientale (Kenya). Une caractérisation plus poussée a montré que ce variant est composée d'un complexe hybride GYPB-A et code pour l'antigène Dantu dans le système de groupes sanguins MNS (Leffler, Band, Busby et al., 2017). L'association de cette région avec le paludisme sévère a été corroborée par une autre étude récente portant sur des populations tanzaniennes (Manjurano, Sepúlveda, Nadjm et al., 2015). Les groupes de gènes de glycophorine, GYPA et GYPB, codent pour le système de groupes sanguins MNS et sont connus pour être des récepteurs de *P. falciparum* lors de l'invasion des globules rouges (Ko, Kaercher, Giombini et al., 2011). GYPA et GYPB servent de récepteur membranaire des érythrocytes pour les protéines EBA-175 et EBL-1 du parasite, respectivement (Mayer, Cofie, Jiang et al., 2009). On sait également que cette région génomique est soumise à une ancienne pression sélective résultant de la course des bras hôte-pathogène entre *P.falciparum* et l'homme 35. Une analyse fonctionnelle plus approfondie est nécessaire pour mieux comprendre comment ces variants affectent l'invasion et / ou le développement des parasites dans les érythrocytes et assurent une protection contre le paludisme grave.

1.3.3.3. *Synthesis of cytochrome C oxidase 1* (SCO1) et *dihydroxyphényl-alanine décarboxylase* (DDC)

Des signaux d'association ont été identifiés par le premier GWAS de prédisposition au paludisme mené dans la population gambienne (Jallow, Teo, Small et al., 2009). Le premier SNP principal (rs6503319) est situé à proximité du gène SCO1 (synthèse de la cytochrome c oxydase) sur le chromosome 17p13. SCO1 est une protéine de signalisation multifonctionnelle qui joue un rôle essentiel dans les voies d'administration du cuivre par la cytochrome c oxydase (COX) mitochondriale36. La COX catalyse le transfert d'électrons du cytochrome c réduit en oxygène et est abondamment exprimée dans les muscles, le cerveau et le foie (Leary, 2010). Une carence en COX provoquée par des mutations du gène SCO1 peut entraîner une détresse respiratoire et une acidose métabolique sévère (Valnot, Osmond, Gigarel et al., 2000), qui sont également les principales complications du paludisme cérébral (Miller, Baruch, Marsh et al., 2002) Des études complémentaires sont nécessaires pour comprendre comment les variants du gène SCO1 sont associés aux voies pathologiques du paludisme cérébral.

Le deuxième signal d'association notable identifié dans cette étude était la dihydroxyphényl-alanine décarboxylase (DDC) sur le chromosome 7p12.2. Une étude récente portant sur des populations tanzaniennes a répliqué l'association de variants de DDC avec le paludisme cérébral (Manjurano, Sepúlveda, Nadjm et al., 2015). Le gène de DDC code pour l'enzyme décarboxlase Aromatic-L-amino-acid qui est impliqué dans la biosynthèse de neurotransmetteurs tels que la dopamine et la sérotonine (Martelle, Raffield, Palmer et al., 2016). Le DDC est une

enzyme essentielle pour le développement du cerveau et du système nerveux et sa déficience est associée à une réduction des fonctions cognitives (Martelle, Raffield, Palmer et al., 2016). Le DDC est impliqué dans l'immunité cellulaire et contribue à la protection contre les maladies parasitaires chez les invertébrés (Zhou, Yang, Wang et al., 2011). En outre, des mutations du gène DDC seraient associées à la réfractarité du moustique Anopheles gambiae contre les parasites P.falciparum41.

1.3.3.4. *MARVEL domain containing 3* (MARVELD3)

En plus de l'ATP2B4, Timmann et al. (Timmann, Thye, Vens et al., 2012) ont identifié une association du SNP (rs2334880) sur le chromosome 16p 22.2 qui est liée à MARVELD3. Cependant, cette association n'a pas été reproduite dans d'autres études. MARVELD3 est l'un des composants des protéines à jonction serrée dans plusieurs tissus épithéliaux et endothéliaux et est exprimé sous la forme de deux variants épissés alternatives (Steed, Rodrigues, Balda et al., 2009). Ces protéines sont impliquées dans l'assemblage, le développement, la maintenance et la régulation des jonctions serrées. Les jonctions serrées jouent un rôle majeur dans les adhérences intracellulaires et sont impliquées dans les mécanismes de signalisation cellulaires (Steed, Rodrigues, Balda et al., 2009).

1.3.3.5. Récepteurs interleukine (IL-12, IL-23)

La plus récente GWAS de prédisposition au paludisme a été réalisée dans la population tanzanienne (Ravenhall, Campino, Sepúlveda et al., 2018). Dans cette étude, des signaux d'association ont été identifiés dans les voies immunitaires, notamment dans les récepteurs de l'interleukine (IL-23R et IL-12RBR2), dans les protéines de type ketch (KLHL3) et dans les régions de l'antigène de leucocyte humain (HLA). L'IL-12 est formé à partir d'un hétérodimère de IL12B (sous-unité ILp40) et de IL-12A (sous-unité ILp35) (Huang, Cancilla et Morahan, 2000). L'IL-12 joue un rôle vital dans la stimulation de la réponse immunitaire à médiation cellulaire contre les agents pathogènes intracellulaires en se liant à des complexes de récepteurs IL-12RB1 et IL-12RBR2 de haute affinité. Les complexes de récepteurs 12RB1 et IL-12RBR2 favorise le développement des lymphocytes T auxiliaires (Th1) et améliore la production d'INF-gamma, connus pour jouer le rôle de médiateur dans la clairance des agents pathogènes intracellulaires (Huang, Cancilla et Morahan, 2000). Dans le paludisme, l'IL-12 a été impliquée dans la médiation de l'immunité protectrice chez les animaux de laboratoire et chez l'homme (Luty, Perkins, Lell et al., 2000). L'IL-23 est une cytokine pro-inflammatoire importante qui partage les sous-unités p40 avec l'IL12 (Duvallet, Semerano, Assier et al., 2011). Il induit la différenciation des cellules T CD4 naïves en IL-17, qui joue un rôle clé dans la pathogenèse des maladies auto-immunes (Blanden, Doherty, Dunlop et al., 1975). le HLA est codé par le CMH, qui représente les gènes les plus polymorphes connus dans le génome humain. On pense que la diversité des CMH résulte de la pression de sélection exercée par des agents pathogènes infectieux et que l'on sait qu'elle est associée au risque de plusieurs maladies infectieuses [(Hill, Allsopp, Kwiatkowski et al., 1991) ; hill_genetic_1997]. Des variants HLA tels que l'antigène HLA de classe I (HLA-Bw53) et le variant HLA de classe II (DRB1

* 1302-DQB1 * 0501) conféraient une protection contre le paludisme grave chez les populations gambiennes47. L'antigène HLA de classe I est exprimé par les cellules hépatiques, ce qui suggère que les réponses des cellules T (CTL) pourraient agir efficacement contre le stade hépatique du parasite du paludisme chez les personnes atteintes de HLA-Bw5347. De même, les individus porteurs du variant DRB1 * 1302-DQB1 * 0501 pourraient posséder un mécanisme de présentation antigénique efficace pouvant conduire à une clairance rapide du parasite au stade sanguin (Hill, Allsopp, Kwiatkowski et al., 1991).

Les variants des voies immunitaires présentent un grand intérêt en raison de leur potentiel d'information pour la mise au point de vaccins efficaces contre le paludisme (Kwiatkowski, 2005).

1.4. Approches bioinformatiques de la régulation transcriptionnelle

1.4.1. La régulation transcriptionnelle

Le transfert d'information à partir d'ADN sur un organisme multicellulaire est une question d'une grande importance en biologie. Comment la cellule a-t-elle le potentiel, à partir du génome, de développer différents tissus, organes et systèmes ? Les connaissances actuelles nous permettent de comprendre comment cette information circule dans le processus biochimique, de l'ADN à l'ADN (réplication), d'ADN à ARN (transcription), d'ARN à ADN (rétro-transcription) et d'ARN à protéine (traduction) (Figure 1.21).

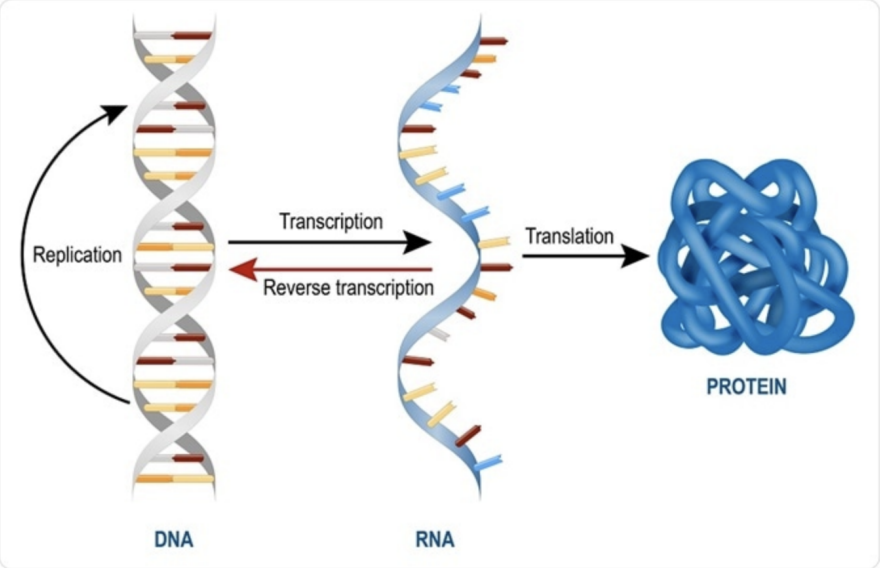


FIGURE 1.21. – Transfert d'information entre macromolécule (Source : https ://www.news-medical.net/life-sciences/-Types-of-RNA-mRNA-rRNA-and-tRNA-(French).aspx).

En 1960, Jacob et Monod ont été les premiers à travailler sur les mécanismes de la régulation des gènes. Ils ont démontré que la synthèse des protéines à partir de la transcription de l'ADN est médiée par les protéines appelées répresseurs, qui exercent un contrôle négatif sur l'expression des gènes, par l'intermédiaire de courtes séquences spécifiques d'ADN appelées opérateurs (JACOB, PERRIN, SANCHEZ et al., 1960). Cette découverte a ouvert un champ de recherche en biologie moléculaire (la régulation transcriptionnelle). Plusieurs années après, on découvert qu'il existait un autre type de protéines régulatrices appelées activateurs, qui contrairement aux répresseurs, peuvent réguler positivement l'expression des gènes. Ces deux types de protéines (répresseurs et activateurs) sont connues de nos jours comme **facteurs de transcription**.

La transcription de l'ADN est la première étape de l'expression génique. La transcription produit un ARN, qui peut servir de modèle pour la synthèse d'une protéine (Figure 1.21), ou bien exercer d'autres fonctions en tant qu'ARN (tARN,

rARN, micro-RNA, lncRNA, ...).

Le niveau de transcription est chez les bactéries et les métazoaires, les régions de l'ADN où se fixent les TFs sont en amont près du gène sont appelées promoteurs, d'où la régulation transcriptionnelle est dictée par davantage d'éléments régulateurs.

1.4.1.1. Les bases de la transcription

L'initiation de la transcription se fait au niveau du site d'initiation (*Transcription Start Site*, TSS) sous le contrôle du promoteur et d'autre éléments de régulation. Le produit de la transcription est un ARN immature qui subit ensuite diverses modifications pour produire un ARNm mature. L'ARNm produit sera ensuite traduit pour produire une protéine.

Dans tout organisme vivant (des bactéries aux métazoaires), la transcription est réalisée par un complexe protéique appelé ARN polymérase (*RNAP*) qui présente une affinité pour les séquences courtes situées dans le promoteur (par exemple, la boîte TATA, les éléments BRE). Ces courtes séquences chez les promoteurs sont capables de recruter l'ARN polymérase. Un promoteur capable de démarrer seul la transcription de son gène en aval (c'est-à-dire de recruter l'ARN polymérase sans l'aide d'autres éléments) est considéré comme promoteur puissant. En revanche, certains promoteurs sont qualifiés de "faibles" (ou co-promoteurs) car leur affinité pour le RNAP n'est pas suffisante pour commencer la transcription et la RNAP nécessite l'aide des TFs pour se stabiliser ou recruter certaines de ses sous-unités (Qin, Zhang, Clift et al., 2010) (Figure 1.22).

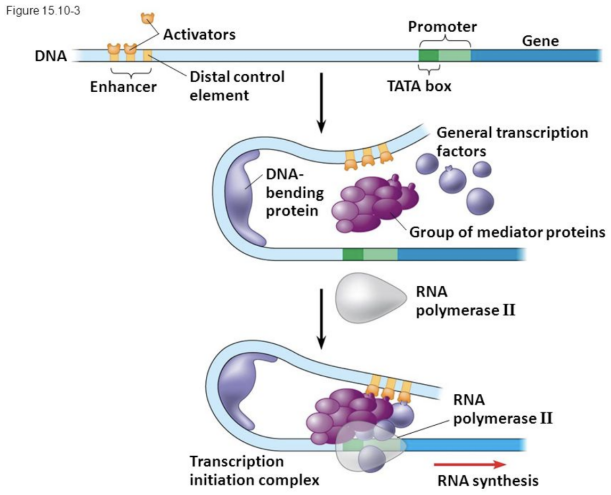


Figure 1.22. – Mécanisme de la transcription. (source : https ://slide-player.com/slide/10605177/).

1.4.1.2. Les facteurs de transcription

Voici la définition des facteurs transcriptionnels dans la *Concise Encyclopaedia of Bioinformatics* (*Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition*) :

"Les facteurs de transcription (TFs) sont de protéines affectant le niveau de transcription d'un ensemble spécifique de gènes. Un TF est qualifié d'activateur ou

de répresseur selon qu'il augmente ou diminue l'expression de son ou de ses gènes cibles. Il convient de noter que le qualificatif activateur / répresseur s'applique à l'interaction entre le TF et un gène donné plutôt qu'au le TF lui-même, puisqu'un facteur peut activer certains gènes et en réprimer d'autres" (voir Figure 1.23).

Par exemple le facteur RUNX1, impliqué dans la différenciation des cellules sanguines, peut agir soit comme activateur dans certaines conditions, soit comme répresseur dans d'autres (Whitfield, Wang, Collins et al., 2012).

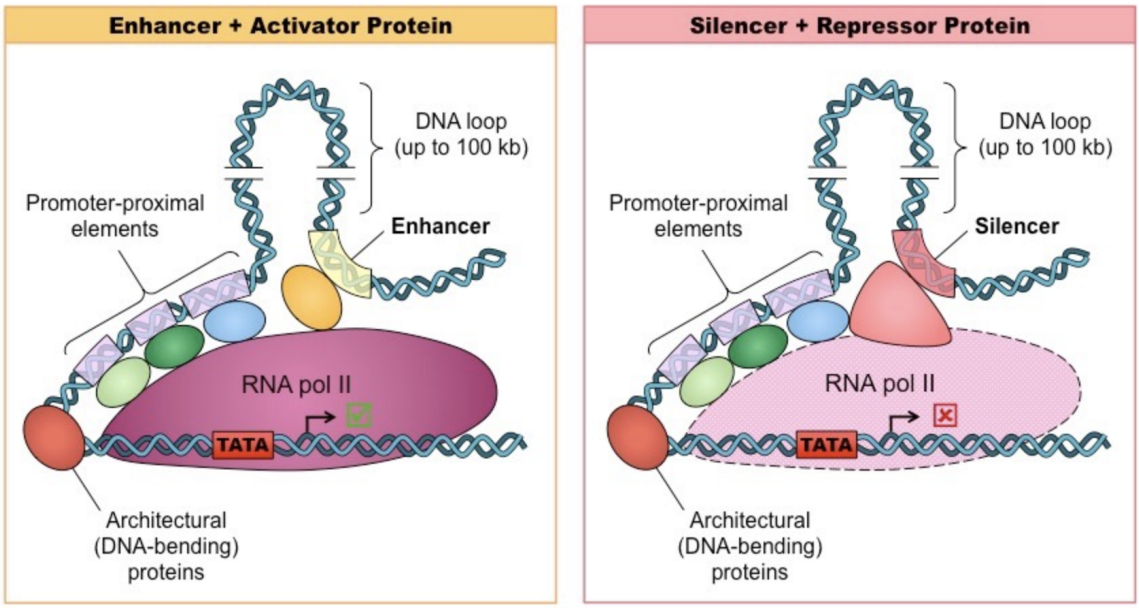


Figure 1.23. – Mécanisme de l'activation et de la repression par les TFs. (source : https ://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/72-transcription-and-gene/gene-expression.html).

On distingue différents types de facteurs transcriptionnels.

"Les TFs sont qualifiés de spécifiques ou globaux selon qu'ils agissent sur un nombre limité de gènes (la limite entre les facteurs spécifiques et les facteurs globaux est quelque peu arbitraire)" (*Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition*).

"Les TFs se liant à l'ADN agissent en se liant à des emplacements génomiques spécifiques (courtes séquences de dont la longueur varie entre 6 et 20 pb), appelés sites de liaison du facteur de transcription (TFBS)".

"Un TF peut également agir indirectement sur l'expression de son gène-cible en interagissant avec d'autres TFs (co-facteurs). Par exemple, le répresseur de levure Gal80p ne se lie pas à l'ADN, mais interagit avec le facteur Gal4p de liaison à l'ADN et l'empêche d'activer ses gènes cibles" (voir Figure 1.24) (*Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition*).

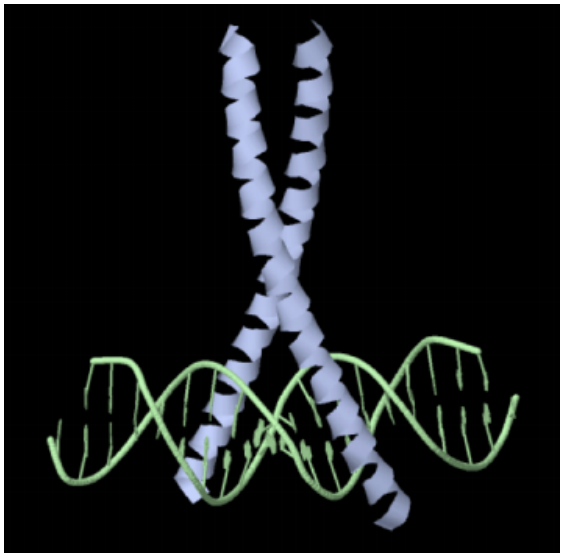


FIGURE 1.24. – Structure 3D de Gcn4p de Saccharomyces cerevisiae. (source : PDB 2DGC https ://www.rcsb.org/structure/2DGC).

Séquences cis-régulatrices Les TFs se lient à des séquences courtes spécifiques dans des régions ouvertes de l'ADN où ils peuvent réguler la transcription. Si ces régions sont situées près du TSS d'un gène, on parle de promoteurs. S'ils activent l'expression génique à distance (par rapport au TSS du gène-cible) et indépendamment de leur orientation, on parle d'amplificateurs (*enhancers*) (BANERJI, RUSCONI et SCHAFFNER, 1981). S'ils répriment l'expression génique à distance, ils sont appelés "silenceurs" (*silencers*) ; s'ils délimitent l'euchromatine (chromatine qui apparaît partiellement décondensée en interphase) de l'hétérochromatine (structure observable de l'ADN, une condensation de la chromatine que l'on distingue de l'euchromatine qui est la structure observable non condensée) et isolent l'activité de l'activateur sont appelés isolants (*silencers*) (KOLOVOS, KNOCH, GROSVELD et al., 2012) (voir Figure 1.23).

Chez les métazoaires, les TFs sont généralement concentrés dans les régions cis-régulatrices (promoteurs et amplificateurs, par exemple). Certaines régions d'ADN présentent un regroupement de sites de TFs : on les appelle modules cis-régulateurs (*Cis-Regulatory Modules*, CRM) (Figure 1.25).

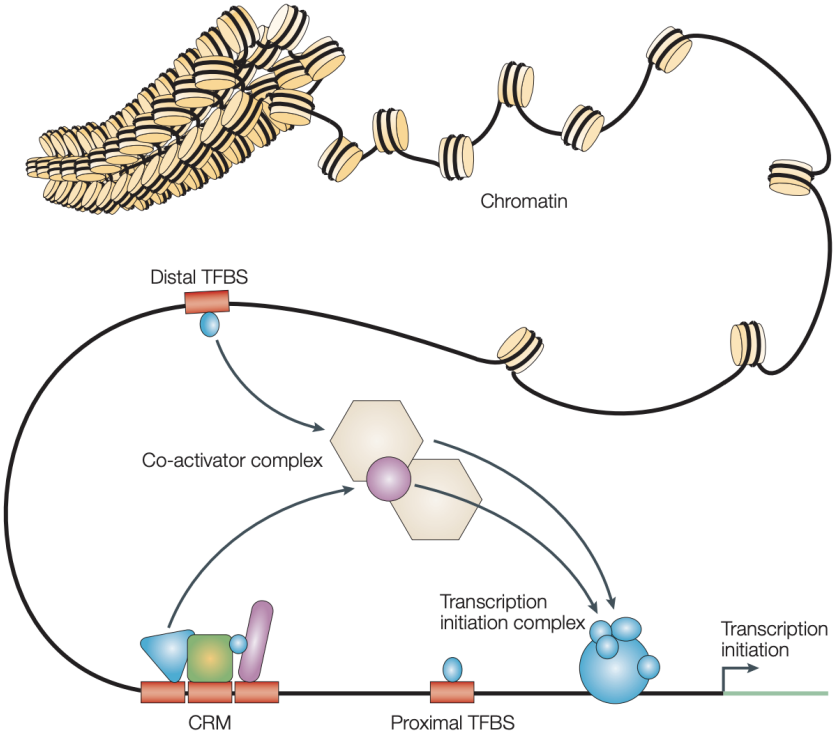


FIGURE 1.25. – Composants de la régulation transcriptionnelle. Les TFs se lient à des sites de liaison du facteur de transcription (TFBS) spécifiques qui sont soit proximaux, soit distaux par rapport au site d'initiation de la transcription. Des ensembles de TFs peuvent former des modules cis-régulateurs (cis-regulatory modules, CRM) qui permettent une régulation fine en intégrant les informations de plusieurs facteurs. Les interactions entre les TFs liés et les cofacteurs stabilisent la machinerie d'initiation de la transcription pour permettre l'expression génique (source : WASSERMAN et SANDELIN, 2004).

1.4.1.3. Caractérisation expérimentale de sites de liaison d'un facteur de transcription

L'identification des sites de liaison des TFs est d'une grande importance dans la compréhension de la régulation transcriptionnelle. Les TFs peuvent reconnaître directement des motifs de séquences spécifiques sur la molécule d'ADN. Certaines méthodes caractérisent des sites de liaison (position génomique) alors que d'autres caractérisent des liaisons qui pourraient ou ne pas exister dans le génome.

Retard sur gel Le retard sur gel (*Electrophoretic Mobility Shift Assays* , EMSA) est une technique d'étude des interactions ADN-ARN ou ADN-protéine (CANN, 1998). Il peut être utilisé pour déterminer si une protéine (ou un ensemble de protéines, par exemple plusieurs TFs regroupés sur un CRM) est liée à une séquence d'ADN d'intérêt. La logique de cette méthode est que la séquence d'intérêt est caractérisée par un poids moléculaire qui peut être visualisé sous la forme d'une

bande détectée par électrophorèse sur un gel de Poly acrylamide ou d'agarose. La bande correspondant au fragment d'ADN "nu" (non lié à des protéines) est utilisée comme contrôle (Cann, 1998). La vitesse de migration étant inversement proportionnelle au poids de la molécule, la fixation de facteurs transcriptionnels ralentit la migration du fragment d'ADN. En comparant la migration de l'ADN complexé avec l'ADN seul, la technique permet d'identifier la région de fixation d'un facteur trans donné (Figure 1.26). La méthode EMSA permet par exemple d'isoler les protéines qui se fixent à un fragment d'ADN, de mesurer l'affinité d'une protéine pour une séquence d'ADN ou encore de déterminer le domaine de fixation de l'ADN sur une protéine. L'approche EMSA est limitée par le fait que seuls quelques sites peuvent être évalués en une seule fois et que cela nécessite une connaissance préalable de la séquence à évaluer.

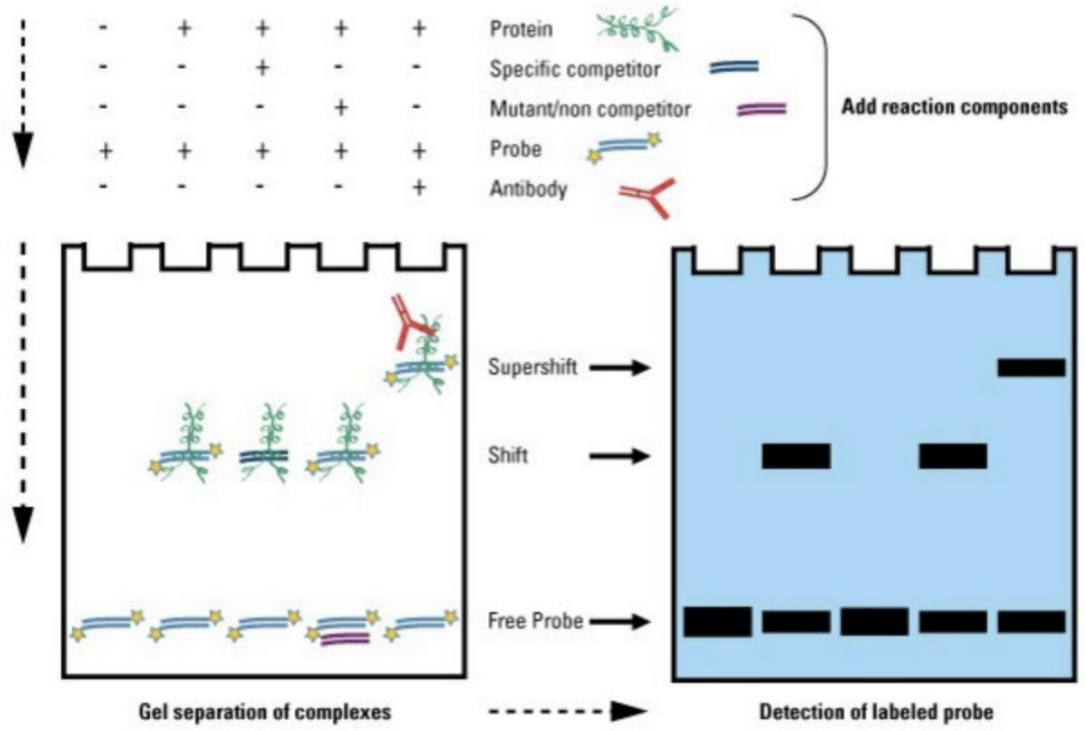


Figure 1.26. – Aperçu de la méthode EMSA. Le test de décalage sur gel comprend trois étapes principales : (1) les réactions de liaison, (2) l'électrophorèse, (3) la détection par sonde (source : https ://www.thermofisher.com).

Empreinte à la DNase Bien que l'EMSA puisse être utilisée pour déterminer la présence / l'absence de protéines liées dans une séquence donnée, il ne détecte pas l'emplacement précis du TFBS. Pour ce faire, une autre méthode peut être utilisée, à savoir le test d'empreinte à la DNAse (DNAse footprinting), qui tire parti des propriétés moléculaires de la désoxyribonucléique (DNAse), une enzyme qui dégrade l'ADN (Galas et Schmitz, 1978). Le principe de la technique repose sur la protection d'une séquence d'ADN contre l'endonucléase I lors de la fixation d'un facteur trans. Celui-ci, par encombrement stérique, empêche la nucléase de

couper l'ADN au niveau de sa séquence de fixation. L'empreinte DNase révèle l'emplacement exact où une protéine interagit avec l'ADN (par exemple, le TFBS). Les fragments d'ADN protégés par les protéines peuvent être davantage isolés et amplifiés afin de détecter la séquence exacte (Galas et Schmitz, 1978) Figure 1.27. En plus d'identifier les TFBS, la méthode de l'empreinte à la DNase peut-être utilisée pour détecter la quantité minimale de protéines (en augmentant sa concentration) nécessaire pour observer le schéma d'empreinte DNase, c'est-à-dire le minimum requis. La concentration d'une protéine donnée doit être liée à l'ADN. La limite de cette méthode est que seuls quelques sites peuvent être évalués en une seule fois.

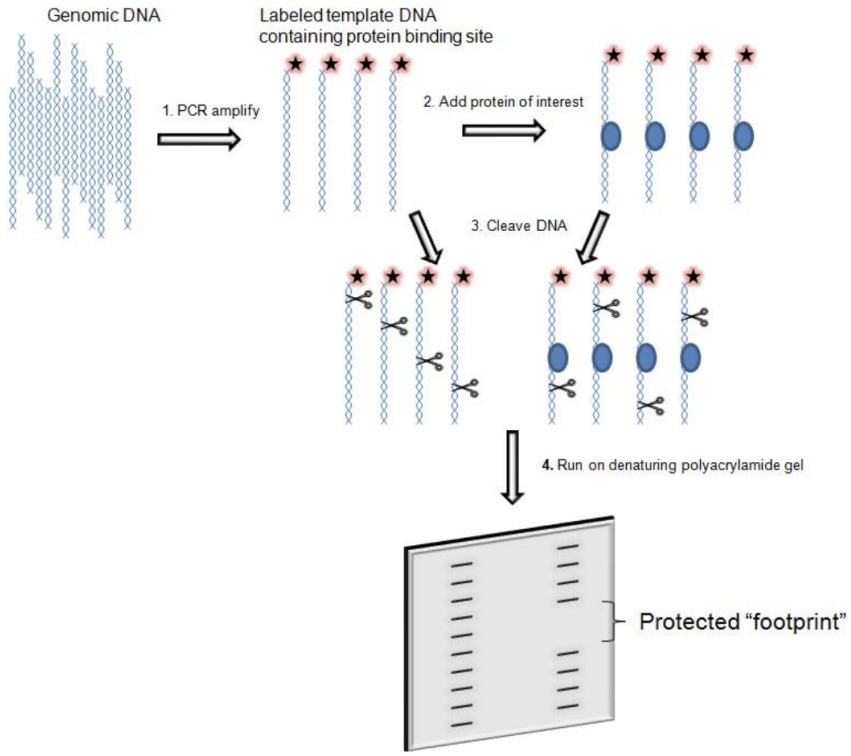


Figure 1.27. – La méthode footprinting de l'ADN d'identification des facteurs de transcription (TFs) se fixant sur une séquence de l'ADN.

SELEX L'évolution systématique de ligands par enrichissement exponentiel (*Systematic Evolution of Ligands by EXponential enrichment*, SELEX) est une méthode qui permet d'isoler des séquences nucléotidiques spécifiques qui interagissent avec une protéine de liaison de l'ADN de choix. Le principe consiste à construire une banque de séquences par synthèse aléatoire d'oligonucléotides ou par fractionnement d'un génome, puis à sélectionner les séquences qui interagissent avec la protéine par EMSA ou chromatographie d'affinité et à les amplifier par PCR.

L'application d'un séquençage massivement parallèle à chaque cycle de SELEX permet de déterminer les affinités relatives à toute séquence d'ADN pour tout TF ou complexe de TFs.

Il faut noter que la méthode SELEX identifie des séquences de liaison de TF mais pas de sites à proprement parler, puisqu'elle se base sur des librairies aléatoires d'oligonucléotides, qui n'ont pas forcément de localisation génomique. Les motifs

issus des collections SELEX sont généralement sur-sélectionnés et sont donc trop spécifiques pour refléter la spécificité de liaison *in vivo* du facteur.

Les séquences obtenues par la méthode SELEX peuvent être alignées pour former des motifs (typiquement, matrices position-poids), qui peuvent ensuite être utilisés pour prédire les sites de liaison d'ADN potentiels dans les génomes et ainsi détecter les gènes-cibles potentiels des TFs (Figure 1.28).

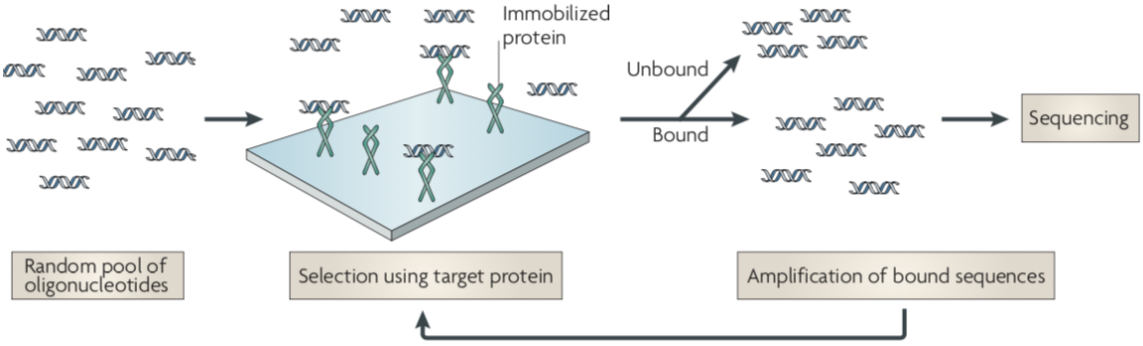


FIGURE 1.28. – Vue d'ensemble de la méthode SELEX. Les molécules d'ADN aléatoires sont exposées au TF ; certaines séquences sont liées, alors que d'autres passent. La fraction liée est séquencée pour déterminer la probabilité d'être lié. Des cycles d'amplification de séquences liées peuvent être utilisés (adapté de Gary D. Stormo and Yue Zhao, 2010).

Protein binding Microarrays (PBM) Les puces de liaison aux protéines (*Protein binding Microarrays*, PBM) représentent la première méthode à haut débit capable de détecter des événements de liaison de TF *in vitro* à l'échelle du génome et peuvent être utilisés pour évaluer la spécificité de séquences de sites de liaison à l'ADN *in vitro* de TF, d'autres protéines de liaison à l'ADN ou de composés synthétiques (BERGER et BULYK, 2006, BERGER et BULYK, 2009) (Figure 1.29).

Dans les expériences PBM, une protéine d'intérêt se liant à l'ADN est exprimée avec une étiquette d'épitope ; cette étiquette a un double objectif : (1) elle permet la purification de la protéine de liaison à l'ADN exprimée et (2) la protéine de liaison à l'ADN marquée par un épitope est ensuite appliquée à une puce à ADN double brin. La micro puce liée aux protéines est lavée doucement pour éliminer toute protéine non spécifiquement liée, puis colorée avec un anticorps primaire spécifique du marqueur épitopique.

La méthode PBM présente l'avantage de détecter et de mesurer la force des événements de liaison du TF dans n'importe quel génome, indépendamment de l'annotation du génome ou si le génome d'intérêt a déjà été séquencé. Toute l'analyse peut être réalisée dans un court laps de temps (environ deux jours). Les PBM sont limités par le nombre de séquences pouvant être représentées sur un microréseau.

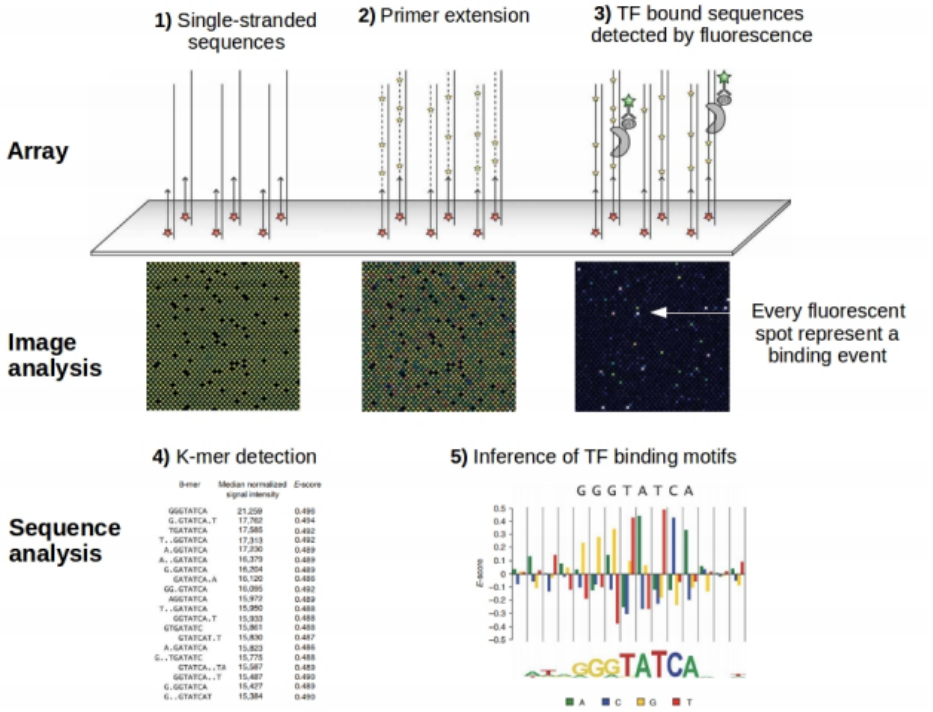


FIGURE 1.29. – Analyse de liaison de TF avec des matrices de liaison. Les séquences simple brin contenant les k-mères sont double brin par extension d'amorce. Les TFs marqués d'un épitope sont ajoutés au puces de liaison aux protéines (PBM). Ensuite, un anticorps marqué avec un fluorophore est ajouté, produisant un signal fluorescent à l'endroit où le TF est lié à l'ADN. Ce signal fluorescent est proportionnel à la force de liaison. Les k-mères de fluorescence la plus élevée sont classés et peuvent être assemblés en tant que motif de liaison au TF (figure adaptée de Berger et al., 2009).

1.4.1.4. Caractérisation génomique des régions de liaison d'un facteur transcriptionnel

Chromatin Immunoprecipitation sequencing (ChIP-seq) Le séquençage de l'ADN de nouvelle génération (*next generation sequencing*, NGS) a été appliqué à l'étude des interactions protéine-ADN. L'immunoprécipitation de la chromatine suivie d'un séquençage à haut débit (ChIP-seq) (FUREY, 2012) est une méthode couramment utilisée pour déterminer les régions de liaison génomique des protéines à l'ADN (y compris les TFs et les histones).

L'approche ChIP capture les protéines en les réticulant de manière covalente en ADN sur leurs sites d'interaction en utilisant des réactifs tels que le formaldéhyde appliqué à des cellules vivantes. L'ADN peut ensuite être extrait des cellules et cisaillé en petits fragments (200-500 pb). Un anticorps spécifique d'une protéine d'intérêt est ensuite utilisé pour immunoprécipiter des complexes protéine-ADN et les fragments d'ADN sont libérés en inversant les liaisons croisées (*Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition*). Les séquences liées sont identifiées par séquençage à haut débit (Figure 1.30). Une

expérience TF ChIP-seq donne généralement plusieurs régions de liaison prévus (souvent appelés « pics ChIP-seq »).

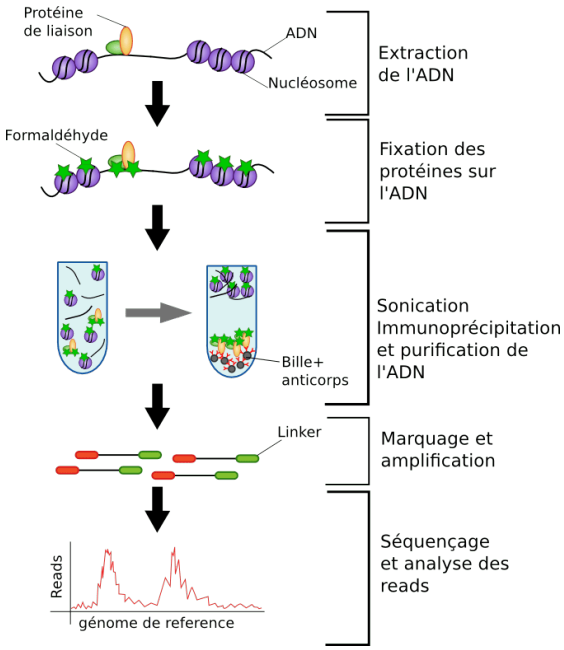


FIGURE 1.30. – Principe de Chip-Seq (source : https ://bioinfo-fr.net/dnase-seq-faire-seq-chip-seq-3-outils-danalyse-de-la-regulation-de-lexpression-des-genes).

Les « pics » du signal ChIP-seq sont généralement déduits par des algorithmes *peak calling* (méthode de calcul utilisée pour identifier les zones d'un génome enrichies en lectures alignées), puis analysés avec un logiciel (par exemple : MEME-ChIP (MACHANICK et BAILEY, 2011), ChIPMunk (KULAKOVSKIY, BOEVA, FAVOROV et al., 2010) ou CISMAPPER (O'CONNOR, BODÉN et BAILEY, 2016)) pour rechercher des motifs enrichis dans les régions liées (Figure 1.31). Des centaines de jeux de données ChIP-seq ont été générés, notamment par le consortium ENCODE (ENCODE PROJECT CONSORTIUM, 2012a), fournissant des données sur les régions de liaison de TFs spécifiques à un type cellulaire ou tissu. Cependant, ChIP-seq présente plusieurs défis clés pour déterminer les motifs des sites de liaison de TF (FUREY, 2012).

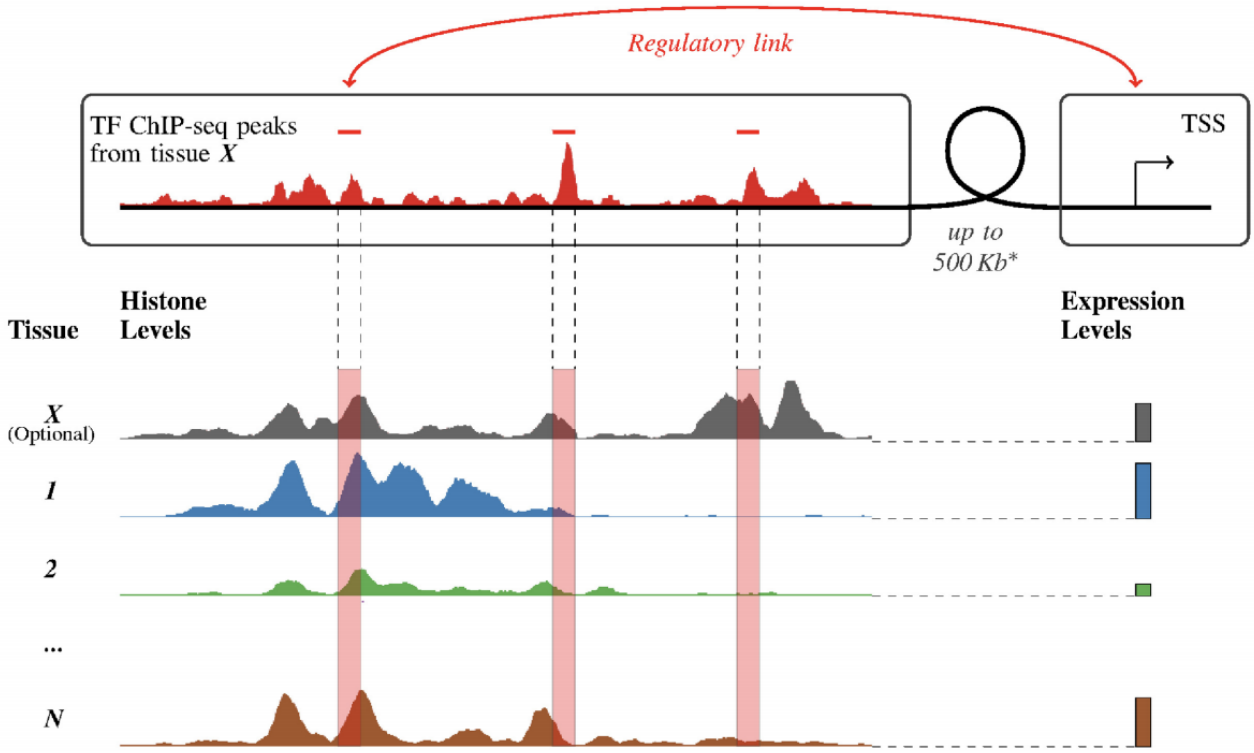


FIGURE 1.31. – Schéma de la méthode CISMAPPER. CISMAPPER prédit les liens de régulation dans le tissu X entre les pics TF ChIP-seq (rouge) et les TSS des gènes en mesurant la corrélation des niveaux d'histone (représentés par des traces colorées) qui chevauchent les pics (surlignés en rouge) avec les niveaux d'expression sur un ensemble (panneau) de tissus. Le tissu X n'a pas besoin d'être présent dans le panneau. La limite de distance est configurable par l'utilisateur avec 500 Kb choisis pour ce travail.

Les pics de ChIP-seq peuvent s'étendre sur des dizaines à des centaines de bases, alors que les motifs des sites de liaison pour la plupart des TF sont inférieurs à 10 pb ; avec l'hétérogénéité de la taille des fragments et l'enrichissement occasionnellement faible en ChIP, ces facteurs peuvent rendre difficile la cartographie précise des sites de liaison, en particulier lorsque les sites de liaison sont regroupés à proximité immédiate (RHEE et PUGH, 2011). Les avancées récentes du protocole ChIP-seq (par exemple, ChIP-exo, ChIP-nexus) résolvent ce problème en éliminant les séquences en excès avec des exonucléases, permettant ainsi une cartographie des sites de liaison à résolution quasi nucléotidique (RHEE et PUGH, 2011, HE, JOHNSTON et ZEITLINGER, 2015). Cependant, ces approches sont encore insuffisantes pour surmonter les difficultés liées à la découverte de motifs de novo robustes pour de nombreux TFs à partir de données ChIP-seq. Par exemple, la spécificité de condition biologique des signaux ChIP-seq ne capture pas tous les sites de liaison possibles et la détection d'évènements de liaison indirecte ou coopérative peut masquer les préférences de liaison directe du TF profilé (FUREY, 2012).

Chromatin ImmunoPrecipitation on Chip (ChIP-chip) La méthode de ChIP-chip est une combinaison des techniques de l'immunoprécipitation de la chromatine (ChIP) et des puces à ADN (Buck et Lieb, 2004). Le ChIP est utilisé pour récuperer des fragments de génome lié à une protéine marqué par un fluorophore et les puces à ADN sont utilisés pour la détection des régions enrichies (fragments immunoprécipités qui interagissent avec une proteine d'interêt) parmi plusieurs milliers de fragments de génomes contenus dans la puce (qui sont généralement les régions inter géniques d'un organisme infecté). Les régions enrichies sont dénaturées en de simples brins d'ADN qui seront en suite hybridés à un microréseau d'ADN simples brins (voir Figure 1.32**A**) contenant un ensemble sélectionné de séquences (par exemple, les promoteurs et les régions inter géniques). L'identification et la caractérisation des fragments par la puce à ADN n'est pas indispensable. On peut utiliser la PCR classique pour la mise en évidence ou l'absence des fragments enrichis (voir Figure 1.32**B**). L'avantage des puces à ADN sur la PCR classique (qui teste un fragment à la fois) est qu'elles accélèrent le processus d'identification des fragments enrichis à l'échelle de milliers de fragments à la fois. Contrairement aux autres techniques classiques d'étude des interactions protéines-ADN, l'ADN utilisé dans la méthode ChIP-chip est directement récupéré *in vivo*.

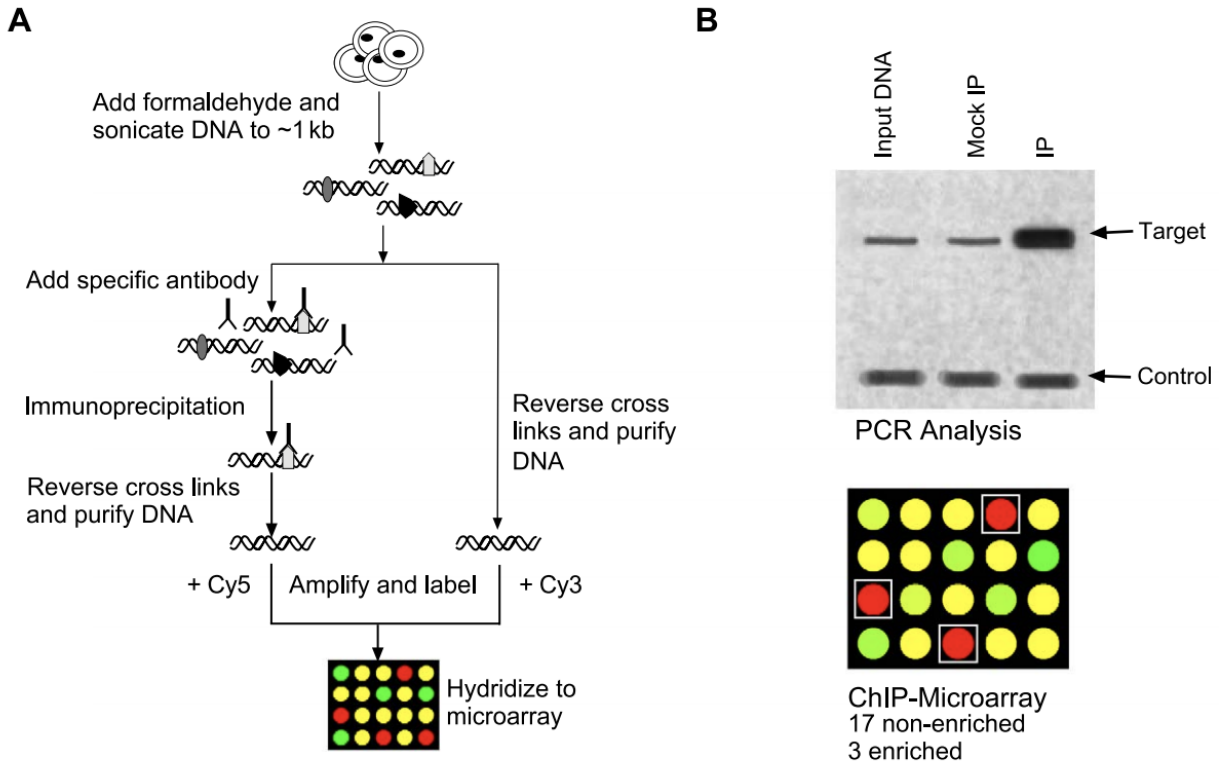


FIGURE 1.32. – Principe de ChIP-chip. L'immuno-précipitation de la chromatine (ChIP) et la détection des séquences nucléotidiques par hybridation sur puce à ADN. (A) dénaturation en simple brin des régions génomiques liées par des protéines d'intérêt et hybridation des régions dénaturées au microréseau d'ADN simple brin. (B) identification des fragment de séquences d'ADN liant les protéines d'intérêt.

1.4.2. Méthodes bioinformatiques pour l'analyse des interactions entre facteur de transcription et ADN

Deux question fondamentales se posent sur la base des découvertes de la liaison de TFs sur l'ADN à savoir : Comment localiser tous les sites de liaison d'un TF donné ? Comment modéliser les affinités de liaison d'un même TF ? Ces questions ont ouvert une voie au développement des méthodes de calcul permettant la découverte, la détection et la représentation des sites de fixation des TFs (Transcription Factor Binding Site, TFBS). Les sites liés par un facteur de transcription donné présentent des similitudes dans leurs séquences nucléotidiques, certaines positions montrant de fortes contraintes sur les résidus (parfois appelées positions « conservées », bien que les différents sites d'un même génome ne soient généralement pas homologues), alors que d'autres positions peuvent être modifiées. sans affecter la liaison de TF (parfois appelée « dégénérée », avec la même remarque sémantique) (*Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition*). Il est donc nécessaire de créer un modèle de séquences qui peuvent lier un même TF. Ces modèles devraient être robuste pour détecter les sites de liaison déjà connus et inconnus. La représentation de séquences reconnues par un même TF est appelée motif de

liaison des TFs (Transcription Factor Binding Motifs, TFBMs) ou simplement appelé **motif** (« Development of bioinformatics methods for the analysis of large collections of transcription factor binding motifs : positional motif enrichment and motif clustering »).

