Université de Montréal

Analyse multiomique pour l'identification de biomarqueurs prédictifs de la réponse aux vaccins

par
Slim FOURATI

Département de biochimie
Faculté de médecine

Thèse présentée à la Faculté de Médecine
en vue de l'obtention du grade de Philosophiæ Doctor (Ph. D.)
en bio-informatique

Avril 2022

Université de Montréal
Faculté de Médecine

Cette thèse intitulée

Analyse multiomique pour l'identification de biomarqueurs prédictifs de la réponse aux vaccins

Présentée par
Slim FOURATI

A été évaluée par un jury composé des personnes suivantes :
Julie HUSSIN
Président-rapporteur

Sylvie MADER
Directrice de recherche

Rafick-Pierre SÉKALY
Codirecteur

Claude PERREAULT
Membre du jury

Arndt BENECKE
Examinateur externe

Résumé

Cette thèse traite de l'utilisation de données large échelle pour élucider les mécanismes de réponse à un vaccin et identifier des biomarqueurs permettant de prédire une réponse immunitaire protectrice induite par la vaccination.

La vaccination s'est avérée très efficace pour prévenir et éradiquer certaines maladies infectieuses. Malgré cela, certaines personnes ne bénéficient pas de la vaccination. De plus, pour certains pathogènes, comme le VIH, le développement d'un vaccin efficace s'avère difficile. Les technologies large échelle fournissent un moyen d'évaluer le système immunitaire dans son intégralité et permettent d'évaluer l'impact de l'hôte et des facteurs environnementaux qui façonnent la réponse vaccinale. Dans cette thèse, nous montrons comment l'analyse bio-informatique de données large échelle a permis d'identifier des gènes et protéines inflammatoires comme marqueurs d'une faible réponse au vaccin contre l'HepB et contre le VIH, ainsi que comme marqueurs associés avec la sévérité des symptômes d'une infection virale respiratoire. De plus, nous démontrons comment l'utilisation de méthodes d'intégration d'ensembles de données provenant de multiples plates-formes multiomiques permet l'identification de mécanismes impliqués dans la modulation de la réponse vaccinale. En effet, une analyse intégrative a révélé le lien entre les cellules sanguines, l'hème et la réponse inflammatoire associée avec une réponse au vaccin contre l'HepB. Une analyse intégrative a également mené à l'identification des cellules dendritiques comme source des interférons et les lymphocytes CD4+ comme cible des interférons; les deux impliqués dans le mécanisme de réponse au vaccin RV144. Ainsi, l'utilisation des données large échelle combinée aux analyses bio-informatiques apportent de nouvelles solutions aux défis actuels de la vaccination.

Mots clefs : bio-informatique, vaccin, immunologie

Abstract

Vaccination has proven highly effective in preventing and eradicating infectious diseases. A significant number of subjects, however, respond poorly to vaccination. In addition, for some pathogens, including HIV, the development of an efficient vaccine has been challenging. Systems vaccinology provides a means to assess the immune system comprehensively and allows one to evaluate the impact of host and environmental factors shaping the vaccine response. This unbiased approach allowed us to identify inflammatory genes (ex. C3AR1, CXCL2, HEBP1) and proteins (ex. TNFβ, TNFR2) as putative markers of the hyporesponse to vaccines (against Hepatitis B and HIV) and severe symptoms following infection with respiratory viruses (Influenza, RSV, Rhinovirus). These biomarkers were measured pre-vaccination (respiratory viruses and Hepatitis B vaccine) or early after vaccination (HIV vaccine). In addition to identifying new biomarkers that can predict a. protective immune response to foreign pathogens, the integrative analysis of those system-wide data allowed a better understanding of the mechanisms triggered by vaccine-induced perturbations of the immune system. As illustrated for the RV144 vaccine against HIV, the integrative analysis allowed us to identify dendritic cells as the source and CD4+ T cells as the target cells of an antiviral program that confers the protection triggered by the vaccine. Systems vaccinology and the integration of datasets from multiple OMICs platforms represent a quantum step towards understanding immunity and vaccine responses; these approaches will provide novel solutions to the current vaccination challenges.

Keywords: bioinformatics, vaccines, immunology

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et des abréviations

ADN	Acide désoxyribonucléique
Acs	Anticorps
Ag	Antigène
ARN	Acide ribonucléique
ARNm	Acide ribonucléique messager
AUC	Aire sous la courbe ROC
CMH	Complexe majeur d'histocompatibilité
FN	Faux négatifs
FP	Faux positifs
HepB	Hépatite B
PCA	Analyse en composantes principales
PCR	Réaction de polymérisation en chaîne
ROC	*Receiver Operating Characteristic*
SIDA	Syndrome d'immunodéficience acquise
VIH	Virus de l'immunodéficience humaine
VN	Vrais négatifs
VP	Vrais positifs
VRS	Virus respiratoire syncytial

à Mounira et au Hajj

Remerciements

Je ne suis pas un homme de mille mots, mais je considère essentiel de remercier certaines personnes qui m'ont permis directement ou indirectement de réussir à terminer cette thèse de doctorat.

Sylvie M. et l'ensemble des membres du laboratoire du Dre Mader passés et présents,
J'ai réellement apprécié discuter de biologie avec vous, connaître vos routines. Cela nous permet d'apprécier les efforts que vous mettez dans votre travail, bonne continuation à tous.

Rafick-Pierre S. et les membres de son laboratoire,
Merci de m'avoir accueilli lors de vos réunions de laboratoire, de m'avoir inclus dans les activités scientifiques et ludiques. Grâce à vous, j'ai eu l'opportunité de côtoyer des experts en immunologie, informatique, mathématiques et statistiques; ce qui m'a permis d'améliorer ma formation.

Mes proches,
Je terminerai par vous remercier du fond du cœur. En premier lieu, mes parents à qui je dois tout. Sans leur soutien et les sacrifices qu'ils ont faits pour moi je ne serais sans doute jamais arrivé jusque-là. Je remercie également mon frère et mes amis de m'avoir soutenu pendant tout ce temps.

Introduction

Au Canada en 2019/2020, 3,898 infections par le virus de l'hépatite B (HepB) (1), 1,639 cas d'infection par le VIH (2) et 55,379 cas d'infection par le virus Influenza (3) ont été diagnostiqués. L'impact économique et humain de ces infections est important. La quasi-totalité de ces cas pourrait être évitée par l'utilisation d'approches préventives comme les vaccins ou par l'administration de thérapies antivirales spécifiques dès les premiers jours d'une infection. La réponse aux vaccins et à ces interventions thérapeutiques peut varier entre individu. Les outils cliniques actuels qui permettraient de personnaliser ces approches prophylactiques ou thérapeutiques avec pour but d'augmenter leur efficacité sont absents. Cette thèse à articles vise à démontrer que des approches bio-informatiques utilisant des données à large échelle peuvent remédier à ces lacunes.

1. Les virus

Cette thèse a comme objectif l'étude de la réponse immunitaire contre trois types de virus : l'HepB, HIV-1 et les virus respiratoires. Cette section décrit cette classe de pathogènes et détaille les différences entre ces trois types de virus (**Figure 1**).

Les virus sont des micro-organismes parasitaires qui répliquent en utilisant la machinerie cellulaire d'une cellule hôte. Ces virus sont classifiés en fonction de leur tropisme cellulaire, la structure de leurs acides nucléiques (ADN ou ARN; simple brin ou double-brin) et la présence ou l'absence d'une enveloppe virale. Nous décrivons ci-dessous cinq familles de virus capables d'infecter l'Homme et qui seront étudiés dans cette thèse.

1.1. Hépatite B

Le virus de l'HepB est un virus circulaire à ADN double-brin. Il est constitué de trois structures : 1) une enveloppe externe composée de lipides, d'hydrates de carbone et des protéines virales, formant l'antigène (Ag) de surface, 2) la capside, une structure interne formée de protéines (ces protéines sont appelées protéines de capside ou Cp, **Figure 1**) et 3) un l'ADN viral contenant l'information génétique nécessaire à la réplication du virus (4).

Le virus de l'HepB se transmet par contact avec le sang, le sperme ou la salive d'une personne infectée. Il existe trois principaux modes de transmission du virus de l'HepB : 1) la transmission de la mère à l'enfant, lors de la grossesse, de l'accouchement, ou de l'allaitement, 2) la transmission par voie sexuelle et 3) la transmission par voie sanguine lors de l'utilisation de

drogues, par transfusion sanguine, ou lors de procédures pouvant mener à un contact sanguin comme les dialyses, l'acupuncture ou les tatouages (5).

Le virus de l'HepB infecte principalement les cellules du foie (*i.e.* hépatocytes). Les évidences actuelles suggèrent que le virus de l'HepB a un tropisme pour les hépatocytes, ces cellules possédant à leurs surfaces les glycoprotéines d'héparane de sulfate HSPG reconnues par l'enveloppe du virus. Ces cellules expriment préférentiellement le récepteur présumé du virus de l'HepB, *NA+/Taurocholate Cotransporting Polypeptide* ou NTCP (encodé par le gène SLC10A1), point d'entrée du virus dans les hépatocytes (**Figure 2**) (6).

L'hépatite aiguë (phase initiale de l'infection) est souvent asymptomatique. L'hépatite chronique (quand elle persiste au-delà de six mois après l'infection initiale) engendre une désorganisation de l'architecture du foie conduisant à un dérèglement de son fonctionnement (*i.e.* cirrhose) et aboutissant au développement d'un carcinome hépatocellulaire (4).

Le diagnostic de l'HepB consiste en la détection de l'Ag de surface du virus de l'HepB (Ag HepB, présent sous trois formes, Ag HepB large, moyenne et petite) dans le sérum d'individus infectés. L'Ag HepB est détecté de 2 à 12 semaines après l'infection et disparait dans la majorité des cas après la 12^{e} semaine. En cas de persistance de l'Ag HepB à la 20^{e} semaine, le risque de chronicité de l'HepB augmente (7).

1.2. Virus de l'Immunodéficience Humaine

Le virus de l'immunodéficience humaine (VIH) est un virus à ARN simple-brin enveloppé. Il possède une enzyme, la transcriptase inverse, qui permet la transcription de l'ARN viral en ADN (8). L'ADN viral s'intègre dans les régions de chromatines accessibles du génome de la cellule hôte infectée (4). Le VIH se présente sous deux formes, le VIH-1 que l'on retrouve dans la très grande majorité des régions du globe et le VIH-2 principalement retrouvé en Afrique de l'Ouest (9).

Le VIH infecte principalement les lymphocytes T CD4+. En effet, ces cellules immunitaires expriment à leur surface le récepteur CD4 et le corécepteur CCR5 (10). Après la liaison de la forme trimérique de la glycoprotéine gp160 du VIH aux molécules CD4 et CCR5, le virus fusionne avec la membrane cellulaire et relâche ses acides nucléiques dans la cellule et utilise la machinerie cellulaire pour se multiplier avant de détruire (dans la majorité des cas) les lymphocytes T CD4+ infectés (11). Une fraction (entre 0.001% et 0.25%) de ces lymphocytes T CD4+ infectés persistent et deviennent un réservoir pour le VIH (12). Les lymphocytes T CD4+ sont un membre important du système immunitaire puisqu'ils fournissent les cytokines nécessaires au développement et à la différenciation des cellules innées et adaptatives de la

réponse immunitaire. Leur destruction par le VIH entraîne un déficit immunitaire cellulaire ayant pour conséquence une vulnérabilité accrue vis-à-vis de certaines infections opportunistes (ex. infection par le cytomégalovirus ou tuberculose (13)). Avec l'affaiblissement progressif du système immunitaire, ces infections opportunistes s'aggravent et s'accumulent résultant au syndrome d'immunodéficience acquise (SIDA) (14).

Le VIH est présent dans les fluides biologiques (sang, sperme, liquide séminal, et sécrétions cervico-vaginales, lait maternel) de personnes infectées. Ainsi, il existe trois principaux modes de transmission du VIH : (1) la transmission par voie sexuelle, (2) la transmission par voie sanguine, par les mêmes procédures que l'infection virale par le virus de l'HepB, (3) la transmission de la mère à l'enfant, lors de la grossesse, de l'accouchement, ou de l'allaitement (15).

La symptomatologie de l'infection aiguë par le VIH est difficile à distinguer d'autres infections virales comme la mononucléose et inclut des signes cliniques tels que la fièvre, la fatigue, une irritation cutanée ou une douleur musculaire. Ces symptômes ne persistent pas généralement au-delà de deux semaines (16). En effet, une infection par le VIH peut rester silencieuse pendant de nombreuses années (due au réservoir du VIH) tout en restant potentiellement transmissible.

L'infection par le VIH peut être diagnostiquée par la détection de l'Ag p24 (protéine de la capside virale), par la présence dans le plasma d'anticorps (Acs) dirigés contre plusieurs protéines virales ou encore par la quantification de l'activité de la transcriptase inverse du VIH. En cas de positivité, une analyse de confirmation visant à détecter l'ADN proviral et/ou l'ARN plasmatique est effectuée. Suite au diagnostic de l'infection par le VIH, il est recommandé de mesurer la quantité d'ARN viral dans le sang (qui croît proportionnellement avec le risque d'infections opportunistes) et le ratio de lymphocytes CD4+/CD8+ (qui décroît proportionnellement avec le risque d'infections opportunistes) (16).

1.3. Virus Influenza

Les virus respiratoires sont responsables d'approximativement 69% des pathologies respiratoires d'origine infectieuse (17). Les virus les plus souvent incriminés sont les virus influenza, les virus respiratoires syncytiaux (VRS), ou encore les rhinovirus.

Les virus Influenza regroupent les virus Influenza A, B et C. Le virus Influenza A et B sont responsables de la grippe saisonnière et le virus Influenza C est généralement bénin (18).

Les virus Influenza sont des virus à ARN simple-brin enveloppé. La structure des virus Influenza se compose de deux parties : (1) une enveloppe lipidique contenant les glycoprotéines de surface hémagglutinine et neuraminidase, dont il existe respectivement 16 et 9 sous-types. Chez l'homme, les sous-types H1, H2, H3, N1 et N2 sont majoritaires. Les sous-types H5 et H9 peuvent

émerger ponctuellement. L'hémagglutinine permet la liaison du virus aux acides sialiques présents à la surface des cellules cibles du virus entraînant la formation d'une cage de clathrines entourant le virus et permettant son endocytose dans la cellule infectée. La neuraminidase facilite la production de nouveaux virions en clivant la liaison entre les acides sialiques et l'hémagglutinine des virions (19).

Les virus Influenza sont transmis par voie respiratoire soit par l'émission de microgouttelettes aérosols ou par proximité avec une personne infectée (20).

Les virus Influenza infectent presque exclusivement des cellules de l'épithélium respiratoire. En effet, les cellules de l'épithélium respiratoire possèdent des acides sialiques à leur surface agissant comme les récepteurs pour les virus Influenza (21).

Les symptômes d'une infection par les virus influenza apparaissent généralement après une période d'incubation de 1 à 2 jours. Ces symptômes incluent une fièvre, des frissons, des maux de tête, des douleurs musculaires (*i.e.* myalgie) sévères et une anorexie. La plupart du temps, l'intensité des maux de tête, la myalgie et la fièvre déterminent la gravité de la maladie. Ces symptômes sont principalement accompagnés d'une symptomatologie des voies respiratoires telles qu'une toux sèche, des écoulements nasaux et des maux de gorge (22).

Typiquement, les infections par les virus Influenza sont diagnostiquées par leurs manifestations cliniques sans le recours à des tests de laboratoire. Dans des circonstances particulières, le diagnostic d'une infection par les virus Influenza nécessite la détection de l'ARN viral par un test de réaction en chaîne par polymérase (PCR), l'Ag viral par un test immuno-enzymatique (ELISA) ou, dans de rares cas, par des méthodes de culture cellulaire (22).

1.4. Virus respiratoires syncytiaux

Le virus respiratoire syncytial (VRS) est un virus enveloppé à ARN simple brin dont il existe 2 sous-types majeurs : le sous-type A est responsable de symptômes plus sévères que le sous-type B (23).

Le VRS se propage par voie respiratoire soit par l'émission de microgouttelettes aérosols ou par proximité avec une personne infectée. Le VRS peut également se propager par contact direct (sur la peau) ou indirect (surface d'objet) où le VRS peut demeurer pendant plusieurs heures (24).

Le VRS infecte principalement les cellules de l'épithélium respiratoire. Les cellules de l'épithélium respiratoire expriment à leur surface le récepteur nucléolin. Le récepteur nucléolin peut se lier aux glycoprotéines présentes dans l'enveloppe virale et sert de point d'entrée pour le virus (25).

Une infection par le virus VRS provoque un rétrécissement des voies respiratoires en raison de la sécrétion excessive de mucus et une obstruction des bronches et des bronchioles due à

l'inflammation locale des voies respiratoires. Une pneumonie résultant d'une infection par le virus VRS peut engendrer des œdèmes et une nécrose de la muqueuse des voies respiratoires (26).
Le diagnostic du VRS repose sur l'identification du virus dans les sécrétions respiratoires d'un patient. Quatre stratégies principales sont disponibles pour le diagnostic du VRS: culture cellulaire, PCR, des tests d'immunofluorescence directe et des tests de détection rapide d'antigènes viraux (27).

1.5. Rhinovirus

Les rhinovirus sont des virus nus à ARN simple brin. Les rhinovirus sont composés d'une capside qui contient quatre protéines virales (VP1, VP2, VP3 et VP4) et l'ARN viral (28).
Le rhinovirus peut-être transmis à la fois par la voie aérosol et par contact direct avec une personne infectée (29).
Les rhinovirus peuvent infecter les cellules ciliées de l'épithélium nasal, mais peuvent également infecter d'autres cellules nasales. L'intégrine ICAM-1 est le récepteur utilisé par le rhinovirus pour pénétrer dans les cellules épithéliales respiratoires (30).
Les rhinovirus sont le pathogène le plus fréquent associé aux symptômes d'un rhume, incluant l'écoulement nasal, le mal de gorge, la congestion nasale, les éternuements, la toux et les maux de tête (31).
Une infection par un rhinovirus est mise en évidence par des tests PCR ou, moins fréquemment, par culture cellulaire (32).

La prévention d'infections par les virus de l'HepB et du VIH-1 est l'objet des deux premiers articles de cette thèse. Les infections par les virus respiratoires incluant le virus Influenza font l'objet du troisième article de cette thèse. Les virus respiratoires induisent divers niveaux de symptômes après l'infection, et ce avec différentes cinétiques. Le troisième article de cette étudie les marqueurs de l'hôte qui sont commun à ces virus et qui sont associés avec l'apparition de symptômes décrits ci-dessus.

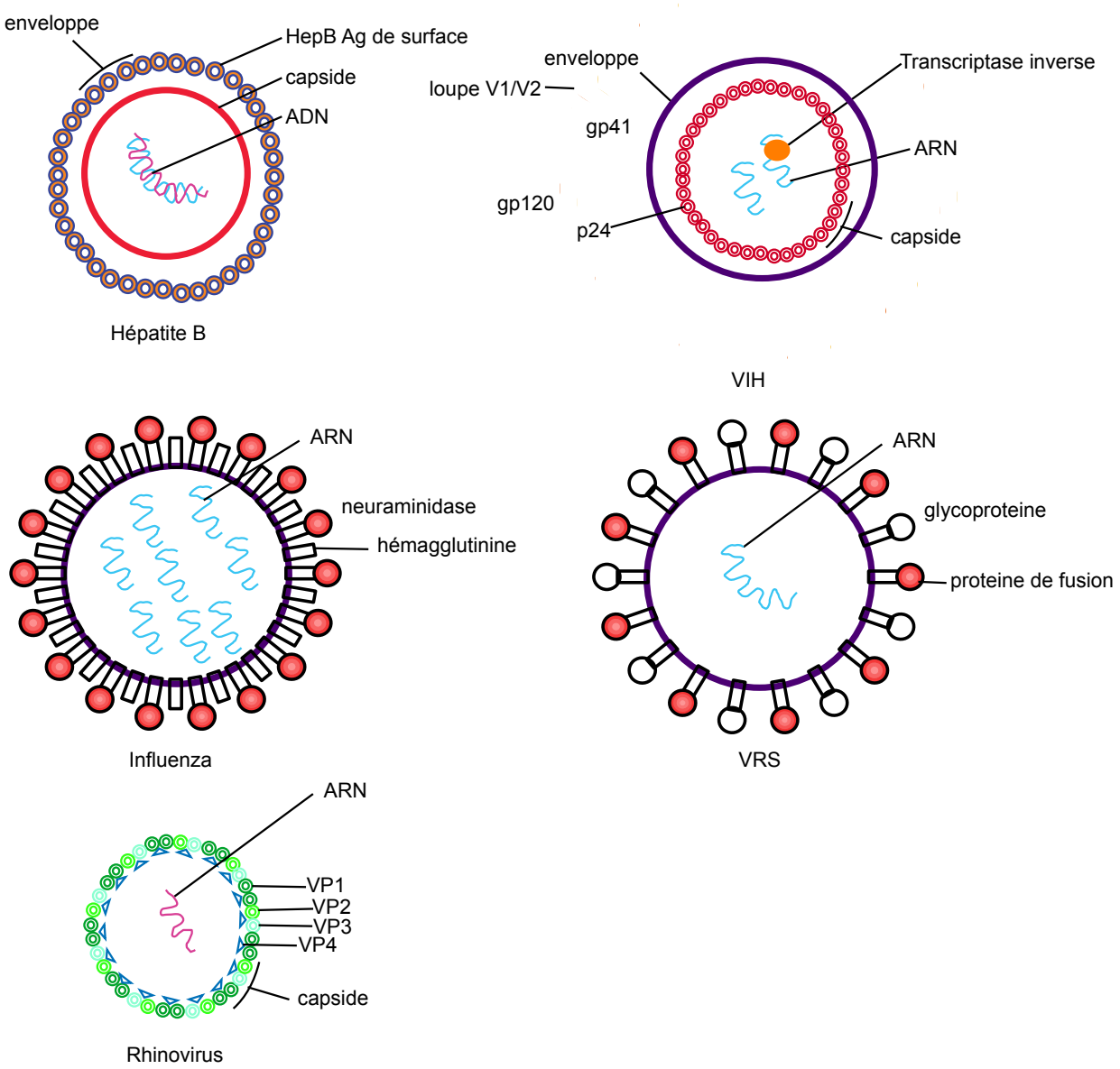


Figure 1. Représentation schématique des virus de l'HepB, VIH et les virus respiratoires

Les composantes structurelles des virus de l'HepB, VIH, Influenza, VRS et Rhinovirus sont indiquées dans la figure. La taille des virus dans la figure n'est pas à l'échelle réelle. Le diamètre de ces virus est 42 nm, 120 nm, 100 nm, 150 nm et 30 nm, respectivement.

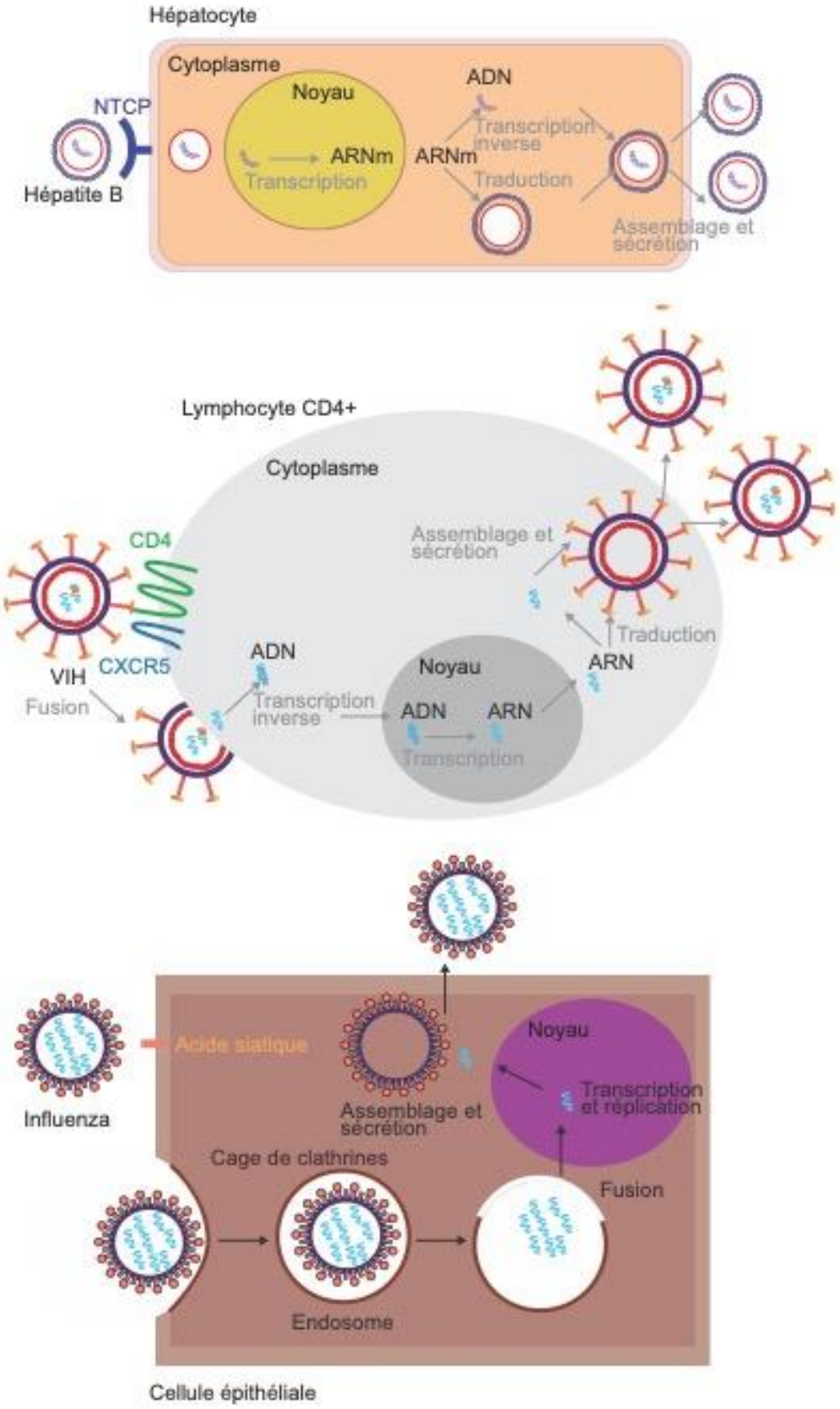

Hépatocyte
Cytoplasme
Noyau
ADN
NTCP
Transcription inverse
ARNm
ARNm
Transcription
Traduction
Hépatite B
Assemblage et sécrétion
Lymphocyte CD4+
Cytoplasme
Assemblage et sécrétion
CD4
Traduction
ADN
Noyau
ARN
VIH
CXCR5
Transcription inverse
ADN
ARN
Fusion
Transcription
Acide sialique
Noyau
Influenza
Assemblage et sécrétion
Transcription et réplication
Cage de clathrines
Fusion
Endosome
Cellule épithéliale

Figure 2. Cycle de réplication virale
(Haut) Cycle de réplication du virus de l'HepB. (Milieu) Cycle de réplication du virus de l'immunodéficience humain. (Bas) Cycle de réplication du virus Influenza.

2. Vaccins et adjuvants

Les vaccins sont des interventions prophylactiques ou à visée thérapeutique. Les vaccins visent à prévenir (vaccin prophylactique) ou à réduire les symptômes (vaccin thérapeutique) d'une infection bactérienne ou virale. Les vaccins ont été récemment utilisés pour prévenir certains cancers (33). Le but premier d'un vaccin est d'induire une immunité humorale et cellulaire spécifiques des molécules qui permettent l'entrée du virus dans les cellules cibles. Ceci réduit ou abolit l'infection (immunité stérilisante) lors d'une subséquente infection par le virus ciblé par le vaccin.

2.1. Vaccins

Les vaccins peuvent être classés en huit groupes; cette classification est basés sur le type d'immunogène utilisé: (1) les vaccins composés d'agents infectieux entiers vivants atténués, (2) les vaccins composés d'agents infectieux tués, (3) les vaccins à base d'anatoxines, (4) les vaccins conjugués combinant un Ag avec une faible immunogénicité à une protéine cargo à forte immunogénicité, (5) les vaccins composés de protéines virales recombinantes, (6) les vaccins composés de peptides synthétiques correspondant à des épitopes viraux immunodominants, (7) les vaccins utilisant des vecteurs viraux et (8) les vaccins à l'ADN ou l'ARN incluant des gènes codant pour des protéines du virus ciblé par le vaccin (**Table 1**).

2.1.1. Vaccins vivants atténués

Les vaccins vivants atténués ont été historiquement les premiers types de vaccins produits. Ils sont constitués de souches de virus ou de bactéries qui ont perdu leur pouvoir pathogénique, mais qui restent capables d'induire une réponse immunitaire. L'atténuation du pouvoir pathogénique d'un agent infectieux est obtenue par son passage sur des cultures cellulaires dans des conditions défavorables (ex. température compromettant la survie et la réplication du virus) ou en utilisant des techniques de virologie moléculaire (ex. insérant des mutations délétères dans le génome de l'agent infectieux) (34). L'utilisation de vaccins vivants atténués provoque une infection dénuée de manifestations pathologiques. Ces vaccins génèrent typiquement une forte réponse immunitaire (cellulaire et humorale) et une protection durable. Cette réponse immunitaire protectrice requiert le plus souvent une seule immunisation (34).

Un exemple de vaccin vivant atténué utilisé au Canada est le vaccin FluMist contre les virus Influenza contenant des souches atténuées des virus H1N1, H3N2 et Influenza B (35) (**Table 2**).

2.1.2. Vaccins tués

Les vaccins tués contiennent des agents infectieux entiers inactivés (tués). Ces vaccins sont de bons immunogènes capables d'induire une réponse humorale satisfaisante et protectrice (36).
Un exemple de vaccin tué utilisé au Canada est le vaccin contre le virus de l'HepB Infanrix (GlaxoSmithKline) contenant trois souches du virus de la poliomyélite inactivé (37) (**Table 2**).

2.1.3. Vaccins à sous-unités

Les vaccins sous-unités sont des vaccins composés d'antigènes isolés et purifiés à partir de l'enveloppe (ou membrane de surface) des agents infectieux et qui constituent les cibles des Acs. Ces antigènes sont constitués soit de polysaccharides, protéines membranaires ou de protéines capsulaires de l'agent infectieux.

2.1.4. Vaccins aux anatoxines

Une anatoxine est une molécule dérivée d'une toxine produite par un agent infectieux bactérien et responsable de la pathogénicité de la bactérie. Une anatoxine est obtenue en modifiant une toxine de sorte qu'elle soit dépourvue de de ses propriétés toxiques tout en conservant sa structure et ses propriétés antigéniques (38).
Un exemple de vaccin à base d'anatoxine utilisé au Canada est le vaccin contre le virus de l'HepB Infanrix (GlaxoKoSmith) qui inclut aussi les anatoxines dérivées des agents infectieux de la diphtérie et du tétanos (37) (**Table 2**).

2.1.5. Vaccins recombinants

Les cultures en laboratoire de certains agents infectieux (virus ou bactéries anaérobiques) peuvent souvent être complexes et ceci pose un problème pour le développement de vaccins. C'est le cas par exemple du virus HepB puisqu'il n'existe pas de lignées cellulaires permettant d'obtenir de hauts titres de virus. Grâce aux avancées en génétique, il est possible de produire des protéines recombinantes issues de ce pathogène. Ainsi, les gènes exprimant des antigènes capables d'induire une réponse immunitaire protectrice sont insérés dans un plasmide. Ce plasmide est ensuite introduit dans un vecteur (ex. une bactérie, une levure). Ces cellules expriment alors une molécule recombinante ayant conservé ses propriétés antigéniques (39).

Deux vaccins utilisés couramment au Canada contre l'HepB, Recombivax HB (Merck) et Engerix-B (GlaxoSmithKline) sont des exemples de vaccins recombinants. Ces deux vaccins sont produits en insérant le gène codant pour l'Ag de surface de l'HepB dans la levure (Saccharomyces cerevisiae). La levure produit l'Ag, qui est ensuite collecté et purifié. Ces deux vaccins recombinants contiennent 95% d'Ag de surface du virus de l'HepB et moins de 5% de protéines de la levure (aucun ADN de la levure n'est détecté dans ces vaccins) (40) (**Table 2**).

2.1.6. Les vaccins à ADN et ARN

Les vaccins à ADN ou ARN consistent à injecter de façon intramusculaire ou intradermique le gène codant ou l'ARN messager pour un Ag. Une fois au niveau des cellules, l'ADN/ARN est traduit en une protéine imitant l'Ag viral. L'utilisation de l'ADN/ARN a comme avantage d'être facile à construire et à produire à large échelle. L'ADN possède l'avantage additionnel d'être stables à température ambiante donc facile à transporter, conserver et distribuer. Les vaccins à ADN et ARN ne sont pas soumis au problème de réversion de l'agent infectieux vers la virulence comme les vaccins atténués, et typiquement se soldent en de plus fortes réponses immunitaires que les vaccins à sous-unités protéiques (41).

Les vaccins à ARN messagers ont récemment été testé chez l'homme pour palier à la pandémie COVID-19. Au paravent, des essais cliniques étaient prévus pour des vaccins à ARN messagers contre les virus Zika et VIH (42). Le séquençage d'isolats du virus SARS-CoV-2, publié en fin 2019, a permis aux compagnies Moderna et Pfizer de pivoter leur plateforme à ARN messagers contre SARS-CoV-2. Les vaccins à ARN messagers contre SARS-CoV-2 sont capables de prévenir une infection symptomatique avec une efficacité moyenne de 95% (43). L'efficacité de ces vaccins est meilleure ou comparable aux vaccins a vecteurs viraux (efficacité moyenne de 67%) (43). Les mécanismes conférant la protection par les vaccins à ARN messagers sont encore mal compris mais implique à la fois une réponse humorale et cellulaire aux vaccins (44). L'effet adjuvant des capsules liposomiques permettant de délivrer les ARN messagers est également mal compris.

2.1.7. Vaccins à vecteurs viraux

Les vaccins à vecteurs viraux utilisent un vecteur viral contenant un ou plusieurs gènes codant pour des antigènes. Le vecteur viral peut être (1) dépourvu de capacité de réplication ou (2) peut se répliquer tout en étant dépourvu de gènes causant la virulence/pathogénicité de l'agent infectieux. De nombreux virus ont été modifiés génétiquement afin d'être utilisés comme vecteurs de vaccination. Les principaux vecteurs viraux utilisés sont les adénovirus, les rétrovirus et les

poxvirus (incluant les canarypox). Ces vecteurs viraux ne sont pas toujours interchangeables; la taille des insertions introduites dans le vecteur, le tropisme cellulaire et l'immunogénicité varient d'un vecteur viral à un autre (45).

Un exemple de vaccin à vecteur viral est le vaccin RV144 contre le VIH. L'immunisation comprend quatre injections avec le vecteur viral ALVAC, un Poxvirus d'origine avienne ou trois gènes du VIH ont été insérés. L'administration de ce vecteur est suivie de deux injections de la forme monomérique de la glycoprotéine de l'enveloppe du VIH (gp120) (46) (**Table 2**).

2.1.8. Vaccins à pseudoparticules virales

Les vaccins à pseudoparticules virales sont des particules vaccinales formées de protéines recombinantes capables de s'assembler en une structure similaire à celle des particules virales. Pour produire les pseudoparticules virales, les gènes codant les protéines structurales sont clonés dans des bactéries ou levures puis exprimés *in vitro*.

2.1.9. Comparaison des différentes plateformes vaccinales

Les vaccins générés en utilisant les plateformes précédemment citées (**Table 1**) sont plus ou moins faciles à produire et confèrent des réponses immunitaires variées. En général, les vaccins atténués réplicatifs engendrent des réponses humorales et cellulaires plus importantes et durables quand on les compare aux vaccins atténués non réplicatifs. De même, les vaccins inactivés génèrent des réponses immunitaires moins importantes que les vaccins atténués s'expliquant par les antigènes perdurant moins longtemps chez l'individu vacciné et pouvant aussi avoir un épitope modifié comparé aux vaccins atténués. Certains vaccins sont plus faciles à produire tel que les vaccins à ADN ou à ARN qui peuvent être modifiés et déployés rapidement grâce aux avancées en termes de séquençage. À l'image de la pandémie COVID-19 qui a touché le monde fin 2019-debut 2020, les vaccins à ADN (Inovio) et à ARN (Moderna et Pfizer) ont été les vaccins les plus rapides à compléter les différentes étapes de fabrication et de test cliniques (pour les vaccins à ARN); plus rapidement que les vaccins à vecteurs viraux (Johnson et Johnson), à sous-unité de protéine virale (Novarax) ou à protéine recombinante (Sanofi) (47).

Table 1. Avantages et inconvénients des différents vaccins

Plateforme	Type de réponse immunitaire (48)	Avantages (47)	Inconvénients (47)
Vaccins inactivés	Humorale et cellulaire	Forte réponse immunitaire (plusieurs épitopes), moins de risque que les vaccins	Le processus d'inactivation du pathogène peut altérer l'épitope

		atténués, ne nécessite pas d'adjuvant	
Vaccins atténués	Humorale et cellulaire	Forte réponse immunitaire (plusieurs épitopes), conserve l'épitope original, simule une infection naturelle, ne nécessite pas d'adjuvant	Risque de virulence partielle (en particulier chez les personnes immunocompromises)
Vaccin à vecteur viral	Humorale et cellulaire	Forte réponse immunitaire, conserve l'épitope original, simule une infection naturelle	Production à large échelle compliquée, risque d'intégration génomique, réponse amoindrie chez les individus possédant des Acs contre le vecteur viral
Vaccin à sous-unité	Humorale	Bien toléré, conserve l'épitope original	Faible réponse immunitaire, nécessitent des adjuvants
Vaccin à particule	Humorale	Bien toléré, conserve la présentation naturelle de l'épitope	Faible réponse immunitaire, Production à large échelle compliquée,
Vaccin à ADN	Humorale et cellulaire	Bien toléré, peuvent être conservé à des températures ambiantes, facile à adapter à des pathogènes émergents, exprime l'Ag naturel	Faible réponse immunitaire (49), administration plus compliquée, risque d'intégration génomique
Vaccin à ARN	Humorale et cellulaire	Bien toléré, facile à adapter à des pathogènes émergents, exprime l'Ag naturel	Faible réponse immunitaire, nécessite d'être conservé et transporter a des températures froides, peuvent potentiellement induire une forte réponse interféron médié par RIG-1 et MDA5

2.2. Adjuvants

Un adjuvant est une molécule inerte qui exerce une activité immunostimulante, sans être elle-même immunogène. Un adjuvant est dépourvu d'activité toxique et est stable chimiquement. Les adjuvants potentialisent la réponse immunitaire, permettant ainsi d'obtenir des quantités plus élevées d'Acs persistant pour une plus longue durée, et ce en utilisant une quantité plus faible d'antigènes et un plus petit nombre de doses d'un vaccin. Les adjuvants peuvent être séparés en deux groupes majeurs : (1) les adjuvants inorganiques (ex. les adjuvants minéraux) et (2) les adjuvants organiques (dérivés de produits végétaux ou animaux).

Les adjuvants les plus largement utilisés sont les composés d'alumine et font partie des adjuvants inorganiques. Les composés d'alumine sont adsorbés avec les vaccins constitués d'anatoxines, de protéines recombinantes ou de peptides synthétiques. Le mécanisme d'action des composés

d'alumine n'est pas complètement résolu, ils formeraient un dépôt d'Ag au site d'inoculation ce qui engendrerait une réaction inflammatoire qui attirerait les cellules immunitaires (50).
Les dérivés du squalène, une substance présente à l'état naturel chez les plantes et les animaux, capable de stimuler la réponse immunitaire par la production de lymphocytes T CD4+ mémoires sont actuellement utilisé au Canada. Un exemple de ce type d'adjuvant est l'adjuvant MF59 de Norvartis (50).
Parmi les vaccins utilisés au Canada contre le virus de l'HepB, Recombivax HB, Engerix B et Infanrix sont tous formulés avec les adjuvants hydroxydephospate sulfate d'aluminium ou hydroxyde d'alumine (40). Un autre vaccin utilisant comme adjuvant hydroxyde d'alumine est le vaccin RV144 comprenant deux injections de la protéine du VIH gp120 formulées avec l'hydroxyde d'alumine (46) (**Table 2**).
Au Canada, le seul vaccin contre les virus Influenza formulé avec un adjuvant est le vaccin Fluad (Seqirus/CSL Limited), un vaccin à sous unité contant l'Ag de surface de l'hémagglutinine de trois souches des virus Influenza. Une seule dose du vaccin est administrée concomitamment avec l'adjuvant MF59, de la classe des dérivés du squalène (**Table 2**).

2.3. Stratégie de vaccination

2.3.1. Voies d'administration

Il existe trois voies principales d'administration des vaccins: intramusculaire, sous-cutanée ou intradermique. La voie d'administration dépend du type de vaccin administré et influence la qualité des réponses vaccinales et la fréquence et l'intensité des effets secondaires locaux éventuels (51). Les vaccins tués contenant des adjuvants sont préférentiellement injectés par voie intramusculaire, car ils engendrent des effets secondaires moins importants en comparaison avec une administration par voie sous-cutanée (51). Les vaccins inactivés sans adjuvant, comme les vaccins contre les virus Influenza, sont bien tolérés, quelle que soit la voie d'administration. Pour les vaccins atténués, la voie d'administration ne semble jouer aucun rôle (52).

2.3.2 Espacement de l'administration de doses vaccinales

Le nombre d'immunogènes, d'immunisations et l'intervalle entre les immunisations d'un vaccin peuvent avoir un impact sur l'efficacité du vaccin à éliciter une protection contre un pathogène. En effet, certains vaccins sont capables de conférer une protection dès la première injection (ex. les vaccins contre la diphtérie et le tétanos) alors que d'autre nécessite multiples doses (ex. le vaccin contre le virus de l'HepB Twinrix nécessite trois immunisations données dans un intervalle d'au moins un mois après chaque immunisation) (53).

La formulation d'un vaccin avec un adjuvant influence aussi l'espacement des doses vaccinales. Le but de l'adjuvant est de rendre le vaccin plus immunogène afin de générer une réponse protectrice en utilisant une dose moindre de l'Ag ou éviter la nécessité d'immuniser à multiple reprise un individu (54).

La combinaison de plusieurs immunogènes ou de plusieurs vecteurs administrés séquentiellement dans la stratégie « prime-boost » peut améliorer la réponse à un vaccin. Cette stratégie est fréquemment utilisée pour des pathogènes capables d'esquiver le vaccin tel que VIH (55).

2.4. Effets secondaires des vaccins

Les vaccins ne sont pas dépourvus d'effets secondaires. Ces symptômes incluent fièvre, fatigue, irritation locale, douleur musculaire, maux de tête. Ces symptômes disparaissent rapidement et ne laissent pas de séquelles (56). Les vaccins incitant des effets secondaires sévères causant une invalidité temporaire ou définitive sont typiquement rejetés lors des essais cliniques et ne sont pas commercialisés (57).

Table 2. Vaccins contre les virus de l'HepB, Influenza et VIH

Nom du vaccin	Pathogène	Type de vaccin	Adjuvant	Pourcentage de protection	Reference
Engerix-B©	HepB	Recombinant	Hydroxyde d'aluminium	81.3%	(58)
Twinrix©	Hépatite A/B	Recombinant	Hydroxyde d'aluminium et aluminium phosphate	95.1%	(59)
Infanrix Hexa©	Hépatite A/B, diphtérie, tétanos, polio	Recombinant	Hydroxyde d'aluminium	100%	(60)
Recombivax HB©	HepB	Recombinant	Aluminium hydrophosphate sulfate	98.5%	(61)
ALVAC-gp120	VIH	Vecteur viral	Hydroxyde d'aluminium	31.2 %	(46)
Flulaval©	Influenza	Tué	Aucun	64.8%-85.5%	(62)
Fluzone©	Influenza	Tué	Aucun	72.8%-96.5%	(63)
Afluria©	Influenza	Tué	Aucun	31.0%-40.9%	(64)
Influvac©	Influenza	Sous-unité	Aucun	50.0%-99.4%	(65)
FluMist©	Influenza	Attenué	Aucun	5.0%-42.4%	(35)
Agriflu©	Influenza	Sous-unité	Aucun	87%-99%	(66)
Fluviral©	Influenza	Tué	Aucun	92%-98%	(67)
Fluad©	Influenza	Sous-unité	MF59	35%-94%	(68)

2.5. Obstacle aux développements de nouveaux vaccins

Certains pathogènes mutants rapidement tels que le VIH et le virus influenza rendent difficile l'élaboration de vaccins prophylactiques. Le VIH possède un obstacle supplémentaire, car les cellules cibles du vaccin sont les cellules CD4+ activées qui elles-mêmes sont nécessaire pour établir une réponse humorale contre le virus.

Les stratégies vaccinales étudiées pour pallier aux problèmes cités ci-dessus incluent l'utilisation de nouveaux des vecteurs viraux (ex. Adénovirus 26 utilisé dans le vaccin Mosaico testé durant l'essai clinique HVTN706) de nouvelles plateformes vaccinales comme les vaccins à ARN (ex. Moderna); l'utilisation de la biologie structurale pour la conception d'immunogènes (l'utilisation de plusieurs antigènes comme le vaccin Mosaico); et l'émergence de nouveaux adjuvants (ex. l'adjuvant ALFQ du centre de recherche de l'armée américaine) (69).