1.4.2.1. Construction d'un motif à partir d'une collection de sites/régions

Chaque TF reconnaît généralement une collection de séquences d'ADN similaires (qui partagent des caractéristiques communes), qui peuvent être représentées sous la forme de motifs de liaison à l'aide de modèles tels que les consensus stricts, consensus "dégénérés"' (code IUPAC), expressions régulières, matrices position-poids (PSSM) (Bernard, Thorsson, Rovira et al., 2012) ou modèles de Markov cachés (HHM). La collection de ces séquences similaires peut se faire par exemple par récupération des séquences en amont des gènes connus pour être co-régulés (185) et de la transformation logarithmique du nombre d'observations de chaque nucléotide à chaque position (Stormo, 2000, Wasserman et Sandelin, 2004). Les PSSM décrivent la probabilité d'apparition d'un nucléotide donné à chaque position du site de liaison à l'ADN (Figure 1.33**a**).

Motifs de liaison de facteurs transcriptionnels "Un motif est une position sur une molécule d'ADN où un TF de lie spécifiquement. C'est une séquence du segment d'ADN lié. Un motif est tout à fait différent d'un site de liaison. Le terme"site" désigne la séquence particulière (génomique ou artificielle) où un facteur se lie, alors que le terme "motif" est utilisé pour désigner une la description générique de la spécificité de liaison, obtenu en synthétisant les informations fournies par un ensemble de sites" (*Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition*).

Modes de représentation Les TFBS identifiés par les méthodes expérimentales telles que EMSA et DNAse footprint peuvent être collectés et alignés pour générer un modèle de TFBMs. On représente généralement la spécificité de liaison d'un TF par un ensemble de positions conservées et variable d'une collection de sites de liaison. Les TFBMs peuvent être représenté selon le modèles de consensus, de matrice de position-poids et de Markov.

La séquence consensus Une séquence consensus est une chaîne de nucléotides qui représente les résidus conservés dans chaque colonne d'un alignement multiple. La séquence consensus est décidée par une procédure de sélection conçue pour déterminer quel résidu ou quel nucléotide est placé à chaque position individuelle, généralement celui qui se trouve le plus souvent à la position spécifique (*Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition*). Par exemple la Figure 1.33**a** (Wasserman et Sandelin, 2004) montre une collection alignée de séquences de liaison validées expérimentalement du facteur MEF2 (Wasserman et Sandelin, 2004). La construction du modèle de séquence consensus est caractérisé par une sélection d'un symbole nucléotidique de dégénérescence pour chaque position (colonne) dans l'alignement (voir Figure 1.33**b**). Des sites de liaison

inhabituels peuvent avoir un effet extrême sur le consensus (voir, par exemple, le site 8 de la Figure 1.33**a**). Pour les motifs définis sur des séquences nucléiques, le consensus dégénéré est basé sur le code IUPAC (Figure 1.33**g**). Par exemple dans le contexte des séquences régulatrices, un consensus est généralement utilisé pour synthétiser les résidus conservés d'un motif de liaison au TF, construit en alignant une collection de sites de liaison. Pour les motifs définis sur des séquences nucléiques, le consensus dégénéré est basé sur le code IUPAC (Figure 1.33**g**) pour les nucléotides ambigus (*Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition*).

Position Weight Matrix (PWM) Une matrice poids-position (PWM), également appelée matrice de poids spécifique à la position (position-specific weight matrix, PSWM) ou matrice de notation spécifique à la position (position-specific scoring matrix, PSSM), est une représentation couramment utilisée pour représenter les motifs de liaison par exemple d'un TF sur la séquence d'ADN. Pour refléter plus précisément les caractéristiques à chaque position, une matrice contenant le nombre de nucléotides observés à chaque position est créée (voir Figure 1.33**c**) (Wasserman et Sandelin, 2004). Par exemple, la deuxième colonne de l'alignement (voir Figure 1.33**a**) ne contient pas de **C**, a 4 **A**, 3 **G** et 1 **T**, donnant ainsi la deuxième colonne correspondante de la matrice {0,4,3,1}. La matrice de fréquence est généralement convertie en une PWM en utilisant les valeurs de fréquence normalisées en une échelle logarithmique (voir Figure 1.33**d** et **(Équation 1)** et **(Équation 2)**). En utilisant un modèle matriciel, un score quantitatif pour toute séquence d'ADN peut être généré en faisant la somme des valeurs correspondant au nucléotide observé à chaque position (voir Figure 1.33**e**).

Formules mathématiques d'analyse des séquences régulatrices La description de ces formules mathématiques a été faite par Wyeth W. Wasserman et Albin Sandelin (Wasserman et Sandelin, 2004).

(Équation 1) : probabilité de l'observation d'un nucléotide, avec une estimation corrigée par un pseudo-poids

$$p_{b,i} = \frac{\check{\mathrm{N}}_{b,i} + s_b}{N + \sum_{b' \in \{A,C,G,T\}} s_{b'}}$$

où $\check{\mathrm{N}}_{b,i}$ = comptage des occurrences du résidu b en position i ; N = nombre de sites de liaison ; $p_{b,i}$ = probabilité corrigée du résidu b en position i ; s_b = fonction de "pseudo-comptage" pour le résidu b.

Le pseudo-poids (plus précisément pseudo-comptage) vise à éviter d'associer une probabilité nulle à un résidu qu'on n'aurait pas observé à une position donnée dans l'ensemble des sites connus. En effet, les matrices position-poids étaient initialement constituées à partir d'un nombre très restreint de sites (quelques dizaines, voire moins de 10), et l'absence d'un résidu à une position donnée ne signifie pas pour autant que ce résidu ne pourrait pas s'y retrouver dans d'autres sites de liaison du facteur transcriptionnel. Le pseudo-poids est particulièrement important quand on ne dispose que d'un petit nombre de sites connus pour construire la matrice. Quand le nombre de sites connus est suffisant, l'impact du pseudo-poids devient négligeable

et le modèle de spécificité repose donc essentiellement sur les sites connus.

(Équation 1) : construction de la PWM en divisant les probabilités nucléotidiques $p_{b,i}$ par probabilités de modèle de fond attendues s_b et conversion des valeurs en une échelle logarithmique.

$$W_{b,i} = log2\frac{p_{b,i}}{p_b}$$

p_b = modèle de fond de probabilité de la base b ; $p_{b,i}$ = probabilité corrigée de la base b en position i ; $W_{b,i}$ = Valeur de la PWM de la base b en position i.

Équation 3 : Les valeurs de probabilité (**Équation 1**) peuvent être utilisées pour déterminer le contenu total de l'information (en bits) dans chaque position.

$$Di = 2 + \sum_d p_{b,i} log2 p_{b,i}$$

Di = contenu informationnel en position i ; $p_{b,i}$ = probabilité de la base b en position i.

Logo Le contenu informationnel de chaque alignement de séquences de liaison détermine la spécificité dans chaque colonne de de l'alignement (voir Figure 1.33**a**).Un logo de séquence (Schneider et Stephens, 1990) met à l'échelle chaque nucléotide en fonction du nombre total de bits d'information multiplié par l'occurrence relative du nucléotide à la position (voir Figure 1.33**f** et **Équation 3**) (Wasserman et Sandelin, 2004). Les logos de séquence fournissent une représentation visuelle rapide et intuitive des caractéristiques du motif.

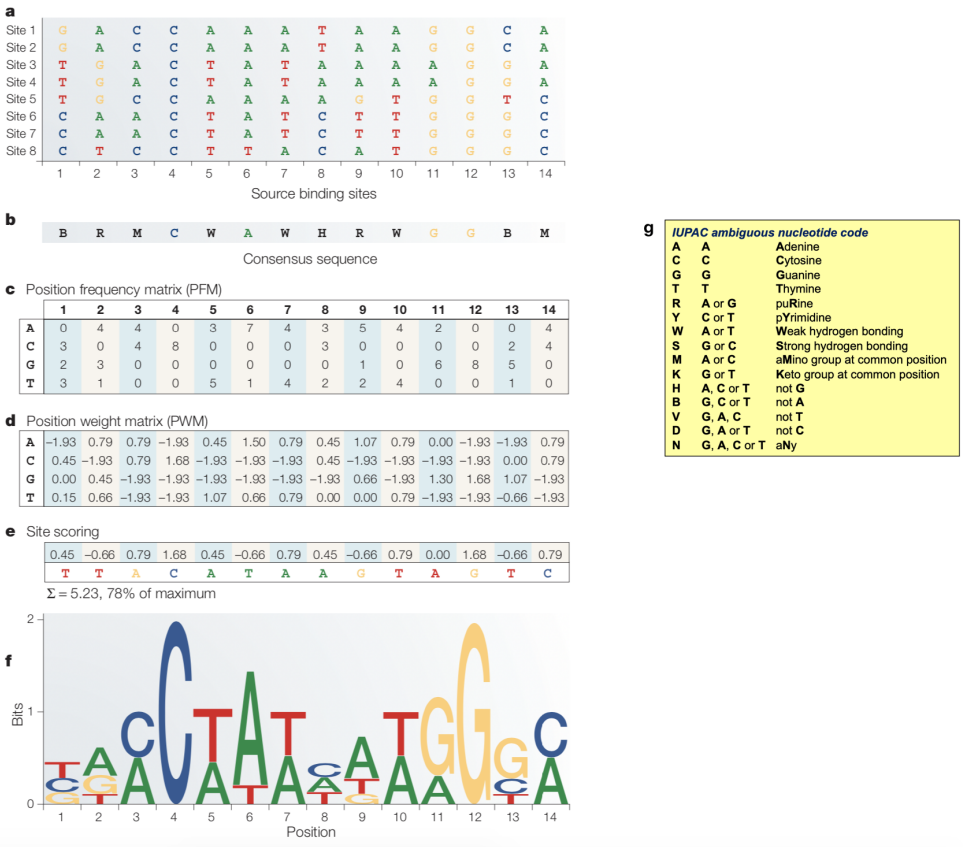


FIGURE 1.33. – Modélisation par matrice des sites liés par le facteur MEF2 (adapté de Wyeth W. Wasserman et Albin Sandelin, 2004). (a) alignement de sites de liaison ; (b) séquence consensus dégénéré basé sur le code IUPAC ; (c) matrice d'occurrences des nucléotides par position ; (d) matrice position-poids ; (e) calcul du score d'un site avec la matrice ; (f) sequence logo dela matrice ; (g) code IUPAC pour les nucléotides partiellement spécifiés.

Modèles de Markov caché Le modèle matriciel se base sur l'hypothèse d'indépendance des colonnes du modèle de Bernoulli. Cependant, l'interdépendance des positions dans les sites de liaison des TFs ne se limite pas qu'aux positions adjacentes, mais peut survenir aussi entre les positions séparées d'un nucléotide de plus. Une PWM ne peut pas représenter l'information sur les insertions et délétions, car un TF qui liant un site avec espacement ne correspond pas au modèle de PWM. On peut alors utiliser les modèles de Markov caché (Hidden Markov Models, HMM) pour compenser les limites du modèle matriciel.

Par définition, un HMM consiste en une série d'états, où chaque état peut décrire l'état suivant basé sur une probabilité d'émission. Les états sont connectés entre eux sous forme d'une chaîne ayant un état initial et un état final. La probabilité de passage d'un état à un autre est appelée probabilité de transition. Plusieurs chemins (donc séquences) peuvent être créés en suivant des passages différents dans la chaîne avec différentes probabilités.

Une chaîne de Markov est une automate stochastique qui présente une grande souplesse pour l'extension de la structure d'un motif afin de pouvoir effectuer des

insertions ou délétions sur le motif élargi. Ceci conduit à la création d'une nouvelle matrice. Mais cette création alourdit les calculs et on utilise un modèle simplifié d'automatisation du HMM qui permet de représenter les dépendances entre les différentes position du site de liaison.

Identification des TFBS altérés par les variants génétiques L'application la plus récente des PSSM concerne la découverte de variants régulateurs (c'est-à-dire des mutations au sein d'une région régulatrice en tant qu'activeur, promoteurs ou TFBS) susceptibles d'affecter la régulation en aval. Bien que certaines de ces variants affectent d'autres aspects de la régulation des gènes tels que le bouclage, d'autres affectent directement la liaison du TF et peuvent être détectés à l'aide de PSSM (voir Figure 1.34). La logique derrière cette approche est que le variant pourrait être reflété dans le score de poids produit par un PSSM. Une différence de poids entre deux allèles pourrait indiquer une véritable variant régulatrice (voir Figure 1.34). Une façon de détecter le TFBS altéré consiste à calculer la distribution complète des scores de poids (ou valeurs p) pour un seul PSSM et à calculer la différence de scores de poids entre les allèles A (requête) et B (contrôle). Cette méthode est utilisée par les outils.is-rSNP (Macintyre, Bailey, Haviv et al., 2010), Regsnps (Teng, Ichikawa, Padgett et al., 2012), spécialisés dans les polymorphismes mononucléotidiques régulateurs (rSNPs). D'autres outils tels que l'analyse par variation de *RSAT* (Medina-Rivera, Defrance, Sand et al., 2015) et le logiciel sTRAP (Thomas-Chollier, Hufton, Heinig et al., 2011) peuvent détecter des SNPs et des insertions ou des délétions au niveau des TFBS. Bien que l'utilisation de PSSM soit la méthode la plus utilisée, d'autres méthodes ont toutefois été proposées, par exemple en combinant la prévision du TFBS et la conservation de la séquence (Andersen, Engström, Lithwick et al., 2008), la prévision du TFBS (Shi, Fornes, Mathelier et al., 2016).

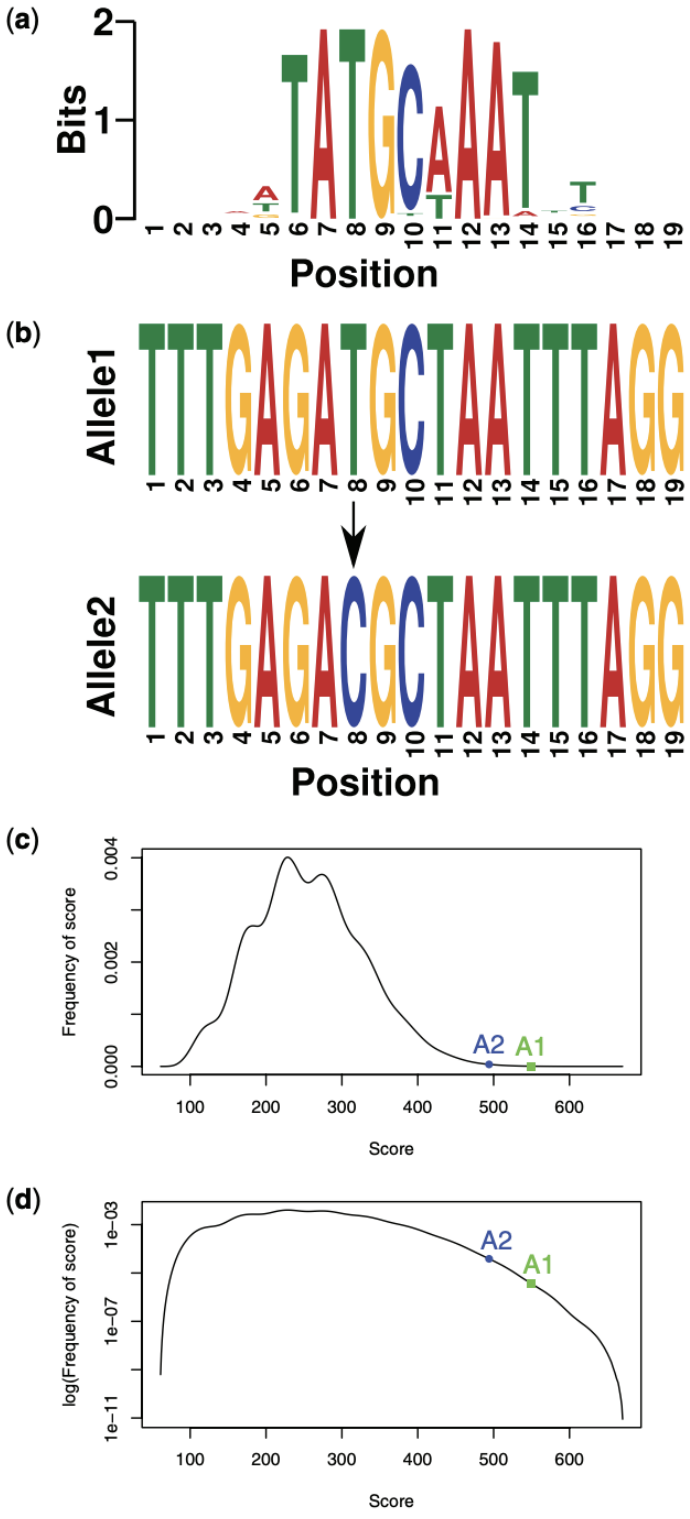


FIGURE 1.34. – Détection des variants régulateurs à l'aide de matrice de notation spécifique à la position (PSSM). (a) Logo pour le facteur OCT1. (b) Les deux allèles analysés, voir position 8. (c) Répartition de tous les scores de poids avec le PSSM OCT1 montre la différence de score entre les allèles. (adapté de Macintyre et al., 2010).

1.4.3. Ressources bioinformatiques pour l'analyse de la régulation et des variants régulateurs

1.4.3.1. *ReMap* : une base de données de pics de liaison de facteurs transcriptionnels

ReMap est une base de données des sites de fixation des facteurs de transcription chez l'homme qui constitue une analyse intégrative des régulateurs de transcription des expériences ChIP-seq issus de jeux de données Public et ENCODE (ENCODE Project Consortium, 2012b). Le catalogue *ReMap* fournit des renseignements sur les facteurs de transcription et les jeux de données analysés. On retrouve des informations sur la classification des facteurs de transcription, une analyse de la conservation des nucléotides au niveau des sites de fixation identifiés et la liste et le niveau de qualité des jeux de données ChiP-seq impliquant chacun des facteurs de transcription.

Le catalogue *ReMap* a été élaboré à partir des données publiques GEO (https://www.ncbi.nlm.nih.gov/geo/) et ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) et des données ENCODE (https://www.encodeproject.org/) (ENCODE Project Consortium, 2012b).

Actuellement, le catalogue *ReMap* comprend 80 millions de pics provenant de 485 TFs, co-activateurs de transcription (TCA) et facteurs de remodelage de la chromatine (CRF) des données Publiques et ENCODE (Chèneby, Gheorghe, Artufel et al., 2018). On dénombre 2829 expériences ChIP-seq de qualité contrôlée effectuées par ENCODE et des sources publiques (GEO, ArrayExpress). Les jeux de données publics ChIP-seq (n = 1 763) ainsi que les données ENCODE ChIP-seq (n = 1 066) ont été mappés vers l'assemblage humain GRCh38 / hg38.

Les données ENCODE et Public ont été analysées pour proposer une intégration unifiée des deux sources de données, produisant un atlas unique de régions régulatrices pour 485 TR. Il existe dans le catalogue 125 TR communes aux deux ensembles, 154 protéines spécifiques à ENCODE et 206 spécifiques au catalogue Public. Pris séparément, les pics ENCODE chevauchent 96% des régions publiques et 87% des pics publics chevauchent les régions ENCODE.

Une identification de 49 millions de pics liés par des facteurs de transcription (80 millions avec les données ENCODE incluses) à partir des données de Pick calling le catalogue. Ces nombres incluent les sites qui se chevauchent pour des TR identiques qui ont été étudiés dans diverses conditions. Pour résoudre ce problème, nous avons fusionné des sites de liaison de TF chevauchant pour des TF similaires, obtenant ainsi un catalogue de 23,7 millions de sites de liaison non redondants (35,5 millions avec des données ENCODE) (Figure 1.35).

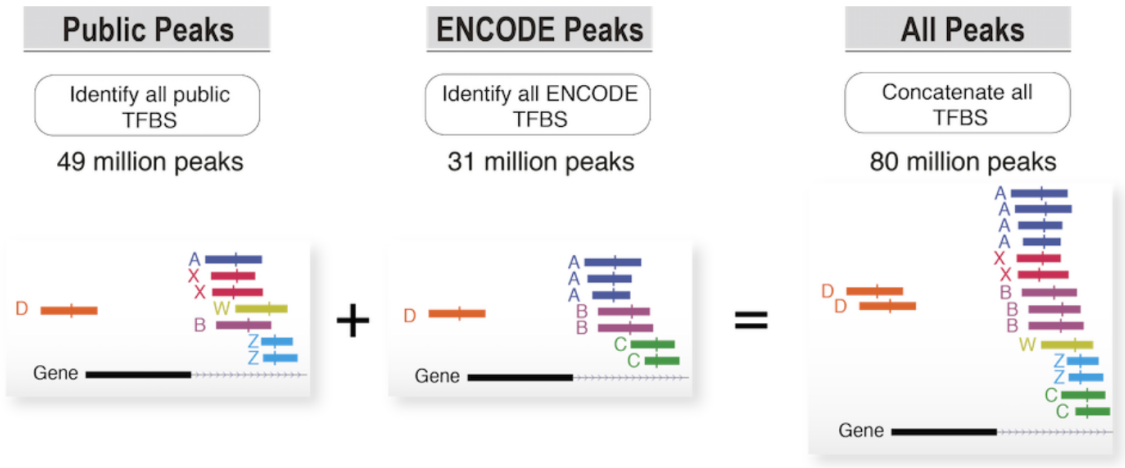


FIGURE 1.35. – Différentes sources de données du catalogue *ReMap* (http ://tagc.univ-mrs.fr/remap/index.php).

Pour faciliter l'accessibilité et l'exploration du catalogue par la communauté scientifique, le site internet *ReMap* a été développé (catalog *ReMap* : http://tagc.univ-mrs.fr/remap/). *ReMap* dispose des informations sur la classification des facteurs de transcription, une analyse de la conservation des nucléotides au niveau des sites fixation identifiés, un lien vers l'outil de découverte de motif *RSAT* (peak-motif) ainsi que la liste et le niveau de qualité des jeux de données ChiP-seq impliquant chacun des facteurs de transcription.

Pour chaque TF, type de cellule et jeu de données, *ReMap* fournit des fichiers pour télécharger des pics de ChiP-seq au format BED et des séquences au format FASTA. L'ensemble du catalogue est ouvertement accessible à la communauté et peut être téléchargé au format BED.

1.4.3.2. Bases de données de motifs de liaison de facteurs transcriptionnels

JASPAR JASPAR est une collection de préférence non redondante de motifs de liaison des TFs à l'ADN modélisés et stockés sous forme de matrices de fréquence de position (position frequency matrices, PFMs) et de modèles flexibles (TFFMs) de TFs pour les espèces multiples dans six groupes taxonomiques des vertébrés, des nématodes, des insectes, des plantes, des champignons et des tuniciers (http://jaspar.genereg.net/). Ces PFMs peuvent être converties en PWM qui sont utilisés pour scanner les séquences génomiques.

JASPAR est structuré en deux catégorie à savoir :

Base de données JASPAR CORE : contient un ensemble de profiles non redondant dérivés de collections publiées de TFBS définis expérimentalement pour des eucaryotes multicellulaires. Les sites de liaison ont été déterminés historiquement soit dans des expériences SELEX, soit par la collecte de données provenant des régions de liaison déterminées expérimentalement de régions de régulation réelles. La version JASPAR 2018. (KHAN, FORNES, STIGLIANI et al., 2018) inclut la collecte des données issues des expériences de ChIP-seq du catalogue *ReMap*

(Khan, Fornes, Stigliani et al., 2018), DAP-seq (O'Malley, Huang, Song et al., 2016), de SMiLE-seq (Isakova, Groux, Imbeault et al., 2017), de PBM (Franco-Zorrilla, López-Vidriero, Carrasco et al., 2014) et de HT-SELEX (Jolma, Yan, Whitington et al., 2013). On compte au total 1404 PFMs non redondant (579 pour les vertébrés, 489 pour les plantes, 176 pour les champignons, 133 pour les insectes, 26 pour les nématodes et 1 pour les urochordata) (Khan, Fornes, Stigliani et al., 2018). Contrairement à TRANSFAC, JASPAR core dispose des matrices non redondant et de qualité en libre accès (*open source*).

— **Collections JASPAR** : ce sont des collections de matrices utiles, mais qui ne rentrent pas dans la portée de JASPAR CORE. Les exemples incluent les formes d'épissage, les modèles de matrices calculés sans TF, les méta-modèles, etc.

Accessibilité **Interface web** : JASPAR est accessible en interface web (http://jaspar.genereg.net/) très interactive

API RESTful : JASPAR dispose d'une API de type web services RESTful (Khan et Mathelier, 2018) qui permet un accès programmatique à JASPAR pour la plupart des langages de programmation (notamment Perl, Python / Biopython, R / Bioconductor et Ruby) et renvoie les données dans sept formats largement utilisés : JSON, JSONP, JASPAR, MEME, PFM, TRANSFAC et YAML. En outre, il fournit une interface navigable et un accès à l'outil JASPAR d'inférence de motifs pour les développeurs d'outils bioinformatiques. L'API RESTful est implémentée dans Python à l'aide de Django REST Framework et est librement accessible à l'adresse http://jaspar.genereg.net/api/. J'ai utilisé l'API RESTful de JASPAR (Khan et Mathelier, 2018) en langage R dans le dévéloppement du workflow en vue de collecter les noms des TFs correspondant à leurs motifs de liaison identifiés par l'outil variation-scan.

1.4.3.3. Bases de données de variants polymorphiques chez l'humain

dbSNP dbSNP est la base de données de polymorphismes nucléotidiques simples des variations génétiques (de différentes espèces) appartenant au Centre national d'information sur la biotechnologie (NCBI) en collaboration avec l'Institut national de recherche sur le génome humain (NHGRI) (Sherry, Ward, Kholodov et al., 2001). Cette base de données (BDD) a été développé en 1999 suite à un besoin de disposer d'un catalogue des variations du génome dans l'objectif de traiter les plans d'échantillonnage des études d'association, la cartographie des gènes et le biologie de l'évolution (Sherry, Ward et Sirotkin, 1999). L'appellation "dbSNP" réside sur la grande majorité des variation de type Single Nucleotide Polymorphism (SNP). La base de données contient aussi des variations des microsatellites et des insertions et délétions à petite échelle, ainsi que des informations sur la publication, la fréquence des populations, les conséquences moléculaires et les informations de cartographie génomique et RefSeq pour les variations courantes et les mutations cliniques. La BDD dbSNP présente un problème de fiabilité. Elle est ouvertes en écriture à l'ensemble de la communauté scientifique. On n'est pas à l'abri que certaines variations ne soient en fait que des erreurs expérimentales ou ne soient pas enregistrées dans la BDD comme il le faudrait (absence du variant de référence,

variant de référence repris parmi les variants alternatifs, coordonnées erronées, etc).

GWAS catalog Les études d'association pangénomique (GWAS) sont devenues un outil indispensable pour élucider les relations génotype – phénotype. GWAS établit une corrélation entre les marqueurs génomiques et les différences phénotypiques dans une population et signale la probabilité de l'association.

GWAS catalog est une BD des études d'association publiées sur le génome de l'Institut national de recherche sur le génome humain (NHGRI) qui fournit une collection publique et organisée des études de GWAS publiées, qui teste au moins 100 000 SNPs.

GWAS catalog[15] est une source accessible au public, classée manuellement, de tous les GWAS et résultats d'études d'association pangénomiques publiés, produits et développés en collaboration par le NHGRI et l'EMBL-EBI. Elle inclut toutes les études GWAS éligibles depuis la première publication de GWAS sur la dégénérescence maculaire liée à l'âge en 2005 (Klein, Zeiss, Chew et al., 2005). La version actuelle de **GWAS catalog** datant du 12 juillet 2019, comporte 4054 publications, 93892 SNPs, 143963 associations, assemblage du génome version GRCh38.p12, dbSNP Build 151 et *Ensembl* Build 96.

Le seuil d'association utilisé pour chaque SNP ($p < 5 \times 10\text{-}8$) (Welter, MacArthur, Morales et al., 2014) correspond au seuil de Bonferroni, une correction qui s'applique pour avoir la certitude qu'une différence de fréquence observée entre une cohorte de témoins et de malades possède une significativité réelle et non liée au hasard.

Les biologistes, bioinformaticiens et chercheurs cliniciens / translationnels utilisent les données du catalogue comme point de départ pour des recherches ultérieures visant à identifier les variants causaux, à comprendre les mécanismes de la maladie et à établir des cibles pour de nouvelles thérapies.

Les données du catalogue GWAS sont également intégrées à de nombreuses ressources bioinformatiques, notamment *Ensembl* (Yates, Akanni, Amode et al., 2016), le navigateur de génomes UCSC (Rosenbloom, Armstrong, Barber et al., 2015), PheGenI (Ramos, Hoffman, Junkins et al., 2014), HuGE Navigator (Yu, Gwinn, Clyne et al., 2008) et GWASdb (Li, Liu, Wang et al., 2016).

Ontologies Chaque trait dans **GWAS catalog** est défini par une *Ontologie Expérimentale des Facteurs (EFO)*, développé par l'équipe *EMBL-EBI Samples, Phenotypes and Ontologies (SPOT)*. EFO fournit une description systématique de nombreuses variables expérimentales disponibles dans les bases de données EBI (`https://www.ebi.ac.uk/efo/`). Il combine des parties de plusieurs ontologies biologiques, telles que l'anatomie, les maladies (par exemple, l'ontologie d'Orphanet des maladies rares) et les composés chimiques. Une cartographie précise des traits signalés avec les termes EFO facilite la recherche, la visualisation et l'intégration des données du catalogue GWAS.

Accessibilité Le catalogue est accessible via une interface Web. Les données sont présentées dans une table téléchargeable. On peut aussi faire appel aux services REST par l'invite des commandes en R, python, perl, etc pour télécharger les

données de **GWAS catalog**. Dans le mode de fonctionnement du workflow **cisreg-GWAS** (voir chapitre 5), j'ai utilisé les services REST en R pour la collecte des SNPs associé à un trait phénotypique.

On recherche les variants associés à une maladie en fournissant soit :

1) le nom de la maladie (example : malaria) ;
2) le nom du premier auteur d'une publication d'une étude GWAS (exemple : Timman) ;
3) l'identifiant du variant qui fournira la maladie associée à ce variant (exemple : rs10900585) ;
4) le numéro d'un locus (exemple : 2q37.1) ;
5) la position génomique (6 :16000000-25000000) ;
6) le nom du gène (exemple : ATP2B4)

***Ensembl* regulation** *Ensembl* (https://www.ensembl.org/info/genome/funcgen/index.html) est un projet conjoint de l'EBI (*European Bioinformatics Institute*) et du WTSI (*Wellcome Trust Institute*). Le but de *Ensembl* est d'annoter automatiquement le génome, d'intégrer cette annotation à d'autres données biologiques disponibles et de rendre toutes ces informations accessibles au public via le Web. La gamme de données disponibles s'est élargie pour inclure la génomique comparative, les données sur la variation et la régulation. La base de données *Ensembl* regroupe les annotations disponibles pour les vertébrés, les plantes, les champignons, les bactéries, les protistes et les archaeas.

Les données *Ensembl* (Flicek, Amode, Barrell et al., 2014) sont accessibles de différentes manières, notamment : le navigateur génomique (genome browser), l'outil d'exploration de données BioMart (Kinsella, Kähäri, Haider et al., 2011), le package Bioconductor R (Gentleman, Carey, Bates et al., 2004), le téléchargement FTP de jeux de données complets, exportations des fonctionnalités ou des séquences directement à partir de pages Web, l'extraire des données de la base de données publique *Ensembl* à l'aide de scripts Perl et par le serveur REST *Ensembl* (Yates, Beal, Keenan et al., 2015) qui permet de récupérer facilement une large gamme de données *Ensembl* dans la plupart des langages de programmation (Perl, Python, Ruby, Java, R, Curl, Wget), en utilisant des formats standard tels que JSON et FASTA, tout en minimisant le travail du client. J'ai travaillé avec les service REST *Ensembl* en R pour la recherche des SNPs en déséquilibre de liaison génétique dans mon workflow.

***RSAT* : une suite logicielle pour l'analyse de séquences régulatrices** *RSAT* (outils d'analyse des séquences de régulation) est une suite modulaire d'outils logiciels dédiés à l'analyse des séquences cis-régulatrices et des motifs de liaison des TFs. Ce projet est dirigé par Jacques van Helden depuis 1998 (Helden, André et Collado-Vides, 1998) et a été constamment mis à jour avec de nouveaux programmes à chaque sortie (Medina-Rivera, Defrance, Sand et al., 2015, Helden, 2000 ;Thomas-Chollier, Defrance, Medina-Rivera et al., 2011). Les programmes de la suite *RSAT* peuvent être utilisés en ligne (http://www.rsat.eu/), en mode autonome via une ligne de commande, à distance

via des services Web SOAP / WSDL et RESTfull, en tant que machine virtuelle VirtualBox ou instanciée sur le cloud de l'Institut Français de Bioinformatique (IFB). La première version comprenait deux programmes de découverte de motifs (Helden, 2000) et trois ans plus tard de nouveaux programmes ont été ajoutés, permettant aux utilisateurs d'effectuer une analyse complète des motifs, y compris la visualisation du résultats et la possibilité de récupérer des séquences (c'est-à-dire des régions régulatrices) pour un grand nombre d'organismes, notamment génomes bactériens (Helden, 2003). La version 2008 comportait 30 programmes, dont l'analyse matricielle, l'un des premiers programmes à balayer des séquences avec PSSM en utilisant les valeurs p comme seuil plutôt que les scores de poids. En outre, cette version comprenait un logiciel permettant de créer des contrôles négatifs et une empreinte phylogénétique. Entre 2008 et la prochaine version en 2011, a commencé l'ère des expériences à haut débit produisant de grandes quantité de données à analyser ; pour cette raison, la version *RSAT* 2011 comprenait plusieurs programmes axés sur la résultats d'expérimentation à haut débit sous forme de motif de pointe constituant un flux de travail permettant d'analyser de grands ensembles de séquences, autres programmes pour télécharger des séquences à partir du serveur UCSC, et un logiciel pour comparer et évaluer la qualité des motifs (Thomas-Chollier, Defrance, Medina-Rivera et al., 2011). La dernière version a été publiée en 2015 et contient 52 outils, parmi lesquels il y a un programme pour détecter et télécharger les variantes réglementaires et un programme pour regrouper les motifs de liaison des TFs. Dans cette version, une partie du travail était axée sur la garantie de la reproductibilité des résultats, par exemple en utilisant des machines virtuelles avec une version particulière de *RSAT*. De plus, depuis le nombre de la croissance des génomes considérablement depuis la dernière version de *RSAT*, les serveurs *RSAT* ont été séparés par des serveurs spécifiques contenant les génomes et d'autres informations relatives à leurs différents organismes (Medina-Rivera, Defrance, Sand et al., 2015).

Les principales applications de *RSAT* sont les suivantes :

— Découverte de motifs, adaptée aux ensembles de données à haut débit tels que ChIP-seq, PBM ou groupes de promoteurs des gènes co-exprimés.
— Analyse du motif de liaison TF (évaluation de la qualité, comparaisons et regroupement).
— Balayage des motifs pour détecter un seul TFBS ou des modules cis-régulateurs.
— Génomique comparative pour un grand nombre d'organismes appartenant à différents taxons.
— Analyse des variations régulateurs
— Création de modèles d'arrière-plan et de contrôles négatifs pour l'analyse des motifs.

RSAT est l'une des suites de programmes les plus complètes pour l'analyse de séquences cis-régulatrices. L'analyse des motifs et inclut également des programmes de génomique comparative et de détection des variantes. En plus de cette polyvalence de programmes, *RSAT* est compatible avec la sortie d'outils externes (grâce aux conversions entre les formats de fichiers). L'accessibilité de *RSAT* s'est imposée comme un choix pour les biologistes avec peu ou pas d'expérience en programmation, de plus, la modularité des programmes permet de créer des flux de travail complexes pour l'analyse des motifs, y compris des outils externes.

On distingue actuellement 5 serveurs publics prenant en charge plus de 500 génomes d'eucaryotes et 9 000 génomes de procaryotes. Axé sur la convivialité et l'accessibilité pour les utilisateurs avec et sans formation en bioinformatique, *RSAT* fournit des outils permettant de récupérer des séquences, d'analyser les motifs, d'évaluer la qualité des motifs de TF, de comparer et de regrouper les motifs, de convertir les formats de fichiers, etc. Dans le développement de mon worflow j'ai utilisé un sous-ensemble d'outils inclus dans *RSAT* (*variation-info*, *retrieve-variation-seq* et *variation-scan*) qui aboutissent à la prédiction des variants régulateurs et à l'évaluation de leur impact potentiel sur les sites de liaison de la TF.

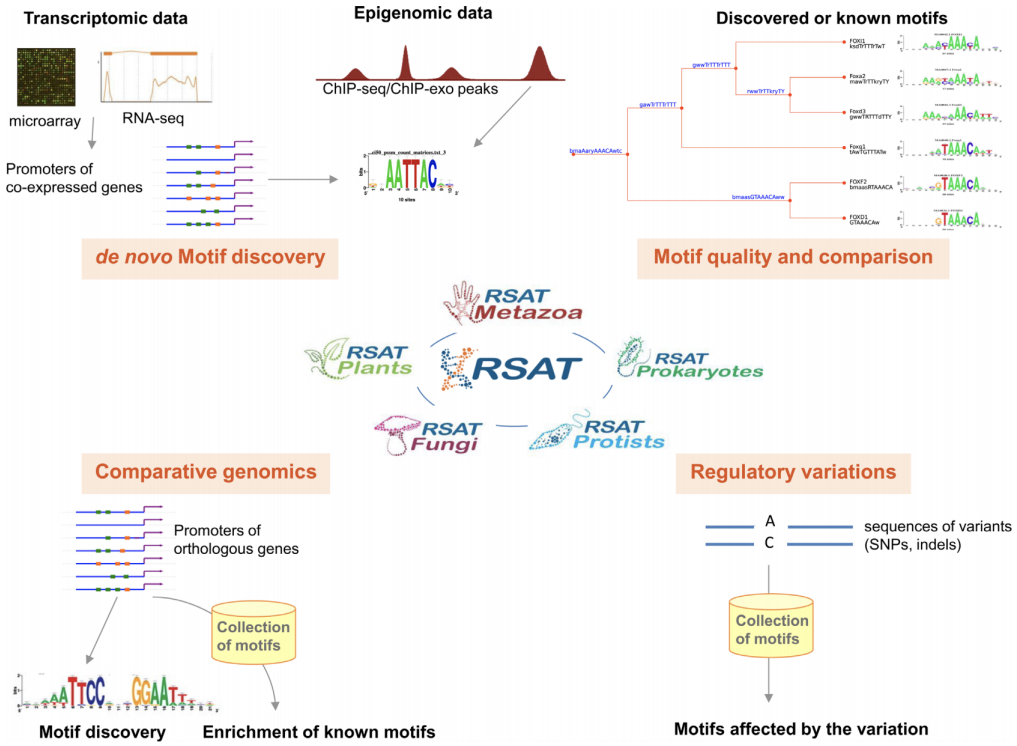


FIGURE 1.36. – Vue d'ensemble des principales applications et différents serveurs et fonctionalités de la suite *RSAT* (Source : MEDINA-RIVERA, DEFRANCE, SAND et al., 2015),

RSAT variation-tools *variation-tools* est un sous-ensemble de *RSAT* composé de quatre programmes (*variation-info, convert-variations, retrieve-variation-seq, variation-scan*) qui prédit l'impact des variants régulateurs sur la liaison aux TFs (Figure 1.37).

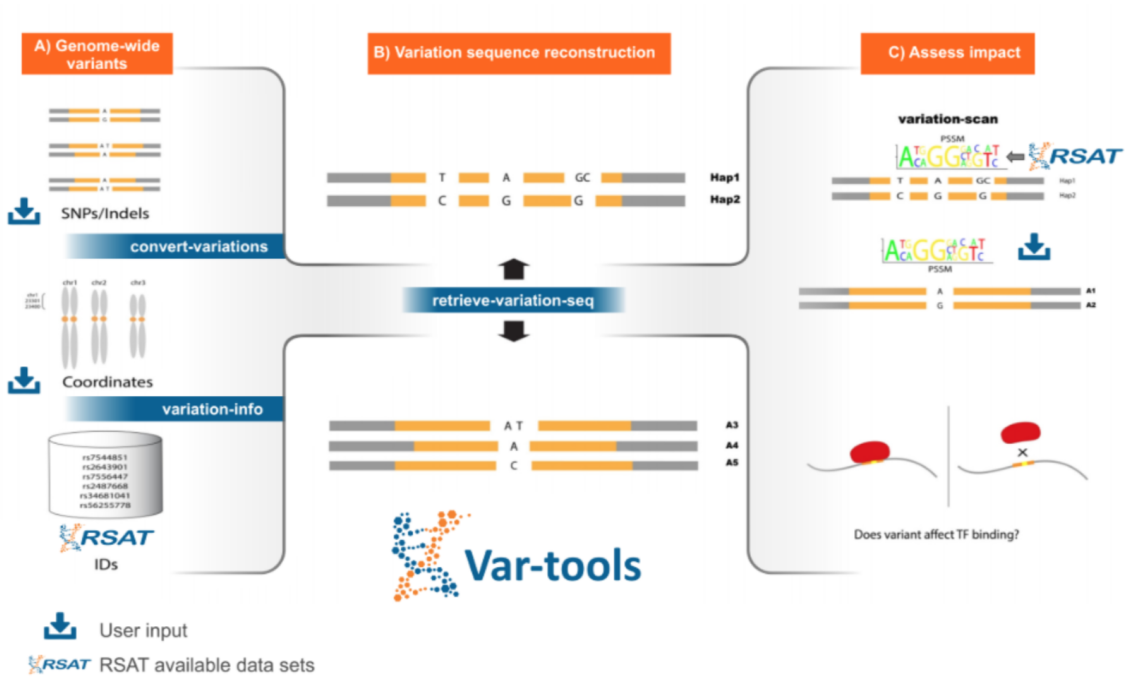


FIGURE 1.37. – Représentation schématique de variation-tools. A. convert-variations permet la conversion d'un fichier de l'utilisateur aux format gff et varBed. variante-info fait une annotation des SNPs à partir des données Ensembl installés sur des serveurs RSAT. B. retrieve-variation-seq récupère la séquence autour du SNPs (y compris les haplotypes possibles) et génère un fichier texte avec une ligne par allèle de chaque SNP ou haplotype au format varSeq. C. L'utilisateur a deux possibilités : il peut soumettre le fichier varSeq soit avec une collection de motifs qu'il peut saisir ou importer dans variation-scan, soit sélectionnées parmi les collections disponibles dans RSAT pour une analyse de variation ; l'outil analysera tous les motifs et comparera par paires de chaque allèle d'un SNP ou d'un haplotype (adapté de Walter Santana-Garcia et al : https ://www.biorxiv.org/content/10.1101/623090v1).

variation-info Cet outil s'appuie sur les informations de variations génétiques d'*Ensembl* (HUNT, MCLAREN, GIL et al., 2018) annotées et installées sur le serveur correspondant à chaque génome particulier (les variantes humaines sont installées sur le serveur Metazoa). Il peut prendre deux entrées différentes : identifiants des variants (rsID) ou locus génomiques au format BED. *Variation-info* récupère les informations des variants correspondant aux identifiants ou les informations des variants situés dans les locus génomiques. Les variants installées sur les serveurs *RSAT* ont été traitées pour supprimer les variants avec des annotations incomplètes (sans allèles) ou des coordonnées ambiguës (les coordonnées des allèles ne correspondent pas). Lorsque les utilisateurs ont leurs propres collections de variants, ils peuvent ignorer cet outil et utiliser directement l'outil *convert-variations* de la suite *RSAT*.

convert-variations Assure l'inter-conversion de différents formats de fichiers tels que VCF, GVF et *varBed*. *varBed* est un format interne de *RSAT* facilitant l'extraction de la séquence entourant le variant.

retrieve-variation-seq Récupère la séquence entourant le variant (par défaut 30 bp de part et d'autre) et produit une séquence pour chaque allèle. L'outil peut prendre en entrée un fichier *varBed* (format produit par *variation-info* ou *convert-variations*). La sortie est fournie dans un format appelé *varSeq*, chaque ligne donnant un allèle avec sa séquence environnante.

variation-scan Effectue le balayage des allèles avec un PSSM et compare les scores et les P-valeurs entre les allèles pour évaluer l'effet putatif sur la liaison du TF. Il nécessite en entrée un fichier *varSeq* (fichier de sortie de *retrieve-variation-seq*), un motif ou une collection de motifs *RSAT* non redondants (Castro-Mondragon, Jaeger, Thieffry et al., 2017), et un modèle de fond (Turatsinze, Thomas-Chollier, Defrance et al., 2008).

Tous les serveurs publics *RSAT* sont accessibles :

1) à partir du portail *RSAT* à l'adresse http://www.rsat.eu/. Les serveurs Web *RSAT* sont librement accessibles à tous les utilisateurs sans obligation de connexion.
2) tous les programmes de *RSAT* peuvent également être utilisés directement à partir du shell Unix. L'accès au shell est moins intuitive que l'interface Web, mais elle permet de réaliser des analyses plus complexes, et il est très pratique pour automatiser des tâches répétitives (http://rsat.sb-roscoff.fr/distrib/tutorial_shell_rsat.pdf).
3) à partir des *RSAT RESTful API*. Une API compatible REST, ou « RESTful », est une interface de programmation d'application qui fait appel à des requêtes HTTP pour obtenir (GET), placer (PUT), publier (POST) et supprimer (DELETE) des données. Actuellement les *RSAT RESTful API* sont en implémentation dans les langages de programmation perl, python et R. Les utilisateurs peuvent déjà avoir accès aux *RSAT RESTful API* pour un sous-ensemble des outils *RSAT* tels que *fetch-squences, retrieve-variation-seq, supported-organisms, variation-info et variation-scan*. En particulier, dans le cadre de ma thèse, nous avons développés des scrits en langage R pour les outils *variation-info, retrieve-variation-seq et variation-scan* qui sont inclus dans le workflow (voir chapitre 6).

1.5. Objectifs de la thèse

Les travaux de cette thèse s'articulent autours de deux parties. La partie expérimentale porte sur la caractérisation moléculaire des infections submicroscopiques (détectées par PCR) à *Plasmodium falciparum* chez les femmes Congolaises asymptomatiques pendant l'accouchement. La partie bioinformatique concerne l'analyse des variants régulateurs associés à la susceptibilité au paludisme sévère.

1.5.1. Partie expérimentale

Le paludisme est un fléau qui sévit dans les régions tropicales, en particulier en Afrique sub-saharienne. Les femmes enceintes et les enfants de moins de 5 ans sont les plus vulnérables. Dans les zones de forte transmission, où les niveaux d'immunité acquise ont tendance à être élevés, l'infection par *P. falciparum* est habituellement asymptomatique pendant la grossesse. Toutefois, les plasmodies indétectables par microscopie, mais qui peuvent néanmoins être détectées à l'aide d'outils moléculaires tels que la PCR peuvent être présentes dans le placenta et contribuer à une anémie maternelle même en l'absence d'une parasitémie périphérique avérée. Une anémie maternelle, comme une parasitémie placentaire, peut être responsable d'un faible poids de naissance, un facteur important dans la mortalité infantile.

Afin de prévenir efficacement l'infection plasmodiale chez la femme enceinte, l'OMS recommande la prise du TPI-SP dès le début du deuxième trimestre de la grossesse. Depuis l'introduction du TPI-SP en République du Congo en 2006, aucune investigation n'a été menée pour évaluer son impact chez la femme enceinte. Des études menées dans le pays entre 2003 et 2005 avaient déjà rapporté des pourcentages élevés (près de 30%) d'échecs thérapeutiques avec la SP pour le traitement du paludisme simple chez les enfants et une prévalence élevée des mutations associées à la résistance du *P. falciparum* à ce médicament.

Entre 2014 et 2015, la fondation Congolaise pour la recherche médicale a conduit une étude chez les femmes enceintes qui venaient accoucher au centre de santé integré de Madibou (quartier sud de Brazzaville). La plupart de ces femmes étaient sous traitement préventif intermittent à base de la sulfadoxine-pyriméthamine (TPI-SP) dès le deuxième trimestre de la grossesse. Chaque dose de sulfadoxine–pyriméthamine (comprimés de 500 mg de sulfadoxine et 25 mg de pyriméthamine) consistait en trois comprimés pris ensemble. Des doses ont été administrées à trois reprises pendant la grossesse lors des consultations prénatales (CPN) suivie d'au moins 1 dose supplémentaire au moins 1 mois plus tard.

Plusieurs études antérieures sur la diversité antigénique de P. falciparum, appariées dans le sang périphérique, le sang placentaire et le cordon ombilical, ont montré des résultats variés selon les pays. Certaines études ont montré un chevauchement parasitaire constaté pour chacun de ces compartiments, tandis que d'autres ont révélé une sous-population de parasites périphériques séquestrés dans le placenta (Kassberger et al., 2002 ; Mayengue et al., 2004 ; Fadlelseed et al., 2017 ; Guitard et al., 2010 ; Cohee et al., 2016). Des travaux récents publiés dans la même région ont signalé 29 allèles MSP-2 dans des isolats de sang périphérique de femmes enceintes présentant une infection asymptomatique à P. falciparum lors de visites prénatales, suggérant une grande diversité de la population de P. falciparum (Ntoumi et al.,

2016) . Par conséquent, la présente étude transversale visait à caractériser les infections à P. falciparum correspondant au sang périphérique, au sang placentaire et au cordon ombilical chez les femmes congolaises à l'accouchement recevant 1, 2 doses ou plus de sulfodoxine-pyriméthamine.

Malgré la mise en place du TPI-SP, le paludisme chez la femme enceinte en République du Congo n'est pas maîtrisée. Dans le but de comprendre les causes de cette persistance, cette étude consistait à :

1) caractériser les populations parasitaires de *P. falciparum* chez les femmes enceintes Congolaises du sud de Brazzaville sous TPI-SP
2) analyser profil génétique dans le sang périphérique, placentaire et du cordon ombilical des souches de *P. falciparum* en utilisant le marqueur génétique *MSP2*.

Les objectifs de cette étude étaient de :1) déterminer la prévalence des mutations dans les gènes dhfr et dhps chez des femmes enceintes ayant reçu 0, 1 ou 2 doses de SP au cours du TPI ; 2) déterminer la multiplicité et la diversité génétiques des souches de P.falciparum dans ces infections.

1.5.2. Partie bioinformatique

La susceptibilité/résistance au paludisme peut s'expliquer par la contribution des facteurs génétiques de l'hôte. L'apport des facteurs génétiques n'a montré que 2% de formes de résistance au paludisme par la variation de l'hémoglobine S. Les bases moléculaire de la susceptibilité/résistance au paludisme sont encore mal connues. Il est donc nécessaire d'étendre le champ des recherches sur la susceptibilité/résistance au paludisme à l'échelle des études d'association du génome entier (GWAS).

Les études GWAS visent à trouver les variants polymorphiques (généralement les SNPs) qui sont liés à un phénotype particulier (par exemple susceptibilité à une maladie). Les SNPs associés à une maladie résidant dans les régions codantes, en particulier les mutations non synonymes, sont généralement faciles à interpréter en termes d'impact sur la séquence protéique, qui peut résulter en un impact fonctionnel. Cependant, identifier l'impact fonctionnel des SNPs des régions non-codantes est plus difficile. Ces SNPs peuvent modifier l'affinité de liaison d'un facteur de transcription (TF) à l'ADN, qui modifie à son tour l'expression des gènes-cibles, contribuant ainsi au phénotype de la maladie. On parle alors de "SNPs régulateurs (rSNPs). Il est difficile de caractériser ces rSNPs, car le SNP détecté dans une étude d'association n'est pas forcément causal du trait phénotypique étudié (par exemple une maladie), et les procédures expérimentales peuvent être coûteuses et demander beaucoup de travail pour la détection des rSNPs. Par conséquent, les approches *in silico* pour identifier les candidats rSNPs sont nécessaires pour améliorer l'efficacité de l'interprétation des maladies associées aux rSNPs.

La partie bioinformatique de ma thèse a comporté une partie méthodologique sur le développement d'un workflow de prédiction des rSNPs, et une application à la détection des rSNPs associés à différents cas d'études. Les objectifs de cette partie sont :

1) développer un workflow qui permet de prédire des SNPs régulateurs

2) tester la pertinence de ce workflow sur base d'une série de cas d'étude de types différents (tuberculose, diabète de type II, sepsis, Alzheimer)
3) appliquer le workflow pour prédire les rSNPs associés au paludisme
4) Développer une interface Web conviviale pour permettre à des biologistes d'utiliser ce workflow.

2. Résultats

2.1. Détection des variants régulateurs à partir d'études GWAS : le workflow cisreg-GWAS

2.1.1. Motivation

Un grand nombre de SNPs dérivés d'études de GWAS sont localisés dans des régions génomiques non-codantes, ce qui accentue la nécessité d'une identification efficace des rSNPs. Cependant, cette tâche reste lourde car l'annotation de la partie régulatrice du génome humain est plus difficile que celle des régions codantes. Je décris dans ce chapitre le workflow nommé **cisreg-GWAS** qui combine différentes ressources bioinformatiques pour prédire l'impact sur la régulation à partir de variants associés à un trait phénotypique (par exemple maladie) documenté dans le études GWAS. Dans le cadre de ma thèse, j'ai appliqué ce workflow pour détecter les variants associés à la susceptibilité au paludisme sévère. De façon plus générale, **cisreg-GWAS** est un outil utilisable par des biologistes pour analyser d'autres traits phénotypiques.

2.1.2. Matériel et méthodes

2.1.2.1. Environnement pour l'analyse informatique et statistique

Le développement de ce workflow et toutes les analyses statistiques sont réalisés avec le langage R dans un environnement de programmation et d'analyse RStudio. Les commandes R sont intégrées dans un rapport au format markdown qui permet de commenter les approches et les résultats, et d'exporter un rapport scientifique en différents formats (pdf, HTML, docx).

2.1.2.2. Outils et bases de données bioinformatiques

La procédure d'identification des variants régulateurs (rSNPs) des régions non codantes associés aux traits (exemple : maladie complexe) implique à la fois des protocoles expérimentaux et des ressources bioinformatiques. Plusieurs outils et bases de données de prédiction et d'annotation des rSNPs existent (Rojano, Seoane, Ranea et al., 2018). Chacun a ses propres avantages, inconvénients et particularités. La particularité de **cisreg-GWAS** est qu'il effectue en direct toutes les étapes de l'analyse à partir d'un identifiant de maladie fourni par l'usager, en permettant de paramétrer chacun des outils invoqués. De plus, il combine essentiellement les outils RSAT Variation-tools (*variation-info*, *retrieve-variation-seq* et *variation-scan*) pour la prédiction des rSNPs et facteurs candidats et ReMap pour corroborer les prédictions de TFs candidats sur base de données de localisation

à échelle génomique (ChIP-seq). Les outils et bases de données inclus dans **cisreg-GWAS** sont *GWAS catalog, bioMart, RSAT, ReMap, Ensembl, dbSNP* et des packages R.

Langage R Le logiciel R a été initialement développé par Ross Ihaka et Robert Gentleman au début des années 1990. R est un logiciel libre tout particulièrement dédié à l'analyse statistique qui permet l'analyse de données, leur visualisation et interprétation. La capacité de R réside dans sa logique d'emploi et sa capacité à produire des tableaux de synthèse et graphique de qualité professionnelle. R est un langage de programmation dérivé du logiciel S. R est disponible sous différents système d'exploitation dont Linux, Mac OS X et Windows. C'est un logiciel bien documenté et est téléchargeable gratuitement à l'adresse : https://cran.r-project.org/.

Plusieurs alternatives auraient été envisageables pour l'implémentation de ce studio, par exemple le langage Python, qui est équipé d'un grand nombre de librairies statistiques et graphiques, et est adopté par un nombre croissant de bioinformaticiens, ou un langage spécialisé pour la gestion de workflows, tel que snakemake.

R peut présenter certaines limitations concernant la gestion de calculs lourds ou le chargement de données massives en mémoire. Cependant, le workflow cisreg-GWAS consiste essentiellement à invoquer des ressources de données et de calcul distantes, via leurs interfaces programmatiques (APIs RESTful ou autres). Les différentes ressources retournent généralement les résultats sous forme de tableaux, qui sont traités de façon efficace par R. R joue donc le rôle d'outil d'intégration et de représentation graphique, et nous permet de générer automatiquement des rapports bien documentés en format markdown.

Environnement de développement RStudio RStudio[16] est un environnement de programmation, libre et multi-plateforme (Microsoft Windows, OS X et GNU/Linux3) pour R. Il est disponible sous la licence libre AGPLv3, ou bien sous une licence commerciale qui offre des fonctionnalités plus avancées. RStudio est disponible en deux versions : RStudio Desktop, pour une exécution locale du logiciel, et RStudio Server qui, lancé sur un serveur Linux, permet d'accéder à RStudio par un navigateur web. RStudio intègre la possibilité d'écrire des notebooks combinant de manière interactive du code R, du texte mis en forme en markdown (http://rmarkdown.rstudio.com/r_notebooks.html) et des appels à du code Python ou Bash5.

Les packages R Un package R est un ensemble de fonctions et de données développées pour renforcer la puissance des fonctionnalités des fonction basiques de R. Un package est composé de code R ou en d'autre langages, ses fonctions et sa documentation.

Interrogation des API par le langage R Certaines des bases de données (BD) rendent les données disponibles via des interfaces accessibles de façon programmatique (Application Programming Interfaces, API), en s'appuyant sur des

standards de services Web (SOAP, RESTful) ou bien sur des librairies spécifiques (ex : modules Perl d'accès à Ensembl).

Je détaille ci-dessous les étapes requises pour la requêtes des données à ces différentes API de service Web à l'aide de R. Plusieurs paquets R ont été développés en tant que clients pour les API de service Web. Celles-ci permettent d'utiliser directement une ressource avec la syntaxe classique de R, sans exiger aucune connaissance des points de terminaison d'API, des requêtes HTTP ou des formats de données tels que XML et JSON.

L'interface de programmation d'application (API) Les API sont un ensemble de procédures permettant à une application logicielle d'interagir avec une autre, qui peut être installée sur le même ordinateur ou invoquée à distance à travers un réseau. Les entrées et les sorties varient d'une API à l'autre, mais le processus est le même : une demande qui suit certaines règles de programmation est soumise et une réponse est renvoyée dans un format standard d'échange de données (JSON, XML, ...).

Dans ce workflow, j'ai utilisé des API de type "service Web", basées sur le standard RESTful. Un client (navigateur) soumet une requête HTTP à un serveur, lequel renvoie une réponse au client. La réponse contient des informations d'état sur la demande et peut également contenir le contenu demandé. Les paramètres d'une requête HTTP sont généralement contenus dans l'URL. Par exemple, pour demander les informations sur les SNPs à l'aide de l'API RSAT RESTful, on soumet la demande RSAT REST.[17]

La demande contient :

— une URL vers le point de terminaison de l'API[18] ;

— une requête et ses paramètres[19].

Les ressources bioinformatiques utilisées dans mon workflow retournent généralement le résultat en format JSON. C'est un format d'échange d'informations structurées, facile d'accès et relativement lisible par l'homme. JSON est un moyen d'encodage qui peut être analysé efficacement par presque tous les langages de programmation. Il est aussi utilisé pour transmettre des données brutes vers ou depuis un composant côté serveur.

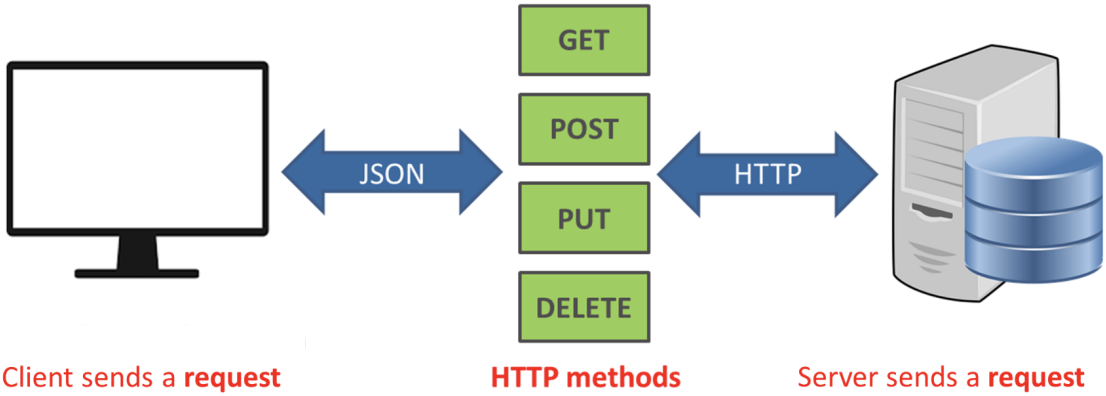


FIGURE 2.1. – L'architecture REST composée d'un client et du serveur en utilisant les méthodes GET, POST, PUT, DELETE. Les requêtes sont basées sur le protocole HTTP universel et les informations sont généralement renvoyées au format JSON que presque tous les langages de programmation peuvent lire (source : https ://phpenthusiast.com/blog/what-is-rest-api).

2.1.3. Workflow cisreg-GWAS d'analyse des variants régulateurs associés aux maladies

Le workflow **cisreg-GWAS** a pour objectif de combiner différentes ressources bioinformatiques (bases de données et outils) afin de détecter les variants non-codants associés à une maladie, susceptibles d'affecter la régulation de la transcription en modifiant les sites de liaison des TFs sur l'ADN. Le workflow repose sur l'intégration d'éléments d'information collectés automatiquement à partir de diverses bases de données génomiques (BioMart, dbSNP, Ensembl), ainsi que sur la sélection des variations susceptibles d'affecter la régulation, en combinant des outils bioinformatiques spécialisés : outils d'analyse de séquences régulatrices (RSAT)[20] et de pics de ChIP-seq (ReMap)[21]. L'outil est conçu de manière générique et peut être adapté à l'étude de variants régulateurs de toute maladie documentée dans le catalogue GWAS. Afin de faciliter son utilisation par un biologiste, le workflow peut être utilisé automatiquement avec une interface graphique en R shiny (en cours de développement). Les résultats sont décrits dans un rapport généré automatiquement (en utilisant le langage à balises R markdown), qui est illustré par des figures et des tableaux.

Le workflow **cisreg-GWAS** se base sur une **approche non supervisée**, c'est-à-dire que la prédiction de rSNPs ne se repose pas sur la base d'un ensemble d'exemples de rSNPs validés expérimentalement. L'utilisateur explore les SNPs associés à une maladie particulière sans aucune connaissance préalable des facteurs transcriptionnels potentiellement impliqués. Le défi consiste à combiner différentes ressources bioinformatiques afin de proposer des rSNPs qui altèrent les sites de liaison des TFs potentiellement pertinents. Le workflow est conçu de manière générique pour l'analyse de n'importe quelle maladie (c'est-à-dire que l'utilisateur peut analyser les SNPs documentés dans le catalogue GWAS ou une liste quelconque des SNPs).

Le workflow présente différentes fonctionnalités pour assurer la reproductibilité scientifique.

— **Automatisation** : toutes les étapes du workflow sont exécutées automatiquement sans intervention manuelle.
— **Reproductibilité** : les résultats peuvent être reproduits à tout moment (pour autant que les BD ne changent pas).
— **Ré-utilisabilité** : le workflow assure la mise à jour des résultats de l'analyse en fonction de la mise à jour des BD. Il peut également être ré-utilisé pour analyser d'autres traits que ceux pour lesquels il a initialement été conçu (susceptibilité au paludisme).
— **Traçabilité** : les paramètres d'analyse sont automatiquement sauvegardés dans un fichier (format YAML) permettant de vérifier a posteriori.

-**Flexibilité** : les paramètres sont chargés via un fichier de configuration (au format YAML) qui est lisible par un être humain et peut être facilement modifié par l'utilisateur (via n'importe quel éditeur de texte), sans avoir besoin d'entrer dans le code R.

— **Portabilité** : le workflow peut-être utilisé sous une machine OS X, Linux (et vraisemblablement, sous Windows, sous réserve de validation).
— **Economie d'installation de ressources locales** : l'utilisateur se connecte à distance et donc n'a pas besoin d'installer les BD et les outils localement.

2.1.4. Les étapes d'analyse

La prédiction des SNPs régulateurs (rSNPs) par le workflow **cisreg-GWAS** est divisée en six étapes. Les sorties d'un programme constituent les entrées du suivant. Le workflow est automatisé dans le langage statistique R qui fait appel à des outils et bases de données à distance via l'interface REST (GWAS catalog, Ensembl, JASPAR, RSAT), par les packages R (BioMart, XGR, TissueEnrich).

Les principales étapes analytiques du workflow sont résumées ci-après et sont illustrés à la Figure 2.2.

1. Collecte de SNPs associés à un trait phénotypique (par exemple susceptibilité à une maladie) dans le catalogue GWAS.
2. Collecte des SNPs en déséquilibre de liaison à partir de l'interface Ensembl RESTful.
3. Sélection des SNPs d'intérêt (SNPs Of Interest, SOIs).
4. Enrichissement de l'ensemble des SOIs pour les maladies.
5. Identification de rSNPs affectant potentiellement la liaison d'un TF (en considérant tous les motifs d'une base de données de TFs).
6. Recherche des pics de ChIP-seq co-localisés avec les SNPs d'intérêt.
7. Prédiction des rSNPs et des TFs potentiellement affectés par chaque rSNP.

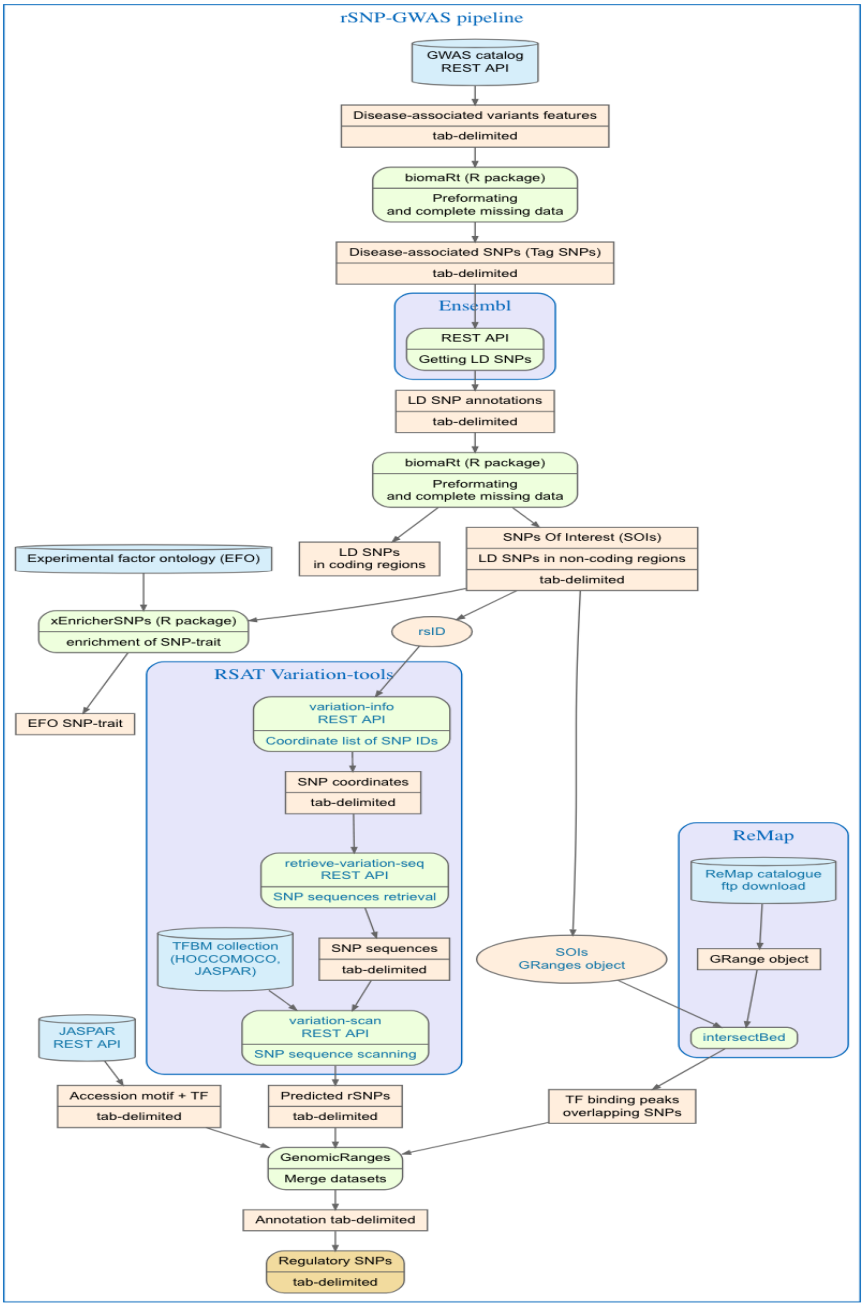


FIGURE 2.2. – Diagramme d'analyse des SNPs régulateurs par le workflow cisreg-GWAS. L'outil interconnecte plusieurs outils bioinformatiques et intègre les informations collectées automatiquement à partir des bases de données GWAS catalog , Ensembl, dbSNP et JASPAR. La prédiction des rSNPs est faite ensuite par une combinaison des l'outils Variation-tools de la suite RSAT (détection des rSNPs) et du catalogue ReMap (prédiction des TFs dont les sites de liaisons sont altérés par les rSNPs).

Le workflow *cisreg-GWAS* a été testé sur différents cas d'étude :

i) Tuberculose, qui a été présenté comme cas d'étude pour une publication sur les outils d'analyse des variants régulateurs (chapitre 10) ;

ii) Diabète de type II (non publiée) ;

iii) Sepsis (collaboration en cours avec Florian Rosier, doctorant du TAGC).

Dans le chapitre suivant, je détaillerai les résultats obtenus étape par étape quand on applique ce workflow au paludisme sévère.

2.2. Paludisme chez la femme enceinte en République du Congo

2.2.1. Article publié : Sub-microscopic *Plasmodium falciparum* infections in matched peripheral, placental and umbilical cord blood samples from asymptomatic Congolese women at delivery

2.2.1.1. Motivation et état de l'art

L'infection plasmodiale, surtout avec *P. falciparum* représente un risque important à la fois fois pour la santé de la mère, du fœtus et du nouveau né. La susceptibilité des femmes enceintes au paludisme s'explique par la présence de souches de parasites se liant spécifiquement à la chondroïtine sulfate A (CSA) présente dans le tissu placentaire.

L'étude de la relation entre génotype et présence des parasites circulants chez la femme enceinte ouvre des voies intéressantes pour le développement de vaccins antipaludiques. La morbi-mortalité liée au paludisme au cours de la grossesse a substantiellement diminué depuis la mise en place des nouvelles politiques de la lutte contre le paludisme recommandées par l'OMS, à savoir : l'accessibilité du diagnostic parasitologique du paludisme, l'utilisation des CTA pour le traitement des cas de paludisme simple et l'utilisation du TPI-SP chez la femme enceinte et aussi chez les enfants de moins de 5 ans dans certains pays endémique.

En effet, pour prévenir efficacement l'infection plasmodiale chez la femme enceinte, l'OMS recommande la prise du TPI-SP dès le début du deuxième trimestre de la grossesse. Au Congo, depuis l'introduction de ce traitement en 2006, aucune investigation n'a été menée pour évaluer son impact. Des études menées dans le pays entre 2003 et 2005 avaient déjà rapporté des pourcentages élevés (près de 30%) d'échecs thérapeutiques avec la SP pour le traitement du paludisme simple chez les enfants et une prévalence élevée des mutations associées à la résistance du *P. falciparum* à ce médicament. Il s'est avéré donc nécessaire de caractériser les populations parasitaires de *P. falciparum* chez les femmes enceintes sous traitement préventif intermittent à base de la sulfadoxine-pyriméthamine (TPI-SP) et d'analyser leur profil génétique dans le sang périphérique, placentaire et du cordon ombilical pendant l'accouchement. Entre 2014 et 2015, la Fondation Congolaise pour la Recherche Médicale a conduit une étude pilote au Centre de Santé intégré de Madibou (CSI, quartiers Sud de Brazzaville).

Cette étude transversale visait à caractériser les infections sub-microscopiques à *P. falciparum* dans des échantillons correspondants au sang périphérique, de placenta et de cordon ombilical chez 370 femmes congolaises asymptomatiques pendant l'accouchement. Des femmes enceintes venant d'accoucher au Centre de santé intégré de Madibou, avaient reçues 1, 2 ou 3 doses de SP pour le TPI. Certaines n'avaient reçu aucune dose. L'infection plasmodiale été recherchée par microscopie en vérifiant la goutte épaisse et le frottis mince. Pour les échantillons qui étaient négative à la microscopie, les infections sub-microscopiques ont été caractérisées par la technique de la PCR en utilisant le marqueur moléculaire du gène *MSP2* de *P. falciparum*. Ensuite, tous les isolats de *P. falciparum* détectés

ont été caractérisés pour la diversité allélique de m de ce gène *MSP2*. Ainsi, la diversité génétique et la multiplicité des infections définie comme le nombre moyen de parasite *P. falciparum* par individu infecté ont été déterminées.

2.2.1.2. Ma contribution

Ma contribution à cette étude a consisté en (1) la collecte des échantillons de sang avant et après l'accouchement au centre de santé intégré de Madibou (quartier sud de Brazzaville), (2) l'extraction d'ADN au laboratoire de Biologie Moléculaire de la FCRM, (3) la caractérisation des infections plasmodiales microscopiques et sub-microscopiques et (4) la rédaction de l'article. Nos résultats ont montré que des infections microscopiques ont été détectées dans des échantillons de sang périphérique, placentaire et de cordon ombilical avec une prévalence respective de 7,3% (27/370), 2,7% (10/370) et 0%. Les échantillons négatifs ont été soumis à une détection sub-microscopique, avec une prévalence respective de 25,4% (87/343), 16,7% (60/360) et 9,4% (35/370) ($P < 0{,}001$). Nous avons également identifié 24 allèles distincts de la famille *3D7* et 20 allèles distincts de la famille *FC27* de *MSP2*. Le nombre total d'allèles pour ces deux familles était de 31, 25 et 19 respectivement dans les échantillons périphériques, placentaires et du cordon. La famille 3D7 était la famille allélique prédominante. La multiplicité des infections dans les échantillons périphériques (1-4 clones), placentaire (1-3 clones) et du cordon (1-3 clones) était similaire ($P = 0{,}9$) et n'étaient pas affectés par l'âge, la gravidité ou la SP.

2.2.1.3. Résumé des résultats

Cette étude qui est la première du genre au Congo, a montré qu'il existe une prévalence élevée d'infections sub-microscopiques et une diversité génétique modérée de *P. falciparum*. La caractérisation du profil génétique des souches circulantes chez les femmes enceintes a montré une diminution de la multiplicité des infections mais avec l'apparition des nouvelles souches plasmodiales. La baisse de la multiplicité des infections suggère que le niveau de transmission du paludisme a aussi baissé. L'âge, la gravidité et les doses de TPI-SP n'interfèrent pas avec la multiplicité des infections. Ces résultats semblent montrer que le TPI dans cette zone de Brazzaville n'est guère bénéfique pour la communauté car les femmes enceintes servent de réservoir de parasites résistants. Cette étude a contribué à la compréhension de la dynamique des parasites dans les types de sang prélevés pendant la grossesse et de la population parasitaire en circulation au Congo.

2.3. Prédiction des SNPs régulateurs associés au paludisme sévère

Les manifestations cliniques du paludisme dépendent des paramètres du parasite, de l'hôte, de l'environnement et des facteurs sociaux. Ceux-ci convergent pour aboutir à des conséquences variables, pouvant aller d'une infection asymptomatique à la maladie grave et la mort. La forte pression de sélection exercée par la diversité génétique de l'hôte a contraint le parasite à une meilleure adaptabilité et a sélectionné la mise en place de la variation antigénique. Plusieurs études démontrent une association entre les variants génétiques et divers phénotypes liés au paludisme (Hill, 1999).

Cette adaptation du parasite a en contrepartie suscité chez l'hôte humain une sélection sur des caractères génétiques qui modifient sa susceptibilité à l'infection paludique. Des études indépendantes témoignent la présence de plusieurs polymorphismes, par exemple dans les gènes ABO, ATP2B4, HBB, MARVELD3 (Timmann, Thye, Vens et al., 2012), TREML4, ZNF536, CSMD1, KLHL3, FAM155A (Timmann, Thye, Vens et al., 2012), TREML4, ZNF536, CSMD1, KLHL3, FAM155A (Ravenhall, Campino, Sepúlveda et al., 2018) associés au paludisme sévère. Parmi ceux-ci, plusieurs polymorphismes du gène ATP2B4 par exemple, ont été caractérisés comme associés à la résistance au paludisme sévère (Timmann, Thye, Vens et al., 2012).

Dans ce chapitre, j'ai appliqué le workflow cisreg-GWAS décrit au chapitre précédent pour détecter des SNPs régulateurs potentiels impliqués dans la susceptibilité susceptibilité au paludisme sévère[22]. Dans les sections suivantes je discuterai les résultats obtenus pour cette maladie aux différentes étapes du workflow.

2.3.1. Collecte de SNPs associés à une maladie (*Disease-Assocuated SNPs*, DA-SNPs)

cisreg-GWAS analyse les SNPs associés à n'importe quel trait (par exemple une maladie) documenté dans le catalogue GWAS[23] (Welter, MacArthur, Morales et al., 2014) ou une liste de SNPs quelconque soumise par l'utilisateur. Tous les SNPs documentés dans le catalogue GWAS sont significativement associés à une maladie avec une P-valeur $< 5x10\text{-}8$.

L'outil prend en entrée l'identifiant **EFO** (Experimental Factor Ontology ou facteur expérimental d'ontology) associé à un trait dans le catalogue GWAS (exemple : EFO_0001068 est associé à la malaria). Pour connaître l'identifiant **EFO** d'un trait, il suffit de taper le nom du trait dans la page de recherche web du site GWAS catalog. **EFO** intègre aussi des ontologie différentes pour certaines maladies. C'est le cas par exemple de la malaria (*EFO_0001068*), du *sepsis* (HP_0100806), de *tuberculosis* (Orphanet_3389). L'outil permet à l'utilisateur de sélectionner un sous-phénotype d'une maladie particulière, par exemple rechercher *tuberculosis* renvoie les catégories *tuberculosis Orphanet_3389 , meningeal tuberculosis EFO_1000039, response to anti-tuberculosis drugs EFO_0007918, susceptibility to Mycobacterium tuberculosis infection measurement EFO_0008407* et bien d'autres phénotypes encore. Le sous-phénotype désiré peut alors être sélectionné au moyen de l'identifiant

EFO.

La collecte d'informations sur les SNPs associés à un trait se fait via l'interface REST qui permet un accès programmatique au catalogue GWAS par la plupart des langages de programmation actuels et renvoie les données au format JSON. Le workflow génère les requêtes via le langage R, et exporte en local les informations concernant l'association SNP-trait dans un fichier tabulaire (format .tsv, qui peut être ouvert par un tableur comme Microsoft Excel ou Libre Office Calc). En cas de redondance des SNPs associés (due aux études GWAS réplicatives pour un SNP dans différentes populations), Le workflow supprime les associations redondantes trait-SNP pour obtenir une liste de SNPs uniques avec des P-valeurs d'association les plus significatives.

Je définis comme SNPs associés à un trait (*Disease-Associated SNPs*, DA-SNPs), tous les SNPs localisés dans les régions codantes ou non codantes de l'ADN et associés au trait phénotypique interrogées dans le catalogue GWAS. Il convient de noter que ces SNPs sont rarement causaux de la maladie, mais sont susceptibles d'être inclus dans un haplotype contenant un variant causal.

J'ai donc testé l'outil **cisreg-GWAS** avec les données GWAS sur la *malaria*. L'outil prend en entrée la requête sous forme de l'identifiant EFO_0001068 en spécifiant la population africaine (AFR). L'outil a retourné 6 études GWAS différentes réalisées en Afrique (Jallow, Teo, Small et al., 2009 ;Ravenhall, Campino, Sepúlveda et al., 2018, Malaria Genomic Epidemiology Network, Band, Rockett et al., 2015, Ding, Andrade, Manolio et al., 2013) qui rapportaient 45 variants associés au paludisme sévère. On distingue 31 variants distincts, car les 14 autres SNPs sont une redondance venant de la mise en évidence d'un même variant associé au paludisme dans ces différentes études (voir Tableau 2.1).

Table 2.1. – SNPs associés susceptibilité au paludisme sévère.

Type	Number.of.results
Nombre d'études	6
SNPs associés à la maladie	45
SNPs non redondant associés à la maladie	31
SNPs avec rsID valide	31
SNPs sans rsID	0
SNPs étudié dans plus d'une étude	14
SNPs étudié dans une étude	17

La majorité des 31 DA-SNPs sont localisés dans les régions non-codantes de l'ADN (Figure 2.3) (Rogano et al., ont prouvé que la majorité des variants associé à un trait phénotypique ou une maladie sont localisés dans les régions non-codantes (Rojano, Seoane, Ranea et al., 2018)).

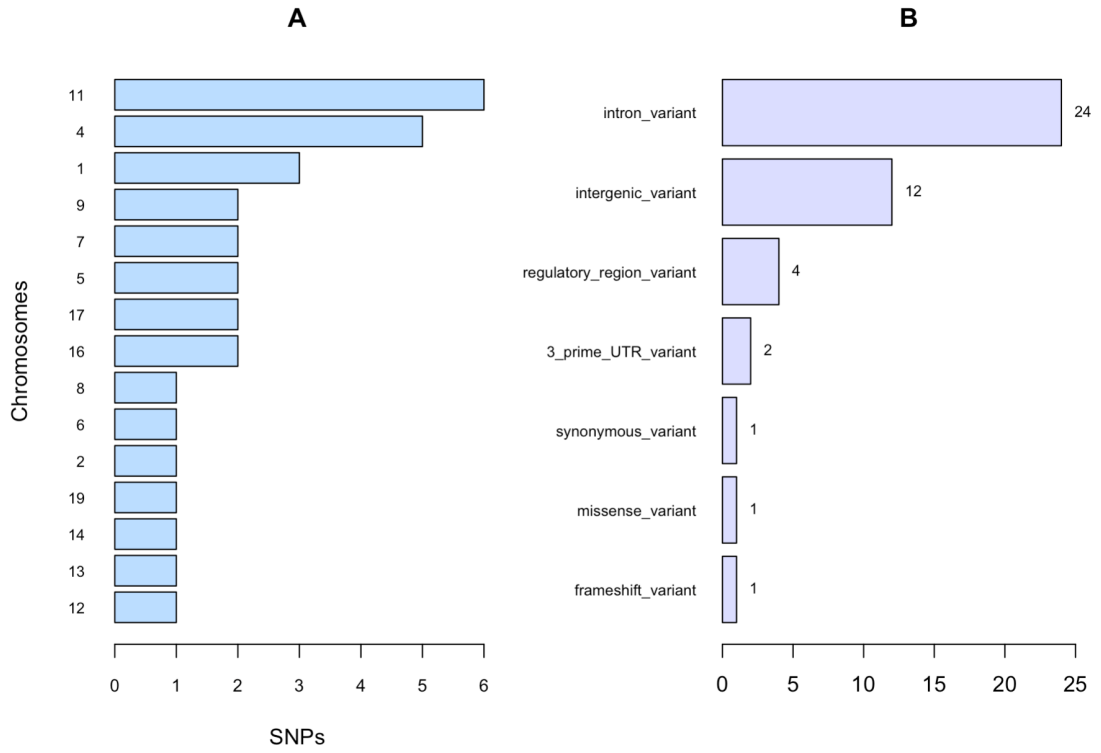


FIGURE 2.3. – Informations génomiques des SNPs associés au paludisme sévère des études GWAS. (A) Nombre de SNPs par chromosome. La plupart des SNPs associés sont dans les chromosomes 11 et 4. (B) Répartition de SNPs selon le contexte génomique. La majorité des SNPs sont localisés dans les régions non-codantes.

Tous les 31 identifiants (rsIDs) des DA-SNPs ont été validés par le package BioMart, qui interroge la base de données Ensembl et que nous avons configuré pour obtenir les coordonnées génomiques des SNPs et le type de région où on les trouve .

2.3.1.1. Les données manquantes

Les SNPs associés au trait ou soumis par l'utilisateur sont automatiquement purgés pour éliminer les identifiant ambigus ou mal annotés (c'est-à-dire l'identifiant du variant n'existe pas dans la nomenclature rsID). C'est le cas de données GWAS du trait *" Susceptibility to Mycobacterium tuberculosis infection measurement"*, où 19 variants associés ont des identifiants non rsID, mais plutôt à des régions HLA associées à la susceptibilité à *Mycobacterium tuberculosis* (MT). Par exemple DRB1*04 :07 est un variant HLA qui n'est pas dans la nomenclature rsID (voir Figure 2.4).

DISEASE.TRAIT	CHR_ID	SNPS	SNP_ID_CURRENT
Tuberculosis	*NA*	DRB1*04:03	*NA*
Tuberculosis	*NA*	DRB1*04:07	*NA*
Tuberculosis	*NA*	DRB1*04:08	*NA*
Tuberculosis	*NA*	DRB1*11:01	*NA*
Tuberculosis	*NA*	DRB1*15:01	*NA*
Tuberculosis	*NA*	chr1:7849205	*NA*
Tuberculosis	*NA*	chr12:41394442	*NA*
Tuberculosis	6	rs114046333	114046333
Tuberculosis	1	rs116245320	116245320
Tuberculosis	20	rs117583211	117583211
Tuberculosis	16	rs117952535	117952535
Tuberculosis	14	rs1243550	1243550
Tuberculosis	20	rs12479598	12479598

FIGURE 2.4. – Extrait de la table des variants associés à la susceptibilité à Mycobacterium tuberculosis (MT) qui montre différents SNPs caractérisés par des identifiants de régions HLA (encadré en rouge) et qui ne sont pas documentés par des identifiants rsID. Ces SNPs sont écartés des analyses ultérieures car leurs identifiants sont ambigüs (mal annotés).

Recherche des données manquantes dans bioMart Il arrive aussi souvent que les informations génomiques manquent pour un variant donné. Cela peut être le numéro du chromosome, la position génomique et même parfois l'identifiant rsID qui manque. Pour éviter ce problème le workflow fait appel au paquet R BioMart qui collecte les informations manquantes pour le variant à partir de la base de données dbSNP [24].

Le fichier de sortie est une liste de SNPs associés à la maladie qui sera le fichier d'entrée dans la recherche des SNPs en LD.

2.3.2. Collecte des SNPs en déséquilibre de liaison

Un variant génotypé statistiquement associé à un phénotype donné, n'est pas forcément un variant causal. Le signal d'association détecté pour ce variant peut être dû à un déséquilibre de liaison avec le variant causal non-génotypé.

Une grande proportion de SNPs associés à des maladies dans des études GWAS (HINDORFF, SETHUPATHY, JUNKINS et al., 2009) sont des marqueurs de risque,

qui ne sont pas directement impliqués dans le déclenchement de la maladie mais sont statistiquement associés à la maladie et sont souvent en déséquilibre de liaison (*linkage disequilibrium*, LD) avec des SNPs causaux de la maladie, qui sont généralement inconnus. L'étude d'association par déséquilibre de liaison est une approche de recherche de causalité entre SNP et phénotype d'intérêt.

L'outil collecte, pour chaque SNP associé à la maladie, les SNPs en déséquilibre de liaison. Cette sélection est limitée aux SNPs proches (distance par défaut $<$ 200 kb, réglable par l'utilisateur) qui semblent être en LD avec des SNPs déjà connus associés à la maladie. Le workflow utilise les web services Ensembl REST[25] qui permettent de récupérer les SNPs en LD en spécifiant la population que l'on souhaite étudier (AFR, AMR, ASN, CEU), le seuil de LD (par défaut $r^2 = 0.8$). Il calcule et renvoie les valeurs LD entre le SNP associé au trait d'intérêt et toutes les autres SNPs en LD (Figure 2.5). La taille de la fenêtre de recherche des LD SNPs est définie par défaut sur 200 kb et un nombre maximal de SNPs en LD par SNP associé est de 500 par défaut. L'outil limite ce nombre de 500 LD pour des contraintes informatiques du temps de calcul et en raison à d'un manque de pertinence des groupes de déséquilibre de liaison comportant plus de 500 SNPs.

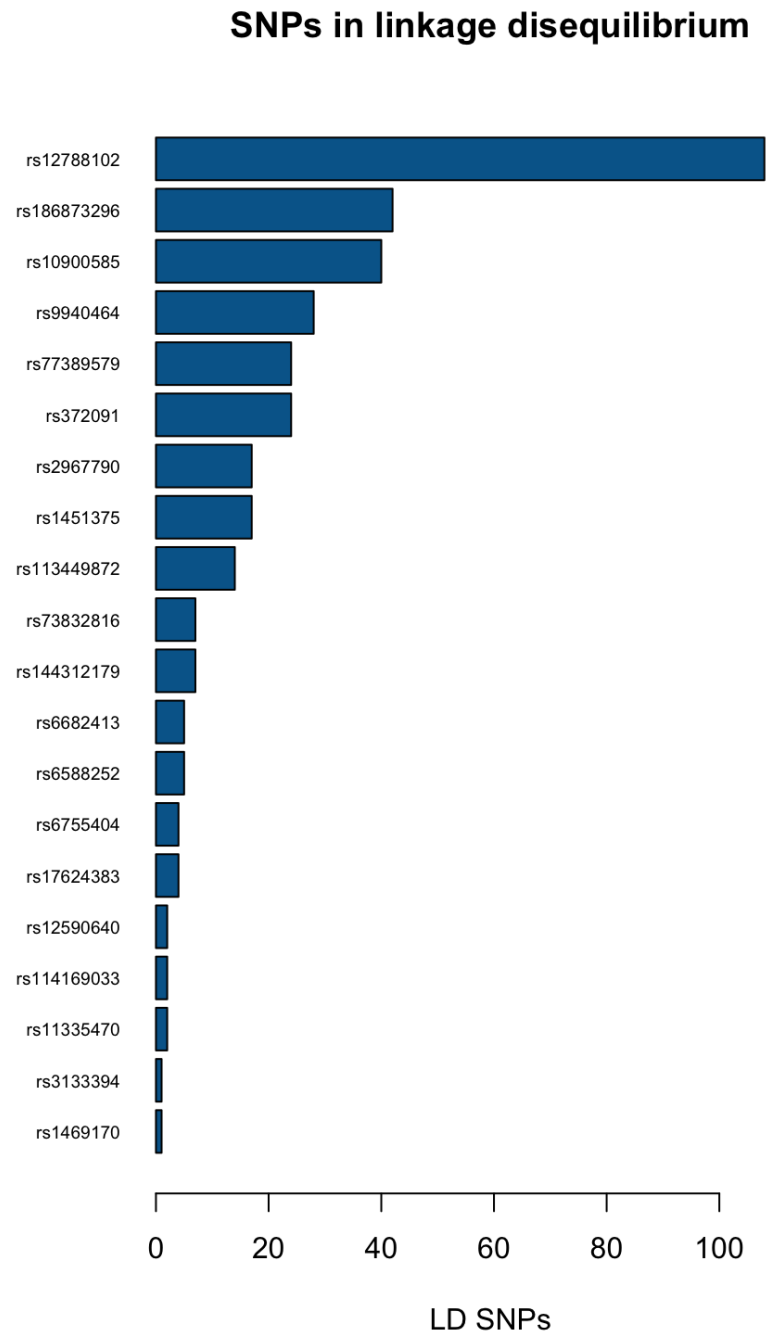


FIGURE 2.5. – Barplot des SNPs en déséquilibre de liaison (LD). Chaque barre représente un groupe de liaison entre le SNP associé associés (en abscisse) à la susceptibilité à Mycobacterium tuberculosis (MT). L'axe des ordonnées représentent le nombre LD SNPs

La recherche des SNPs en LD avec les DA-SNPs via *Ensembl* (DURINCK, MOREAU, KASPRZYK et al., 2005) à partir des services web REST[26] a été faite dans les populations Africaines (AFR) avec un seuil de coefficient de régression ($r^2 = 0,8$) et une distance maximale de 200 bp. Trois cent cinquante (350) SNPs ont été trouvés

en LD avec les 31 DA-SNPs de départ. Au final, ceci donne un ensemble de 379 SNPs (union des SNPs DA et LD SNPs). L'union des SNPs DA et LD diffère de la somme car certains DA SNPs sont mutuellement en LD.

2.3.3. Sélection des SNPs d'intérêt (SOIs)

Pour l'analyse des variations régulatrices, nous écartons les SNPs localisés dans des séquences codantes. Dans le cas de la malaria, 4 des 379 SNPs (DA + LD) sont codants, nous nous retrouvons donc avec 375 SNPs (DA-SNPs + LD SNPs) localisés dans les régions non-codantes (voir Figure 2.6). Nous les appelons ci-dessous "SNPs d'intérêt" (dans le workflow et les figures : *SNPs of Interest*, SOIs).

Les proportions des SOIs ne correspondent pas aux taille des chromosomes, car les comptages des SOIs par chromosome prennent en compte tous les SNPs d'un même groupe de liaison. La majorité de SOIs sont localisés dans le chromosome 11. Des études sur les facteurs génétiques liés au paludisme ont mis en évidence notamment l'hémoglobine S (HbS), l'hémoglobine C (HbC) et l'alpha-thalassémie, qui sont les troubles humains monogéniques les plus courants associés au chromosome 11. Ils confèrent un degré de protection remarquable contre le paludisme sévère à *P. falciparum* (Taylor, Cerami et Fairhurst, 2013, Rees, Williams et Gladwin, 2010).

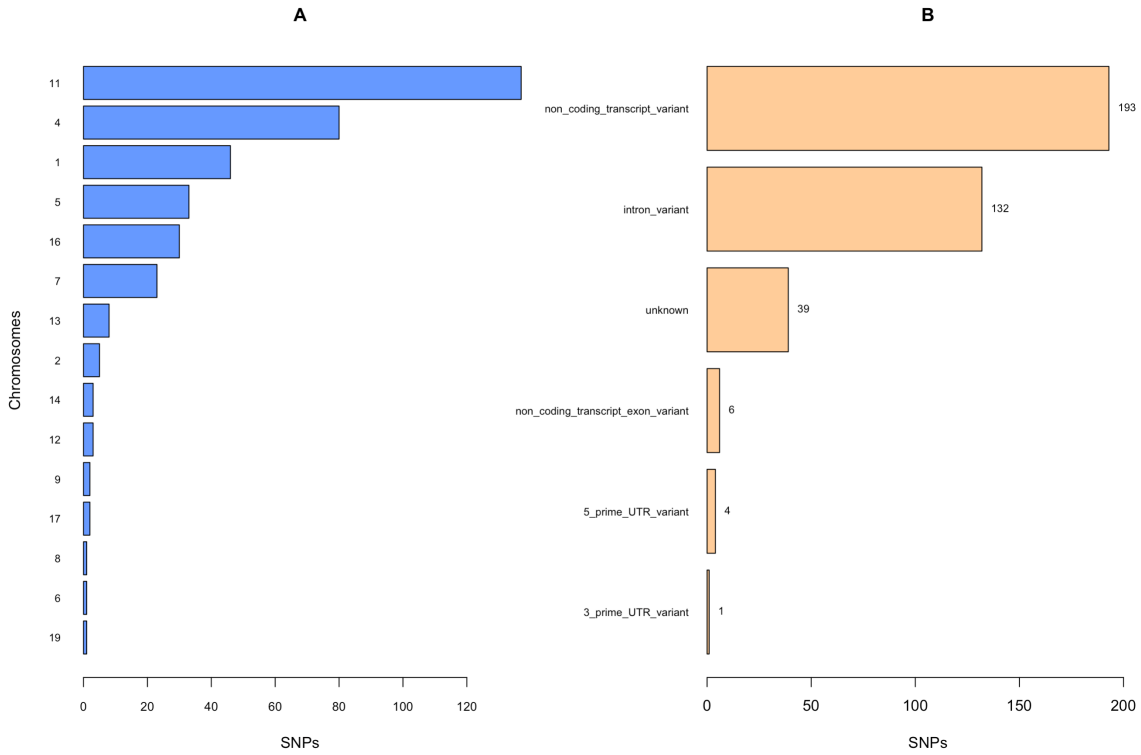


Figure 2.6. – SNPs d'intérêt (SOI) associés au paludisme sévère localisés dans les régions non-codantes de l'ADN (figure B). Les annotations génomiques ont été faites par le package biomaRt. Trente neuf SOIs ont une localisation génomique inconnue. Les SOIs associés au paludisme sévère sont identifiés dans 15 chromosomes parmi lesquels le chromosome 11 contient la majorité de SOIs.

2.3.3.1. Groupes de liaison

Je définis des régions génomiques qui couvrent le DA-SNP et l'ensemble des SNPs en LD correspondants. Ces régions sont nommées groupe de liaison. Un total de 29 groupes de liaison sont identifiés à partir des groupes de SNPs en LD.

La figure Figure 2.7 indique le nombre de SNPs par groupe de déséquilibre de liaison (20 premiers résultats). Le groupe le plus important comporte 109 SNPs. Ceci illustre l'intérêt de collecter les groupes de LD, car chacun de ces 109 SNPs pourrait avoir un effet régulateur responsable de l'association détectée avec le SNP associé au paludisme sévère. Les SNPs associés peuvent être considérés comme des hameçons permettant de collecter un plus grand nombre de candidats pour identifier l'effet causal ("Tag SNP").

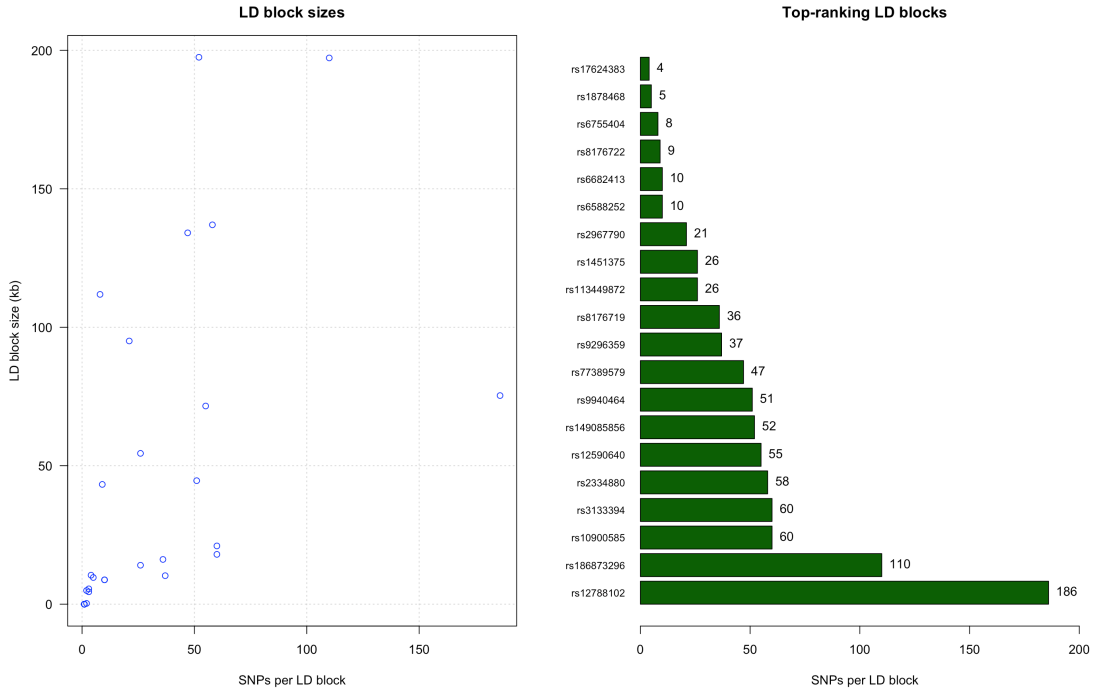


FIGURE 2.7. – Représentation des groupes de liaisons des SNPs associés au trait malaria. Le panneau de droite est un sous-ensemble de 20 groupes ayant plus de SNPs. Au total, on a identifié 29 groupes de liaison (groupe LD), soit un pour chaque SNP associé au paludisme sévère (DA-SNP) sur la base des paires de SNP LD extraites de *Ensembl*. Un calcul de coordonnées des blocs LD en tant que limites de la région englobant tous les SNPs en LD avec un SNP DA donné. Le panneau de gauche représente le nombre de SNPs par taille de blocs en kilobase (kb). Le SNP rs12788102 en LD avec 186 SNPs a une taille de 60 kb alors que le bloc de 180 kb ne possède que 52 LD SNPs.

2.3.4. Enrichissement des SNPs d'intérêt définis pour les facteurs d'ontologies expérimentales

La fonction *xEnricherSNPs* du package XGR (Fang, Knezevic, Burnham et al., 2016) effectue une analyse d'enrichissement en maladie à base des SNPs en utilisant les traits du catalogue GWAS correspondant à un EFO (de la base de donnée *European Bioinformatics Institute*, EBI ou Institut Européen de Bioinformatique). Cette fonction est utilisée pour évaluer l'enrichissement en maladies des groupes de SOIs retournés par le workflow (Figure 2.8). L'intérêt de cette analyse réside dans le fait qu'il existe des SNPs associés à plus d'une maladie à la fois.

Les 375 SOIs ont montré un enrichissement significatif pour 3 maladies : la malaria (voir Figure 2.8), dans les maladies infectieuses et infections parasitaires. Cet enrichissement confirme la pertinence de ces SNPs.