La stratégie vaccinale la plus prometteuse pour ces virus à fort taux de mutation consiste en l'administration d'Acs reconnaissant un large spectre de variants viraux. Ces Acs ont toutefois une courte demi-vie une fois administrés. Des stratégies alternatives visent à entraîner et stimuler les lymphocytes B de la moelle osseuse à produire ces Acs directement. À ce jour, ces stratégies n'ont pas démontré leur efficacité dans des essais cliniques de phase III (70).

2.6. Approches prophylactiques et thérapeutiques antivirales

2.6.1. Traitement pour l'HepB

Les méthodes de prévention contre l'HepB autre que les vaccins incluent le dépistage du virus (dans les donations de sang, plasma, organes, tissus et sperme), les programmes d'échanges de seringues et les campagnes d'éducation sur la vaccination (71).

La thérapie recommandée pour une personne ayant été exposée au virus de l'HepB est le vaccin contre l'HepB ou être traité avec l'immunoglobuline anti-HepB (IgHB) dans les deux semaines après une exposition. IgHB fournit une protection immédiate, mais de courtes durées contre l'infection à l'HepB. L'IgHB contient une grande quantité d'Acs contre le virus de l'HepB provenant de dons de sang (72).

Si une infection par l'HepB est confirmée, le traitement pour une HepB aiguë consiste à traiter les symptômes de la maladie. Le traitement de l'HepB chronique vise à diminuer la quantité de virus circulant à des niveaux indétectables, prévenir la progression de la maladie hépatique vers la cirrhose; tout en tentant d'améliorer de la survie et de la qualité de vie des patients. Le traitement de l'HepB chronique repose sur les analogues nucléotidiques (ex. lamivudine, un analogue de la cytosine, agissant principalement par inhibition de l'activité de la polymérase du l'HepB entraînant une diminution de la réplication virale) et l'interféron alpha (73). L'efficacité limitée de ces

traitements, leur disponibilité et leur coût limitent leur utilisation et sont des arguments pour le développement et l'utilisation de vaccins.

2.6.2. Traitement pour le VIH

Les méthodes de prévention du VIH incluent le dépistage du virus, la circoncision, les programmes de distribution de contraceptifs, d'échanges de seringues et les campagnes d'éducation sur les risques de transmission du VIH (74).

Si une infection par le VIH est confirmée, un traitement antirétroviral (ART) est recommandé. L'objectif est de restaurer ou de maintenir un taux de lymphocytes CD4+ supérieur à 500/mm3 tout en rendant la charge virale VIH plasmatique indétectable (inférieure à 50 copies par millilitre). ART est une thérapie composée de trois molécules antirétrovirales, habituellement deux inhibiteurs nucléotidiques/nucléotidiques associés à un inhibiteur non nucléotidique ou un inhibiteur de protéase (75). Les traitements antiviraux sont utilisés avec une administration quotidienne par voie orale, mais doivent être pris durant la vie entière. En effet les traitements antiviraux ne sont pas des traitements curatifs pour le SIDA car ils ne permettent pas d'éliminer le réservoir du VIH (car les cellules du réservoir du VIH ont une faible activité transcriptionelle et proliférative et sont situées dans des tissues difficilement accessibles par les traitements antiviraux). En cas de SIDA et/ou d'immunodépression profonde, des mesures préventives telles que la prophylaxie des infections opportunistes sont préconisées (76).

Pour accélérer la lutte contre le VIH/SIDA, une nouvelle stratégie de la prévention a été développée : la Prophylaxie Pré-Exposition au VIH ou PrEP. Le principe de la PrEP repose sur l'utilisation, par des personnes non infectées par le virus, d'un ou de plusieurs traitements antirétroviraux pour diminuer le risque de transmission du virus en cas d'exposition (77). Il existe différentes formes de PrEP: PrEP locale utilisant des gels vaginaux et/ou anals, des anneaux vaginaux, PrEP injectable et PrEP orale. La PrEP administrée par voie orale par Truvada est la plus avancée (78) et est capable de réduire 70% à 90% des infection par le VIH (79). Elle a été approuvée au Canada depuis février 2016. Les autres formulations de PrEP sont encore à des aux stades expérimentaux de leur développement. À l'heure actuelle, aucune agence de réglementation dans le monde n'a approuvé l'usage d'autres formes de PrEP (incluant le Canada).

Le coût de la PrEP est remboursé par la plupart des régimes d'assurance maladie publics et certains régimes privés au Canada. Sans assurance, les coûts de la PrEP quotidienne s'élèvent à environ 220$ par mois (80).

2.6.3. Traitement pour la grippe

La méthode la plus efficace pour la prévention d'une infection par les virus Influenza est la vaccination. Les virus VRS et Rhinovirus n'ont pas de vaccin, le développent de vaccins contre ces virus s'avère difficile en raison de la variabilité antigénique de ces virus (81). Un vaccin contre les virus VRS ciblant la protéine F du virus (la protéine F permet au virus de fusionner et d'infecter les cellules épithéliales respiratoires) est couramment testé dans un essai clinique de phase 3 (82). Les autres méthodes de prévention d'une infection par les virus Influenza, VRS ou Rhinovirus consistent au dépistage du virus, l'utilisation de certains antiviraux et l'isolement des personnes infectées/symptomatiques pour réduire la propagation des virus respiratoires (83).

Les personnes diagnostiquées pour une infection par les virus Influenza vont typiquement recevoir des antiviraux (ex. l'inhibiteur de la neuraminidase Tamiflu). Les antiviraux contre les virus Influenza ont pour objectif de prévenir les formes sévères de la grippe tout en inhibant la réplication virale. Trois classes d'antiviraux spécifiquement actifs contre le virus influenza existent (1) les inhibiteurs de la pompe à proton virale M2 qui limitent la réplication virale une fois que le virus a pénétré dans la cellule cible en bloquant le canal ionique viral M2 et la libération du matériel génétique viral nécessaire à sa réplication dans la cellule infectée (83); (2) les inhibiteurs de la neuraminidase, des analogues structurels de l'acide sialique qui constituent le substrat de la neuraminidase. Ils ont pour rôle de limiter la dissémination du virus en inhibant la neuraminidase, impliquée dans la séparation entre des virions néoformés et leur attache sur la membrane cellulaire (83); (3) les inhibiteurs de la polymérase acidique du virus influenza permettant aux virus d'utiliser l'ARN messager des cellules infectées comme amorce pour la transcription de l'ARN messager viral (84). Certaines souches des virus Influenza peuvent être résistantes aux médicaments antiviraux, d'où l'importance de prévenir (par la vaccination) la dissémination des virus Influenza.

Les traitements pour une infection par le VRS ou Rhinovirus sont typiquement des médicaments atténuant les symptômes de l'infection virale (ex. acétaminophène ou l'ibuprofène pour réduire la fièvre et les douleurs causées par l'infection) (81).

La réponse au vaccin recombinant Twinrix donné avec l'adjuvant Alum est l'objet du premier article de cette thèse. Le vaccin à vecteur viral RV144, le seul vaccin ayant montré une efficacité pour prévenir les infections par le VIH fait l'objet du deuxième article de cette thèse. De plus, le deuxième article compare les profiles transcriptionnels, protéiques, cellulaires, et humorales pour identifier des marqueurs modulés après la vaccination et liés avec la réduction du risque d'infection conféré par le vaccin. Cet article souligne entre autres le lien étroit entre la réponse

immunitaire innée aux vaccins (liés aux cellules dendritiques), l'induction de la réponse aux interférons et la réponse cellulaire CD4+ importante pour protéger les individus vaccinés.

3. Mécanismes de la réponse aux vaccins

L'objectif de la vaccination est de permettre à l'individu vacciné de développer une réponse mémoire (i.e. une réponse immunitaire de plus forte amplitude et plus rapide que la réponse immunitaire qui suit le premier contact avec un pathogène) spécifique à un agent infectieux. L'introduction d'un Ag du « non-soi » (Ag viral, Ag bactérien…) chez l'hôte déclenche deux types de réponses immunitaires (1) une réponse immunitaire innée (immédiate) qui implique principalement les cellules myéloïdes (monocytes/macrophages et cellules dendritiques) et cellules NK; et (2) une réponse immunitaire spécifique ou adaptative (qui suit la réponse innée d'une semaine), qui peut être soit humorale (lymphocytes B producteurs d'Acs), cellulaire (lymphocytes T CD8+ tueurs ou CD4+ auxiliaires) ou, le plus souvent, les deux menant à l'élaboration d'une réponse immunitaire pouvant éliminer l'agent infectieux. La réponse adaptative repose sur une reconnaissance par le système immunitaire de l'Ag par le biais de récepteurs spécifiques (les récepteurs BCR à la surface des lymphocytes B et les récepteurs TCR à la surface des lymphocytes T) menant à l'activation et la différenciation des cellules exprimant ces récepteurs. Trois sous-types de cellules orchestrent la réponse immunitaire soit les cellules présentatrices d'Ag (les cellules dendritiques et les monocytes/macrophages), les lymphocytes (B et T) et les cellules lymphoïdes innées (NK, cellules lymphoïdes innées, **Figure 3** et **Figure 4**).

3.1. Réponse innée

Les cellules présentatrices d'antigènes ont pour fonction d'apprêter l'Ag pour promouvoir sa reconnaissance par les lymphocytes T CD4+ et CD8+.

3.1.1. Cellules dendritiques

Les cellules dendritiques sont des cellules d'origine hématopoïétique jouant un rôle fondamental dans le développement des réponses immunitaires spécifiques à des antigènes du « non-soi ». La principale fonction des cellules dendritiques est d'apprêter et de présenter un Ag aux lymphocytes T; pour cette raison les cellules dendritiques font partie de la famille de cellules présentatrices d'Ag professionnelles. On distingue les cellules dendritiques immatures et les cellules dendritiques matures (qui peuvent être sous-divisées en cellules dendritiques conventionnelles qui arrivent à maturité dans les organes lymphoïdes ou cellules dendritiques

plasmacytoïdes qui arrivent à maturité dans la moelle osseuse; les cellules dendritiques et plasmacytoïdes peuvent à leur tour être séparés en multiple sous-groupes de cellules basés sur leurs profils transcriptionnels (85)). Les cellules dendritiques immatures sont spécialisées dans la capture de l'Ag et sa dégradation en peptides antigéniques. Ces derniers sont ensuite chargés sur le complexe majeur d'histocompatibilité de classe I (peptides endogènes ou exogènes provenant de protéines intracellulaires) ou de classe II (peptides exogènes ou endogènes provenant de protéines membranaires ou sécrétées). Les cellules dendritiques immatures ne sont pas capables de stimuler les cellules T de façon efficace, car elles n'expriment que de faibles niveaux de molécules stimulatrices des lymphocytes T (86). Les cellules dendritiques immatures sont également impliquées dans la tolérance aux antigènes du « soi » (87). Quant aux cellules dendritiques matures, elles stimulent les lymphocytes T naïfs et B naïfs qui circulent dans les organes lymphoïdes, sécrètent des cytokines agissant sur les lymphocytes et amorcent la réponse immunitaire spécifique en leur présentant le peptide lié au complexe majeur d'histocompatibilité de classe I ou de classe II (86, 88). Les cellules dendritiques peuvent également présenter des antigènes aux lymphocytes T mémoires sans avoir à co-stimuler ces cellules (la reconnaissance des antigènes par les cellules mémoires ne requiert pas de co-stimulation) (89).

3.1.2. Monocytes et macrophages

Les monocytes (du sang) et macrophages (dans les tissus) sont des cellules initiatrices de la réponse immunitaire innée. Leur rôle est majeur dans la dégradation de l'Ag en peptides et sa présentation aux lymphocytes T. Ils participent à la réponse immunitaire grâce à la synthèse de nombreux produits de sécrétion (les cytokines) qui sont des médiateurs biologiquement actifs sur les lymphocytes T : ils produisent en particulier certaines cytokines nécessaires à l'initiation de la réponse immunitaire comme l'interleukine 1 (IL–1) qui active les cellules T, tandis que d'autres cytokines modulent la polarisation de la réponse immunitaire, par exemple l'IL–10 et l'IL–12. Ils interviennent également comme modérateurs de la coopération entre les lymphocytes T et B. À l'inverse, les macrophages reçoivent des informations des lymphocytes T toujours par l'intermédiaire des cytokines qui confèrent aux macrophages une activité cytolytique ou suppressive (90). Enfin, les macrophages peuvent démontrer une activité cytotoxique et tuer des cellules infectées par des virus en produisant des molécules endogènes comme le monoxyde d'azote (NO) et la cytokine TNF, en englobant les cellules a tuée ou en recrutant d'autres cellules capables de tuer les cellules infectées (91).

3.2. Réponse immunitaire cellulaire

3.2.1 Lymphocytes T

Les lymphocytes incluent les sous-populations cellulaires responsables de l'immunité adaptative. Ils tirent cette spécificité de l'existence de récepteurs spécifiques de l'Ag sur leur surface membranaire. La diversité des récepteurs à la surface des lymphocytes a été acquise à travers un mécanisme de réarrangement génique et de sélection positive/négative dans le thymus. On distingue deux sous-populations : les lymphocytes issus de cellules souches originaires de la moelle osseuse, mais dont la maturation dépend du thymus (lymphocytes T) et les lymphocytes qui se différencient dans la moelle osseuse (lymphocytes B). Selon leur durée de vie, il existe deux sous-types de lymphocytes: ceux ayant une courte durée de vie, en moyenne quatre à cinq jours, et ceux à durée de vie longue, dite « lymphocytes mémoires » qui jouent un rôle important dans les réponses secondaires après la vaccination. Les lymphocytes T jouent un rôle essentiel dans la régulation des réponses immunitaires, en coopération étroite avec les autres acteurs cellulaires principalement les lymphocytes B et interviennent de façon capitale au cours des réponses humorales. Le récepteur T à l'Ag (ou TCR) exprimé à la surface membranaire des lymphocytes T est doté d'une variabilité importante (10^{15} combinaisons) ce qui lui permet de reconnaitre un nombre important de peptides antigéniques (9 à 15 acides aminés) qui sont présentées par les molécules du Complexe majeur d'histocompatibilité (CMH) du « soi » et qui sont exprimées à la surface des cellules présentatrices d'antigènes. Le récepteur T est constitué d'un module de reconnaissance constitué par l'assemblage du TCRα et TCRβ. Ce module de reconnaissance est associé aux chaines du complexe CD3 qui sont responsables de l'assemblage du TCR et de son transport à la surface cellulaire. Les chaines du complexe CD3 sont aussi responsables de la signalisation par le TCR. La reconnaissance par le récepteur TCR à la surface des lymphocytes T du complexe CMH–peptide antigénique active les lymphocytes T et, grâce à l'action de l'IL–2, ces cellules se divisent et se différencient en cellules effectrices qui exercent des fonctions immunitaires (cytotoxicité, production de cytokines). Les lymphocytes T matures comportent deux sous-populations essentielles, qui se distinguent par l'expression de deux récepteurs exclusifs : CD4 ou CD8. Schématiquement, les cellules CD4+ dites « auxiliaires » ont une fonction régulatrice d'amplification des réponses immunitaires, par leur capacité à produire de grandes quantités de diverses cytokines. En fonction du profil de cytokines produit, on les subdivise en cellules Th1 (T helper de type 1 impliquée dans l'immunité cellulaire), Th2 (T helper de type 2 impliqué dans les réactions humorales), et autres (Th21, Th9, Tregs, Th17), au stade ultime de leur différenciation. Les cellules CD8+ produisent également des cytokines. L'activité cytotoxique (i.e. tuer les cellules infectées par les virus) est la fonction principale

attribuée aux cellules CD8+ (92). Il existe d'autres sous-populations de lymphocytes, tels que les lymphocytes T exprimant le récepteur gamma delta. Ces cellules sont rares dans le sang et sont principalement trouvées dans les tissus périphériques (ex. muqueuse intestinale). Leur rôle est encore mal compris à la fois dans la réponse immunitaire et la réponse vaccinale.

3.2.2. Cellules NK

Les cellules Natural Killer (NK) sont une composante clef de la réponse immunitaire cellulaire. Les cellules NK sont recrutées rapidement après une infection virale. Les cellules NK activées sont capables de tuer les cellules cibles. Les cellules NK interagissent avec la molécule HLA-E via le récepteur NKG2A (menant à leur inhibition) et le récepteur NKG2C (menant à leur activation). Les cellules NK activées produisent l'IFN-γ. qui participe à l'orientation de la réponse immunitaire adaptative en activant les cellules dendritiques. La sécrétion de chimiokines inflammatoires lors d'une infection virale ou bactérienne permet la co-localisation des cellules NK avec d'autres cellules hématopoïétiques, comme les cellules dendritiques. L'interaction des cellules NK avec les cellules dendritiques et la production d'IFN-γ par les cellules NK induisent l'expression chez ces cellules dendritiques de molécules de surface qui facilitent la formation de synapses immunologiques (ex. intégrines, molécules CD80 et CD86) ainsi que de la machinerie intracellulaire responsable de la présentation de l'Ag (ex. le complexe immunoprotéasome, molécules du système majeur d'histocompatibilité de type I ou II). Les cellules NK contribuent ainsi à façonner la réponse adaptative exercée par les lymphocytes T et B. En tuant des cellules infectées et stressées, les cellules NK participent aussi au développement de la réponse adaptative en fournissant des débris cellulaires qui peuvent être cross présentés aux lymphocytes T CD8+ cytotoxiques (93) par les cellules dendritiques.

3.3. Réponse humorale

3.3.1. Lymphocytes B

Les lymphocytes B sont les cellules effectrices de l'immunité humorale par leur capacité de produire des Acs spécifiques du pathogène. Leur récepteur de reconnaissance pour l'Ag est formé des CD79 (fragment de transduction du signal) et d'une immunoglobuline (fragment de liaison au ligand) exprimée à la membrane lymphocytaire (BCR). Au contact de l'Ag, les cellules B quiescentes se différencient dans les centres germinatifs des organes lymphatiques secondaires (ex. la rate et ganglions lymphatiques) en plasmocytes qui sont hautement spécialisés dans la synthèse et l'excrétion des immunoglobulines ou Acs. Les lymphocytes B sécrètent en fonction des cytokines en circulation, différents types d'immunoglobuline (Ig) : IgM,

IgG, IgA, IgD, IgE. Il existe des interrelations étroites entre les lymphocytes T et B. Une fois activés par l'Ag, les lymphocytes T auxiliaires spécifiques vont être recrutés par les lymphocytes B qui peuvent aussi présenter l'Ag spécifique. Leur interaction déclenche la production de diverses cytokines et l'expression de récepteurs de membranes spécialisés (par exemple, le ligand de CD40) à la surface de la cellule T auxiliaire (Th2 et T folliculaires). C'est l'activation des lymphocytes B, résultant de l'action conjointe des cytokines et des signaux traduits par les récepteurs membranaires spécialisés, qui aboutit à leur prolifération et leur différenciation en cellules mémoires ou productrices d'Acs (94).

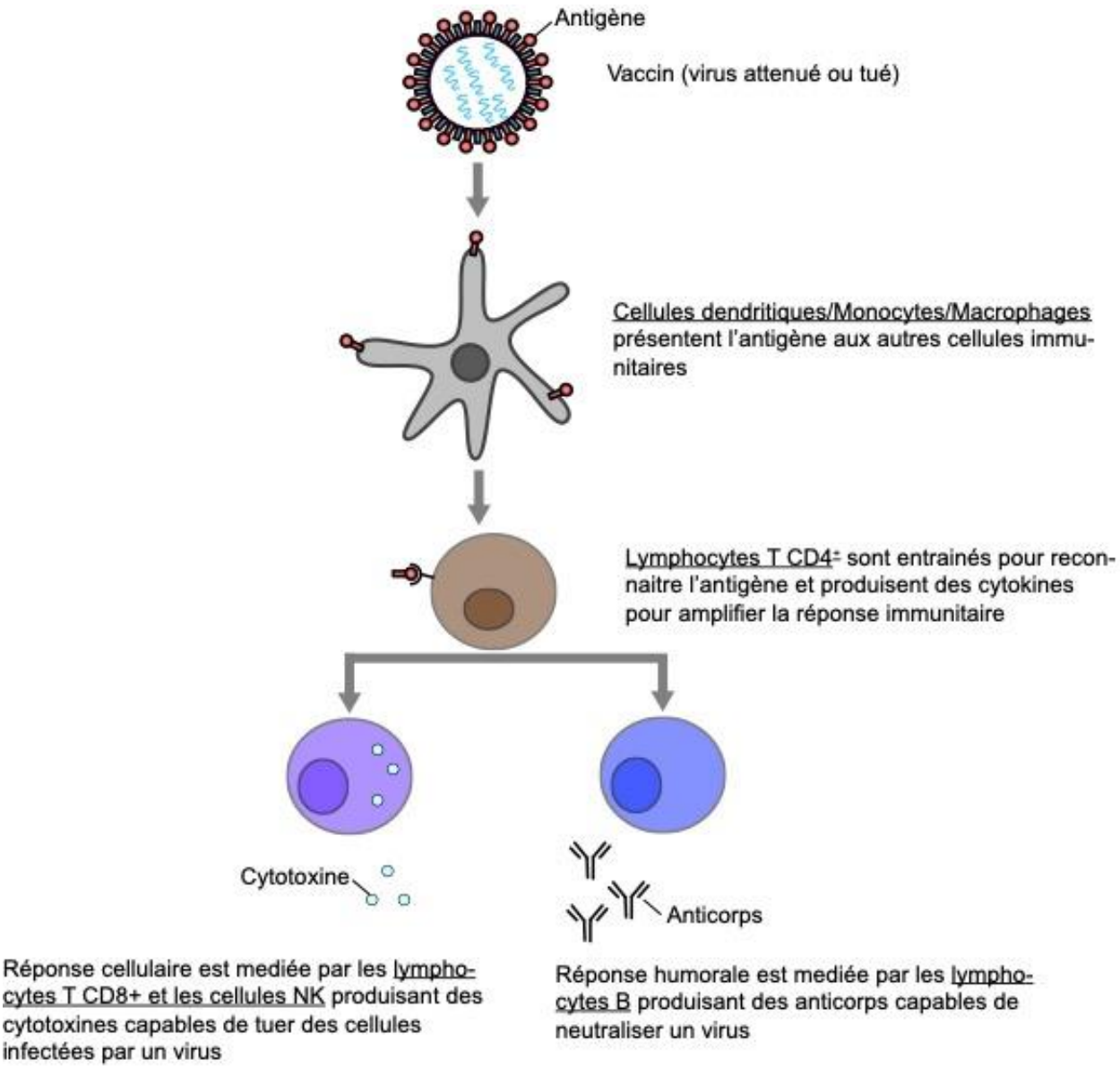


Figure 3. Représentation schématique des cellules immunitaires impliquées dans la réponse à un vaccin

La nature et la fonction des cellules immunitaires impliquées dans la réponse innée et adaptative à un vaccin sont indiquées dans la figure.

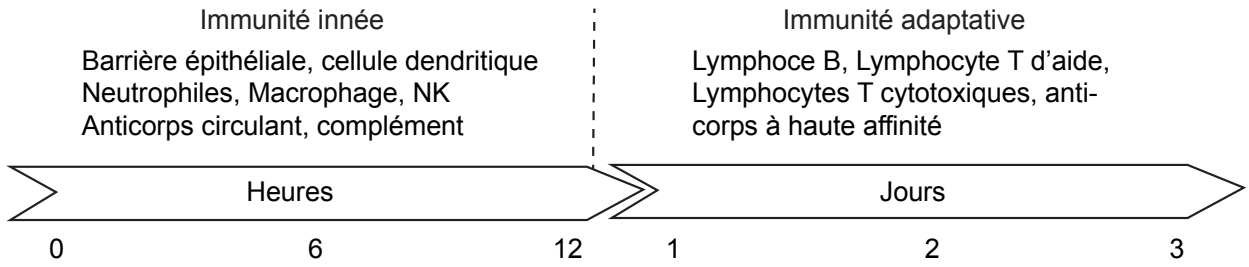


Figure 4. La dynamique de la réponse immunitaire aux infections virales et aux vaccins

3.3.2. Formation des anticorps

La première administration d'un vaccin entraîne, après une période de latence plus ou moins longue, la production d'Acs à un taux faible. L'administration ultérieure du même Ag induit une réponse plus rapide et de plus forte amplitude tel que mesurée par des titres d'Acs plus (de 10 fois plus) élevés; cette réaction est appelée la réponse mémoire et est attribuée à la présence de cellules sensibilisées ayant gardé la mémoire antigénique (**Figure 5**).

La réponse primaire est le résultat d'un ensemble d'interactions entre lymphocytes T et B observées après la première injection vaccinale par opposition aux réactions secondaires, qui sont observées lors de la répétition des injections. On distingue trois phases qui suivent une première injection vaccinale; ce sont: la période de latence, la période de croissance et la période de contraction (ces trois périodes surviennent également pour les lymphocytes T).

La période de latence est l'intervalle de temps qui s'écoule entre l'administration du vaccin et l'apparition des Acs. Elle varie entre vingt-quatre heures et deux semaines, et dépend de plusieurs facteurs environnementaux et épigénétiques (ex. niveau et type d'inflammation, répertoire des lymphocytes T et B), ainsi que de la nature (Ag protéique ou polysaccharide), de la forme (Ag monomérique tel que les anatoxines versus Ag multimérique comme les pseudo-particules virales (95)) et de la dose de l'Ag utilisé.

La période de croissance qui suit la période de latence est caractérisée par l'augmentation exponentielle des titres d'Acs qui sont le résultat de la prolifération des lymphocytes B spécifiques à l'Ag qui a lieu dans les centres germinatifs dans les ganglions lymphoïdes, juxtaposés aux lymphocytes T Tfh qui produisent l'IL-21 (IL21 est une cytokine essentielle à la prolifération des lymphocytes B). Les titres d'Acs atteignent un plateau en un temps variable allant de quatre jours à quatre semaines. La production d'Acs IgM précède celle des Acs IgG lors d'une réponse primaire. Les titres d'Acs plafonnent à un niveau élevé pendant quelques jours puis, en l'absence de nouvelle stimulation antigénique, ces titres décroissent rapidement puisque la demi-vie plasmatique des Acs est de 21 jours.

La période de contraction de la réponse immunitaire est définie par la diminution progressive des titres d'Acs. La durée de cette période de contraction varie en fonction du taux de synthèse des Acs, de leur demi-vie qui est différente selon l'isoforme d'Acs et du titre final des Acs. On note que les IgA et les IgM décroissent plus rapidement que les IgG. Le déclin des Acs plasmiques lors de la période de contraction peut être expliqué par une réduction des plasmocytes. Des faibles niveaux des Acs sont détectables à la fin de la période de contraction, ces Acs demeurent détectables plusieurs mois/années après la vaccination; ceci peut être expliqués par la présence de lymphocytes B mémoire et plasmocytes à longue survie.

Une nouvelle exposition à l'Ag déclenche pour les antigènes de nature protéique, une réponse de type secondaire. Cette réponse se caractérise à la fois par la production rapide des Acs spécifiques et la quantité importante des Acs sécrétés qui sont de type IgG. Les titres maximums d'Acs sont atteints en quelques jours (ce qui s'explique, premièrement, par la présence de lymphocytes B mémoires et, deuxièmement, par la production de nouveaux plasmocytes aux centre germinatifs). La phase d'augmentation reste exponentielle, mais sa croissance est plus rapide, alors que la phase de décroissance est plus prolongée. Néanmoins, une deuxième exposition à l'Ag trop rapprochée de la première pourrait être inefficace du fait de l'élimination de l'Ag par des titres importants d'Acs sériques. Les Acs vont persister beaucoup plus longtemps, parfois indéfiniment (ex. les Acs induits par le vaccin contre la fièvre jaune (96)). Ceci peut être expliqué par la présence de lymphocytes mémoire, qui une fois stimulées par l'Ag, se différencient en cellules sécrétrices d'Acs. La mémoire immunologique existe pour les deux types de lymphocytes T et B (97).

La réponse secondaire s'observe avec un maximum d'intensité, lors de stimulations ultérieures, si l'on augmente les doses d'antigènes. La mémoire immunologique persiste très longtemps chez l'homme même quand la concentration sérique d'Acs descend en dessous du seuil de détection. Elle dépend de la qualité et de la quantité de l'Ag inoculé ainsi que comme déjà évoqué, de la fréquence des expositions à l'Ag (98).

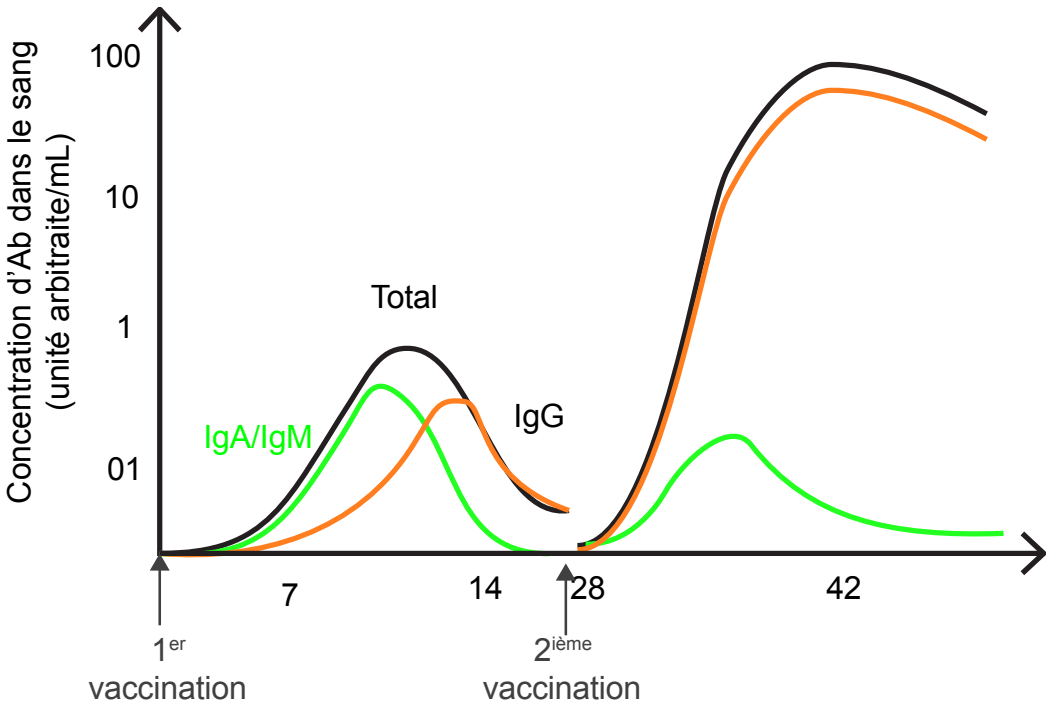


Figure 5. Représentation schématique de la formation d'Acs subséquente à la vaccination
L'isoforme et les titres d'Acs détectés à la suite de la vaccination sont présentés en fonction du temps écoulé à la suite d'une immunisation.

3.4. Vaccins et interaction entre les cellules immunitaires

3.4.1. Interaction entre les cellules immunitaires

Les cellules innées interagissent avec les cellules de la réponse adaptative via des interactions directes ou indirectes. Les interactions directes incluent la liaison des antigènes présentés par les cellules innées aux cellules adaptatives ou à la création de synapses immunitaire entre les cellules. Les interactions indirectes incluent la production de cytokines et chimiokines par les cellules innées recrutant les cellules adaptatives ou modifiant le type de lymphocytes d'aide recrutés (**Figure 6** et **Figure 7**).

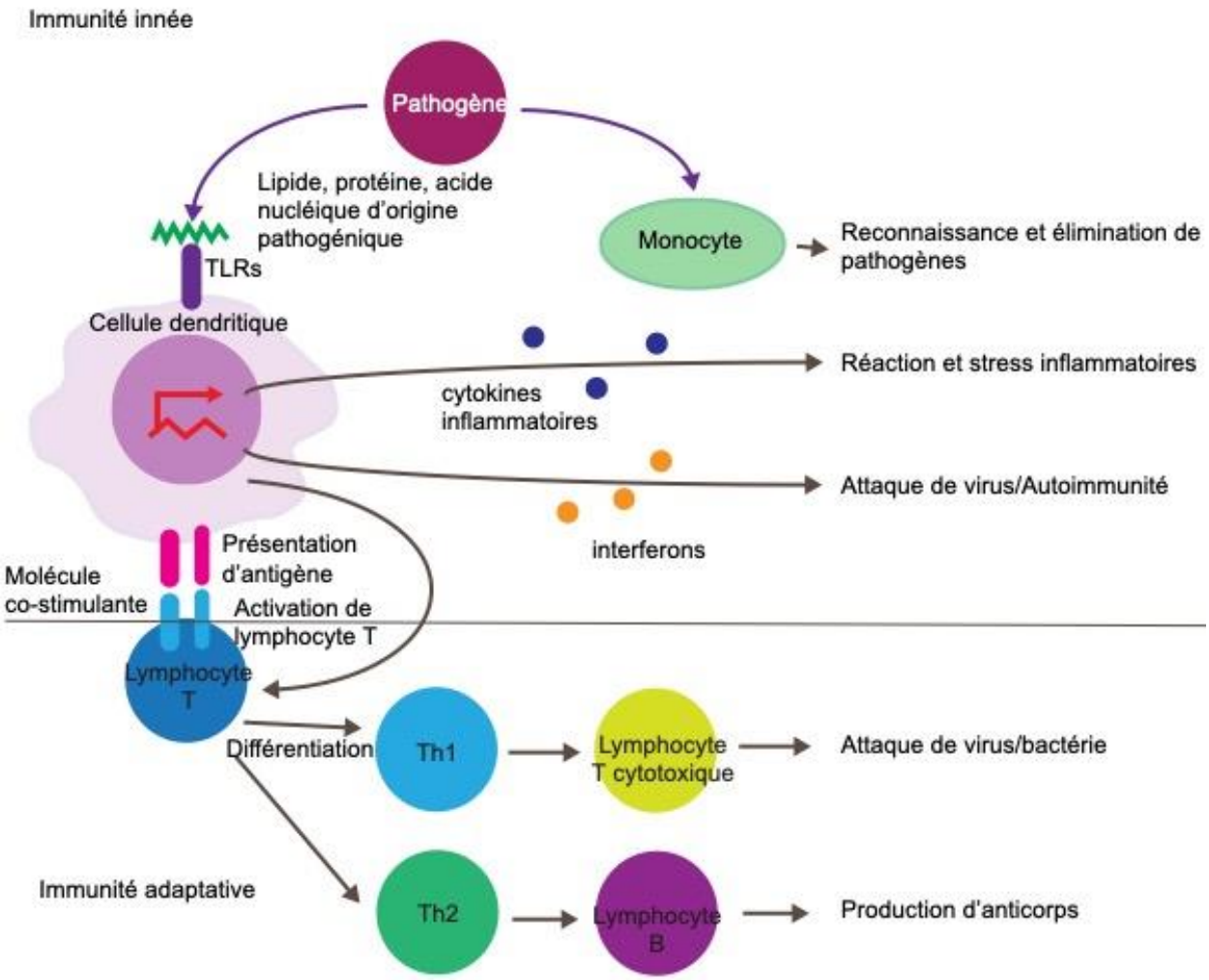


Figure 6. Intégration des différentes composantes cellulaires de la réponse immunitaire
Représentation schématique des moyens de communication entre cellules immunitaires impliquées dans la réponse à un pathogène.

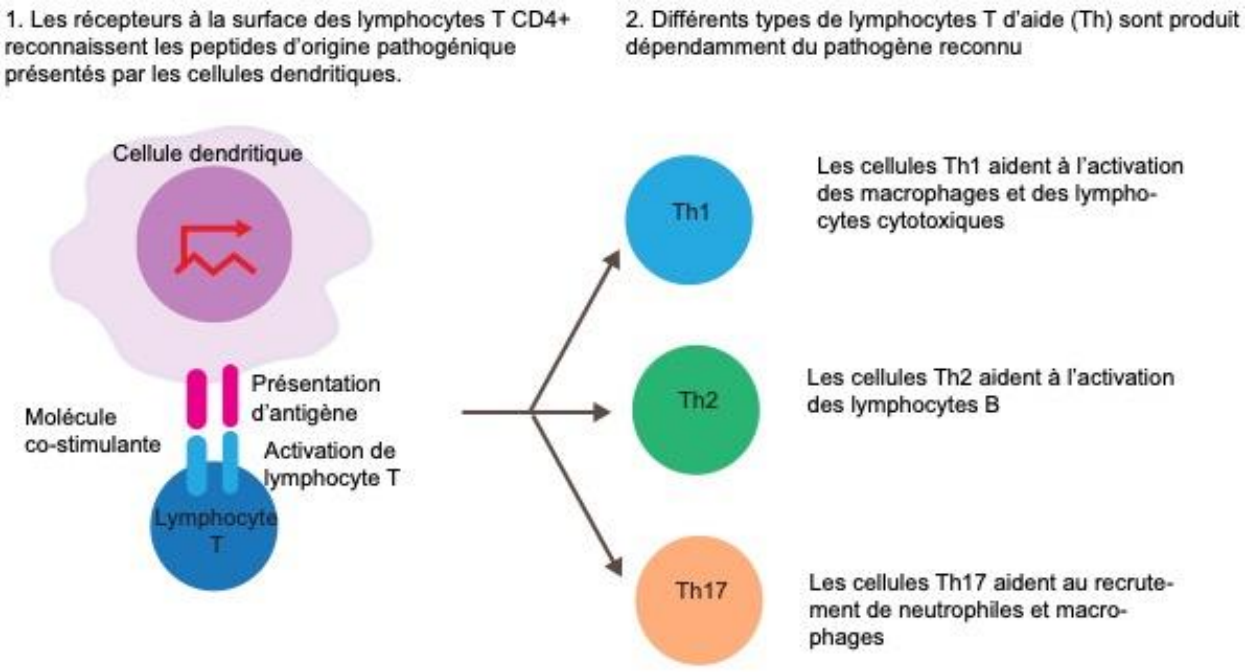


Figure 7. L'aide des lymphocytes T et B

En présence de cytokines (IL-12 pour les Th1, IL-4 pour les Th2, TGF-β/IL-6/IL-23 pour les Th17) et l'activation de facteurs de transcription clefs (T-bet pour les Th1, STAT6/GATA3 pour les Th2, STAT3/RORγt pour les Th17) mènent à la différentiation des lymphocytes T CD4+ naïfs en lymphocytes d'aide (Th). Les cellules Th produisent des cytokines (IFN-γ pour les Th1, IL-4/IL-5/IL-13 pour les Th2, IL-17 pour les Th17) qui sont impliquées dans l'immunité à médiation cellulaire contre les bactéries, parasites et les virus.

3.4.2. Stratégies de vaccination qui miment l'interaction entre les cellules immunitaires

La vaccination met en jeu un processus actif comprenant, d'une part, la présence d'Acs circulants et d'autre part, des lymphocytes effecteurs.

Les cellules présentatrices d'antigènes jouent un grand rôle dans l'induction des réponses immunitaires. Ces cellules assurent la capture et la présentation des antigènes qui sont des étapes décisives pour l'immunogénicité du vaccin. Il existe de nos jours de nombreuses stratégies cherchant à délivrer spécifiquement les antigènes vers les cellules présentatrices d'antigènes. Ce processus consiste à coupler les antigènes soit, à des Acs reconnaissant spécifiquement les molécules de surface des cellules présentatrices d'antigènes ou à des toxines bactériennes qui sont capables de se fixer sur des molécules de surface exprimées par les cellules présentatrices d'antigènes. Plus récemment, il a également été proposé de cibler spécifiquement les cellules présentatrices d'antigènes à l'aide de vecteurs viraux recombinants (lentivirus) pseudotypés avec une enveloppe mutée pour se lier spécifiquement aux cellules dendritiques. De manière globale

ces stratégies favorisent la présentation antigénique par les cellules présentatrices d'antigènes et l'induction des réponses cellulaires (**Figure 8**) (99).

3.5. Comparaison de la réponse immunitaire naturelle et celle induite par les vaccins

La réponse immunitaire naturelle à une infection virale qui mène dans la majorité des cas à la production d'Acs neutralisants et/ou au recrutement de cellules T CD8+ cytotoxiques. La majorité des vaccins tentent d'induire ce type de réponse immunitaire, mais il existe des exceptions. C'est le cas du vaccin RV144 contre le VIH qui induit des Acs non neutralisants, des intermédiaires nécessaires d'une activité cytotoxique par les cellules NK et macrophages. Ceci démontre que les mécanismes conférant la protection par les vaccins peuvent être distincts de la réponse immunitaire naturelle (99).

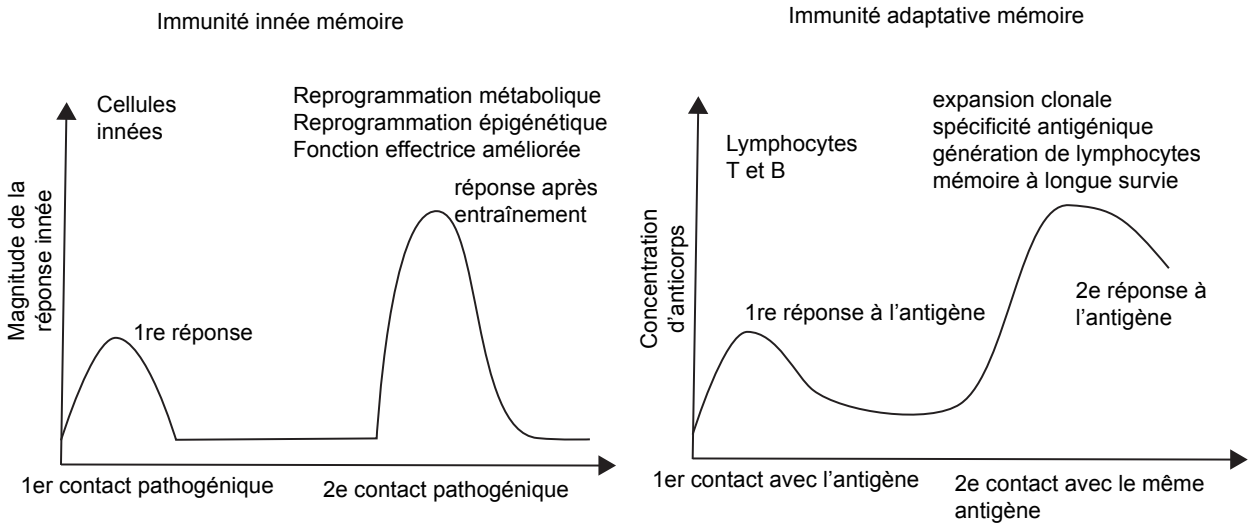


Figure 8. Les différents mécanismes de la mémoire immunitaire

Les trois articles de cette thèse soulignent l'importance de la réponse immunitaire innée (médiée par les monocytes et cellules dendritiques) dans la réponse aux vaccins contre l'HepB et RV144 (1er et 2e article de cette thèse) et à la réponse à une infection par les virus respiratoires. Ces articles soulignent les voies de signalisation spécifiques (interféron, inflammatoire) de ces cellules qui impactent par la suite les lymphocytes T et B.

4. Marqueurs associés à la réponse aux vaccins

Les vaccins ont éradiqué plusieurs maladies infectieuses (ex. variole et poliomyélite). Pourtant, pour certaines infections telles que le VIH, malaria ou la tuberculose, le développement d'un

vaccin efficace demeure élusif. Aujourd'hui la majorité des vaccins candidats pour ces pathogènes, testés lors d'essais cliniques de phase III, ont échoué ou ont permis qu'une protection modeste contre ces infections (100-102). Il est important de définir ce qui différencie un vaccin protecteur d'un vaccin qui ne confère pas une protection contre l'infection par le pathogène ciblé par ce vaccin.

De plus, même pour les vaccins protecteurs, certains individus ne parviennent pas établir une réponse mémoire protectrice adéquate. L'identification de biomarqueurs capables de discerner les individus bénéficiant de la vaccination de ceux nécessitant une stratégie de prévention alternative est d'un grand intérêt clinique.

4.1. Efficacité des vaccins

L'efficacité d'un vaccin est déterminée lors d'essais cliniques de phase 3 où l'objectif premier est de démontrer que l'administration d'un vaccin permet de réduire le nombre d'individus infectés par un pathogène (ex. virus) en comparaison à des individus témoins ayant reçu un placebo ou la thérapie préventive standard contre un pathogène. Une fois les essais cliniques phase 3 complétés, des études additionnelles sont typiquement conduites pour identifier des marqueurs (ou corrélats) pouvant définir ou quantifier la réponse aux vaccins. Habituellement, ces marqueurs sont la fréquence et le type de lymphocytes T CD8+ cytotoxiques (réponse cellulaire) ou les titres sériques/plasmatiques d'Acs spécifique à l'Ag (réponse humorale). Un niveau seuil séparant les individus vaccinés n'ayant pas été infectés des individus vaccinés ayant été infectés est déterminé durant ces essais cliniques phase 3 (103).

Par exemple, les vaccins contre l'HepB sont efficaces à plus de 90% chez les adultes. Les Acs reconnaissant l'Ag de surface de l'HepB apparaissent environ un mois après la première injection. Un titre de ces Acs supérieur ou égal à 10 mUI/ml après la troisième injection est assumé comme étant protecteur.