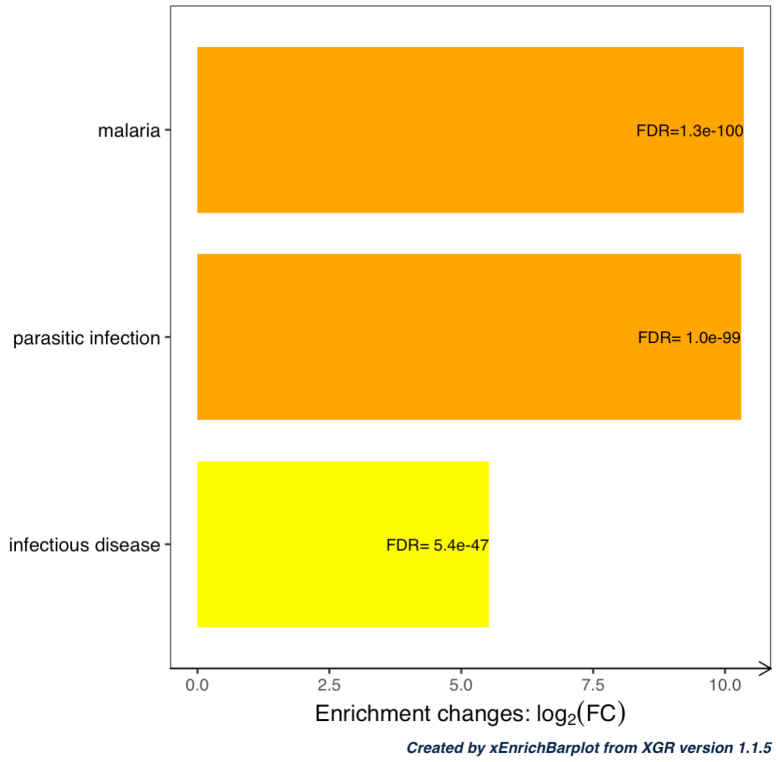


Figure 2.8. – Enrichissement des SNPs d'inteérêt (SOIs) en differents ontologies expérimentales des maladies. Chaque barre du graphique comparant les termes les ontologies expérimentales significatives associées aux SOIs.

2.3.5. Prédiction des SNPs régulateurs affectant potentiellement la liaison aux facteurs de transcription

J'ai utilisé un sous-ensemble d'outils de la suite *RSAT Variation-tools* pour prédire les SNPs qui altèrent les sites de liaisons des TFs, en passant par les services REST de RSAT[27].

2.3.5.1. Développement d'interfaces REST pour les outils d'analyse des variations de la suite RSAT

Les outils RSAT Variation-tools n'étaient pas initialement disponible sous REST. Les services Web RSAT étaient à l'origine basés sur SOAP / WSDL, qui n'est plus supporté par R, alors que le développement du workflow est totalement fait en R. Suite aux besoins d'interopérabilité entre différentes BD et ressources bioinformatiques, nous avons eu l'idée de développer les services REST de quelques outils de RSAT dans le but d'établir une interopérabilité entre les ordinateurs sur Internet. L'interface REST a été développée par Thi-Thuy-Nga Nguyen de l'Ecole Normale Supérieur de Paris, Rosny Thomas de l'IFB-CORE et moi-même.

Les besoins immédiats pour workflow étaient d'interfacer les services REST de Variation-tools (*variation-info, retrieve-variation-seq, variation-scan*) pour le développement de mon workflow. Thi-Thuy-Nga Nguyen et Thomas Rosny ont développé les interfaces REST côté serveur en langage Python et j'ai ensuite développé des scripts clients de test en langage R, qui sont inclus dans mon workflow disponible sur github. Actuellement on compte 7 outils de la suite RSAT interfacés côté serveur en langage Python : *convert-variations, fetch-sequences, peak-motifs, retrieve-variation-seq, supported-organisms, variation-info, variation-scan*[28]. Cette liste sera prochainement complétée pour couvrir les outils RSAT les plus populaires.

Je décris ci-dessous les fonctionnalités des outils de RSAT Variation-tools. Les requêtes RSAT Variation-tools REST sont faites par la méthode POST et le transfert de données au format JSON.

Variation-info : récupère les informations (chromosome, position, SNP, allèle sauvage, allèle muté, fréquence de l'allèle mineur, terme de séquence ontologie) des variants correspondant aux identifiants rsID soumis, en partant d'une liste d'identifiants de variations génétiques d'Ensembl (Hunt, McLaren, Gil et al., 2018). Les coordonnées des variations sont annotées et installées sur le serveur RSAT Metazoa (contient les variants humains). Alternativement, l'outil peut prendre en entrée : une liste de régions génomiques (format BED), et collecte alors tous les SNPs compris dans ces régions, mais je n'ai pas utilisé cette fonctionnalité pour mon workflow, qui prend uniquement les identifiants rsID. Les SNPs installés sur les serveurs RSAT ont été pré-traités pour supprimer les annotations incomplètes (sans allèles) ou les coordonnées ambiguës (qui ne correspondent pas aux coordonnée disponibles dans Ensembl).

retrieve-variation-seq : récupère la séquence entourant le variant et produit une séquence pour chaque allèle. L'outil prend en entrée un fichier de coordonnées de SNPS (format *varBed*, produit par variation-info), et pour les organismes avec des SNPs annotées Ensembl, une liste d'identifiants ou un fichier BED répertoriant les loci génomiques. La sortie est fournie dans un format appelé *varSeq*, chaque ligne donnant un allèle avec sa séquence environnante. Chaque variant à un identifiant interne spécifique permettant de gérer plusieurs SNPs avec différents allèles dans le même fichier.

variation-scan : prend en entrée le fichier de séquences de variants (format varSeq, retourné par retrieve-variation-seq) et un modèle de fond (Turatsinze, Thomas-Chollier, Defrance et al., 2008). Il analyse les séquences des allèles de

chaque SOI avec une matrice poids-position (*position weight matrix*, PWM) donnée ou une collection de motifs et compare les scores et les P-valeur entre les allèles pour évaluer l'effet potentiel sur la liaison du TF. Les PWMs sont utilisés pour évaluer l'affinité de liaison d'un TF pour une séquence d'ADN donnée. L'affinité de liaison est estimée au moyen d'un score de poids (Ws) (*RSAT Var-tools : an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding | bioRxiv*). *variation-scan* compare le Ws obtenu (Ws = Ws_Allele1 - Ws_Alelle2) et la valeur P (P-value = P-value_Allele1 / P-value_Allele2) de chacun des allèles, position par position dans la fenêtre de balayage. Le fichier de sortie est une table délimitée des SNPs localisés dans les motifs de liaison des TFs.

J'ai analysé 375 SOIs associés au paludisme sévère avec les outils *RSAT variation-tools* inclus dans le workflow. Le workflow soumet la liste des rsID des SOIs à l'outil *variation-info* avec le paramètre *Homo sapiens GRCh38* (*Genome Reference Consortium Human Build 38* ou assemblage du génome humain version 38) pour récupérer les informations génomiques de Ensembl de chaque SOI au format *varBed* supporté par l'outil *retrieve-variation-seq* que le workflow utilisé pour récupérer les séquences de 30pb (par défaut) autour de chaque allèle des SOIs. L'analyse et la prédiction de l'impact de la variation sur la liaison du TF pour chaque motif de la collection JASPAR core vertebrate non redondant 2018 inclue dans RSAT (Castro-Mondragon, Jaeger, Thieffry et al., 2017) a été faite avec l'outil *variation-scan* avec les paramètres : P-valeur = 1e-4, P-valeur ratio = 100 (ratio entre les p-valeurs du meilleur et du plus mauvais score du motif avec les allèles) et le modèle de Markov d'ordre 1. Le scan des allèles de chaque SOI par *variation-scan* avec 597 matrices actuelles de JASPAR a permis de prédire 3,386 modifications potentielles de liaison des TFs, couvrant 344 SOIs et 546 motifs distincts. La Figure 2.9 représente les 20 meilleurs motifs modifiés par SNPs régulateurs prédits.

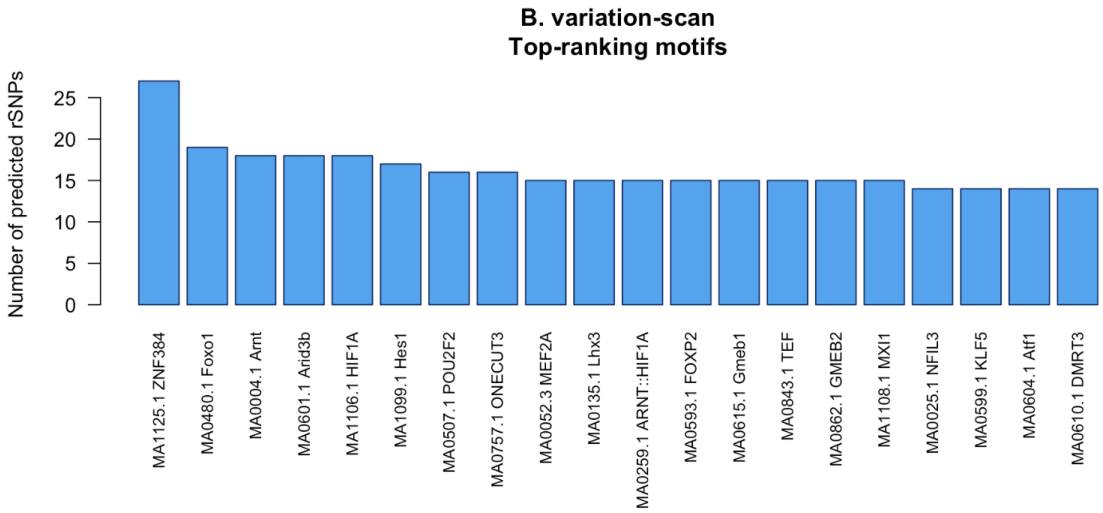


Figure 2.9. – Distribution du meilleur classement des TFBS potentiellement affecté par chaque SNPs d'intérêt (SOI) à partir des résultats de l'analyse de variation-scan. Chaque barree correspond à un motif de liaison d'un TF (source : JASPAR), et la hauteur indique le nombre de rSNPs prédits par variation-scan.

2.3.5.2. Assignation de TFs de JASPAR avec les motifs correspondants dans *RSAT variation-scan*

A cette étape, l'outil **cisreg-GWAS** télécharge la collection de motifs JASPAR CORE non redondant (ensemble de profils dérivés des sites de liaison aux TFs publiés et définis expérimentalement pour les eucaryotes) et les noms de TFs correspondants. A l'aide du service web JASPAR API RESTful, l'outil **cisreg-GWAS** soumet une requête HTTP pour toute la collection des matrices de JASPAR core non redondant au serveur de JASPAR qui retourne des résultats au format JSON qui sont en suite convertis au format tab-délimité (tab-separated values, TSV).

L'outil intègre alors la table des résultats de *variation-scan* et de JASPAR pour attribuer à chaque motif de *variation-scan* le TF correspondant dans JASPAR. Le fichier de sortie est un tableau composé de rSNPs potentiels, annotés avec des identifiants de motifs JASPAR et les noms TFs correspondants (Figure 2.10).

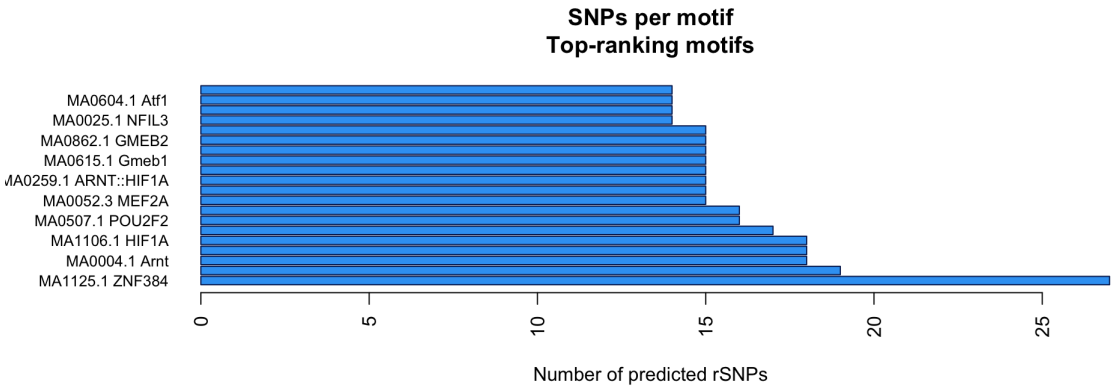


FIGURE 2.10. – Distribution des sites de liaison des facteurs de transcription (TFBS) potentiellement affectés par chaque SNPs d'intérêt (SOI) à partir des résultats de l'analyse de variation-scan. Chaque barree correspond à un motif de liaison d'un TF (source : JASPAR), et la hauteur indique le nombre de rSNPs prédits par variation-scan dans le cas du paludisme sévère.

2.3.6. Recherche des pics de ChIP-seq co-localisés avec les SNPs d'intérêt

Dans cette section, je décris l'annotation en pics de ChIP-seq du catalogue *ReMap* pour les blocs d'haplotypes des SOIs.

Le catalogue *ReMap* (GRIFFON, BARBIER, DALINO et al., 2015) est une analyse intégrative des expériences de ChIP-seq à partir de jeux de données puliques de GEO et d'ENCODE (ENCODE PROJECT CONSORTIUM, 2012c). L'outil **cisreg-GWAS** utilise le package R ReMapEnrich pour identifier la région enrichie

significative du catalogue *ReMap* correspondant au TFBS modifié par les rSNPs potentiels.

Le principe du test est de mesurer la significativité de l'intersection entre les régions d'intérêt et chaque ensemble de pics *ReMap* (un jeu de pics pour chaque expérience ChIP-seq dans *ReMap*). Cette signification est mesurée avec une valeur p, qui représente la probabilité d'obtenir une intersection au moins aussi importante sous hypothèse nulle, c'est-à-dire si nous avions choisi des régions de la même taille de manière aléatoire.

Les utilisateurs peuvent soumettre une liste de régions génomiques qui sont comparées aux régions des pics de ChIP-seq du catalogue afin d'identifier les facteurs enrichis dans ces régions. L'outil calcul en premier le nombre de chevauchement entre les régions génomiques requêtes de l'utilisateur (jeux de données utilisateur) et les pics de fixation de chacun des TFs du catalogue. L'outil génère ensuite des jeux de données des régions génomiques aléatoires ayant les mêmes caractéristiques que le jeu de données de l'utilisateur (même nombre et taille de régions). Un calcul du nombre de chevauchement entre ces régions aléatoires et les pics de fixation de chaque TF est également effectué. Un minimum de 100 jeux de données aléatoires permet de construire, pour chaque TF, la distribution théorique du nombre de chevauchements entre les pics de fixation et les régions génomiques. La valeur d'enrichissement pour chacun des TFs est calculée à partir de ces distributions théoriques (Figure 2.11).

Le principe de cette annotation est de mesurer l'enrichissement en pics de ChIP-seq, c'est-à-dire la mesure de l'intersection entre les régions des SOIs et chaque ensemble de pics *ReMap* (un jeu de pics pour chaque expérience ChIP-seq dans *ReMap*). Cet enrichissement est mesuré avec une valeur p, qui représente la probabilité d'obtenir une intersection au moins aussi importante dans une hypothèse nulle, c'est-à-dire si nous choisissons des régions de la même taille de manière aléatoire. La recherche des pics de ChIP-seq co-localisés avec les SOIs se fait par un calcule d'enrichissement en TFs dans les régions génomiques des SOIs en se basant sur les TFBS.

Cette analyse est réalisée par les étapes suivantes :

1) Téléchargement du catalogue *ReMap* Tous les TBFS du catalogue *ReMap* sont regroupés dans un fichier au format BED téléchargé automatiquement par la fonction *download.file()* appelé "catalogue *ReMap*".
2) Création des fichiers au format *GRange* Le catalogue *ReMap* (*catalog*) et le fichier BED des SOIs (*query*) sont convertis au format *GRange*. Le *catalog* et la *query* sont définis par les attributs : chromosome, position de début, position de fin et brin. Dans le *catalog*, la colonne *name* indique les noms des TFs.
3) Intersection entre le *catalog* et la *query*

Le workflow utilise la fonction *intersectBed()* pour l'intersection des régions génomiques des SOIs et les régions des pics de ChIP-seq.

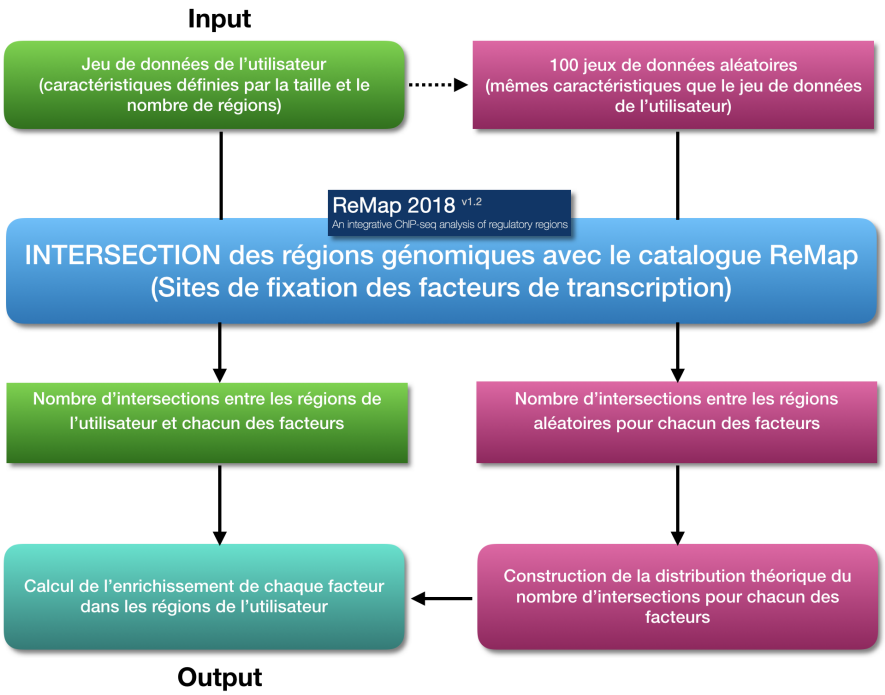


FIGURE 2.11. – Organigramme de l'outil d'annotation et de la méthode de calcul des enrichissements en TFs dans les régions génomiques soumises par l'utilisateur.

L'intersection des régions de 29 groupes de liaison des SOIs avec les régions de pics ChIP-seq du catalog *ReMap* a produit 2,230 chevauchements montrant un ensemble de TFs enrichis dans les régions de ChIP-seq (voir Tableau 2.3).

Le fichier de sortie est un table qui contient les positions génomiques des SNPs qui chevauchent les régions de pics de ChIP-seq documentées dans le catalogue ReMap, les noms des TFs et les rsID. La Figure 2.12 montre les 20 meilleurs régions enrichies en pics de ChIP-seq qui chevauchent les les SOIs.

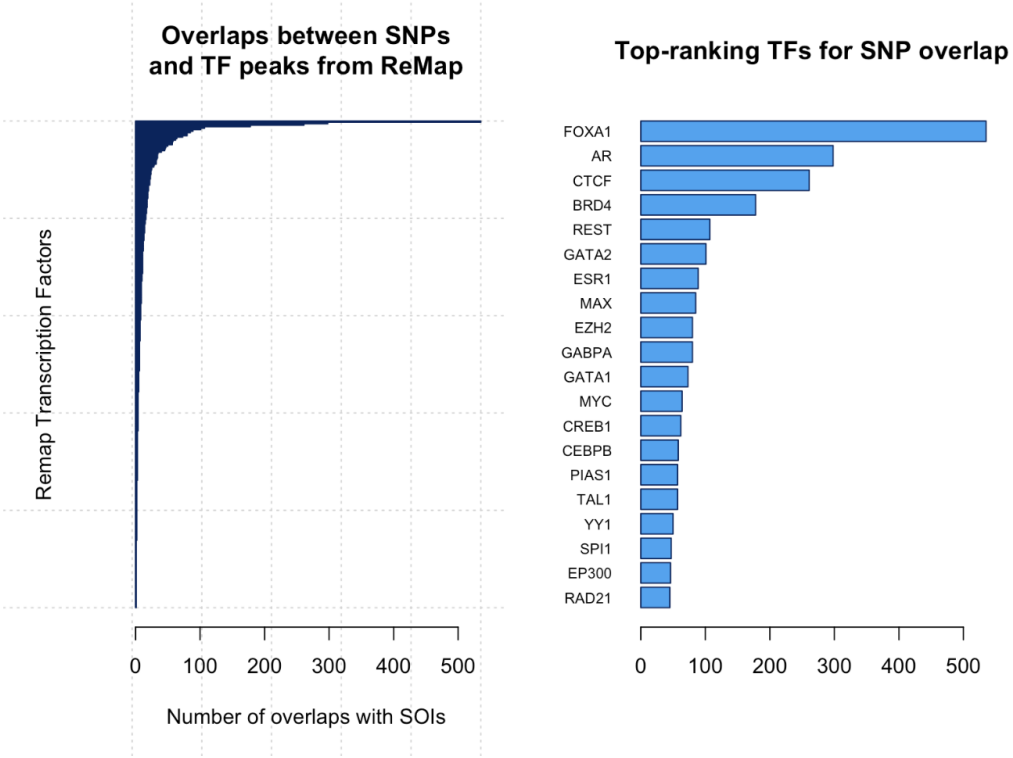


FIGURE 2.12. – Facteurs de transcription (TF) enrichis dans les régions ChIP-seq qui chevauchent avec les groupes de liaison de SNPs d'intérêt (SOI).L'axe des abscisses représente le nombre des SNPs qui chevauchent les régions de ChIP-seq. L'axe des ordonnées représente les TFs enrichis (exemple d'analyse de la susceptibilité au paludisme sévère par cisreg-GWAS).

2.3.7. Prédiction des SNPs régulateurs (rSNPs)

Les interactions entre les TFs et les sites-cibles constituent un mécanisme fondamental dans les réseaux de régulation des gènes, qui déterminent les niveaux d'expression des gènes cibles à un moment donné du développement, dans un tissu donné et dans des conditions spécifiques. Les SNPs localisés dans les TFBS peuvent modifier la régulation de la transcription des gènes cibles en modifiant l'affinité de liaison au TF. L'effet d'un SNP sur la liaison d'un TF est estimé sur la base d'un modèle de matrice poids-position (Position-Weight Matrix, PWM) pour la spécificité de liaison du facteur correspondant.

1) Le workflow **cisreg-GWAS** produit une table de résultats qui inclut toutes les annotations de 170 SOIs co-localisés avec les régions de pics de ChIP-seq de *ReMap* et de 344 SNPs prédits par *variation-scan*.

Le workflow combine les résultats de *variation-scan* et de *ReMap* en vu de proposer des SNPs potentiellement candidats régulateurs, pour lesquels on observe une cohérence entre l'analyse de motifs et la présence de pics de ChIP-seq (Tableau 2.2).

TABLE 2.2. – Potentiels candidats rSNPs associés à la susceptibilité au paludisme sévère

seqnames	var_id	best_variant	worst_variant	ac_motif	remap.tf.name	pval_ratio
chr16	rs12597147	A	G	MA0466.2	CEBPB	209.09
chr16	rs12597597	A	G	MA0148.3	FOXA1	50.00
chr1	rs1541253	C	T	MA0511.2	RUNX2	18.46
chr1	rs1541253	C	T	MA0684.1	RUNX3	45.00
chr1	rs1541254	C	G	MA0473.2	ELF1	33.85
chr1	rs1541254	C	G	MA0474.2	ERG	17.46
chr1	rs1541254	C	G	MA0098.3	ETS1	14.12
chr1	rs1541254	C	G	MA0764.1	ETV4	30.42
chr1	rs1541255	T	A	MA0102.3	CEBPA	13.25
chr1	rs1541255	G	T	MA0473.2	ELF1	30.00
chr1	rs1541255	G	T	MA0474.2	ERG	57.14
chr1	rs1541255	G	T	MA0098.3	ETS1	34.12
chr1	rs1541255	G	A	MA0764.1	ETV4	104.17
chr4	rs181832459	C	G	MA0036.3	GATA2	18.67
chr1	rs35014299	A	G	MA1125.1	ZNF384	863.64
chr1	rs6677188	T	A	MA0507.1	POU2F2	28.33
chr11	rs71480720	G	T	MA0148.3	FOXA1	287.88
chr4	rs77335039	C	A	MA0466.2	CEBPB	14.44
chr4	rs77389579	G	T	MA0507.1	POU2F2	19.68

Un récapitulatif des résultats statistiques d'analyse des SNPs associés au paludisme sévère est rapporté dans la Tableau 2.3.

TABLE 2.3. – Résultats statistiques de l'analyse avec le workflow **cisreg-GWAS** des SNPs associés au paludisme sévère.

workflowSummary	total
Association Studies for the disease	6
Disease SNP associations	45
Disease-Associated SNPs (unique IDs)	31
SNPs with rs identifier	31
SNPs with non-rs identifier	0
SNPs demonstrated by >1 studies	14
SNPs demonstrated by 1 study	17
SNPs with rsID	31
valid rsIDs for BioMart	31
LD : SNPs in Linkage Disequilibrium with DA SNPs	350
Union of DA and LD SNPs	379
SOIs : SNPs of interest (non-coding DA or LD)	375
variation-scan : SNP-motif pairs returned by variation-scan	3386
variation-scan : Number of TFBM altered by the SOIs	546
variation-scan : predicted rSNPs	344
ReMap : overlaps between SOIs and ReMap peaks	2230

workflowSummary	total
ReMap : number of SNPs with ReMap peaks	168
rSNPs : predicted potential rSNPs	11
TFs : TFs whose the TFBS are affected by the rSNPs	12

2) Après comparaisons des résultats de *variation-scan* et *ReMap*, seuls 11 SNPs (voir Tableau 2.2) montrent une correspondance entre le TF immunoprécipité pour l'expérience de pics ChIP-seq et ceux associés au motif JASPAR utilisé par de *variation-scan* (voir Figure 2.13A). Un total de 12 TFs (GATA2, ETS1, CEBPA, CEBPB, FOXA1, ELF1, ERG, POU2F2, RUNX2, RUNX3, ETV4, ZNF384) ont été identifié dans différents types de tissus et cellules documenté dans le catalogue *ReMap* (voir Figure 2.13B). Parmi les 11 potentiels candidats SNPs régulateurs, 4 sont associés au gène ATP2B4 (*ATPase plasma membrane Ca2+ transporting 4*), 2 au gène CDH13 (*cadherin 13*), 1 au gène INPP4B (*inositol polyphosphate-4-phosphatase type II B*), et 3 sans annotation génomique (Figure 2.13C). Trois SNPs (rs1541253, rs1541254, rs1541255) du gène ATP2B4 ont modifie l'activité du promoteur de ce gène (étude en cours par la doctorante Samia NIZAR de l'équipe de Pascal RIHET au laboratoire des technologies avancées pour la génomique et la complexité à Marseille).

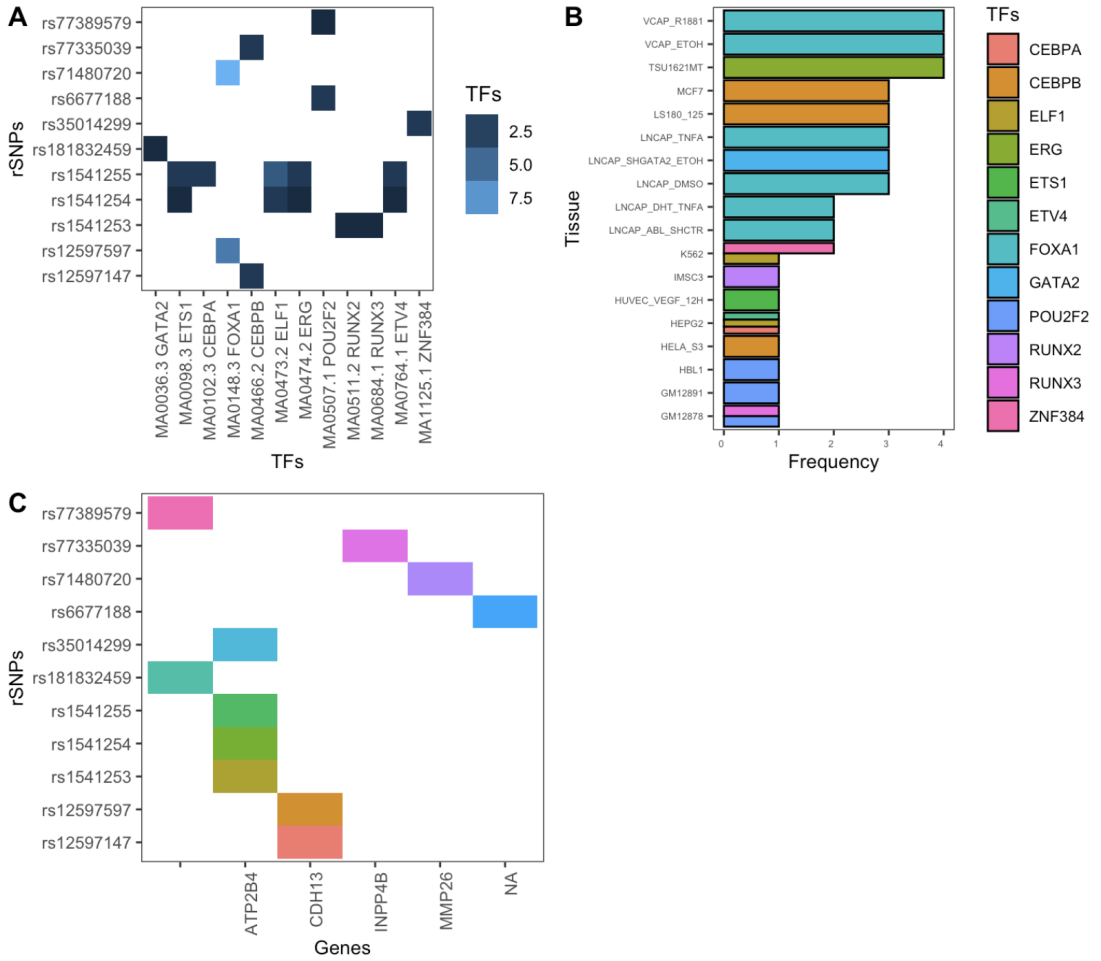


FIGURE 2.13. – Prédiction des potentiels candidats rSNPs associés au paludisme sévère. (A) rSNPs prédits et les facteurs de transcription (TFs) correspondants dont les sites de liaison à l'ADN sont affectés. (B) Types de tissus / cellules dans lesquels on a identifié les TFs (disponible dans le catalogue ReMap). (C) rSNPs et gènes associés.

Une étude expérimentale récente (ARENSBERGEN, PAGIE, FITZPATRICK et al., 2019), présentant une nouvelle technologie à haut débit (*"SuRE" : Survey of Regulation Elements*) pour l'identification des SNPs affectant l'activité d'éléments régulateurs, a permis de démontrer un effet cis-régulateur pour 30 000 SNPs (parmi une librairie de 5,9 millions de SNP testés. Les 30.000 SNPs positifs modifient l'activité d'éléments de régulation putatifs. Les SNPs rs1541253, rs1541254, rs1541255 que nous détecté associé au paludisme sévère figurent parmi ces SNPs qui exercent un effet transcriptionnel sur le gène en *cis*. Ceci confirme la pertinence de l'approche intégrative **cisreg-GWAS** pour la prédiction de SNPs régulateurs.

Par contre, nous n'avons pas retrouvé dans les rSNPs prédits le gène NCR3 (*Natural Cytotoxicity Triggering Receptor 3*), pour lequel des études individuelles ont cependant indiqué un effet de SNPs régulateurs sur la susceptibilité au paludisme simple. Ceci s'explique par le fait qu'aucune des 6 études GWAS menées pour le paludisme sévère n'a rapporté d'association avec ces SNPs de NCR3, ce qui fait

qu'ils ne se trouvaient pas dans les 350 SOIs qui ont été traités par le workflow **cisreg-GWAS**. Les rSNPs de NCR3 semblent donc affecter la susceptibilité au paludisme simple mais pas forcément celle au paludisme sévère.

2.4. Contribution à une publication sur les outils *RSAT variation-tools*

2.4.1. Article accepté moyennant modifications : *RSAT variation-tools* : an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding

L'article résumé dans ce chapitre a été soumis le 26 avril 2019 à la revue Computational and Structural Biotechnology Journal (Elsevier). Il a été accepté le 25 Septembre 2019 et publié le 07 Novembre 2019.

2.4.1.1. Motivation et état de l'art

Dans ce chapitre je présente de manière concise les résultats d'un cas d'étude sur la prédiction des SNPs régulateurs associés à la susceptibilité à *Mycobacterium tuberculosis* en utilisant les outils de *RSAT variation-tools* qui sont inclus le workflow cisreg-GWAS que j'ai développé durant ma thèse de doctorat. J'ai utilisé ce cas d'étude comme modèle d'implémentation du workflow cisreg-GWAS décrit dans le chapitre 3 de manière détaillée.

2.4.1.2. Ma contribution

Actuellement, quelques services web RSAT RESTful développés en langage Python sont disponibles [29].

J'ai contribué au développement de l'accès REST de RSAT en définissant le cahier de charges des outils *variation-scan*, en interagissant avec Thi-Thui-Nga Nguyen, la développeuse des services RSAT REST pendant le développement, en développant en R des outils logiciels clients qui se connectent au serveur pour interroger, via l'interface programmatique RESTful de RSAT, trois logiciels du groupe Variation-tools : *variation-info*, *retrieve-variation-seq*, et *variation-scan*, et je les ai intégrés dans le workflow cisreg-GWAS pour la prédiction des rSNPs.

Pour le cas d'étude présenté dans l'article joint, l'outil cisreg-GWAS collecte les SNPs associés au trait phénotypique « susceptibilité à la mesure de l'infection à *Mycobacterium tuberculosis* » (numéro d'identification de la maladie EFO_0008407) à partir de la dernière version (1.0.2) du catalogue GWAS [30]. Cette requête identifie une étude d'association (Tian, Hromatka, Kiefer et al., 2017) comportant 67 variants distincts, dont 48 portent un identifiant de référence de SNP valide (rsID) et peuvent donc être utilisées pour la suite du workflow (désignées ci-après SNP associées à une maladie ou DA-SNP).

Après la collecte des DA-SNP, il s'en suit une collecte auprès de Ensembl à partir des services web REST [31] 564 SNPs en déséquilibre de liaison (LD-SNP) dans la population Européenne, avec un seuil sur le coefficient de régression (r2 = 0,8) et une distance maximale de 200 pb. Des annotations (localisation chromosomique, type de région génomique) des 612 SNPs obtenus (48 DA + 564 LD) ont été collectées auprès de Ensembl BioMart (Durinck, Moreau, Kasprzyk et al., 2005).

L'outil sélectionne ensuite les SNPs des régions non-codants de l'ADN, donnant lieu à un ensemble de 572 SNPs d'intérêt (SOI) pour la détection de variants régulateurs. En utilisant des SNPs en LD, l'outil détermine les coordonnées des blocs d'haplotypes, qui sont ensuite filtré par chevauchement avec les régions de pics ChIP-seq en vue de calculer l'enrichissement des pics de liaison du TF en utilisant *ReMap* (Chèneby, Gheorghe, Artufel et al., 2018). Le calcul de l'enrichissement pour les annotations de maladie est fait en utilisant le paquet R *XGR*.

Enfin, cisreg-GWAS utilise les outils RSAT *variation-info* pour obtenir les coordonnées des SNPS d'intérêt, *retrieve-variation-seq* pour récupérer les séquence autour de chaque variant (de -30 à +30 pb) et *variation-scan* pour prédire l'impact de la variation sur la liaison de la TF pour chaque motif de la collection JASPAR de motifs non redondants de vertébrés (Khan, Fornes, Stigliani et al., 2018), avec un seuil de 1e-4 sur la p-valeur et de 100 sur le rapport de p-valeurs entre les variants le moins favorable et le plus favorable à la liaison.

L'analyse des SOIs par *variation-scan* avec 579 matrices de la collection JASPAR core vertebrate non redondant a permis de prédire 66 SNPs régulateurs qui modifient potentiellement 107 liaisons de TFs. Les analyses faites avec 80 millions pics de ChIP-seq du catalog *ReMap* ont retourné 4847 chevauchement des pics de ChIP-seq avec 48 régions des blocs d'haplotype dans lesquelles on a identifié 263 SNPs co-localisés avec les pics de ChIP-seq.

Seuls deux des SOIs co-localisés avec les pics de ChIP-seq montrent une correspondance entre le TF du pic de ChIP-seq et le TF associé au motif de *variation-scan*, à savoir : CEBPB pour le SNP rs3131071, ELF1 pour rs3132397. Le facteur CEBPB a été associé à la pathogénie de la tuberculose. Ce facteur est impliqué dans la différenciation et à l'activation des macrophages et dans la régulation de la réponse immunitaire et inflammatoire. Il joue aussi un rôle crucial dans les stimuli des composés immuns des IgG (Lin, Duan, Xu et al., 2017).

2.4.1.3. Conclusion

La convergence des résultats d'analyse de motifs par *variation-scan* et des données de ChIP-seq de *ReMap* a permis de prioriser deux SNPs parmi les 66 SNPs régulateurs candidats prédits. La même approche peut-être appliquée à d'autres études d'association ou pour un jeu de données de SNPs quelconques pour prédire les variations régulateurs potentiellement impliquées dans des maladies spécifiées par l'utilisateur.

2.5. Articles

Yvon Mbouamboua, Félix Koukouikila-Koussounda, Francine Ntoumi et al. (mai 2019). « Sub-microscopic Plasmodium falciparum infections in matched peripheral, placental and umbilical cord blood samples from asymptomatic Congolese women at delivery ». eng. In : *Acta Tropica* 193, p. 142-147. ISSN : 1873-6254. DOI : 10.1016/j.actatropica.2019.03.001

Acta Tropica 193 (2019) 142–147

Contents lists available at ScienceDirect

Acta Tropica

journal homepage: www.elsevier.com/locate/actatropica

Sub-microscopic *Plasmodium falciparum* infections in matched peripheral, placental and umbilical cord blood samples from asymptomatic Congolese women at delivery

Yvon Mbouamboua[a,b], Félix Koukouikila-Koussounda[a,b], Francine Ntoumi[a,b,c,*], Selorme Adukpo[c,d], Michael Kombo[a], Christevy Vouvoungui[a,b], Jacques van Helden[e], Simon Charles Kobawila[b]

[a] Fondation Congolaise pour la Recherche Médicale, Brazzaville, Congo
[b] Faculty of Sciences and Techniques, University Marien Ngouabi, Brazzaville, Congo
[c] Institute for Tropical Medicine, University of Tübingen, Tübingen, Germany
[d] Immunology Department, Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Ghana
[e] Aix-Marseille Univ, INSERM UMR_S 1090, Theory and Approaches of Genome Complexity (TAGC), F-13288 Marseille, France

ARTICLE INFO

Keywords:
Asymptomatic pregnant women
Plasmodium falciparum
Genetic diversity
Multiplicity of infections
Congo

ABSTRACT

In malaria-endemic areas, most pregnant women are susceptible to asymptomatic *Plasmodium falciparum* infections. We present here the results of a cross-sectional study conducted in Madibou, a southern district of Brazzaville in the Republic of Congo, between March 2014 and April 2015. The main aim was to characterize *P. falciparum* infections. Blood samples corresponding to peripheral, placental and cord from 370 asymptomatic malaria women at delivery were diagnosed for plasmodium infection by thick blood smears (microscopic infection). Sub-microscopic infection was detected by PCR, using the *MSP-2* gene as marker. Microscopic infections were detected in peripheral, placental and cord blood samples with a prevalence of respectively 7.3% (27/370), 2.7% (10/370) and 0%. The negative samples were submitted to sub-microscopic detection, with respective prevalence of 25.4% (87/343), 16.7% (60/360) and 9.4% (35/370) ($P < 0.001$). We further investigated the genetic diversity of the parasite by characterizing MSP2 allelic families 3D7 (24 distinct alleles) and FC27 (20 distinct alleles). The total number of alleles for these two families were 31, 25 and 19 in peripheral, placental and cord samples respectively. The 3D7 *MSP-2* was the predominant allelic family. The multiplicity of infections (MOI) in peripheral (mean 1.4 ± 0.01; range 1–4), placental (mean 1.2 ± 0.01; range 1–3) and cord samples (1.4 ± 0.01; range 1–3) were similar ($P = 0.9$) and are unaffected by age, gravidity or sulfadoxine-pyrimethamine. These results shown a high prevalence of sub-microscopic infection and a high genetic diversity of *Plasmodium falciparum* strains in Congo. Age, gravidity and doses of preventive treatment based on sulfadoxine-pyrimethamine do not interfere with the multiplicity of infections.

1. Introduction

Malaria infection during pregnancy is a serious public health issue in sub-Saharan Africa with approximately 30 million pregnant women being at risk of pregnancy-associated-malaria (Dellicour et al., 2010). The symptoms and complications of malaria infection during pregnancy may vary according to the level of malaria transmission in different geographical areas and the level of acquired immunity (Okell et al., 2009). In high-transmission areas, most infections during pregnancy are asymptomatic but still result in poor pregnancy outcome (Steketee et al., 1996; Mockenhaupt et al., 2000; Nosten et al., 2004). A high proportion of individuals living in malaria-endemic areas harbour *P. falciparum* parasites undetectable by microscopy — therefore called sub-microscopic infection — which can nevertheless be identified using molecular tools like PCR (Snounou et al., 1993). Sub-microscopic infection is common during pregnancy and at delivery (Cottrell et al., 2015). To reduce the burden of malaria infections in pregnant women, the World Health Organization (WHO) has recommended a package of

* Corresponding author at: Fondation Congolaise pour la Recherche Médcale, Villa D6, Campus WHO/AFRO, Djoué, Brazzaville, Congo.
E-mail addresses: fridyvon@gmail.com (Y. Mbouamboua), felixkoukouikila@yahoo.fr (F. Koukouikila-Koussounda), fntoumi@fcrm-congo.com (F. Ntoumi), sadukpo812@gmail.com (S. Adukpo), michael.kombo@yahoo.fr (M. Kombo), vjchristevy@gmail.com (C. Vouvoungui), Jacques.van-Helden@univ-amu.fr (J. van Helden), kobawila.simon@gmail.com (S.C. Kobawila).

https://doi.org/10.1016/j.actatropica.2019.03.001
Received 7 August 2018; Received in revised form 26 February 2019; Accepted 1 March 2019
Available online 02 March 2019
0001-706X/

interventions, including the use of long-lasting insecticidal nets (LLINs) and administration of sulfadoxine-pyrimethamine (SP) as intermittent preventive treatment of malaria during pregnancy (IPTp) (WHO, 2005). In the Republic of Congo, a malaria-endemic country in Central Africa with perennial transmission (Trape and Zoulani, 1987), the administration of IPTp-SP (2 or 3 doses) during pregnancy has been in place since 2006. The study conducted in asymptomatic pregnant women in Brazzaville revealed a reduction of the prevalence of *P. falciparum* infection (Ntoumi et al., 2016). However, high resistance to SP has been found in parasites collected from Congolese pregnant women (Koukouikila-Koussounda et al., 2015). The development of an effective vaccine is the one of the major challenges of the fight against malaria. However, this approach is limited by the polymorphic nature of Plasmodium (Mayengue et al., 2004; Ntoumi et al., 2016; Guitard et al., 2010; Cohee et al., 2016). The susceptibility of pregnant women to malaria is explained by the presence of parasite strains that specifically bind to the chondroitin sulphate A (CSA) present in the placental tissue (Fried and Duffy, 1996; Duffy and Fried, 1999). Investigating of genotype and the circulating parasites in pregnant women is interesting for the malaria vaccine (Genton et al., 2002; Takala et al., 2007; Takala and Plowe, 2009). Several previous studies on the antigenic diversity of *P. falciparum* matched in peripheral, placental and umbilical cord blood have shown varied results in different country. Some studies have shown a parasitic overlap found for each of these compartments while others have revealed a subpopulation of peripheral parasites sequestered in the placenta (Kassberger et al., 2002; Mayengue et al., 2004; Fadlelseed et al., 2017; Guitard et al., 2010; Cohee et al., 2016). Recent published work conducted in the same area reported 29 *MSP-2* alleles in isolates from peripheral blood of pregnant women with asymptomatic *P. falciparum* infection at antenatal visits, suggesting a wide diversity of the *P. falciparum* population (Ntoumi et al., 2016). Therefore, the present cross-sectional study was conducted to characterize *P. falciparum* infections matched in peripheral, placental and cord blood among Congolese women at delivery receiving 1, 2 or more doses of sulfadoxine-pyrimethamine.

2. Materials and methods

2.1. Study site

The study was carried out at Madibou Integrated Health Center, located in the semi-urban area of Makélékélé district, Southern part of Brazzaville, the capital of the Republic of Congo. This area with about 6.000 inhabitants is situated along the Congo River, where malaria transmission is intense and continuous throughout the year. The entomological inoculation rate is approximately 23 infective bites/person/year (Trape and Zoulani, 1987). *P. falciparum* is the predominant malaria species and *Anopheles gambiae s.s* the main mosquito vector.

2.2. Study population

Between March 2014 and April 2015, a total of 370 pregnant women were recruited at the maternity ward. The inclusion criteria were as follows: signed informed consent obtained from the participant at admission before delivery, no episodes of clinical malaria during the last two weeks, no notion of fever at least 48 h before enrolment and an axillary temperature $\leq$ 37.5 °C during the examination. Socio-demographic data were recorded using a structural questionnaire and from medical records.

2.3. Ethics

The study was approved by the Institutional Ethics Committee of the Fondation Congolaise pour la Recherche Médicale (ethical clearance No 001/CEI/FCRM/2012). Written informed consent was obtained from all pregnant women before enrolment.

2.4. Blood collection and treatment of samples

A total volume of 5 ml of peripheral blood was collected before delivery, 5 ml placental intervillous and umbilical cord blood were collected within 30 min after delivery into EDTA tubes for the detection of and quantification of malaria parasites. Thick blood smears were prepared for each blood compartment, stained with 10% Giemsa solution for 15 min, and examined by an experienced microscopist. Trophozoites were counted concomitantly with 200 leukocytes, and expressed in number of asexual parasites per µL of blood, assuming the leukocyte count of 8.000/µL of blood. The rest of the blood was spun to separate cells from plasma and stored separately at −20 °C for future use.

2.5. DNA extraction and characterization of P. falciparum isolates

Genomic DNA was extracted separately from peripheral, placental and cord blood using the QIAmp DNA Blood Mini kit (QIAGEN GmbH, Hilden, Germany) according to manufacturer instructions. Parasite isolates were genotyped for *P. falciparum* using the nested PCR technique, targeting the highly polymorphic locus *MSP-2* central region as previously described (Ntoumi et al., 2016). The initial amplifications were followed by individual nested PCR reactions, using specific primers for FC27 and 3D7 allelic families of the *MSP-2* gene. Allele-specific positive controls (3D7 and Dd2) and DNA-free water as a negative control were included in each set of reactions. Each PCR product was stained with Sybr Green, loaded on a 2% agarose gel, and visualized under ultraviolet trans-illumination. Various alleles were determined by their amplicon sizes and the corresponding allele-specific primers used. The sizes of PCR product were estimated using a 100 bp DNA ladder marker (Invitrogen, Karlsruhe, Germany). It was assumed that one band represented one amplified PCR fragment derived from a single copy of *P. falciparum MSP-2* gene. Alleles in each family were considered as one if fragment sizes were within a 20 bp interval. The number of different *MSP-2* fragments identified per isolate was considered as the minimum number of parasite genotypes present in this isolate. The amplified PCR products from the different blood compartments for the one woman were run in the same gel to ensure a better comparison.

2.6. Case definitions

Asymptomatic *P. falciparum* infection was defined as the presence of malaria parasites in blood in the absence of any clinical symptom. Peripheral, placental and cord blood isolates were grouped into a) *P. falciparum* microscopy positive infection; b) *P. falciparum* infection that were negative by microscopy but detected by PCR (sub-microscopic infection); c) negative individuals who were negative by both microscopy and PCR. The genetic diversity of *MSP-2* alleles was determined as the total number of different *MSP-2* fragments amplified. The multiplicity of infection (MOI) was defined as the mean number of *P. falciparum* genotypes per infected woman. Gravidity was grouped as paucigravidae (First and second full-term pregnancies), and multigravidae (3rd full term pregnancies).

2.7. Statistical analysis

All statistical analyses were performed using R (version 3.5.1) under the RStudio environment (1.1.456). To evaluate the impact of different factors (age, gravidity, SP), we applied non-parametric tests, since the multiplicity of infections takes discrete values, and does thus not comply with the normality assumption required for parametric tests. For two-group comparisons (age, gravidity) we applied Wilcoxon test. For multi-group comparisons (number of SP doses) we ran a Kruskal-

Table 1
Characteristics of the Congolese women at delivery. The age and gestational age are given as mean plus or minus standard deviation. N = number of women in each group.

Parameters	Values (*N = 370*)
Mean age ± SD (years)	25.9 ± 6.45.
Mean gestational age ± SD (weeks)	38.1 ± 1.9
Gravidity, n (%)	
Paucigravidae	152 (41%)
Multigravidae	218 (59%)
Number of IPTp-SP doses	
No	60 (16.2%)
1 dose	106 (28.6)
2 doses	95 (25.7%)
3 doses	112 (30.3%)

Wallis test. For multi-group comparisons between compartments (prevalence of infection; allelic diversity), we applied the Kruskal-Wallis test. The statistical significance threshold was set to $P < 0.05$.

3. Results

3.1. Characteristics of the study population

From March 2014 to April 2015, a total of 370 women were enrolled in the study. The characteristics of the women are shown in Table 1.

3.2. Prevalence of P. falciparum asymptomatic infection

Overall, matched peripheral, placental and cord blood samples from all study participants were analysed. The prevalence of *P. falciparum* microscopic infection in peripheral blood was 7.3% and that of placental blood was 2.7% (Table 2). Not a single sample was found microscopic positive in cord blood (Table 2). The prevalence of *P. falciparum* sub-microscopic infections was 25.4% in peripheral blood, 16.7% in placental blood and 9.4% in cord blood (Table 2). All the microscopy-positive peripheral and placental blood samples were positive by PCR. The overall (microscopy-positive and sub-microscopic infections put together) prevalence of *P. falciparum* sub-microscopic infection was 30.8% in peripheral blood and 19.4% in placental blood (Table 2).

Out of 87 (25.4%) of women with peripheral sub-microscopic infections, 9 (4.6%) had placental infections and 11 (12.6%) had infections in cord. Out of 60 (16.7%) of women with placental sub-microscopic infections, 4 (6.7%) samples were matched in cord. Furthermore, 62 (71.3%), 42 (70%) and 15 (42.8%) positive samples were detected exclusively in peripheral, placental and cord respectively. Overall, 5 cases of *P. falciparum* sub-microscopic infections matched in peripheral, placental and cord samples (Fig. 1).

3.3. Distribution of P. falciparum MSP-2 alleles

With regard to the genetic diversity, the distribution of different *MSP-2* alleles is shown in Table 3. A total, 30, 25 and 19 ($P = 0.29$) alleles were detected respectively in the peripheral, placental and cord samples. The 3D7 family was predominant whatever sample types. The mixed 3D7 and FC27 alleles were more prevalent in peripheral samples (26.3%) (Table 3).

Table 2
Prevalence of *MSP-2 P. falciparum* infections in Congolese women at delivery. N: number of women recruited in this study. For each woman, we collected three blood samples corresponding for peripheral, placenta and cord.

N = 370	Peripheral, n (%)	Placenta, n (%)	Cord, n (%)	*P*
Infections				
Microscopy	27/370 (7.9)	10/370 (2.7)	0/370 (0.0)	0.005
Sub-microscopy	87/343 (25.4)	60/360 (16.7)	35/370 (9.5)	< 0.001
Total PCR	114/370 (30.8)	70/370 (19.4)	35/370 (9.5)	< 0.001

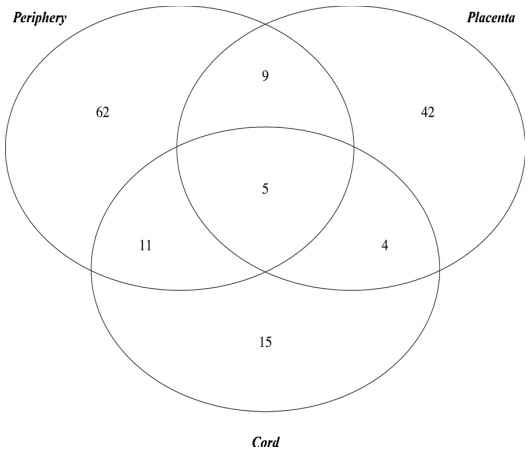


Fig. 1. Sub-microscopic *P. falciparum* infections in Congolese women at delivery matched in peripheral, placental and cord blood.