Pour les vaccins contre les virus Influenza, deux paramètres sont souvent utilisés comme marqueurs de la réponse au vaccin : (1) la séroprotection et (2) la séroconversion. À la suite de la vaccination contre les virus Influenza, la séroprotection est mesurée en quantifiant les Acs sériques capables d'inhiber l'hémagglutination du virus avant et après la vaccination. Ceci est particulièrement utile, car il a été établi que les titres d'Acs spécifiques de l'hémagglutinine HI sont associés avec une protection réelle contre l'infection par les virus Influenza et peuvent être utilisés comme corrélat de protection (104). Habituellement, un titre protecteur de 1:40 est utilisé (correspondant à une réduction du risque d'infection par 45% (104)). La séroconversion quant à elle calcule le ratio du titre des Acs HI après vaccination comparé au titre prévaccination. Chaque

année, le virus de l'influenza réapparait (avec une combinaison différente des protéines hémagglutinine et neuraminidase de l'enveloppe virale) et infecte la population humaine (typiquement en automne et en hiver). Les sous-types du virus d'influenza circulant diffèrent d'année en année. Les vaccins contre le virus d'influenza sont modifiés pour cibler le sous-type de virus circulant l'année en cours. Un individu vacciné ou infecté les années précédentes peut déjà avoir des Acs capables de reconnaître le virus circulant (si similaire à une variante du virus circulant ultérieurement). Pour ces individus, le but du vaccin durant l'année en cours est d'augmenter la quantité et la qualité (ex. Acs à large spectre) d'Acs.

La séroconversion est la métrique permettant de mesurer l'augmentation des titres d'Acs à la suite de la vaccination, prenant en prenant compte des titres d'Acs présents chez un individu pré-vaccination. Une augmentation de 4 fois des titres d'Acs après la vaccination est considérée comme une réponse protectrice par le vaccin (105) (les vaccins contre influenza ont une efficacité entre 19% et 60% à ces titres (106)).

Le seul vaccin spécifique au VIH ayant démontré une protection chez l'Homme est le vaccin RV144. Les titres d'Acs IgG reconnaissant la boucle V1/V2 du VIH est le marqueur humoral associé avec un risque réduit d'infection par le VIH. Les titres d'Acs IgG reconnaissant la boucle V1/V2 du VIH ne constituent pas un marqueur de protection de l'acquisition, mais un marqueur de risque réduit d'une infection par le VIH. En effet, ces deux types de marqueurs doivent être distingués, un marqueur de protection par le vaccin nécessite que les individus vaccines ait été exposés au virus alors qu'un marqueur associé à un risque réduit incluses tous les individus vaccinés négatifs pour le virus qu'ils aient ou non été exposés au virus (107).

Les marqueurs de protection par un vaccin ne sont pas tous impliqués dans le mécanisme de protection conféré par un vaccin. C'est le cas de IgG reconnaissant V1/V2 qui ne perdure pas en circulation plus de quelques semaines après la vaccination (108).

4.2. Critères cliniques

Certains critères cliniques sont associés à la réponse aux vaccins. Ces critères incluent l'âge, le sexe, le poids, fumer la cigarette et seront discutés dans les prochaines sections de cette thèse.

4.2.1. Âge

Les âges extrêmes (nourrissons et personnes âgées) présentent un risque plus élevé de morbidité et de mortalité liées aux maladies infectieuses et seraient les plus grands bénéficiaires de la vaccination. Les déficits immunitaires (absence de maturation du système immunitaire chez les nouveaux nés ou encore l'inflammation associée aux comorbidités chez les personnes âgées)

rendent ces individus plus à risque d'infection et ont également des réponses plus faibles à la vaccination. En effet, les études chez les enfants prématurés démontrent que 9 à 21% moins d'enfants prématurés sont protégés par le vaccin contre l'HepB que les enfants nés à terme (109). D'autres études chez les enfants démontrent que les enfants nés de mères possédant des Acs contre l'Ag de surface de l'HepB ont 10% moins de chance d'être protégés par le vaccin contre l'HepB que les enfants de mères séronégatives pour les Acs contre l'Ag de surface de l'HepB (110). Une observation similaire a été faite pour les vaccins contre les virus influenza, ainsi de 3% à 70% moins d'enfants répondent au vaccin s'ils possèdent des Acs maternels contre l'influenza en comparaison avec ceux qui ont des niveaux indétectables de ces Acs après la naissance (111). Ceci suggère que le jeune âge du système immunitaire et les Acs hérités de la mère (pouvant interférer avec un vaccin) jouent un rôle prépondérant dans la réponse aux vaccins. Notre connaissance du développement du système immunitaire au début de la vie est limitée; ceci est dû au fait que les échantillons de nourrissons sont difficiles à obtenir et par conséquent l'accès aux matériels biologiques (ex. sang, plasma) est plus limité. Les nouvelles technologies qui permettent d'acquérir le profil transcriptionnel et fonctionnel de cellules immunitaires individuelles pourront accélérer notre connaissance du système immunitaire des nouveaux nés et des enfants en bas âge.

L'augmentation de la prévalence et de la gravité des maladies infectieuses constitue l'un des principaux problèmes de santé associés au vieillissement. Ceci est au moins en partie dû à un déclin de la fonction immunitaire, appelé immunosénescence, qui se manifeste à la fois par des déficits de la réponse innée et adaptative du système immunitaire (ainsi qu'une augmentation du risque de comorbidités) (112). L'immunosénescence est caractérisée par un déséquilibre entre les mécanismes inflammatoires et anti-inflammatoires, qui serait le résultat d'une stimulation antigénique chronique et de stress oxydatif induit par l'accumulation de radicaux libres oxygène et de produits toxiques (112). Si la prévention des maladies infectieuses par la vaccination est une mesure importante pour assurer un vieillissement en bonne santé, et que plusieurs pays ont adopté des recommandations pour la vaccination des personnes âgées (ex. Grippe, Streptococcus pneumoniae et virus de l'herpès zoster), il est bien connu que l'immunosénescence affecte la réponse aux vaccins, défectueux et moins protecteurs chez les personnes âgées (113). En effet, la réponse au vaccin contre l'HepB est 61% moins importante chez les personnes âgées de 65 et plus que chez les adultes (18-50 ans) (114). Plus de 90% des décès attribuables aux virus Influenza surviennent chez des adultes de plus de 65 ans, et la réponse à la vaccination dans ce groupe d'âge est 17% à 83% moins importante que chez les adultes (115).

4.2.2. Sexe

En général, les femmes présentent des réponses immunitaires quantitativement plus importantes que les hommes, tant à la suite d'une infection ou d'une vaccination, et sont simultanément plus sujettes aux maladies auto-immunes. Ainsi, l'efficacité du vaccin contre l'HepB est ~40% plus élevée chez les femmes que les hommes (116). Pareillement, une étude canadienne a observé que les vaccins contre les virus Influenza sont au moins 11% plus efficaces chez les femmes que chez les hommes (117). L'implication des hormones sexuelles (œstrogènes and androgènes) dans le contrôle de la réponse immunitaire est supportée par certaines études précliniques utilisant des modèles animaux, mais leur implication chez les hommes reste à être démontrée. Les cellules B expriment les récepteurs des œstrogènes ERα / β, permettant aux œstrogènes de moduler directement la fonction lymphocytaire. Le traitement in vitro des PBMC humains mâles et femelles avec de l'œstradiol (E2) a considérablement amélioré la différenciation des cellules B [PMID: 6459399]. Les œstrogènes semblent affecter diverses fonctions des cellules B comme la signalisation par le BCR, l'expansion des cellules B et la maturation des cellules B (118). La liaison directe des récepteurs aux œstrogènes a plusieurs effets sur les cellules B activées, y compris l'augmentation de la production d'Acs, l'hypermutation somatique, la recombinaison par commutation de classe ainsi que le développement et la persistance de la mémoire des cellules B (119). Dans cette lignée, Furman et collab. ont démontré que des niveaux élevés de testostérone sérique étaient associés à de plus faibles concentrations d'Acs après la vaccination contre Influenza chez les sujets masculins, ce qui étaye l'hypothèse selon laquelle la testostérone pourrait être un facteur clef dans l'atténuation des réponses immunitaires. Pour vraiment prouver que la testostérone elle-même inhibe la réponse immunitaire à la vaccination, il serait nécessaire de comparer les réponses vaccinales dans le traitement pré- et post-testostérones d'individus ayant subi une castration (120).

4.2.3. Autres marqueurs cliniques

D'autres marqueurs cliniques tels que la grossesse (pas associée à la réponse aux vaccins contre l'HepB (121), mais réduit de 31.4% 32.5% l'efficacité des vaccins contre Influenza (122)), les maladies chroniques telles qu'une infection par l'hépatite C ou des problèmes cardiaques (réduisent l'efficacité des vaccins contre l'HepB par 22% (123) et par 50% pour les vaccins contre Influenza (124)), les personnes immunocompromises (réduis l'efficacité des vaccins contre l'HepB par 20% à 70% (125) et par 5-68% pour les vaccins contre l'Influenza (126)), les personnes ayant reçu une transplantation sont associés à la réponse aux vaccins (une

transplantation réduis l'efficacité des vaccins contre l'HepB par 72.4% (127) et réduis l'efficacité des vaccins contre l'influenza par 0% à 68% (126)).

L'ethnicité a également été associée à la réponse aux vaccins contre l'HepB. La réponse immunitaire des lymphocytes CD8+ spécifiques de l'Ag de l'HepB de sujets caucasiens est plus importante que celle de participants d'origine asiatique (128). Le taux de vaccination contre le virus Influenza varie au Canada en fonction de l'ethnicité des participants (participants d'origine asiatique plus à même d'être vaccinés que les participants d'origine européenne (129)), toutefois le lien entre l'ethnicité et la réponse humorale n'a pas été établi ou n'est pas significatif pour les vaccins contre le virus Influenza.

4.3. Biomarqueurs génétiques, transcriptionnels et protéiques

L'utilisation des technologies de criblage large échelle, permettant d'interroger des milliers de gènes et protéines simultanément, a accéléré la recherche de biomarqueurs putatifs de la réponse aux vaccins.

4.3.1 Biomarqueurs génétiques

Plusieurs études ont tenté d'identifier des marqueurs génétiques pour la réponse aux vaccins (130). La majorité de ces études se sont concentrées sur le chromosome 6 qui comprend le locus du système majeur d'histocompatibilité et qui contrôle, par le biais de la collaboration entre les lymphocytes T CD4+ et lymphocytes B, les titres d'immunoglobulines pouvant être produits après la stimulation par les antigènes d'un vaccin. Ainsi, dans une étude incluant 20 personnes vaccinées contre l'HepB (vaccin Heptavax-B) ayant des titres d'Acs spécifique de la protéine de surface de l'HepB 33 fois inférieurs, un haplotype (*i.e.* ensemble de polymorphismes de la séquence primaire d'acides nucléiques) particulier des gènes HLA-B, HLA-B, HLA-C, et HLA-DRB1 (haplotype HLA-B8,SC01,DR3) a été observé chez 5 des 20 individus (25%) alors que dans la population générale moins de 0.9% des personnes sont homozygotes pour cet haplotype (131). Une autre étude basée sur le génotypage de locus du système majeur d'histocompatibilité de 134 personnes vaccinées contre l'HepB qui ont répondu faiblement au vaccin (titres d'Acs spécifique de la protéine de surface de l'HepB en dessous de 10 mIU/mL), a identifié un polymorphisme nucléotidique dans le gène HLA-DPB1 (rs7770370) présent dans 113 des 134 individus (79%) alors qu'uniquement 92 de 151 individus (42%) vaccinés ayant produit des niveaux d'Acs contre l'Ag de surface de l'HepB possèdent ce polymorphisme. Aucun polymorphisme dans le gène HLA-DRB1 n'était significativement associé à la réponse au vaccin dans cette étude (132). Une troisième étude génotypant le locus du système majeur

d'histocompatibilité et comprenant 108 personnes hyporépondantes et 77 répondants au vaccin contre l'HepB a identifié un polymorphisme du gène HLA-DRB1 (rs477515 dans la région non-codante en amont du gène) associé à une faible réponse au vaccin. Une étude indépendante incluant 79 hyporépondants et 85 répondants au vaccin contre l'HepB a identifié à la fois un haplotype du gène HLA-DRB1 (HLA-DRB1*07) et une délétion de 4 nucléotides dans le promoteur du gène IL-12B comme étant associés à l'hyporéponse au vaccin (133) Une étude incluant 665 participants hyporépondants et 981 participants répondants a identifié deux polymorphismes comme étant associés à une faible réponse aux vaccins contre l'HepB, soit le premier dans le gène HLA-DRA et un second en aval du gène codant pour le facteur de transcription FOXP1 (134). Un polymorphisme dans le gène CXCR5 (rs3922), est associé à une plus forte expression de l'ARN messager du récepteur de chimokine CXCR5 est associé à l'hyporéponse au vaccin contre HepB. Le récepteur CXCR5 est exprimé à la surface de lymphocytes T CD4+ est responsable de fournir les signaux de survie et de différenciation des lymphocytes B. D'autres polymorphismes à proximité des gènes IL1, IL10, TLR2 ont été identifiés (135). Toutefois, l'impact de ces polymorphismes sur les gènes qui incluent ces polymorphismes ou qui sont à proximité de ces polymorphismes demeure inconnu.

De nombreux polymorphismes des gènes du système majeur d'histocompatibilité (ex. HLA-B, HLA-DRB1), des cytokines (ex. IL6, IL18, IFNG) et des récepteurs de cytokines (ex. IL1R, IFNAR2, TNFRSF1A) ont aussi été associés à la réponse aux vaccins spécifique des virus Influenza (136).

Pour le vaccin contre le VIH RV144, une seule étude a tenté d'identifier des polymorphismes dans les gènes du système majeur d'histocompatibilité. Aucun de ces polymorphismes n'est directement associé au risque d'infection par le VIH. Toutefois cette étude qui a inclus 6987 participants a identifié l'allèle HLA-DPB1*13 associé à une plus forte expression des Acs IgG reconnaissant la boucle V1/V2 du VIH en comparaison aux personnes vaccinées qui n'exprimaient pas cet allèle (137).

En résumé, plusieurs études ont identifié une association entre des marqueurs génétiques du gène HLA-DRB1 et la réponse au vaccin de l'HepB. Ces études ont identifié des polymorphismes à proximité de gènes impliqués dans la réponse immunitaire, la quasi-totalité de ces polymorphismes n'ont été décrits que dans une seule de ces études et n'ont pas été reproduits sur des cohortes indépendantes ou par des groupes de recherche différents de ceux identifiant ces polymorphismes. La majorité de ces polymorphismes ont une fonction inconnue nécessitant plusieurs études avec de larges cohortes pour définir leur impact sur la réponse immunitaire aux vaccins.

4.3.2. Biomarqueurs transcriptionnels

Avec l'émergence des technologies large-échelle comme les micropuces à ADN et le séquençage d'ARN, l'utilisation des données transcriptionnelles pour identifier des biomarqueurs de la réponse aux vaccins est devenu plus accessible. Approximativement 60% des données transcriptionnelles relatives à la réponse aux vaccins ont été générées pour les vaccins contre l'Influenza et la quasi-totalité de ces études a utilisé des échantillons sanguins qui incluent majoritairement des lymphocytes de sang périphérique pour identifier ces biomarqueurs. Un nombre restreint de ces études ont utilisé des cohortes indépendantes pour confirmer la validité de ces biomarqueurs transcriptionnels. Les Tableaux 2 et 3 se concentrent sur ces études ayant tenté de confirmer leurs résultats sur des échantillons indépendants ou en utilisant des techniques d'apprentissage machine telles que la validation croisée pour évaluer la robustesse de ces marqueurs.

Ces marqueurs transcriptionnels, incluant des gènes uniques ou un ensemble de gènes faisant partie d'une une signature transcriptionnelle, peuvent être séparés sur la base du temps écoulé entre la vaccination et le prélèvement sanguin. En effet, un nombre limité de ces études ont démontré que des marqueurs transcriptionnels prévaccination peuvent être identifiés alors que d'autres études se sont concentrées sur des échantillons prélevés peu après la vaccination (entre 24h et 7 jours) pour prédire la réponse aux vaccins (typiquement évalué en mesurant les titres d'Acs un mois après la vaccination).

Une liste extensive des études transcriptomiques ayant identifié des marqueurs transcriptionnels pré-vaccination est fournie dans la Table 2. Bartholomeaus et collab. ont construit un classificateur naïf bayésien basé sur l'expression de l'ensemble du transcriptome et ont évalué leur classificateur par validation croisée (entrainé et testé sur des parties du jeu d'entrainement sans avoir recours à un jeu test indépendant). Les gènes dans ce classificateur sont pondérés sur la base de leur capacité à distinguer hyporépondants et répondants aux vaccins; des gènes impliqués dans la transcription d'ARN et dans la réponse immunitaire (138). Le groupe HIPC (comprenant 8 universités aux États-Unis) a effectué une méta-analyse de plusieurs jeux de données transcriptionnelles de sang d'individus avant d'avoir reçu des vaccins contre les virus Influenza. Ils ont identifié 15 gènes qui reflètent l'activation de la réponse inflammatoire (ex. DPP3, MVP codant pour des protéines bloquant l'apoptose et activant le facteur de transcription NFκB, respectivement), ils ont utilisé la moyenne géométrique de ces 15 gènes comme score et évalué le pouvoir prédictif sur une cohorte indépendante de 223 individus (139).

Table 3. Biomarqueurs transcriptionnels mesurés avant la vaccination

Référence	Vaccin (pathogène)	Voies de signalisation	Précision
(138)	Engerix-B (HepB)	Réponse immunitaire, transcription d'ARN	AUC=0.66
(140)	Twinrix (HepB)	Lymphocytes B, cytokines inflammatoires	Précision=0.81
(139)	Fluzone, Fluvirin, Fluarix (Influenza)	Inflammation	AUC=0.79
(141)	MVA85A (*Mycobacterium tuberculosis*)	TLR1	Précision=0.80
(142)	Fluzone (Influenza)	Apoptose	AUC=0.84
(143)	Fluvirin, Fluzone, Fluarix (influenza), YF-17D (Fièvre jaune)	Cellules dendritiques, lymphocytes, interférons	AUC=0.83

Une liste extensive des études transcriptomiques ayant identifié des marqueurs transcriptionnels après la vaccination est fournie dans la Table 3. Nakaya et collaborateurs ont construit un ensemble de modèles prédictifs de la protection conférée par les vaccins contre les virus influenza en utilisant des données transcriptionnelles à 1, 3 et 7 jours après vaccination. En utilisant une règle de majorité sur un jeu indépendant, ils ont démontré que leurs prédicteurs ont une précision de 67.6%. Les marqueurs associés aux cellules B produisant des Acs sont associés à une réponse humorale 1 mois après la vaccination.

Table 4. Biomarqueurs transcriptionnels mesurés après la vaccination

Référence	Vaccin (pathogène)	Temps après la vaccination	Voies de signalisation	Précision/Corrélation
(144)	SV-AS03 (Influenza)	7 jours	Genes Ig et Ki-67 (prolifération)	MSE=3.32
(140)	Twinrix (HepB)	1 jour	Interférons, cytokines inflammatoires	Précision=1.00
(145)	rVSV-ZEBOV (Ebola)	1 jour	IP10	MSE=1.01
(146)	RTS,S (Malaria)	56 jours	Marqueurs de NK	Précision=0.80
(147)	SV-AS03 (Influenza)	1 jour	Interférons	MSE=0.05
(148)	Menveo, ACWYVax (*Neisseria meningitidis)*	7 jours	Lymphocytes B	R=0.38
(149)	Fluarix (Influenza)	28 jours	Réparation de l'ADN et transport du gaz carbonique	MSE=0.33
(150)	Fluarix, Fluvirin, Fluzone (Influenza)	1/3/7 jours	Lymphocytes B	Précision=0.70
(151)	Fluvirin (Influenza)	7 jours	Stress du réticulum endoplasmique, synthèse des N-glycan, cycle cellulaire	AUC=0.70
(152)	Fluariz, Fluvirin, FluMist (Influenza)	3/7 jours	Monocytes et cellules dendritiques	Précision=0.90
(153)	Fluzone (Influenza)	1/3 jours	STAT1 et E2F2	Précision=0.75
(154)	YF-17D (Fièvre jaune)	7 jours	Lymphocytes B	Précision=1.00

AUC : aire sous la courbe ROC, MSE= Erreur quadratique moyenne, R=Corrélation de Pearson

En résumé, plusieurs études ont identifié une association entre l'expression de gènes codant pour des protéines impliquées dans la réponse inflammatoire ou la réponse humorale et la réponse aux vaccins. Toutefois, ces études n'ont que brièvement décrit le potentielle rôle de ces gènes dans le mécanisme de réponse aux vaccins.

4.3.3. Biomarqueurs protéiques

Les technologies large échelle permettant d'interroger la fréquence des cellules immunitaires dans le sang (par cytométrie en flux) et des protéines sécrétées dans le plasma (protéomique plasmatique) ont également permis d'identifier de nouveaux biomarqueurs putatifs de la réponse aux vaccins.

La fréquence de cellules immunitaires a été évaluée comme des biomarqueurs putatifs de la réponse aux vaccins. En effet, dans une étude visant à identifier des marqueurs de la réponse au vaccin spécifique du virus Influenza Fluvirin, la fréquence des cellules B et des plasmablastes (cellules B différentiées produisant les Acs) était positivement associée à la réponse au vaccin avec un AUC=0.78 (151). L'expression transcriptionnelle prévaccination n'a résulté en aucun modèle prédictif significatif et n'a ajouté aucun gain en termes de prédiction lorsque combinée avec les données de cytométrie en flux. Similairement, la fréquence des lymphocytes T CD4+ capables de produire multiples cytokines, un marqueur associé à la protection contre la progression d'infection virales, a aussi été identifié comme marqueur de la protection conférer par le vaccin contre le VIH RV144, et ce avec une aire sous la courbe ROC AUC=0.67 (155).

L'utilisation de techniques permettant de mesurer les protéines sécrétées dans le plasma a également permis l'identification de biomarqueurs putatifs de la réponse aux vaccins.

Dans une étude cherchant à identifier des marqueurs de la réponse au vaccin contrer les virus Influenza Fluzone, l'expression de certaines cytokines telles que le ligand Fas soluble (sFasL) et l'IL-12p40, mesuré avant la vaccination, constitue un prédicteur associé négativement à la réponse au vaccin. sFasL empêche le ligand FasL de se lier au récepteur Fas à la surface de cellules, diminuant ainsi la sensibilité à l'apoptose contrôle Fas suggérant que la présence d'inhibiteurs solubles de l'apoptose pourrait augmenter les titres d'Acs en réponse à la vaccination. Dans la même étude, les lymphocytes T CD4+ mémoire central et les cellules lymphocytes T CD8+ mémoire effectrice, et les niveaux de base de STAT1 dans les cellules T CD8+ ont aussi une corrélation négative avec la réponse au vaccin Fluzone. Un prédicteur basé sur la fréquence des lymphocytes T et des protéines dans le plasma avant la vaccination a permis de prédire la réponse au vaccin Fluzone avec une aire sous la courbe ROC AUC=0.844 (142).

En résumé, plusieurs études ont identifié une association entre la fréquence de lymphocytes B et T et la réponse aux vaccins contre le virus Influenza. Le pouvoir prédictif de ces marqueurs cellulaires pour d'autre vaccins reste à être établit.

4.3.4. Biomarqueurs métaboliques et microbiome

Les études utilisant des données large échelle ont également démontré le rôle des facteurs environnementaux, tels que les co-infections, les microbes commensaux et les métabolites dans la réponse immunitaire après la vaccination. Par exemple, les infections chroniques telles que l'infection par le cytomégalovirus (CMV) sont associées positivement à la séropositivité de jeunes adultes (20 à 32 ans) après la vaccination avec le vaccin contre les virus Influenza Fluzone. Toutefois, cette association n'est pas observée chez les personnes âgées (60-89 ans) (156). Les auteurs ont suggéré que la protection croisée médiée par le CMV chez les jeunes est médiée par l'IFNγ et pourrait être le résultat d'une réactivité croisée des lymphocytes T (pour les virus Influenza et CMV) ou d'une immunité hétérologue entraînée (par le virus CMV permettant une réponse accrue pour Fluzone); deux hypothèses intéressantes qui devraient être davantage validées. Dans les zones économiquement sous-développées, la présence de co-infections au moment de la vaccination constituerait un obstacle majeur à la réalisation d'une réponse protectrice à la vaccination.

Un grand nombre de microbes commensaux et symbiotiques colonisent le corps humain. La communauté scientifique a récemment commencé à apprécier l'impact du microbiome de l'hôte sur l'homéostasie des cellules du système immunitaire. L'impact du microbiome dans la réponse aux vaccins a été étudié pour les vaccins contre les virus Influenza et de la poliomyélite où il a été démontré que le microbiome altèrerait la réponse humorale à la suite de la vaccination. L'impact du microbiome sur la réponse au vaccin contre la fièvre jaune n'a pas été observé dans cette même étude suggérant que l'effet du microbiome dépend de la nature du vaccin donné. Dans une étude clinique contrôlée randomisée, l'utilisation d'antibiotiques pour altérer la composition du microbiome n'a modifié que légèrement la réponse immunitaire au vaccin contre les virus Influenza (157).

Enfin, il a également été démontré que le métabolome, c'est-à-dire l'ensemble de petites molécules dérivées de notre alimentation ou produit par les bactéries du microbiome commensal, affecte la réponse immunitaire et ont le potentiel d'affecter la réponse vaccinale. Aucun métabolite avant la vaccination n'a encore été associé à des différences dans la réponse aux vaccins, mais le métabolisme du phosphate d'inositol au jour 1 suivant la vaccination s'est avéré être un

prédicteur de la réponse des lymphocytes T au vaccin vivant atténué contre le zona, Zostavax (158).

Compte tenu de l'immense complexité des antécédents d'infection, ainsi que des variations majeures de la composition du microbiome et du métabolome dans les communautés médicalement mal desservies et dans les régions à forte charge de morbidité, seule une approche systémique exhaustive, non biaisée et intégrative permettra de quantifier la contribution spécifique de ces paramètres sur l'issue de la vaccination.

Les deux premiers articles de cette thèse identifient plusieurs biomarqueurs potentiels pour les vaccins contre l'HepB et le vaccin RV144 contre le VIH. Certains de ces biomarqueurs ont déjà été décrits dans la littérature scientifique (ex. réponse aux interférons) alors que d'autres sont de nouveaux biomarqueurs (ex. l'hème pour le vaccin contre l'HepB).

5. Technologies large échelle

Malgré les avancées en Immunologie, nos connaissances sur les interactions hôtes pathogènes sont limitées et les mécanismes associés à la réponse protectrice d'un vaccin sont encore inconnus. La plupart des vaccins testés sont dérivés de concepts empiriques. L'intégration de technologies à large échelle pour déterminer des corrélats de protection spécifiques aux vaccins a eu comme conséquence immédiate une approche plus rationnelle pour le développement de de vaccins. Ces approches multiomiques ont aussi permis de comprendre davantage la réponse immunitaire à la suite de la vaccination.

5.1. Transcriptomique

5.1.1. Biopuces

Le transcriptome est défini comme l'ensemble des gènes exprimés par une cellule. La technologie des biopuces (ou puces à ADN) repose sur la détection simultanée de plusieurs milliers d'ARN messagers, les intermédiaires dans la traduction des gènes en protéines (159). Des brins d'ADN simple brin appelé sondes sont fixés à un support fixe (en silicone, verre ou plastique). Chaque sonde est conçue pour être spécifique à un ARNm. Une fois l'ARN extrait d'un échantillon de cellules, il est amplifié, rétrotranscrit en ADN complémentaire, couplé à un marqueur fluorescent/chimiluminescent, fragmenté et hybridé avec la biopuce. Une fois l'hybridation complétée, des lavages répétés de la biopuce sont effectués pour éliminer tous les fragments non hybridés. La biopuce est ensuite lue par un scanneur à la longueur d'onde d'excitation des marqueurs fluorescents afin de générer une image pixélisée de la biopuce. L'intensité du signal

de fluorescence de chaque couple transcrit/sonde est proportionnelle à l'intensité d'hybridation, donc à l'expression du gène ciblé (**Figure 9**).

Les biopuces ne servent pas uniquement à mesurer l'expression génique, mais peuvent également être employées pour le génotypage, pour identifier les régions génomiques liant une protéine tels que les facteurs de transcription, la détection de petits ARN ou pour l'hybridation génomique comparative. Cependant, cette thèse ne considère les biopuces que dans le contexte de la mesure de l'expression génique.

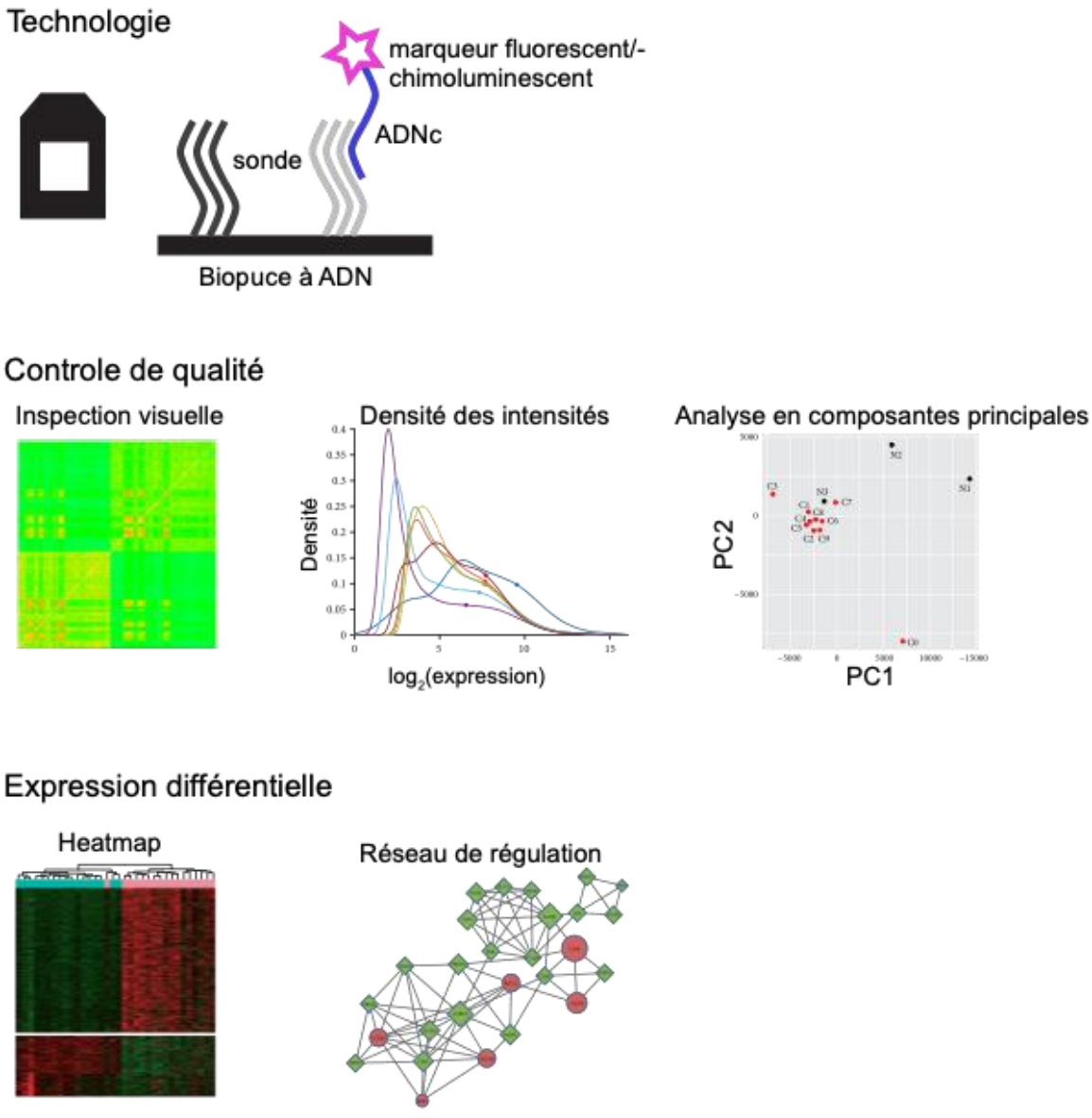


Figure 9. Principe de fonctionnement d'une biopuce, contrôle de qualité et analyse différentielle (Haut) La technologie des biopuces est basée sur l'hybridation d'une sonde nucléique avec un ADN complémentaire (ADNc) d'intérêt marqué avec un fluorochrome. La biopuce peut être lue

pour quantifier le nombre de sondes liées à l'ADNc. (Milieu) Exemples de test de qualité effectué après la lecture de la biopuce, incluant l'inspection visuelle de la puce, quantifier la densité des intensités lumineuses et l'utilisation de techniques de réduction de dimension pour identifier des biopuces aberrantes. (Bas) une fois la qualité des données évaluée, elles sont analysées pour identifier des gènes différentiellement exprimés entre groupes d'échantillons d'intérêt et annotés en termes de fonctions biologiques partagées par ces gènes.

L'ARNm, correspondant aux transcrits des cellules à analyser, est isolé et jumelé à un marqueur fluorescent/chimiluminescent. Les ARNm sont ensuite déposés sur la biopuce. Les fragments d'ARNm se fixent sur les sondes lorsqu'ils possèdent la séquence complémentaire. La disposition des taches fluorescentes sur la biopuce est ensuite analysée et traduite en intensité d'expression des ARN messagers de la cellule. Adaptée de (160).

Les biopuces peuvent être séparées en fonction de la méthode de synthèse des sondes (présynthétisées ou synthétisées directement sur la puce) et du nombre d'échantillons pouvant être hybridés par plaque (simple ou double canal de lecture).
Les biopuces basées sur des sondes présynthétisées sont surtout utilisées dans la fabrication de biopuces personnalisées. Ce type de biopuces sont souvent imprimées en petite quantité et n'ont pas toujours la même efficacité de synthèse que les biopuces commerciales (161). Les biopuces commerciales utilisent des techniques comme la photolithographie pour synthétiser les sondes directement sur plaque (162). Selon les compagnies ces sondes peuvent être de différentes tailles (généralement des 25-mer pour les biopuces Affymetrix, 60-mer pour les biopuces Agilent et Illumina). Plus une sonde est longue et plus elle est spécifique pour un transcrit. Plus une sonde est courte et plus elle peut être imprimée en grand nombre de copies sur la puce augmentant ainsi la sensibilité de la réaction.
Les biopuces possédant un canal de lecture unique ne permettent l'hybridation que d'un seul échantillon biologique par biopuce. Étant donné que les sondes sur une même biopuce peuvent présenter des affinités différentes pour leurs transcrits respectifs, les intensités des sondes sur une même biopuce ne sont pas comparables entre elles. Toutefois, les intensités d'une même sonde sur deux biopuces différentes peuvent être comparées (mesure d'expression relative) (163). Les biopuces possédant des canaux de lecture doubles permettent d'hybrider deux échantillons sur une même biopuce (164). Le second échantillon est généralement une solution référence ou un tissu contrôle pour l'échantillon d'intérêt. Les intensités des sondes d'une même biopuce sont comparables entre elles si elles sont normalisées en fonction des intensités de

l'échantillon contrôle; ceci permet d'éliminer certaines sources de variabilité telles que des affinités différentes des sondes utilisées pour leur transcrits respectifs (mesure d'expression absolue). L'inconvénient majeur de ce type de biopuces est que deux jeux de données utilisant différentes solutions de référence sont difficilement comparables entre eux.

5.1.2. Séquençage de l'ARN

Les technologies de séquençage de l'ARN sont basées sur le séquençage de l'ADN complémentaire et n'incluent pas des approches d'hybridation de l'ADN complémentaire comme la technologie des puces à ADN. Cette technique peut être utilisée pour séquencer directement l'ADN d'un individu (DNA-seq) afin de créer un génome de référence ou pour détecter des altérations de la séquence d'ADN ou pour séquencer l'ARN messager dans une cellule (RNA-seq). La stratégie de séquençage (la quantité de fragments d'ARN séquencés, la longueur des fragments d'ARN séquencés, la stratégie d'amplification des fragments) dépend du contexte biologique étudié. Par exemple, augmenter le nombre de fragments d'ARN séquencés par échantillons permet de quantifier les ARNm à faibles niveaux d'expression (ARMm avec moins de dix copies par échantillon). Séquencer des fragments plus longs facilite l'analyse de l'épissage alternatif; alors que l'utilisation d'amorces aléatoires permet à la fois d'étudier l'expression de gènes codant pour des protéines et des ARN non-codants. Différentes analyses/étapes permettent de s'assurer de la qualité des données de séquençage. Ces étapes incluent l'exclusion de fragments d'ARN avec un nombre élevé de nucléotides ambigus, l'exclusion de fragments d'ARN ne pouvant pas être alignés au génome de l'hôte et l'exclusion de fragments d'ARN alignés de façon ambiguë au génome de l'hôte ou présent dans des régions répétées dans le génome. Dans cette thèse, nous avons utilisé exclusivement des données issues de la technologie de puces à ADN.

Les technologies de séquençage sont de plus en plus utilisées pour identifier des biomarqueurs de vaccins, mais sont encore restreintes à de petites cohortes.

5.2. Protéomique

5.2.1. Cytométrie en flux

La cytométrie en flux a été conçue d'abord pour permettre le phénotypage des cellules du système immunitaire (**Figure 10**). Par le biais de mesures optiques, la cytométrie en flux compte et analyse les particules d'une taille préspécifiée. L'échantillon est aspiré à travers un tube en silicone. Ensuite, l'échantillon est injecté dans une solution saline. Ainsi, cette solution englobe l'échantillon qui est au centre d'une cuvette d'écoulement. La nature du cette solution dépend de

l'échantillon et doit avoir un indice de réfraction similaire à celui du liquide englobant les cellules à analyser. Ainsi, l'échantillon est progressivement étiré jusqu'à la formation d'une "ligne de particules". Les particules sont séparées et alignées par une focalisation hydrodynamique le long de la ligne de courant du fluide. Des Acs jumelés aux fluorochromes permettent de mesurer des protéines à la surface de la cellule. Les particules sont excitées par un laser et l'absorbance mesurée par un capteur. Un fluorochrome excitable à une longueur d'onde spécifique est ainsi mesuré pour chaque cellule, d'où la mesure de l'expression de plusieurs protéines à la surface d'une même cellule. Une perméabilisation de la membrane cytoplasmique permet également de mesurer des protéines intracellulaires (technique de cytométrie intracellulaire ou ICS).

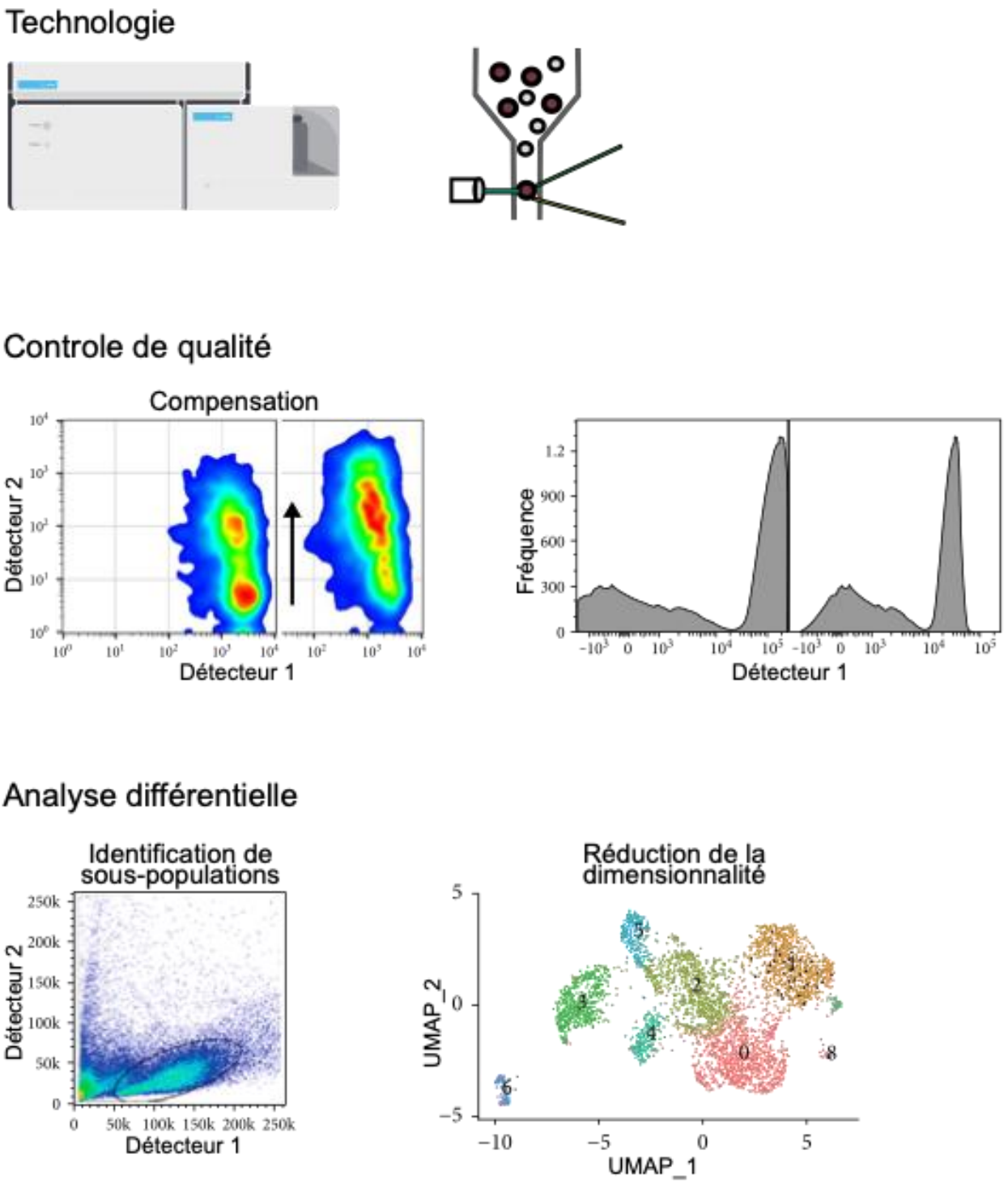


Figure 10. Principe de fonctionnement de la cytométrie en flux, contrôle de qualité et analyse différentielle

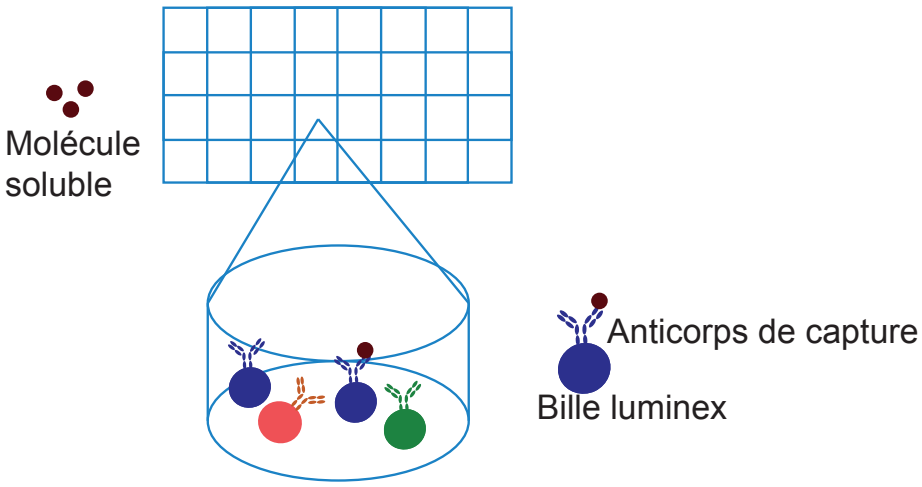
(Haut) La technologie des cytométrie est basée sur la lecture d'une cellule marqué par des Acs jumelés à des fluorochromes. (Milieu) La transformation mathématique des intensités permet de

mieux differentier l'expression des marqueurs (Bas) Les données sont ensuite analysées pour identifier des sous-populations de cellules d'intérêt.

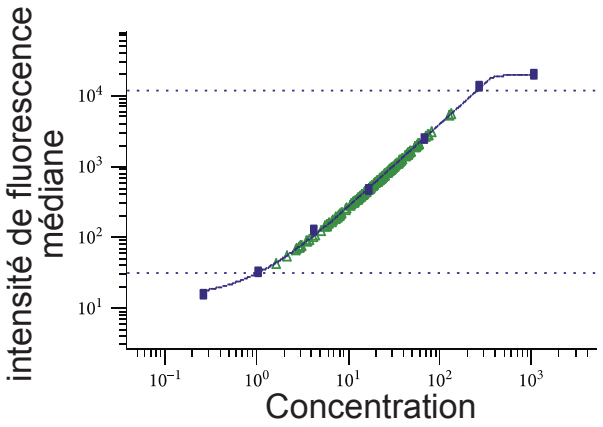
5.2.2. Protéomique du plasma

Les mesures de cytokines (facteurs solubles produits par les cellules du système immunitaire) et de chimiokines (facteurs solubles qui contrôlent la migration des cellules du système immunitaire) dans un échantillon sont essentielles pour la caractérisation de l'homéostasie et des fonctions des cellules effectrices du système immunitaire (**Figure 11**). Les techniques de protéomique du plasma permettent de mesurer ces protéines qui sont produites par les lymphocytes B et T ainsi que les cellules de la réponse immunitaire innée et qui s'accumulent dans le sang. Un échantillon de plasma ou de sérum est ajouté à un mélange de billes de couleur auxquelles sont liés par liaison covalente des Acs de capture spécifique aux protéines à mesurer. Les Acs se lient aux protéines d'intérêt. Des Acs de détection biotinylés spécifiques des protéines d'intérêt sont ajoutés et forment un sandwich Acs-Ag. La streptavidine conjuguée est couplée à un fluorochrome comme la phycoérythrine et se lie aux Acs de détection qui sont biotinylés. Les billes sont lues sur un instrument de détection basé sur des flux à double laser, tel que l'analyseur Luminex. Un laser classe la bille et la protéine qui y est associée. Le deuxième laser détermine l'amplitude du signal dérivé de la phycoérythrine, qui est directement proportionnelle à la quantité de protéine liée.