The 3D7 and FC27 alleles identified in the paired peripheral / placental, peripheral / cord and placental / cord blood samples were compared. The proportion of samples matched with identical and partially concordant alleles is shown in Fig. 2, which summarizes the number of alleles found in peripheral, placental and cord blood. Some samples revealed the same alleles in the three compartments. In other cases, one or both samples had alleles not detected in the other ones.

3.4. Multiplicity of P. falciparum infections

The MOI was 1.4 ± 0.01 (ranged 1–4 clones) in peripheral, 1.2 ± 0.01 (range 1–3 clones) in placental and 1.4 ± 0.01 (ranged 1–3 clones) in cord samples. Of 114 infected samples from peripheral blood, 110 (96.5%) contained single and 40 (35.1%) multiples infections. Of the 70 infected samples from placental blood, 46 (65.7%) harboured single and 24 (34.3%) multiples infections. In cord, 32 (91.4%) of 35 infected samples contained multiples infections and 3 (8.6%) samples have single infection. The MOI was neither affected by maternal age, nor gravidity or SP doses administered (Table 4).

4. Discussion

The current study is the first of its kind to evaluate the frequency of sub-microscopic *P. falciparum* infection among malaria asymptomatic Congolese women at delivery. The main findings of this study are the high prevalence of sub-microscopic infection at delivery, high diversity of parasites circulating in the population, and differential patterns of *MSP-2* allelic families in matched peripheral, placental and umbilical cord parasite isolates from even the same individual.

The prevalence of sub-microscopic *P. falciparum* infection in the peripheral and placental blood was higher than microscopic infection in all compartments. In the cord compartment, the *P. falciparum* infection was found only by PCR. This indicates that the threshold of detection of *P. falciparum* by microscopy in pregnant women strongly under-estimates the actual infection rates (Ntoumi et al., 2016; Mockenhaupt et al., 2000). Thus, reliance on microscopy may lead to miss diagnosis and failure to avert adverse effect of malaria during pregnancy (Mockenhaupt et al., 2000; Ntoumi et al., 2013; Ntoumi et al., 2016).

The prevalence of sub-microscopic infection was generally high at delivery. This suggests that pregnant women remained at increased risk

Table 3
Genetic diversity of *P. Falciparum MSP-2* gene. Total Nb. of alleles: total number of 3D7 and FC27 alleles of each compartment.

	Nb. of samples (%)				Fragment size range (bp)			Total Nb. of alleles (%)		
Alleles	Periphery	Placenta	Cord	*P*	Periphery	Placenta	Cord	Periphery	Placenta	Cord
3D7	62 (54.4)	42 (60)	16 (45.7)	< 0.001	180-680	140-600	120-480	16 (51.6)	14 (56)	12 (63.2)
FC27	22 (19.3)	20 (28.6)	11 (31.4)	0.14	260-620	260-540	240-520	14 (48.4)	11 (44)	7 (36.8)
3D7 + FC27	30 (26.3)	8 (11.4)	8 (22.9)	< 0.001						
Total	114 (100)	70 (100)	35 (100)	< 0.001				30 (100)	25 (100)	19 (100)

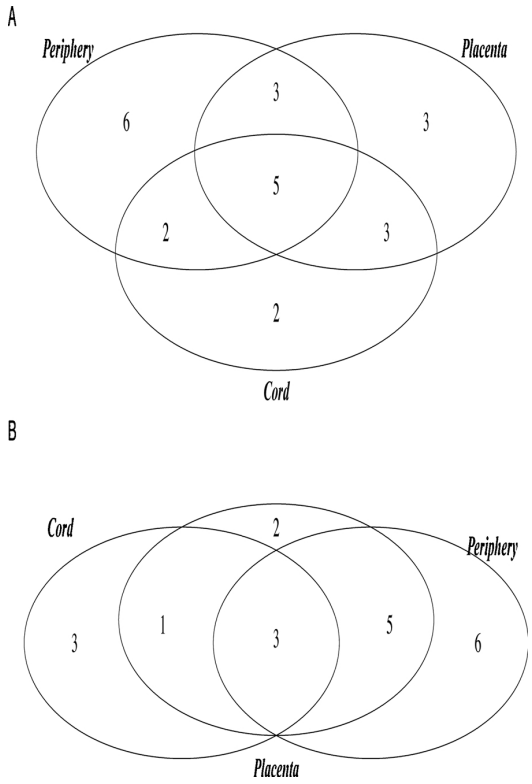


Fig. 2. comparison of *P. falciparum MSP-2* alleles detected in paired peripheral, placental and cord blood from Congolese women at delivery. (A) 3D7 family. (B) FC27 family.

of new infections during the late period of their pregnancies after receiving IPTp-SP, which is mostly administered during the second trimester. This increased danger may explain why despite the risk of infection being significantly reduced among delivered women who received 2 or 3 doses of IPTp-SP, there were still about 20% who harboured sub-microscopic infection. The proportion of sub-microscopic infection in our study is quite small compared to those reported from Gabon (Tshibola Mbuyi et al., 2014). This difference might be due to the transmission intensity of the parasite at the different study sites and possibly differential host factors such as a degree of immunity that might restrict parasites growth to microscopically undetectable levels. Alternatively, these low-density infections may represent recently acquired infections, which could reach detectable levels by microscopy if infections were not treated.

All microscopic placental blood infections were microscopically detectable in the peripheral blood. However, 42 women with sub-microscopic infection of the placental blood showed no evidence of peripheral blood infection. During pregnancy, parasites expressing a ligand for chondroitin sulphate A tend to adhere to placenta and sequester in this unique compartment (Fried and Duffy, 1996) and are rarely detected in the peripheral blood. This might have accounted for the observation made in these 42 participants. The prevalence of placental *P. falciparum* infection was lower than peripheral, which is consistent with some of the previous reports (Omalu et al., 2012; Hviid, 2011) but not with others (Mayengue et al., 2004; Ouédraogo et al., 2012). The high prevalence of infection in peripheral than placenta blood may explain the new *P. falciparum* infections in third trimester of pregnancy.

Asymptomatic *P. falciparum* infection is highly prevalent in malaria endemic regions, which raises a serious concern about the efforts at eliminating the parasite (Riley et al., 2001). Particularly, sub-patent malaria is still transmissible and will complicate the elimination of malaria in high-transmission regions (Sturrock et al., 2013). Most studies suggested that exposed individuals are likely to carry chronic asymptomatic *P. falciparum* infections (Sturrock et al., 2013; Ouédraogo et al., 2016), a situation in which the majority of this population can act

Table 4
Multiplicity of *P. falciparum* infections according to age of mother, gravidity and SP doses administered. Prevalence of infection determined by PCR, including positive peripheral, placental and/or cord samples. Multiplicity of infections = mean number of *P. falciparum* genotypes per infected woman. The columns labelled *P* indicate the P-value of a mean comparison test, based on either the Wilcoxon test or Kruskal-Wallis test depending on the number of groups. The null hypotheses are explicated in the column header.

Multiplicity of P. *falciparum* infections (range) in the different compartments						
	Periphery	*P*	Placenta	*P*	Cord	*P*
Overall	1.4 (1-4)		1.2 (1-3)		1.4 (1-3)	
Age (year). Wilcoxon test with H_o: $m (< 18) = m (\geq 18)$.						
< 18	1.4 (1-2)	0.79	1.3 (1-2)	0.78	2.0 (1-2)	0.88
≥ 18	1.4 (1-4)		1.2 (1-3)		1.4 (1-3)	
Gravidity. Wilcoxon test with H_o: *m(Paucigravidae) = m(Multigravidae)*.						
Paucigravidae	1.4 (1-3)	0.26	1.1 (1-3)	0.94	1.3 (1-3)	0.26
Multigravidae	1.3 (1-4)		1.3 (1-3)		1.4 (1-3)	
IPTp-SP. Kruskal-Wallis test with H_o: *m(no dose) = m(1 dose) = m(2 doses) = m(3 doses)*						
No	1.2 (1-2)	0.39	1.1 (1-2)	0.39	1.7 (1-3)	0.39
1 dose	1.4 (1-3)		1.2 (1-3)		1.7 (1-3)	
2 doses	1.6 (1-4)		1.1 (1-2)		1.6 (1-3)	
3 doses	1.1 (1-3)		1.3 (1-3)		1.1 (1-3)	

as a reservoir for malaria transmission.

Our results showed 15 women with sub-microscopic infection in the cord samples with no evidence of infection in peripheral and/or placental samples. The unique presence of certain parasites in the cord can be explained by parasite clearance in the placenta and / or maternal-foetal transfusions before or during birth (Riley et al., 2001).

The *MSP-2* marker is very polymorphic. We detected 24 distinct 3D7 alleles and 20 distinct FC27 alleles. Similar results were observed in Gabon (Kassberger et al., 2002; Mayengue et al., 2004) in Senegal (Schleiermacher et al., 2001). The 3D7/MSP2 allele was predominant such as reported in the previous report in the same site (Ntoumi et al., 2016). Similar results are observed in Senegal (Jafari-Guemouri et al., 2005), Sudan (Fadlelseed et al., 2017). In Gabon, the predominant allele was FC27/MSP2 (Kassberger et al., 2002; Mayengue et al., 2004). Allelic frequencies are very different from one family to another and the distribution of these alleles reveals the genetic variability present in the parasite population and might reflect the geographical variation of parasite genotype and the level of transmission.

Multi-clonal infection was detected respectively in 40 (35.1%), 24 (34.3%) and 32 (91.4%) of the peripheral, placental and cord samples. These results are the characteristic of the stable malaria transmission area (Mayengue et al., 2004).

The MOI could be used in the evaluation of malaria interventions as it is an indicator of malaria infection in humans and may, to some extent, reflect transmission and immunity malaria (Smith et al., 1999). In this cohort the MOI is lower than that reported (about 1.6 MOI) in the previous study in the same site (Ntoumi et al., 2016) and in other malaria endemic area (Beck et al., 2001; Walker-Abbey et al., 2005). Furthermore, the low MOI could be explained by the use of impregnated mosquito nets and the degree of immunity. As a result, the low parasite density of this study, which results in low MOI, can be observed, but this did not affect the diversity of the parasites in circulation (Ibara-Okabande et al., 2012).

The high prevalence of asymptomatic malaria in young pregnant women and those in their first and second pregnancies than those in their third or more pregnancies have reported (Smith et al., 1999). With successive pregnancies, women acquire a gravidity-dependent antimalarial immunity, resulting in a decrease in both prevalence and severity of infection (Smith et al., 1999). Studies have demonstrated a correlation between the MOI and parasite density as well as age of the study participants (Mayengue et al., 2004; Omer et al., 2011). In our study, MOI was associated with neither maternal age nor gravidity, which is in line with other study (Kassberger et al., 2002; Walker-Abbey et al., 2005). Age and the acquisition of pregnancy-related immunity do not appear to result in elimination of parasites or reduction of the number of different infecting strains. The SP doses have not reduced the MOI. This might result from the carriage of SP-resistant parasites in such women as suggested earlier (Moussiliou et al., 2013; Koukouikila-Koussounda et al., 2015; Ojurongbe et al., 2018).

The cross-sectional nature of the study did not allow participants to be followed until delivery. The lack of current entomological data during the collection of samples and the lack of data on the use of mosquito nets by pregnant women do not allow us to deepen the analyses on the low multiplicity of infections.

5. Conclusions

In summary, this study showed that there is a high prevalence of sub-microscopic infection and a moderate genetic diversity of *Plasmodium falciparum*. Age, gravidity and doses of preventive treatment based on sulfadoxine-pyrimethamine do not interfere with the multiplicity of infections. These results contribute to the understanding of the dynamic of parasites in blood sample types during pregnancy and parasite population circulating in Congo.

Funding

This study was financially supported by Total E&P Congo and YM received support from Campus France. YM received a financial training support from (CANTAM) RegNet 2015 funded by EDCTP and European member states. YM was funded by a grant from Campus France for his stays in Aix-Marseille Université.

Competing interests

The authors declare that they have no competing interests.

Data availability

Raw data of the work presented here are available upon request.

Acknowledgements

We are highly indebted to all women who donated blood samples for this study. We are also sincerely grateful to the staff of Madibou Integrated health center for their kind assistance. We thank Mr Benjamin Galloy, Mr Dagène Ebouroumbi and Ms Sylvia Nkombo-Nkoula for lab assistance. FN and FKK are members CANTAM (EDCTP-RegNet2015-1045) and PANDORA-ID-Net (Grant Agreement RIA2016E-1609) networks supported by EDCTP and European member states.

References

Beck, S., Mockenhaupt, F.P., Bienzle, U., et al., 2001. Multiplicity of *Plasmodium falciparum* infection in pregnancy. Am. J. Trop. Med. Hyg. 65, 631–636. https://doi.org/10.4269/ajtmh.2001.65.631.

Cohee, L.M., Kalilani-Phiri, L., Mawindo, P., et al., 2016. Parasite dynamics in the peripheral blood and the placenta during pregnancy-associated malaria infection. Malar. J. 15, 483. https://doi.org/10.1186/s12936-016-1541-x.

Cottrell, G., Moussiliou, A., Luty, A.J.F., et al., 2015. Sub-microscopic *plasmodium falciparum* infections are associated with maternal Anemia, Premature births, and low birth weight. Clin. Infect. Dis. 60, 1481–1488. https://doi.org/10.1093/cid/civ122.

Dellicour, S., Tatem, A.J., Guerra, C.A., et al., 2010. Quantifying the number of pregnancies at risk of malaria in 2007: a demographic study. PLoS Med. 7, e1000221. https://doi.org/10.1371/journal.pmed.1000221.

Duffy, P.E., Fried, M., 1999. Malaria during pregnancy: parasites, antibodies and chondroitin sulphate A. Biochem. Soc. Trans. 27, 478–482.

Fadlelseed, O.E., Osman, M.E., Shamseldin, N.M., et al., 2017. *Plasmodium falciparum* genotypes in matched peripheral, placental and umbilical cord blood in an area characterised by unstable malaria transmission in eastern Sudan. Heliyon 3, e00326. https://doi.org/10.1016/j.heliyon.2017.e00326.

Fried, M., Duffy, P.E., 1996. Adherence of *plasmodium falciparum* to chondroitin sulfate a in the human placenta. Science 272, 1502–1504. https://doi.org/10.1126/science.272.5267.1502.

Genton, B., Betuela, I., Felger, I., et al., 2002. A recombinant blood-stage malaria vaccine reduces *Plasmodium falciparum* density and exerts selective pressure on parasite populations in a phase 1-2b trial in Papua New Guinea. J. Infect. Dis. 185, 820–827. https://doi.org/10.1086/339342.

Guitard, J., Andersen, P., Ermont, C., et al., 2010. *Plasmodium falciparum* population dynamics in a cohort of pregnant women in Senegal. Malar. J. 9, 165. https://doi.org/10.1186/1475-2875-9-165.

Hviid, L., 2011. The case for PfEMP1-based vaccines to protect pregnant women against *Plasmodium falciparum* malaria. Expert Rev. Vaccines 10, 1405–1414. https://doi.org/10.1586/erv.11.113.

Ibara-Okabande, R., Koukouikila-Koussounda, F., Ndounga, M., et al., 2012. Reduction of multiplicity of infections but no change in msp2 genetic diversity in *Plasmodium falciparum* isolates from Congolese children after introduction of artemisinin-combination therapy. Malar. J. 11, 410. https://doi.org/10.1186/1475-2875-11-410.

Jafari-Guemouri, S., Ndam, N.T., Bertin, G., et al., 2005. Demonstration of a high level of parasite population homology by quantification of *Plasmodium falciparum* alleles in matched peripheral, placental, and umbilical cord blood samples. J. Clin. Microbiol. 43, 2980–2983. https://doi.org/10.1128/JCM.43.6.2980-2983.2005.

Kassberger, F., Birkenmaier, A., Khattab, A., et al., 2002. PCR typing of *Plasmodium falciparum* in matched peripheral, placental and umbilical cord blood. Parasitol. Res. 88, 1073–1079. https://doi.org/10.1007/s00436-002-0715-3.

Koukouikila-Koussounda, F., Bakoua, D., Fesser, A., et al., 2015. High prevalence of sulphadoxine–pyrimethamine resistance-associated mutations in *Plasmodium falciparum* field isolates from pregnant women in Brazzaville, Republic of Congo. Infect. Genet. Evol. 33, 32–36. https://doi.org/10.1016/j.meegid.2015.04.007.

Mayengue, P.I., Rieth, H., Khattab, A., et al., 2004. Sub-microscopic *Plasmodium falciparum* infections and multiplicity of infection in matched peripheral, placental and

umbilical cord blood samples from Gabonese women. Trop. Med. Int. Health 9, 949–958. https://doi.org/10.1111/j.1365-3156.2004.01294.x.

Mockenhaupt, F.P., Rong, B., Till, H., et al., 2000. Sub-microscopic *Plasmodium falciparum* infections in pregnancy in Ghana. Trop. Med. Int. Health 5, 167–173. https://doi.org/10.1046/j.1365-3156.2000.00532.x.

Moussiliou, A., De Tove, Y.S.-S., Doritchamou, J., et al., 2013. High rates of parasite recrudescence following intermittent preventive treatment with sulphadoxine-pyrimethamine during pregnancy in Benin. Malar. J. 12, 195. https://doi.org/10.1186/1475-2875-12-195.

Nosten, F., Rogerson, S.J., Beeson, J.G., et al., 2004. Malaria in pregnancy and the endemicity spectrum: what can we learn? Trends Parasitol. 20, 425–432. https://doi.org/10.1016/j.pt.2004.06.007.

Ntoumi, F., Vouvoungui, J.C., Ibara, R., et al., 2013. Malaria burden and case management in the Republic of Congo: limited use and application of rapid diagnostic tests results. BMC Public Health 13. https://doi.org/10.1186/1471-2458-13-135.

Ntoumi, F., Damien, B., Farnert, A., et al., 2016. Characterization of asymptomatic *Plasmodium falciparum* infection and its risk factors in pregnant women from the Republic of Congo. Acta Trop. 153, 111–115. https://doi.org/10.1016/j.actatropica.2015.10.009.

Ojurongbe, O., Nguetse, C.N., Fayemiwo, S.A., et al., 2018. High prevalence of dihydrofolate reductase gene mutations in *Plasmodium falciparum* parasites among pregnant women in Nigeria after reported use of sulfadoxine-pyrimethamine. Pathog. Glob. Health 112, 86–92. https://doi.org/10.1080/20477724.2017.1422615.

Okell, L.C., Ghani, A.C., Lyons, E., Drakeley, C.J., 2009. Sub-microscopic Infection in *Plasmodium falciparum*–Endemic Populations: A Systematic Review and Meta-Analysis. J. Infect. Dis. 200, 1509–1517. https://doi.org/10.1086/644781.

Omalu, I.C.J., Mgbemena, C., Mgbemena, A., et al., 2012. Prevalence of congenital malaria in Minna, North Central Nigeria. J. Trop. Med. 2012, 1–5. https://doi.org/10.1155/2012/274142.

Omer, S., Khalil, E., Ali, H., Sharief, A., 2011. Sub-microscopic and multiple *plasmodium falciparum* infections in pregnant Sudanese women. N. Am. J. Med. Sci. 3, 137–141. https://doi.org/10.4297/najms.2011.3137.

Ouédraogo, A., Tiono, A.B., Diarra, A., et al., 2012. Transplacental transmission of *plasmodium falciparum* in a highly malaria endemic area of Burkina Faso. J. Trop. Med. 2012, 1–7. https://doi.org/10.1155/2012/109705.

Ouédraogo, A.L., Gonçalves, B.P., Gnémé, A., et al., 2016. Dynamics of the human infectious reservoir for malaria determined by mosquito feeding assays and ultrasensitive malaria diagnosis in Burkina Faso. J. Infect. Dis. 213, 90–99. https://doi.org/10.1093/infdis/jiv370.

Riley, E.M., Wagner, G.E., Akanmori, B.D., Koram, K.A., 2001. Do maternally acquired antibodies protect infants from malaria infection? Parasite Immunol. 23, 51–59. https://doi.org/10.1046/j.1365-3024.2001.00364.

Smith, T., Beck, H.P., Kitua, A., et al., 1999. Age dependence of the multiplicity of *Plasmodium falciparum* infections and of other malariological indices in an area of high endemicity. Trans. R. Soc. Trop. Med. Hyg. 93 (Suppl 1), 15–20.

Snounou, G., Viriyakosol, S., Zhu, Xin Ping, et al., 1993. High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction. Mol. Biochem. Parasitol. 61, 315–320. https://doi.org/10.1016/0166-6851(93)90077-B.

Steketee, R.W., Wirima, J.J., Campbell, C.C., 1996. Developing effective strategies for malaria prevention programs for pregnant African women. Am. J. Trop. Med. Hyg. 55, 95–100.

Sturrock, H.J.W., Hsiang, M.S., Cohen, J.M., et al., 2013. Targeting asymptomatic malaria infections: active surveillance in control and elimination. PLoS Med. 10, e1001467. https://doi.org/10.1371/journal.pmed.1001467.

Takala, S.L., Plowe, C.V., 2009. Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria.'. Parasite Immunol. 31, 560–573. https://doi.org/10.1111/j.1365-3024.2009.01138.x.

Takala, S.L., Coulibaly, D., Thera, M.A., et al., 2007. Dynamics of polymorphism in a malaria vaccine antigen at a vaccine-testing site in Mali. PLoS Med. 4, e93. https://doi.org/10.1371/journal.pmed.0040093.

Trape, J.F., Zoulani, A., 1987. Malaria and urbanization in central Africa: the example of Brazzaville. Part II: results of entomological surveys and epidemiological analysis. Trans. R. Soc. Trop. Med. Hyg. 81 (suppl 2), 10–18.

Tshibola Mbuyi, M.L., Bouyou-Akotet, M.K., Mawili-Mboumba, D.P., 2014. Molecular detection of *plasmodium falciparum* infection in matched peripheral and placental blood samples from delivering women in Libreville, Gabon. Malar. Res. Treat. 2014, 1–6. https://doi.org/10.1155/2014/486042.

Walker-Abbey, A., Djokam, R.R., Eno, A., Leke, R.F., Titanji, V.P., Fogako, J., Sama, G., Thuita, L.H., Beardslee, E., Snounou, G., Zhou, A., Taylor, D.W., 2005. Malaria in pregnant Cameroonian women: the effect of age and gravidity on submicroscopic and mixed-species infections and multiple parasite genotypes. Am. J. Trop. Med. Hyg. 72 (3), 229–235.

WHO, 2005. World Malaria report, 2005. World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/43213/9241593199_eng.pdf;jsessionid=072CD8222AB26F16289AB555A435027D?sequence=1.

Walter Santana-Garcia, Maria Rocha-Acevedo, Lucia Ramirez-Navarro et al. (nov. 2019). « RSAT variation-tools : An accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding ». en. In : *Computational and Structural Biotechnology Journal*. ISSN : 20010370. DOI : 10.1016/j.csbj.2019.09.009. URL : https://linkinghub.elsevier.com/retrieve/pii/S2001037019301898 (visité le 02/12/2019)

Computational and Structural Biotechnology Journal 17 (2019) 1415–1428

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

RSAT variation-tools: An accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding

Walter Santana-Garcia [a,b], Maria Rocha-Acevedo [b], Lucia Ramirez-Navarro [b], Yvon Mbouamboua [c,f], Denis Thieffry [a], Morgane Thomas-Chollier [a], Bruno Contreras-Moreira [d,e], Jacques van Helden [f,g,*], Alejandra Medina-Rivera [b,*]

[a] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
[b] Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Campus Juriquilla, Blvd Juriquilla 3001, Santiago de Querétaro 76230, Mexico
[c] Fondation Congolaise pour la Recherche Médicale, Brazzaville, People's Republic of Congo
[d] Estación Experimental de Aula Dei-CSIC, Zaragoza, Spain
[e] Fundación ARAID, Zaragoza, Spain
[f] Aix-Marseille Univ, INSERM UMR S 1090, Theory and Approaches of Genome Complexity (TAGC), F-13288 Marseille, France
[g] CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Evry, France

ARTICLE INFO

Article history:
Received 27 April 2019
Received in revised form 22 September 2019
Accepted 25 September 2019
Available online 7 November 2019

Keywords:
Regulatory variants
Transcription factors
Position specific scoring matrix
SNPs
Binding motifs

ABSTRACT

Gene regulatory regions contain short and degenerated DNA binding sites recognized by transcription factors (TFBS). When TFBS harbor SNPs, the DNA binding site may be affected, thereby altering the transcriptional regulation of the target genes. Such regulatory SNPs have been implicated as causal variants in Genome-Wide Association Study (GWAS) studies. In this study, we describe improved versions of the programs *Variation-tools* designed to predict regulatory variants, and present four case studies to illustrate their usage and applications. In brief, *Variation-tools* facilitate i) obtaining variation information, ii) interconversion of variation file formats, iii) retrieval of sequences surrounding variants, and iv) calculating the change on predicted transcription factor affinity scores between alleles, using motif scanning approaches. Notably, the tools support the analysis of haplotypes. The tools are included within the well-maintained suite Regulatory Sequence Analysis Tools (RSAT, http://rsat.eu), and accessible through a web interface that currently enables analysis of five metazoa and ten plant genomes. *Variation-tools* can also be used in command-line with any locally-installed Ensembl genome. Users can input personal collections of variants and motifs, providing flexibility in the analysis.

1. Introduction

Genomic DNA sequence harbors the gene regulatory information necessary spatial and temporal gene expression patterns [38,31]. Gene regulatory regions encompass short, highly redundant DNA motifs recognized by transcription factors (TF) [36]. These regulatory regions may contain genetic variants, Single Nucleotide Polymorphisms (SNPs) or indels, that alter the DNA TF binding site (TFBS), and thereby the binding of TF [20]. Moreover, it has been reported that 93.7% of variants that have been associated with human traits or diseases have been found to be located in non-coding regions [43,40], and particularly enriched in open chromatin regions [57], indicating that these variants

Abbreviations: RSAT, Regulatory Sequence Analysis Tools; SNP, Single Nucleotide Polymorphism; TF, Transcription Factor; TFBS, Transcription Factor Binding Site; PSSM, Position Specific Scoring Matrix; MPRA, Massively Parallel Reporter Assays: MPRA; LD, Linkage Disequilibrium; rsID, Reference SNP Identifier; SOIs, SNPs of Interest; GWAS, Genome Wide Association Studies; CRM, *Cis*-Regulatory Module; eQTL, Expression Quantitative Trait Loci; ROC, Receiver Operating Characteristic; CEU, Northern Europeans from Utah.

* Corresponding authors at: Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Campus Juriquilla, Blvd Juriquilla 3001, Santiago de Querétaro 76230, México (Medina-Rivera). Aix-Marseille Univ, INSERM UMR S 1090, Theory and Approaches of Genome Complexity (TAGC), F-13288 Marseille, France (J. van Heldenf).

E-mail addresses: Jacques.van-Helden@univ-amu.fr (J. van Helden), amedina@liigh.unam.mx (A. Medina-Rivera).

https://doi.org/10.1016/j.csbj.2019.09.009

may affect transcriptional regulatory mechanisms, and thereby explain the observed phenotypes.

The Regulatory Sequence Analysis Tools (RSAT, http://rsat.eu) [47,26] has established itself in the last 20 years as a major software suite dedicated to the analysis of regulatory regions, with five public servers supporting more than 500 eukaryote and 9,000 prokaryote genomes. With a major focus on usability and accessibility to users with or without formal bioinformatics training, RSAT provides tools to retrieve sequences, perform motifs analysis, evaluate TF motif quality, compare and cluster motifs, convert file formats, etc. Here we describe *Variation-tools,* a subset of tools included in RSAT that enable users to analyse regulatory variants and assess their putative impact on TF binding sites.

1.1. Current approaches for detecting potential regulatory variants

The fact that many variants are located in non-coding regions triggered the development of bioinformatic tools to identify the regulatory potential of these genetic variants. Starting from a list of SNPs, computational analyses can help formulating hypotheses on which TF may be impacted by a genetic variant. However, there are numerous challenges for *in silico* analysis to unravel the impact of genetic variations in gene regulatory regions. Several tools and resources have been published, providing alternative methods to tackle this problem (Table 1). Most of them are either based on pattern-matching approaches to evaluate the impact of alleles on TF binding, or on machine learning models built using functional annotations of the regulatory regions, *e.g.* epigenomics and transcriptomics data. Still, these resources and tools have limitations hampering their usage in several organisms [68,28,35], on new annotated variants [5,63,54], and/or on analyses with personal collections of TF motifs [68,28,41].

All tools in the Pattern Matching category, (labeled PM in Table1) use Position-Specific Scoring Matrices (PSSMs) to evaluate the affinity of a TF to a given sequence with an allele. Major differences between these tools can be found in (i) their availability: web pages [5], command line [12] or both [69]; (ii) flexibility for the user to input their own data [61]; (iii) usability: the possibility to use several variant formats [28]; (iv) results representation: figures and/or tables [63]; (v) available organisms: only human [62], or other organisms [61]; and (vi) the possibility to calculate results on-the-fly [41] or access pre-calculated ones [5].

Another set of tools (labeled ML in Table1) aim for the identification of potential regulatory variants by integrating several types of data, beyond taking into account potential disruption of TF binding. Particularly, Lee, *et al.* [37] integrated DNaseI-seq data with SVM approaches to identify variants that could potentially disrupt TF binding. DeepSea [68] integrates functional genomic data from ChIP-seq, DNaseI-seq, RNA-seq and other functional genomic high-throughput data to assess the potential damage of variants across the human genome. Precalculated results for annotated variants can be accessed on their website.

Both tools can be trained on other organisms, provided that functional genomic data are available. The main limitation of these resources is the required expertise in bioinformatics and/or computational resources for users to analyse their own data sets. Other tools identify potential regulatory effects of a variant by comparing the measured affinity of a TF to the different possible alleles. Our tool, named *variation-scan,* falls within this category.

1.2. Variation-tools

In this context, we have developed *Variation-tools* to address the main limitations identified in existing programs (Table 1). *Variation-tools* are composed of four programs that enable (i) retrieval of information of Ensembl annotated variants when available for a given genome in RSAT (*variation-info*), (ii) conversions between variant file formats (*convert-variations*), (iii) retrieval of the sequences surrounding variants (*retrieve-variation-seq*), and (iv) scanning of different alleles of a variant with one or several motifs, comparing the scores and p-values in order to identify affected TFBS (*variation-scan*) (Fig. 1). Earlier versions of these programs were reported in 2015 as part of a RSAT update article [45], these first versions were developed in perl and were refactored and improved for the 2018 update [47]. In this article we present the latest versions of the tools, with optimized memory usage, and novel support for the inclusion of haplotype information.

In summary, RSAT *Variation-tools* provide an accessible resource for experienced and non-expert users to analyze regulatory variants in a web interface for fifteen organisms (five metazoa (http://metazoa.rsat.eu) and ten plants (http://plants.rsat.eu), with flexibility to upload personal variant and PSSM collections. We describe here *Variation-tools* methodology, along with four case studies demonstrating the flexibility of the tools, enabling the analysis of data sets from different origin (Ensembl variants, Genome-Wide Association Study (GWAS) data, ChIP-seq regions, etc.), complexity, and organisms.

2. Methods

2.1. Variation-tools: from variants to identification of regulatory effects

Variation-tools consist in a subset of four tools within RSAT devoted to the identification of genetic variants putatively affecting TF binding

1) *variation-info*: this tool relies on the Ensembl genetic variation information [29] annotated and installed on the corresponding server for each particular genome (*i.e.* human variants are installed in the Metazoa server). It can take two different inputs: 1) variant rsID or 2) genomic loci in bed format. This tool will retrieve the information of the variants matching the IDs or the information of the variants located in the genomic loci. Variants installed in RSAT servers have been processed to remove variants with incomplete annotations (no alleles) or ambiguous coordinates (non matching alleles coordinates). When users have their own variants collections, they can skip this tool and use directly *convert-variations.*
2) *convert-variations*: enables the interconversion of variant file formats such as VCF, GVF and varBed. varBed is an internal format of RSAT that facilitates the retrieval of the sequence surrounding the variant (Supplementary Fig. 1A).
3) *retrieve-variation-seq*: retrieves the sequence surrounding the variant, and produces one sequence for each allele (Supplementary Fig. 1B). The tool can take as input a varBed file (see *convert-variations*). For organisms with Ensembl annotated variants, it can take a list of IDs or a bed file listing genomic loci. The output is provided in a format named varSeq, with each row giving one allele with its surrounding sequence. Each variant has a specific internal ID to accommodate several variants with various alleles in the same file.
4) *variation-scan*: performs the scanning of alleles with a PSSM and compares the scores and p-values between alleles to assess the putative effect on TF binding (see details below) (Supplementary Fig. 2). It requires as input a varSeq file (see *retrieve-variation-seq*), a motif or collection of motifs (over twenty supported file formats), and a background model (for methodological details on background model, refer to [59] Box n°3). Different background models are read-

Table 1
Tools similar to variation-scan with available implementation. PM stands for Pattern Matching, ML stands for Machine Learning.

Name	PMID	Source	Approach	Organism	Input	Output	Matrix flexibility	Type	Last update
deltaSVM	26075791	http://www.beerlab.org/deltasvm/	Gapped k-mer SVM classifier.	Any organism	DNaseI-seq data; putative regulatory regions as positive training set and randomized sequences as negative training set.	deltaSVM, predicted impact of a variant in chromatin accessibility which is measured by adding up the contribution of all 10-mers in which the SNP is present for chromatin accessibility.	It can only be trained for one TF at a time.	ML, non-static.	Last update Sept 2015.
DeepSea	26301843	http://deepsea.princeton.edu/job/analysis/create/	Deep convolutional network.	Human	SNPs in VCF format.	Chromatin feature probabilities for reference and alternative alleles, chromatin feature probability log fold changes for each variant, chromatin feature probability differences for each variants, e-values for chromatin feature effects, functional significance score for each variant. There are 919 chromatin features evaluated.	It contains 690 TF binding profiles for 160 different TFs, but does not support the addition of new matrices.	ML, non-static.	Last update May 2017.
atSNP	26092860	https://github.com/keleslab/atSNP	Importance sampling algorithm for p-value calculation, first-order Markov Model to generate random background sequences.	Any organism whose genome is included in the Bioconductor BSGenome package.	SNP list, motif file.	p-value for binding affinity with alternative and reference allele, p-value for binding affinity change based on log-likelihood ratio and log-rank ratio. It also provides composite logo plots for directly visualizing the SNP effects on motif matches.	It accepts several matrices, and several different formats. It includes a motif library of 2,065 PSSMs from ENCODE and JASPAR, but also allows user-defined motif libraries.	PM, non-static.	Last update Nov 2018.
BayesPI-BAR	26202972	http://folk.uio.no/junbaiw/BayesPI-BAR/	Biophysical modeling of protein-DNA interaction, estimation of TF chemical potential (through a bayesian nonlinear regression model) and differential binding affinity.	Any organism	ChIP-seq experiment for TFs to be tested, DNA sequences for selected SNPs,PSSMs for selected TFs.	Given a SNP and a PSSM list, it produces two lists sorted by significance: one composed of binding motifs disrupted by the SNP, and one by sites with an increased affinity to the TF caused by the SNP.	Can use several PSSMs simultaneously.	PM, biophysical modeling. Non-static.	No updates listed, software created July 2015.
GWAS4D	29771388	http://mulinlab.tmu.edu.cn/gwas4d/gwas4d/gwas4d/gwas4d_server	Variant prioritization method, followed by an integrative analysis of genome-wide association.	Human	Accepts VCF-like, coordinate only, dbSNP ID and PLINK-like formats.	Regulatory variant prioritization table: includes the most likely affected motif by alternative variant effect.	The model includes motifs of 1,480 transcriptional regulators from 13 different resources. It is not possible to upload user-specified matrices.	PM, static	Last update Sept 2018.
sTRAP	20127973	http://trap.molgen.mpg.de/cgi-bin/home.cgi	Prediction of local binding affinity followed by a normalization of binding affinities to determine difference between reference allele and SNP.	Organisms available in TRANSFAC.	Accepts only two sequences in FASTA format.	List of TFs ranked according to changes induced by the SNP.	There is no option for user-specified matrices, matrices from TRANSFAC versions can be selected.	PM, non-static	No updates listed, software created in 2011.

(*continued on next page*)

Table 1 (*continued*)

Name	PMID	Source	Approach	Organism	Input	Output	Matrix flexibility	Type	Last update
SNP2TFBS	27899579	https://ccg.epfl.ch//snp2tfbs/	Estimation based on PSSM model.	Human.	When working with the code, the input required is the reference genome, a SNP catalogue and a PSSM collection. The web interface accepts SNP IDs and VCF format, as well as a specification of a genomic region through a bed file or by specifying the start and end positions.	List of affected TFBSs, sorted by the magnitude of the effects.	On the web interface, only matrices from JASPAR can be used. Nonetheless, it is possible to download the code used to generate the database and use a different input.	PM, static.	Last update July 2017.
atSNP Search	30534948	http://atsnp.biostat.wisc.edu/	Used atSNP algorithm with dbSNP build 144 for human genome assembly 38 against JASPAR and Encode motifs to create a repository with all the SNP-motif combinations resulting from the previous resources.	Human.	It can receive a set of rsIDs, a rsID and a window size around the SOI, genomic coordinates, a gene symbol and a window size around the gene of interest, or a TF name.	Table including p-values for motif matches for both reference and alternate alleles, as well as the change in the motif matching and the direction of said change. Output includes logo plots, displaying the sequence logos aligned to best motif matches with reference and SNP alleles.	Only JASPAR or ENCODE matrices can be selected, and it is possible to select only one transcription factor at a time.	PM, static.	Last update Jan 2018.
HaploReg	22064851, 26657631	https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php	It contains data from multiple genome annotation resources. PSSMs are scored against reference and alternative alleles, and change in log-odds is calculated.	Human	Users can provide a list of rsIDs or chromosome regions. Users can also select GWAS studies from the NHGRI catalog.	Provides data on allelic frequencies, conservation, chromatin states, and near genes. For each of the regulatory motifs altered by the SNP, it provides the change in log-odds and a logo.	HaploReg contains a library created from literature sources, TRANSFAC, JASPAR and PBM experiments. There is no option for user-specified matrices.	PM, static.	Last update November 2015.
RegulomeDB	22955989	http://www.regulomedb.org/	RegulomeDB uses information from several datasets, as well as manual curation and a heuristic method to distinguish between functional and non-functional variants.	Human.	Users can provide a list of dbSNP IDs, hg19 coordinates in BED, VCF or GFF3 format, or hg19 chromosomal regions in the same formats.	Table sorted by likely functionality, containing variant coordinates, score assigned by the algorithm, and evidence of function including protein binding, motifs, chromatin structure, eQTLs and histone modifications.	RegulomeDB includes all PSSMs from TRANSFAC, JASPAR CORE, and UniProbe. There is no option for user-specified matrices.	PM, static.	No updates, listed, software created in Sept 2012.
motifbreakR	26272984	https://github.com/Simon-Coetzee/MotifBreakR	It has three options of algorithms: the standard sum of log probabilities, weighted sum, and an information content method.	Organisms included in BSgenome.	SNPs can be imported from an R package or provided to the algorithm in BED or VCF format. PSSMs can be selected from the MotifDb package or be user-specified.	Table containing statistics describing the percent of maximum score for a matrix and matrix values for both alleles, as well as the strand. It also reports whether the TFBS is disrupted strongly or weakly.	PSSMs can be imported from the MotifDb package or be user-specified. More than one matrix can be used at a time.	PM, non-static.	Last update Jul 2018.
variation-scan		http://rsat.eu	Estimation based on PSSM model.	web interface: installed Ensembl organisms. command-line: any locally installed organism.	A collection of PSSMs and a set of variants in varSeq format. This format can be obtained using retrieve-variation-seq.	A table with one line per pair of alleles per motif (if there are more than two, there will be one line per possible pair) reporting the position, weight and p-value of each allele, weight difference and p-value ratio.	Users can select for the collections available in RSAT (JASPAR, HOCOMOCO, CisBP), but they can also use personal collections.	PM-non static.	April 2019.

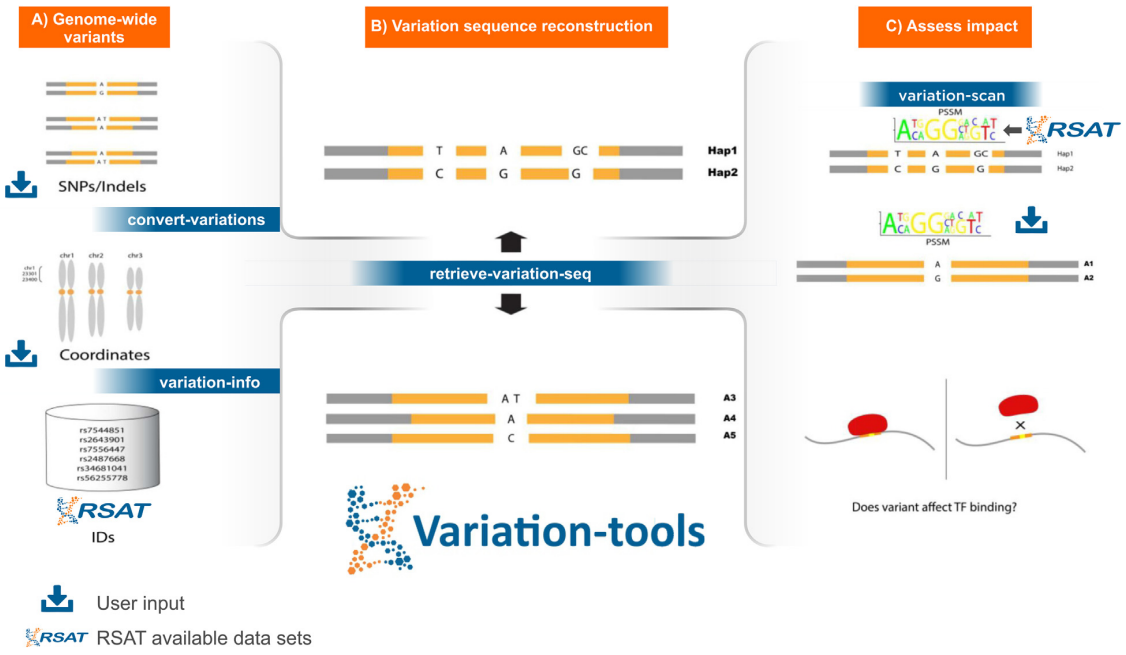


Fig. 1. Schematic representation of Variation-tools: This set of tools, included in the Regulatory Sequence Analysis Tools (RSAT), focuses on assessing the impact of different allelic variants on Transcription factor binding sites. A) *convert-variations* allows users to input their own variants and convert them to other formats (VCF, GVF and varBed, the latter is the format used in the next step), while *variation-info* retrieves the annotated information of Ensembl variants installed in RSAT servers. B) The tool *retrieve-variation-seq* retrieves the surrounding sequence of variants (including possible haplotypes) and generates a text file with one line per allele and per variant or haplotype (varSeq format). C) Users can input their variants in varSeq format and a collection of motifs (direct input by the user or selected from RSAT available collections) to *variation-scan*; the tool then scans the corresponding sequences with all motifs and perform pairwise comparisons between the binding scores of each transcription factor onto all alleles of a variant or haplotype.

ily available through the web interface. However, depending on the biological question and related potential biases, we recommend the creation of a dedicated background model, which can be done using the RSAT tool *create-background*, also available via the RSAT web interface.

2.2. Haplotype processing

Genetic variants can be detected using high-throughput techniques. This has enabled the identification of millions of variants in the HapMap [30] and 1000 genomes projects [1]. However, the information on the variants alone is less useful than knowing which groups of alleles are co-located on the same chromosome (haplotype). The process of identifying the variants that belong to each chromosome is known as phasing. Including haplotype phasing information facilitates the identification of relations between variants [6].

VCF files can include haplotype phasing information. The tool *convert-variations* identifies and retrieves the phasing information of the variants, while the tool *retrieve-variation-seq* builts the corresponding haplotype with all the SNPs that lay within a defined window (default: 30 bp).

2.3. Computing binding specificity of a transcription factor to a DNA sequence

variation-scan uses PSSMs to assess the binding specificity of a TF to a DNA sequence with different alleles in a given position. The first step of *variation-scan (i.e.,* scanning of the sequences with a given PSSM) is delegated to the RSAT tool *matrix-scan.* The scoring scheme and p-value calculation are described in detail in [59], Box n°1 and Box n°2, respectively. In brief:

PSSM are used to assess the binding specificity of a TF. This affinity is calculated as a weight score (*Ws*). The *Ws* of a site in *variation-scan* is calculated using [27]:

$$Ws = ln\left(\frac{P(S|M)}{P(S|B)}\right)$$

where S is a sequence segment of the same length of M, M is the PSSM, and B is the background model. Hence, P(S|M) is the probability of the sequence given the PSSM and P(S|B) is the probability of the sequence given the background model. *Ws* has been related to the affinity of the TF to the sequence, as it assesses similarity of a sequence to a known set of binding sites, providing information about the probability of a sequence to be a new instance of a binding site [56].

Moreover, it is possible to calculate the p-value of a given score as:

$$P - value = P(W \geq w|B)$$

where the P-value is calculated as the probability of observing a score of at least W given a background model $w|B$.

When a sequence is longer than the PSSM, the PSSM is shifted base by base until the full sequence has been scored. This scanning step is performed on the sequences of all reported alleles, so that each allele is compared with all the positions of a given motif.

Background models represent the nucleotide composition of a set of sequences (whole genome, all promoter sequences, etc.). These models are used to estimate the expectancy of a nucleotide being found. Background models can represent dependency between nucleotides in sequences (*e.g.* taking into account the fre-

quencies of dinucleotides to build a Markov model of order 1 [59] Box n°3). As background models are used to calculate weight scores for a binding site (P(S|B)), it is important to select an appropriate model for each analysis. Examples of selected background models are presented in the different study cases matching each particular biological question.

2.4. Assessment of allele effect on transcription factor binding

In the second step, *i.e.*, evaluating the impact of SNPs, *variation-scan* compares the obtained Ws (Ws difference = Ws_Allele1 – Ws_Alelle2) and the *P-value* (*P-value* ratio = P-value_Allele1/P-value_Allele2) of each of the alleles, position by position throughout the scanning window. To evaluate indels, *variation-scan* compares the highest Ws and its corresponding *P-value* for each sequence of the reported alleles. When more than two alleles of a variant are reported, all alleles are compared to all alleles in a pairwise manner.

2.5. variation-scan performance test

2.5.1. Computing efficiency

The tools *variation-info* and *convert-variation* are coded in Perl, while *retrieve-variation-seq* and *variation-scan* are coded in C, to enable the analysis of large numbers of variants from eukaryotic genomes in a reasonable time. To further improve performance, we reduced the data transfer from the hard drive to memory.

variation-scan performance was assessed by randomly selecting a variant from the 1000 genomes project [1] and a motif from the RSAT non-redundant motifs collection [8]. The randomly selected variant was used to create sets with different numbers of replicates, ranging from one thousand to nine millions, to estimate the relation between running time and the amount of evaluated variants. The processes were run on a Dell PowerEdge C6145 server with 2 AMD Opteron(tm) Processor 6386 SE, 16 cores each, Processor speed of 2.8–3.5 Ghz, RAM 256 Gb and with an operating system CentOS 7 (7.6.1810).

2.5.2. Dataset: experimentally-determined regulatory variants in red blood cells

The regulatory activity of 2,756 red blood cell variants has been systematically measured using Massively Parallel Reporter Assays (MPRA) [60]. MPRA is a high-throughput assay in which a library of putative regulatory elements, each followed by a unique barcode, is inserted into a plasmid, then transfected into a cell, and transcripts are then quantified through the abundance of barcodes. These variants are known to be in strong linkage disequilibrium (LD) with 75 variants associated with common traits of this cell type. Three sliding windows per variant (left, right, and center) were synthesized, barcoded and used to study the effect of slight changes in their genomic context. Following methods described by Ulirsch, *et al.* [60], for each sequence mRNA/DNA ratio was computed to obtain a quantitative evaluation of the regulatory effect of a sequence variant.

2.5.3. Evaluation of variation-scan

The variant dataset was used as input for *variation-scan*; the variants assessed in the red blood cell assay were annotated with the Ensembl GRCh37 human genome release, and given as input to *convert-variations* followed by *retrieve-variation-seq*. Since three sliding windows were used for each variant in the MPRA, the corresponding windows were merged before computing a background model using the *create-background-model* tool.

According to the original study [60], binding sites for the following TF were enriched in the sequences of interest: GATA1, KLF1, DHS, TAL1, ETS, FLI1 and AP-1. Therefore, a total of 48 PSSMs annotated as related to these TF were retrieved from the non-redundant RSAT motif collection [8], and given as input to *variation-scan*.

A negative control set of motifs was created using the RSAT tool *permute-matrix* [47]; five pemuted motifs were created for each of the 48 motifs, generating a collection of 240 control motifs.

For a variant to be reported in *variation-scan* as positive, we requested that at least one of the allele sequences was evaluated as a binding sites with a p-value of at most 10^{-4} (using the parameter -uth pval 1e-4 in the command line), and that the p-value ratio was greater or equal to ten (a change of one order of magnitude between the best and the worst allele p-values) (-lth pval_ratio 10).

We compared *variation-scan* to two other tools previously used to assess the same set of variants by Ulirsch, et al. [60]: DeepSea [50] and [37] deltaSVM. In order to avoid personal biases when calibrating tool parameters, we decided to rely on the published ones [60]. For this analysis *variation-scan* was run without thresholds to identify the impact of the parameters, particularly the threshold on p-value ratio.

2.6. Case studies

2.6.1. Case study 1: Identification of regulatory variants in the "Platinum" genomes haplotypes

The set of high-confidence variants from the two CEU (Northern Europeans from Utah) human Platinum Genomes NA12877 and NA12878 [23] were downloaded through the Amazon Web Service (AWS) Command Line Interface from the Illumina Platinum Genomes AWS S3 bucket (https://github.com/Illumina/PlatinumGenomes). The downloaded VCF files contained phasing information of each CEU individual haplotype configuration. The genome version used was GRCh37.

We selected SNPs intersecting with the annotated DNAseI-seq clustered peaks V3 from the ENCODE project [4]. The VCF file with the selected SNPs was processed using *convert-variations* with the option *phased* and then the haplotype sequences were reconstructed with *retrieve-variation-seq*.

For a haplotype SNP set or single position variants to be reported in *variation-scan*, we requested that at least one of the sequences was evaluated as a binding site with a p-value of at most 10^{-4} (-uth pval 1e-4) and that the p-value ratio between the two alleles was greater or equal to 100 (a change of two orders of magnitude between the best and the worst alleles p-values) (-lth pval_ratio 100). In addition, we require a change of sign between the best and worst score as an additional filter.

We annotated the predicted disrupted TFBS with the TF ChIP-seq non-redundant peak collection and with the *Cis*-Regulatory Modules (CRM) regions from ReMap [10] using bedtools intersect version 2.27 [49]. We also calculated the enrichment for annotations in the provenance sequence segments of the predicted haplotypes sites.

2.6.2. Case study 2: prediction of regulatory variants associated with susceptibility to Mycobacterium tuberculosis infection

We collected SNPs associated with the phenotypic trait "*susceptibility to Mycobacterium tuberculosis infection measurement*" (disease ID EFO_0008407) from the 1.0.2 version of the GWAS catalog [40] (https://www.ebi.ac.uk/gwas/). This query returned one study [58] with 67 distinct variants, of which 48 had a valid reference SNP identifier (rsID) and could be further used (denoted hereafter as disease-associated SNPs, or DA-SNPs). To predict the TF binding sites putatively affected by these selected SNPs, we designed an approach combining *Variation-tools* with different external resources. We further collected from Ensembl REST interface (http://rest.ensembl.org/) 564 SNPs in linkage disequilibrium

(LD-SNPs) in the European population [62], with a threshold on the regression coefficient ($r^2 \geq 0.8$) and a maximal distance of 200 bp.

Annotations (chromosomal location, type of genomic region) of the resulting 612 SNPs (48 DA + 564 LD) were collected from Ensembl BioMart [22,21]. We then restricted the selection to SNPs in non-coding regions, resulting in a set of 572 SNPs of interest (SOIs) for the detection of regulatory variants. Using SNPs in LD, we determined LD-Block regions. These were then annotated based on overlaps with ChIP-seq peaks collected from the ReMap database [10]. We also calculated enrichment for disease annotations using the R XGR package [24].

Finally, we used *retrieve-variation-seq* to retrieve the sequence variants around each SOI, and predicted the impact of the variation on TF binding for each motif of the JASPAR non-redundant RSAT motif collection [8] using *variation-scan*, with the thresholds of 1e-4 on the p-value and 100 on the p-value ratio.

2.6.3. Case study 3: Assessment of the regulatory effect of GWAS reported variants in promoters with enhancer function

The STARR-seq assay [2] is in its principle similar to the MPRA, and helps identify self-transcribing active regulatory regions that have enhancer potential. Using this approach Dao et al. [17], analysed the enhancer potential of annotated RefSeq promoters [48]. In the two cell lines K562 and HELA, they identified 632 and 493 promoters with enhancer potential (ePromoters), respectively. Moreover, the authors identified enrichment of eQTL variants reported by GTEx [25].

To identify ePromoters variants that could be affecting TF binding, we retrieved the GWAS catalog version 1.0 (downloaded on 7/01/19) [40]. Using bedtools overlap version 2.26.0 [49], we computed the overlap between SNPs and the ePromoter coordinates reported in [17]. PSSMs representing TF enriched in ePromoters were also obtained from [17], corresponding to SMRC1, JUN, FOS, ATF:MAF:NEF2, YY1, ETS family, Creb and USF1/2.

Using the selected GWAS variants that fall within ePromoters and the TF motifs enriched in these regions, we applied *variation-scan* to assess the potential regulatory effect of these variants. *variation-scan* was run with the parameters – lth w_diff 1 – lth pval_ratio 10, with a background model built using *create-background* with all RefSeq promoter sequences. In order to filtrate variants with the highest putative regulatory disruption, we further selected variants that showed a change of sign in the weight score between alleles.

2.6.4. Case study 4: identification of regulatory variants affecting VRN1 binding in barley

The latest version of *Hordeum vulgare* (barley) reference genome [42] and a panel of mapped genetic variants were imported from Ensembl Genomes release 42 [34] and installed in the RSAT Plants server (http://plants.rsat.eu). We obtained experimentally determined binding sites (ChIP-seq) for VRN1 from [19]. Since these peaks were originally positioned within contigs of the 2012 genome assembly [14], they had to be matched to the corresponding regions of the current assembly with BLAST + v2.9.0 (blastn) local alignments against the repeat-masked genome sequence (perfect matches) [7]. Using bedtools overlap version 2.26.0 [49], we selected variants falling within the VRN1 reported binding peaks. The selected variants in VCF format were then processed using *convert-variations* and *retrieve-variation-seq* to obtain the sequences with the alternative alleles.

The VRN1 DNA motif used to scan the variants was obtained from the footprintDB plant collection [16] version: 2018-06 (http://floresta.eead.csic.es/footprintdb/index.php?motif=AY750993:VRN1:EEADannot). *variation-scan* was used with a precomputed background Markov model (order 1) for barley to assess the effect of variants in TF binding, with the following parameters: – lth score 1 – lth w_diff 1 – lth pval_ratio 10 – uth pval 1e-3.

2.7. Availability

Variation-tools are available on the web (Metazoa: http://metazoa.rsat.eu/, Plants: http://plants.rsat.eu/, Teaching: http://teaching.rsat.eu/). The tools can be also installed for command-line usage with the RSAT suite (http://download.rsat.eu/).

The code and material to reproduce the results presented in the article can be accessed through GitHub (https://github.com/RSAT-doc/supp-material-publications.git).

3. Results

The *Variation-tools* provide complementary programs enabling the retrieval of variants (*variation-info*) and of their surrounding sequences (*retrieve-variation-seq*), as well as interconversion between file formats (*convert-variation*). The main predictive program is *variation-scan*, which can be used with any set of variants provided by the user (in VCF or GVF formats) or annotated in Ensembl (from a list of rsIDs or a bed file to identify overlapping variants in genome coordinates), with any set of motifs selected from the collections available in RSAT, or provided by the user.

3.1. variation-scan accurately assesses the effect of experimentally validated regulatory variants

The original version of *variation-scan* [45] required approximately five hours to assess the allele effect of nine millions variants. The novel version [47] significantly reduces the processing time to about one hour (Supplementary Fig. 3).

To evaluate the performance of *variation-scan*, we used an experimentally validated regulatory variant set obtained from a MPRA experiment [60]. For all of the assessed allele pairs, we compared the weight score differences computed with *variation-scan* with the mRNA/DNA ratio of the MPRA (see methods). As shown in Supplementary Fig. 4A, we are able to recover only 9.37% of the experimentally validated variants with *variation-scan,* as we requested at least one of the alleles to have a binding site of high confidence (p-value $\leq 10^{-4}$). Focusing on the variants reported as positive in the MPRA data set, we observed a weak correlation between the weight difference and the MPRA mRNA/DNA ratio in positive variants. However, this correlation is not significant, as MPRA values do not scale with the *variation-scan* weight differences. Nevertheless, all variants show a p-value ratio indicative of allele binding effects, showing that *variation-scan* gives accurate measurements of the impact of regulatory variants (Fig. 2A).

With the proposed thresholds, we can confidently reject 96.35% of MPRA negative sequences, which could be improved using more restrictive parameter, with a concomitant reduction in true positives. Noteworthy, as any high-throughput assay, MPRA has its limitations [51] and sequencing biases could increase the number of false negatives.

We performed a negative control, consisting of 240 permuted matrices (five permuted versions of the 48 motifs). With this collection, it was still possible to recover a group of variants, but it only represented 31.2% of the MPRA positive variants (Supplementary Fig. 4B).

We compared the performance of *variation-scan* to two other tools that had been previously used by Ulirsch, *et al* [60] to assess the same set of MPRA variants: DeepSea [68] and deltaSVM [37]. We decided to use the same parameters in order to avoid personal biases when calibrating the tools. Therefore, training weights for DNAse I hypersensitivity sites were used in the deltaSVM analysis.

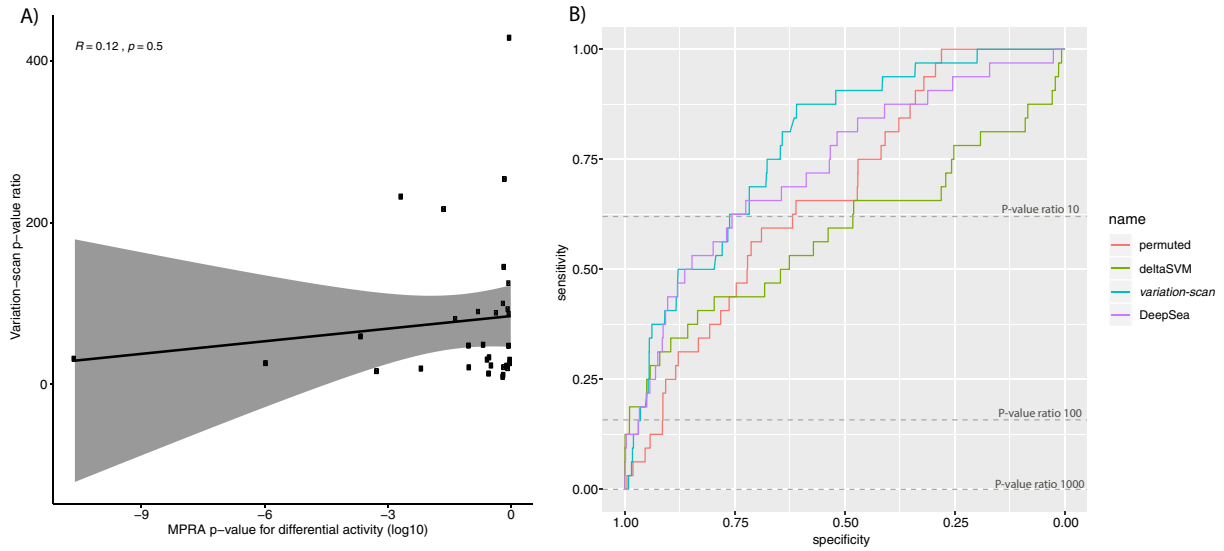


Fig. 2. Identification of experimentally validated regulatory variants using *variation-scan*. A) Correlation of the Massively Parallel Reporter Assays (MPRA) p-value of the mRNA/DNA ratio of positive variants and the *variation-scan* weight difference for the MPRA variants with significant change. B) Receiver Operating Characteristic (ROC) curve comparing the performance when aiming to classify MPRA experimentally analyzed variants using *variation-scan* (turquoise), DeepSea (purple), deltaSVM (green), and a negative control which consists of permuted motifs scored with *variation-scan* (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As for DeepSea, the web implementation of the tool was used, with the metric Functional Significance Score.

Tools were compared based on ROC curves (Fig. 2B), we additionally ran *variation-scan* using a set of permuted matrices as negative control (Fig. 2B, red line). The three tools show very similar sensitivity vs specificity at the beginning of the curves, but only *variation-scan* and DeepSea further remain separated from the negative control. As expected DeepSea performs slightly better than *variation-scan* at the beginning of the curve, nevertheless this tool requires training using epigenetic data, while *variation-scan* requires only a motif and a set of variants.

3.2. Variation-tools case studies

To illustrate the diverse applications of *Variation-tools* to tackle various biological questions, we designed four different case studies:

1. Impact of regulatory variants in the same haplotype on TF binding sites.
2. Identification of the regulatory potential of variants reported in GWAS.
3. Assessment of the regulatory potential of GWAS variants within experimentally determined regulatory regions.
4. Determination of regulatory variants within TF binding regions identified using ChIP-seq [19].

3.2.1. Genome-wide haplotype variant information can be used to identify sets of regulatory variants affecting the same TFBS

The lowering costs in sequencing have made it possible to obtain whole genome sequences of more individuals, opening the possibility of knowing, not only the variants of a genome, but also the haplotypes, and determining which variants are passed linked within the same chromosome. This enables the assessment of the regulatory effects of sets of variants within the same haplotype in a given TFBS.

Using the high-confidence SNPs from two "Platinum" Genomes [23], we determined haplotype variants that are likely to affect one TFBS. We selected variants 30bps apart, located in open chromatin, to be analysed with *variation-scan* using the non-redundant motif collection at RSAT [8]. We detected 7,406 haplotype sites with at least two heterozygous variants and a probable effect in binding of 361 TFs. Overall the number of heterozygous variants within a haplotype increases the measured weight difference. This is expected as more changes in the binding sites are more likely to change TF affinity (Fig. 3A).

To assess the biological relevance of all the putative disrupted TFBS predictions, we annotated 7,485 predicted haplotypes sites containing two or more variants with at least one heterozygous variant and 15,396 predicted sites containing a SNP (singletons) with the TF ChIP-seq peaks and the *Cis*-Regulatory Modules (CRM) regions from ReMap [10]. We found that almost all the predicted disrupted TFBS (~85%) contain a CRM or peak annotation or both (Fig. 3B). Interestingly, we found enrichment of CRM and peak annotations in the provenance sequence segments of the 7,485 predicted haplotypes sites compared to the provenance sequence segments of the single variants (Fisher exact test, p-value < 2.2e-16).

One of these annotated haplotypes is composed of the minor alleles of two SNPs (rs2732317 and rs2732318), where we observed a potential regulatory effect likely affecting three binding motifs, for EHF/ELF2, ETV4/ELK1/ETS1/FLI1/ELK4/ETS2/FEV/GABP1, and ELK3/ELF1/ERG/GABPA (Fig. 3C).

3.2.2. Genetic variants associated with Mycobacterium tuberculosis infection show potential regulatory effects

The second case study illustrates a knowledge-free use of *Variation-tools* to identify regulatory variants from GWAS studies for a user-specified disease, without prior indication about the potentially involved transcription factors or binding motifs. The approach is based on the prediction of regulatory variants with RSAT *Variation-tools,* narrowed down by selecting the regulatory SNPs that overlap ChIP-seq peaks in ReMap [10], in order to identify convergent indications for a potential impact of the variants on the binding of a TF.

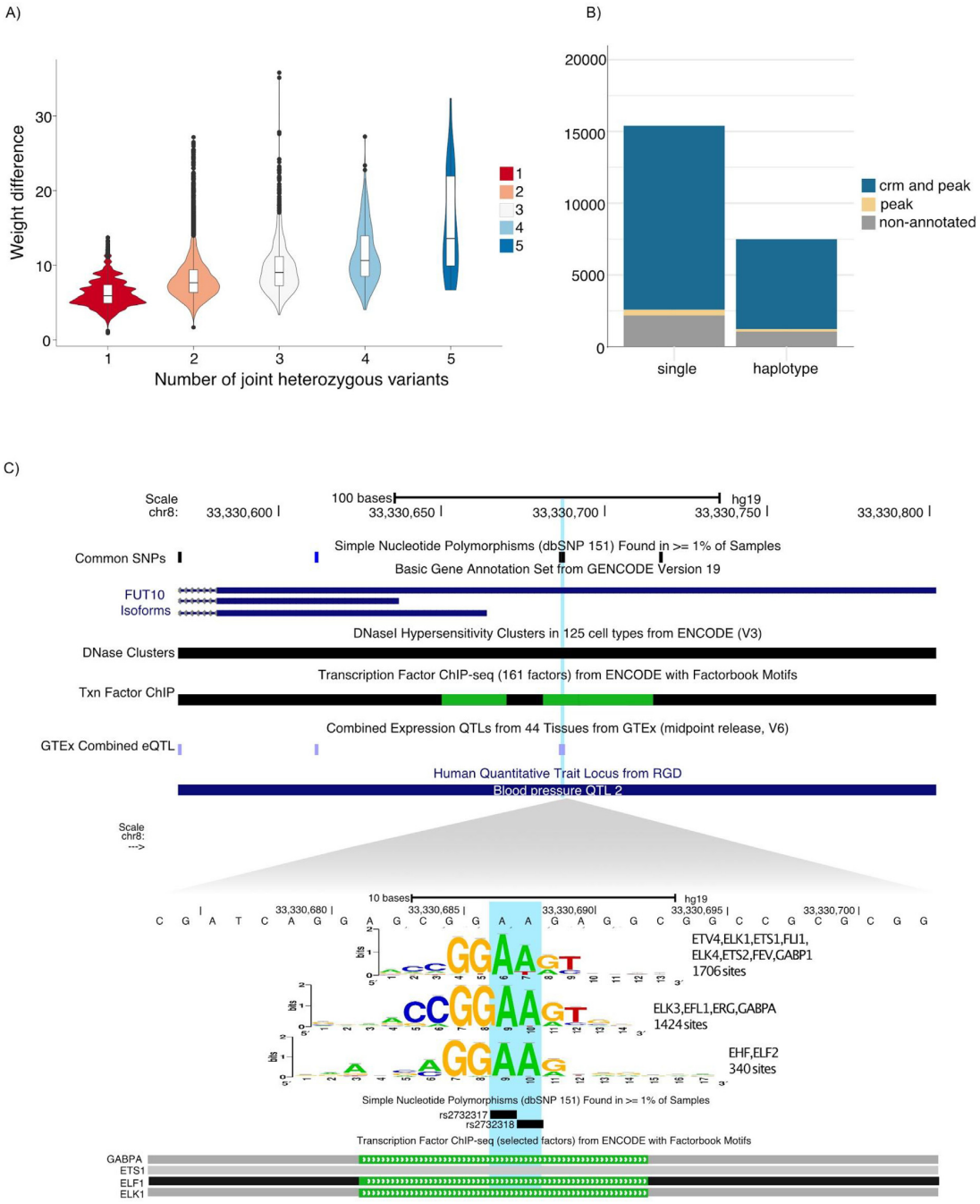


Fig. 3. Haplotype analysis in high-quality human genomes. A) The number of heterozygous variants (X-axis) within the same putative binding site tend to have a greater impact on the TF binding probability. This is expected as the increase of weight difference observed on the violin plot corresponds to the expected cumulated impact of variations affecting different positions of the same binding site. B) Number of predicted disrupted Transcription Factor Binding Sites (TFBSs) with *Cis*-Regulatory Modules (CRMs) and TF ChIP-seq peak annotation (blue), with only peak annotation (yellow), and non-annotated predictions (grey). C) University of California Santa Cruz (UCSC) browser [48] screen shot, showing a locus encompassing two SNPs that compose an heterozygous haplotype in one of the Northern Europeans from Utah (CEU) individuals. The figure shows the reference genome haplotype. The variants are located in the FUT10 promoter (top). *variation-scan* predicts an effect in three motifs that represent binding sites for GABPA, ETS1 and ELF2, factors that have been proven to have binding sites in this region by the ENCODE project. The variant rs2732317 has been associated with effects in gene expression by the GTEx project. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Interestingly, the 572 SNPs of interest (SOIs, see methods) show a significant enrichment for diseases related to respiratory functions (lung carcinoma, respiratory neoplasm, squamous cell carcinoma, lung disease), as well as for schizophrenia (Fig. 4), confirming the relevance of the collection of SNPs.

The scanning of the SOIs with the 579 matrices of Jaspar core non-redundant Vertebrate collection predicted 107 modifications of TF binding, covering 66 distinct SNPs and 181 distinct motifs. There are 4,847 overlaps between the 80 million ChIP-seq peaks of the ReMap catalog and 263 of the 572 initial SOIs, but only two of them show a match between the TF of the ChIP-seq peak and that corresponding to the motif returned by *variation-scan*: CEBPB for reference SNP rs3131071 and ELF1 for rs3132397. Noticeably, CEBPB has been reported as the main regulator for genes differentially expressed between tuberculosis patients and control cases [39]. CEBPB has also been associated with the pathogenesis of tuberculosis. This factor is involved in the differentiation and activation of macrophages and in the regulation of the immune and inflammatory response. It also plays a crucial role in the stimulation of IgG immune compounds [39]. In summary, the convergence between ChIP-seq and motif scanning results enabled the identification of two promising candidates among the 66 candidate regulatory SNPs. The same approach can be applied to other association studies in order to predict regulatory variations potentially involved in user-specified diseases.

3.2.3. Assessment of the regulatory effect of GWAS reported variants in promoters with enhancer function

ePromoters are regulatory regions with dual functions: as promoters, they regulate the gene downstream, but they also show enhancer potential according to [17]. ePromoters have been described to be enriched for eQTLs, suggesting that their function

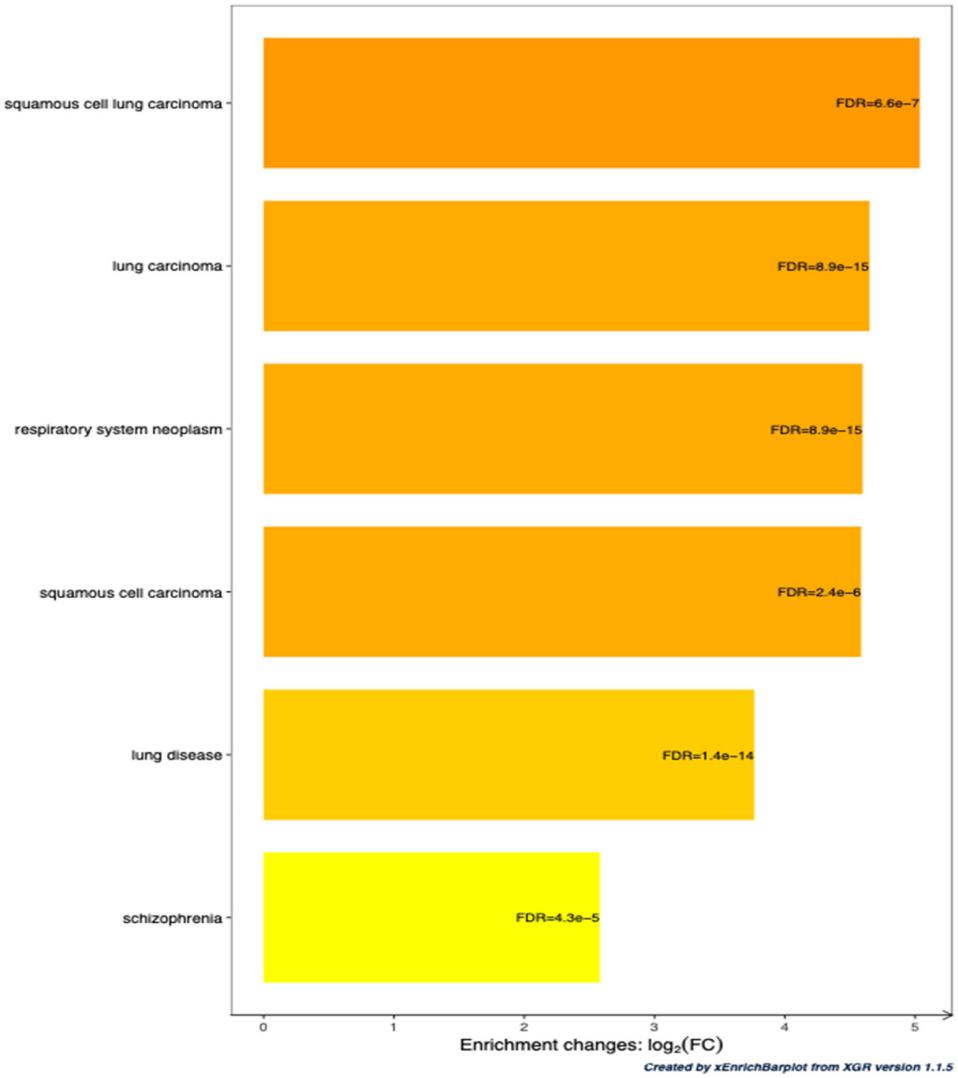


Fig. 4. Enrichment of the set of SNPs of Interest (SOIs) for diseases. The SNPs of interest includes SNPs reported by a GWAS to be associated with resistance to *Mycobacterium tuberculosis* infection and the SNPs in linkage disequilibrium with those. The genes associated with these SNPs were compared to each term of a catalogue of diseases.

could be affected by genetic variants. We thus set out to investigate if GWAS reported variants could be affecting TF binding in ePromoters.

We identified five and twelve GWAS reported variants falling within the reported coordinates of ePromoters corresponding to the human cell lines HELA and K562, respectively. Using *variation-scan* with the motifs of TF reported as enriched in ePromoters, we were able to detect two variants (rs3771180, rs3822259) affecting two TF binding motifs in HELA (MAF::NFE2, FLI1/FEV/ETS2/ELK4/ELK4/GABP1/Gabpa), and two variants (rs147997200, rs62229372) affecting three binding motifs in K562 (Creb3l2, Atf3/MAFG::NFE2L1/MAFG/NF2L1, FOS::JUN).

Noteworthy, in HELA ePromoters, we found the SNP rs3771180, which is also described as an eQTL in the whole blood dataset by the GTEx project. This variant has been associated with asthma and is upstream the Interleukin Receptor 1 gene, which is consistent with ePromoters related to inflammatory response in HELA [17].

3.2.4. Population genetic variations in barley can potentially affect VRN1 binding

Variation in traits in crops can be due to changes in TF binding that likely affect gene regulation. As a proof of concept that *Variation-tools* can be used for this purpose, we set out to identify reported variants in *Hordeum vulgare* (barley) that can putatively affect binding of VRN1, a TF involved in vernalization response. We focused this analysis on VRN1 ChIP-seq reported regions [19], selecting only variants that overlapped them (n = 1604).

Using the VRN1 motif annotated in footprintDB and the barley variants annotated in Ensembl Plants, we identified a total of thirteen variants likely to affect VRN1 binding. Of these, two are proximal to genes MLOC_73196 and MLOC_79452, which belong to a set of 38 genes known to change their expression level upon VRN1 binding in RNA-seq experiments [19].

3.3. Limitations and parameter selection

One issue arising from the analysis of big data sets is the number of false positives [44], due to the weak information content of TF binding motifs relative to genome sizes. This is one of the limitations of any bioinformatics approach to predict TF binding sites, and thus also affects the performances of *variation-scan* (Supplementary Fig. 4A and B).

Taking advantage of specific biological insights (*e.g.* identification of relevant TF, reduction of genomic regions using functional genomic information) can significantly improve results, reducing the number of false positives [9,55]. In this respect, Case Studies 3 and 4 focused on the analysis on TF known to bind on the regions of interest, which enabled us to consistently assess the performance of the tool in the evaluation set, and further helped us to identify biologically relevant regulatory variants, affecting ePromoters function in Case Study 3, or affecting VRN1 binding in barley in Case Study 4. Regarding Case Study 2, the usage of ChIP-seq information enabled the identification of potentially relevant variants related to tuberculosis.

Furthermore, the selection of adequate thresholds to select variants with *variation-scan* has an impact on reducing the number of false positives:

- P-value (-uth pval): This option refers to the upper threshold set on the p-value; this criteria has to be valid for the binding site prediction associated to at least one of the alleles; this means that at least one of the alleles allows for the prediction of a reliable binding site.
- Weight difference (-w_diff): This option determines minimal allowed weight differences between the predicted binding sites of two alleles (see methods for a description of how the weight difference is calculated).
- P-value ratio (-lth pval_ratio): This option determines the lower threshold for the p-value ratio between the predicted binding sites of two alleles (see methods for a description of how the weight difference is calculated).

Depending on the biological question, users should decide to use more or less restrictive thresholds. As shown in Case Studies 3 and 4, when the biological hypothesis is well defined, lower thresholds return manageable numbers of predictions with interesting biological insights. For more general biological questions, as in Case Studies 1 and 2, using a larger number of data, we recommend to select more stringent thresholds to reduce the number of false positives, and thereby focus the analysis on the best predictions.

Regarding the selection of a particular set of motifs, there are multiple databases installed within RSAT, which provide easy access to several reference collections (*i.e.* JASPAR, HOCOMOCO, etc.). The selection will depend on the biological question. For some TF and TF families there are structural descriptions of the protein-DNA interfaces. In some cases these structures can be used to map TF residues to particular bases within the DNA motif. Motifs from 3D-footprint [15], which are part of the footprintDB collection (http://floresta.eead.csic.es/footprintdb), allow users to further investigate the effect of variants in the light of structural information.

4. Discussion

The lowering costs of sequencing technologies has facilitated the identification of genetic variants associated with traits and diseases in humans and other species [64]. For this reason, the identification of variants affecting TF binding sites has become mainstream [3,53,32], calling for efficient computational approaches to analyse large sets of variants [18].

The case studies presented here demonstrate the application of RSAT *Variation-tools* to a diverse selection of real-world problems. Current genotype information facilitates the characterization of haplotypes, but this requires tools designed to take advantage of this information [13,52,11]. In Case Study 1, we show how the tool *convert-variations* facilitates the usage of this information, by enabling users to analyse the impact of combinations of several variants located within a 30 bp window of a chromosome. Indeed, a specific combination of variations in the same haplotype may have a synergic impact on a given TF binding site, whilst the analysis of individual variations may fail to reveal some actual regulatory impact. In the absence of information enabling TF preselection, we decided to use the complete collection of motifs. Nevertheless, by requiring more than one SNP affecting one TF binding site, we were able to identify haplotypes with potential regulatory effects. In the advent of new genome-wide characterization in population studies, this function will facilitate the integration of phasing information in the search of regulatory variants.

The identification of causal regulatory variants is a real challenge, for several reasons: (i) GWAS, which typically cover one million SNPs ("tag SNPs"), only represent a small fraction of the actual variants (150 millions currently known); (ii) the information content of a TF binding motif is relatively small, so that testing the potential impact of hundreds – or thousands – of candidate SNPs on the binding of hundreds of TF will unavoidably return an impor-

tant number of false positives. A strategy to circumvent this intrinsic limitation is to take into account TF binding regions evidenced by ChIP-seq peak experiments, in order to prioritize the predictions of regulatory variants. Case Study 2 shows that this approach ranks first SNPs highly relevant for susceptibility to *Mycobacterium tuberculosis* infection. It has to be noted that the absence of ChIP-seq peak does not preclude a predicted regulatory variant from being valid. Indeed, ChIP-seq data are only available for a subset of transcription factors, and only indicate the TF binding locations for the specific cell types or tissues in which the experiments have been performed, which may differ from those involved in the aetiology of the considered disease. The consistency between ChIP-seq peaks and predicted regulatory variants should thus be considered as a way to identify the most promising candidates rather than as a strict requirement to consider a prediction as valid.

The third case study focuses on ePromoters, defined as are regulatory regions with the capacity to act both as promoters and as enhancers, and thus the potential to affect more than one gene. Hence, regulatory variants associated with human diseases in these loci can have complex effects on gene regulation. We were able to identify four variants affecting TF binding within ePromoters in HELA and K562, ePromoters function has been linked to quick response gene expression related to inflammation and stress [46]. We were able to identify four SNPs (rs3771180, rs3822259, rs147997200 and rs62229372) putatively affecting TF binding, which are associated with traits related to inflammation and stress, supporting the relevance of ePromoters in inflammatory response.

The fourth case study takes published barley data and enlights natural variations in two regulatory regions bound by transcription factor VRN1 that are predicted to have an effect on the expression of two downstream genes. One of them is annotated as an amino acid permease (MLOC_73196), but the other one (MLOC_79452) is a protein-coding gene of unknown function. Further work would be required to confirm whether these natural variants display relevant phenotypes.

Finally, while *Variation-tools* provide a flexible framework to assess the effect of variants in TF gene regulation, there are other factors affecting regulatory mechanisms that may be taken into account, such as i) DNA accessibility [33], ii) DNA shape [66], iii) DNA methylation (Xuan [65] and iv) TF protein availability [67].

5. Conclusions

Variation-tools enables the prediction of the effect of sequence variants on TF binding. In addition to reasonable computing time, the focus is put on usability and high flexibility: annotated variants can be retrieved from specific genomic loci, as well as from personal collections of variants, motifs (provided as PSSMs) can be chosen from the collections available in RSAT (JASPAR, HOCOMOCO, CisBP, etc.), as well as from user-provided PSSM sets. The tools supports various organisms in selected RSAT servers: currently Metazoa, Plants and Teaching. In addition to the web interface, *Variation-tools* can also be used on the command line to facilitate analysis of custom data sets. *Variation-tools* can be used in combination with external databases, as exemplified with the study of GWAS data. Finally, as part of the long-lasting RSAT suite, *Variation-tools* programs are continuously maintained and updated.

Acknowledgements

We thank the persons contributing to the maintenance of the RSAT servers, in particular Laboratorio Nacional de Visualización Científica Avanzada (Mexico) specially Luis Alberto Aguilar Bautista and Jair Garcia Sotelo, the ABims platform in Roscoff, France, Pierre Vincens at the ENS, Paris and Aurora Martín Cotaina from EEAD-CSIC for her help on managing the Plants server. We acknowledge Salvatore Spicuglia for useful comments during the development of the tools. We thank Lambert Moyon and Swann Floc'hlay for providing feedback on the use of *Variation-tools*. We thank Alejandra Castillo and Carina Uribe for technical assistance. We thank Mauricio Guzman for styling the figures.

Funding

A.M.-R.'s laboratory is supported by a Consejo Nacional de Ciencia y Tecnologia (CONACYT) grant [269449]; Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica – Universidad Nacional Autónoma de México (PAPIIT-UNAM) grant [IA206517-IA201119]; M.T.-C., A.M.-R and D.T. further acknowledge SEP-CONACYT-ECOS-ANUIES [291235] support. M. T.-C. and W. S.-G. are supported by the Institut Universitaire de France. W. S.-G. benefits from a Master fellowship of the Institut de Convergences Q-life of PSL. B.C.M. was supported by Gobierno de Aragón grant A08_17R ("Genética, genómica, biotecnología y mejora de cultivos").

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2019.09.009.

References

[1] 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. A Global Reference for Human Genetic Variation. Nature 526 (7571): 68–74.
[2] Arnold, Cosmas D., Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, Alexander Stark. 2013. Genome-Wide quantitative enhancer activity maps identified by STARR-Seq. Science, March. https://doi.org/10.1126/science.1232542.
[3] Behera Vivek, Evans Perry, Face Carolyne J, Hamagami Nicole, Sankaranarayanan Laavanya, Keller Cheryl A, et al. Exploiting genetic variation to uncover rules of transcription factor binding and chromatin accessibility. Nat Commun 2018;9(1):782.
[4] Bernstein Bradley E, Birney Ewan, Dunham Ian, Green Eric D, Gunter Chris, Snyder Michael. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57–74.
[5] Boyle, Alan P., Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc a. Schaub, Maya Kasowski, Konrad J. Karczewski, et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research 22 (9): 1790–97.
[6] Browning Sharon R, Browning Brian L. Haplotype phasing: existing methods and new developments. Nat Rev Genet 2011;12(10):703–14.
[7] Camacho Christiam, Coulouris George, Avagyan Vahram, Ma Ning, Papadopoulos Jason, Bealer Kevin, et al. BLAST+: architecture and applications. BMC Bioinf 2009;10(December):421.
[8] Castro-Mondragon, Jaime Abraham, Sébastien Jaeger, Denis Thieffry, Morgane Thomas-Chollier, Jacques Van Helden. 2017. RSAT Matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res 45 (13). https://doi.org/10.1093/nar/gkx314.
[9] Chen, Chih-Yu, I-Shou Chang, Chao A. Hsiung, and Wyeth W. Wasserman. 2014. On the identification of potential regulatory variants within genome wide association candidate SNP sets. BMC Med Genom 7 (June): 34.
[10] Chèneby Jeanne, Gheorghe Marius, Artufel Marie, Mathelier Anthony, Ballester Benoit. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-Seq experiments. Nucleic Acids Res 2018;46(D1):D267–75.
[11] Choi Yongwook, Chan Agnes P, Kirkness Ewen, Telenti Amalio, Schork Nicholas J. Comparison of phasing strategies for whole human genomes. PLoS Genet 2018;14(4):e1007308.
[12] Coetzee Simon G, Coetzee Gerhard A, Hazelett Dennis J. motifbreakR: an R/bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics 2015;31(23):3847–9.
[13] International Hapmap Consortium et al. A haplotype map of the human genome. Nature 2005;437(7063):1299.
[14] The International Barley Genome Sequencing Consortium, The International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. Nature 2012. https://doi.org/10.1038/nature11543.

[15] Contreras-Moreira, Bruno. 2010. 3D-Footprint: A database for the structural analysis of protein-DNA complexes. Nucleic Acids Res 38 (Database issue): D91–97.
[16] Contreras-Moreira Bruno, Sebastian Alvaro. FootprintDB: analysis of plant cis-regulatory elements, transcription factors, and binding interfaces. Methods Mol Biol 2016;1482:259–77.
[17] Dao, Lan T. M., Ariel O. Galindo-Albarrán, Jaime A. Castro-Mondragon, Charlotte Andrieu-Soler, Alejandra Medina-Rivera, Charbel Souaid, Guillaume Charbonnier, et al. 2017. Genome-wide characterization of mammalian promoters with distal enhancer functions. Nat Genetics 49 (7): 1073–81.
[18] D'Argenio, Valeria. 2018. The high-throughput analyses era: are we ready for the data struggle? High-Throughput 7 (1). https://doi.org/10.3390/ht7010008.
[19] Deng Weiwei, Cristina Casao M, Wang Penghao, Sato Kazuhiro, Hayes Patrick M, Jean Finnegan E, et al. Direct links between the vernalization response and other key traits of cereal crops. Nat Commun 2015;6(January):5882.
[20] Deplancke Bart, Alpern Daniel, Gardeux Vincent. The genetics of transcription factor DNA binding variation. Cell 2016;166(3):538–54.
[21] Durinck Steffen, Moreau Yves, Kasprzyk Arek, Davis Sean, De Moor Bart, Brazma Alvis, et al. BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005;21 (16):3439–40.
[22] Durinck Steffen, Spellman Paul T, Birney Ewan, Huber Wolfgang. Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomaRt. Nat Protoc 2009;4(8):1184–91.
[23] Eberle Michael A, Fritzilas Epameinondas, Krusche Peter, Källberg Morten, Moore Benjamin L, Bekritsky Mitchell A, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res 2017;27 (1):157–64.
[24] Fang Hai, Knezevic Bogdan, Burnham Katie L, Knight Julian C. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. Genome Med 2016;8(1):129.
[25] GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, et al. 2017. Genetic Effects on Gene Expression across Human Tissues. Nature 550 (7675): 204–13.
[26] van Helden Jacques. Regulatory sequence analysis tools. Nucleic Acids Res 2003;31(13):3593–6.
[27] Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 1999;15(7–8):563–77.
[28] Huang Dandan, Yi Xianfu, Zhang Shijie, Zheng Zhanye, Wang Panwen, Xuan Chenghao, et al. GWAS4D: multidimensional analysis of context-specific regulatory variant for human complex diseases and traits. Nucleic Acids Res 2018;46(W1):W114–20.
[29] Hunt Sarah E, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, et al. 2018. Ensembl variation resources.Database: J Biol Databases Curat 2018 (January). https://doi.org/10.1093/database/bay119.
[30] International HapMap Consortium. The international HapMap project. Nature 2003;426(6968):789–96.
[31] Inukai Sachi, Kock Kian Hong, Bulyk Martha L. Transcription factor–DNA binding: beyond binding site motifs. Curr Opin Genet Dev 2017;43 (April):110–9.
[32] Kalita Cynthia A, Brown Christopher D, Freiman Andrew, Isherwood Jenna, Wen Xiaoquan, Pique-Regi Roger, et al. High-throughput characterization of genetic effects on DNA–protein binding and gene transcription. Genome Res 2018;28(11):1701–8.
[33] Kaplan Tommy, Li Xiao-Yong, Sabo Peter J, Thomas Sean, Stamatoyannopoulos John A, Biggin Mark D, et al. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early drosophila development. PLoS Genet 2011. https://doi.org/10.1371/journal.pgen.1001290.
[34] Kersey Paul Julian, Allen James E, Allot Alexis, Barba Matthieu, Boddu Sanjay, Bolt Bruce J, et al. Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res 2018;46(D1):D802–8.
[35] Kumar Sunil, Ambrosini Giovanna, Bucher Philipp. SNP2TFBS-a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res 2017;45(D1):D139–44.
[36] Lambert Samuel A, Jolma Arttu, Campitelli Laura F, Das Pratyush K, Yin Yimeng, Albu Mihai, et al. The human transcription factors. Cell 2018;175 (2):598–9.
[37] Lee Dongwon, Gorkin David U, Baker Maggie, Strober Benjamin J, Asoni Alessandro L, McCallion Andrew S, et al. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet 2015;47(8):955–61.
[38] Lelli Katherine M, Slattery Matthew, Mann Richard S. Disentangling the many layers of eukaryotic transcriptional regulation. Annu Rev Genet 2012;46 (August):43–68.
[39] Lin Yan, Duan Zipeng, Feng Xu, Zhang Jiayuan, Shulgina Marina V, Li Fan. Construction and analysis of the transcription factor-microRNA co-regulatory network response to mycobacterium tuberculosis: a view from the blood. Am J Transl Res 2017;9(4):1962–76.
[40] MacArthur Jacqueline, Bowler Emily, Cerezo Maria, Gil Laurent, Hall Peggy, Hastings Emma, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 2017;45(D1): D896–901.
[41] Manke Thomas, Heinig Matthias, Vingron Martin. Quantifying the effect of sequence variation on regulatory interactions. Hum Mutat 2010;31:477–83.
[42] Mascher Martin, Gundlach Heidrun, Himmelbach Axel, Beier Sebastian, Twardziok Sven O, Wicker Thomas, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature 2017;544 (7651):427–33.
[43] Maurano, Matthew T., Hao Wang, Tanya Kutyavin, John A. Stamatoyannopoulos. 2012. Widespread site-dependent buffering of human regulatory polymorphism. PLoS Genetics 8 (3): e1002599.
[44] Medina-Rivera Alejandra, Abreu-Goodger Cei, Thomas-Chollier Morgane, Salgado Heladia, Collado-Vides Julio, van Helden Jacques. Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Res 2011;39(3):808–24.
[45] Medina-Rivera Alejandra, Defrance Matthieu, Sand Olivier, Herrmann Carl, Castro-Mondragon Jaime A, Delerce Jeremy, et al. RSAT 2015: regulatory sequence analysis tools. Nucleic Acids Res 2015;43(W1):W50–6.
[46] Medina-Rivera Alejandra, Santiago-Algarra David, Puthier Denis, Spicuglia Salvatore. Widespread enhancer activity from core promoters. Trends Biochem Sci 2018;43(6):452–68.
[47] Nguyen Nga Thi, Thuy Bruno Contreras-Moreira, Castro-Mondragon Jaime A, Santana-Garcia Walter, Ossio Raul, Robles-Espinoza Carla Daniela, et al. RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res 2018;46(W1):W209–14.
[48] O'Leary Nuala A, Wright Mathew W, Rodney Brister J, Ciufo Stacy, Haddad Diana, McVeigh Rich, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44(D1):D733–45.
[49] Quinlan Aaron R, Hall Ira M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–2.
[50] Ramirez Fidel, Ryan Devon P, Gruning Bjorn, Bhardwaj Vivek, Kilpert Fabian, Richter Andreas S, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 2016;44(April):160–5.
[51] Santiago-Algarra, David, Lan T. M. Dao, Lydie Pradel, Alexandre España, Salvatore Spicuglia. 2017. Recent advances in high-throughput approaches to dissect enhancer function. F1000Research 6 (June): 939.
[52] Seo Jeong-Sun, Rhie Arang, Kim Junsoo, Lee Sangjin, Sohn Min-Hwan, Kim Chang-Uk, et al. De Novo assembly and phasing of a Korean human genome. Nature 2016;538(7624):243–7.
[53] Sewell, Jared Allan, Shaleen Shrestha, Clarissa Stephanie Santoso, Elena Forchielli, Sebastian Carrasco Pro, Melissa Martinez, and Juan Ignacio Fuxman Bass. 2018. Uncovering human transcription factor interactions associated with genetic variants, Novel DNA motifs, and repetitive elements using enhanced yeast one-hybrid assays. bioRxiv. https://doi.org/10.1101/459305.
[54] Shin Sunyoung, Hudson Rebecca, Harrison Christopher, Craven Mark, Keles Sündüz. atSNP search: a web resource for statistically evaluating influence of human genetic variation on transcription factor binding. Bioinformatics 2018 (December). https://doi.org/10.1093/bioinformatics/bty1010.
[55] Shi Wenqiang, Fornes Oriol, Mathelier Anthony, Wasserman Wyeth W. Evaluating the impact of single nucleotide variants on transcription factor binding. Nucleic Acids Res 2016;44(21):10106–16.
[56] Stormo Gary D. Modeling the specificity of protein-DNA interactions. Quantitative Biology (Beijing, China) 2013;1(2):115–30.
[57] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature 2012;488 (7414):75–82.
[58] Tian Chao, Hromatka Bethann S, Kiefer Amy K, Eriksson Nicholas, Noble Suzanne M, Tung Joyce Y, et al. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat Commun 2017;8(1):599.
[59] Turatsinze Jean-Valery, Thomas-Chollier Morgane, Defrance Matthieu, Van Helden Jacques. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc 2008;3(10):1578–88.
[60] Ulirsch Jacob C, Nandakumar Satish K, Wang Li, Giani Felix C, Zhang Xiaolan, Rogov Peter, et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell 2016;165(6):1530–45.
[61] Wang Junbai, Batmanov Kirill. BayesPI-BAR: a new biophysical model for characterization of regulatory sequence variations. Nucleic Acids Res 2015;43 (21):e147.
[62] Ward Lucas D, Kellis Manolis. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 2012;40(Database issue):D930–4.
[63] Ward Lucas D, Kellis Manolis. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res 2016;44(D1):D877–81.
[64] Wetterstrand KA. 2019. DNA Sequencing Costs: Data. Genome.gov. July 23, 2019. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.
[65] Lin Xuan, Xiao Quy, Sian Stephanie, An Omer, Thieffry Denis, Jha Sudhakar, et al. MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. Nucleic Acids Res 2019;47(D1):D145–54.
[66] Yang, Lin, Yaron Orenstein, Arttu Jolma, Yimeng Yin, Jussi Taipale, Ron Shamir, Remo Rohs. 2017. Transcription factor family-specific DNA shape readout

revealed by quantitative specificity models. Mol Syst Biol 13 (2). https://doi.org/10.15252/msb.20167238.
[67] Zabet Nicolae Radu, Adryan Boris. Estimating binding properties of transcription factors from genome-wide binding profiles. Nucleic Acids Res 2015. https://doi.org/10.1093/nar/gku1269.
[68] Zhou Jian, Troyanskaya Olga G. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015;12(10):931–4.
[69] Zuo Chandler, Shin Sunyoung, Keleş Sündüz. atSNP: transcription factor binding affinity testing for regulatory SNP detection. Bioinformatics 2015;31 (20):3353–5.

3. Discussion générale et perspectives

Le travail de ma thèse a consisté à analyser les interactions entre *P. falciparum* et son hôte humain, en combinant des méthodes expérimentales et bioinformatiques.

La première partie concernait une étude transversale de 370 femmes congolaises asymptomatiques ayant accouché au centre de santé de Madibou dans la zone sud de Brazzaville, et bénéficiant de prophylaxie anti-paludéenne intermittente à base de la sulfadoxine pyriméthamine (SP). L'analyse des infections microscopiques et submicroscopiques dans le sang périphérique, placentaire et du cordon ombilical en utilisant les techniques de la goutte épaisse et la PCR ont montré que la PCR permet de détecter des parasites du sang périphérique chez 1/4 des femmes ainsi que des parasites dans le sang placentaire dans environ 17% des échantillons. Cette prévalence dépasse largement celle des infections détectées par microscopie (7,3%, $p < 0.001$).

La microscopie optique a une faible sensibilité dans les pays à faible transmission et chez les patients asymptomatiques, ce qui entraîne une sous-estimation de la prévalence de la maladie par rapport à l'outil de diagnostic moléculaire de référence, la PCR (Okell, Ghani, Lyons et al., 2009 ; Mwingira, Genton, Kabanywanyi et al., 2014).

La microscopie ne permet pas de détecter plus d'infection dans le placenta que dans le sang circulant. En outre, la PCR a permis de détecter les parasites dans 42 placentas des femmes avec PCR négative dans le sang périphérique, ce qui confirme le fait que la sensibilité des femmes enceintes au paludisme s'explique par la présence de souches de *P. falciparum* dans le tissu placentaire (Fried et Duffy, 1996). La séquestration placentaire ne peut être correctement appréciée qu'en utilisant des techniques de détection moléculaires comme la PCR. Je souhaite rappeler l'importance de la PCR en tant qu'un des outils de détection moléculaire (Fried, Muehlenbachs et Duffy, 2012) dans la surveillance du paludisme car il permet de mettre en évidence le changement de profil (sensible ou résistant) des souches parasitaires avant même qu'il soit visible par l'échec thérapeutique. Cependant, en clinique, on continue de pratiquer essentiellement la goutte épaisse et le test de diagnostic rapide, car la PCR est très sensible mais reste trop complexe pour une mise en œuvre sur le terrain en raison de la nécessité de conditions de laboratoire sophistiquées, d'une formation avancée et d'un délai relativement long avant l'obtention des résultats.

Le traitement préventif intermittent à base de la SP n'a d'effet ni sur les infections asymptomatiques ni sur la multiplicité des génotypes de parasites hébergés par les femmes enceintes. Une étude menée par Koukouikila-Koussounda et al. (2015) a montré que la SP devient inefficace dans la zone géographique du sud de Brazzaville probablement du fait de la présence de souches de résistance de *P.*

falciparum (Koukouikila-Koussounda, Bakoua, Fesser et al., 2015). Ces résultats semblent montrer que le TPI-SP dans cette zone de Brazzaville n'est guère bénéfique pour la communauté car les femmes enceintes servent de réservoir de parasites résistants.

Cette étude expérimentale, qui contribue à la compréhension de la dynamique de la population parasitaire en circulation en République du Congo, portait sur un nombre d'échantillons limités à 370. Pour compléter cette étude, il serait utile de mener des études longitudinales pendant la grossesse et d'élargir l'échantillonnage à l'ensemble du territoire congolais. L'équipe de Francine Ntoumi poursuit cette études sur les gènes de résistance *dhfr (dihydrofolate reductase)* et *dhps (dihydropteroate synthase)* à la SP sont en cours.

La seconde partie de cette étude était axée sur l'approche bioinformatique d'analyse des variants régulateurs (rSNPs) associés au paludisme sévère m'a amené à développer le workflow cisreg-GWAS, qui combine différents outils d'analyse (*RSAT Variation-tools*) et bases de données (*GWAS catalog, dbSNP, Ensembl, ReMap*) pour la prédiction de SNPs qui ont un impact sur la régulation transcriptionnelle (candidats rSNPs). L'outil est conçu de façon générique, et peut s'adapter pour l'étude des rSNPs de n'importe quel trait phénotypique documenté dans *GWAS catalog*. Il sera prochainement adapté pour permettre également à un utilisateur d'entrer une liste de SNPs de son choix (en court-circuitant la collecte des SNPs de *GWAS catalog*).

Afin de faciliter l'interprétation des résultats, l'outil génère automatiquement un rapport d'analyse illustré par des figures et tableaux, à partir d'un modèle en format R markdown. Je collabore avec Ferran Moratalla Navaro de *Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology (ICO), L'Hospitalet del Llobregat, 08908 Barcelona, Spain* pour développer une interface Web conviviale avec Shiny (librairie d'interfaçage Web de R), qui permettra de déployer le workflow sur un serveur Web, afin de le rendre accessible à des biologistes qui ne maîtrisent pas le langage statistique R.

Le workflow a été testé sur une série de cas d'étude qui ont permis de vérifier sa pertinence et sa ré-utilisabilité : malaria, susceptibilité à la tuberculose, sensibilité au sepsis, diabète de type II.

Le cas d'étude des SNPs associés à la susceptibilité à *Mycobacterium tuberculosis* a fait l'objet d'une contribution dans un article sur les outils *RSAT Variation-tools*. L'analyse des SNPs de la tuberculose a révélé que le SNP rs3131071 modifie le site de liaison du facteur CEBPB. En effet, Le facteur CEBPB a été associé à la pathogénie de la tuberculose. Ce facteur est impliqué dans la différenciation et à l'activation des macrophages et dans la régulation de la réponse immunitaire et inflammatoire. Il joue aussi un rôle crucial dans les stimuli des composés immuns des IgG (Lin, Duan, Xu et al., 2017).