Technologie



Controle de qualité



Analyse différentielle

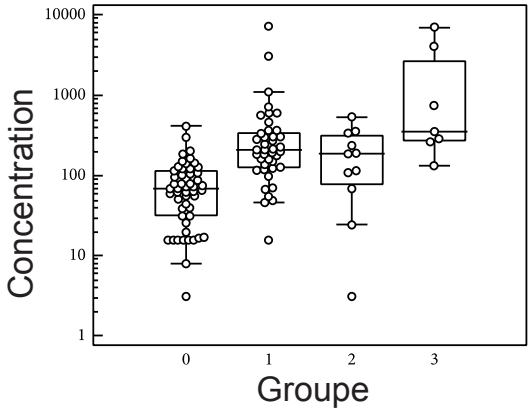


Figure 11. Principe de fonctionnement de la plateforme Luminex, contrôle de qualité et analyse différentielle

(Haut) la technologie Luminex est basée sur la lecture de protéines solubles liées par des Acs jumelés à des fluorochromes. (Milieu) L'utilisation d'échantillons gabarits permet de quantifier la concentration des protéines par échantillon (Bas) les données sont ensuite analysées pour identifier des protéines différentiellement exprimées entre groupes d'échantillons d'intérêt.

5.2.3. Sérologie

Les techniques large échelle peuvent aussi être utilisées pour déterminer la quantité et la qualité des Acs dans le sang. Historiquement les mesures sérologiques se sont limitées à la mesure du titre d'Acs et l'activité de neutralisation spécifique à des antigènes sélectionnés ou des virus d'intérêt. Toutefois, des Acs non neutralisants peuvent également avoir une activité effectrice antivirale. Jumelés à des cellules effectrices telles que les monocytes ou les cellules NK, ces Acs peuvent induire la cytotoxicité cellulaire dépendante de l'anticorps (ADCC), la phagocytose cellulaire dépendante de l'anticorps (ADCP) et le dépôt de complément anticorps-dépendant (ADCD). Par exemple, IgG capable de reconnaitre la loupe V1/V2 du VIH, peut lier des cellules infectées et être reconnue par des monocytes et NK menant à la mort des cellules infectées par le VIH (165).

Grâce aux avancées technologiques en sérologie, il est maintenant possible de mesurer la quantité et la fonction de ces Acs à large échelle. Basé sur une technologie similaire à la protéomique plasmatique (Luminex), il est possible d'avoir des billes sur lesquelles sont attachées différents antigènes, exprimant différents récepteurs aux Acs et utiliser des essaies in vitro avec différents sous-types cellulaires pour évaluer la quantité et la qualité des Acs démontrant ces activités effectrices dans un échantillon (166).

Les données transcriptionnelles sont utilisées dans l'ensemble des articles de cette thèse. Les deux premiers articles montrent la synergie que peuvent apporter des mesures de protéines dans le plasma et cellulaires pour l'identification de biomarqueurs et agents mécanistiques de la réponse aux vaccins de l'HepB et du VIH.

6. Analyses bio-informatiques des données large échelle

6.1. Prétraitement des données

Les principales difficultés engendrées par l'analyse de données de biopuces concernent (1) la haute dimensionnalité des données due au fait que les expressions de plusieurs milliers de gènes

sont mesurées simultanément dans un nombre restreint d'échantillons (2) le fort bruit de fond des données associé aux sources de variation de chaque étape du processus menant à l'obtention des données brutes et (3) la colinéarité entre les mesures vu que les niveaux d'expression des gènes sont fortement interdépendants.

Heureusement, il est possible d'estimer *a posteriori* certains effets non spécifiques engendrés lors du processus expérimental et de réduire leur impact en employant diverses approches de prétraitement des données.

Les expériences de biopuces sont coûteuses, nécessitent beaucoup de temps et produisent des données avec un bruit de fond non spécifique. De nombreux contrôles de qualité sont réalisés à ce stade de l'analyse, incluant une évaluation de l'homogénéité des pixels pour une même sonde, l'élimination des artéfacts physiques sur la biopuce, l'évaluation de l'homogénéité et intensité du bruit de fond. Une fois ces contrôles de qualité effectués, la normalisation des données est nécessaire pour éliminer tout biais systématique dans les données (p. ex. effet de *batch*, mauvaise calibration du scanneur). De nombreuses méthodes permettent de normaliser ces données, mais aucune ne fait l'unanimité au sein de la communauté bio-informatique (**Figure 9**). Le choix de la méthode de normalisation dépend essentiellement du type de biopuce utilisé et du biais attendu ou observé. Ces différentes méthodes (p. ex. MAS 5.0 (167), RMA (168)) ne seront pas détaillées dans cette thèse, mais n'en restent pas moins importantes pour l'exploitation des données de biopuces (169).

6.2. Analyses différentielles

Une fois les données large échelle sont prétraitées, elles sont soumises à une analyse différentielle visant à identifier des gènes différentiellement exprimés entre des groupes d'échantillons d'intérêt (ex. participants qui répondent ou non à un vaccin) ou des gènes qui sont corrélés à une variable continue (ex. réponse aux Acs après la vaccination). Les analyses statistiques pour identifier ces gènes différentiellement exprimés/corrélés reposent sur les régressions linéaires où les gènes sont utilisés comme variables dépendantes et la variable catégorique/continue est utilisée comme variable indépendante. Un t-test (standard or modéré) est utilisé pour tester que le coefficient de régression est différent de zéro. Étant donné qu'une petite fraction des gènes du génome sont attendus d'être différentiellement exprimés, le t-test est souvent modéré pour prendre en considération le nombre attendu de gènes différentiellement exprimée. Cette méthode est implémentée dans le paquet R LIMMA.

6.3. Enrichissement de terme biologique

L'une des principales limitations d'une signature d'expression transcriptionnelle résultant d'une approche supervisée (utilisant les groupes/variables d'intérêt pour identifier les marqueurs géniques) est que souvent elle ne représente aucun aspect biologique de la maladie étudiée. Il est toutefois possible *a posteriori* de déterminer si certains groupes de gènes prédéfinis sont associés aux classes (groupes d'échantillons présentant une expression similaire des marqueurs géniques) identifiées par une signature génique, ce qui facilite l'interprétation biologique des données. Ces groupes de gènes peuvent englober des gènes impliqués dans une même voie de signalisation moléculaire, des gènes associés à une même fonction biologique, des gènes cibles d'un même facteur de transcription, etc. Cette étape de l'analyse de biopuces à ADN est souvent appelée "analyse d'enrichissement".

Plusieurs méthodes statistiques ont été proposées pour effectuer des analyses d'enrichissement. Khatri P. *et collab.* ont proposé à cet égard une classification des méthodes d'enrichissement de voies de signalisation; de plus ils effectuent une revue assez complète des méthodes disponibles pour ces analyses(170). On y distingue notamment trois types d'approches : (1) les approches de surreprésentation basées sur l'évaluation statistique de l'intersection entre deux listes de variables binaires (p. ex. gènes régulés/non régulés vs gènes impliqués/non impliqués dans une voie de signalisation; **Figure 12**); (2) les approches de niveau-gènes qui prennent la statistique d'expression différentielle comme point de départ pour l'analyse; (3) les approches basées sur la topologie des réseaux de régulations.

La seconde catégorie d'approches, les approches de niveau-gènes, est la plus utilisée en pratique. Les différentes implémentations de ces approches diffèrent entre elles par les critères suivants : (1) la statistique d'expression différentielle employée pouvant être le ratio des expressions (*Fold Change*), la statistique *t*, la statistique *t* régularisée, le coefficient de régression, le coefficient de corrélation, etc. (2) la transformation appliquée à la statistique : rang, valeur *p*, etc. (3) la statistique pour le groupe de gènes : moyenne des rangs, *maxmean* , moyenne, médiane, etc. (4) le type d'hypothèse nulle testant soit que l'association entre un groupe de gènes et la variable d'intérêt est différente de celle du reste des gènes (p. ex., test de permutation sur les gènes) ou soit que l'association entre un groupe de gènes et la variable d'intérêt n'est pas due au hasard (p. ex., test de permutation sur les phénotypes) (**Figure 12**).

Quoiqu'il s'agisse d'un sujet très intéressant, il serait difficile de fournir des explications détaillées de l'ensemble des méthodes d'analyse d'enrichissement proposées dans la littérature à ce jour. Les principes clefs et les limites des approches niveau-gènes seront toutefois décrits brièvement

afin de pouvoir apprécier davantage l'utilité de ces approches dans l'interprétation biologique d'une classification basée sur une approche supervisée.

Contrairement aux approches de surreprésentation, les approches niveau-gènes (l'approche la plus populaire étant GSEA) ne reposent pas sur un seuil de détection arbitrairement qui discrimine les gènes "associés" ou non à un phénotype, elles utilisent l'ensemble des gènes pour évaluer statistiquement une association avec un phénotype donné. Les approches niveau-gènes utilisent l'intensité d'expression d'un gène pour quantifier la force de l'association d'un groupe de gènes avec un phénotype alors que les approches de surreprésentation ne décrivent un gène que par des états discrets (p. ex. induit, réprimé, non altéré). Une des assomptions des approches de surreprésentation est que l'expression d'un gène est indépendante de l'expression d'un autre gène du jeu de données alors qu'aucune assomption de ce genre n'est effectuée avec les approches niveau-gènes. Les approches niveau-gènes ne sont toutefois pas dépourvues de limites. En effet, les approches niveaux-gènes sont basées sur les assomptions que l'information contenue dans un groupe de gènes "corégulés" peut être résumée par une seule statistique d'expression différentielle et que chaque voie de signalisation est indépendante des autres. Un exemple biologique où ces assomptions ne sont pas respectées et les répercussions possibles sur l'interprétation des données sont présentés dans l'article #1 et élaborés dans la discussion de cette thèse.

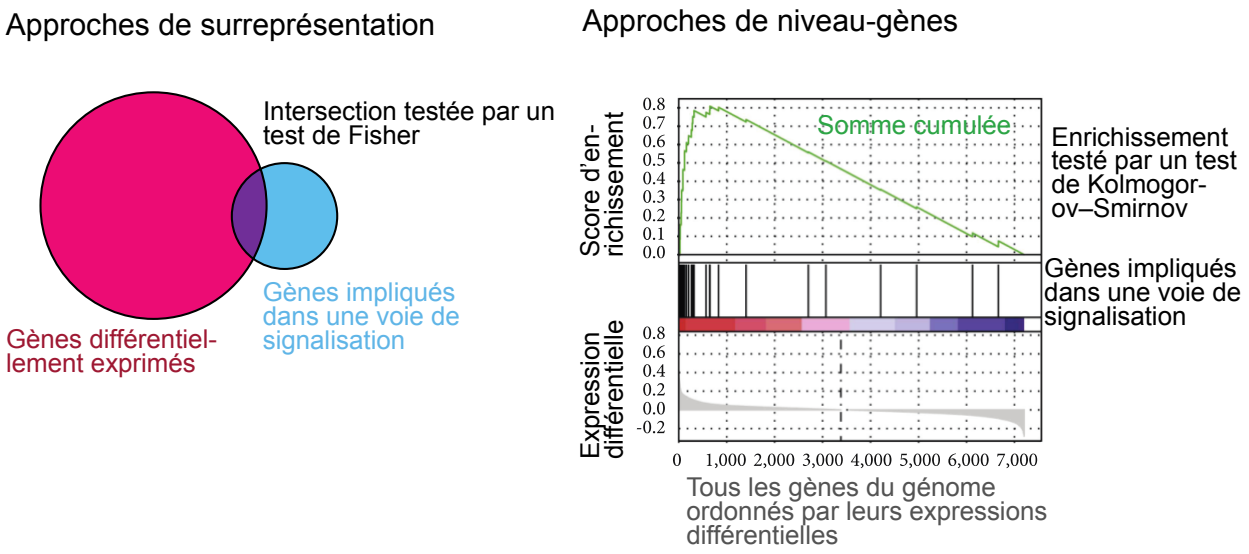


Figure 12. Approches pour évaluer l'enrichissement de terme biologique

(Gauche) Approches de surreprésentation testant l'intersection entre la liste de gènes différentiellement exprimés (nécessite de fixer un seuil considéré significatif) et une liste de gènes impliqués dans une voie de signalisation. (Droite) Approches de niveau-gènes, testant pour un

enrichissement de gènes impliqués dans une voie de signalisation en amont out en aval de tous les gènes ordonnés par leurs expressions différentielles.

Déterminer la fonction biologique partagée d'une liste de gènes n'est souvent pas suffisant pour déterminer les mécanismes biologiques subséquents à une perturbation; l'identification des cellules qui expriment les gènes exprimés dans ces signatures peut également apporter des informations déterminantes pour l'interprétation des données transcriptionnelles.

6.4. Déconvolution de sous populations cellulaires

L'expression d'un gène dans une population cellulaire hétérogène dépend du niveau d'expression spécifique du gène dans chacun des types cellulaires inclus dans cet échantillon et de la fréquence de ce type cellulaire dans l'échantillon en question. Dans le cas où les proportions des différents types cellulaires d'un échantillon ne sont pas connues *a priori*, il est donc nécessaire de les estimer afin de pouvoir définir l'impact de la distribution des différents types cellulaires dans le profil transcriptionnel observé.

Soit un mélange de k types cellulaires différents et ψ_k les profils moyens d'expression. On peut alors utiliser ces profils pour estimer les quantités q_k des différents types cellulaires du mélange. Une première approche consiste à identifier pour chaque type cellulaire un sous-ensemble S_k de gènes caractéristiques de ce type cellulaire tel que le niveau des gènes appartenant à S_k dans les types cellulaires autres que le type k soit négligeable par rapport au niveau observé de ces gènes dans le type k. On peut ensuite utiliser une méthode d'enrichissement de gènes comme un indicateur de la quantité du type k dans cet échantillon. Cette approche demeure imparfaite pour plusieurs raisons :

- Elle postule qu'il est possible d'identifier des transcrits qui soient spécifiques de chaque type cellulaire. Or, on observe généralement de très fortes corrélations des niveaux des transcrits entre les différents types cellulaires sanguins. De plus, la majeure partie des transcrits considérés ici comme spécifiques d'un type cellulaire sont en réalité exprimés à des niveaux parfois élevés dans d'autres types cellulaires. Cette approche a donc tendance à surestimer les fréquences inférées de chacune des sous-populations cellulaires.
- Cette méthode ne fournit qu'un indicateur relatif de la quantité des différents types cellulaires entre les individus et ne permet pas d'estimer la fréquence absolue d'un type cellulaire dans un échantillon.

Afin de corriger ces défauts, une autre approche peut être utilisée. Cette approche consiste à écrire la matrice d'expression X comme un produit matriciel

$$X \approx \Psi \times Q$$

où Ψ est la matrice des profils d'expressions ψ_k des différents types cellulaires et Q est la matrice des quantités q_k. On applique une procédure en 2 étapes : 1. On estime les profils d'expression ψ_k de chaque type cellulaire, à partir des atlas d'expression publics. On utilise la matrice des profils estimés $\widehat{\Psi}$ pour estimer les quantités Q par des estimateurs des moindres carrés ordinaires du type :

$$Q = \left(\widehat{\Psi}'\widehat{\Psi}\right)^{-1}\widehat{\Psi}X$$

Pour la première étape qui consiste à estimer les profils ψ_k, nous pouvons utiliser les données de Nakaya et coll. Une étape où la moyenne des profils d'expression génique des différents échantillons est calculée pour chacun des types cellulaires considérés afin d'obtenir les profils types $\widehat{\psi_k}$. On applique ensuite l'algorithme de déconvolution à la matrice $\widehat{\Psi}$ formée. Notons les quantités $\widehat{q_{kj}}$ estimées pour chaque individu, qui une fois sommées totales 1.

On estime dans un deuxième temps les proportions Q des différents types cellulaires à partir des profils ψ_k. Pour atteindre cet objectif, des contraintes de positivité sont imposées sur les quantités estimées par un processus itératif :

1. Estimer pour chaque échantillon x_j les quantités des différents types cellulaires

$$\widehat{q_j} = \left(\widehat{\Psi}'\widehat{\Psi}\right)^{-1}\widehat{\Psi}x_j$$

2. Si certaines quantités sont négatives:
    a. Fixer à zéro la quantité du type cellulaire, dont la quantité estimée $\widehat{q_{kj}}$ est minimale
    b. Répéter l'estimation pour les autres quantités $\widehat{q_{kj}}$.
3. Répéter l'opération précédente jusqu'à ce que toutes les quantités du vecteur q_j soient positives ou nulles.

En résumé, ces méthodes *in silico* sont capables d'estimer la fréquence de sous-population cellulaires dans un échantillon hétérogène et nous permettent de quantifier l'impact de la fréquence de sous-populations de cellules sur l'expression génique. Jumelé avec les analyses d'enrichissement de termes biologiques, ces méthodes nous ont permis, dans les articles de cette thèse, de proposer des mécanismes biologiques pouvant expliquer les variations d'expression génique associées à la réponse aux infections virales et aux vaccins.

6.5. Analyse intégrative

Les analyses intégratives visent à analyser simultanément des données multiomiques afin d'identifier des sous-groupes d'échantillons présentant des profils d'expression homogènes, d'identifier des biomarqueurs et de définir les mécanismes biologiques identifier par les données large échelle.

Les méthodes intégratives peuvent être séparées en six groupes bases sur les modèles mathématiques utilisés pour analyser conjointement des données larges échelles : (1) les méthodes de corrélations (2) méthodes de similarité (3) méthodes basées sur les réseaux (4) méthodes bayésiennes (5) les modes de fusion (6) les méthodes multivariées (**Figure 13**).

Les méthodes intégratives sont empiriquement avantageuses comparer à l'analyse d'un seul type de donnes ou compare à l'analyse en série des données large échelle.

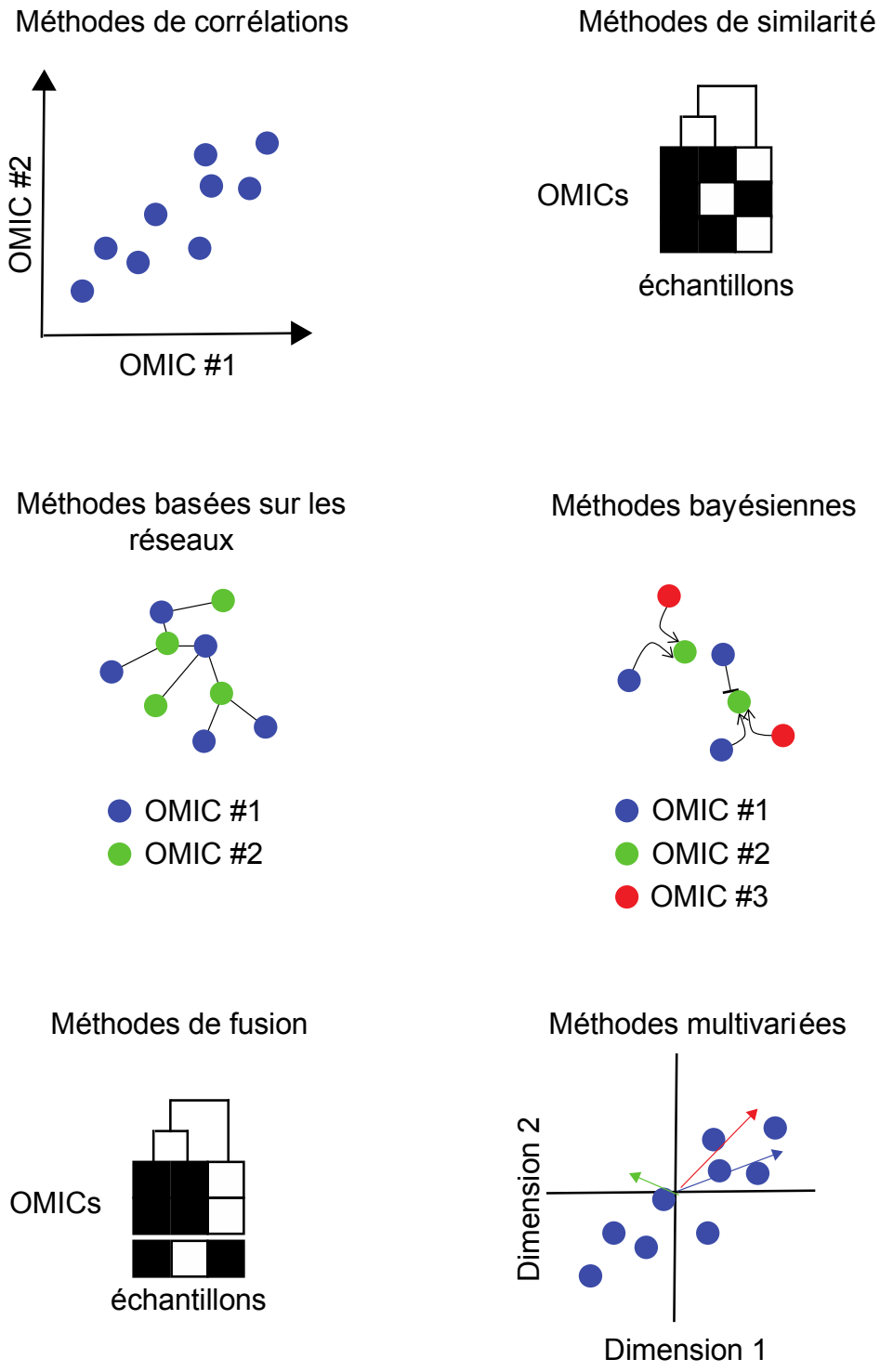


Figure 13. Illustration des six types de méthodes d'intégration multiomiques

Les méthodes d'intégration multiomiques peuvent être classifiées dans six groupes. Ces six groupes de méthodes sont décrit ci-dessous.

Chacun des six types de méthodes visant à explorer simultanément des données multiomiques pour identifier des sous-groupes d'échantillons aux profils similaires ou à connecter les éléments biologiques (ex. gènes, protéines) évalués par différentes plates-formes à large échelle sont présentés de façon simplifiée dans la **Figure 12**.

6.5.1. Méthodes de corrélations

Les méthodes de corrélations sont basées sur l'utilisation de la corrélation de Pearson ou la corrélation non paramétrique de Spearman (base sur le rang) pour combiner les multiples jeux de données mesurés sur les mêmes échantillons. Ces méthodes sont les plus simples utilisées pour les analyses multiomiques. Un exemple d'implémentation d'une méthode de corrélation pour l'intégration de données de variation de copies de gènes, de méthylation d'ADN et de transcription est la méthode CNAmet (171). Toutefois ces méthodes ne permettent pas de corriger pour les biais liés aux technologies utilisées pour générer ces données.

6.5.2. Méthodes de similarité

Les méthodes de similarité sont essentiellement utilisées pour l'identification de sous-groupes d'échantillons définis par un ensemble de des données multiomiques. Deux individus sont regroupés dans un même sous-groupe, si les individus sont regroupés ensemble lorsque 1) chaque type de données est partitionné séparément et 2) en utilisant différentes méthodes de partitionnement (ex. partitionnement hiérarchique, agglomératif). Une implémentation d'une méthode de similarité pour l'intégration de données de variation de copies de gènes, de polymorphismes nucléotidiques et de transcription est la méthode PINSplus (172).

6.5.3. Méthodes basées sur les réseaux

Les méthodes basées sur les réseaux d'interactions existants entre gènes, l'ARN, protéines pour combiner différents jeux de données. Des algorithmes permettant de traverser le réseau sont utilisés pour connecter différents éléments par échantillon (ex. algorithme de Floyd pour identifier les chemins les plus courts entre toutes les paires de nœuds). Une implémentation d'une méthode basée sur les réseaux pour l'intégration de données de variation de copies de gènes, de polymorphismes nucléotidiques et de transcription est la méthode NetICS (173). Les limites de ces méthodes incluent le fait qu'elles reposent sur l'établissement préalable de réseaux d'interaction. Ces méthodes tentent d'identifier des interactions entre gènes, ARN, protéines impliquant des nœuds ayant une forte connectivité.

6.5.4. Méthodes bayésiennes

Les méthodes bayésiennes utilisent également les réseaux d'interactions existant entre gènes, l'ARN, protéines pour combiner différents jeux de données. Ces méthodes utilisent le théorème de Bayes pour déterminer l'activité d'un gène dans le réseau, où la probabilité d'activation d'un gène est décrite mathématiquement comme le produit des probabilités d'induction du nombre de copies d'ADN, le niveau de l'ARN et le niveau de protéine. Un exemple d'implémentation de cette méthode est PARADIGM (174). Ces méthodes sont aussi limitées par la qualité des réseaux d'interaction et tentent d'identifier des nœuds avec une forte connectivité.

6.5.5. Méthodes de fusion

Les méthodes de fusion sont essentiellement utilisées pour l'identification de sous-groupes d'échantillons bases sur des données multiomiques. Les méthodes de fusion sont proches des méthodes de similarité et combinent les éléments (ex. expression d'ARN et protéique) d'un échantillon uniquement si l'individu est partitionné dans le même sous-groupe pour différents types de données. Une implémentation d'une méthode de similarité pour l'intégration de données de variation de copies de gènes, de polymorphismes nucléotidiques et de transcription est la méthode PFA (175).

6.5.6. Méthodes multivariées

Les méthodes multivariées reposent sur les techniques de réduction de dimension de données. En effet, une fois la dimensionnalité de diffèrent types de données est réduite (ex. en utilisant la technique de PCA où chaque dimension correspond à la combinaison linéaire des éléments) ces dimensions peuvent être combinées à travers types de données. Une implémentation de ce type d'approches est disponible dans l'application mixOmics (176).

Une autre méthode d'intégration entre données d'accessibilité de chromatine et de transcriptomique implémentant une méthode multivariée est la méthode d'ancrage décrite dans Seurat (177). Cette méthode repose sur la réduction de dimensionnalités et sur l'identification d'ancres qui partagent des voisins similaires entre les données à intégrer.

6.5.7. Critères de sélection de méthodes d'intégration

Le choix d'une stratégie d'intégration repose principalement sur l'objectif de l'analyse intégrée et ce soit pour l'identification de sous-groupes d'échantillons ou encore pour identifier l'association entre variables de différents essais mesurés sur de mêmes donneurs ou différents donneurs. Certaines métriques peuvent être utilisées pour comparer différentes approches, par exemple en

testant l'association de marqueurs connus a priori pour être fortement associés (ex. fréquences de lymphocyte B et l'expression ARN et protéique des marqueurs de surface CD79A et CD79B). Dans une étude comparative de différentes méthodes d'intégration, les méthodes multivariées (tel que RGCCA, mixOmics) sont parmi les deux meilleures méthodes pour la création de sous-groupes d'échantillons et pour identifier des connexions biologiques (178). Finalement, la facilite d'implémenter/d'utiliser ses méthodes et d'interpréter les résultats doit être également pris en compte. Dans cette thèse, l'approche multivariée implémentée dans le paquet mixOmics a été sélectionnée sur la base des critères listés ci-dessus (facile à utiliser/implémenter, résultats informatifs pour l'interprétation des données multiomiques et une des meilleures approches dans des études comparatives (178)).

6.6. Prédiction (modèles)

Les approches supervisées visent à identifier les liens entre les données d'entrée (p. ex. données d'expression génique) et une variable d'intérêt (p. ex. survie sans récidive ou la réponse à un traitement). Les approches supervisées nécessitent donc de connaître *a priori* le nombre d'instances de la variable d'intérêt.

Les approches supervisées débutent par une phase d'apprentissage au cours de laquelle un modèle est construit à l'aide de données d'entraînement pour lesquels on connaît la valeur de la variable d'intérêt. Puis, le modèle est utilisé avec d'autres données (les données tests) dont la valeur de la variable d'intérêt est inconnue. Le type de modèles dépend de la nature de la variable d'intérêt. On parle de classification pour les variables discrètes, de régression pour les variables continues et d'analyse de survie pour les données de survie. Nous traiterons de la classification dans le reste de cette section.

Un grand nombre des méthodes de classification ont été testées sur les données d'expression génique et la qualité des résultats s'est avérée être très dépendante des jeux de données utilisés et des différents paramètres expérimentaux (179). Le choix d'une méthode de classification reste donc aujourd'hui un choix subjectif basé essentiellement sur le principe d'essais et erreurs.

Les différentes méthodes de classification sont un sujet d'étude à part entière. Nous limiterons donc cette présentation introductive aux méthodes utilisées dans les articles de cette thèse.

6.6.1. Évaluation de la qualité de prédiction

Un classificateur peut être évalué sur la base de (1) la précision de la classification, (2) à quel point les résultats de la classification sont interprétables et (3) la facilité d'utilisation du classificateur.

La qualité de la prédiction du classificateur est caractérisée par deux critères complémentaires qui sont la sensibilité (fraction des prédictions positives qui sont détectées correctement) et la spécificité (fraction des prédictions négatives qui sont détectées correctement). Plus la sensibilité et la spécificité d'un classificateur sont importantes et plus la précision (*accuracy*) sera élevée conduisant alors à un faible nombre d'erreurs de classification.
L'analyse ROC (*Receiver Operating Characteristic*) est utilisée pour mesurer la précision d'un classificateur binaire quand le seuil de discrimination est variable. Graphiquement, on représente souvent les résultats d'une analyse ROC sous la forme d'une courbe sensibilité/(1 — spécificité). Les courbes sensibilité/(1 — spécificité) sont des outils graphiques permettant de représenter la capacité d'un classificateur à discriminer les populations de deux classes.
Un index simple et quantitatif de la précision d'un classificateur est de mesurer l'aire sous une courbe sensibilité/(1 — spécificité) (AUC). L'AUC indique la probabilité que la fonction de discrimination d'un classificateur permette de classifier les échantillons dans la bonne classe. Lorsque l'AUC est égale à 1, le classificateur ne commet aucune erreur de classification. Lorsqu'elle est égale à 0.5, le classificateur est non informatif et fait autant de prédictions correctes qu'incorrectes.

Pour un problème d'estimation linéaire, la performance d'un prédicteur est souvent évaluée par l'erreur quadratique moyenne (MSE). Le MSE est la mesure de la dispersion des estimes autour de la vraie valeur.

$$MSE_{\hat{\theta}} = E[(\hat{\theta} - \theta)^2]$$

ou $\hat{\theta}$ est l'estimateur et θ est le vecteur d'observation.

6.6.2. Choix de la technique d'apprentissage machine

Il peut être difficile de déterminer l'importance d'un gène dans le classificateur global, si une méthode complexe (i.e. non-linaire) est adoptée. Une méthode plus simple est alors préférée même si elle a une précision réduite. Vu les ressources nécessaires pour réaliser des expériences de biopuces, il est parfois important de développer un classificateur basé sur le moins de gènes possible tout en maintenant une précision correcte (180).
Dudoit S. *et collab.* ont testé plusieurs techniques d'apprentissage machine pour la prédiction de classes de différents cancers et ont démontré l'efficacité des méthodes les plus simples comme les *k*-plus proches-voisins ou l'analyse discriminante diagonale linéaire (181). Lee *et collab.* ont comparé 21 méthodes sur sept jeux de données, et ont conclu au contraire à la supériorité des méthodes plus sophistiquées telles que les machines à vecteur de support (182). Ces deux

travaux présentant des résultats contradictoires et reflètent la difficulté d'identifier *a priori* quelles méthodes produisent les meilleurs résultats.

Plusieurs problèmes engendrés par l'utilisation des données issues de biopuces pour la classification (approches supervisées) ont été soulignés par plusieurs groupes de recherche (183, 184). Ces difficultés incluent les problèmes de surapprentissage dû à la haute dimensionnalité des données par rapport au nombre d'échantillons analysés (185); l'absence de réelle validation due à l'absence de jeux de données indépendants ou de l'utilisation erronée de la validation croisée (186). Dupuy A. et Simon R. ont estimé que près de la moitié des articles produisant un classificateur basé sur les données de biopuces commettent une des erreurs citées plus haut (187).

Différentes études ont démontré que ces problèmes peuvent engendrer des signatures instables, lorsqu'entraînées sur des jeux indépendants (179). Pour Ein-Dor *et collab.,* cette instabilité est attribuée au fait que les jeux étudiés sont hétérogènes (du point de vue clinique, pathologique), que les plates-formes utilisées pour analyser les profils d'expression sont variables (Affymetrix, Agilent, Illumina) et qu'une même méthode d'analyse peut aboutir à de nombreuses signatures géniques différentes présentant des précisions similaires. Ils montrent que le nombre d'échantillons nécessaires afin d'obtenir une signature stable (minimum de 50 % de gènes communs) entre deux jeux de données transcriptionnelles indépendantes serait de plusieurs milliers (188).

Malgré qu'aucun article de cette thèse ne possède un tel nombre d'échantillons, il est nécessaire d'essayer d'évaluer la robustesse de signatures tout en reconnaissant la possibilité que d'autres signatures puissent être découvertes dans le même jeu de données.

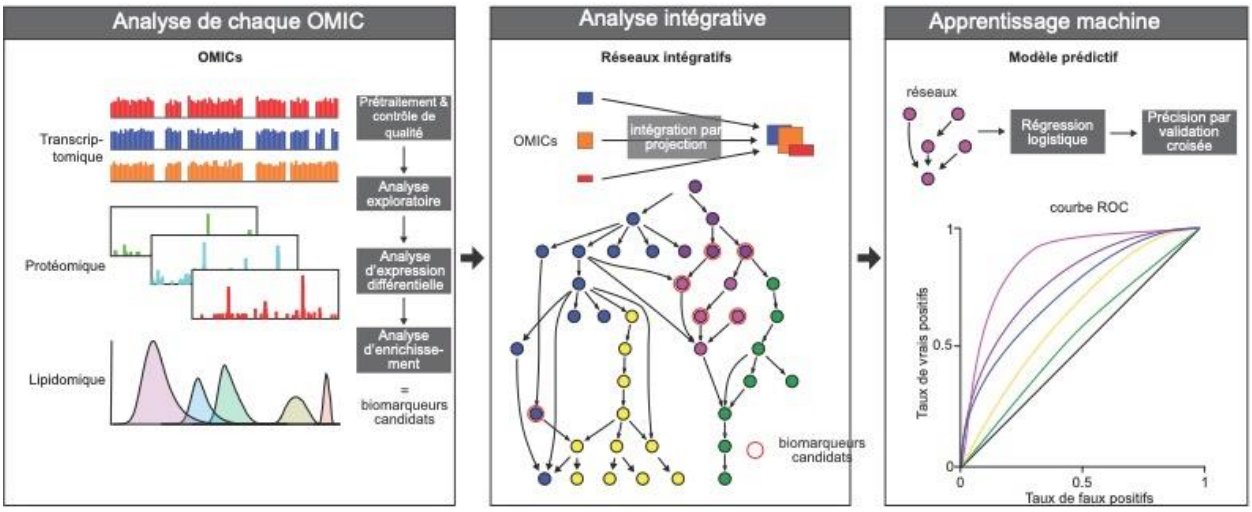


Figure 14. Schéma décrivant la stratégie développée dans la thèse pour répondre aux différentes

questions et hypothèses testées

Les trois articles de cette thèse montrent comment des approches d'apprentissage machine semi-supervisé (1er article) et supervisé (2e et 3e articles) peuvent être utilisés pour identifier des marqueurs de la réponse aux vaccins capables de prédire la sévérité des symptômes après une infection virale. Le troisième article de cette thèse inclut une analyse comparative des différentes méthodes ainsi que des modèles mathématiques. Ceci a permis d'identifier les méthodes d'apprentissage les plus efficaces permettant de définir des biomarqueurs de la réponse immunitaire (**Figure 14**).

7. Objectifs et hypothèses

La thèse repose sur l'hypothèse que des profils transcriptionnels générés sur les globules blancs du sang périphérique prévaccination peuvent identifier des mécanismes cellulaires et moléculaires pouvant prédire la réponse aux vaccins. De plus, une analyse bio-informatique intégrative de plusieurs types de données multidimensionnelles peuvent permettre d'approfondir nos connaissances des mécanismes d'action des vaccins.

8. Problématique

La médecine personnalisée repose sur le principe de prédiction de la réponse à une thérapie en tenant compte de la diversité entre individus. En pratique, certains outils manquent afin de réaliser ce but. La prédiction de la réponse vaccinale est un exemple d'outil clinique qui serait d'un apport conséquent, mais qui est aujourd'hui plus du domaine de la théorie sans exemple concret en application clinique. De plus, la meilleure stratégie analytique permettant l'obtention de tels prédicteurs n'a pas été résolue. À cette fin, les articles de cette thèse visent à définir une méthodologie qui permettrait l'identification de prédicteurs de la réponse vaccinale. L'étude de ces prédicteurs pourrait également avoir un apport important sur nos connaissances mécanistiques de la réponse aux vaccins.

Résultats

Article #1 : Pre-vaccination inflammation and B-cell signalling predict age-related hyporesponse to hepatitis B vaccination

Mise en contexte

Dans ce premier article nous tenterons de définir les mécanismes qui font que certains facteurs sociodémographiques tels que le vieillissement aboutissent à une faible réponse vaccinale. Plusieurs voies de signalisations sont affectées lors du vieillissement. Notre connaissance des voies de signalisations qui sont responsables de cette faible réponse aux vaccins demeure limitée.

Dans cette étude, des adultes âgés de plus de 65 ans, naïfs pour l'infection par le virus de l'HepB ont reçu trois vaccins, dont un contre l'HepB.

L'utilisation de données transcriptionnelles et cytométriques du sang d'individus prévaccination a été explorée pour déterminer s'il était possible de développer un classificateur de la réponse au vaccin contre l'HepB, et de tenter d'identifier les mécanismes qui définissent une réponse protectrice à la suite de la vaccination.

Mon rôle dans ce projet s'est focalisé sur l'analyse complète des données transcriptionnelles, en commençant par la normalisation des jeux de données de micropuces et en finissant par l'analyse d'enrichissement des voies de signalisation au sein des gènes corrélés au taux d'Acs en réponse à la vaccination. J'ai aussi construit des classificateurs supervisés basés sur les données transcriptionnelles, de cytométrie en flux et des niveaux de cytokines et de chimiokines. L'approche semi-supervisée a été effectuée en collaboration avec un autre bio-informaticien. La rédaction des résultats et méthodes concernant les analyses bio-informatiques et statistiques m'a été assignée. La discussion et le résumé de l'article ont été écrits en collaboration avec les deux derniers auteurs de l'article. En résumé, 90% de l'analyse bio-informatique, et 75% de la rédaction de l'article sont le fruit de mon effort.

Pre-vaccination inflammation and B-cell signalling predict age-related hyporesponse to hepatitis B vaccination.

Slim Fourati[1,2], Razvan Cristescu[3], Andrey Loboda[3], Aarthi Talla[1,2], Ali Filali[1,2], Radha Railkar[3], Andrea K. Schaeffer[3], David Favre[4], Dominic Gagnon[4], Yoav Peretz[4], I-Ming Wang[3], Chan R. Beals[3], Danilo R. Casimiro[3], Leonidas N. Carayannopoulos[3], and Rafick-Pierre Sékaly[a,1,2]

Affiliations

[1]Vaccine and Gene Therapy Institute of Florida, Port St Lucie, Florida 34987, USA

[2]Department of Pathology, University Hospitals of Cleveland and Case Western Reserve University, 2103 Cornell Road, Cleveland, Ohio 44106, USA

[3]Department of Discovery Medicine, Merck Research Laboratories, Rahway, New Jersey 07065, USA

[4]Caprion/ImmuneCarta, Montréal, Québec, Canada, H1W 4A4

[a]Email: rafick-pierre.sekaly@case.edu

This work was originally published in Nature Communications:
Nat Commun. 2016 Jan 8;7:10369. doi: 10.1038/ncomms10369. PubMed PMID: 26742691; PubMed Central PMCID: PMC4729923.

Abstract

Aging is associated with hyporesponse to vaccination, whose mechanisms remain unclear. Hepatitis B virus (HBV) naïve older adults received three vaccines, including one against HBV. Transcriptional and cytometric profiling of whole blood collected before vaccination revealed that heightened expression of genes that augment B-cell responses and higher memory B-cell frequencies correlated with stronger responses to HBV vaccine. In contrast, higher levels of inflammatory response transcripts and increased frequencies of pro-inflammatory innate cells correlated with weaker responses to this vaccine. Increased numbers of erythrocytes and the heme-induced response also correlated with poor response to the HBV vaccine. A transcriptomics-based pre-vaccination predictor of response to HBV vaccine was then built and validated in distinct sets of older adults. This moderately accurate (AUC≈65%) but robust signature was supported by flow cytometry and cytokine profiling. This study is the first that identifies baseline predictors and mechanisms of response to the HBV vaccine.

Introduction

Aging confers elevated risk of illness and death from infection[1]. As the number and proportion of older individuals increase worldwide[2], prevention of severe or poorly treatable infections among the elderly is ever more pressing. Though vaccination offers a proven approach to such prevention, well-described age-related vaccine hyporesponses (VHR) blunt its potential benefit[3].

To date, studies of responses to vaccination have revealed associations with human leukocyte antigen polymorphisms[4], innate and adaptive immune cell phenotypes[5], suggesting that VHR results from the simultaneous interplay of many elements[6]. The impracticality of sampling human lymphoid organs coupled with limited translatability of animal models[7] poses a formidable challenge to the discovery of these mechanisms as well as biomarkers of VHR using single-hypothesis approaches. To overcome this challenge, high dimensionality studies of cellular and molecular responses to vaccines have been proposed to speed convergence to actionable mechanistic and biomarker hypotheses; an approach termed “systems vaccinology”[8]. This approach has been used to discover early post-vaccination mRNA signatures of responses to yellow fever vaccine YF17D and to correlate strong immunogenicity with cross-lineage cellular activation[9,10]. Studies of influenza vaccination have yielded insights into new signaling pathways for B cell regulation, the potential importance of early neutrophil responses to interferon exposure, and early predictors of post-vaccination immunity[11,12]. Systems vaccinology was applied to multi-vaccine comparisons[11,13] as well as to identification of pre-vaccination transcriptomic[14] and flow

cytometric (FCM) correlates of response[15]. Nevertheless, biomarkers predictive of the response to vaccination have yet to be validated and more importantly mechanisms that control the response to vaccination remain to be defined.

Here, we first identified a mRNA signature of aging that associated with the seroresponse to the hepatitis B virus surface antigen (anti-HBsAg) in naïve older adults. Then transcriptomics, polychromatic FCM and serum cytokine profiling were used to generate an integrated model to inform on potential involvement of specific cellular and molecular pathways in nascent anti-HBsAg responses in older adults.

Results

Clinical characteristics of the study cohort

Socio-demographic and serological characteristics of this study's cohort (designated as the EM131 cohort) appear in **Supplemental Table A1.1**. One hundred and seventy four healthy adults received vaccines against Hepatitis A/Hepatitis B viruses (HAV/HBV), Diphtheria/Tetanus toxoids and Cholera bacteria/toxin (**Fig. A1.1a**). The percentage of study participants that mounted protective Ab titers (anti-HBsAg: 10 mIU/mL, anti-Tetanus toxoid: 0.1 IU/mL, anti-Diphtheria toxoid: 0.1 IU/mL and anti-cholera toxin: dilution 1:40)[16-18] were, in decreasing order, 95% for Tetanus, 84% for Diphtheria, 71% for Cholera and 37% for HBV (**Fig. A1.1b**). Only the HBV vaccine involved an antigen to which all study participants were naïve and presented the lowest response rate (37%, after two injections) even when we restricted the analysis to participants with undetectable Ab titers pre-vaccination (**Supplementary Fig. A1.1a**). We focused our analysis on the HBV vaccine response (primary immunization). Kernel density estimation of anti-HBsAg titers showed that a cut-off of 5 mIU/mL demarcated study participants as HBV vaccine responders (Ab titers ≥ 5 mIU/mL) and poor-responders (Ab titers below detection threshold) (**Supplementary Fig. A1.1b**). Logistic regression between demographic variables and HBV vaccine response revealed that age and gender were significantly associated with response to the HBV vaccine on the EM131 data set (**Supplementary Table A1.2**). Here, we investigated the association of age-related transcriptomic changes with the response to HBV vaccination.

Impact of age on the human immune transcriptome

The impact of age on the transcriptome of peripheral blood mononuclear cells obtained from healthy participants was studied under controlled conditions (morning, after fasting) in the San Antonio Family Heart Study (SAFHS, n=1,240)[19]. Linear regression models were used to identify

transcripts correlated to age (**Methods** formula 1). Using a cutoff of adjusted $p \leq 0.05$ (with the Benjamini-Hochberg method to correct for multiple testing), we identified 1,143 and 1,142 transcripts positively and negatively correlated to age, respectively, in the SAFHS dataset (**Fig. A1.2a**). The Gap statistic method estimated as twenty the optimal number of clusters that explained gene expression variability (**Supplementary Fig. A1.2a**). *k*-means clustering was used to regroup the 2,285 transcripts in twenty modules (M1 to M20) based on their co-expression in the SAFHS dataset.

Pro-inflammatory pathways prevail in elderly participants

To identify the biological pathways associated to older age, we assessed the overlap between each of these modules and Ingenuity canonical pathways (Fisher's exact test). Using a cutoff of $p \leq 0.05$, we identified 250 biological pathways enriched in the twenty modules. On the one hand, pathways positively associated with older age included pathways that regulate cell motility (M12 and M14) and genes downstream of integrin signaling (M15). A type II interferon signaling module (M17) characterized by the induced expression of *TNF* and *IFNG* was also enriched in elderly participants. Inflammatory responses that include complement genes (M16) downstream of interferons, as well as T cell/NK cells mediated cytolysis (M19) markers such as perforin and granzyme B, were also positively associated with older age. Module M19 was highly enriched in NK cell markers such as the killer cell immunoglobin-like receptors (*KIR2DL2*, *KIR3DL1* and *KIR3DL3*). On the other hand, pathways enriched in younger participants included immunological pathways such as B-cell signaling (M1) comprising several B cell markers (*CD79A*, *CD79B*, *CD19*, *CD20* and *CD22*), T-cell receptor signaling (M6) that included *ZAP70*, *TRAF1*, *TCF7* a marker of memory T cells[20] and antiviral response (M8) that included the antiviral signaling gene *MAVS*. MYC signaling (M3), that triggers metabolic pathways such as amino acid metabolism (M2 and M4), fatty acid biosynthesis (M9) and lipid metabolic pathways (M11), all critical for the induction of T and B cell memory, was enriched in younger participants. These results indicate that each of the twenty modules reflect a specific biological process that is positively or negatively associated to aging. These results also confirm that elderly show constitutive upregulation of several pro-inflammatory pathways downstream of type I/II interferons which are suggested to favor immunosenescence[21].

An age-related gene-expression signature is observed prior to vaccination

We developed an aggregate score that integrates transcriptomic changes associated with older age by subtracting the average expression of the eleven modules negatively associated with age

from the average expression of the nine modules positively associated with age (**Methods** formula 2). We calculated the BioAge score on the SAFHS cohort and as expected observed a significant correlation between the BioAge score and chronological age (Pearson correlation: r=0.421, $p < 2.2e-16$; **Supplementary Fig. A1.2b**). The same modules that discriminated elderly and young participants in the SAFHS cohort also distinguished the two age groups in the EM131 cohort (**Fig. A1.2a-2b**). There was a statistically significant association between chronological age and BioAge on the EM131 cohort (**Fig. A1.2c**, Wilcoxon rank-sum test: p=1.01e-05), confirming the validity of this score to reflect the impact of age on gene-expression in whole blood cells. While most young study participants had a low BioAge score, the BioAge predictor divided elderly patients (≥ 65, EM131 elderly cohort) into two groups designated as BioAge young (n=62, age ∈ [65, 78]) and BioAge old (n=76, age ∈ [65, 83]) (**Fig. A1.2c**).

BioAge score associated with HBV vaccine response prior to immunization

We determined if different BioAge scores correlated with different responses to HBV vaccination. Both the BioAge score (**Supplementary Fig. A1.3**) and the two groups of elderly identified by the BioAge signature (Fisher's exact test: OR=2.14, p=0.0357; **Fig. A1.3a**) were associated with the rate of response to the HBV vaccine). Importantly, in a multivariate analysis including available clinical parameters, the BioAge was both a better predictor of the HBV response than chronological age and independent of gender (**Supplementary Table A1.2**). Among the twenty gene modules defined in the BioAge signature, two modules M1 (lower expression in BioAge old) and M16 (high expression in BioAge old) allowed the most accurate prediction of the response to the HBV vaccine (**Fig. A1.3b**). Pathway enrichment analysis showed that module M1 was enriched in genes encoding two critical components of the B cell receptor *i.e*. *CD79A* and *CD79B*, B cell activation markers such as *CD19*, *CD22*, the transcription factor *POU2AF1* (a protein essential for the response of B-cells to antigens and required for the formation of germinal centers)[22] and *MZB-1*, a chaperone involved in assembly of IgM[23] (**Fig. A1.3c**). Module M16 includes genes of the acute phase response *i.e.* the complement gene C1, the mediator of type II interferon response *MYD88* and the transcription factor *IRF7*. IRF7 target genes such as *LILRA5* (a marker expressed by TNF producing macrophages), *S100* (a molecule associated to cellular senescence), *IL-15* (a proinflammatory cytokine that triggers CD8 T cell and NK cell expansion), *PILRA* (an inhibitory receptor that acts through SHP-1) and *PRELID-1* (a gene that negatively regulates the development of TH2 responses) were also present in module M16 (**Fig.**

A1.3c). Overall, module M1 includes mostly genes involved in B cell activation while module M16 includes genes linked to inhibition of the B cell response.

To test the accuracy of this signature, we performed a receiver operating characteristic (ROC) analysis that indicated that the BioAge signature could predict the response to the HBV vaccine with an accuracy of 60.0% in the EM131 elderly cohort (**Fig. A1.3b**, permutation test: p=0.0163). Collectively these results show for the first time that transcriptional profiling allows the identification before vaccination, albeit with a moderate accuracy, of participants that will mount a poor response to the HBV vaccine. This signature highlights the interplay between the innate inflammatory pathways and B cells in the response to the HBV vaccine.

A 15-gene signature predicts the HBV vaccine response prior to immunization

We used an independent, supervised approach to identify other gene sets/pathways important for predicting the response to the HBV vaccine. The EM131 elderly cohort was randomly split into a training cohort (2/3 of the study participants) and a test cohort (1/3 of the participants). We were not able to identify a single transcript that was significantly differentially expressed between HBV vaccine responders and poor-responders of the EM131 training set (LIMMA: adjusted $p \leq 0.05$). We then tested if a combination of genes (multivariate model) could distinguish the two classes of responders. To that end, a naïve Bayes classifier based on the top 15 differentially expressed genes between responders and poor-responders to the HBV vaccines was built on the EM131 training set (**Fig. A1.4a**). The accuracy of the resulting 15-gene signature was assessed on the EM131 training set by 10-fold cross-validation and on the EM131 test set and found to be 62.6% and 62.2%, respectively (**Fig. A1.4b-4c**, **Supplementary Fig. A1.4**). A permutation test was performed to assess the probability of building a signature of the same size (15 genes) with better or equal accuracy to the 15-gene signature on the EM131 test set. That probability was 0.0381 suggesting that building a better predictive model randomly is improbable. Network inference revealed that several members of this 15-gene signature included markers of B cells (ex. *CD20*, *IGHG1*) as well as downstream targets of B-cell receptor signaling (ex. *BANK1*) and molecules known to have functional interactions with IgG (ex. *C1*, *FCGR3B*; **Fig. A1.4d**). Network inference revealed that both the 15-gene signature and the BioAge signature shared members of the B-cell signaling (*CD20*, *BANK1;* Fisher's exact test: p=7.99e-07) and Inflammation pathways (**Fig. A1.4e**). Both signatures showed upregulated expression of genes directly associated with inflammation. In the BioAge M16, *IRF7* (the master switch of type I interferon) and its downstream targets *IL15* and *OAS1* as well as *NFkB* and *C1* (hallmarks of inflammation) were associated with lower Ab titers. In the 15-gene signature, haptoglobin (*HP*), an acute phase protein and also a

target gene of type I interferon signaling, was associated with low Ab titers. Pathway analysis performed on a list of genes derived with a less stringent statistical cutoff of nominal $p \leq 0.05$ confirmed the positive association of B cell signaling and the response to HBV vaccine. Collectively, two independent bioinformatic approaches showed that B-cells and inflammation are important predictors of HBV vaccine response in the elderly.

FCM and red blood cell counts show that B-cells and inflammation are predictors of HBV vaccine response

A set of 96 FCM markers was tested for their association with the HBV vaccine response (**Supplementary Fig. A1.5a**). Logistic regression was used to determine the association of each FCM marker (frequencies of cells) to antibody titers to the HBV vaccine (**Table A1.1**). Two cell subsets defined by FCM, namely frequencies of switched IgG+ memory B cells (CD3-, CD19+, HLADR+, CD27+, IgG+) and of IgG+ memory B cells (CD3-, CD19+, HLADR+, CD27+, CD10-, CD20+, IgG+) were identified as significant predictors of the response to the HBV vaccine (**Fig. A1.5a** and **Table A1.1**). Frequencies of CD4 T-effector memory 2 (CD28-, marker of immunosenescence[24]) and CD40 on plasmacytoid dendritic cells (pDCs), another byproduct of chronic immune activation and inflammation[25], correlated negatively with HBV vaccine response (**Fig. A1.5a**). A multivariate regression model combining four FCM markers (% switched IgG+ memory B cells, % IgG+ memory B cells, % CD4 T-effector memory 2 and the median fluorescence intensity of CD40 in pDCs; **Table A1.1**), trained on the EM131 training set, was able to predict the response to HBV vaccine in the test set with an accuracy of 63.3% (**Fig. A1.5b**, permutation test: p=0.0468).

We also screened a panel of 47 serum cytokines and chemokines for their association with the response to HBV vaccination. Three inflammation-associated proteins IL-15, IL-1ra and TNFR2 were negatively associated with HBV vaccine response on the EM131 training set (**Supplementary Fig. A1.5b**). This model was not significantly associated to the HBV vaccine response in the EM131 test set (**Supplementary Fig. A1.5c,** permutation test: p=0.250).

Similarly, a set of hematologic markers was screened for their association with the HBV vaccine response. Interestingly, higher red blood cell (RBC) counts were associated with low Ab titers in elderly (**Table A1.1** and **Fig. A1.5c-5d**). Collectively our data show for the first time that baseline measures of immune function as defined by elevated frequencies of effector cells of the response

to vaccines (B cells and T helper cells) and inflammation (downstream of infections or other metabolic pathways) are associated to the response to vaccination.

Integrated model confirms the role of CD4+ T cells, B cells, RBC and inflammation in shaping the response to the HBV vaccine

We used least square regression models[26] to demonstrate that the gene-expression signatures underlying the poor response to HBV vaccine in elderly participants were correlated to levels of effector cytokines, frequencies of specific subsets of CD4 cells and innate immune cells (**Fig. A1.6**). We included the transcriptomic, FCM, cytokines/chemokines and RBC counts datasets in the integrative analysis. Results of this analysis showed that the two independently derived B cell gene-expression signatures (*i.e.* the BioAge module M1 and the 15-gene signature) were correlated with frequencies of memory B cells identified by FCM and with increased CD4 counts. Conversely the gene-expression signature of inflammation (ex. *C1QB*, *C3*, *IRF7*) correlated positively with FCM markers of activated innate immune cells and immunosenescent CD4 T cells (TEM2 in CD4s, CD40 in mDC1s and CD40 in pDCs, respectively) (**Fig. A1.6**). This integrated analysis showed the correlation between heightened levels of pro-inflammatory cytokines (ex. SCF, TNFR2 and VEGF) and increased frequencies of activated innate immune cells (TEM2 in CD4s, CD40 in mDC1s and CD40 in pDCs, respectively). Overall, the gene-expression signatures that highlight the contrasting roles of B cells and inflammation on the development of the response to HBV vaccine were confirmed by the analysis of cellular subsets and of cytokine profiles.

Discussion

In this study we provide an integrated model and propose mechanisms that could explain the response to vaccines prior to primary vaccination. This study cohort included a training set of 95 elderly participants (**Supplementary Table A1.1**) that was used to identify biomarkers of the response to vaccines; a distinct set of 49 participants sharing similar demographic and clinical characteristics served as an independent test set. This study design allowed the assessment of the accuracy and robustness of biomarkers that can predict the response to HBV vaccine. Two types of signaling pathways, B cell signaling and inflammation, were thus identified using two independent bioinformatic approaches as important regulators/predictors of the response to HBV vaccination. Both classifiers, the BioAge signature and the 15-gene signature, were confirmed in an independent test cohort and showed for the first time that it was possible to distinguish high responders from low responder to HBV vaccination with accuracies greater than 60%. While this is lower than accuracies previously reported post-vaccination[10], it is well within the range of those

reported for baseline predictors to cancer treatments (AUC = 67.7% [50.0%, 80.4%])[27]. The opposing impacts of B cell signaling and inflammation reflected in both classifiers were confirmed in independent set of participants and using three different experimental approaches: gene expression (**Fig. A1.3** and **Fig. A1.4**), flow cytometry (**Fig. A1.5a-b**) and cytokine profiling (**Supplementary Fig. A1.5b-c**). The fact that all three approaches identified these two major pathways as being associated to a successful or a failed response to vaccination highlight the robustness of these signatures.

Ab titers are the established correlates of vaccine efficacy for the HBV vaccine, implying that the presence of the appropriate subset of B cells prior to vaccination could favor a strong response to the vaccine. Our results confirm this observation. First, FCM results show that the percentage of memory B cell subsets (% switched IgG+ memory B cells and % IgG+ memory B cells) are significant univariate predictors of HBV vaccine response (**Fig. A1.5a**) although absolute counts of total B cells were not a significant predictor of the response to HBV vaccine. Similar results were reported by Tsang *et al.* where a specific B cell phenotype, namely % of CD38+CD27+ memory B cells, was the main feature of a model that predicted the Ab response to H1N1 vaccine in participants with undetectable Ab titers to H1N1 prior to vaccination[15]. Transcriptomic profiling shows that several B cell markers including components of the BCR complex (i.e. *IGH*, *CD79A*, *CD79B* and *CD19*) were also increased in responders and were positive predictors of the HBV vaccine response in our training and test cohorts. Transcriptional profiles show that a strong response to HBV vaccine additionally requires that B cells express transcription factors such as POU2AF1 and MZB1 that program B cells for Ab production and trigger the formation of germinal centers[22,23]. Transcriptional profiling further highlighted the role of *TNFRSF13B*, the receptor for BAFF and APRIL cytokines involved in the maturation of the humoral response, and a positive correlates of strong responses to vaccination[13]. Our results show that increased frequencies of CD4+ T cells, which can provide help to B cells in white blood cells, are associated with higher anti-HBsAg titers (**Fig. A1.5a**)[28].

Integration of FCM and cytokine profiles highlights the role of elements of the pro-inflammatory response which is negatively correlated to antibody production. Increased frequencies of activated innate immune cells (DR+CD40+ mDC and CD40+ pDCs) were associated with transcriptional signatures and cytokines that predicted the poor Ab response (**Fig. A1.5a**). These cells produce type I interferons in response to inflammatory signals and this is confirmed by the induction of the transcription factors IRF7 and IRF9 in poor-responders to the HBV vaccine (**Fig.**

A1.4). Activation of the interferon(s) transcriptional pathway was confirmed by the upregulation of several interferon-induced genes downstream of *IRF7* including *OASL*[29,30], and *LILRA5*[29] in participants who responded poorly to the HBV vaccine. The induction of the proinflammatory cytokine IL15, downstream of type I interferon[29,30], was confirmed at the gene and protein levels in poor-responders to the HBV vaccine (**Fig. A1.6**). The pro-inflammatory complement complex (e.g. *C3*, *C1QA*, *C1QB*) and VEGF, a positive regulator of angiogenesis that promotes the chronic inflammatory process[31], are also negatively associated to antibody production (**Fig. A1.6**). Increased levels of *MYD88* and *TRAF*, downstream of TLR and IL1 signaling, are observed in poor-responders to HBV vaccination. Both *MYD88* and *TRAF* are negative correlated to the frequency of CD4 T cells, a marker of good response to HBV vaccination (**Supplementary Fig. A1.6b**, Pearson correlation: r=-0.347 $p < 0.05$ and r=-0.353 $p < 0.05$, respectively). Moreover, *C3* and *WARS*, two target genes of type I interferons, were negatively associated to the frequencies of the two memory B cell subsets. In contrast, anti-inflammatory proteins coding genes like *CD200*[32] and *BATF3*[33] were repressed in HBV vaccine hyporesponders (**Fig. A1.6**). Collectively our results highlight the importance of type I interferon and other pro-inflammatory pathways in the development of a poor response to HBV vaccine.

Poor-responders to HBV vaccine showed higher numbers of RBCs than responders. Interestingly genes and pathways that correlate positively with HBV Ab titers including genes within module M1 were negatively correlated with RBC counts. Conversely genes included in module M16 (**Fig. A1.6**) and proinflammatory genes included in pathways enriched in poor responders (**Fig. A1.5e**) were positive correlates of RBC counts. The HIF-1 α pathway. which controls erythropoietin expression, was significantly upregulated in poor-responders (Fisher's exact test: p=0.0360); target genes of HIF-1α were induced in poor-responders and their expression levels correlated with higher RBC counts (Pearson t-test: p=0.0462, **Fig. A1.5e**). RBC counts were positively correlated to the *HMOX1* (heme oxidase catalyzing the degradation of heme), *HP* (scavenger of free Hb) and *CD163* expressions (scavenger receptor of the Hb-HP complex), all of which are interferon regulated genes and were induced in HBV vaccine poor-responders. The heme-induced response pathway was significantly enriched among genes differentially expressed between HBV vaccine responders and poor-responders (Fisher's exact test: p=0.0461). *EIF2AK1* (Heme-Regulated Inhibitor) and several of its downstream targets (ex. *EIF2, EIF3 and EIF4*) were negatively associated with Ab titers (**Supplementary Fig. A1.6a**). These results highlight a potential role for the heme-induced response and for hypoxia in the poor response to the HBV vaccine.

These results suggest a mechanism whereby this aging-associated inflammation leads to induction of the HIF-1α pathway and other pro-inflammatory effector molecules. This may upregulate erythropoietin, a known target of HIF-1α, leading to increased RBC counts, release of cell free hemoglobin as suggested from the heightened levels of haptoglobin mRNA[34], and the downstream upregulation of HRI (**Supplementary Fig. A1.6a**), a trigger of type I interferon production by pDC and other innate immune cells (**Supplementary Fig. A1.7**). This will further enhance inflammatory pathways that downregulate B-cell activation and hence hyporesponse to *de novo* vaccination with a vaccine such as HBV (**Supplementary Fig. A1.7**). Alternatively, these perturbations in RBC homeostasis and B-cell response could be mark of a global hyperimmune inflammation observed in elderly ("inflammaging")[35]. The interplay between inflammation and B-cell signaling is supported by all the experimental approaches in this study (**Supplementary Fig. A1.7**).

Several strategies may help overcome the mechanisms impeding optimal response to vaccines. Subjects having a pro-inflammatory signature prior to vaccination could benefit from vaccine regimens conferring improved immunogenicity. For example, inflammatory bowel disease patients that did not respond to a first HBV vaccine course were more likely to mount an antibody response upon receiving a second complete immunization schedule[36]. Strategies used to overcome desensitization of the innate immune response including the use of more potent adjuvants[37], or anti-inflammatory drugs could also improve immune responses to HBV vaccination in the elderly. For example, the anti-inflammatory drug rapamycin improved Ab responses to influenza vaccination[38] when used at low dose prior to vaccination. Our results show that different "inflammatory" pathways are at cross-purposes; these differential effects of inflammation will need to be further clarified to improve the response to vaccines. Accordingly, the predictive gene-expression signatures described herein might allow tailoring of vaccine regimens to older persons predisposed to VHR.

Methods

Study Design and Conduct

This was an observational open-label study (**Fig. A1.1a**), performed between July 2010 and November 2011. This study was registered on ClinicalTrials.gov (NCT01119703), was overseen by Institutional Review Board Services (Ontario Canada), and was conducted according to cGCP and applicable laws.

One hundred and seventy four (174) generally healthy, Hepatitis B virus (HBV) naïve, adult residents of Québec were divided into two groups to participate in this study. One group was aged 65 or over (n=144); participants in the other group were aged between 25 and 40 years (n=30). A demographic summary appears in **Supplementary Table A1.1**. All participants provided written informed consent.

All participants were vaccinated with two doses of Twinrix® (HBsAg and Hepatitis A virus - Glaxo Smith-Kline), and single doses of generic Tetanus-diphtheria booster (tetanus and diphtheria - Sanofi-Pasteur), and Dukoral (recombinant cholera toxin B subunit and whole killed Vibrios - Sanofi Pasteur) according to the respective product labels. Intramuscular vaccines were administered to opposite arms and all doses of Twinrix were administered in the same arm. HBV titers were collected after the second injection based on published results suggesting that all three injections would result in very high rates of seroresponse[39,40]. As we desired a wide spectrum of responses against which to regress biomarker results, titer responses were only recorded after two injections – leading to an unexpectedly large number of poor-responders. The observed range of titer responses sufficed for the required discovery effort.

Antibody titers

Anti-cholera toxin B subunit (CTB) IgG were quantified in serum samples according to a GM1-ELISA method adapted from Svennerholm *et al.*[41]. GM1 is a membrane ganglioside which is found in cell membranes of the gastro-intestinal tract and that acts as a receptor for CTB. The GM1 ganglioside was used in the ELISA assay to increase binding of CTB to the plate wells. Briefly, high protein binding microplates were first coated overnight with GM1 (1.5 µM in PBS). Plates were washed with PBS and were blocked with a 0.5% BSA solution for 1 hour at room temperature. The plates were then coated with CTB (1 µg/mL in 0.5% BSA) for 1 hour at room temperature and washed. Serum samples were serially diluted (six serial dilutions) with a 0.5% BSA + Tween solution. Diluted sera were added in duplicates to the blocked GM1/CTB-coated plate and incubated for 2 hours at room temperature. A positive control serum was also included on each assay plate and was assayed at six serial dilutions, in duplicate. After successive washes to remove unbound antibodies, a goat anti-human IgG antibody conjugated to horseradish peroxidase (HRP) was added and incubated for 1 hour at room temperature. Unbound conjugate was removed by washing, and a tetramethylbenzidine (TMB) substrate solution was added. Following a 10 minutes incubation period the stopping solution was added and optical density

was measured at 450/620 nm using a spectrophotometer. Titration curves were drawn by plotting averaged optical density (OD) values versus the dilution factor and fitted using a 5-parameter curve fit. The antibody titer was defined as the midpoint (or IC50) of the titration curve. Titers were normalized against the positive control serum titer.

Anti-hepatitis B surface antigen (anti-HBsAg) total antibodies were quantified in serum samples using the MONOLISA Anti-HBs PLUS kit (Bio-Rad catalogue number 72566). Briefly, serum samples were diluted at the recommended dilution factor in the dilution buffer provided by the kit. Diluted serum samples as well as calibration standards and QC high and QC low controls were added to HBsAg-coated plates, and incubated for 60 min at 37°C. After a washing step, the conjugate was added to the plates and incubated for another 60 min at 37C. The plates were washed and the TMB substrate solution was added. Following a 30 minutes incubation period the stopping solution was added and optical density was measured at 450/620 nm using a spectrophotometer. The determination of anti-HBs levels has been standardized by use of the WHO Anti-HBs reference preparation expressed in milli-International Units per milliliter (mIU/mL). Though HBV titers were measured one month after the second dose, all patients were offered the third dose to conform to the approved regimen.

Anti-Diphtheria IgG were quantified in serum samples using a commercially available kit (Sekisui Virotech, catalogue number EC129.00). Briefly, serum samples were diluted at the recommended dilution factor in the dilution buffer provided by the kit. Diluted serum samples as well as calibration standards and IgG high and IgG low controls were added to the antigen-coated microplates and incubated for 30 min at 37°C. After a washing step, the anti-IgG HRP conjugate was added to the plates and incubated for another 30 min at 37°C. The plates were washed and the TMB substrate solution was added. Following a 30 minutes incubation period the stopping solution was added and optical density was measured at 450/620 nm using a spectrophotometer. The Anti-Diphtheria toxin IgG concentrations were expressed in International Units (IU/mL) following the WHO Standard. The standard curve of the Diphtheria ELISA has been verified using the Diphtheria Antitoxin Human Serum (00/496) of the Institute for Biological Standards and Control, WHO International Laboratory for Biological Standards in Great Britain.

Anti-Tetanus IgG were quantified using a commercially available kit (Sekisui Virotech, catalogue number EC124.00). Briefly, serum samples were diluted at the recommended dilution factor in the dilution buffer provided by the kit. Diluted serum samples as well as calibration standards and

IgG high and IgG low controls were added to the antigen-coated microplates and incubated for 30 min at 37°C. After a washing step, the anti-IgG HRP conjugate was added to the plates and incubated for another 30 min at 37°C. The plates were washed and the TMB substrate solution was added. Following a 30 minutes incubation period the stopping solution was added and optical density was measured at 450/620 nm using a spectrophotometer. The Anti-Tetanus toxin IgG concentrations were expressed in International Units (IU/mL) following the WHO Standards. The standard curve of the Tetanus ELISA has been verified using the international standard of the WHO for human Tetanus Immunoglobulin (TE-3).

Development of the BioAge signature

The Gene Expression Omnibus and Arrays Express databases were searched for the largest publically available microarray dataset studying blood samples and including age as clinical endpoint. The San Antonio Family Heart Study (SAFHS) comprising of 1240 samples hybridized on Illumina Sentrix Human Whole Genome 6 BeadChips[19] was used to train the BioAge signature. Raw files containing background-subtracted intensities, chip annotation and sample annotation were downloaded from ArrayExpress database with accession number E-TABM-305. Analysis of the raw data was conducted using R/Bioconductor software[42]. Missing values (corresponding to less than 1% of all the intensity values) were imputed using nearest neighbor averaging method[43]. Quantile normalization was performed followed by log2 transformation for variance stabilization. Background value of 0.1 was used for surrogate replacement prior to log2 transformation in order to prevent infinite intensities values.

The LIMMA package[44] was used to fit a linear regression model (1) to each probe where the log2-expression of transcript (*x*) was used as independent variable and chronological age as dependent variable (*age*). Age was kept as a continuous variable ranging from 15 years up to 94 years.

$$(1)\ age \sim ax + b$$

For each regression model, a (moderated) *t*-test was performed in order to test the significance of the association of expression of the transcript and the chronological age. The *p*-values were adjusted for multiple comparisons using the Benjamini and Hochberg method[45]. A *p*-value, corrected for false-positive rate, less than or equal to 0.05 was used as cutoff to identify transcripts significantly associated with chronological age. 2,285 transcripts passed this specified cutoff. Once the genes correlated with age were identified (LIMMA: adjusted $p \leq 0.05$), we evaluate of in how many sets of correlates genes (*i.e.* modules) they could be separated. To that end, the log2-

expression was transformed to *z*-score (i.e. for each transcript, its average expression was subtracted and divided by its standard deviation across samples). The optimal number of modules (*k*) was estimated via the gap statistic[46] (**Supplementary Fig. A1.2a**).

We then defined the BioAge signature as two sets of modules (M): *up arm* with modules overexpressed in elderly (M12 to M20) and *down arm* with modules overexpressed in the non-elderly (M1 to M11). The BioAge score (2) is defined as the difference between the average gene-expression (μ) of modules in the *up arm* minus the average gene-expression of modules in the *down arm*.

$$(2)\ BioAge = \mu(M12, \ldots, M20) - \mu(M1, \ldots, M11)$$

Establishing gene-expression signature predictive of HBV vaccine response

RNA was isolated from PAXgene blood samples according to the manufacturer's instructions. Isolated total RNA samples were assayed for quality metrics (Agilent Bioanalyzer) prior to amplification. Three hundred and thirty five (335) samples passing quality control were then amplified using the NuGEN amplification protocol and hybridized to Affymetrix HuRSTA-2a520709 chips (Affymetrix, Santa Clara, CA). The chips contain 60,607 probesets representing 22,580 unique genes (GEO: GPL15048).

Analysis of the CEL files was conducted using R/Bioconductor software packages[42]. The expression data was normalized using the RMA method[47]. Five technical duplicates were included in the microarray experiments; each pair of technical duplicates was averaged after the microarrays normalization.

The elderly dataset was separated randomly into a training set composed of 2/3 (91 samples) of the samples and a test set composed of 1/3 (47 samples) of the samples (**Supplementary Table A1.1**). Samples from subjects under 50 (n=55) years were excluded from the dataset. The randomization of the remaining samples (n=280) was blocked by patients' age and gender (i.e. controlling that the proportion of male/female and the age range are similar between training set and test set). In order to identify pre-vaccination biomarkers of vaccine response, the dataset was restricted to pre-vaccination samples (138/280 samples). The resulting training set and test set contained 91 and 47 samples, respectively (**Supplementary Table A1.1**). Mathematical modeling of the anti-HBsAg titers performed using kernel density estimation as implemented in the function density of R package stats revealed a bimodal distribution of the titers. Participants who had anti-HBsAg titers one month after the 2nd immunization below detection threshold (< 5 mIU/mL) were considered poor-responders to the HBV vaccine (n=52) while participants with anti-HBsAg titers

equal or above 5 mIU/mL were considered responders to the HBV vaccine (n=38). The R package LIMMA was used to fit a linear model to each probe and to perform (moderated) *t*-tests on the comparison HBV vaccine responders versus poor-responders. The *p*-values were adjusted for multiple comparisons using the Benjamini and Hochberg method[45]. A naïve Bayes classifier was build on the training set based on the expression of top differentially expressed transcripts (ordered by LIMMA moderated *t*-test *p*-values). The number of features in the naive Bayes classifier was established by performing 10-fold cross-validation on the EM131 training set. The resulting classifier was based on the expression of the top 15 differentially expressed transcripts between HBV vaccine responders and poor-responders. ROC analysis and permutation test were performed on EM131 test set to evaluate the performance of the classifier. A 10000-fold permutation test was performed to assess the probability of obtaining an equal or superior classifier performance.

The microarray data presented in this article have been submitted to the National Center for Biotechnology Information Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession number GEO: GSE65834.

Pathway enrichment analysis

Fisher's exact test was used to assess significance of the intersect between the modules of genes composing the BioAge signature and *a priori* built genesets (IPA canonical pathways). Significantly enriched pathways were organized in sets of related pathways using the enrichment map strategy[48]. Overlap between significant genesets was computed according to the Jaccard Index. A Jaccard index cutoff of 0.25 was used.

Statistical analysis

Clinical characteristics including age, gender, height, weight, body mass index and pre-vaccination cytomegalovirus antibody titers were tested as predictors of HBV vaccine response using univariate and multivariate logistic regression analysis. *z*-tests were used to test whether the regression coefficient of each predictor was different from zero. Statistical tests were two-sided, adjusted for multiple testing using the Benjamini-Hochberg correction and considered significant when the adjusted *p*-value ≤ 0.05.

Polychromatic Flow Cytometry

Phenotypic analysis of T cell and innate cell subsets was performed by staining whole blood

collected in EDTA Vacutainer® tubes. Blood for B cell analysis was collected in Na-Heparin Vacutainer® tubes and further processed by density gradient centrifugation through Ficoll to obtain PBMC prior to performing the following steps. Processing and staining steps for cytometry were performed within 8 hours of blood draw. Briefly, 50uL of whole blood were added to TruCount tubes (Becton-Dickinson) for the phenotypic analysis and absolute enumeration of cellular subsets per microliter of blood. Following 30 minutes incubation at room temperature with surface staining antibodies, BD FACSLyze was added for 10min to lyse red blood cells. Cells were then washed and transferred to a 96 well plate for cytometric data acquisition. A similar procedure was implemented for the phenotypic analysis of innate cell subsets. However, given the lower frequencies of target populations, 100uL of whole blood were processed and stained prior to analyzing by flow cytometry (FCM). An example of gating strategies for B cell populations is presented in **Supplementary Fig. A1.5a**.

Three FCM panels were respectively used to determine frequencies of T cells, B cells, and innate immune cells obtained from the 174 pre-vaccination (Visit 2) samples of the study cohort. Logistic regression was used to assess the association between individual FCM markers and the response to HBV vaccine on the EM131 training set. A multivariate model combining the best markers (z-test: $p \leq 0.1$) was built using the forward selection approach. Briefly, the forward selection process starts with no variable in the model and tests the addition of each variable using the Akaike information criterion. Only those variables that provide a superior improvement of the model compared to the remaining variable are selected in the final model. The best model obtained included four FCM markers. ROC analysis was performed on EM131 test set to evaluate the performance of the multivariate model.

Cytokine Measurements

A panel of 45 cytokines/chemokines was measured on the 174 pre-vaccination (V2) samples of the EM131 cohort. Samples were assayed to determine the concentration of inflammatory proteins according to multianalyte profiling (MAP) protocols established by Rules-Based Medicine (RBM) (Austin, Tex). Serum samples were shipped to RBM for
analysis using the multiplex ELISA protocol Human Inflammation MAP. Logistic regression was used to assess the association between individual cytokine markers and the response to the HBV vaccine and a multivariate model combining the best markers was built on the EM131 training set using a forward selection approach, as described above. The best model obtained included three

cytokines. ROC analysis was performed on EM131 test set to evaluate the performance of the multivariate model.

Hematologic Measurements

A panel of 3 standard hematologic markers (complete blood count) was measured on the 174 pre-vaccination samples taken 28 days prior to vaccination (V1) of the EM131 cohort. Logistic regression was used to assess the association between individual hematologic markers and the response to HBV vaccine and a multivariate model combining the best markers was built on the EM131 training set and tested on the EM131 test set.

Integrative analysis of all multi-dimensional data sets

For every pair of omics data, the one containing the most features was used as the independent variable and the other omics was used as dependent variable. A partial least-square regression as implemented by the function *spls* of the R package mixOmics was used to project the omics data in the same space[26]. Once the two omics are projected in the same space, a Pearson correlation between features of the first and second omics are calculated. All absolute Pearson correlations $|r| \geq 0.188$ corresponding to a *t*-test *p*-value below 0.05 were used to generate a correlation network between omics.

Code availability

All the source codes are available at https://sites.google.com/a/case.edu/fouslim/publication/em131.

References

1 Gavazzi, G. & Krause, K. H. Ageing and infection. *Lancet Infect Dis* **2**, 659-666 (2002).

2 United Nations, D. o. E. a. S. A., Population Division World population Ageing 2013. *ST/ESA/SER.A/348* (2013).

3 Clements, M. L. *et al.* Effect of age on the immunogenicity of yeast recombinant hepatitis B vaccines containing surface antigen (S) or PreS2 + S antigens. *J Infect Dis* **170**, 510-516 (1994).

4 Hohler, T. *et al.* Differential genetic determination of immune responsiveness to hepatitis B surface antigen and to hepatitis A virus: a vaccination study in twins. *Lancet* **360**, 991-995 (2002).

5 Weihrauch, M. R. *et al.* T cell responses to hepatitis B surface antigen are detectable in non-vaccinated individuals. *World J Gastroenterol* **14**, 2529-2533 (2008).

6 Goronzy, J. J. & Weyand, C. M. Understanding immunosenescence to improve responses to vaccines. *Nat Immunol* **14**, 428-436 (2013).

7 Seok, J. *et al.* Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A* **110**, 3507-3512 (2013).

8 Pulendran, B., Li, S. & Nakaya, H. I. Systems vaccinology. *Immunity* **33**, 516-529 (2010).

9 Gaucher, D. *et al.* Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. *J Exp Med* **205**, 3119-3131 (2008).

10 Querec, T. D. *et al.* Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat Immunol* **10**, 116-125 (2009).

11 Obermoser, G. *et al.* Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity* **38**, 831-844 (2013).

12 Tan, Y. *et al.* Gene signatures related to B-cell proliferation predict influenza vaccine-induced antibody response. *Eur J Immunol* **44**, 285-295 (2014).

13 Li, S. *et al.* Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat Immunol* **15**, 195-204 (2014).

14 Furman, D. *et al.* Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination. *Proc Natl Acad Sci U S A* **111**, 869-874 (2014).

15 Tsang, J. S. *et al.* Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* **157**, 499-513 (2014).

16 Atkinson, W. L. *et al.* General recommendations on immunization. Recommendations of the Advisory Committee on Immunization Practices (ACIP) and the American Academy of Family Physicians (AAFP). *MMWR Recomm Rep* **51**, 1-35 (2002).

17 Scerpella, E. G. *et al.* Serum and Intestinal Antitoxin Antibody Responses after Immunization with the Whole-Cell/Recombinant B Subunit (WC/rBS) Oral Cholera Vaccine in North American and Mexican Volunteers. *J Travel Med* **3**, 143-147 (1996).

18 Van Buren, R. C. & Schaffner, W. Hepatitis B Virus: A Comprehensive Strategy for Eliminating Transmission in the United States Through Universal Childhood Vaccination: Recommendations of the Immunization Practices Advisory Committee (ACIP). *MMWR Recomm Rep* (1991).

19 Goring, H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* **39**, 1208-1216 (2007).

20 Ioannidis, V., Beermann, F., Clevers, H. & Held, W. The beta-catenin--TCF-1 pathway ensures CD4(+)CD8(+) thymocyte survival. *Nat Immunol* **2**, 691-697 (2001).

21 Panda, A. *et al.* Human innate immunosenescence: causes and consequences for immunity in old age. *Trends Immunol* **30**, 325-333 (2009).

22 Teitell, M. A. OCA-B regulation of B-cell development and function. *Trends Immunol* **24**, 546-553 (2003).

23 Anelli, T. & van Anken, E. Missing links in antibody assembly control. *Int J Cell Biol* **2013**, 606703 (2013).

24 Parish, S. T., Wu, J. E. & Effros, R. B. Sustained CD28 expression delays multiple features of replicative senescence in human CD8 T lymphocytes. *J Clin Immunol* **30**, 798-805 (2010).

25 Takagi, H. *et al.* Plasmacytoid dendritic cells are crucial for the initiation of inflammation and T cell immunity in vivo. *Immunity* **35**, 958-971 (2011).

26 Liquet, B., Le Cao, K. A., Hocini, H. & Thiebaut, R. A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC Bioinformatics* **13**, 325 (2012).

27 Shi, L. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol* **28**, 827-838 (2010).

28 Vieira, P. & Rajewsky, K. Persistence of memory B cells in mice deprived of T cell help. *Int Immunol* **2**, 487-494 (1990).

29 Goubau, D. *et al.* Transcriptional re-programming of primary macrophages reveals distinct apoptotic and anti-tumoral functions of IRF-3 and IRF-7. *Eur J Immunol* **39**, 527-540 (2009).

30 Lazear, H. M. *et al.* IRF-3, IRF-5, and IRF-7 coordinately regulate the type I IFN response in myeloid dendritic cells downstream of MAVS signaling. *PLoS Pathog* **9**, e1003118 (2013).

31 Monaco, C., Andreakos, E., Kiriakidis, S., Feldmann, M. & Paleolog, E. T-cell-mediated signalling in immune, inflammatory and angiogenic processes: the cascade of events leading to inflammatory diseases. *Curr Drug Targets Inflamm Allergy* **3**, 35-42 (2004).

32 Gorczynski, R. M. CD200: CD200R-mediated regulation of immunity. *ISRN Immunology* **2012** (2012).

33 Iacobelli, M., Wachsman, W. & McGuire, K. L. Repression of IL-2 promoter activity by the novel basic leucine zipper p21SNFT protein. *J Immunol* **165**, 860-868 (2000).

34 Glass, G. A., Gershon, D. & Gershon, H. Some characteristics of the human erythrocyte as a function of donor and cell age. *Exp Hematol* **13**, 1122-1126 (1985).

35 Shaw, A. C., Goldstein, D. R. & Montgomery, R. R. Age-dependent dysregulation of innate immunity. *Nat Rev Immunol* **13**, 875-887 (2013).

36 Yolima, C. G. *et al.* Immunogenicity of hepatitis B vaccine in patients with inflammatory bowel disease and the benefits of revaccination. *J Gastroenterol Hepatol* (2014).

37 Janssen, R. S. *et al.* Immunogenicity and safety of an investigational hepatitis B vaccine with a toll-like receptor 9 agonist adjuvant (HBsAg-1018) compared with a licensed hepatitis B vaccine in patients with chronic kidney disease. *Vaccine* **31**, 5306-5313 (2013).

38 Mannick, J. B. *et al.* mTOR inhibition improves immune function in the elderly. *Sci Transl Med* **6**, 268ra179 (2014).

39 Treadwell, T. L. *et al.* Immunogenicity of two recombinant hepatitis B vaccines in older individuals. *Am J Med* **95**, 584-588 (1993).

40 Thoelen, S. *et al.* The first combined vaccine against hepatitis A and B: an overview. *Vaccine* **17**, 1657-1662 (1999).

41 Svennerholm, A. M., Holmgren, J., Black, R., Levine, M. & Merson, M. Serologic differentiation between antitoxin responses to infection with Vibrio cholerae and enterotoxin-producing Escherichia coli. *J Infect Dis* **147**, 514-522 (1983).

42 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).

43 Hastie, T. *et al.* Imputing missing data for gene expression arrays. (1999).

44 Smyth, G. K. in *Bioinformatics and computational biology solutions using R and Bioconductor* 397-420 (Springer, 2005).

45 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300 (1995).

46 Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**, 411-423 (2001).

47 Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).

48 Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).

Figures

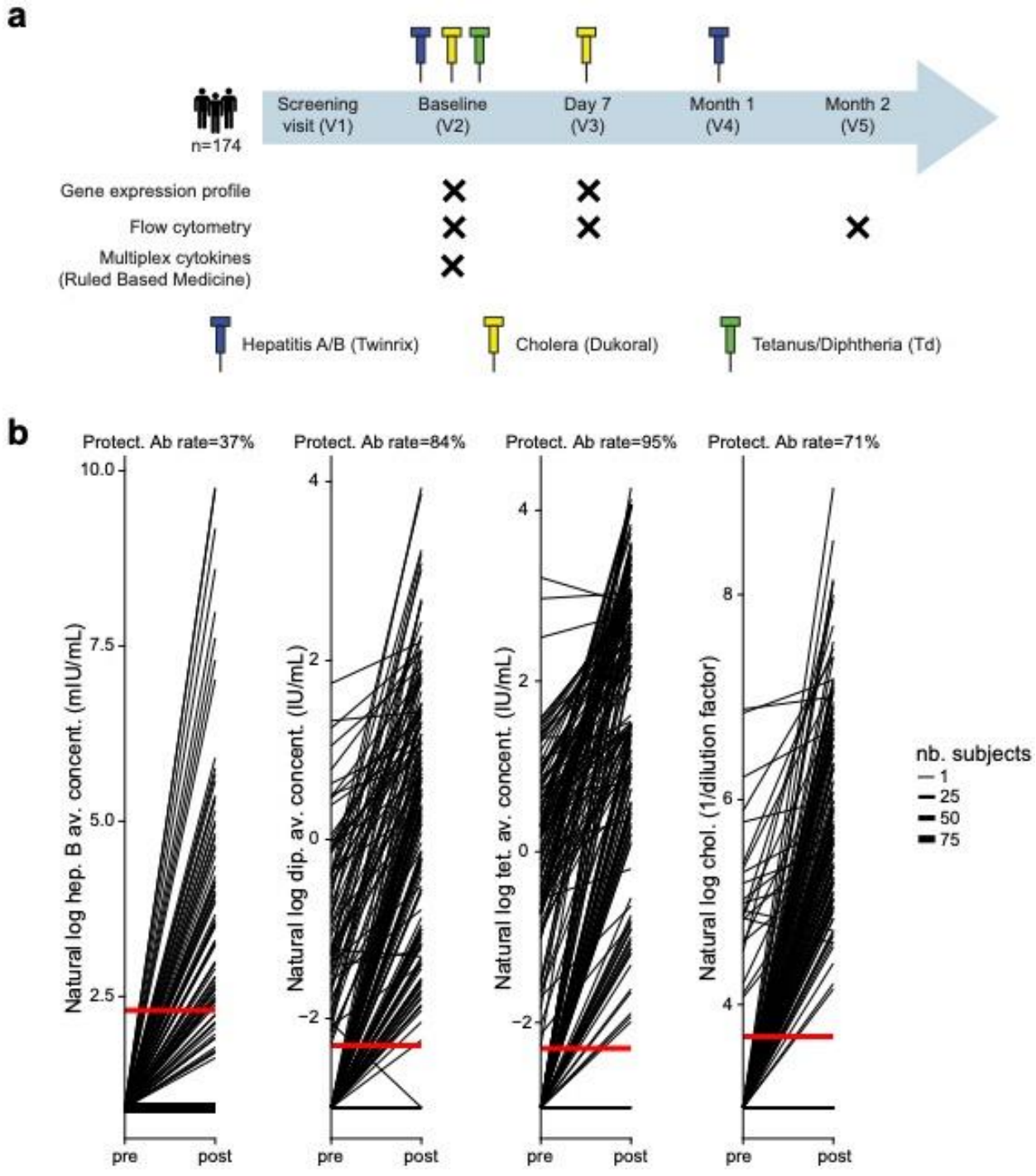


Figure A1.1. Study design and antibody titers for the three vaccines used in the EM131 study

(**a**) Schematic representation of the study design indicating blood collections and assays performed. All analyzed participants were HBV-naïve at time of recruitment and received vaccines for hepatitis A/B, cholera and tetanus/diphtheria. (**b**) Response plots showing antibody titers for HBsAg (hep. B), diphtheria toxin (dip.), tetanus toxin (tet.) and cholera toxin (chol.) for the 174 study participants as function of the vaccination status (x-axis: before and after vaccination). Red horizontal lines indicate standard titer thresholds and the percentage of participants above the protective thresholds are indicated above each plot (Protect. Ab.).