L'analyse de 375 SNPs associés au paludisme sévère des études de GWAS par le workflow a révélé 344 SNPs avec prédiction d'une modification pour la liaison d'un facteur transcriptionnel (avec RSAT variation-scan), 169 SNPs chevauchant des pics ReMap, parmi lesquels 11 SNPs régulateurs potentiels, avec une cohérence de TFs entre les prédictions de RSAT et les pics de ReMap. Ces 11 rSNPs candidats sont associés à 3 gènes (ATP2B4, CDH13, INPP4B) et les variations modifieraient les sites de liaison de 12 TFs (GATA2, ETS1, CEBPA, CEBPB, FOXA1, ELF1,

ERG, POU2F2, RUNX2, RUNX3, ETV4, ZNF384).

Il faut noter que les prédictions de variants régulateurs par RSAT peuvent éventuellement être ou ne pas être consistantes avec les pics de ChIP-seq (ReMap) mais si ce n'est pas le cas ça ne veut pas forcément dire qu'ils sont faux pour plusieurs raisons : i) Il y a des tas de facteurs pour lesquels on n'a pas fait d'analyse ChIP-seq. ii) Même si on a fait des analyses CHIP-seq, ce n'est pas forcément dans le tissu pertinent. La plupart des résultats de ReMap proviennent d'ENCODE, qui repose sur des lignées cellulaires. iii) Même si on a fait une étude ChIP-seq avec le bon facteur et dans le bon tissu, il se peut que les pics ne couvrent pas l'ensemble des sites de liaison du TF. La présence de pics de ChIP-seq constitue donc une confirmation de la pertinence d'une prédiction de rSNPs, mais leur absence laisse la porte ouverte à d'éventuelles validations par d'autre méthodes.

Parmi les rSNPs prédits pour la susceptibilité au paludisme, on notera en particulier 3 SNPs (rs1541253, rs1541254, rs1541255) associés au gène ATP2B4 (qui code pour l'enzyme ATPase 4 transportant le calcium dans la membrane plasmique), sont en cours de validation expérimentale pour leur impact sur l'activité du promoteur (analyse menée par la doctorante Samia NIZAR de l'équipe de Pascal RIHET). Je collabore aussi avec Florian ROSIER, un autre doctorant de l'équipe de Pascal RIHET pour tester son jeu de données des SNPs associés au sepsis.

Une étude publiée pendant la rédaction de ma thèse (Arensbergen, Pagie, FitzPatrick et al., 2019), présente une nouvelle technologie à haut débit pour l'identification des SNPs affectant l'activité d'éléments régulateurs. Cette technologie, dénommée SuRE (*Survey of Regulation Elements*) a été appliquée à un ensemble de 5,9 millions de SNPs humains. Cette étude a démontré un effet cis-régulateur de 30 000 SNPs qui modifient l'activité d'éléments de régulation putatifs. A ma connaissance, c'est la première méthode à haut débit permettant de tester de façon directe l'effet régulateur d'un SNP. En effet, les autres méthodes à haut débit pour l'analyse de la régulation fournissent soit des indications générales sur le niveau de transcription (biopuces à expression, RNA-seq) soit sur la liaison de facteurs transcriptionnels à l'ADN (ChIP-chip, ChIP-seq) mais n'apportent pas d'informations concernant l'impact potentiel de variants sur la liaison de ces TFs ou sur l'expression des gènes-cibles. De façon intéressante, les SNPs rs1541253, rs1541254, rs1541255 identifiés par mon workflow figurent parmi ceux détectés par SuRE.

L'incorporation des résultats de ce type d'études apporterait un élément complémentaire à mon workflow, qui se concentre pour l'heure sur les effets trans-régulateurs de l'interaction entre le facteur transcriptionnel et ses sites de liaison (impact des variations sur la liaison de TF, coïncidence entre les SNPs et les pics de liaison de TF). J'inclurai dans une prochaine version du workflow les études systématiques d'effets cis-régulateurs des variations.

Le nombre de SNPs d'intérêt est un facteur limitant sur l'efficacité actuelle du workflow. Quand on analyse un trait phénotypique avec un grand nombre de SNPs associés (DA) ou en déséquilibre de liaison (LD), certains des services Web invoqués à distance commencent à poser des problèmes. C'est le cas par exemple du diabète de type 2 où on a 1.153 DA SNPs et ~18.000 LD SNPs. Les problèmes peuvent provenir soit d'une limitation de l'interface programmatique (par exemple, l'interface REST d'Ensembl ne permet d'obtenir les SNPs en DL que pour un SNP

à la fois) soit du temps de calcul requis pour effectuer des analyses (par exemple le scanning de tous les SNPs d'intérêt avec tous les motifs JASPAR avec RSAT variation-scan).

Plusieurs possibilités sont envisageables pour réduire fortement la charge de ces serveurs quand le nombre de SNPs est excessif. Parmi elles :

— Restreindre l'analyse de motifs (RSAT variation tools) aux SNPs qui sont chevauchés par au moins un pic de ReMap. Ceci occasionnera bien entendu une baisse de sensibilité pour la partie motifs, mais permettra dans certains cas de gagner un temps de calcul significatif. Et si l'utilisateur désire uniquement retenir les candidats croisés entrée ReMap et RSAT variation tools, cela ne présentera pas d'intérêt pour lui d'analyser les SNPs qui ne chevauchent pas les pics de ReMap. Ceci doit cependant rester une option car un utilisateur peut très bien désirer analyser tous les SNPs avec RSAT variation-scan (par exemple parce qu'il travaille sur des types tissulaires qui n'ont pas été très bien couverts dans les études ChIP-seq, et donc il s'attend à ce qu'une bonne partie des variations régulatrices ne soient pas couvertes par des pics dans l'état actuel des données ChIP-seq).

— Restreindre l'analyse de motifs (RSAT variation tools) aux SNPs connus pour avoir un effet régulateur (données supplémentaires de l'article publié en 2019 par de Joris van Arensbergen). Ici aussi, c'est une option à double tranchant, car cette étude n'a porté que sur les variations de 4 individus de 4 populations différentes (qui n'incluent pas de "Caucasien"), et n'échantillonnent que 5,9 millions de SNPs, alors qu'Ensembl en contient aujourd'hui'hui $>$ 150 millions. De plus l'étude de Joris van Arensbergen a été menée sur 2 lignées cellulaires (K562 et HedG2) correspondant respectivement à des leucocytes et des cellules de foie, et les 30.000 SNPs régulateurs qu'ils détectent ne couvrent donc vraisemblablement qu'un sous-ensemble des SNPs régulateurs qu'on peut trouver dans l'ensemble des tissus humains.

Ces deux stratégies sont donc des heuristiques qui permettent d'accélérer les temps de calcul, de diminuer les taux de faux-positifs (en réduisant la multiplicité des hypothèses à tester) mais elles s'accompagneront forcément d'une perte de sensibilité.

Une approche plus intéressante serait de paralléliser la détection de SNPs régulateurs sur les serveurs RSAT. Ceci fait partie des projets de l'équipe RSAT, et permettrait un gain de temps important pour les requêtes adressées via les interfaces Web ou l'API REST.

L'application de mon workflow aux différents cas d'étude (paludisme sévère, tuberculose, sepsis et diabète de type 2) donne dans chaque cas des résultats pertinents. Cependant, certains des paramètres ont été fixés de façon empirique, soit en se basant sur les habitudes du domaine (par exemple : coefficient $r^2 \geq 0.8$ pour la collecte des SNPs LD, seuils sur la p-valeur et le poids pour RSAT), soit en restreignant le nombre de données pour éviter une augmentation excessive des SNPs à analyser (par exemple : fenêtre de 200 pb pour la collecte des SNPs LD). Un étude systématique de l'impact de ces paramètres pourrait contribuer à augmenter la fiabilité des prédictions. Une telle étude est envisagée en vue de la prochaine publication d'un article sur le workflow cisreg-GWAS.

Bibliographie

Abba, Katharine, Jonathan J Deeks, Piero L Olliaro et al. (juil. 2011). « Rapid diagnostic tests for diagnosing uncomplicated *P. falciparum* malaria in endemic countries ». en. In : *Cochrane Database of Systematic Reviews*. Sous la dir. de Cochrane Infectious Diseases Group. issn : 14651858. doi : 10.1002/14651858.CD008122.pub2. url : http://doi.wiley.com/10.1002/14651858.CD008122.pub2 (visité le 29/08/2019) (cf. p. 41).

Abyzov, Alexej, Jessica Mariani, Dean Palejev et al. (déc. 2012). « Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells ». eng. In : *Nature* 492.7429, p. 438-442. issn : 1476-4687. doi : 10.1038/nature11629 (cf. p. 52).

Adams, Yvonne, Pongsak Kuhnrae, Matthew K. Higgins et al. (mar. 2014). « Rosetting Plasmodium falciparum-infected erythrocytes bind to human brain microvascular endothelial cells in vitro, demonstrating a dual adhesion phenotype mediated by distinct P. falciparum erythrocyte membrane protein 1 domains ». eng. In : *Infection and Immunity* 82.3, p. 949-959. issn : 1098-5522. doi : 10.1128/IAI.01233-13 (cf. p. 65).

Albert, Frank W. et Leonid Kruglyak (avr. 2015). « The role of regulatory variation in complex traits and disease ». eng. In : *Nature Reviews. Genetics* 16.4, p. 197-212. issn : 1471-0064. doi : 10.1038/nrg3891 (cf. p. 60).

Amoah, L. E., S. V. Nuvor, E. K. Obboh et al. (déc. 2017). « Natural antibody responses to Plasmodium falciparum MSP3 and GLURP(R0) antigens are associated with low parasite densities in malaria patients living in the Central Region of Ghana ». en. In : *Parasites & Vectors* 10.1. issn : 1756-3305. doi : 10.1186/s13071-017-2338-7. url : http://parasitesandvectors.biomedcentral.com/articles/10.1186/s13071-017-2338-7 (visité le 15/06/2019) (cf. p. 33).

Andersen, Malin C., Pär G. Engström, Stuart Lithwick et al. (jan. 2008). « In silico detection of sequence variations modifying transcriptional regulation ». eng. In : *PLoS computational biology* 4.1, e5. issn : 1553-7358. doi : 10.1371/journal.pcbi.0040005 (cf. p. 86).

Arensbergen, Joris van, Vincent D. FitzPatrick, Marcel de Haas et al. (fév. 2017). « Genome-wide mapping of autonomous promoter activity in human cells ». In : *Nature biotechnology* 35.2, p. 145-153. issn : 1087-0156. doi : 10.1038/nbt.3754. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5498152/ (visité le 09/09/2019) (cf. p. 57).

Arensbergen, Joris van, Ludo Pagie, Vincent D. FitzPatrick et al. (juil. 2019). « High-throughput identification of human SNPs affecting regulatory element activity ». eng. In : *Nature Genetics* 51.7, p. 1160-1169. issn : 1546-1718. doi : 10.1038/s41588-019-0455-2 (cf. p. 126, 154).

Argy, Nicolas et Sandrine Houzé (mar. 2018). « Épidémiologie et cycle parasitaire d'un fléau mondial, le paludisme ». fr. In : *Actualités Pharmaceutiques* 57.574, p. 18-20. issn : 05153700. doi : 10.1016/j.actpha.2018.01.004. url : https://linkinghub.elsevier.com/retrieve/pii/S0515370018300041 (visité le 28/05/2019) (cf. p. 28, 29).

Baaklini, Sabrina, Sarwat Afridi, Thy Ngoc Nguyen et al. (nov. 2017). « Beyond genome-wide scan : Association of a cis-regulatory NCR3 variant with mild malaria in a population living in the Republic of Congo ». en. In : *PLOS ONE* 12.11. Sous la dir. de Claudio Romero Farias Marinho, e0187818. issn : 1932-6203. doi : 10.1371/journal.pone.0187818. url : https://dx.plos.org/10.1371/journal.pone.0187818 (visité le 14/06/2019) (cf. p. 63).

Band, Gavin, Quang Si Le, Luke Jostins et al. (mai 2013). « Imputation-based meta-analysis of severe malaria in three African populations ». eng. In : *PLoS genetics* 9.5, e1003509. issn : 1553-7404. doi : 10.1371/journal.pgen.1003509 (cf. p. 65, 66).

Banerji, J., S. Rusconi et W. Schaffner (déc. 1981). « Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences ». eng. In : *Cell* 27.2 Pt 1, p. 299-308. issn : 0092-8674. doi : 10.1016/0092-8674(81)90413-x (cf. p. 72).

Barber, Bridget E., Timothy William, Matthew J. Grigg et al. (avr. 2013). « Evaluation of the sensitivity of a pLDH-based and an aldolase-based rapid diagnostic test for diagnosis of uncomplicated and severe malaria caused by PCR-confirmed Plasmodium knowlesi, Plasmodium falciparum, and Plasmodium vivax ». eng. In : *Journal of Clinical Microbiology* 51.4, p. 1118-1123. issn : 1098-660X. doi : 10.1128/JCM.03285-12 (cf. p. 35).

Barker, R. H., T. Banchongaksorn, J. M. Courval et al. (avr. 1992). « A simple method to detect Plasmodium falciparum directly from blood samples using the polymerase chain reaction ». eng. In : *The American Journal of Tropical Medicine and Hygiene* 46.4, p. 416-426. issn : 0002-9637. doi : 10.4269/ajtmh.1992.46.416 (cf. p. 41).

Beeson, J G, S J Rogerson, B M Cooke et al. (2000). « Adhesion of Plasmodium falciparum-infected erythrocytes to hyaluronic acid in placental malaria ». In : *Nature medicine* (cf. p. 30, 31).

Berger, Michael F. et Martha L. Bulyk (2006). « Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins ». eng. In : *Methods in Molecular Biology (Clifton, N.J.)* 338, p. 245-260. issn : 1064-3745. doi : 10.1385/1-59745-097-9:245 (cf. p. 76).

Berger, Michael F et Martha L Bulyk (mar. 2009). « Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors ». en. In : *Nature Protocols* 4.3, p. 393-411. issn : 1754-2189, 1750-2799. doi : 10.1038/nprot.2008.195. url : http://www.nature.com/articles/nprot.2008.195 (visité le 26/06/2019) (cf. p. 76).

Bernard, Brady, Vesteinn Thorsson, Hector Rovira et al. (août 2012). « Increasing Coverage of Transcription Factor Position Weight Matrices through Domain-level Homology ». In : *PLoS ONE* 7.8. issn : 1932-6203. doi : 10.

1371/journal.pone.0042779. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3428306/ (visité le 08/07/2019) (cf. p. 82).

Blanden, R. V., P. C. Doherty, M. B. Dunlop et al. (mar. 1975). « Genes required for cytotoxicity against virus-infected target cells in K and D regions of H-2 complex ». eng. In : *Nature* 254.5497, p. 269-270. ISSN : 0028-0836. DOI : 10.1038/254269a0 (cf. p. 67).

Bourée, Patrice (oct. 2008). « [The Institut Pasteur Hospital in Paris : a model in infectiology] ». fre. In : *La Revue Du Praticien* 58.15, p. 1732-1737. ISSN : 0035-2640 (cf. p. 38).

Bourée, Patrice et Francine Bisaro (jan. 2007). « [Parasitic diseases and pregnancy] ». fre. In : *La Revue Du Praticien* 57.2, p. 137-147. ISSN : 0035-2640 (cf. p. 36).

Brabin, B. J. (1983). « An analysis of malaria in pregnancy in Africa ». eng. In : *Bulletin of the World Health Organization* 61.6, p. 1005-1016. ISSN : 0042-9686 (cf. p. 36).

Brisebarre, Audrey, Brice Kumulungui, Serge Sawadogo et al. (mai 2014). « A genome scan for Plasmodium falciparum malaria identifies quantitative trait loci on chromosomes 5q31, 6p21.3, 17p12, and 19p13 ». eng. In : *Malaria Journal* 13, p. 198. ISSN : 1475-2875. DOI : 10.1186/1475-2875-13-198 (cf. p. 62, 64).

Brown, H., G. Turner, S. Rogerson et al. (nov. 1999). « Cytokine expression in the brain in human cerebral malaria ». eng. In : *The Journal of Infectious Diseases* 180.5, p. 1742-1746. ISSN : 0022-1899. DOI : 10.1086/315078 (cf. p. 35).

Bruder, Carl E. G., Arkadiusz Piotrowski, Antoinet A. C. J. Gijsbers et al. (mar. 2008). « Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles ». eng. In : *American Journal of Human Genetics* 82.3, p. 763-771. ISSN : 1537-6605. DOI : 10.1016/j.ajhg.2007.12.011 (cf. p. 52).

Buck, Michael J. et Jason D. Lieb (mar. 2004). « ChIP-chip : considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments ». eng. In : *Genomics* 83.3, p. 349-360. ISSN : 0888-7543 (cf. p. 80).

Campbell, Catarina D., Elizabeth L. Ogburn, Kathryn L. Lunetta et al. (août 2005). « Demonstrating stratification in a European American population ». eng. In : *Nature Genetics* 37.8, p. 868-872. ISSN : 1061-4036. DOI : 10.1038/ng1607 (cf. p. 55).

Cann, J. R. (fév. 1998). « Theoretical studies on the mobility-shift assay of protein-DNA complexes ». eng. In : *Electrophoresis* 19.2, p. 127-141. ISSN : 0173-0835. DOI : 10.1002/elps.1150190202 (cf. p. 73, 74).

Carlson, J., H. Helmby, A. V. Hill et al. (déc. 1990). « Human cerebral malaria : association with erythrocyte rosetting and lack of anti-rosetting antibodies ». eng. In : *Lancet (London, England)* 336.8729, p. 1457-1460. ISSN : 0140-6736. DOI : 10.1016/0140-6736(90)93174-n (cf. p. 32).

Carnevale, Pierre, Vincent Robert, Sylvie Manguin et al. (2009). *Les anophèles : biologie, transmission du Plasmodium et lutte antivectorielle*. fre. Didactiques. Marseille : IRD. URL : http://www.documentation.ird.fr/hor/fdi:010047862 (visité le 02/09/2019) (cf. p. 26).

Carosella, E. D., P. Paul, P. Moreau et al. (nov. 2000). « HLA-G and HLA-E : fundamental and pathophysiological aspects ». eng. In : *Immunology Today* 21.11, p. 532-534. ISSN : 0167-5699 (cf. p. 37).

Castro-Mondragon, Jaime Abraham. « Development of bioinformatics methods for the analysis of large collections of transcription factor binding motifs : positional motif enrichment and motif clustering ». en. In : (), p. 130 (cf. p. 82).

Castro-Mondragon, Jaime Abraham, Sébastien Jaeger, Denis Thieffry et al. (juil. 2017). « RSAT matrix-clustering : dynamic exploration and redundancy reduction of transcription factor binding motif collections ». en. In : *Nucleic Acids Research* 45.13, e119-e119. ISSN : 0305-1048, 1362-4962. DOI : 10.1093/nar/gkx314. URL : https://academic.oup.com/nar/article/45/13/e119/3862068 (visité le 25/06/2019) (cf. p. 96, 119).

Charrier, Cécile, Kaumudi Joshi, Jaeda Coutinho-Budd et al. (mai 2012). « Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation ». eng. In : *Cell* 149.4, p. 923-935. ISSN : 1097-4172. DOI : 10.1016/j.cell.2012.03.034 (cf. p. 52).

Chèneby, Jeanne, Marius Gheorghe, Marie Artufel et al. (jan. 2018). « ReMap 2018 : an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. » In : *Nucleic acids research* 46.D1, p. D267-D275 (cf. p. 57, 88, 129).

Clark, H. C. (oct. 1915). « The diagnostic value of the placenta blood film in aestivo-autumnal malaria ». In : *The Journal of Experimental Medicine* 22.4, p. 427-444. ISSN : 0022-1007. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2125361/ (visité le 26/07/2019) (cf. p. 31).

Clark, I. A. et K. A. Rockett (avr. 1994). « Sequestration, cytokines, and malaria pathology ». eng. In : *International Journal for Parasitology* 24.2, p. 165-166. ISSN : 0020-7519. DOI : 10.1016/0020-7519(94)90022-1 (cf. p. 36).

Clark, Ian A. et William B. Cowden (août 2003). « The pathophysiology of falciparum malaria ». eng. In : *Pharmacology & Therapeutics* 99.2, p. 221-260. ISSN : 0163-7258 (cf. p. 35).

Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition. en-fr. URL : https://www.wiley.com/en-fr/Concise+Encyclopaedia+of+Bioinformatics+and+Computational+Biology%2C+2nd+Edition-p-9780470978719 (visité le 17/07/2019) (cf. p. 70, 71, 77, 81-83).

Conrad, Donald F., Dalila Pinto, Richard Redon et al. (avr. 2010). « Origins and functional impact of copy number variation in the human genome ». In : *Nature* 464.7289, p. 704-712. ISSN : 0028-0836. DOI : 10.1038/nature08516. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3330748/ (visité le 02/10/2019) (cf. p. 51).

Dahlbäck, Madeleine, Morten A. Nielsen et Ali Salanti (mai 2010). « Can any lessons be learned from the ambiguous glycan binding of PfEMP1 domains ? » en. In : *Trends in Parasitology* 26.5, p. 230-235. ISSN : 14714922. DOI : 10.1016/j.pt.2010.02.002. URL : https://linkinghub.elsevier.com/retrieve/pii/S1471492210000231 (visité le 12/06/2019) (cf. p. 19, 31).

Davison, B. B., F. B. Cogswell, G. B. Baskin et al. (oct. 2000). « Placental changes associated with fetal outcome in the Plasmodium coatneyi/rhesus monkey model of malaria in pregnancy ». eng. In : *The American Journal of Tropical*

Medicine and Hygiene 63.3-4, p. 158-173. ISSN : 0002-9637. DOI : 10.4269/ajtmh.2000.63.158 (cf. p. 36).

DAY, Nicholas P. J., Tran Tinh HIEN, Tineke SCHOLLAARDT et al. (oct. 1999). « The Prognostic and Pathophysiologic Role of Pro- and Antiinflammatory Cytokines in Severe Malaria ». en. In : *The Journal of Infectious Diseases* 180.4, p. 1288-1297. ISSN : 0022-1899. DOI : 10.1086/315016. URL : https://academic.oup.com/jid/article/180/4/1288/845197 (visité le 04/09/2019) (cf. p. 36).

DEBETTE, Stéphanie (mai 2012). « How to interpret a genome-wide association study (GWAS) ? » In : *Sang Thrombose Vaisseaux* 24.5, p. 240-247. DOI : 10.1684/stv.2012.0692. URL : http://doi.org/10.1684/stv.2012.0692 (visité le 15/05/2019) (cf. p. 60).

DELAHAYE, Nicolas F., Mathieu BARBIER, Francis FUMOUX et al. (fév. 2007). « Association analyses of NCR3 polymorphisms with P. falciparum mild malaria ». en. In : *Microbes and Infection* 9.2, p. 160-166. ISSN : 12864579. DOI : 10.1016/j.micinf.2006.11.002. URL : https://linkinghub.elsevier.com/retrieve/pii/S1286457906003856 (visité le 29/05/2019) (cf. p. 63).

DELLICOUR, Stephanie, Andrew J TATEM, Carlos A GUERRA et al. (jan. 2010). « Quantifying the Number of Pregnancies at Risk of Malaria in 2007 : A Demographic Study ». In : *PLOS Medicine* 7.1, e1000221 (cf. p. 24).

DEPLANCKE, Bart, Daniel ALPERN et Vincent GARDEUX (juil. 2016). « The Genetics of Transcription Factor DNA Binding Variation ». eng. In : *Cell* 166.3, p. 538-554. ISSN : 1097-4172. DOI : 10.1016/j.cell.2016.07.012 (cf. p. 50).

DEROOST, Katrien, Thao-Thy PHAM, Ghislain OPDENAKKER et al. (mar. 2016). « The immunological balance between host and parasite in malaria ». eng. In : *FEMS microbiology reviews* 40.2, p. 208-257. ISSN : 1574-6976. DOI : 10.1093/femsre/fuv046 (cf. p. 62).

DESAI, Meghna, Feiko O. ter KUILE, François NOSTEN et al. (fév. 2007). « Epidemiology and burden of malaria in pregnancy ». eng. In : *The Lancet. Infectious Diseases* 7.2, p. 93-104. ISSN : 1473-3099. DOI : 10.1016/S1473-3099(07)70021-X (cf. p. 36).

DING, Keyue, Mariza de ANDRADE, Teri A. MANOLIO et al. (juil. 2013). « Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans : an electronic medical record-based genome-wide association study ». eng. In : *G3 (Bethesda, Md.)* 3.7, p. 1061-1068. ISSN : 2160-1836. DOI : 10.1534/g3.113.006452 (cf. p. 110).

DUFFY, P E et M FRIED (août 1999). « Malaria during pregnancy : parasites, antibodies and chondroitin sulphate A. » In : *Biochemical Society transactions* 27.4, p. 478-482 (cf. p. 19).

DUNST, Josefine, Faustin KAMENA et Kai MATUSCHEWSKI (2017). « Cytokines and Chemokines in Cerebral Malaria Pathogenesis ». eng. In : *Frontiers in Cellular and Infection Microbiology* 7, p. 324. ISSN : 2235-2988. DOI : 10.3389/fcimb.2017.00324 (cf. p. 35).

DURINCK, Steffen, Yves MOREAU, Arek KASPRZYK et al. (août 2005). « BioMart and Bioconductor : a powerful link between biological databases and microarray data analysis ». eng. In : *Bioinformatics (Oxford, England)* 21.16, p. 3439-3440. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bti525 (cf. p. 114, 128).

Duvallet, Emilie, Luca Semerano, Eric Assier et al. (nov. 2011). « Interleukin-23 : a key cytokine in inflammatory diseases ». eng. In : *Annals of Medicine* 43.7, p. 503-511. issn : 1365-2060. doi : 10.3109/07853890.2011.577093 (cf. p. 67).

Eastman, Richard T. et David A. Fidock (déc. 2009). « Artemisinin-based combination therapies : a vital tool in efforts to eliminate malaria ». eng. In : *Nature Reviews. Microbiology* 7.12, p. 864-874. issn : 1740-1534. doi : 10.1038/nrmicro2239 (cf. p. 44).

Edwards, Stacey L., Jonathan Beesley, Juliet D. French et al. (nov. 2013). « Beyond GWASs : Illuminating the Dark Road from Association to Function ». In : *The American Journal of Human Genetics* 93.5, p. 779-797. issn : 0002-9297. doi : 10.1016/j.ajhg.2013.10.012. url : http://www.sciencedirect.com/science/article/pii/S0002929713004692 (visité le 07/10/2019) (cf. p. 59).

ENCODE Project Consortium (sept. 2012a). « An integrated encyclopedia of DNA elements in the human genome ». eng. In : *Nature* 489.7414, p. 57-74. issn : 1476-4687. doi : 10.1038/nature11247 (cf. p. 78).

– (sept. 2012b). « An integrated encyclopedia of DNA elements in the human genome. » In : *Nature* 489.7414, p. 57-74 (cf. p. 88).

– (sept. 2012c). « An integrated encyclopedia of DNA elements in the human genome. » In : *Nature* 489.7414, p. 57-74 (cf. p. 120).

Fang, Hai, Bogdan Knezevic, Katie L Burnham et al. (2016). « XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits ». In : *Genome Medicine* 8 (1), p. 129. doi : 10.1186/s13073-016-0384-y. url : http://www.ncbi.nlm.nih.gov/pubmed/27964755 (cf. p. 117).

Flicek, Paul, M. Ridwan Amode, Daniel Barrell et al. (jan. 2014). « Ensembl 2014 ». eng. In : *Nucleic Acids Research* 42.Database issue, p. D749-755. issn : 1362-4962. doi : 10.1093/nar/gkt1196 (cf. p. 92).

Flick, Kirsten, Sanjay Ahuja, Arnaud Chene et al. (déc. 2004). « Optimized expression of Plasmodium falciparum erythrocyte membrane protein 1 domains in Escherichia coli ». eng. In : *Malaria Journal* 3, p. 50. issn : 1475-2875. doi : 10.1186/1475-2875-3-50 (cf. p. 32).

Flori, L, N F Delahaye, F A Iraqi et al. (sept. 2005). « TNF as a malaria candidate gene : polymorphism-screening and family-based association analysis of mild malaria attack and parasitemia in Burkina Faso ». en. In : *Genes & Immunity* 6.6, p. 472-480. issn : 1466-4879, 1476-5470. doi : 10.1038/sj.gene.6364231. url : http://www.nature.com/articles/6364231 (visité le 08/09/2019) (cf. p. 63, 64).

Flori, Laurence, Serge Sawadogo, Christelle Esnault et al. (fév. 2003). « Linkage of mild malaria to the major histocompatibility complex in families living in Burkina Faso ». eng. In : *Human Molecular Genetics* 12.4, p. 375-378. issn : 0964-6906. doi : 10.1093/hmg/ddg033 (cf. p. 62, 63).

Francine, Ntoumi, Bakoua Damien, Fesser Anna et al. (jan. 2016). « Characterization of asymptomatic Plasmodium falciparum infection and its risk factors in pregnant women from the Republic of Congo ». In : *Acta Tropica* 153, p. 111-115 (cf. p. 39).

Franco-Zorrilla, José M., Irene López-Vidriero, José L. Carrasco et al. (fév. 2014). « DNA-binding specificities of plant transcription factors and their

potential to define target genes ». eng. In : *Proceedings of the National Academy of Sciences of the United States of America* 111.6, p. 2367-2372. ISSN : 1091-6490. DOI : 10.1073/pnas.1316278111 (cf. p. 90).

FRIED, M et P E DUFFY (1996). « Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta ». In : *Science* (cf. p. 19, 30, 31, 152).

FRIED, M, F NOSTEN, A BROCKMAN et al. (1998). « Maternal antibodies block malaria ». In : *Nature* (cf. p. 31, 37).

FRIED, Michal, Atis MUEHLENBACHS et Patrick E. DUFFY (oct. 2012). « Diagnosing malaria in pregnancy : an update ». eng. In : *Expert Review of Anti-Infective Therapy* 10.10, p. 1177-1187. ISSN : 1744-8336. DOI : 10.1586/eri.12.98 (cf. p. 41, 152).

FUREY, Terrence S. (déc. 2012). « ChIP-seq and beyond : new and improved methodologies to detect and characterize protein-DNA interactions ». eng. In : *Nature Reviews. Genetics* 13.12, p. 840-852. ISSN : 1471-0064. DOI : 10.1038/nrg3306 (cf. p. 77-79).

FUTURA. *L'étude des Snips*. fr-FR. URL : https://www.futura-sciences.com/sante/dossiers/genetique-diversite-genetique-questions-races-786/page/6/ (visité le 10/07/2019) (cf. p. 50).

GALAS, D. J. et A. SCHMITZ (sept. 1978). « DNAse footprinting : a simple method for the detection of protein-DNA binding specificity ». eng. In : *Nucleic Acids Research* 5.9, p. 3157-3170. ISSN : 0305-1048. DOI : 10.1093/nar/5.9.3157 (cf. p. 74, 75).

GARCIA, A., M. COT, J. P. CHIPPAUX et al. (avr. 1998). « Genetic control of blood infection levels in human malaria : evidence for a complex genetic model. » en. In : *The American Journal of Tropical Medicine and Hygiene* 58.4, p. 480-488. ISSN : 0002-9637, 1476-1645. DOI : 10.4269/ajtmh.1998.58.480. URL : http://www.ajtmh.org/content/journals/10.4269/ajtmh.1998.58.480 (visité le 09/09/2019) (cf. p. 63).

GARDNER, Malcolm J., Neil HALL, Eula FUNG et al. (oct. 2002). « Genome sequence of the human malaria parasite Plasmodium falciparum ». eng. In : *Nature* 419.6906, p. 498-511. ISSN : 0028-0836. DOI : 10.1038/nature01097 (cf. p. 32, 33).

GAZARINI, Marcos L., Andrew P. THOMAS, Tullio POZZAN et al. (avr. 2003). « Calcium signaling in a low calcium environment : how the intracellular malaria parasite solves the problem ». en. In : *The Journal of Cell Biology* 161.1, p. 103-110. ISSN : 0021-9525, 1540-8140. DOI : 10.1083/jcb.200212130. URL : http://www.jcb.org/lookup/doi/10.1083/jcb.200212130 (visité le 09/09/2019) (cf. p. 65).

GENTILINI, M., M. DANIS et J. MOUCHET (jan. 1990). « [Current strategies for the prevention of malaria] ». fre. In : *Bulletin De l'Academie Nationale De Medecine* 174.1, 147-158 ; discussion 158-160. ISSN : 0001-4079 (cf. p. 29).

GENTLEMAN, Robert C., Vincent J. CAREY, Douglas M. BATES et al. (2004). « Bioconductor : open software development for computational biology and bioinformatics ». eng. In : *Genome Biology* 5.10, R80. ISSN : 1474-760X. DOI : 10.1186/gb-2004-5-10-r80 (cf. p. 92).

González, Aitor, Marie Artufel et Pascal Rihet. « TAGOOS : genome-wide supervised learning of non-coding loci associated to complex phenotypes ». en. In : *Nucleic Acids Research* (). doi : 10.1093/nar/gkz320. url : https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz320/5482505 (visité le 03/08/2019) (cf. p. 57).

Gosi, Panita, Charlotte A. Lanteri, Stuart D. Tyner et al. (nov. 2013). « Evaluation of parasite subpopulations and genetic diversity of the msp1, msp2 and glurp genes during and following artesunate monotherapy treatment of Plasmodium falciparum malaria in Western Cambodia ». eng. In : *Malaria Journal* 12, p. 403. issn : 1475-2875. doi : 10.1186/1475-2875-12-403 (cf. p. 33).

Grau, G. E., T. E. Taylor, M. E. Molyneux et al. (juin 1989). « Tumor necrosis factor and disease severity in children with falciparum malaria ». eng. In : *The New England Journal of Medicine* 320.24, p. 1586-1591. issn : 0028-4793. doi : 10.1056/NEJM198906153202404 (cf. p. 35).

Gravel, Simon, Brenna M. Henn, Ryan N. Gutenkunst et al. (juil. 2011). « Demographic history and rare allele sharing among human populations ». eng. In : *Proceedings of the National Academy of Sciences of the United States of America* 108.29, p. 11983-11988. issn : 1091-6490. doi : 10.1073/pnas.1019276108 (cf. p. 54).

Gregson, Aric et Christopher V. Plowe (mar. 2005). « Mechanisms of resistance of malaria parasites to antifolates ». eng. In : *Pharmacological Reviews* 57.1, p. 117-145. issn : 0031-6997. doi : 10.1124/pr.57.1.4 (cf. p. 44).

Griffon, Aurélien, Quentin Barbier, Jordi Dalino et al. (fév. 2015). « Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. » In : *Nucleic acids research* 43.4, e27-e27 (cf. p. 120).

Grobusch, M. P. et P. G. Kremsner (2005). « Uncomplicated malaria ». eng. In : *Current Topics in Microbiology and Immunology* 295, p. 83-104. issn : 0070-217X (cf. p. 35).

Gudbjartsson, Daniel F., Hannes Helgason, Sigurjon A. Gudjonsson et al. (mai 2015). « Large-scale whole-genome sequencing of the Icelandic population ». En. In : *Nature Genetics* 47.5, p. 435. issn : 1546-1718. doi : 10.1038/ng.3247. url : https://www-nature-com.gate2.inist.fr/articles/ng.3247 (visité le 27/07/2019) (cf. p. 54).

Guerini, Danilo, Bin Pan et Ernesto Carafoli (oct. 2003). « Expression, Purification, and Characterization of Isoform 1 of the Plasma Membrane Ca $^{2+}$ Pump : FOCUS ON CALPAIN SENSITIVITY ». en. In : *Journal of Biological Chemistry* 278.40, p. 38141-38148. issn : 0021-9258, 1083-351X. doi : 10.1074/jbc.M302400200. url : http://www.jbc.org/lookup/doi/10.1074/jbc.M302400200 (visité le 09/09/2019) (cf. p. 65).

Gusev, Alexander, S. Hong Lee, Gosia Trynka et al. (nov. 2014). « Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases ». eng. In : *American Journal of Human Genetics* 95.5, p. 535-552. issn : 1537-6605. doi : 10.1016/j.ajhg.2014.10.004 (cf. p. 50).

Haines, Jonathan L., Michael A. Hauser, Silke Schmidt et al. (avr. 2005). « Complement factor H variant increases the risk of age-related macular degeneration ». eng. In : *Science (New York, N.Y.)* 308.5720, p. 419-421. issn : 1095-9203. doi : 10.1126/science.1110359 (cf. p. 60).

Hanssen, S., P. Collinet, E. Leblanc et al. (mai 2013). « Immunological analogies between ovarian cancer and pregnancy ». fre. In : *Journal De Gynecologie, Obstetrique Et Biologie De La Reproduction* 42.3, p. 217-226. ISSN : 1773-0430. DOI : 10.1016/j.jgyn.2012.10.007 (cf. p. 36).

He, Qiye, Jeff Johnston et Julia Zeitlinger (avr. 2015). « ChIP-nexus enables improved detection of in vivo transcription factor binding footprints ». eng. In : *Nature Biotechnology* 33.4, p. 395-401. ISSN : 1546-1696. DOI : 10.1038/nbt.3121 (cf. p. 79).

Helden, J. van, B. André et J. Collado-Vides (sept. 1998). « Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies ». eng. In : *Journal of Molecular Biology* 281.5, p. 827-842. ISSN : 0022-2836. DOI : 10.1006/jmbi.1998.1947 (cf. p. 92).

Helden, J. v. (avr. 2000). « Discovering regulatory elements in non-coding sequences by analysis of spaced dyads ». In : *Nucleic Acids Research* 28.8, p. 1808-1818. ISSN : 13624962. DOI : 10.1093/nar/28.8.1808. URL : https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.8.1808 (visité le 23/06/2019) (cf. p. 92, 93).

Helden, Jacques van (juil. 2003). « Regulatory sequence analysis tools ». eng. In : *Nucleic Acids Research* 31.13, p. 3593-3596. ISSN : 1362-4962. DOI : 10.1093/nar/gkg567 (cf. p. 93).

Hernandez-Valladares, Maria, Pascal Rihet et Fuad A. Iraqi (jan. 2014). « Host susceptibility to malaria in human and mice : compatible approaches to identify potential resistant genes ». en. In : *Physiological Genomics* 46.1, p. 1-16. ISSN : 1094-8341, 1531-2267. DOI : 10.1152/physiolgenomics.00044.2013. URL : http://www.physiology.org/doi/10.1152/physiolgenomics.00044.2013 (visité le 06/09/2019) (cf. p. 62).

Hill, A. V. (1999). « Genetics and genomics of infectious disease susceptibility ». eng. In : *British Medical Bulletin* 55.2, p. 401-413. ISSN : 0007-1420. DOI : 10.1258/0007142991902457 (cf. p. 62, 109).

Hill, A. V., C. E. Allsopp, D. Kwiatkowski et al. (août 1991). « Common west African HLA antigens are associated with protection from severe malaria ». eng. In : *Nature* 352.6336, p. 595-600. ISSN : 0028-0836. DOI : 10.1038/352595a0 (cf. p. 67, 68).

Hindorff, Lucia A., Praveen Sethupathy, Heather A. Junkins et al. (juin 2009). « Potential etiologic and functional implications of genome-wide association loci for human diseases and traits ». eng. In : *Proceedings of the National Academy of Sciences of the United States of America* 106.23, p. 9362-9367. ISSN : 1091-6490. DOI : 10.1073/pnas.0903103106 (cf. p. 58, 112).

Huang, D., M. R. Cancilla et G. Morahan (déc. 2000). « Complete primary structure, chromosomal localisation, and definition of polymorphisms of the gene encoding the human interleukin-12 p40 subunit ». en. In : *Genes & Immunity* 1.8, p. 515-520. ISSN : 1476-5470. DOI : 10.1038/sj.gene.6363720. URL : https://www.nature.com/articles/6363720 (visité le 09/09/2019) (cf. p. 67).

Hunt, Sarah E, William McLaren, Laurent Gil et al. (jan. 2018). « Ensembl variation resources ». en. In : *Database* 2018. ISSN : 1758-0463. DOI : 10.1093/database/bay119. URL : https://academic.oup.com/database/article/doi/10.1093/database/bay119/5255129 (visité le 10/05/2019) (cf. p. 95, 118).

Hviid, Lars (oct. 2011). « The case for PfEMP1-based vaccines to protect pregnant women against Plasmodium falciparum malaria ». eng. In : *Expert Review of Vaccines* 10.10, p. 1405-1414. ISSN : 1744-8395. DOI : 10.1586/erv.11.113 (cf. p. 31).

Isakova, Alina, Romain Groux, Michael Imbeault et al. (2017). « SMiLE-seq identifies binding motifs of single and dimeric transcription factors ». eng. In : *Nature Methods* 14.3, p. 316-322. ISSN : 1548-7105. DOI : 10.1038/nmeth.4143 (cf. p. 90).

Jacob, F., D. Perrin, C. Sanchez et al. (fév. 1960). « [Operon : a group of genes with the expression coordinated by an operator] ». fre. In : *Comptes Rendus Hebdomadaires Des Seances De l'Academie Des Sciences* 250, p. 1727-1729. ISSN : 0001-4036 (cf. p. 69).

Jallow, Muminatou, Yik Ying Teo, Kerrin S. Small et al. (juin 2009). « Genome-wide and fine-resolution association analysis of malaria in West Africa ». eng. In : *Nature Genetics* 41.6, p. 657-665. ISSN : 1546-1718. DOI : 10.1038/ng.388 (cf. p. 64, 66, 110).

Jepson, A. P., W. A. Banya, F. Sisay-Joof et al. (juil. 1995). « Genetic regulation of fever in Plasmodium falciparum malaria in Gambian twin children ». eng. In : *The Journal of Infectious Diseases* 172.1, p. 316-319. ISSN : 0022-1899. DOI : 10.1093/infdis/172.1.316 (cf. p. 62).

Jolma, Arttu, Jian Yan, Thomas Whitington et al. (jan. 2013). « DNA-Binding Specificities of Human Transcription Factors ». English. In : *Cell* 152.1, p. 327-339. ISSN : 0092-8674, 1097-4172. DOI : 10.1016/j.cell.2012.12.009. URL : https://www.cell.com/cell/abstract/S0092-8674(12)01496-1 (visité le 08/07/2019) (cf. p. 90).

Kakuru, Abel, Prasanna Jagannathan, Mary K. Muhindo et al. (mar. 2016). « Dihydroartemisinin–Piperaquine for the Prevention of Malaria in Pregnancy ». en. In : *New England Journal of Medicine* 374.10, p. 928-939. ISSN : 0028-4793, 1533-4406. DOI : 10.1056/NEJMoa1509150. URL : http://www.nejm.org/doi/10.1056/NEJMoa1509150 (visité le 13/06/2019) (cf. p. 46).

Kang, Hyun Min, Jae Hoon Sul, Susan K. Service et al. (avr. 2010). « Variance component model to account for sample structure in genome-wide association studies ». eng. In : *Nature Genetics* 42.4, p. 348-354. ISSN : 1546-1718. DOI : 10.1038/ng.548 (cf. p. 55).

Kanunfre, K. A., F. M. S. Leoratti, E. H. E. Hoffmann et al. (sept. 2003). « Differential Recognition of Plasmodium falciparum Merozoite Surface Protein 2 Variants by Antibodies from Malaria Patients in Brazil ». en. In : *Clinical and Vaccine Immunology* 10.5, p. 973-976. ISSN : 1556-6811. DOI : 10.1128/CDLI.10.5.973-976.2003. URL : http://cvi.asm.org/cgi/doi/10.1128/CDLI.10.5.973-976.2003 (visité le 14/05/2019) (cf. p. 33).

Khan, Aziz, Oriol Fornes, Arnaud Stigliani et al. (jan. 2018). « JASPAR 2018 : update of the open-access database of transcription factor binding profiles and its web framework ». en. In : *Nucleic Acids Research* 46.D1, p. D260-D266. ISSN : 0305-1048, 1362-4962. DOI : 10.1093/nar/gkx1126. URL : http://academic.oup.com/nar/article/46/D1/D260/4621338 (visité le 05/05/2019) (cf. p. 89, 90, 129).

KHAN, Aziz et Anthony MATHELIER (2018). « JASPAR RESTful API : accessing JASPAR data from any programming language ». eng. In : *Bioinformatics (Oxford, England)* 34.9, p. 1612-1614. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btx804 (cf. p. 90).

KHURANA, Ekta, Yao FU, Dimple CHAKRAVARTY et al. (fév. 2016). « Role of non-coding sequence variants in cancer ». eng. In : *Nature Reviews. Genetics* 17.2, p. 93-108. ISSN : 1471-0064. DOI : 10.1038/nrg.2015.17 (cf. p. 55).

KILPINEN, Helena, Sebastian M. WASZAK, Andreas R. GSCHWIND et al. (nov. 2013). « Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription ». eng. In : *Science (New York, N.Y.)* 342.6159, p. 744-747. ISSN : 1095-9203. DOI : 10.1126/science.1242463 (cf. p. 59).

KIMBATSA, Francelet Gildas, Evelyne MAHOUNGOU et Yolande Berton OFOUEME (juin 2018). « L'importance de l'horticulture dans la lutte contre l'insécurité alimentaire, la pauvreté et la protection de l'environnement à Brazzaville (République du Congo) ». fr. In : *Études caribéennes* 39-40. ISSN : 1779-0980. DOI : 10.4000/etudescaribeennes.12382. URL : http://journals.openedition.org/etudescaribeennes/12382 (visité le 04/08/2019) (cf. p. 25).

KINSELLA, Rhoda J., Andreas KÄHÄRI, Syed HAIDER et al. (2011). « Ensembl BioMarts : a hub for data retrieval across taxonomic space ». eng. In : *Database : The Journal of Biological Databases and Curation* 2011, bar030. ISSN : 1758-0463. DOI : 10.1093/database/bar030 (cf. p. 92).

KIRCHER, Martin, Daniela M. WITTEN, Preti JAIN et al. (mar. 2014). « A general framework for estimating the relative pathogenicity of human genetic variants ». eng. In : *Nature Genetics* 46.3, p. 310-315. ISSN : 1546-1718. DOI : 10.1038/ng.2892 (cf. p. 57).

KISZEWSKI, Anthony, Andrew MELLINGER, Andrew SPIELMAN et al. (mai 2004). « A global index representing the stability of malaria transmission ». eng. In : *The American Journal of Tropical Medicine and Hygiene* 70.5, p. 486-498. ISSN : 0002-9637 (cf. p. 27).

KLEIN, Robert J., Caroline ZEISS, Emily Y. CHEW et al. (avr. 2005). « Complement factor H polymorphism in age-related macular degeneration ». eng. In : *Science (New York, N.Y.)* 308.5720, p. 385-389. ISSN : 1095-9203. DOI : 10.1126/science.1109557 (cf. p. 91).

KO, Wen-Ya, Kristin A. KAERCHER, Emanuela GIOMBINI et al. (juin 2011). « Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations ». eng. In : *American Journal of Human Genetics* 88.6, p. 741-754. ISSN : 1537-6605. DOI : 10.1016/j.ajhg.2011.05.005 (cf. p. 66).

KOLOVOS, Petros, Tobias A. KNOCH, Frank G. GROSVELD et al. (jan. 2012). « Enhancers and silencers : an integrated and simple model for their function ». eng. In : *Epigenetics & Chromatin* 5.1, p. 1. ISSN : 1756-8935. DOI : 10.1186/1756-8935-5-1 (cf. p. 72).

KOUKOUIKILA-KOUSSOUNDA, Felix, Damien BAKOUA, Anna FESSER et al. (juil. 2015). « High prevalence of sulphadoxine–pyrimethamine resistance-associated mutations in Plasmodium falciparum field isolates from pregnant women in Brazzaville, Republic of Congo ». In : *Infection, Genetics and Evolution* 33, p. 32-36 (cf. p. 153).

Kublin, James G., Fraction K. Dzinjalamala, Deborah D. Kamwendo et al. (fév. 2002). « Molecular markers for failure of sulfadoxine-pyrimethamine and chlorproguanil-dapsone treatment of Plasmodium falciparum malaria ». eng. In : *The Journal of Infectious Diseases* 185.3, p. 380-388. ISSN : 0022-1899. DOI : 10.1086/338566 (cf. p. 45).

Kulakovskiy, I. V., V. A. Boeva, A. V. Favorov et al. (oct. 2010). « Deep and wide digging for binding motifs in ChIP-Seq data ». eng. In : *Bioinformatics (Oxford, England)* 26.20, p. 2622-2623. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btq488 (cf. p. 78).

Kwiatkowski, Dominic P. (août 2005). « How malaria has affected the human genome and what human genetics can teach us about malaria ». eng. In : *American Journal of Human Genetics* 77.2, p. 171-192. ISSN : 0002-9297. DOI : 10.1086/432519 (cf. p. 68).

Langhorne, Jean, Francis M Ndungu, Anne-Marit Sponaas et al. (juil. 2008). « Immunity to malaria : more questions than answers ». en. In : *Nature Immunology* 9.7, p. 725-732. ISSN : 1529-2908, 1529-2916. DOI : 10.1038/ni.f.205. URL : http://www.nature.com/articles/ni.f.205 (visité le 02/12/2019) (cf. p. 38, 39).

Leary, Scot C. (nov. 2010). « Redox regulation of SCO protein function : controlling copper at a mitochondrial crossroad ». eng. In : *Antioxidants & Redox Signaling* 13.9, p. 1403-1416. ISSN : 1557-7716. DOI : 10.1089/ars.2010.3116 (cf. p. 66).

Leffler, Ellen M., Gavin Band, George B. J. Busby et al. (2017). « Resistance to malaria through structural variation of red blood cell invasion receptors ». eng. In : *Science (New York, N.Y.)* 356.6343. ISSN : 1095-9203. DOI : 10.1126/science.aam6393 (cf. p. 52, 66).

Li, Mulin Jun, Zipeng Liu, Panwen Wang et al. (jan. 2016). « GWASdb v2 : an update database for human genetic variants identified by genome-wide association studies ». en. In : *Nucleic Acids Research* 44.D1, p. D869-D876. ISSN : 0305-1048, 1362-4962. DOI : 10.1093/nar/gkv1317. URL : https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1317 (visité le 08/03/2019) (cf. p. 91).

Lin, Yan, Zipeng Duan, Feng Xu et al. (2017). « Construction and analysis of the transcription factor-microRNA co-regulatory network response to Mycobacterium tuberculosis : a view from the blood ». eng. In : *American Journal of Translational Research* 9.4, p. 1962-1976. ISSN : 1943-8141 (cf. p. 129, 153).

Luty, A. J., D. J. Perkins, B. Lell et al. (juil. 2000). « Low interleukin-12 activity in severe Plasmodium falciparum malaria ». eng. In : *Infection and Immunity* 68.7, p. 3909-3915. ISSN : 0019-9567. DOI : 10.1128/iai.68.7.3909-3915.2000 (cf. p. 67).

López, Carolina, Carolina Saravia, Andromeda Gomez et al. (nov. 2010). « Mechanisms of genetically-based resistance to malaria ». en. In : *Gene* 467.1-2, p. 1-12. ISSN : 03781119. DOI : 10.1016/j.gene.2010.07.008. URL : https://linkinghub.elsevier.com/retrieve/pii/S0378111910002933 (visité le 06/09/2019) (cf. p. 62).

Machanick, Philip et Timothy L. Bailey (juin 2011). « MEME-ChIP : motif analysis of large DNA datasets ». eng. In : *Bioinformatics (Oxford, England)*

27.12, p. 1696-1697. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btr189 (cf. p. 78).