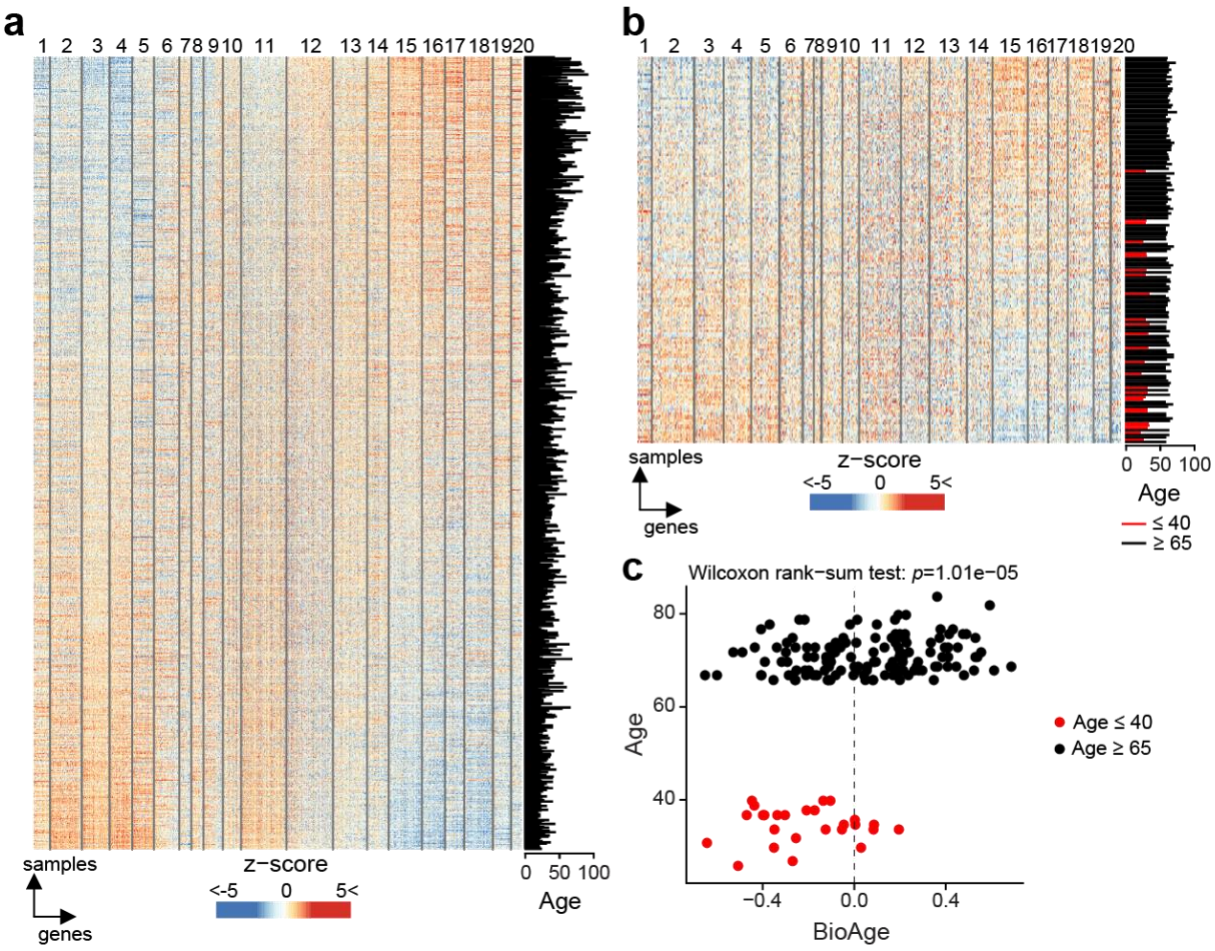


Figure A1.2. Development of the BioAge signature and application to the EM131 cohort (**a**) Expression of the 2,285 age-related transcripts derived from SAFHS data set (n=1240)[19]; all transcripts correlating with chronological age (moderated t-test: adjusted $p \leq 0.05$) were clustered in 20 modules using *k*-means clustering. In the heatmap, transcripts (columns) were ordered by their membership to the 20 modules; the 1,240 samples/participants (row) were ordered by their BioAge score (signed average expression of the age-related transcripts). Transcript expression was transformed to z-score and is depicted in blue to white to red color scale. The chronological

age is given in the barplot at the right. (**b**) Expression of the 2,285 transcripts of the BioAge signature in the EM131 data set (n=174). The chronological age of the 174 participants is given in the barplot at the right. (**c**) Scatter plot showing the chronological age as a function of the BioAge. The vertical line indicates a BioAge score of 0. We observe that young participants (red circles) have significantly younger BioAge; and among elderly (black circles) about half of participants have a young BioAge (< 0) while the other half have an old BioAge (≥ 0).

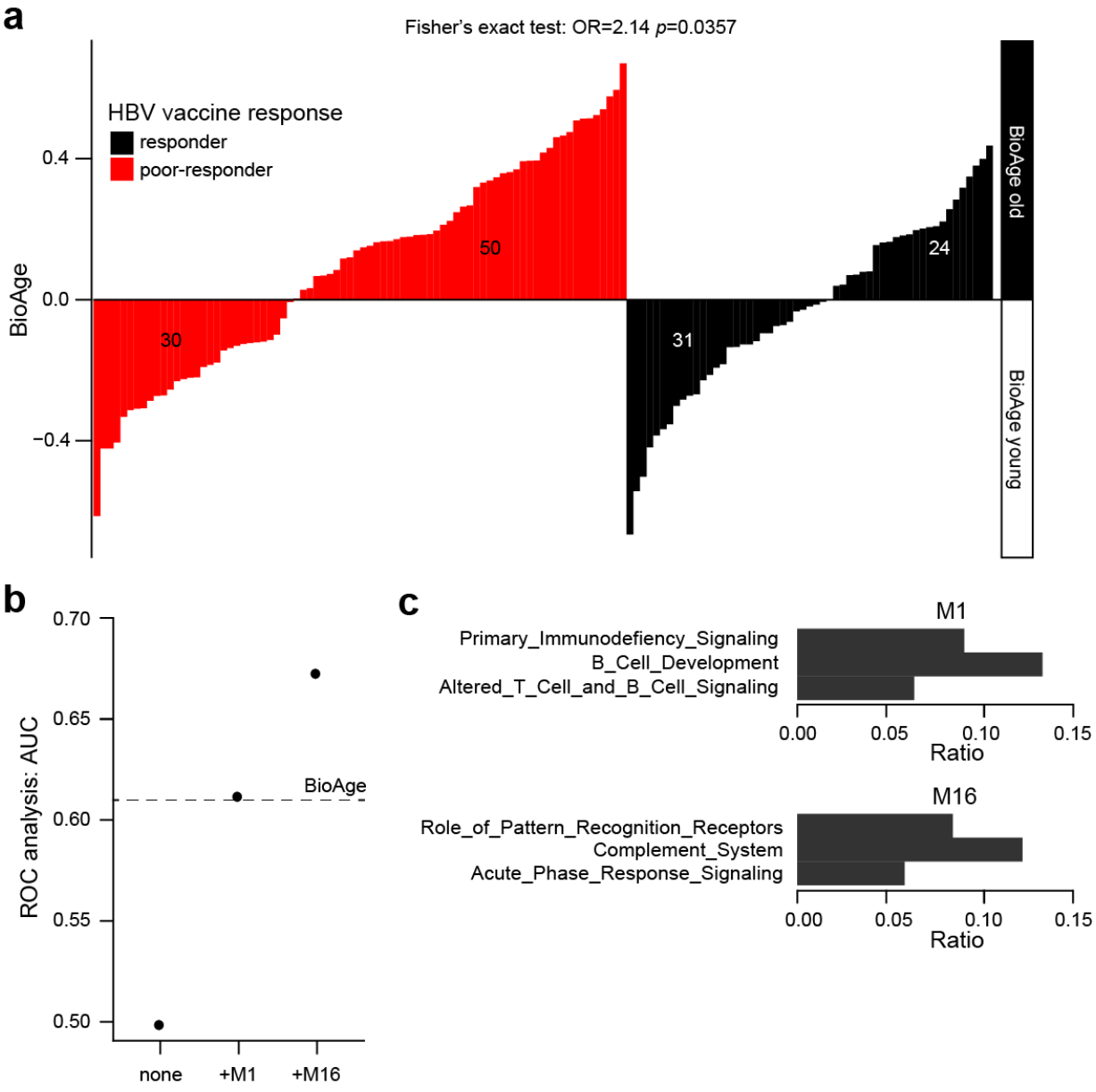


Figure A1.3. BioAge predicts HBV vaccine response

(**a**) BioAge classification of the 135 elderly patients based on gene expression prior to vaccination is significantly associated with the response to HBV vaccination. Each bar of the barplot represents one of the elderly participants in the EM131 cohort. The height of the bar indicates the BioAge score of that donor. Bars were ordered by increasing level of the BioAge, separately for the HBV poor-responders (in red) and HBV responders (in black). Fisher's exact test *p*-values are given on the plot. (**b**) Forward selection among the modules composing the BioAge signature resulted in the selection of modules M1 and M16 as optimal signatures predicting HBV vaccine response in the EM131 elderly cohort (age ≥ 65). (**c**) Pathway enrichment analysis on the genes included in modules M1 and M16 using the IPA canonical pathway database. Fisher exact test was performed to assess statistical enrichment and genesets with FDR-corrected *p*-value ≤ 0.05 are presented.

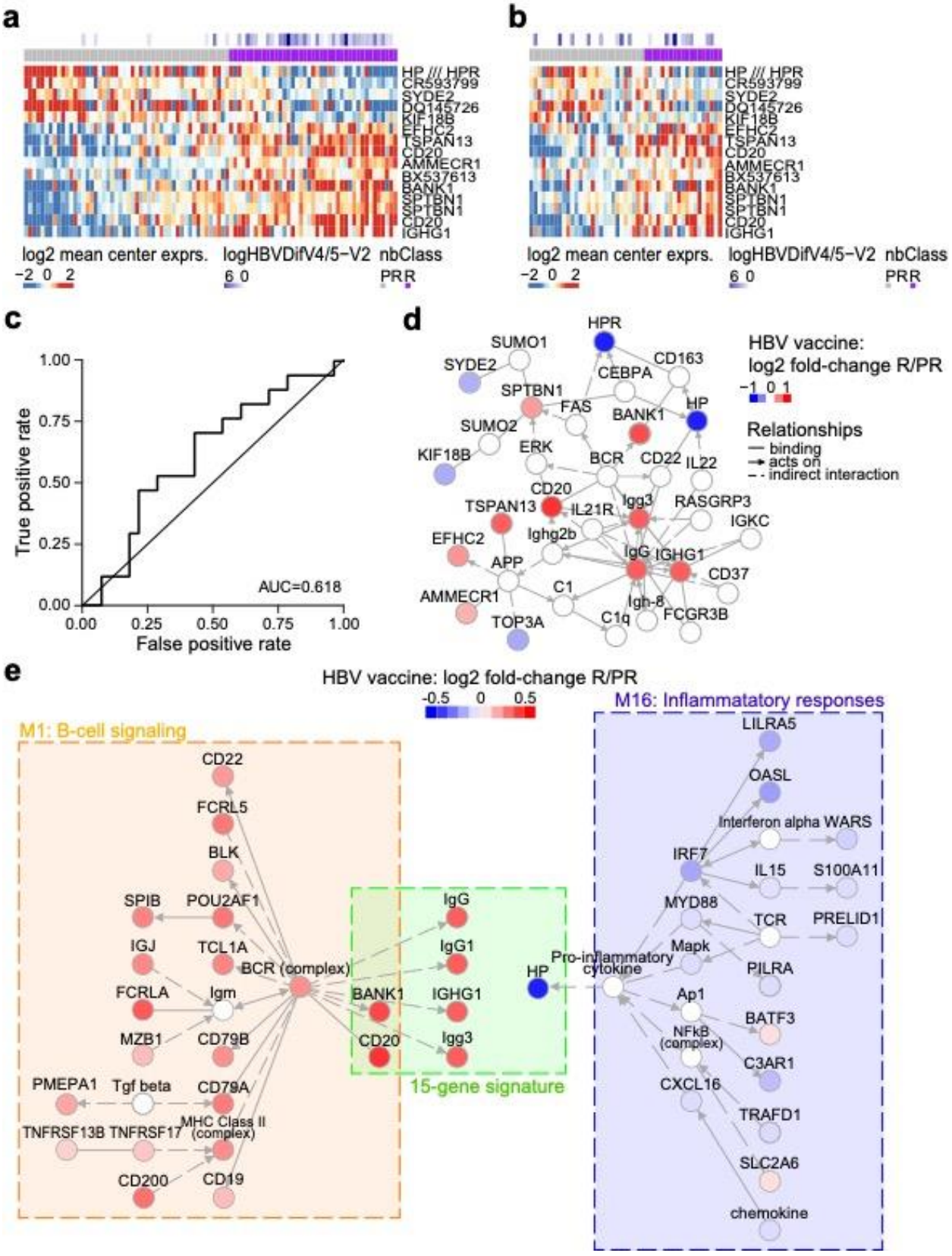


Figure A1.4. Identification of gene-expression signature predicting the HBV vaccine response

(**a**) Expression of 15 genes identified as predictors of the response to the HBV vaccine in the EM131 training set (n=95). The mean-centered gene-expression is represented using a blue to white to red color scale. Rows and columns correspond to the genes and the profiled samples, respectively. Samples were ordered by increasing levels of their predicted probability of responding to the vaccine *(i.e.* posterior probability). Antibody response to the HBV vaccine (log(HepBDifV4/5-V2)) and the response group predicted by the 15-gene signature (nbClass) are presented in colored squares above each sample. (**b**) Expression of the 15-gene signature on the EM131 test set (n=49). (**c**) ROC curve and area under the curve (AUC) for the prediction of the HBV vaccine response using the 15-gene signature on the EM131 test set (n=49). (**d**) Network inference based on the 15 markers identified as predictors of the response to the HBV vaccine. Red and blue nodes represent genes induced or repressed in HBV vaccine responders (R) compared to poor-responders (PR), respectively. (**e**) Networks were inferred for the BioAge module 1 combined to the 15-gene signature and the BioAge module 16 combined to the 15-gene signature, respectively. Nodes included in the BioAge or the 15-gene signature are colored by their fold-change between R vs. PR to the HBV vaccine in the EM131 training set.

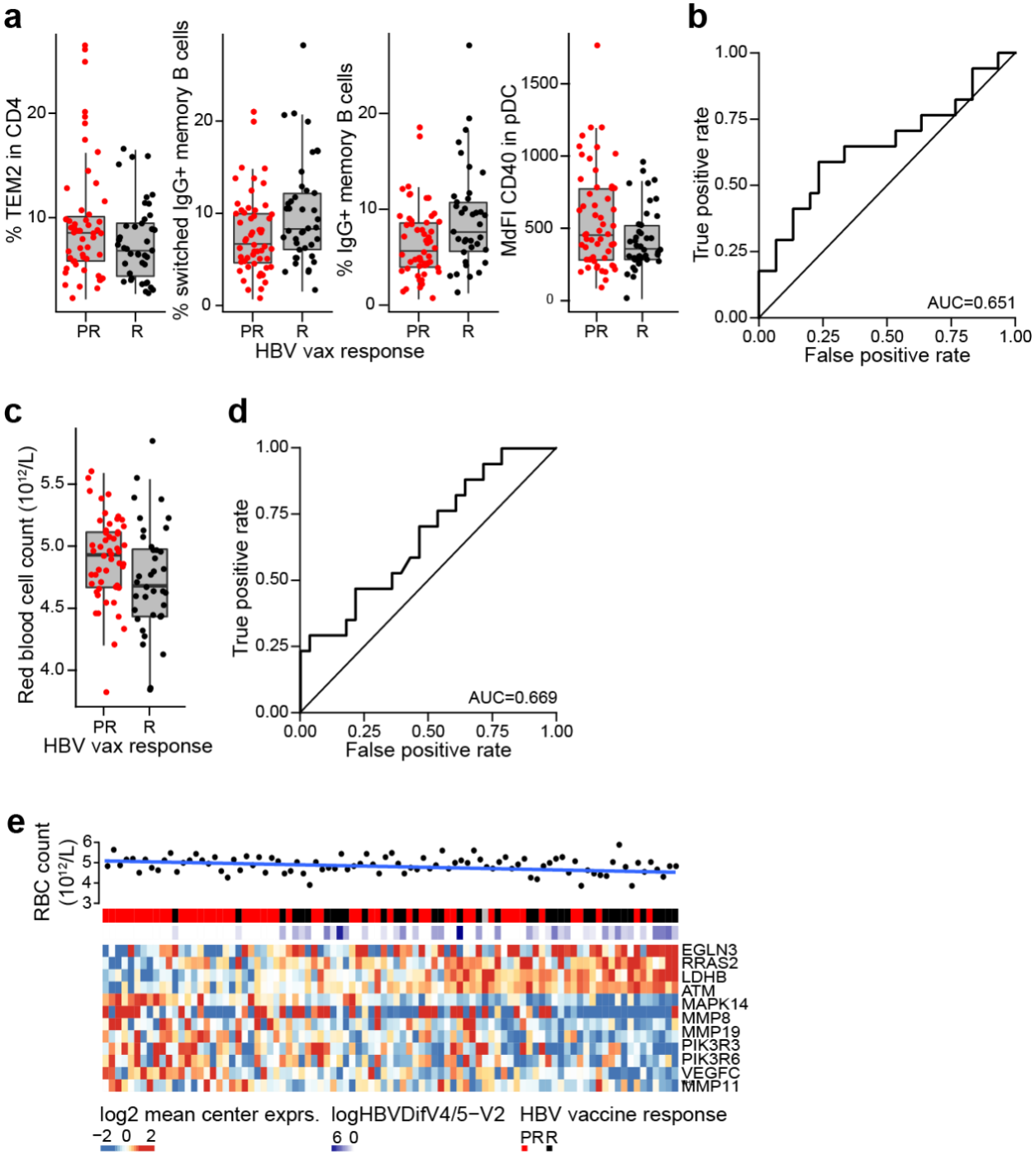


Figure A1.5. B-cell subsets, innate immune cell subsets and RBC counts are predictive of the response to the HBV vaccine

Polychromatic FCM was used to identify cell surface markers of the response to the HBV vaccine. (**a**) Boxplots presenting the frequencies of the four FCM markers selected in the forward selection model in responders and poor-responders to the HBV vaccine on the EM131 training set (n=95). (**b**) ROC curves for the prediction of the HBV vaccine response using the FCM data on the EM131

test set. (**c**) Boxplots presenting the RBC counts (measured in a 28 days window prior to HBV vaccination) in responders and poor-responders to the HBV vaccine on the EM131 training set (n=95). (**d**) ROC curves illustrating the prediction of the HBV vaccine response using RBC counts on the EM131 test set. (**e**) Heatmap representation of the genes differentially expressed between HBV vaccine R and PR overlapping the HIF-1a canonical pathway in the training set. The mean-centered gene-expression is represented using a blue-red color scale. Rows and columns correspond to the genes and the profiled samples, respectively. Samples were ordered by increasing level of their expression of the genes associated to the vaccine response (mean-rank ordering). Antibody response to the HBV vaccine (logHepBDifV4/5-V2) and HBV vaccine response group (HBV vaccine response) and red blood cells counts (RBC in 10^{12}/L) are presented in colored squares above each sample. The *p*-value of *t*-test between RBC and the ordering of the samples is 0.0462.

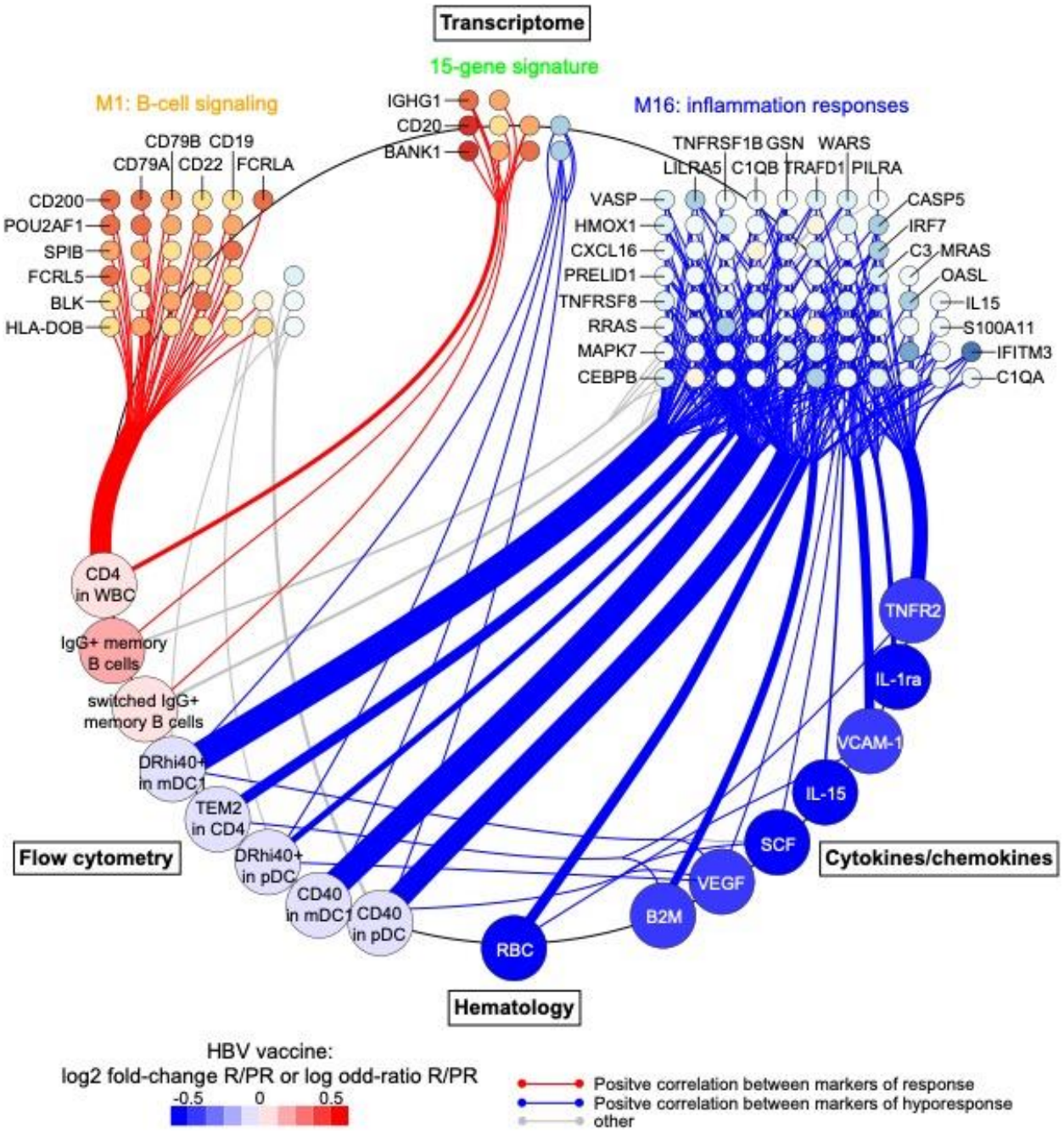


Figure A1.6. Integrative analysis of the transcriptomic, FCM, hematologic and cytokine/chemokine expression data reveals positive correlations between biomarkers of HBV vaccine response

A projection-based multivariate approach was adopted to assess the correlation between transcriptomic, FCM, hematologic and cytokine/chemokine expression datasets. Least square regressions between pairs of datasets was performed using the R package “mixOmics”. The resulting regression coefficients were converted to Pearson correlations between pairs of features

of the different datasets (presented at each quadrant of the figure). Significant (*t*-test: $r \geq 0.188$, $p \leq 0.05$) positives correlations are presented as edges and the features as vertices. The vertices are colored by fold-change between HBV vaccine responders vs. poor-responders of the EM131 training set.

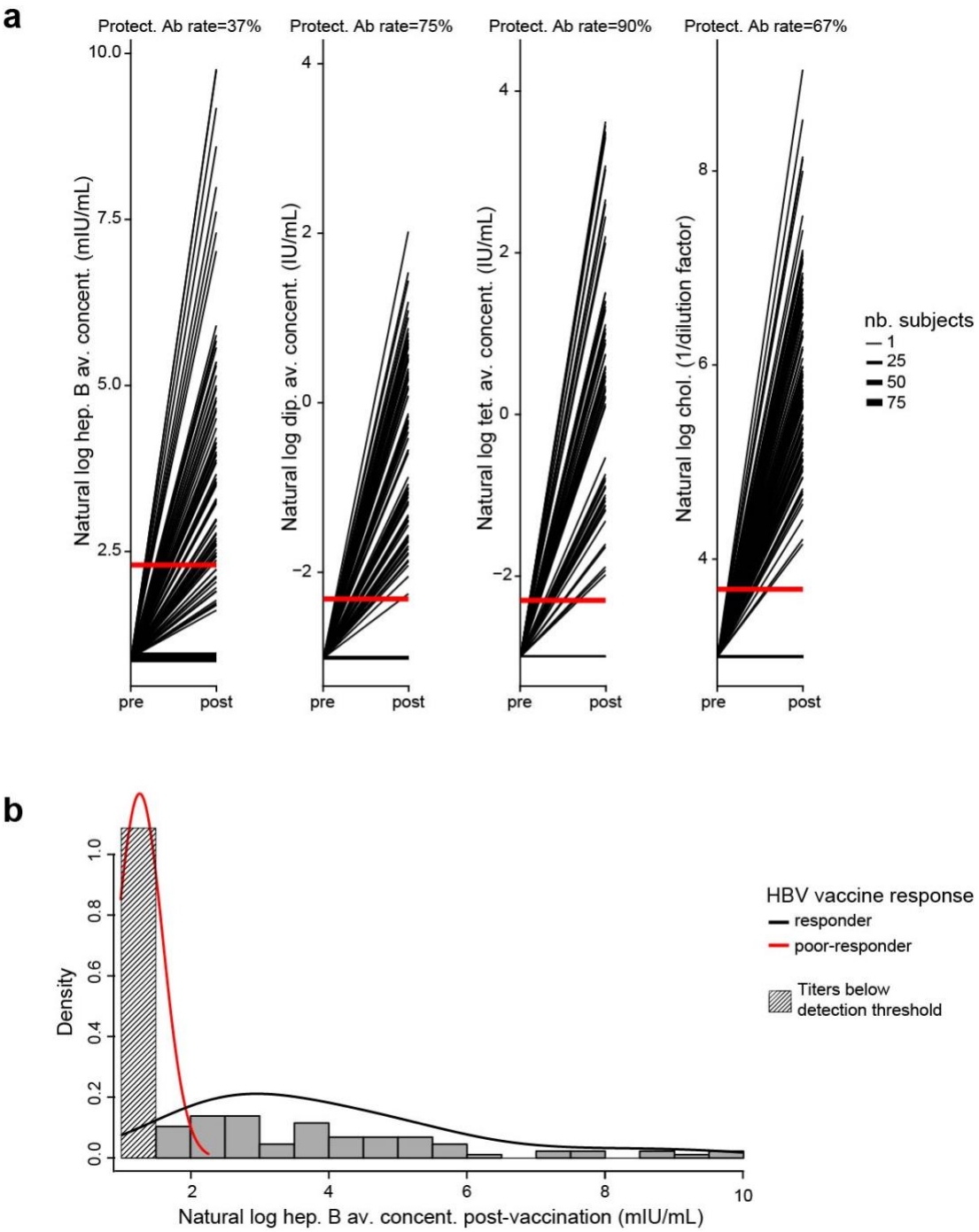
Tables

Table A1.1. Identification of FCM markers and hematologic markers associated with response to with HBV vaccine

Marker	Type of marker	Unit	Univariate logistic regression			Multivariate logistic regression		
			OR	95% CI	*p*-value	OR	95% CI	*p*-value
Flow marker								
TEM2 in CD4	T-cells	%	1.09	[0.997, 1.21]	0.0773	1.12	[1.00, 1,28]	0.0619
switched IgG+ memory B cells	B-cells	%	0.907	[0.825, 0.987]	**0.0310**	1.90	[0.933, 4.24]	0.0932
IgG+ memory B cells	B-cells	%	0.883	[0.790, 0.973]	**0.0177**	0.429	[0.172, 0.948]	**0.0494**
CD4 in WBC	Innate	%	0.939	[0.872, 1.01]	0.0852	0.953	[0.874, 1.03]	0.248
DRhi40+ in pDC	Innate	%	1.03	[0.999, 1.07]	0.0679	1.01	[0.998, 1.08]	0.725
DRhi40+ in mDC1	Innate	%	1.02	[0.998, 1.05]	0.0924	0.945	[0.845, 1.05]	0.307
CD40 in pDC	Innate	MdFI	1.00	[1.00, 1.00]	**0.0332**	1.00	[0.998, 1.00]	0.478
CD40 in mDC1	Innate	MdFI	1.00	[1.00, 1.00]	0.0670	1.00	[0.998, 1.01]	0.264
Hematologic marker								
Red blood cell count (RBC)		10^{12}/L	3.03	[1.01, 10.0]	0.0555			

Univariate and multivariate logistic regression between marker levels pre-vaccination (Visit 2) and response to HBV vaccine in the EM131 training set was performed, separately for FCM markers and hematologic markers. P-values of a normality test (*z*-test) testing the statistical significance of the association are given in the table. Separate multivariate model was build based for FCM markers and for hematologic markers. Forward selection method based on AIC criteria was used to select the markers (highlighted in grey) for each multivariate model.

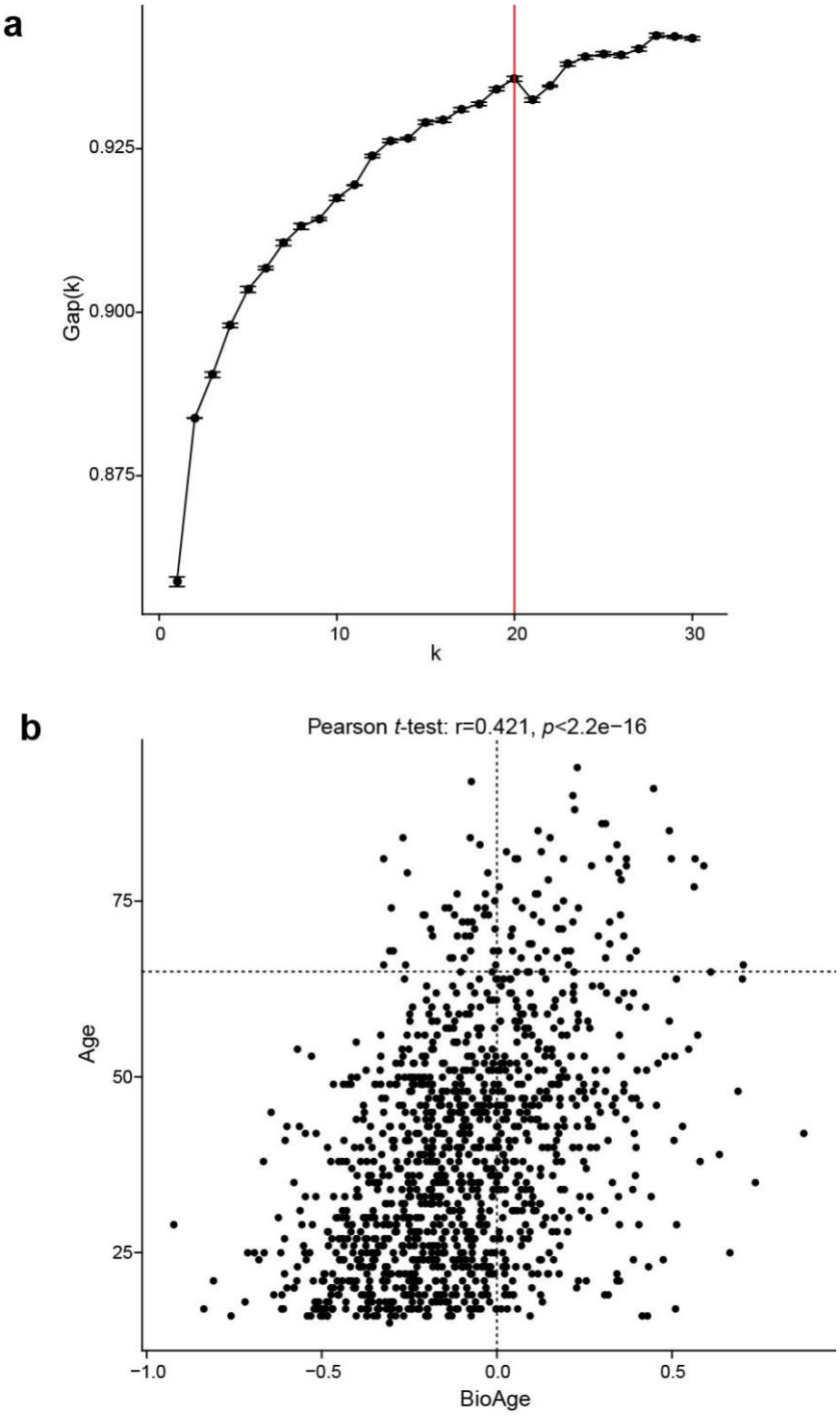
Supplemental Data



Supplementary Figure A1.1. Two groups of responders to the HBV vaccine can be identified based on the antibody response titers

(a) Response plots showing antibody titers as function of the vaccination status (x-axis: before and after receiving the vaccines). The analysis was restricted to subjects with undetectable titers pre-vaccination (hep. B: n= 170, dip: n=105, tet: n=67 and chol: n=151). The titers for HBsAg (hep. B), diphtheria toxin (dip.), tetanus toxin (tet.) and cholera toxin (chol.) are given on natural

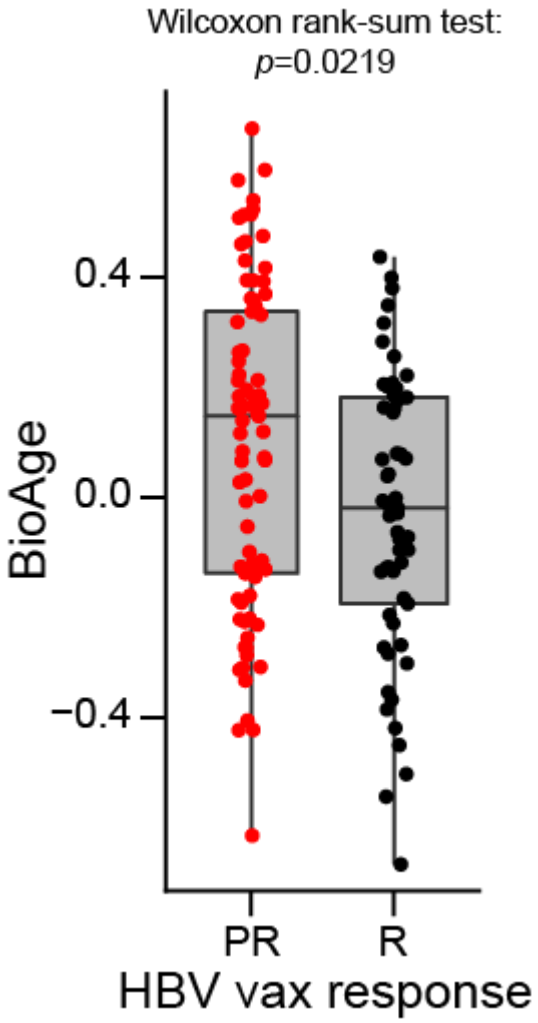
log scale. The unit of every titer is given in the y-axis label. Red horizontal lines indicated standard titer cutoffs used to define protective antibody levels. The percentage of patients having antibody titers above the protective thresholds (Protect. Ab) post-vaccination is indicated above each plot. (b) Density of the anti-HBsAg levels detected post-vaccination. Ab titers are given in mIU/mL in natural log scale space. 94 of the 174 subjects (54%) present Ab titers below the detection threshold of 5 mIU/mL while 79 subjects presented Ab titers ranging between 5.07 to 17,311.4 mIU/mL. Kernel density estimation of the Ab titers revealed two groups of responders based on Ab titers, HBV vaccine poor-responders (red line) with Ab titers below the detection threshold and HBV vaccine responders (black line) with Ab titers above 5 mIU/mL.



Supplementary Figure A1.2. Development of the BioAge signature on the SAFHS cohort

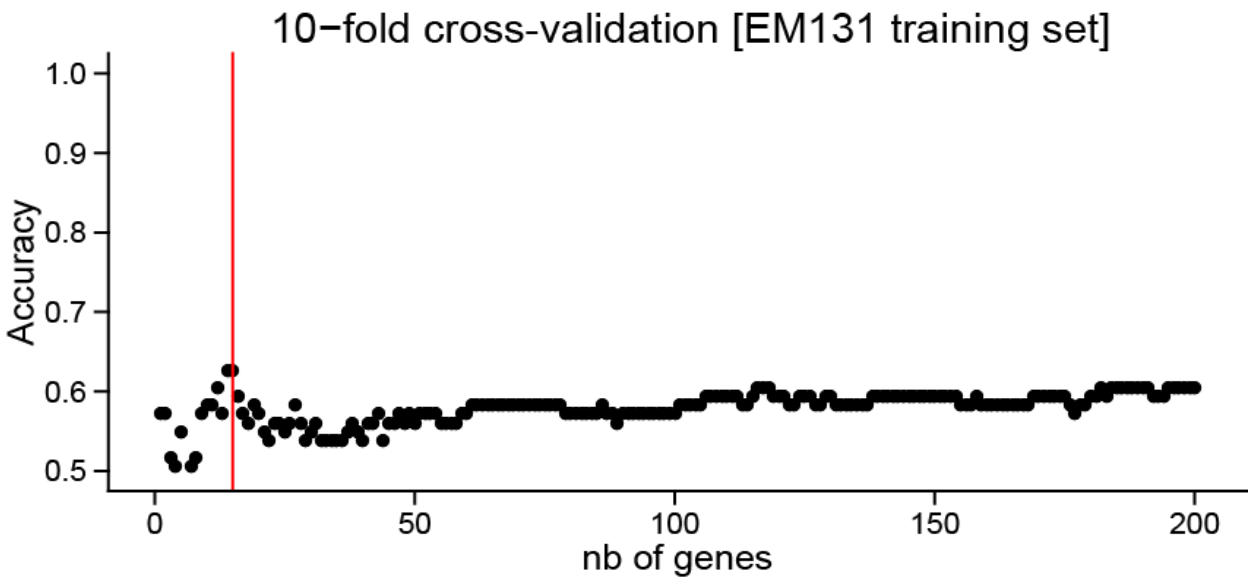
(a) Scatter plot showing the gap statistic estimation of the optimal number of cluster of genes among the 2,285 genes correlated to age on the SAFHS cohort. The x-axis corresponds to the different number of clusters tested and the y-axis corresponds to the gap statistic (Gap) and its 95% confidence interval. The gap statistic corresponds to the between-clusters variance divided

by the intra-cluster variance (the greater the gap statistic the better the fit). The optimal number of clusters was obtained following the rule described in Tibshirani *et al.* defined as the smallest k such that Gap(k) ≥ Gap(k+1) – sd(k+1). Consequently, twenty clusters were identified as the optimal number of clusters of genes. (b) Scatter plot showing the chronological age as a function of the BioAge on the SAFHS dataset. Among the 109 donors aged 65 or over (which is the population with the same age range as the elderly in EM131), 42 (39%) have a young BioAge (BioAge < 0) while 67 (61%) have an old BioAge (BioAge ≥ 0).



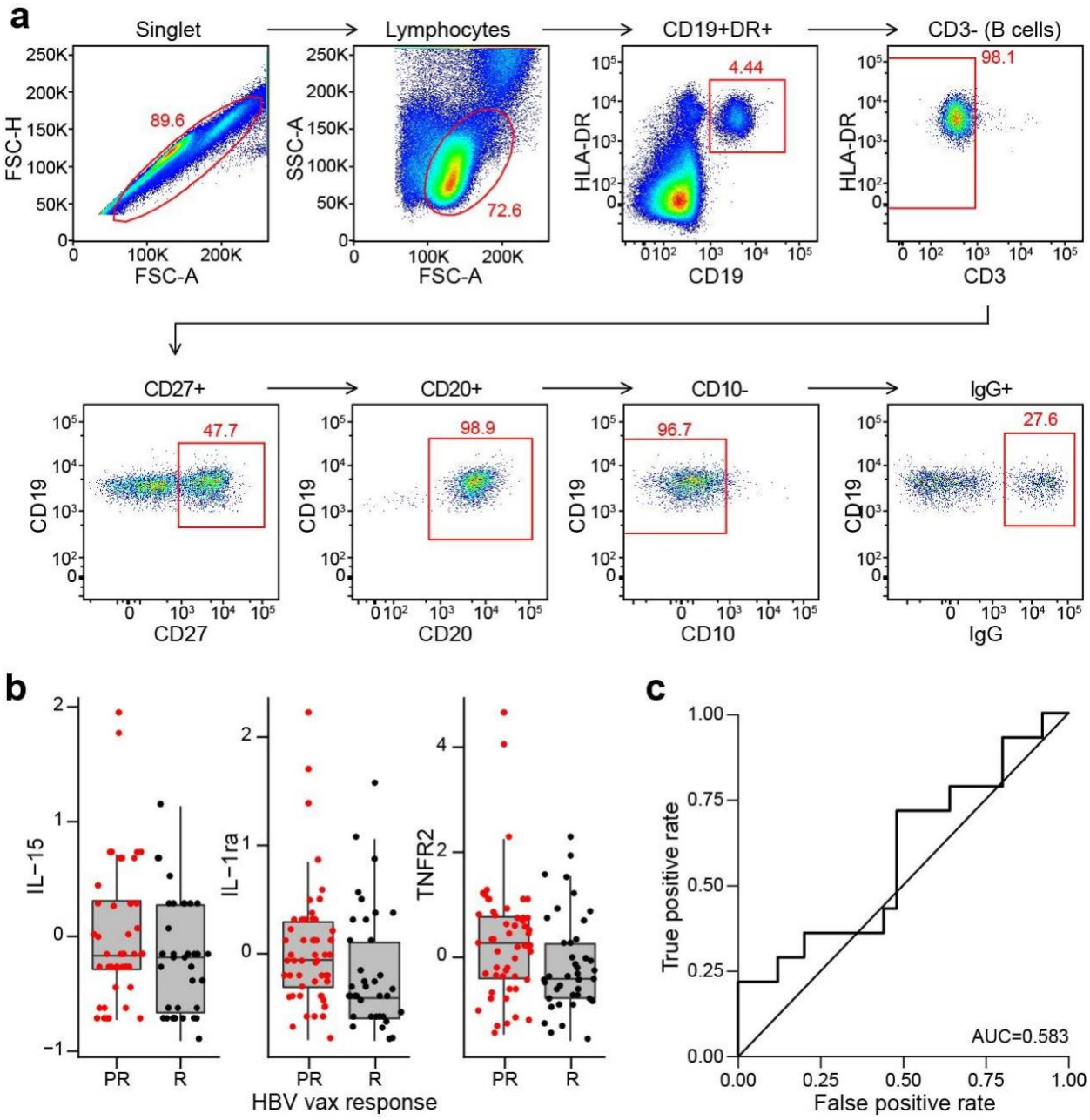
Supplementary Figure A1.3. Boxplot showing the BioAge score as a function of HBV vaccine response on the EM131 elderly cohort

A Wilcoxon rank-sum test was used to test for statistical difference in BioAge between HBV vaccine responders (R) and poor-responders (PR), the *p*-value is presented above the boxplot.



Supplementary Figure A1.4. Scatter plot showing the results of the 10-fold cross-validation on the EM131 training set.

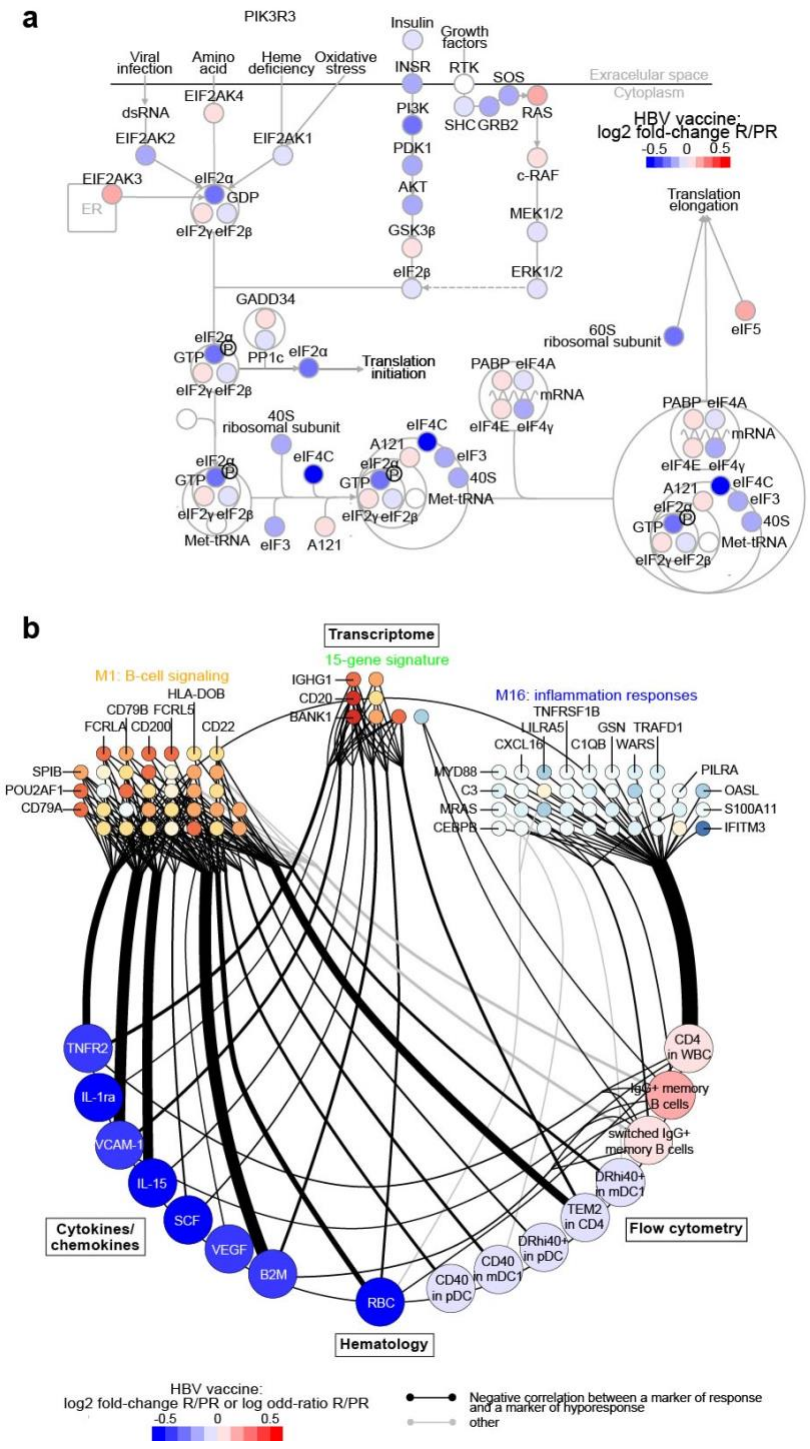
The red line corresponds to naïve Bayes classifier of 15 genes, which gave an estimated accuracy of 62.6% accuracy.



Supplementary Figure A1.5. FCM and cytokines associated with response to the HBV vaccine on the EM131 training set

(a) Examples of our gating procedure for obtaining the percentage of memory IgG+ B cell populations (CD3-CD19+HLA-DR+CD20+CD27+CD10-IgG+) among total B cells (CD3-CD19+HLA-DR+) for subject 2087 at visit 2 (value=12.6%). (b-c) Cytokine-expression was analyzed in order to identify markers of response to the HBV vaccine. (b) Boxplots presenting the levels of three cytokines selected in the forward selection model between responders and poor-responders to the HBV vaccine on the EM131 training set. The three cytokines selected are

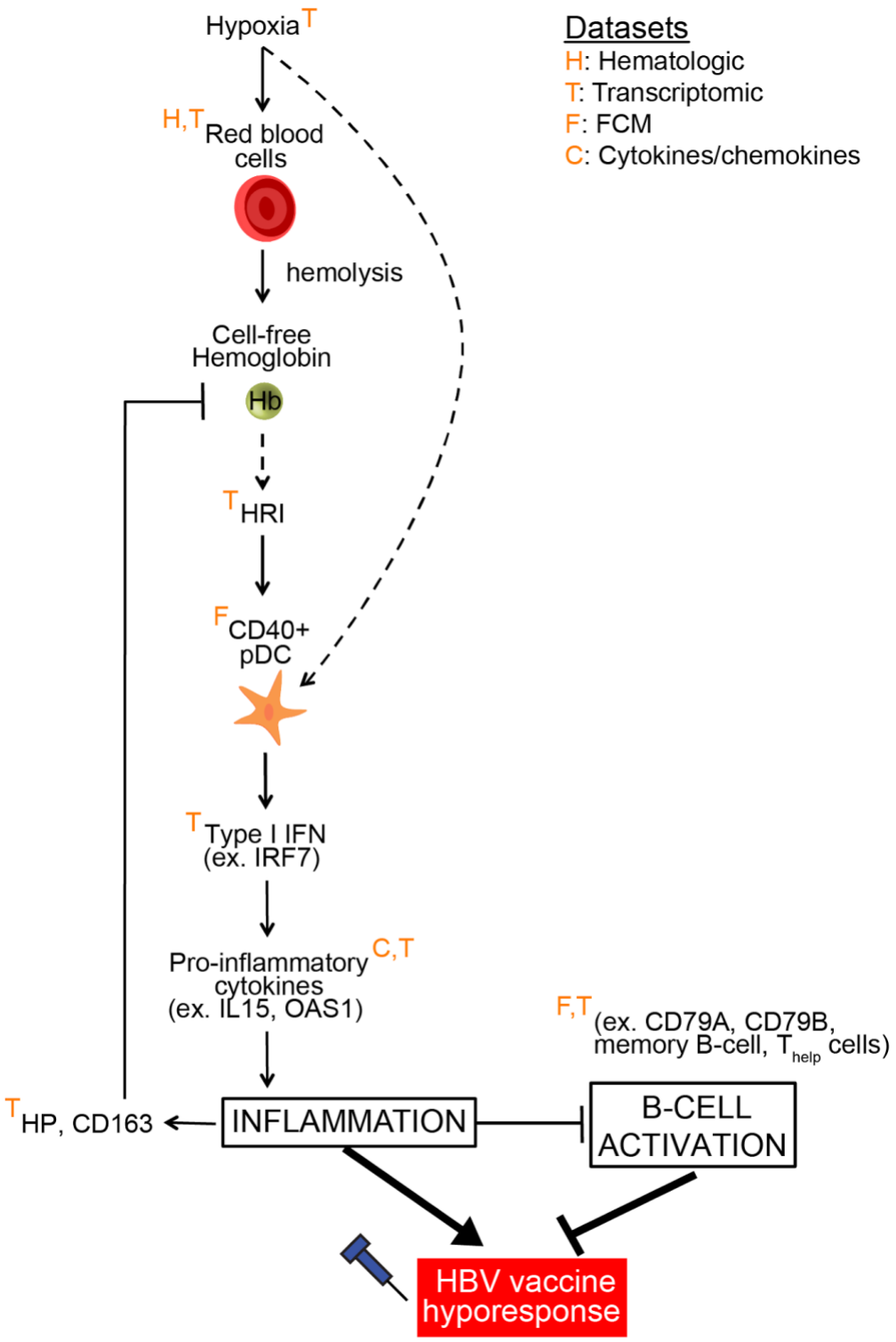
implicated in the inflammatory mechanism. (c) ROC curves for the prediction of the HBV vaccine response using the cytokine data on the EM131 test set.



Supplementary Figure A1.6. Integrative analysis of the transcriptomic, FCM, hematologic and cytokines/chemokines expression data reveals negative correlations between biomarkers of response and hyporesponse to the HBV vaccine

(a) EIF2 signaling pathway is associated with hyporesponse to the HBV vaccine. Ingenuity EIF2 canonical pathway where nodes are colored by the fold-change difference between HBV vaccine

responders (R) compared to poor-responders (PR) as calculated on the EM131 training set. (b) Least square regressions between pairs of “omics” was performed using the R package “mixOmics”. The results regression coefficients were converted to Pearson correlation between pairs of features of the different “omics”. Significant (*t*-test: $p \leq 0.05$) negatives correlations are presented as edges and the features as vertices. The vertices are colored by fold-change between HBV vaccine responders vs. poor-responders of the EM131 training set.



Supplementary Figure A1.7. Proposed mechanism leading to hyporesponse to HBV vaccine Hypoxia (HIF-1α signaling pathway) leads to the increase in red blood cells. Turnover of red blood cells will release free-hemoglobin in the circulation. Hemoglobin, a source of heme-complex, will trigger the activation of the heme-regulated eIF2α kinase (HRI), In turn, HRI will stimulate the

production of type I interferon by pDCs and promote the release of pro-inflammatory cytokines in the system. The resulting inflammatory response will hinder response to HBV vaccination. A negative feedback loop involving haptoglobin (HP) and CD163 will be activated to catalyze the degradation of cell-free hemoglobin. In parallel, the absence of B-cell activation or the low frequency of memory B-cell will result in hyporesponse to HBV vaccine. The datasets (transcriptomic, FCM, cytokines/chemokines and hematologic) supporting the elements of the proposed mechanism are indicated on the figure.

Supplementary Table A1.1. Clinical characteristics of the study cohort

	Complete data set	Training set	Test set	Fisher's exact test/ Wilcoxon test p-value
n	174	95	49	
Age				
range	[25, 83]	[65, 81]	[65, 83]	0.896
[25-40]	30	0	0	0.810
[65-75[	122	81	41	
[75-83]	22	14	8	
Gender				
male	79	44	23	1.00
female	95	51	26	
Race				
white	174	95	49	N/A
Height (cm)				
range	[139, 190]	[144, 182]	[139, 181]	0.666
Weight (kg)				
range	[40.3, 112.4]	[47.7, 102.5]	[40.3, 94.5]	0.634
BMI (kg/m2)				
range	[17.0, 34.9]	[21.5, 33.9]	[17.0, 33.5]	0.452
< 30	140	78	41	0.813
≥ 30	32	16	7	
N/A	2	1	1	
CMV at V2 (IU/mL)				
< 0.75	94	49	24	0.861
≥ 0.75	80	46	25	
HepB at V2 (mIU/mL)				
< 5	171	94	48	N/A
≥ 5	3	1	1	
HepB at V5 (mIU/mL)				
< 5	94	54	30	0.390
≥ 5	79	41	19	
NA	1	0	1	
Cholera at V2 (1/dilution factor)				
< 40	152	83	43	1.00
≥ 40	22	12	6	
Cholera at V4 (1/dilution factor)				
< 40	50	25	13	1.00
≥ 40	123	70	35	
N/A	1	0	1	
Dip at V2 (IU/mL)				
range	]0, 5.72]	]0, 5.72]	]0, 1.61]	0.401
Dip at V4 (IU/mL)				
range	]0, 51.1]	]0, 24.3]	]0, 51.1]	0.377
Tet at V2 (IU/mL)				
range	]0, 24.9]	]0, 19.4]	]0, 24.9]	0.628
Tet at V4 (IU/mL)				
range	]0, 71.3]	]0, 71.3]	]0, 41.6]	0.696

All participants were of European descent with BMI below 35 kg/m^2; ~55% were women; and none had histories of autoimmunity, immunomodulating medications, cancer, recent vaccination or recent infection. The Fisher's exact test (categorical variable) and the Wilcoxon Rank sum test (continuous variable) were used to assess difference between the EM131 training set and EM131 test set. No significant difference in term of clinical (BMI: body mass index) or serological (HepB: hepatitis B, Dip: diphtheria, Tet: tetanus) variables was observed between those two sets.

Supplementary Table A1.2. Logistic regression with poor-response to HBV vaccination

	Univarite			Multivariate		
	OR	95% CI	*p*-value	OR	95% CI	*p*-value
Age ([65, 83] *v* [25, 65[)	2.80	[1.24, 6.69]	**0.0158**	1.52	[0.593, 4.03]	0.385
Sex (male *v* female)	2.80	[1.50, 5.33]	**0.00142**	2.44	[1.26, 4.80]	**0.00847**
Height (per cm)	1.03	[0.994, 1.07]	0.105	---	---	---
Weight (per kg)	1.02	[0.998, 1.05]	0.0832	---	---	---
BMI (per kg/m2)	1.06	[0.959, 1.17]	0.272	---	---	---
CMV (≥0.75 *v* <0.75 IU/ml)	1.26	[0.688, 2.33]	0.453	---	---	---
BioAge (per unit)	5.62	[1.82, 18.6]	**0.00343**	4.14	[1.21, 15.2]	**0.0270**

Univariate and Multivariate logistic regression against response to HBV vaccine was performed using clinical variables and the BioAge signature. Only variables presenting a significant association with response to HBV vaccine at the univariate level (z-test: *p*-value ≤ 0.05) were included in the multivariate level. Age, gender and the BioAge are significantly associated with HBV vaccine response in the univariate analysis. The BioAge signature maintains significant association with HBV vaccine response even after correcting for clinical variable like age and gender.

Article #2 : Integrated systems approach defines the antiviral pathways conferring protection by the RV144 HIV vaccine

Mise en contexte

Dans le premier article de cette thèse, nous avons démontré qu'il était possible d'identifier des classificateurs basés sur des biomarqueurs mesurés avant la vaccination pour prédire la réponse au vaccin contre l'HepB. De plus, l'analyse intégrative de données large échelle nous a permis d'identifier un mécanisme potentiel par lequel ces biomarqueurs peuvent contribuer à une meilleure réponse au vaccin.

En outre, cet article démontre que seules certaines voies de signalisation impactées par l'âge sont prédictives de l'hypo-réponse aux vaccins. Dans le cas du vaccin contre l'HepB une réponse mémoire (pas forcément spécifique aux vaccins, car les patients étaient vaccin-naïfs) ainsi qu'un faible niveau inflammation sont prédictifs de l'amplitude de la réponse vaccinale. Ces résultats sont inattendus, car un fort niveau de cellules B mémoire suggèrerait une réduction de la quantité de cellules B naïves nécessaire à la production de plasmablastes et d'Acs. Une forte inflammation prévaccination, caractérisée par une réponse aux interférons et un niveau élevé de monocytes, est aussi surprenante, car ces individus étaient en pleine santé. Méthodologiquement, une approche semi-supervisée et supervisée ont toutes deux mené à l'identification de signatures avec peu de gènes en commun, mais reflétant les mêmes voies de signalisation.

Dans le 2^{e} article de cette thèse, nous avons tenté d'appliquer les mêmes méthodes pour identifier des biomarqueurs de la réponse au vaccin contre le VIH RV144 en utilisant des données large échelle mesurées 2 semaines après la vaccination. Nous avons également appliqué les mêmes méthodes intégratives pour confirmer le mécanisme biologique induit par le vaccin et capable de conférer une protection contre le VIH.

En effet, l'essai clinique du vaccin RV144 a démontré une réduction du risque d'acquisition du VIH-1 de près de 31.2%. Toutefois les mécanismes qui ont conduit à cette protection partielle n'ont pas été identifiés. Dans cet article, nous avons tenté d'identifier des corrélats transcriptionnels du risque d'infection par le VIH-1 après la vaccination. Les profils transcriptionnels du sang prélevé sur 223 participants ont été utilisés avec des méthodes d'apprentissage machine pour identifier ces corrélats potentiels. De plus, une approche intégrative multiomique a été utilisée pour combiner des données transcriptionnelles, de cytométrie en flux et de protéomique plasmatique pour déterminer les mécanismes induits par le vaccin RV144 qui permettent cette protection partielle.

L’intégralité de l’analyse bio-informatique des données transcriptionnelles, de cytométrie de flux et de cytokines a été ma responsabilité. Le développement de classificateurs et l’analyse intégrative multiomiques ont aussi été ma responsabilité. La normalisation des données du jeu RV144 pilote et des données de cytométrie en flux ont été effectuées par d’autres bio-informaticiens. J’ai généré la totalité des figures et rédigé la totalité de l’article sous la supervision du dernier auteur.

Integrated systems approach defines the antiviral pathways conferring protection by the RV144 HIV vaccine.

Slim Fourati[1], Susan Pereira Ribeiro[1], Filipa Blasco Tavares Pereira Lopes[1], Aarthi Talla[1], Francois Lefebvre[2], Mark Cameron[3], J Kaewkungwal[4], P Pitisuttithum[4], S Nitayaphan[5], S Rerks-Ngarm[6], Jerome H Kim[7,8], Rasmi Thomas[7], Peter B Gilbert[9], Georgia D Tomaras[10], Richard A Koup[11], Nelson L Michael[7], M Juliana McElrath[9], Raphael Gottardo[9], Rafick-Pierre Sékaly[12]

Affiliations

[1]Department of Pathology, Case Western Reserve University, Cleveland, OH, 44106, USA.

[2]Canadian Center for Computational Genomics, Montréal, QC, H3A 0G1, Canada.

[3]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, 44106, USA.

[4]Faculty of Tropical Medicine, Mahidol University, Bangkok, 10400, Thailand.

[5]Royal Thai Army, Armed Forces Research Institute of Medical Sciences, Bangkok, 10400, Thailand.

[6]Department of Disease Control, Ministry of Public Health, Nonthaburi, 11000, Thailand.

[7]Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, MD, 20910, USA.

[8]International Vaccine Institute, Seoul, 08826, Korea.

[9]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA.

[10]Duke Human Vaccine Institute, Duke University, Durham, NC, 27710, USA.

[11]Vaccine Research Center, US National Institutes of Health, Bethesda, MD, 20892, USA.

[12]Department of Pathology, Case Western Reserve University, Cleveland, OH, 44106, USA. rafick.sekaly@case.edu.

This work was originally published in Nature Communications:

Nat Commun. 2019 Feb 20;10(1):863. doi: 10.1038/s41467-019-08854-2. PubMed PMID: 30787294; PubMed Central PMCID: PMC6382801.

Abstract

The RV144 vaccine trial showed reduced the risk of HIV-1 acquisition by 31.2%, although mechanisms that led to protection remain poorly understood. Here we identified transcriptional correlates for reduced HIV-1 acquisition after vaccination. We assessed the transcriptomic profile of blood collected from 223 participants and 40 placebo recipients. Pathway-level analysis of HIV-1 negative vaccinees reveals that type I interferons that activate the IRF7 antiviral program and type II interferon-stimulated genes implicated in antigen-presentation are both associated with a reduced risk of HIV-1 acquisition. In contrast, genes upstream and downstream of NF-κB, mTORC1 and host genes required for viral infection are associated with an increased risk of HIV-1 acquisition among vaccinees and placebo recipients defining a vaccine independent association with HIV-1 acquisition. Our transcriptomic analysis of RV144 trial samples identifies IRF7 as a mediator of protection and the activation of mTORC1 as a correlate of the risk of HIV-1 acquisition.

Introduction

The RV144 trial evaluated the efficacy of ALVAC-HIV (vCP1521) prime and AIDSVAX B/E (gp120) boost strategy adjuvanted in alum to prevent Human Immunodeficiency Virus 1 (HIV-1) acquisition. Participants enrolled in the RV144 clinical trial were followed up to 3 years after a series of four immunizations. The vaccine reduced the risk of HIV-1 acquisition at 3 years following completion of the vaccination series by 31.2% when compared to placebo (modified intention-to-treat analysis, Likelihood-ratio test: p=0.0385)[1]. Correlates of risk studies showed that two non-neutralizing antibody responses measured two weeks after vaccination were associated with HIV-1 acquisition: levels of IgA recognizing the Envelope (Env) region of HIV-1 associated with a higher risk of HIV-1 acquisition (similar to the risk of placebo recipients) and levels of IgG recognizing the V1/V2 regions of HIV-1 Env associated with a decreased risk of HIV-1 acquisition[2]. More recently, CD4+ T cells polyfunctionality measured in response to Env stimulation (i.e. polyfunctionality score; PFS) was associated with a decreased risk of HIV-1 acquisition in RV144 vaccinees[3], while two host Human leukocyte antigen (HLA) alleles (DBQ1*06 and DPB1*13) were shown to modulate the HIV-specific antibody responses associated with HIV-1 acquisition[4]. The benefit of combining these correlates, that underlie different arms of the immune response (humoral and cellular), to predict the risk of HIV-1 acquisition among RV144 vaccinees has not been assessed. Moreover, a specific innate immune response that can help prime cellular and humoral immune effector mechanisms are yet to be defined.

Despite the identification of correlates of risk for the RV144 vaccine that could potentially be correlates of protection[5], the mechanisms that lead to vaccine conferred-protection are still unknown. IgG antibodies elicited by the vaccine have been suggested to trigger antibody-dependent cell-mediated cytotoxicity (ADCC) by binding Fc receptors on the surface of natural killer cells, whereas IgA antibodies compete with IgG antibodies for binding to HIV-1 Env and thus abrogate ADCC in vaccinees[6]. Conversely, the ALVAC vector has been reported to trigger cytosolic pattern-recognition receptors sensing double-stranded DNA leading to the activation of IRF3/IRF7 and the induction of an innate immune antiviral response that could prevent HIV-1 infection[7]. Understanding the mechanisms that led to RV144 vaccine-conferred protection could offer new insights into the development of more effective HIV vaccines.

In this study, we assessed the transcriptomic profile of in vitro stimulated peripheral blood mononuclear cells (PBMCs) taken from 223 vaccinees and 40 placebo-recipient two weeks after the last immunization with the RV144 vaccine (or placebo). We identified that IFNγ stimulated genes are associated with reduced risk of HIV-1 acquisition among vaccinees. Genes downstream of NF-κB and mTORC1 required for viral infection or replication are instead associated with increased risk of HIV-1 acquisition in both vaccinees and placebo recipients, with a mechanism independent of the vaccine-induced immune response.

Results

The RV144 vaccine induces IFNγ, NF-κB and mTORC1 pathways

As an initial step, a pilot study was conducted to identify (through differential gene-expression analysis) and down-select (through clustering) RV144 vaccine-induced transcriptomic signatures that would then be tested for their association with HIV-1 acquisition[2] (**Fig. A2.1**). We compared the transcriptomic profile of in vitro HIV-1 Env-stimulated PBMCs obtained pre-immunization and two weeks after the last immunization from 40 vaccine recipients and 10 placebo recipients; all 50 participants were HIV-1 negative at the last follow-up visit (**Supplementary Tables A2.1-2**)[2]. Linear regression models followed by Gene Set Enrichment Analysis (GSEA) were used to identify pathways that were differentially regulated when comparing the vaccine and placebo groups (see **Methods**). Linear regression revealed that the expression of 2946 genes was altered post-immunization (differentially expressed between the vaccine and placebo groups, LIMMA: moderated t-test $p \leq 0.05$); none of these genes showed differential expression after correction for multiple testing. GSEA identified eleven pathways significantly enriched among genes induced post-immunization compared to pre-immunization in Env-stimulated PBMCs of vaccinees while none could be detected in Env-stimulated PBMCs of placebo recipients (DB: Hallmark, GSEA:

False Discovery Rate (FDR) ≤ 5%, **Fig. A2.2a**). Sample Level Enrichment Analysis (SLEA), a method allowing to average expression of genes of a given pathway per sample, was used to quantify the enrichment of those 11 pathways for each sample of the pilot study. These 11 pathways were separated into four groups based on their correlation across samples (**Fig. A2.2b** and **Supplementary Fig. A2.1**). In vaccinees, Env stimulation led to the induction of genes part of the IFNγ response pathway (englobing as well IFNα response genes), genes implicated in NF-κB signaling, genes downstream of mTORC1 as well as genes associated to allograft rejection, i.e. genes triggered by T cell activation. Known IFNγ response genes specifically induced in RV144 vaccinees included genes implicated in antigen presentation by the major histocompatibility complex (MHC) class I (FCGR1A, HLA-B, HLA-G, ICAM1[8], TAP1) and by MHC class II (CD74, HLA-DQA1, HLA-DMA, HLA-DRB1). Several of those mediate antiviral responses (IFIH1[9], MYD88[10], TRAFD1[11]) or have described antiviral effects against HIV-1 (APOL6, LGALS3BP, RSAD2). Stimulation of PBMCs with HIV-1 Env was not associated with increased expression of these genes in placebo recipients (**Fig. A2.2c**). Likewise, several genes coding for members of the NF-κB transcriptional complex (NFKBIA, REL), genes encoding upstream activators of the NF-κB transcriptional activity (BIRC3, IL1B, RIPK2, TANK, TNF) as well as several NF-κB transcriptional targets (CXCL2, ICAM1, IL6, IL8, SDC4) were specifically induced in Env stimulated cells from RV144 vaccinees. Several mTORC1 downstream targets (GCLC, SCD, TFRC) were induced as well by Env stimulation of PBMCs from RV144 vaccinees. Together these results highlight IFNγ-, NF-κB- and mTORC1-regulated genes as major transcriptional targets of the RV144 vaccine in Env-stimulated PBMCs.

IFNγ pathway is associated with a reduced risk of acquisition

To determine if changes in gene expression were associated with HIV-1 acquisition, we analyzed the transcriptomic profiles of HIV-1 Env in vitro stimulated PBMCs obtained two weeks after the last immunization from 183 vaccine recipients part of the case-control study[2] that included 30 participants that acquired HIV-1 after vaccination (cases) and 153 participants that remained HIV-1 negative during the 3 year follow-up (controls). In addition to case-control samples, this dataset also included 30 placebo recipients, 17 of which acquired HIV-1 after vaccination and 13 that remained HIV-1 negative (**Supplementary Tables A2.3-4**). We used linear regression models and GSEA to assess the association of the four pathways induced by the RV144 vaccine with HIV-1 acquisition; this analysis was performed separately for vaccinees and placebo recipients. 2058 genes were differentially expressed between participants that acquired HIV-1 and those that remained HIV-1 negative within the placebo group, while 3009 genes were differentially

expressed between those two groups of participants in the vaccine group (LIMMA: t-test $p \leq 0.05$); none of these genes remained significative after correction for multiple testing. Our analysis revealed that 3 out of the 4 pathways induced by the RV144 vaccine (described above) were significantly associated with HIV-1 acquisition in the vaccine group, namely genes part of the IFNγ response pathway, genes implicated in NF-κB signaling and genes downstream of mTORC1 (GSEA: FDR ≤ 0.05; **Fig. A2.3**). The IFNγ response pathway was unique in its association with a lower risk of HIV-1 acquisition. In contrast, the remaining two pathways, NF-κB and mTORC1 signaling, were associated with higher risk of HIV-1 acquisition for RV144 vaccinees (**Fig. A2.3**, **Supplementary Table A2.5**).

The association between the induction of the IFNγ response pathway and lower risk of HIV-1 acquisition was not observed for the placebo group suggesting that the IFNγ response pathway is associated with a vaccine-conferred reduced risk of HIV-1 acquisition (**Fig. A2.3**). These IFNγ response genes included genes involved in the maturation of the MHC class II complex (AP2A1 coding for the AP2 vesicle, CTSA, CTSB, CTSD coding for cathepsins A, B and D) and genes involved in MHC class II antigen processing (LGMN, IFI30) suggesting that these genes are critical for the class II MHC restricted Env-specific response. Activation of the IRF7 innate antiviral program (IFIH1[9], IRF7; **Supplementary Fig. A2.2**) was also associated with vaccine-conferred protection of HIV-1 acquisition (**Supplementary Fig. A2.2**). We also observed the enrichment of putative binding sites for IRF7 in the promoter regions of genes associated with a reduced risk of HIV-1 acquisition (12/53 of IFNγ response genes identified in the transcriptomic analysis are also part of the geneset V$IRF7_01. This geneset includes genes that have a putative IRF7 binding site within ± 2000 base pairs of their transcriptional starting sites), genes transcriptionally repressed in IRF7 siRNA transfection experiments[12] and genes induced by the overexpression of IRF7[13,14]. These results highlight the role of IRF7 as a transcriptional regulator of the reduced risk of HIV-1 acquisition conferred by the RV144 vaccine.

While the IFNγ response pathway was associated with reduced risk of acquisition in vaccinees, the NF-κB and mTORC1 signaling pathways were associated with increased risk of HIV-1 acquisition in both vaccine and placebo recipients (**Fig. A2.3**). Genes implicated in NF-κB signaling that were associated with an increased risk of HIV-1 acquisition included genes of the NF-κB complex (NFKB1, RELA, RELB) and activators of the NF-κB complex (BMP2, RIPK2) (**Supplementary Fig. A2.3**). Genes implicated in mTORC1 signaling that were associated with an increased risk of HIV-1 acquisition included regulators of the mTORC1 complex (CXCR4, DDIT4, NAMPT, XBP1) and its downstream targets (CDKN1A[15], SLC2A3[16], TFRC[17]). These results suggest that induction of IFNγ response genes is a vaccine-induced correlate of reduced

risk of HIV-1 acquisition while induction of NF-κB related and mTORC1-related genes are vaccine-independent mechanisms associated with increased risk of HIV-1 acquisition.

IFNγ pathway associated with HIV-specific antibodies and CD4s

The six antibody and cellular assays which were previously assessed for their association with HIV-1 risk in RV144 vaccinees[2] as well as cell counts for seven cell subsets measured by flow cytometry (FCM), seven Luminex markers, six intracellular cytokine staining (ICS) and 31 MHC class II alleles measurements[4] were included in an integrative analysis. A projection-based approach that minimizes the technical effect specific to each dataset by generating a unique scale (i.e. projecting) for every dataset allowing to assess the correlation between different datasets was used for the integrative analysis[18]. This integrative analysis revealed that the IFNγ signaling pathway was significantly positively correlated to the frequency of Env-specific CD4+ T cells (including IFNγ producing T cells) and to titters of IgG against V1/V2, a described correlate of low risk of HIV-1 acquisition in RV144 vaccinees (**Supplementary Fig. A2.4**). The IFNγ signaling pathway was also associated (through Env-specific CD4+ T cells) to the frequency of pDCs, the primary source of type I interferons (interferons α/β) in blood as well as to the polyfunctionality score (PFS), another previously described correlate of low risk in RV144 vaccinees that integrates the cytokine response of CD4+ T cells in response to HIV-1 Env. The integrative analysis highlighted the relevance of the pathway identified above as they show their association with previously identified correlates of risk that underlie the major effector pathways of the immune response (IgG against V1/V2 and PFS).

This integrative analysis also revealed that the NF-κB pathway was significantly correlated with a higher frequency of monocytes and with heightened levels of the pro-inflammatory cytokines IL2 and IL3 (**Supplementary Fig. A2.4**). To further investigate the association between the NF-κB pathway and specific cell subsets including monocytes, we performed a deconvolution[19] (i.e. separation) of the PBMC gene expression profile into six major immune subsets (B cells, T cells, NK, monocytes, mDC and pDC) using the frequencies of those subsets measured by FCM (see **Methods** and **Supplementary Fig. A2.5a**). Analysis of the deconvoluted gene-expression profiles revealed that genes of the NF-kB pathway were expressed at significantly higher levels in monocytes compared to the other five subsets (**Supplementary Fig. A2.5b**, Wilcoxon rank-sum test: p=1.64e-06), thereby supporting the results of the integrative analysis (**Supplementary Fig. A2.4**).

The mTORC1 pathway is a marker of HIV-1 acquisition

To test if the IFNγ signaling pathway could be a novel marker of low infection risk in vaccinees, we built a logistic regression model combining gender, behavior risk and previously described correlates of risk (IgA against Env, IgG against V1/V2, DQB1*06 allele, DPB1*13 allele and PFS) to predict acquisition of HIV-1 among RV144 vaccinees. This model was then compared to one that included results from gene-expression profiling. The balanced accuracy of both models was assessed by 10-fold cross-validation. The best model built without gene-expression showed a balanced accuracy of 62.3% while the best model that included results from gene-expression had a balanced accuracy of 67.9% (**Fig. A2.4a**). Receiver operating characteristic (ROC) analysis was performed to compare these two models; no statistically significant gain in accuracy resulted from the addition of gene-expression results to previously described correlates (**Fig. A2.4b**). A multivariate logistic regression was built using all the candidate markers of protection to assess their relative contribution to the prediction of HIV-1 acquisition among RV144 vaccinees (**Table A2.1**). Only the PFS, the interaction term IgA:DQB1*06, the interaction term IgG:DPB1*13 (i.e. association between IgG level with HIV-1 acquisition separately for DPB1*13- and DPB1*13+ vaccinees) and mTORC1 signaling remained statistically significantly associated to HIV-1 acquisition in a multivariate model. The IFNγ pathway identified by the transcriptional profiling (univariate analysis: Odd ratio=0.883 p=0.00837; **Supplementary Table A2.6**) did not bring an independent contribution to the prediction of the risk of HIV-1 acquisition among RV144 vaccinees (multivariate analysis: Odd ratio=0.974 p=0.677; **Table A2.1**). This analysis suggests that the IFNγ pathway identified by gene-expression was likely confounded (i.e. bring similar predictive information) with the cellular (PFS) and serological/genetic (IgG:DPB1*13) correlates of the RV144 vaccine-conferred low risk of HIV-1 acquisition (**Supplementary Table A2.6**).

We then assessed whether the IFNγ pathway was confounded with IgG:DPB1*13. Thus, we stratified the RV144 vaccinees by the DPB1*13 allele and evaluated the association between the IFNγ pathway and IgG antibodies binding to V1/V2. Stratifying RV144 vaccinees by the DPB1*13 allele revealed that the IFNγ signaling pathway and IgG antibodies binding to V1/V2 were positively correlated to each other only in DPB1*13+ vaccinees (**Fig. A2.5a-b**); moreover, the association of IFNγ signaling pathway with low risk was more pronounced in DPB1*13+ vaccinees (**Fig. A2.5c**). Conversely, the IFNγ signaling pathway was not correlated to IgG antibodies binding to V1/V2 nor HIV-1 acquisition in DPB1*13- vaccinees. A significant overlap of 38 genes was observed between IFNγ response genes correlated with IgG antibodies binding to V1/V2 in DPB1*13+ vaccinees (103 IFNγ response genes) and IFNγ response genes negatively associated with HIV-1 acquisition (53 IFNγ genes; Fisher's exact test: p=0.00367). Those 38 genes included the transcription factor IRF7 and its target genes known to block viral entry (XCL1)

or prevent HIV-1 virion assembly in infected cells (IFITM3, ISG15, MX2, TRIM26). Several genes encoding components of the killing machinery required for ADCC function i.e. CASP3[20], FAS[21], TNFSF10[21] were included among the genes that correlated with titers of IgG antibodies binding to V1/V2 suggesting that ADCC responses could correlate with the decreased rate of virus acquisition after vaccination. Concomitantly, we observed within the IRF7 signature the expression of genes that could be involved in dampening Th1 cell development (IL18BP) or in driving the response towards Th2 cells (PARP14). Genes with a potent anti-inflammatory activity that can suppress global immune activation (SERPING1, LY6E) were also expressed in the protective transcriptomic signature. ITGB7, an integrin that is expressed by T cells and NK cells known for homing the gut was included in the IRF7 target genes that were associated to protection from simian immunodeficiency virus (HIV-like virus) acquisition[22]. While several of the 38 genes included in the IRF7 signature were known to be regulated by IFNγ, 24 out of 38 were also IFNα stimulated genes. In vitro experiments performed on healthy blood samples confirmed that IRF7 phosphorylation (measured by flow cytometry) was induced by more than 1.5-fold upon treatment with interferons (IFNα, IFNβ or IFNγ). Interferon treatment rendered host cells on average 8 times more resistant to in vitro HIV-1 infection (**Fig. A2.5d-e** and **Supplementary Fig. A2.6**). These results show for the first time the possible contribution of innate antiviral immune responses to lower the risk of HIV-1 acquisition among RV144 vaccinees and similar to what was previously reported by Haynes et al.[2], they highlight the contribution of innate cellular functions (NK/ADCC) as correlates of risk in RV144, suggesting that these innate immune functions may play an essential role as correlates of RV144 vaccine-protection.

The IFNγ signaling pathway that correlated with PFS, a correlate of reduced risk of HIV-1 acquisition distinct from IgG antibodies against V1/V2, included a different set of genes with the transcription factor STAT1 as their key regulator (**Supplementary Fig. A2.7**). Several genes involved in the upregulation of MHC class I (TAP1, TAPBP) and class II antigen presentation (HLA-DRB1, HLA-DMA), the initial step in the priming of T cell responses, as well as genes involved in the development of helper T cell functions (IL15, IL15RA, IL4R) were involved in the positive correlation observed between the STAT1 target genes part of the IFNγ pathway and the PFS score.

TNFα signaling via NF-κB (41 genes) and mTORC1 (39 genes) signaling pathways were not associated with previously identified correlates of the RV144 vaccine response. Herein we provide evidence that these two pathways were associated with an increased risk of HIV-1 acquisition both in placebo- and vaccine-recipients (**Fig. A2.6**). The NF-κB signaling pathway included markers of activated T cells and their survival (BCL2A1, IL12B, TNFSF9), cell migration (EFNA1,

CXCL2, CCL20) and induction of proinflammatory prostaglandins (PTGS2). The mTORC1 pathway, with BHLHE40 as the upstream transcription factor, included genes known to be important for HIV-1 entry into target cells (co-receptor for HIV-1 on CD4+ T cells CXCR4, SLC2A1, UNG) as well as genes involved in HIV replication (ETF1, HMGCS1, PGM1). Both mTORC1 and NF-κB pathways included genes that were positive regulators of cell cycle (CCDN1, CCNG1), genes that can inhibit cell cycle progression and downstream of the immune suppressive TGF-beta signaling pathway (TGIF1, PPP1R15A). These results suggest that controlling the balance between pro- and anti-inflammatory pathways trigger the development of protective vaccine responses.

Discussion

Transcriptional profiling of PBMCs from RV144 vaccinees stimulated with Env peptides was characterized by the upregulated expression of genes associated with antigen presentation, maturation of MHC class II complex and genes endowed with antiviral functions; these genes and pathways were induced only in vaccinees that remained HIV-1 negative at their last follow up (control) compared to vaccinees that acquired HIV-1 (cases). Induction of these pathways was not observed in placebo recipients or in the absence of stimulation with Env peptides indicating that the vaccine triggered those pathways. Ex vivo experiments showed that IRF7 (a key regulator of an antiviral innate immune response) associated with low risk in vaccinees, was expressed by T cells. Expression of IRF7, and more importantly genes with an antiviral activity that are regulated by IRF7 (XCL1, IFITM3, ISG15, MX2, TRIM26), in T cells can render these cells less susceptible to HIV-1 infection[23–28]. These results provide a mechanism whereby CD4+ T cells from subjects immunized with the RV144 vaccine will mount an Env specific type II interferon (interferon γ) response that could, in turn, trigger the expression of these antiviral genes (ex. XCL1, IFITM3, ISG15, MX2, TRIM26), thereby rendering these cells and bystander cells resistant to HIV-1 infection. Attenuated viruses which are all known to be very efficacious vaccines (YF17D, measles, smallpox) are known to trigger these innate antiviral immune response pathways[29,30].

Integration of gene-expression, antibody responses and ICS datasets revealed that the IFNγ pathway correlated with IgG antibodies binding to V1/V2 (explaining 5% of the variance of the IFNγ pathway) and with cytokine production by Env-specific CD4+ T cells (explaining 8% of the variance of the IFNγ pathway). Herein we show that different signatures can independently predict immunogenicity of the vaccine (STAT1 regulated genes) and low risk of acquisition (IRF7 target genes) confirming the non-redundant roles of IgG antibodies binding to V1/V2 and T cells polyfunctionality. Indeed, antiviral genes regulated by IRF7 were significantly associated with IgG

response while STAT1-target genes implicated in antigen presentation via MHC class I and class II were specifically correlated with the PFS. The IRF7 signature was enriched in IFNα-stimulated genes (24/38 genes) compared to the STAT1 signature (34/97 genes, X2 test: p=0.00555) suggesting that different stimuli triggered the IRF7 and the STAT1 gene signatures; type I interferons stimulated the IRF7 gene signature while type II interferons could specifically induce the STAT1 gene signature. In support of our observations, knock out of IRF7 in murine models did not abrogate CD4+ T cell responses[31] while it led to enhanced viral replication. In contrast, knock out of the STAT1 gene in mice did abrogate antigen-specific CD4+ T cell responses. In line with these reports, STAT1 regulated genes were specifically correlated to the heightened polyfunctional CD4+ T cell response linked to low risk in RV144 vaccinees and not to a general "non-protective" T cell response (ex. monofunctional CD4+ T cells that can secrete only one cytokine). The IRF7 signature was specifically correlated to an IgG response linked to low risk in RV144 vaccinees and was not correlated to other "non-protective" antibody markers such as IgG response to non-V1/V2 epitopes. Further evidence suggests that both IRF7 and STAT1 activation are required for the development of the protective immune response also supported by the fact that knock-out of STAT1 and induction of IRF7 is lethal in virally infected animal models as it leads to an uncontrolled cytokine storm[32]. Our data shows that the antiviral innate immune response and the HIV specific CD4+ T cell response are two independent correlates of low risk of HIV-1 acquisition.

The IRF7 antiviral transcriptomic program and the IgG antibodies binding to V1/V2 were heightened in vaccinees that express DPB1*13 allele. This association may result from a T cell response to specific T cell epitopes present in Env antigen and restricted by MHC class II of DPB1*13. The protective role of DPB1*13 could also be attributed to the poor IgA responses previously shown to occur in DPB1*13+ subjects[33]. IgA antibodies compete with IgG antibodies for binding to HIV-1 Env and thus abrogate ADCC in vaccinees[6]. IgG responses have been shown to mediate antibody-dependent cellular cytotoxicity ADCC[6]. Type I interferon (including IRF7) induces the expression of FCGR that triggers ADCC; the latter has been suggested to be a mechanism of RV144 vaccine-mediated protection[2]. Binding of IgG antibodies (specific for V1/V2) to FCGR will trigger ADCC and will induce IRF7 and type I interferons[34]. Gene expression analysis confirmed the potential role of ADCC in response to the RV144 vaccine since several genes implicated in ADCC are induced by the vaccine; they include markers of NK cells (CD48, KIR2DL1, KIR3DL1, KIR3DL2, NKG2C, FCGR3A) as well as effector molecules of ADCC (NCR1, NCR2, FAS, GZMB, PRF1, TNFSF10); of note only FCGR3A, TNFSF10 were associated with a reduced risk of HIV-1 acquisition. In addition, IRF7 induced genes important for cell trafficking

(ex. CXCL10 and ITGB7) may promote migration of effector cells to the mucosal sites where HIV-1 infection will occur. Of note, ITGB7 was confirmed in an animal model as a correlate of protection of an RV144-like vaccine[22].

Integration of all these datasets suggests a model whereby ALVAC vector, used to prime RV144 vaccinees and know to be able to infect dendritic cells, trigger pDCs to produce type I interferons as ALVAC interact the innate sensor STING[35]. Type I interferon will upregulate antiviral genes in bystanders cells including Env-specific activated CD4 T cells. These cells now can provide help to B cells to produce IgG and to other CD4 T cells. This model can be deduced only by integrating multiple OMICs and will be validated in subsequent clinical trials.

Our analysis indicated that genes downstream of the proinflammatory transcription factor NF-κB as well as genes downstream of mTORC1 that are required for HIV-1 life-cycle (CXCR4, ETF1[36]) were associated with the risk of HIV-1 acquisition; those associations were observed in the placebo and vaccine arms of the study and thus were independent of vaccination. Hence, participants expressing the mTORC1 and NF-κB signatures did not benefit from the RV144 vaccine. Moreover, those results are not supportive of vaccine-related enhancement of HIV-1 acquisition in RV144 vaccinees.

Both the integrative analysis between gene-expression and frequency of cells measured by FCM as well as in the deconvolution of the gene-expression confirmed monocytes as the cellular subset that expressed high levels of NF-κB and its downstream targets. Activated monocytes that express high levels of NF-κB genes would trigger the production of proinflammatory cytokines/chemokines such as IL2 and IL3 that can enhance the survival of activated T cells thereby providing HIV-1 with potential target cells to infect and lead to HIV-1 acquisition susceptibility. Moreover, the association with increased acquisition of genes that regulate TGFβ signaling or genes that are downstream of the anti-proliferative cytokine TGFβ could result from the immunosuppressive activity of TGFβ on the development of protective HIV-1 specific CD4+ or CD8+ T cell responses[36]. TGFβ is also known to regulate IgA class switch; IgA against Env was shown to be associated with a higher risk of HIV-1 acquisition2.

In addition, host proteins required for HIV-1 life cycle downstream of mTORC1 signaling pathways (CXCR4, ETF1[37]) were elevated in both placebo recipients and vaccinees. Rapamycin, that downregulates mTORC1 pathway, could improve the effectiveness of the RV144 vaccine as it did enhance the response to flu vaccination in a cohort of elderly subjects[38].

Our study highlights the important contributions of unbiased system biology approaches in defining mechanisms underlying vaccine-mediated protection. Similar approaches could lead to the identification of host-related markers associated with vaccine-conferred protection by

investigating pre-vaccination gene-expression profiling of participants receiving the RV144 vaccine. Moreover, the contribution of the mucosal immune response needs to be assessed since we have previously shown using similar unbiased approaches that integrate mucosal and systemic immune responses can inform us on mechanisms leading to vaccine-conferred protection[22]. Differential gene-expression could also result from polymorphisms in the coding and regulatory regions of those genes, alternative splicing, chromatin accessibility or non-coding RNA expression[39]. The platform used in this study, microarrays, does not provide us with the ability to investigate all those regulatory elements. Further studies using such high-dimensional data types would further complement the mechanistic insights identified in this study that lead to the vaccine-conferred protection by the RV144 vaccine. Finally, follow-up studies using in vitro experiments and animal models are required to confirm the functional/mechanistic contribution of the aforementioned pathways.

In conclusion, we have shown that the establishment of a productive HIV-1 infection in participants depends on the balance between innate antiviral and proinflammatory responses. The proinflammatory responses mediated by mTORC1 and NF-κB signaling can lead to the activation and proliferation of HIV-1 target cells. Immune modulators that boost innate antiviral responses and suppress pro-inflammatory detrimental immune responses may decrease the risk of HIV-1 infection or replication. The HVTN 702 trial, the follow-up efficacy trial of a pox-protein vaccine regimen initiated in South Africa in Q4 2016, will allow us to evaluate and confirm the mechanisms identified in this study as being associated with the RV144 correlates.

Methods

Study Design

Fifty participants of the RV144 clinical trial were part of the transcriptomic pilot study, randomly sampled within each (gender × treatment arm) strata (50% for each gender, 80% vaccine recipients) among subjects completing follow-up HIV negative. This ensured that baseline characteristics of subjects enrolled in the transcriptomic pilot study were similar to the original RV144 cohort (16,402 participants) except for greater proportions of vaccinees and participants that completed the trial protocol in the transcriptomic pilot study cohort (**Supplementary Table A2.1**). Separately, 183 participants of the RV144 clinical trial were selected for the transcriptomic case/control study. Baseline characteristics of subjects enrolled in the transcriptomic case/control study were similar to the original RV144 cohort (16,402 participants) except for greater proportions of vaccinees, participants that completed the trial protocol and participants that acquired HIV-1 in the transcriptomic pilot study cohort (**Supplementary Table A2.3**). No

imbalance was observed in term of clinicopathological characteristics between the vaccinees and placebo recipients included in the transcriptomic pilot study cohort (**Supplementary Table A2.2**) or the transcriptomic case/control study cohort (**Supplementary Table A2.4**). The RV144 trial protocol was reviewed by the ethics committees of the Ministry of Public Health, the Royal Thai Army, Mahidol University, and the Human Subjects Research Review Board of the U.S. Army Medical Research and Materiel Command. All participants gave their informed consent. Written informed consent was obtained from all volunteers.

Vaccine

ALVAC-HIV (vCP1521) is a recombinant canarypox genetically engineered to express HIV-1 gag and pro (subtype B, LAI strain) and CRF01_AE (subtype E) HIV-1 gp120 (92TH023) linked to the transmembrane 3 anchoring portion of gp41 (LAI). AIDSVAX B/E is an HIV gp120 envelope glycoprotein vaccine containing a subtype E envelope from the HIV-1 strain A244 (CM244) and a subtype B envelope from the HIV-1 MN. The envelope glycoproteins, 300 µg of each, are co-formulated with 600 µg of alum adjuvant. ALVAC-HIV placebo consisted of virus stabilizer and freeze-drying medium in 1 ml sodium chloride. AIDSVAX placebo was 600 µg alum adjuvant.