MACINTYRE, Geoff, James BAILEY, Izhak HAVIV et al. (sept. 2010). « is-rSNP : a novel technique for in silico regulatory SNP detection ». en. In : *Bioinformatics* 26.18, p. i524-i530. ISSN : 1460-2059, 1367-4803. DOI : 10.1093/bioinformatics/btq378. URL : https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq378 (visité le 24/06/2019) (cf. p. 86).

MACKINTOSH, Claire L., James G. BEESON et Kevin MARSH (déc. 2004). « Clinical features and pathogenesis of severe malaria ». en. In : *Trends in Parasitology* 20.12, p. 597-603. ISSN : 14714922. DOI : 10.1016/j.pt.2004.09.006. URL : https://linkinghub.elsevier.com/retrieve/pii/S147149220400251X (visité le 06/09/2019) (cf. p. 35).

MALARIA GENOMIC EPIDEMIOLOGY NETWORK, Gavin BAND, Kirk A. ROCKETT et al. (oct. 2015). « A novel locus of resistance to severe malaria in a region of ancient balancing selection ». eng. In : *Nature* 526.7572, p. 253-257. ISSN : 1476-4687. DOI : 10.1038/nature15390 (cf. p. 110).

MALLICK, Swapan, Heng LI, Mark LIPSON et al. (oct. 2016). « The Simons Genome Diversity Project : 300 genomes from 142 diverse populations ». eng. In : *Nature* 538.7624, p. 201-206. ISSN : 1476-4687. DOI : 10.1038/nature18964 (cf. p. 54).

MANJURANO, Alphaxard, Nuno SEPÚLVEDA, Behzad NADJM et al. (oct. 2015). « USP38, FREM3, SDC1, DDC, and LOC727982 Gene Polymorphisms and Differential Susceptibility to Severe Malaria in Tanzania ». eng. In : *The Journal of Infectious Diseases* 212.7, p. 1129-1139. ISSN : 1537-6613. DOI : 10.1093/infdis/jiv192 (cf. p. 66).

MANOLIO, Teri A. (juil. 2010). « Genomewide Association Studies and Assessment of the Risk of Disease ». en. In : *New England Journal of Medicine* 363.2. Sous la dir. de W. Gregory FEERO et Alan E. GUTTMACHER, p. 166-176. ISSN : 0028-4793, 1533-4406. DOI : 10.1056/NEJMra0905980. URL : http://www.nejm.org/doi/10.1056/NEJMra0905980 (visité le 15/05/2019) (cf. p. 58).

MANOLIO, Teri A., Francis S. COLLINS, Nancy J. COX et al. (oct. 2009). « Finding the missing heritability of complex diseases ». eng. In : *Nature* 461.7265, p. 747-753. ISSN : 1476-4687. DOI : 10.1038/nature08494 (cf. p. 60).

MARQUET, S, A GARCIA, N FIEVET et al. (juin 1998). « Linkage analysis of blood Plasmodium falciparum levels : interest of the 5q31-q33 chromosome region. » en. In : *The American Journal of Tropical Medicine and Hygiene* 58.6, p. 705-709. ISSN : 0002-9637, 1476-1645. DOI : 10.4269/ajtmh.1998.58.705. URL : http://www.ajtmh.org/content/journals/10.4269/ajtmh.1998.58.705 (visité le 09/09/2019) (cf. p. 63).

MARSH, K., D. FORSTER, C. WARUIRU et al. (mai 1995). « Indicators of life-threatening malaria in African children ». eng. In : *The New England Journal of Medicine* 332.21, p. 1399-1404. ISSN : 0028-4793. DOI : 10.1056/NEJM199505253322102 (cf. p. 35).

MARTELLE, Susan E., Laura M. RAFFIELD, Nichole D. PALMER et al. (2016). « Dopamine pathway gene variants may modulate cognitive performance in the DHS - Mind Study ». eng. In : *Brain and Behavior* 6.4, e00446. ISSN : 2162-3279. DOI : 10.1002/brb3.446 (cf. p. 66, 67).

Mavoungou, Elie, Jana Held, Ludovic Mewono et al. (mai 2007). « A Duffy Binding—Like Domain Is Involved in the NKp30-Mediated Recognition of Plasmodium falciparum—Parasitized Erythrocytes by Natural Killer Cells ». en. In : *The Journal of Infectious Diseases* 195.10, p. 1521-1531. ISSN : 0022-1899. DOI : 10.1086/515579. URL : https://academic.oup.com/jid/article/195/10/1521/2192041 (visité le 08/09/2019) (cf. p. 63).

Mayengue, Pembe Issamou, Horst Rieth, Ayman Khattab et al. (sept. 2004). « Submicroscopic Plasmodium falciparum infections and multiplicity of infection in matched peripheral, placental and umbilical cord blood samples from Gabonese women ». In : *Tropical Medicine & International Health* 9.9, p. 949-958 (cf. p. 36).

Mayer, D. C. Ghislaine, Joann Cofie, Lubin Jiang et al. (mar. 2009). « Glycophorin B is the erythrocyte receptor of Plasmodium falciparum erythrocyte-binding ligand, EBL-1 ». eng. In : *Proceedings of the National Academy of Sciences of the United States of America* 106.13, p. 5348-5352. ISSN : 1091-6490. DOI : 10.1073/pnas.0900878106 (cf. p. 66).

Mbouamboua, Yvon, Félix Koukouikila-Koussounda, Francine Ntoumi et al. (mai 2019). « Sub-microscopic Plasmodium falciparum infections in matched peripheral, placental and umbilical cord blood samples from asymptomatic Congolese women at delivery ». eng. In : *Acta Tropica* 193, p. 142-147. ISSN : 1873-6254. DOI : 10.1016/j.actatropica.2019.03.001 (cf. p. 38, 39, 130).

McCarthy, Mark I. et Joel N. Hirschhorn (oct. 2008). « Genome-wide association studies : past, present and future ». eng. In : *Human Molecular Genetics* 17.R2, R100-101. ISSN : 1460-2083. DOI : 10.1093/hmg/ddn298 (cf. p. 58).

McConnell, Michael J., John V. Moran, Alexej Abyzov et al. (avr. 2017). « Intersection of diverse neuronal genomes and neuropsychiatric disease : The Brain Somatic Mosaicism Network ». eng. In : *Science (New York, N.Y.)* 356.6336. ISSN : 1095-9203. DOI : 10.1126/science.aal1641 (cf. p. 52).

McDougall, Ian, Francis H. Brown et John G. Fleagle (fév. 2005). « Stratigraphic placement and age of modern humans from Kibish, Ethiopia ». eng. In : *Nature* 433.7027, p. 733-736. ISSN : 1476-4687. DOI : 10.1038/nature03258 (cf. p. 54).

McGregor, I. A. (juil. 1984). « Epidemiology, malaria and pregnancy ». eng. In : *The American Journal of Tropical Medicine and Hygiene* 33.4, p. 517-525. ISSN : 0002-9637. DOI : 10.4269/ajtmh.1984.33.517 (cf. p. 36).

Medina-Rivera, Alejandra, Matthieu Defrance, Olivier Sand et al. (juil. 2015). « RSAT 2015 : Regulatory Sequence Analysis Tools ». eng. In : *Nucleic Acids Research* 43.W1, W50-56. ISSN : 1362-4962. DOI : 10.1093/nar/gkv362 (cf. p. 86, 92-94).

Milet, Jacqueline, Gregory Nuel, Laurence Watier et al. (juil. 2010). « Genome Wide Linkage Study, Using a 250K SNP Map, of Plasmodium falciparum Infection and Mild Malaria Attack in a Senegalese Population ». In : *PLoS ONE* 5.7. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0011616. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2904701/ (visité le 09/09/2019) (cf. p. 64).

Miller, Louis H., Dror I. Baruch, Kevin Marsh et al. (fév. 2002). « The pathogenic basis of malaria ». eng. In : *Nature* 415.6872, p. 673-679. ISSN : 0028-0836. DOI : 10.1038/415673a (cf. p. 29, 66).

Mockenhaupt, Frank P., George Bedu-Addo, Claudia Junge et al. (jan. 2007). « Markers of sulfadoxine-pyrimethamine-resistant Plasmodium falciparum in placenta and circulation of pregnant women ». eng. In : *Antimicrobial Agents and Chemotherapy* 51.1, p. 332-334. ISSN : 0066-4804. DOI : 10.1128/AAC.00856-06 (cf. p. 45).

Moody, Anthony (jan. 2002). « Rapid diagnostic tests for malaria parasites ». eng. In : *Clinical Microbiology Reviews* 15.1, p. 66-78. ISSN : 0893-8512. DOI : 10.1128/cmr.15.1.66-78.2002 (cf. p. 41).

Muanza, K., B. Traoré, F. Gay et al. (juil. 1999). « Circulating receptors implicated in the cyto-adherence occurring in severe Plasmodium falciparum malaria in Thailand ». eng. In : *Annals of Tropical Medicine and Parasitology* 93.5, p. 449-455. ISSN : 0003-4983. DOI : 10.1080/00034989958186 (cf. p. 30).

Mwingira, Felista, Blaise Genton, Abdu-Noor M. Kabanywanyi et al. (nov. 2014). « Comparison of detection methods to estimate asexual Plasmodium falciparum parasite prevalence and gametocyte carriage in a community survey in Tanzania ». eng. In : *Malaria Journal* 13, p. 433. ISSN : 1475-2875. DOI : 10.1186/1475-2875-13-433 (cf. p. 152).

Naidoo, Inbarani et Cally Roper (oct. 2013). « Mapping 'partially resistant', 'fully resistant', and 'super resistant' malaria ». eng. In : *Trends in Parasitology* 29.10, p. 505-515. ISSN : 1471-5007. DOI : 10.1016/j.pt.2013.08.002 (cf. p. 45).

Narlikar, L. et I. Ovcharenko (juil. 2009). « Identifying regulatory elements in eukaryotic genomes ». en. In : *Briefings in Functional Genomics and Proteomics* 8.4, p. 215-230. ISSN : 1473-9550, 1477-4062. DOI : 10.1093/bfgp/elp014. URL : https://academic.oup.com/bfg/article-lookup/doi/10.1093/bfgp/elp014 (visité le 25/06/2019) (cf. p. 56).

Newbold, C., P. Warn, G. Black et al. (oct. 1997). « Receptor-specific adhesion and clinical disease in Plasmodium falciparum ». eng. In : *The American Journal of Tropical Medicine and Hygiene* 57.4, p. 389-398. ISSN : 0002-9637. DOI : 10.4269/ajtmh.1997.57.389 (cf. p. 30).

Nosten, F., F. ter Kuile, L. Maelankirri et al. (août 1991). « Malaria during pregnancy in an area of unstable endemicity ». eng. In : *Transactions of the Royal Society of Tropical Medicine and Hygiene* 85.4, p. 424-429. ISSN : 0035-9203. DOI : 10.1016/0035-9203(91)90205-d (cf. p. 36).

Nosten, F., R. McGready, J. A. Simpson et al. (août 1999). « Effects of Plasmodium vivax malaria in pregnancy ». eng. In : *Lancet (London, England)* 354.9178, p. 546-549. ISSN : 0140-6736. DOI : 10.1016/s0140-6736(98)09247-2 (cf. p. 36, 37).

Nosten, Francois et Rose McGready (juin 2015). « Intermittent presumptive treatment in pregnancy with sulfadoxine-pyrimethamine : a counter perspective ». eng. In : *Malaria Journal* 14, p. 248. ISSN : 1475-2875. DOI : 10.1186/s12936-015-0765-5 (cf. p. 45).

Ntoumi, F., H. Contamin, C. Rogier et al. (jan. 1995). « Age-dependent carriage of multiple Plasmodium falciparum merozoite surface antigen-2 alleles in asymptomatic malaria infections ». eng. In : *The American Journal of Tropical Medicine and Hygiene* 52.1, p. 81-88. ISSN : 0002-9637. DOI : 10.4269/ajtmh.1995.52.81 (cf. p. 33).

O'Connor, Timothy, Mikael Bodén et Timothy L. Bailey (oct. 2016). « CisMapper : predicting regulatory interactions from transcription factor ChIP-seq data ». en. In : *Nucleic Acids Research*, gkw956. issn : 0305-1048, 1362-4962. doi : 10.1093/nar/gkw956. url : https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw956 (visité le 26/06/2019) (cf. p. 78).

Ohta, Keisuke, Hideo Goto, Yusuke Matsumoto et al. (oct. 2016). « Graf1 Controls the Growth of Human Parainfluenza Virus Type 2 through Inactivation of RhoA Signaling ». en. In : *Journal of Virology* 90.20, p. 9394-9405. issn : 0022-538X, 1098-5514. doi : 10.1128/JVI.01471-16. url : https://jvi.asm.org/content/90/20/9394 (visité le 09/09/2019) (cf. p. 64).

Okell, Lucy C., Azra C. Ghani, Emily Lyons et al. (nov. 2009). « Submicroscopic infection in Plasmodium falciparum-endemic populations : a systematic review and meta-analysis ». eng. In : *The Journal of Infectious Diseases* 200.10, p. 1509-1517. issn : 1537-6613. doi : 10.1086/644781 (cf. p. 152).

O'Malley, Ronan C., Shao-Shan Carol Huang, Liang Song et al. (mai 2016). « Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape ». eng. In : *Cell* 165.5, p. 1280-1292. issn : 1097-4172. doi : 10.1016/j.cell.2016.04.038 (cf. p. 90).

Organisation mondiale de la santé (2018). *World malaria report 2018*. en. OCLC : 1088512397. isbn : 978-92-4-156565-3 (cf. p. 23, 24).

Pang, Andy W., Jeffrey R. MacDonald, Dalila Pinto et al. (2010). « Towards a comprehensive structural variation map of an individual human genome ». eng. In : *Genome Biology* 11.5, R52. issn : 1474-760X. doi : 10.1186/gb-2010-11-5-r52 (cf. p. 51).

Parise, M. E., J. G. Ayisi, B. L. Nahlen et al. (nov. 1998). « Efficacy of sulfadoxine-pyrimethamine for prevention of placental malaria in an area of Kenya with a high prevalence of malaria and human immunodeficiency virus infection ». eng. In : *The American Journal of Tropical Medicine and Hygiene* 59.5, p. 813-822. issn : 0002-9637. doi : 10.4269/ajtmh.1998.59.813 (cf. p. 44).

Pende, D., S. Parolini, A. Pessino et al. (nov. 1999). « Identification and molecular characterization of NKp30, a novel triggering receptor involved in natural cytotoxicity mediated by human natural killer cells ». eng. In : *The Journal of Experimental Medicine* 190.10, p. 1505-1516. issn : 0022-1007. doi : 10.1084/jem.190.10.1505 (cf. p. 63).

Pieroni, P., C. D. Mills, C. Ohrt et al. (avr. 1998). « Comparison of the ParaSight-F test and the ICT Malaria Pf test with the polymerase chain reaction for the diagnosis of Plasmodium falciparum malaria in travellers ». eng. In : *Transactions of the Royal Society of Tropical Medicine and Hygiene* 92.2, p. 166-169. issn : 0035-9203. doi : 10.1016/s0035-9203(98)90730-1 (cf. p. 41).

Polley, S. D., Y. Mori, J. Watson et al. (août 2010). « Mitochondrial DNA Targets Increase Sensitivity of Malaria Detection Using Loop-Mediated Isothermal Amplification ». en. In : *Journal of Clinical Microbiology* 48.8, p. 2866-2871. issn : 0095-1137. doi : 10.1128/JCM.00355-10. url : http://jcm.asm.org/cgi/doi/10.1128/JCM.00355-10 (visité le 10/06/2019) (cf. p. 43).

Poon, Leo L. M., Bonnie W. Y. Wong, Edmund H. T. Ma et al. (fév. 2006). « Sensitive and inexpensive molecular test for falciparum malaria : detecting Plasmodium falciparum DNA directly from heat-treated blood by loop-mediated

isothermal amplification ». eng. In : *Clinical Chemistry* 52.2, p. 303-306. ISSN : 0009-9147. DOI : 10.1373/clinchem.2005.057901 (cf. p. 43).

PRICE, Alkes L., Nick J. PATTERSON, Robert M. PLENGE et al. (août 2006). « Principal components analysis corrects for stratification in genome-wide association studies ». eng. In : *Nature Genetics* 38.8, p. 904-909. ISSN : 1061-4036. DOI : 10.1038/ng1847 (cf. p. 55).

PUBLICATION DATA, WHO Library Cataloguing in (2007). « Technical Expert Group meeting on intermittent preventive treatment in pregnancy (IPTp) ». eng. In : *WHO HEADQUARTERS, GENEVA* 1 (cf. p. 44).

PÖSCHL, Birgit, Jarurin WANEESORN, Oriel THEKISOE et al. (juil. 2010). « Comparative diagnosis of malaria infections by microscopy, nested PCR, and LAMP in northern Thailand ». eng. In : *The American Journal of Tropical Medicine and Hygiene* 83.1, p. 56-60. ISSN : 1476-1645. DOI : 10.4269/ajtmh.2010.09-0630 (cf. p. 43).

QIN, Jane Yuxia, Li ZHANG, Kayla L. CLIFT et al. (mai 2010). « Systematic comparison of constitutive promoters and the doxycycline-inducible promoter ». eng. In : *PloS One* 5.5, e10611. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0010611 (cf. p. 70).

RAMENSKY, V. (sept. 2002). « Human non-synonymous SNPs : server and survey ». In : *Nucleic Acids Research* 30.17, p. 3894-3900. ISSN : 13624962. DOI : 10.1093/nar/gkf493. URL : https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkf493 (visité le 15/06/2019) (cf. p. 55).

RAMOS, Erin M., Douglas HOFFMAN, Heather A. JUNKINS et al. (jan. 2014). « Phenotype-Genotype Integrator (PheGenI) : synthesizing genome-wide association study (GWAS) data with existing genomic resources ». eng. In : *European journal of human genetics : EJHG* 22.1, p. 144-147. ISSN : 1476-5438. DOI : 10.1038/ejhg.2013.96 (cf. p. 91).

RAVENHALL, Matt, Susana CAMPINO, Nuno SEPÚLVEDA et al. (2018). « Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania ». eng. In : *PLoS genetics* 14.1, e1007172. ISSN : 1553-7404. DOI : 10.1371/journal.pgen.1007172 (cf. p. 64, 67, 109, 110).

REDDY, Timothy E., Jason GERTZ, Florencia PAULI et al. (mai 2012). « Effects of sequence variation on differential allelic transcription factor occupancy and gene expression ». eng. In : *Genome Research* 22.5, p. 860-869. ISSN : 1549-5469. DOI : 10.1101/gr.131201.111 (cf. p. 59).

REES, David C., Thomas N. WILLIAMS et Mark T. GLADWIN (déc. 2010). « Sickle-cell disease ». eng. In : *Lancet (London, England)* 376.9757, p. 2018-2031. ISSN : 1474-547X. DOI : 10.1016/S0140-6736(10)61029-X (cf. p. 115).

Reporter Gene Assays - FR. fr. URL : https://www.thermofisher.com/fr/fr/home/references/gibco-cell-culture-basics/transfection-basics/reporter-gene-assays.html (visité le 03/08/2019) (cf. p. 57).

RHEE, Ho Sung et B. Franklin PUGH (déc. 2011). « Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution ». eng. In : *Cell* 147.6, p. 1408-1419. ISSN : 1097-4172. DOI : 10.1016/j.cell.2011.11.013 (cf. p. 79).

RIHET, Pascal, Yves TRAORÉ, Laurent ABEL et al. (août 1998). « Malaria in Humans : Plasmodium falciparum Blood Infection Levels Are Linked to Chro-

mosome 5q31-q33 ». en. In : *The American Journal of Human Genetics* 63.2, p. 498-505. ISSN : 00029297. DOI : 10.1086/301967. URL : https://linkinghub.elsevier.com/retrieve/pii/S0002929707614931 (visité le 06/09/2019) (cf. p. 62, 63).

RILEY, E M, G E WAGNER, B D AKANMORI et al. (fév. 2001). « Do maternally acquired antibodies protect infants from malaria infection ? » In : *Parasite Immunology* 23.2, p. 51-59 (cf. p. 38).

RITCHIE, Graham R. S., Ian DUNHAM, Eleftheria ZEGGINI et al. (mar. 2014). « Functional annotation of noncoding sequence variants ». eng. In : *Nature Methods* 11.3, p. 294-296. ISSN : 1548-7105. DOI : 10.1038/nmeth.2832 (cf. p. 57).

ROJANO, Elena, Pedro SEOANE, Juan A G RANEA et al. (juin 2018). « Regulatory variants : from detection to predicting impact ». en. In : *Briefings in Bioinformatics*. ISSN : 1467-5463, 1477-4054. DOI : 10.1093/bib/bby039. URL : https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby039/5035219 (visité le 07/06/2019) (cf. p. 100, 110).

ROSENBERG, Noah A., Jonathan K. PRITCHARD, James L. WEBER et al. (déc. 2002). « Genetic structure of human populations ». eng. In : *Science (New York, N.Y.)* 298.5602, p. 2381-2385. ISSN : 1095-9203. DOI : 10.1126/science.1078311 (cf. p. 54).

ROSENBLOOM, Kate R., Joel ARMSTRONG, Galt P. BARBER et al. (jan. 2015). « The UCSC Genome Browser database : 2015 update ». eng. In : *Nucleic Acids Research* 43.Database issue, p. D670-681. ISSN : 1362-4962. DOI : 10.1093/nar/gku1177 (cf. p. 91).

ROWE, J. A. et S. A. KYES (août 2004). « The role of Plasmodium falciparum var genes in malaria in pregnancy ». eng. In : *Molecular Microbiology* 53.4, p. 1011-1019. ISSN : 0950-382X. DOI : 10.1111/j.1365-2958.2004.04256.x (cf. p. 30).

RSAT Var-tools : an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding | bioRxiv. URL : https://www.biorxiv.org/content/10.1101/623090v1 (visité le 03/08/2019) (cf. p. 57, 119).

SAINT PIERRE, Aude et Emmanuelle GÉNIN (sept. 2014). « How important are rare variants in common disease ? » eng. In : *Briefings in Functional Genomics* 13.5, p. 353-361. ISSN : 2041-2657. DOI : 10.1093/bfgp/elu025 (cf. p. 59).

SANTANA-GARCIA, Walter, Maria ROCHA-ACEVEDO, Lucia RAMIREZ-NAVARRO et al. (nov. 2019). « RSAT variation-tools : An accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding ». en. In : *Computational and Structural Biotechnology Journal*. ISSN : 20010370. DOI : 10.1016/j.csbj.2019.09.009. URL : https://linkinghub.elsevier.com/retrieve/pii/S2001037019301898 (visité le 02/12/2019) (cf. p. 137).

SCHANTZ-DUNN, J et N M NOUR (sept. 2009). « Malaria and pregnancy : a global health perspective ». In : *Reviews in obstetrics and gynecology* 2.3, p. 1-7 (cf. p. 36).

SCHLEBUSCH, Carina M., Pontus SKOGLUND, Per SJÖDIN et al. (oct. 2012). « Genomic variation in seven Khoe-San groups reveals adaptation and complex

African history ». eng. In : *Science (New York, N.Y.)* 338.6105, p. 374-379. ISSN : 1095-9203. DOI : 10.1126/science.1227721 (cf. p. 54).

SCHNEIDER, T D et R M STEPHENS (oct. 1990). « Sequence logos : a new way to display consensus sequences. » In : *Nucleic Acids Research* 18.20, p. 6097-6100. ISSN : 0305-1048. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC332411/ (visité le 19/07/2019) (cf. p. 84).

SCHRIDER, Daniel R., Matthew W. HAHN et David J. BEGUN (2016). « Parallel Evolution of Copy-Number Variation across Continents in Drosophila melanogaster ». eng. In : *Molecular Biology and Evolution* 33.5, p. 1308-1316. ISSN : 1537-1719. DOI : 10.1093/molbev/msw014 (cf. p. 52).

SCHULTZ, L. J., R. W. STEKETEE, A. MACHESO et al. (nov. 1994). « The efficacy of antimalarial regimens containing sulfadoxine-pyrimethamine and/or chloroquine in preventing peripheral and placental Plasmodium falciparum infection among pregnant women in Malawi ». eng. In : *The American Journal of Tropical Medicine and Hygiene* 51.5, p. 515-522. ISSN : 0002-9637. DOI : 10.4269/ajtmh.1994.51.515 (cf. p. 44).

SERRA-CASAS, Elisa, Paulo MANRIQUE, Xavier C. DING et al. (oct. 2017). « Loop-mediated isothermal DNA amplification for asymptomatic malaria detection in challenging field settings : Technical performance and pilot implementation in the Peruvian Amazon ». In : *PLoS ONE* 12.10. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0185742. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5628891/ (visité le 04/09/2019) (cf. p. 43).

SHERMAN, I.W., I.E. CRANDALL, N. GUTHRIE et al. (oct. 1995). « The sticky secrets of sequestration ». en. In : *Parasitology Today* 11.10, p. 378-384. ISSN : 01694758. DOI : 10.1016/0169-4758(95)80006-9. URL : https://linkinghub.elsevier.com/retrieve/pii/0169475895800069 (visité le 03/09/2019) (cf. p. 30).

SHERMAN, Rachel M., Juliet FORMAN, Valentin ANTONESCU et al. (jan. 2019). « Assembly of a pan-genome from deep sequencing of 910 humans of African descent ». In : *Nature genetics* 51.1, p. 30-35. ISSN : 1061-4036. DOI : 10.1038/s41588-018-0273-y. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6309586/ (visité le 09/09/2019) (cf. p. 52).

SHERRY, S. T., M. WARD et K. SIROTKIN (août 1999). « dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation ». eng. In : *Genome Research* 9.8, p. 677-679. ISSN : 1088-9051 (cf. p. 90).

SHERRY, S. T., M.-H. WARD, M. KHOLODOV et al. (jan. 2001). « dbSNP : the NCBI database of genetic variation ». en. In : *Nucleic Acids Research* 29.1, p. 308-311. ISSN : 0305-1048. DOI : 10.1093/nar/29.1.308. URL : https://academic.oup.com/nar/article/29/1/308/1116004 (visité le 18/07/2019) (cf. p. 90).

SHI, Wenqiang, Oriol FORNES, Anthony MATHELIER et al. (déc. 2016). « Evaluating the impact of single nucleotide variants on transcription factor binding ». In : *Nucleic Acids Research* 44.21, p. 10106-10116. ISSN : 0305-1048. DOI : 10.1093/nar/gkw691. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5137422/ (visité le 09/07/2019) (cf. p. 86).

SHIHAB, Hashem A., Mark F. ROGERS, Julian GOUGH et al. (mai 2015). « An integrative approach to predicting the functional effects of non-coding and coding

sequence variation ». eng. In : *Bioinformatics (Oxford, England)* 31.10, p. 1536-1543. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btv009 (cf. p. 57).

SINGH, Balbir, Lee KIM SUNG, Asmad MATUSOP et al. (mar. 2004). « A large focus of naturally acquired Plasmodium knowlesi infections in human beings ». eng. In : *Lancet (London, England)* 363.9414, p. 1017-1024. ISSN : 1474-547X. DOI : 10.1016/S0140-6736(04)15836-4 (cf. p. 27).

SMEDLEY, Damian, Max SCHUBACH, Julius O. B. JACOBSEN et al. (2016). « A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease ». eng. In : *American Journal of Human Genetics* 99.3, p. 595-606. ISSN : 1537-6605. DOI : 10.1016/j.ajhg.2016.07.005 (cf. p. 57).

SMITH, Joseph D, Gangadharan SUBRAMANIAN, Benoit GAMAIN et al. (oct. 2000). « Classification of adhesive domains in the Plasmodium falciparum Erythrocyte Membrane Protein 1 family ». en. In : *Molecular and Biochemical Parasitology* 110.2, p. 293-310. ISSN : 01666851. DOI : 10.1016/S0166-6851(00)00279-6. URL : https://linkinghub.elsevier.com/retrieve/pii/S0166685100002796 (visité le 15/06/2019) (cf. p. 30).

SMYTHE, J. A., M. G. PETERSON, R. L. COPPEL et al. (mar. 1990). « Structural diversity in the 45-kilodalton merozoite surface antigen of Plasmodium falciparum ». eng. In : *Molecular and Biochemical Parasitology* 39.2, p. 227-234. ISSN : 0166-6851 (cf. p. 33).

SNOUNOU, G., S. VIRIYAKOSOL, X. P. ZHU et al. (oct. 1993). « High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction ». eng. In : *Molecular and Biochemical Parasitology* 61.2, p. 315-320. ISSN : 0166-6851 (cf. p. 41, 42).

SOBOLEWSKI, Peter, Irene GRAMAGLIA, John FRANGOS et al. (sept. 2005). « Nitric oxide bioavailability in malaria ». eng. In : *Trends in Parasitology* 21.9, p. 415-422. ISSN : 1471-4922. DOI : 10.1016/j.pt.2005.07.002 (cf. p. 65).

SPRONSEN, Joanne H. M. van, Tom A. J. SCHNEIDER et Stephen ATASIGE (avr. 2012). « Placental malaria and the relationship to pregnancy outcome at Gushegu District Hospital, Northern Ghana ». eng. In : *Tropical Doctor* 42.2, p. 80-84. ISSN : 1758-1133. DOI : 10.1258/td.2011.110431 (cf. p. 45).

STEED, Emily, Nelio T. L. RODRIGUES, Maria S. BALDA et al. (déc. 2009). « Identification of MarvelD3 as a tight junction-associated transmembrane protein of the occludin family ». eng. In : *BMC cell biology* 10, p. 95. ISSN : 1471-2121. DOI : 10.1186/1471-2121-10-95 (cf. p. 67).

STEKETEE, R W, B L NAHLEN, M E PARISE et al. (jan. 2001). « The burden of malaria in pregnancy in malaria-endemic areas. » In : *The American journal of tropical medicine and hygiene* 64.1-2 Suppl, p. 28-35 (cf. p. 19, 36).

STORMO, G. D. (jan. 2000). « DNA binding sites : representation and discovery ». eng. In : *Bioinformatics (Oxford, England)* 16.1, p. 16-23. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/16.1.16 (cf. p. 82).

STURROCK, Hugh J W, Michelle S HSIANG, Justin M COHEN et al. (juin 2013). « Targeting Asymptomatic Malaria Infections : Active Surveillance in Control and Elimination ». In : *PLOS Medicine* 10.6, e1001467 (cf. p. 38).

TAYLOR, R. R., D. B. SMITH, V. J. ROBINSON et al. (nov. 1995). « Human antibody response to Plasmodium falciparum merozoite surface protein 2 is

serogroup specific and predominantly of the immunoglobulin G3 subclass ». eng. In : *Infection and Immunity* 63.11, p. 4382-4388. ISSN : 0019-9567 (cf. p. 33).

TAYLOR, Steve M., Carla CERAMI et Rick M. FAIRHURST (2013). « Hemoglobinopathies : slicing the Gordian knot of Plasmodium falciparum malaria pathogenesis ». eng. In : *PLoS pathogens* 9.5, e1003327. ISSN : 1553-7374. DOI : 10.1371/journal.ppat.1003327 (cf. p. 115).

TAYLOR, T. E., A. BORGSTEIN et M. E. MOLYNEUX (fév. 1993). « Acid-base status in paediatric Plasmodium falciparum malaria ». eng. In : *The Quarterly Journal of Medicine* 86.2, p. 99-109. ISSN : 0033-5622 (cf. p. 35).

TENG, Mingxiang, Shoji ICHIKAWA, Leah R. PADGETT et al. (juil. 2012). « regSNPs : a strategy for prioritizing regulatory single nucleotide substitutions ». eng. In : *Bioinformatics (Oxford, England)* 28.14, p. 1879-1886. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/bts275 (cf. p. 86).

THE 1000 GENOMES PROJECT CONSORTIUM (oct. 2015). « A global reference for human genetic variation ». en. In : *Nature* 526.7571, p. 68-74. ISSN : 0028-0836, 1476-4687. DOI : 10.1038/nature15393. URL : http://www.nature.com/articles/nature15393 (visité le 09/06/2019) (cf. p. 49, 51, 53).

The role of regulatory variation in complex traits and disease. - PubMed - NCBI. URL : https://www-ncbi-nlm-nih-gov.gate2.inist.fr/pubmed/?term=Albert%2C+F.+W.+%26+Kruglyak%2C+L.+Te+role+of+regulatory+variation+in+complex+traits+and+disease.+Nat.+Rev.+Genet.+16%2C+197%E2%80%93212+(2015) (visité le 09/07/2019) (cf. p. 50).

THOMAS-CHOLLIER, Morgane, Matthieu DEFRANCE, Alejandra MEDINA-RIVERA et al. (juil. 2011). « RSAT 2011 : regulatory sequence analysis tools ». eng. In : *Nucleic Acids Research* 39.Web Server issue, W86-91. ISSN : 1362-4962. DOI : 10.1093/nar/gkr377 (cf. p. 92, 93).

THOMAS-CHOLLIER, Morgane, Andrew HUFTON, Matthias HEINIG et al. (nov. 2011). « Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs ». eng. In : *Nature Protocols* 6.12, p. 1860-1869. ISSN : 1750-2799. DOI : 10.1038/nprot.2011.409 (cf. p. 86).

THORLEIFSSON, Gudmar, Kristinn P. MAGNUSSON, Patrick SULEM et al. (sept. 2007). « Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma ». eng. In : *Science (New York, N.Y.)* 317.5843, p. 1397-1400. ISSN : 1095-9203. DOI : 10.1126/science.1146554 (cf. p. 60).

TIAN, Chao, Bethann S. HROMATKA, Amy K. KIEFER et al. (2017). « Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections ». eng. In : *Nature Communications* 8.1, p. 599. ISSN : 2041-1723. DOI : 10.1038/s41467-017-00257-5 (cf. p. 128).

TIMMANN, Christian, Thorsten THYE, Maren VENS et al. (sept. 2012). « Genome-wide association study indicates two novel resistance loci for severe malaria ». eng. In : *Nature* 489.7416, p. 443-446. ISSN : 1476-4687. DOI : 10.1038/nature11334 (cf. p. 64, 65, 67, 109).

TODOROVIC, Vesna (mar. 2016). « Genetics. Predicting the impact of genomic variation ». eng. In : *Nature Methods* 13.3, p. 203. ISSN : 1548-7105 (cf. p. 59).

TRAPE, J F et A ZOULANI (jan. 1987). « Malaria and urbanization in Central Africa : the example of Brazzaville. Part II : Results of entomological surveys and

epidemiological analysis ». In : *Transactions of The Royal Society of Tropical Medicine and Hygiene* 81.Supplement$_2$, p. 10-18 (cf. p. 26).

Trovoada, Maria de Jesus, Madalena Martins, Riadh Ben Mansour et al. (mar. 2014). « NOS2 Variants Reveal a Dual Genetic Control of Nitric Oxide Levels, Susceptibility to Plasmodium Infection, and Cerebral Malaria ». In : *Infection and Immunity* 82.3, p. 1287-1295. issn : 0019-9567. doi : 10.1128/IAI.01070-13. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3958015/ (visité le 09/09/2019) (cf. p. 65).

Turatsinze, Jean-Valery, Morgane Thomas-Chollier, Matthieu Defrance et al. (oct. 2008). « Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules ». en. In : *Nature Protocols* 3.10, p. 1578-1588. issn : 1754-2189, 1750-2799. doi : 10.1038/nprot.2008.97. url : http://www.nature.com/articles/nprot.2008.97 (visité le 08/03/2019) (cf. p. 96, 118).

Turner, G. D., H. Morrison, M. Jones et al. (nov. 1994). « An immunohistochemical study of the pathology of fatal malaria. Evidence for widespread endothelial activation and a potential role for intercellular adhesion molecule-1 in cerebral sequestration ». eng. In : *The American Journal of Pathology* 145.5, p. 1057-1069. issn : 0002-9440 (cf. p. 65).

Uneke, Chigozie J. et Anayo Ogbonna (août 2009). « Malaria and HIV co-infection in pregnancy in sub-Saharan Africa : impact of treatment using anti-malarial and antiretroviral agents ». eng. In : *Transactions of the Royal Society of Tropical Medicine and Hygiene* 103.8, p. 761-767. issn : 1878-3503. doi : 10.1016/j.trstmh.2008.06.017 (cf. p. 37).

Vainberg, I E, W D Tap, S A Lewis et al. (1995). « Adherence of Plasmodium falciparum to Chondroitin Sulfate A in the Human Placenta ». In : *Nature* (cf. p. 31).

Valnot, I., S. Osmond, N. Gigarel et al. (nov. 2000). « Mutations of the SCO1 gene in mitochondrial cytochrome c oxidase deficiency with neonatal-onset hepatic failure and encephalopathy ». eng. In : *American Journal of Human Genetics* 67.5, p. 1104-1109. issn : 0002-9297. doi : 10.1016/S0002-9297(07)62940-1 (cf. p. 66).

Vásquez, Ana María, Lina Zuluaga, Alberto Tobón et al. (juil. 2018). « Diagnostic accuracy of loop-mediated isothermal amplification (LAMP) for screening malaria in peripheral and placental blood samples from pregnant women in Colombia ». In : *Malaria Journal* 17. issn : 1475-2875. doi : 10.1186/s12936-018-2403-5. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6044080/ (visité le 04/09/2019) (cf. p. 43).

Wang, D. G. (mai 1998). « Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome ». In : *Science* 280.5366, p. 1077-1082. issn : 00368075, 10959203. doi : 10.1126/science.280.5366.1077. url : http://www.sciencemag.org/cgi/doi/10.1126/science.280.5366.1077 (visité le 15/06/2019) (cf. p. 49, 50).

Wang, Jun, Abu Z. Dayem Ullah et Claude Chelala (mai 2018). « IW-Scoring : an Integrative Weighted Scoring framework for annotating and prioritizing genetic variations in the noncoding genome ». eng. In : *Nucleic Acids Research* 46.8, e47. issn : 1362-4962. doi : 10.1093/nar/gky057 (cf. p. 57).

WARRELL, D. A. (oct. 1997). « Cerebral malaria : clinical features, pathophysiology and treatment ». eng. In : *Annals of Tropical Medicine and Parasitology* 91.7, p. 875-884. ISSN : 0003-4983. DOI : 10.1080/00034989760644 (cf. p. 30).

WASSERMAN, Wyeth W. et Albin SANDELIN (avr. 2004). « Applied bioinformatics for the identification of regulatory elements ». eng. In : *Nature Reviews. Genetics* 5.4, p. 276-287. ISSN : 1471-0056. DOI : 10.1038/nrg1315 (cf. p. 73, 82-84).

WELTER, Danielle, Jacqueline MACARTHUR, Joannella MORALES et al. (jan. 2014). « The NHGRI GWAS Catalog, a curated resource of SNP-trait associations ». eng. In : *Nucleic Acids Research* 42.Database issue, p. D1001-1006. ISSN : 1362-4962. DOI : 10.1093/nar/gkt1229 (cf. p. 91, 109).

WHITE, Nicholas J., Sasithon PUKRITTAYAKAMEE, Tran Tinh HIEN et al. (fév. 2014). « Malaria ». eng. In : *Lancet (London, England)* 383.9918, p. 723-735. ISSN : 1474-547X. DOI : 10.1016/S0140-6736(13)60024-0 (cf. p. 19, 27).

WHITE, Tim D., Berhane ASFAW, David DEGUSTA et al. (juin 2003). « Pleistocene Homo sapiens from Middle Awash, Ethiopia ». eng. In : *Nature* 423.6941, p. 742-747. ISSN : 0028-0836. DOI : 10.1038/nature01669 (cf. p. 54).

WHITFIELD, Troy W., Jie WANG, Patrick J. COLLINS et al. (sept. 2012). « Functional analysis of transcription factor binding sites in human promoters ». eng. In : *Genome Biology* 13.9, R50. ISSN : 1474-760X. DOI : 10.1186/gb-2012-13-9-r50 (cf. p. 71).

WILSON, Mary E., Leisa H. WELD, Andrea BOGGILD et al. (juin 2007). « Fever in returned travelers : results from the GeoSentinel Surveillance Network ». eng. In : *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America* 44.12, p. 1560-1568. ISSN : 1537-6591. DOI : 10.1086/518173 (cf. p. 35).

World Health Organization , 2015. Malaria Treatment Guidelines. Geneva, Switzerland : World Health Organization (cf. p. 19).

YAMAN, F. al, B. GENTON, D. MOKELA et al. (fév. 1995). « Human cerebral malaria : lack of significant association between erythrocyte rosetting and disease severity ». eng. In : *Transactions of the Royal Society of Tropical Medicine and Hygiene* 89.1, p. 55-58. ISSN : 0035-9203. DOI : 10.1016/0035-9203(95)90658-4 (cf. p. 32).

YANG, Jian, S. Hong LEE, Michael E. GODDARD et al. (jan. 2011). « GCTA : a tool for genome-wide complex trait analysis ». eng. In : *American Journal of Human Genetics* 88.1, p. 76-82. ISSN : 1537-6605. DOI : 10.1016/j.ajhg.2010.11.011 (cf. p. 60).

YATES, Andrew, Wasiu AKANNI, M. Ridwan AMODE et al. (jan. 2016). « Ensembl 2016 ». eng. In : *Nucleic Acids Research* 44.D1, p. D710-716. ISSN : 1362-4962. DOI : 10.1093/nar/gkv1157 (cf. p. 91).

YATES, Andrew, Kathryn BEAL, Stephen KEENAN et al. (jan. 2015). « The Ensembl REST API : Ensembl Data for Any Language ». en. In : *Bioinformatics* 31.1, p. 143-145. ISSN : 1460-2059, 1367-4803. DOI : 10.1093/bioinformatics/btu613. URL : https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu613 (visité le 10/05/2019) (cf. p. 92).

YU, Wei, Marta GWINN, Melinda CLYNE et al. (fév. 2008). « A navigator for human genome epidemiology ». eng. In : *Nature Genetics* 40.2, p. 124-125. ISSN : 1546-1718. DOI : 10.1038/ng0208-124 (cf. p. 91).

ZHOU, Zhi, Jialong YANG, Lingling WANG et al. (avr. 2011). « A dopa decarboxylase modulating the immune response of scallop Chlamys farreri ». eng. In : *PloS One* 6.4, e18596. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0018596 (cf. p. 67).

ZÁMBÓ, Boglárka, György VÁRADY, Rita PADÁNYI et al. (juil. 2017). « Decreased calcium pump expression in human erythrocytes is connected to a minor haplotype in the ATP2B4 gene ». en. In : *Cell Calcium* 65, p. 73-79. ISSN : 01434160. DOI : 10.1016/j.ceca.2017.02.001. URL : https://linkinghub.elsevier.com/retrieve/pii/S0143416017300040 (visité le 09/09/2019) (cf. p. 65).

Notes

1. . . . https://apps.who.int/iris/bitstream/handle/10665/275867/9789241565653-eng.pdf?ua=1
2. . . . http://malaria.who.int/pregnantwomenandinfants.html
3. . . . World malaria report 2015
4. . . . WHO, MHP (Ministère de la Santé et de la Population). Profil entomologique du Congo, rapport technique OMS/Ministère de la Santé et de la Population, Congo 2006
5. . . . NMCP (PNLP). Rapport d'activités du programme national de lutte contre le paludisme, année 2015. Brazzaville 2016
6. . . . Rapport d'activités du programme national de lutte contre le paludisme, année 2015. Brazzaville 2016
7. . . . Centers for Disease Control and Prévention. Malaria : Disease. http://www.cdc.gov/malaria/about/disease.html (Accessed on July 28, 2015)
8. . . . Institute for Health Metrics and Evaluation. Global Burden of Disease (GBD) Study 2015. http://ghdx.healthdata.org/gbd-2015/data-input-sources. Accessed 12 Mar 2017
9. . . . Prise en charge et prévention du paludisme d'importation, mise à jour 2017, SPILF) (http://www.infectiologie.com/UserFiles/File/spilf/recos/2017-palu-texte-final.pdf)
10. . . . http://whqlibdoc.who.int/afro/2004/AFR_MAL_04.01.pdf
11. . . . https://apps.who.int/medicinedocs/en/d/Jh2923f/2.5.1.html
12. . . . www.who.int/malaria/areas/preventive_therapies/pregnancy/en
13. . . . http://www.who.int/malaria/publications/atoz/istp-and-act-in-pregnancy.pdf
14. . . . https://www.ensembl.org/info/genome/variation/prediction/classification.html#seq
15. . . . www.ebi.ac.uk/gwas
16. . . . https://www.rstudio.com/products/RStudio/
17. . . . http://rsat-tagc.univ-mrs.fr/rsat/rest.wsgi/variation-info/Homo_sapiens/GRCh38?i_string=rs141426589&i_string_type=text&format=id&col=1&content-type=text%2Fplain
18. . . . http://rsat-tagc.univ-mrs.fr/rsat/rest.wsgi/variation-info/Homo_sapiens/GRCh38?
19. . . . i_string=rs141426589&i_string_type=text&format=id&col=1&content-type=text%2Fplain
20. . . . http://tagc.univ-mrs.fr/rsa-tools/
21. . . . http://tagc.univ-mrs.fr/remap/
22. . . . https://www.ebi.ac.uk/gwas/efotraits/EFO_0001068
23. . . . https://www.ebi.ac.uk/gwas/)
24. . . . https://www.ncbi.nlm.nih.gov/snp/

25. . . . http://rest.ensembl.org/documentation/info/ld_id_get

26. . . . http://rest.ensembl.org/

27. . . . http://rsat-tagc.univ-mrs.fr/rest/

28. . . . http://rsat-tagc.univ-mrs.fr/rest/

29. . . . http://rsat-tagc.univ-mrs.fr/rest/

30. . . . https://www.ebi.ac.uk/gwas/

31. . . . http://rest.ensembl.org/

A. ANNEXES

Annexe1. Performance des tests de diagnostic rapide du sang périphérique et placentaire pour le diagnostic du paludisme gestationnel (Fried D. et al, 2012). FS = frottis sanguin; ICT = Immunochromatographic test (test immunochromatographique).

Test (sang périphérique)	Kit (source)	Référence	Sensibilité et spécificité	Etude (année)
HRP2	ICT Malaria Pf (Amrad ICT, Sydney, NSW, Australia)	FS périphérique	94.4, 90.6	Leke et al. (1999)
HRP2		FS placentaire	89, 94.9	Leke et al. (1999)
FS		FS placentaire	52.1, 92.7	Mankhambo et al. (2002)
LDHp	OptiMAL (Flow, Inc. Portland, OR, USA)	FS périphérique	70.7, 93.8	Mankhambo et al. (2002)
	OptiMAL (Flow, Inc. Portland, OR, USA)	FS placentaire	38.4, 90.8	Mankhambo et al. (2002)
FS		FS placentaire	42, 97	Mankhambo et al. (2002)
HRP2	ICT Malaria Pf/Pv (BD, Heidelberg, Germany)	FS placentaire	80, 90	Mankhambo et al. (2002)
FS		PCR placentaire	27, 100	Mankhambo et al. (2002)
HRP2	ICT Malaria Pf/Pv (BD, Heidelberg, Germany)	PCR placentaire	56, 97	Mankhambo et al. (2002)
FS		FS placentaire	82, 86	Singer et al. (2004)
HRP2	MAKROmed Pty, Ltd., (Johannesburg, South Africa)	FS périphérique	96, 67	Singer et al. (2004)
HRP2	MAKROmed Pty, Ltd., (Johannesburg, South Africa)	FS placentaire	95, 61	Singer et al. (2004)
HRP2	MAKROmed Pty, Ltd., (Johannesburg, South Africa)	PCR placentaire	92, 59	Singer et al. (2004)
LDHp	OptiMAL (Flow, Inc., OR, USA)	FS périphérique	15, 98	Vanderjagt et al. (2005)
FS		FS placentaire	50, 98	Mockenhaupt et al. (2006)
HRP2	ICT Malaria Pf/Pv (BD, Heidelberg, Germany)	FS placentaire	78, 89	Mockenhaupt et al. (2006)
LDHp	OptiMAL (DiaMed AG, Cressier, Swizerland)	FS périphérique	96.6, 85.4	Tagbor et al. (2008)
	Diagnosticks, Malaria Pf cassette (SSA Diagnostics and Biotech Systems, Goa, India)	FS périphérique	96.8, 73.5	Kyabayinze et al. (2011)
FS		Histologie placentaire	80.9, 87.5	Kyabayinze et al. (2011)
		PCR périphérique	36.4, 99.6	Dhorda et al. (2012)
HRP2	Paracheck Pf (Orchid, Goa, India)	PCR périphérique	31.8, 100	Dhorda et al. (2012)
		Histologie placentaire	65.2, 97.8	Mayor et al. (2012)
HRP2	SD Bioline (Standard Diagnostics)	Histologie placentaire	78.3, 93.4	Mayor et al. (2012)
Test (sang placentaire)	.	.	.	.
HRP2	MAKROmed Pty, Ltd. (Johannesburg, South Africa)	FS placentaire	95, 72	Singer et al. (2004)
		PCR placentaire	89, 76	Singer et al. (2004)
HRP2	Paracheck Pf (Orchid, Goa, India)	FS placentaire	93.3, 84.4	Singh et al. (2005)
HRP2	ParaHITf (Span Diagnostics, Inc., Surat, India)	FS placentaire	87.5, 97	Singh et al. (2005)
HRP2	MAKROmed Pty, Ltd., (Johannesburg, South Africa)	FS placentaire	100, 68	Sarr et al. (2006)
HRP2	Diagnosticks, Malaria Pf cassette (SSA Diagnostics and Biotech Systems, Goa, India)	FS placentaire	80.9, 87.5	Kyabayinze et al. (2011)

Annexe 2. Résumé de la capacité à haut débit et de l'applicabilité sur le terrain de la microscopie, du TDR et de nouvelles plates-formes d'amplification d'acides nucléiques (Sumudu B et al., 2016).

Technique de diagnostic	Plateforme	Ref	Limite de détection (parasites/µL)	Validé sur des échantillons symptomatiques ou positifs au microscope	Validé sur des échantillons asymptomatiq ues ou de surveillance	
Microscopie	Light microscopy	Coleman R et al., 2002	50–499	Oui	Oui	
TDR	Immunochromatogra p hic ateral flow assay	Tiono A et al., 2014	200.0	Oui	Oui	
PCR	Ultrasensitive qPCR	Imwong M et al., 2014	0.022 (sang)	Oui	Non	
PCR	Multiplex Malaria Sample Ready	Kamau E et al., 2014	0.244 (sang)	Uncertain	Uncertain	
PCR	Mobile PCR	Canier L et al., 2013	2 (tache de sang séché)	Uncertain	Oui	
PCR	PET-PCR	Lucchi N et al., 2013	3.2 (sang)	Uncertain	Oui	
PCR	Gelcycler	Taylor B et al., 2014	2 (sang)	Oui	Oui	
PCR	Multiplex microarray assay	Tan J et al., 2014	uncertain	Oui	Uncertain	
PCR	CLIP-PCR	Cheng Z et al., 2015	0.3 (tache de sang séché)	Oui	Oui	
LAMP	RealAmp	Lucchi N et al., 2010	0.4–40 (sang)	Oui	Oui	
LAMP	HtLAMP	Britton S et al., 2015	2.5 (sang)	Oui	Non	
LAMP	LAMP-LFD	Yongkiettrakul S et al., 2014	Uncertain	Oui	Non	
LAMP	NINA-LAMP	Sema M et al., 2015	Uncertain	Oui	Oui	
LAMP	LoopAmp	Hopkins H et al., 2013	5.0 (sang)	Oui	Oui	