Primary endpoint

HIV infection was monitored every six months, from month 6 to month 36 after the initial immunization. HIV infection established from repeated positive results on enzyme immunoassay and Western blots, with two confirmatory HIV nucleic acid tests: the Amplicor HIV Monitor (version 1.5) assay (Roche) in Thailand and the Procleix HIV discriminatory assay (Novartis) in the United States. Correlates analyses defined the primary endpoint as the diagnosis of HIV-1 infection any time after the month 6 visit post initial immunization.

Transcriptomic analysis

PBMC samples taken pre-immunization and two weeks (window: -2 to +14 days) after the last immunization, were either stimulated in vitro for 15 hours with Env peptides or with the vehicle (dimethyl sulfoxide). The Env peptides consisted of 15 amino acids spanning the Env 92TH023 sequence expressed in vCP1521 and overlapping by 11 amino acids (Biosynthesis, Lewisville, TX) were combined into one pool at a final concentration of 1 µg/ml per peptide, and used to stimulate 106 PBMC ex vivo; as further detailed in Haynes et al.[2]. The transcriptomic profile of the stimulated PBMC was assessed using Illumina Human HT-12 beadchips. RNA was isolated using the Rneasy micro kit (Qiagen) and the quantity and quality of the RNA were confirmed using

a NanoDrop 2000c (Thermo Fisher Scientific) and an Experion Electrophoresis System. Samples (50 ng) were amplified using Illumina TotalPrep RNA amplification kits (Ambion). The microarray analysis was conducted using 750 ng of biotinylated complementary RNA hybridized to Human HT-12 version 4 beadchips (Illumina) at 58 °C for 20 h. The chips were scanned using Illumina's iSCAN and quantified using Genome Studio (Illumina).

Raw beadchips intensities were quantile-normalized and log2-transformed. The LIMMA framework was used to fit linear regression model with the log2 gene-expression as dependent variable and the groups of interest (vaccination group, HIV-1 infection status or antibody response) as independent variables in order to identify genes differentially expressed between vaccination group (vaccine or placebo), HIV-1 infection status (control versus case) or genes correlated to antibody-response (IgG antibodies binding to V1/V2). A moderated t-test was used to assess the statistical significance of the association between gene-expression and the groups of interest. Benjamini & Hochberg correction was applied to adjust for multiple testing.

Genecards[40], Reactome[41], GeneRIF[42] and Literome[43] were used to annotate the function of genes. The HIV-1 host factors were obtained from the NCBI HIV-1 interaction database[44].

GSEA was used to identify pathways modulated after Env stimulation and/or associated with HIV-1 acquisition[45]. In GSEA, the most varying probe across samples was used as representative of redundant probes annotated to the same gene. The gene list ranked by LIMMA moderated t-statistic were used as input for the GSEA analysis. The pathways (i.e. genesets) database used for all GSEA analysis were the Molecular Signatures Database (version 5.1) hallmark genesets[46], canonical pathways (module C2.CP), transcription factor targets (module C3.TFT) and blood cells markers[47]. The GSEA Java desktop program was downloaded from 'http://www.broadinstitute.org/gsea/index.jsp [http://www.broadinstitute.org/gsea/index.jsp]' and the default parameters of GSEA preranked module (number of permutations: 1000; enrichment statistic: weighted; seed for permutation: 101, 15 ≤ gene set size ≤ 500) were applied for analyses. Putative transcription factor binding sites were identified in the regulatory region of genes associated with HIV-1 acquisition using HOMER version 4.9 using default parameters[48].

Sample-level enrichment analysis (SLEA) was used following GSEA analysis to investigate the enrichment of pathways in the different samples[49]. Briefly, the expression of all the genes in a specific pathway was averaged across samples and compared to the average expression of 1000 randomly generated genesets of the same size. The resulting z-score is then used to reflect the overall perturbation of a pathway in a sample.

Intracellular cytokine staining

PBMC were plated in a 96-well plate (106 cells per well). PBMC stimulation was performed in 10% FBS/RPMI media in the presence of 1µg/ml anti-CD28 and anti-CD49d and Brefeldin A (BD Biosciences, San Diego, CA) and stimulated with HIV peptides (New England Peptide, Gardner, MA) of 15-mer overlapping by 11 amino acids representing HIV subtype E-Env (TH023; 162 peptides) and HIV subtype B-Gag (LAI; 120 peptides). PBMC supplemented with DMSO was used as a negative control. After 6 hours of stimulation at 37°C, 5% CO2 EDTA (20mM, Sigma) was added and incubated for 15 min. Subsequently, PBMC were fixed and permeabilized using FACS Lysing Solution and FACS Permeabilizing Solution 2 (BD) according to the manufacturer's instructions. The following antibodies were added for 60 min at room temperature in the dark: CD4–fluorescein isothiocyanate (FITC), CD3-allophycocyanin (APC), IFNγ-phycoerythrin (PE), Il-2-phycoerythrin (PE) and CD8-PerCP-Cy5.5 (all BD Biosciences). PBMC were washed and fixed with 1% paraformaldehyde. The analysis was performed using a FACSCalibur flow cytometer (BD Immunocytometry Systems). Intracellular cytokine staining (ICS) data was provided to us by the trial investigators in either an ICS positivity score (call) format and aggregate value format[2]. ICS analytes included CD154, IFNγ, IL-4, IL-2, IL-17α and TNFα.

Multiplex cytokine bead array

Cryopreserved PBMC were thawed and rested overnight. 5 × 105 PBMC each were stimulated with Env 92TH023 peptides at 37°C and 0.5% DMSO served as negative control. After 48 hours, supernatants were harvested and frozen at -80°C until analysis. Analyte concentrations were measured using a MILLIPLEX MAP Human Cytokine/Chemokine – Custom-12-Plex kit (Millipore, Billerica, MA) following instructions provided by the manufacturer. All samples were acquired on a Luminex 200 instrument (Millipore) and data analyses were performed using MasterPlex software. Multiplex cytokine (Luminex) data was provided to us by the trial investigators as normalized values (mean of 0, a standard deviation of 1)[2]. Luminex analytes included GM-CSF, IFNγ, IL-2, IL-3, IL-4, IL-5, IL-9, IL-10, IL-13, MIP1β, TNFα and TNFβ.

Flow cytometry phenotyping panel

Samples were stained for extracellular markers for FACS analysis following the manufacturer's recommendations (Becton Dickinson Cytofix/Cytoperm Kit). Samples were surface stained with the following antibodies to distinguish cell subsets: CD11c (Pe-Cy5), CD14 (FITC), HLA-DR (Allophycocyanin-Cy7), CD3 (Qdot 800), CD19 (Qdot 605), and CD123 (Brilliant Violet). They were acquired using a Becton Dickinson LSR II flow cytometer and analyzed using FlowJo

(TreeStar). Antibodies were from Becton Dickinson unless otherwise stated. The flow cytometry (FCM) phenotyping data was provided to us by the trial investigators in raw cell counts.

Ex vivo stimulation and in vitro HIV-1 infectability assays

Cryopreserved or fresh isolated PBMC samples from healthy donors (n=10) were thaw and enriched for CD4 memory T cells by negative selection, according to manufacturer's protocol, with the EasySep Human CD4 memory T cell Enrichment Kit (StemCell Technologies, 10157). Isolated CD4 memory T cells were cultured at concentration of 2×106 cells/mL with logarithmic increasing concentrations of IFNα [0.02, 200 ng/mL], IFNβ [0.002, 200 ng/mL], IFNγ [0.005, 50 ng/mL] or left unstimulated. The efficacy of the cytokines to induce IRF7 and STAT1 phosphorylation was evaluated in 100,000 cells from 5 donors at the lowest and the highest concentration of the respective cytokines by flow cytometry. The remaining cells were kept in culture for 18 hours. Following incubation, HIV-1 infection through spinoculation with 89.6 viral supernatant (NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: p89.6 from Ronald G. Collman, MD) was performed at 200 ng/mL p24/million CD4+ T cells in the presence of 4 ug/mL polybrene (Sigma, H9268), at 2500 rpm for approximately 2.5 hours at 30○C. After spinoculation, viral supernatants were removed and infected cells were cultured at a concentration of 2×106 cells/mL in cRPMI supplemented with 30 IU/mL IL-2 (R&D Systems, 202-IL), 5 μM saquinavir (NIH Aids Reagent Program, 4658) and the respective cytokines at 37○C, 5% CO2 for three days. On day 4, HIV-1 p24 levels were evaluated by flow cytometry on CD4 negative cells. The data is represented as the frequency of infected p24+ cells, HIV-1 per cell level (MFI: median of fluorescence intensity) per condition normalized by unstimulated cells. A paired Wilcoxon-rank sum was used to compared the frequencies/intensities of cells after interferon stimulation to the unstimulated condition.

Integrative analysis

A projection-based approach implemented in the R package mixOmics was adopted to assess the correlations between gene-expression and other data types (ICS, Luminex and FCM). For each pair of data type, a sparse-least square regression was used to project the first element of the pair onto the second element of the pair. Once the two data types are projected on the same scale, the Pearson correlation between the features of the two data types was calculated. To assess the probability of obtaining a Pearson correlation equal to or greater than the one observed, we derived a p-value based on the distribution of the Pearson correlations calculated

between all pair of features of the two data types (i.e. the statistical universe). Pearson correlations corresponding to a p-value cutoff of 0.05 were considered significant.

Deconvolution of blood transcriptome

The function lsfit of the R package CellMix[19] was used to model the PBMC gene expression measures across samples as the function of contributions of immune subset-specific gene expression weighted by the corresponding cell frequencies of those subsets measured by FCM. To verify that the deconvolution was successful, we investigated the gene-expression of the cell surface markers used for cell-sorting. We observed a concordant expression of several protein markers specifically expressed at the surface of those subsets (**Supplementary Fig. A2.5a**).

To obtain immune subset-specific gene expression estimates for RV144 controls and cases, we applied linear regression separately to each group's gene expression. The linear regression coefficient estimates are taken as surrogates for estimated cell type-specific average gene expression while a difference between these cell type-specific estimates was used as the level of gene expression change between the controls and cases in a given cell type.

HIV-1 acquisition classifiers

Logistic regression models were built using the function glm of the R package stats. Gender and behavior risk were included in all regressions as independent variables. HIV-1 acquisition, the dependent variable was coded as a binary variable while PFS, gene-expression pathways, IgA and IgG titers were coded as continuous variables. A Wald test was used to test if the coefficients of regression are statistically different zero (i.e. null hypothesis being no association between a marker of interest and HIV-1 acquisition). The accuracy of each regression model was estimated by 10-fold cross-validation. Receiver operating characteristic curves and Delong's test was used to compare the accuracy of two regression models.

Statistical analyses

Student t test was used to test for the significance of pathway and antibody response while a non-parametric test, the Wilcoxon rank-sum test, was used to test for difference in pathway expression between HIV-1 cases and controls. The Benjamini & Hochberg correction was used to adjust for multiple testing. for all gene expression analysis. A 5% cutoff on the probability of false-positive (i.e. p-value) was used as a statistical stringency for all analyses presented in this work.

Code availability

All the source code used to generate the figures part of this manuscript is available at https://github.com/sekalylab/rv144. The authors declare that all other data supporting the findings of this study are available from the authors upon request.

Data availability

The microarray data have been submitted to the National Center for Biotechnology Information Gene Expression Omnibus ('https://www.ncbi.nlm.nih.gov/geo [https://www.ncbi.nlm.nih.gov/geo]') under accession number GEO: 'GSE103740 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103740]'.

References

1. Rerks-Ngarm, S. et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med 361, 2209–2220 (2009).
2. Haynes, B. F. et al. Immune-correlates analysis of an HIV-1 vaccine efficacy trial. N Engl J Med 366, 1275–1286 (2012).
3. Lin, L. et al. COMPASS identifies T-cell subsets correlated with clinical outcomes. Nat Biotechnol 33, 610–616 (2015).
4. Prentice, H. A. et al. HLA class II genes modulate vaccine-induced antibody responses to affect HIV-1 acquisition. Sci Transl Med 7, 296ra112 (2015).
5. Plotkin, S. A. & Gilbert, P. B. Nomenclature for immune correlates of protection after vaccination. Clin. Infect. Dis. 54, 1615–1617 (2012).
6. Tomaras, G. D. et al. Vaccine-induced plasma IgA specific for the C1 region of the HIV-1 envelope blocks binding and effector function of IgG. Proc Natl Acad Sci U S A 110, 9019–9024 (2013).
7. Harenberg, A., Guillaume, F., Ryan, E. J., Burdin, N. & Spada, F. Gene profiling analysis of ALVAC infected human monocyte derived dendritic cells. Vaccine 26, 5004–5013 (2008).
8. Lebedeva, T., Dustin, M. L. & Sykulev, Y. ICAM-1 co-stimulates target cells to facilitate antigen presentation. Curr Opin Immunol 17, 251–258 (2005).
9. Schoggins, J. W. & Rice, C. M. Interferon-stimulated genes and their antiviral effector functions. Curr Opin Virol 1, 519–525 (2011).
10. Li, J. et al. Inhibition of hepatitis B virus replication by MyD88 involves accelerated degradation of pregenomic RNA and nuclear retention of pre-S/S RNAs. J Virol 84, 6387–6399 (2010).

11. Schoggins, J. W. et al. Pan-viral specificity of IFN-induced genes reveals new roles for cGAS in innate immunity. Nature 505, 691–695 (2014).
12. Bosco, A., Wiehler, S. & Proud, D. Interferon regulatory factor 7 regulates airway epithelial cell responses to human rhinovirus infection. BMC Genomics 17, 76 (2016).
13. Bidwell, B. N. et al. Silencing of Irf7 pathways in breast cancer cells promotes bone metastasis through immune escape. Nat Med 18, 1224–1231 (2012).
14. Kim, T. H. & Zhou, H. Functional Analysis of Chicken IRF7 in Response to dsRNA Analog Poly(I:C) by Integrating Overexpression and Knockdown. PLoS One 10, e0133450 (2015).
15. Yohn, N. L., Bingaman, C. N., DuMont, A. L. & Yoo, L. I. Phosphatidylinositol 3'-kinase, mTOR, and glycogen synthase kinase-3beta mediated regulation of p21 in human urothelial carcinoma cells. BMC Urol 11, 19 (2011).
16. Yu, J. et al. IGF-1 induces hypoxia-inducible factor 1α-mediated GLUT3 expression through PI3K/Akt/mTOR dependent pathways in PC12 cells. Brain Res. 1430, 18–24 (2012).
17. Zheng, Y. et al. A role for mammalian target of rapamycin in regulating T cell activation versus anergy. J Immunol 178, 2163–2170 (2007).
18. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Comput. Biol. 13, e1005752 (2017).
19. Gaujoux, R. & Seoighe, C. CellMix: A comprehensive toolbox for gene expression deconvolution. Bioinformatics 29, 2211–2212 (2013).
20. Lieberman, J. The ABCs of granule-mediated cytotoxicity: New weapons in the arsenal. Nature Reviews Immunology 3, 361–370 (2003).
21. Smyth, M. J. et al. Activation of NK cell cytotoxicity. Molecular Immunology 42, 501–510 (2005).
22. Vaccari, M. et al. Adjuvant-dependent innate and adaptive immune signatures of risk of SIV acquisition. Nat Med (2016).
23. Dicks, M. D. et al. Oligomerization Requirements for MX2-Mediated Suppression of HIV-1 Infection. J Virol 90, 22–32 (2015).
24. Durfee, L. A., Lyon, N., Seo, K. & Huibregtse, J. M. The ISG15 conjugation system broadly targets newly synthesized proteins: implications for the antiviral function of ISG15. Mol Cell 38, 722–732 (2010).
25. Guzzo, C., Fox, J. C., Miao, H., Volkman, B. F. & Lusso, P. Structural Determinants for the Selective Anti-HIV-1 Activity of the All-beta Alternative Conformer of XCL1. J Virol 89, 9061–9067 (2015).

26. Okumura, A., Lu, G., Pitha-Rowe, I. & Pitha, P. M. Innate antiviral response targets HIV-1 release by the induction of ubiquitin-like protein ISG15. Proc Natl Acad Sci U S A 103, 1440–1445 (2006).
27. Uchil, P. D., Quinlan, B. D., Chan, W. T., Luna, J. M. & Mothes, W. TRIM E3 ligases interfere with early and late stages of the retroviral life cycle. PLoS Pathog 4, e16 (2008).
28. Yu, J. et al. IFITM Proteins Restrict HIV-1 Infection by Antagonizing the Envelope Glycoprotein. Cell Rep 13, 145–156 (2015).
29. Gaucher, D. et al. Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. J Exp Med 205, 3119–3131 (2008).
30. Pulendran, B. & Ahmed, R. Immunological mechanisms of vaccination. Nat Immunol 12, 509–517 (2011).
31. Zhou, S., Cerny, A. M., Fitzgerald, K. A., Kurt-Jones, E. A. & Finberg, R. W. Role of interferon regulatory factor 7 in T cell responses during acute lymphocytic choriomeningitis virus infection. J Virol 86, 11254–11265 (2012).
32. Li, W., Hofer, M. J., Jung, S. R., Lim, S. L. & Campbell, I. L. IRF7-dependent type I interferon production induces lethal immune-mediated disease in STAT1 knockout mice infected with lymphocytic choriomeningitis virus. J Virol 88, 7578–7588 (2014).
33. MacHulla, H. K. et al. HLA-A, B, Cw and DRB1, DRB3/4/5, DQB1, DPB1 frequencies in German immunoglobulin A-deficient individuals. Scand J Immunol 52, 207–211 (2000).
34. McEwan, W. A. et al. Intracellular antibody-bound pathogens stimulate immune signaling via the Fc receptor TRIM21. Nat Immunol 14, 327–336 (2013).
35. Liu, F. et al. Priming and Activation of Inflammasome by Canarypox Virus Vector ALVAC via the cGAS/IFI16–STING–Type I IFN Pathway and AIM2 Sensor. J. Immunol. (2017).
36. Barouch, D. H. et al. Rapid Inflammasome Activation following Mucosal SIV Infection of Rhesus Monkeys. Cell 165, 656–667 (2016).
37. Besnard, E. et al. The mTOR Complex Controls HIV Latency. Cell Host Microbe 20, 785–797 (2016).
38. Mannick, J. B. et al. mTOR inhibition improves immune function in the elderly. Sci Transl Med 6, 268ra179 (2014).
39. Zak, D. E., Tam, V. C. & Aderem, A. Systems-level analysis of innate immunity. Annu Rev Immunol 32, 547–577 (2014).
40. Stelzer, G. et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinforma. 54, 1 30 1-1 30 33 (2016).

41. Fabregat, A. et al. The Reactome pathway Knowledgebase. Nucleic Acids Res 44, D481-7 (2016).
42. Lu, Z., Cohen, K. B. & Hunter, L. GeneRIF quality assurance as summary revision. Pac Symp Biocomput 269–280 (2007).
43. Poon, H., Quirk, C., DeZiel, C. & Heckerman, D. Literome: PubMed-scale genomic knowledge base in the cloud. Bioinformatics 30, 2840–2842 (2014).
44. Ako-Adjei, D. et al. HIV-1, human interaction database: current status and new features. Nucleic Acids Res 43, D566-70 (2015).
45. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102, 15545–15550 (2005).
46. Liberzon, A. et al. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 1, 417–425 (2015).
47. Nakaya, H. I. et al. Systems biology of vaccination for seasonal influenza in humans. Nat Immunol 12, 786–795 (2011).
48. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576–589 (2010).
49. Gundem, G. & Lopez-Bigas, N. Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. Genome Med 4, 28 (2012).

Acknowledgments

We would like to thanks Petra Stafova for performing the gene-expression arrays experiments; Nicole Frahm and Stephen De Rosa for providing us with the intracellular and Luminex data; Barton Haynes for comments on the protocol and the paper. This work made use of the High-Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. Grants from the Bill and Melinda Gates Foundation (OPP1032325 and OPP1147555) supported this work. S.F. received a travel fellowship from the Bill and Melinda Gates Foundation (OPP1084285).

Author contributions

J.K., P.P., S.N., S.R.-N., J.H.K., N.L.M. and M.J.M. were involved in the conceptualization and oversight of the study; R.T. and G.D.T provided the experimental data, S.F. was involved in conception of the methodology; S.F. was involved in software development; S.F., A.T. and F.L.

were involved in formal analysis; F.B.T.P.L. and S.P.R. performed the ex vivo experiments; S.F., R.G. and R.-P.S. prepared the original draft; R.A.K., M.C., P.B.G., N.L.M. provided critical inputs.

Competing interests

The authors declare no competing interests.

Figures

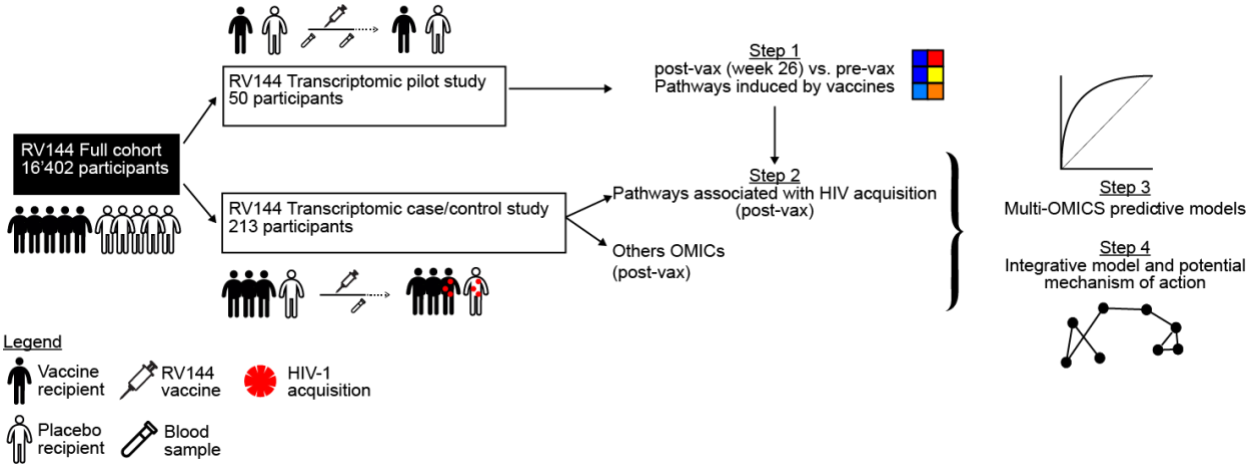


Figure A2.1. Study overview

Four analysis steps were used to identify transcriptomic markers of risk of HIV-1 acquisition among RV144 vaccinees. A first transcriptomic dataset of blood collected from 40 HIV-1 negative vaccinees and 10 HIV-1 negative placebo recipients pre-vaccination and two weeks after vaccination was used to identify pathways modulated by the RV144 vaccine (step 1). A second independent transcriptomic dataset of blood collected from 183 case-control vaccinees (including 31 infected participants) and 30 placebos (including 17 infected participants), two weeks after vaccination was used to identify pathways associated with HIV-1 acquisition. Logistic regression was used to build a multi-OMICS classifier of HIV-1 acquisition among RV144 vaccinees (step 3) and a projection-based integrative analysis was used to associate the different OMICS to identify mechanistic mediators of vaccine response (step 4).

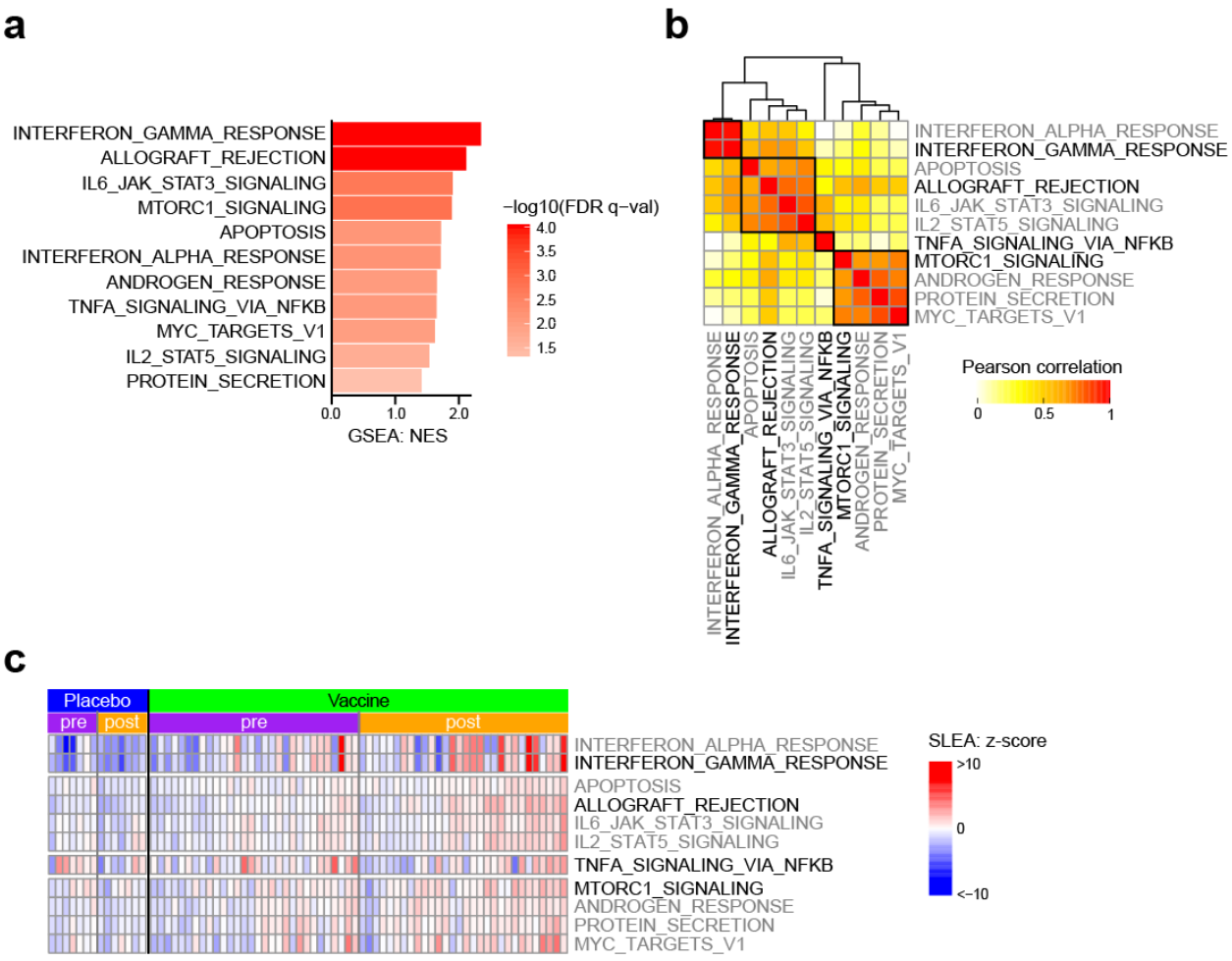


Figure A2.2. IFNγ response is strongly induced by the in RV144 vaccine

a Barplot presenting the pathways modulated by the RV144 vaccine two weeks after the last immunization compared to pre-immunization. A normalized enrichment score (NES) greater than 0 corresponds to a pathway for which member genes are up-regulated in vaccinees. Eleven pathways were significantly modulated after immunization in RV144 vaccinees but not in placebo recipients (DB: Hallmark, GSEA: FDR ≤ 5%). **b** Sample-enrichment analysis (SLEA) of those 11 pathways followed by clustering revealed that those pathways could be separated into 4 groups of highly correlated pathways (indicated by the black boxes). The representative pathway of each of the four groups (the most significantly enriched) is indicated in black while the remaining pathways are labeled in grey. **c** Heatmap presenting the SLEA z-score of each of the 11 pathways among the 40 vaccinees and 10 placebo recipients included in the transcriptomic pilot study at both timepoints investigated (pre: pre-vaccination, post: 2 weeks after the last immunization). An SLEA z-score greater than 0 corresponds to a pathway for which member genes are up-regulated

while an SLEA z-score inferior to 0 corresponds to a pathway with genes down-regulated in that sample.

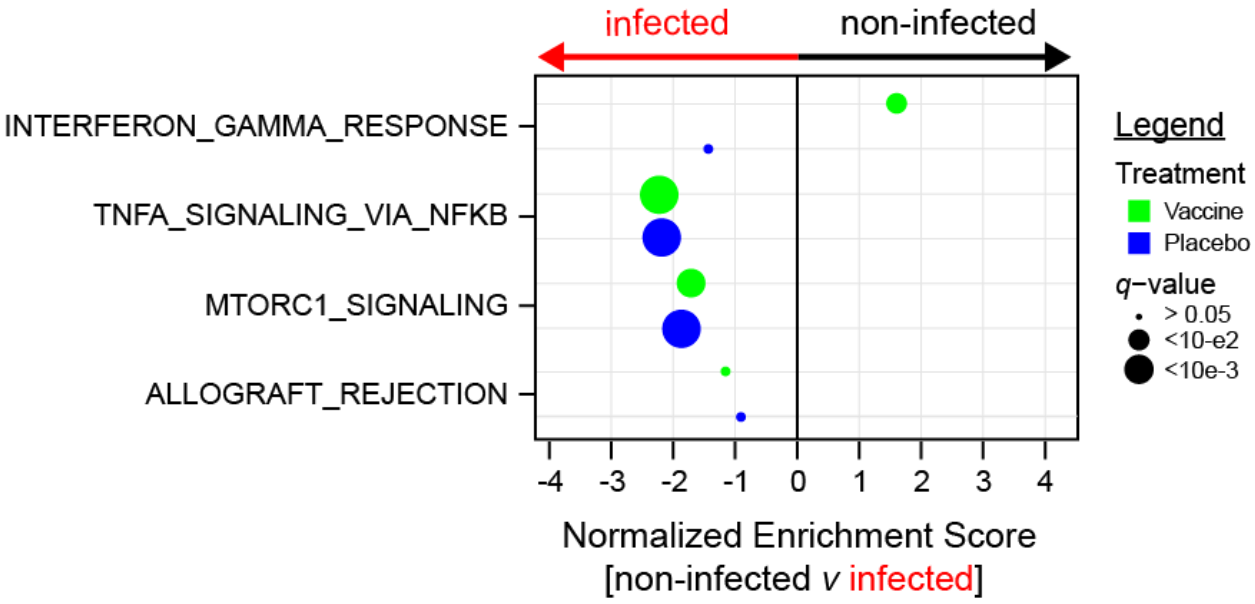


Figure A2.3. IFNγ response associated with the reduction of the risk of HIV-1 infection in vaccinees

Dotplot presenting the association between the pathways induced by the RV144 vaccine and HIV-1 infection status, separately for vaccinees and placebo recipients. Gene-expression of 183 vaccine recipients, 30 cases and 153 controls, and 30 placebo recipients, of which 17 were infected, were used for this analysis. GSEA was performed and identified one pathway associated with the reduction of the risk of HIV-1 acquisition in vaccinees and the two pathways associated with a higher risk of HIV-1 acquisition both in vaccinees and placebo recipients. The normalized enrichment scores (NES) of those pathways are presented on the plot. An NES greater than 0 suggests that participants with higher expression of the genes in that pathway are less likely to be infected by HIV-1 while an NES below 0 corresponds to participants with higher expression of the genes in that pathway and more likely to acquire HIV-1. The size of the dots is proportional to the false-discovery rate (q-value) of the enrichment.

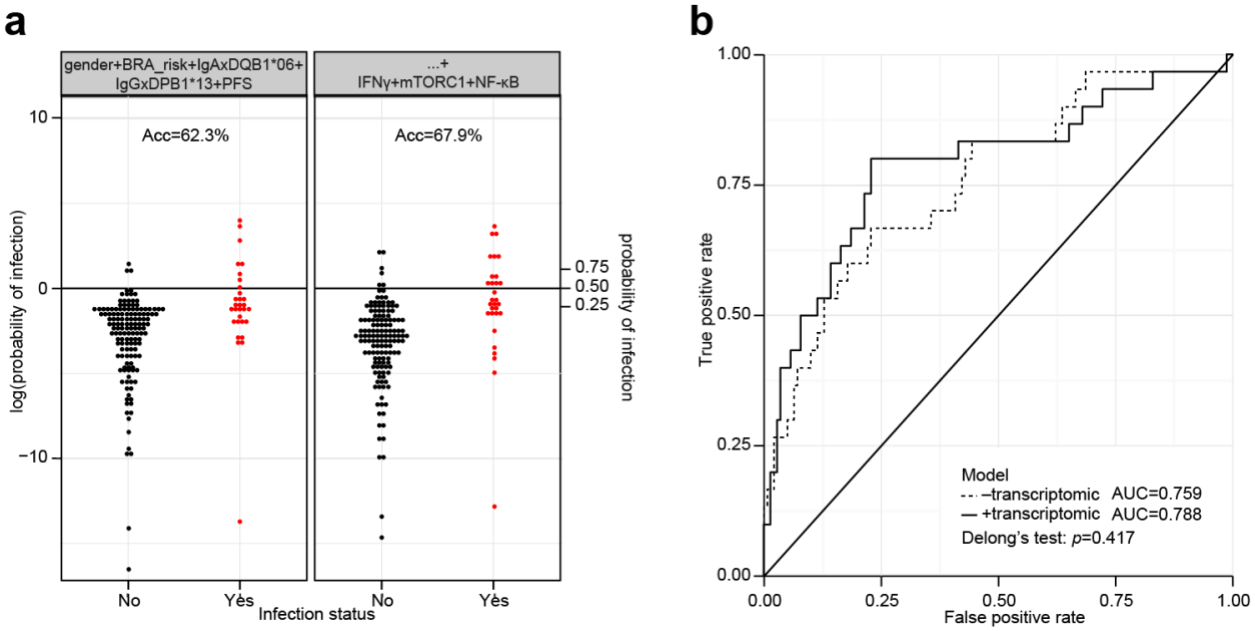


Figure A2.4. Prediction of the response does not improve by adding transcriptomic data

a Logistic regression models were built to predict HIV infection status of RV144 vaccinees (142 vaccinees that were HIV negative at last follow-up and 30 vaccinees that acquired HIV). The accuracy of each model was assessed by 10-fold cross-validation. The first model (left panel) included IgA against V2, IgG against V1/V2 and the polyfunctionality score (PFS) previously identified as markers of response to RV144 vaccine. The second model (right panel) included the same markers as the first model but with the addition of the three pathways associated with HIV status in the transcriptomic analysis (IFNγ response, MTORC1 signaling and TNFα signaling via NF-κB). The balanced accuracy (Acc) of each model is given on the plot. **b** Corresponding ROC curves based 10-fold cross-validation for the model without the three genesets and the model with the three genesets identified in the transcriptomic analysis.

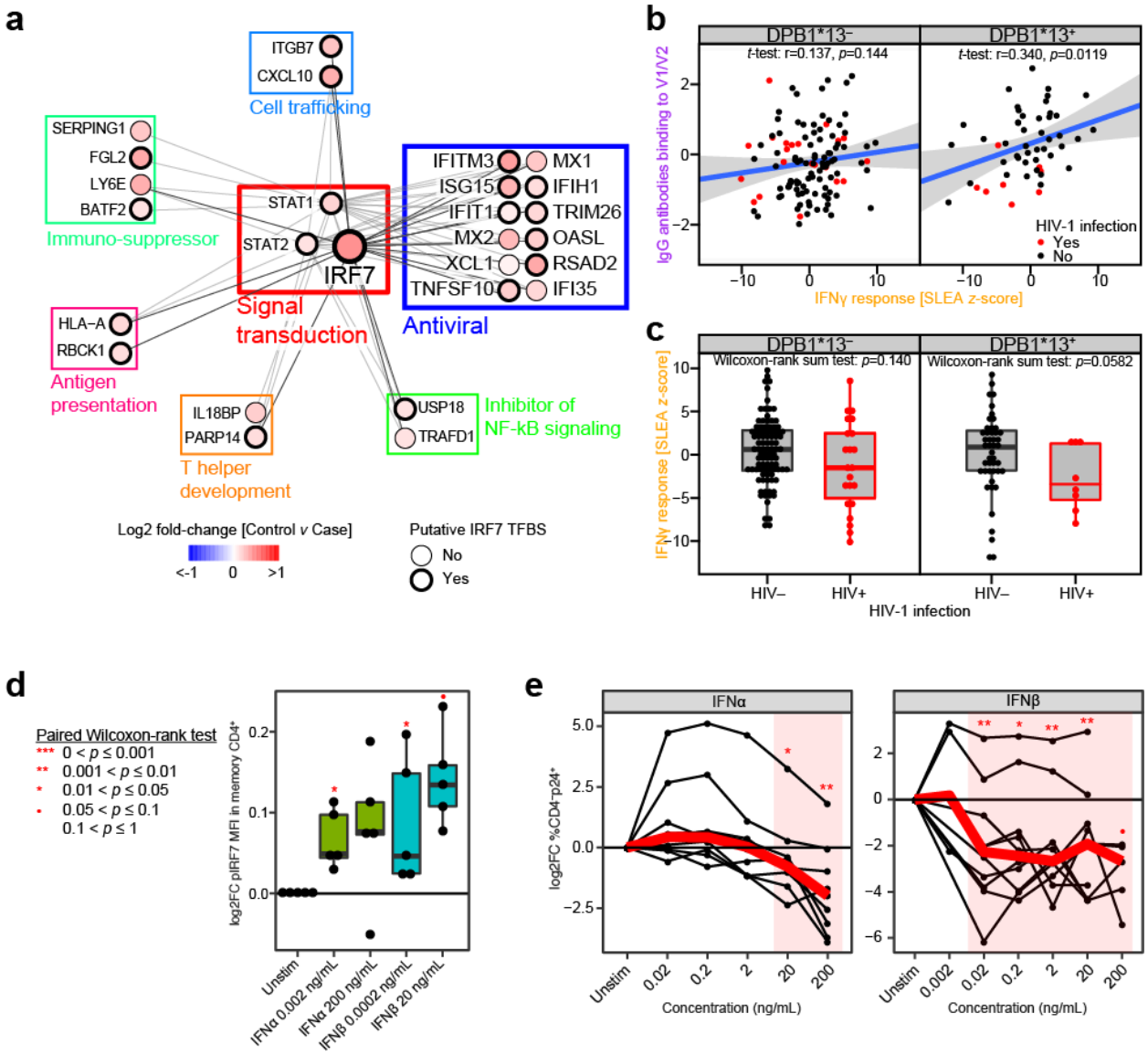


Figure A2.5. Mechanisms associated with a reduced risk of HIV-1 acquisition among RV144 vaccinees

a Network showing the genes implicated in IFNγ signaling with annotated functions. Nodes correspond to genes; the color of a node is proportional to the log2 fold-change between controls and HIV-1 cases. Edges are inferred by GeneMANIA and correspond to physical interactions, colocalization or co-expression. **b** Scatter plot presenting the expression of IFNγ responsive genes as a function of the levels of IgG antibodies binding to V1/V2 and DPB1*13 alleles. The average expression of the IFNγ genes was calculated using the SLEA z-score method. A linear regression model (blue line), and its 95% confidence interval (grey zone), was fit between SLEA z-score and IgG antibodies against V1/V2, and this separately for each DPB1*13 allele. A Pearson correlation and a t-test were performed to assess the significance of the correlation

between the transcriptomic data and antibody response. **c** Scatter plot presenting the association of IFNγ target genes and HIV-1 acquisition, separately for patients DPB1*13− and DPB1*13+. Wilcoxon-rank sum test was performed to assess the significance of the association between the transcriptomic data and HIV-1 acquisition. **d** Boxplot of the ratio of phosphorylated IRF7 in memory CD4+ cells stimulated with interferon compared to unstimulated memory CD4+ cells. The ex vivo experiments were performed on cells from five healthy donors. The fold-change in the Median Fluorescence Intensity (MFI) between interferon stimulated samples and the unstimulated condition is presented as a function of the concentration of interferon α and β used. **e** Lines plot showing the ratio of the frequency of CD4—p24+ after interferon stimulation over the unstimulated levels as a function of interferon concentration. The red lines indicate the median frequencies of CD4—p24+ across 10 healthy donors. e-f A paired Wilcoxon rank-sum test was used to assess the statistical significance of the fold-change (***: $p \leq 0.001$, **: $0.001 < p \leq 0.01$, *: $0.01 < p \leq 0.05$, •: $0.05 < p \leq 0.1$).

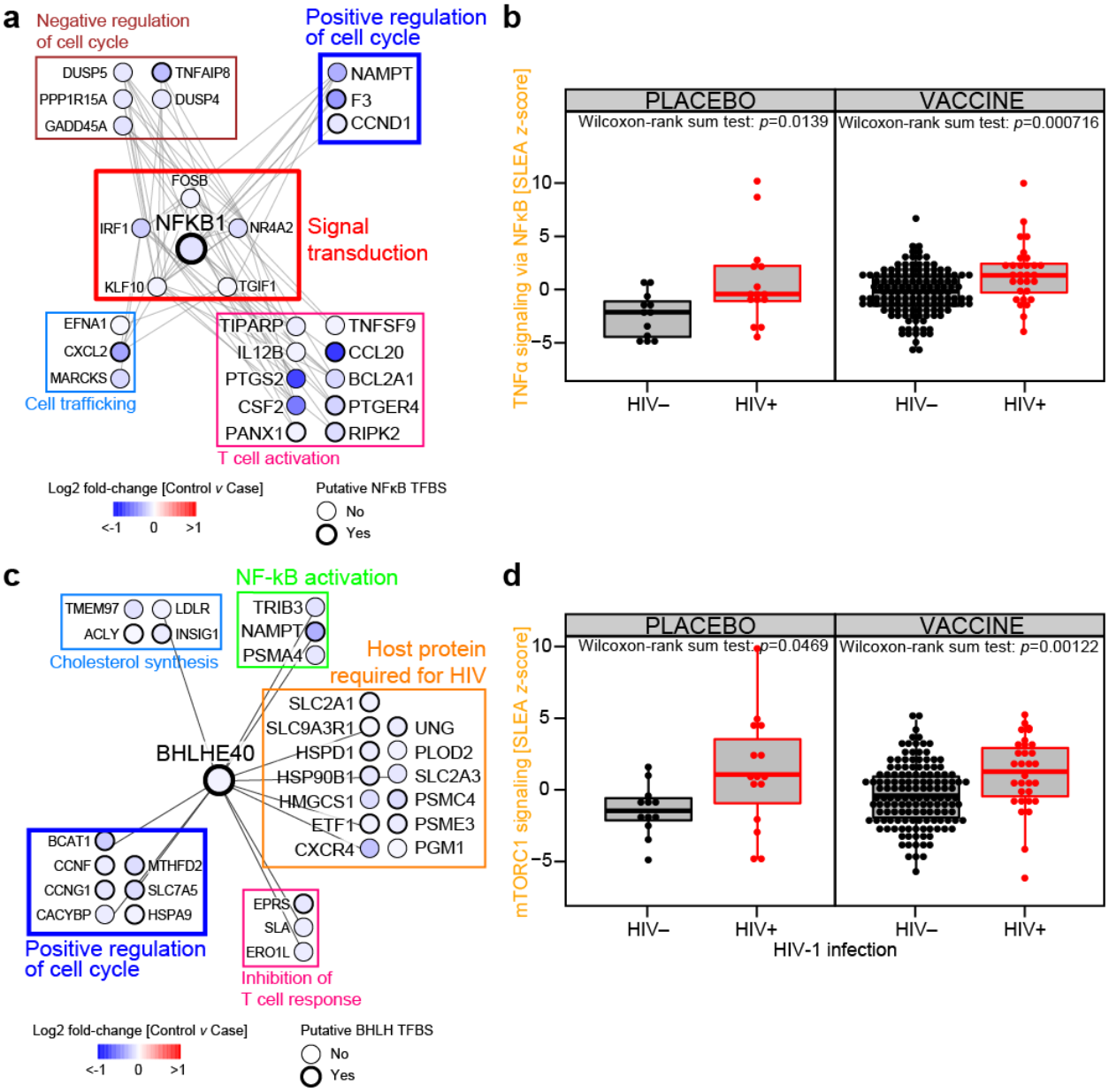


Figure A2.6. Mechanisms associated with increased risk of HIV-1 acquisition

a Network showing the genes implicated in NF-κB signaling. Nodes correspond to genes; the color of a node is proportional to the log2 fold-change between controls and HIV-1 cases. Edges are inferred by GeneMANIA and correspond to physical interactions, colocalization or co-expression. **b** Boxplot presenting the association of genes implicated in NF-κB signaling and HIV-1 acquisition, separately for placebo recipients and vaccinees. Wilcoxon-rank sum test was performed to assess the significance of the association between the transcriptomic data and HIV-1 acquisition. On the boxplot, the lower whisker, the lower hinge, the midhinge, the upper hinge and the upper whisker correspond to the interquartile (IQR) from the 1st quartile, the 1st quartile,

the median, the 3rd quartile and the IQR from the 3rd quartile, respectively. **c** Network showing the genes implicated in mTORC1 signaling. Nodes correspond to genes; the color of a node is proportional to the log2 fold-change between controls and HIV-1 cases. Edges are inferred by GeneMANIA and correspond to physical interactions, colocalization or co-expression. **d** Boxplot presenting the association of genes implicated in mTORC1 signaling and HIV-1 acquisition, separately for placebo recipients and vaccinees. Wilcoxon-rank sum test was performed to assess the significance of the association between the transcriptomic data and HIV-1 acquisition.

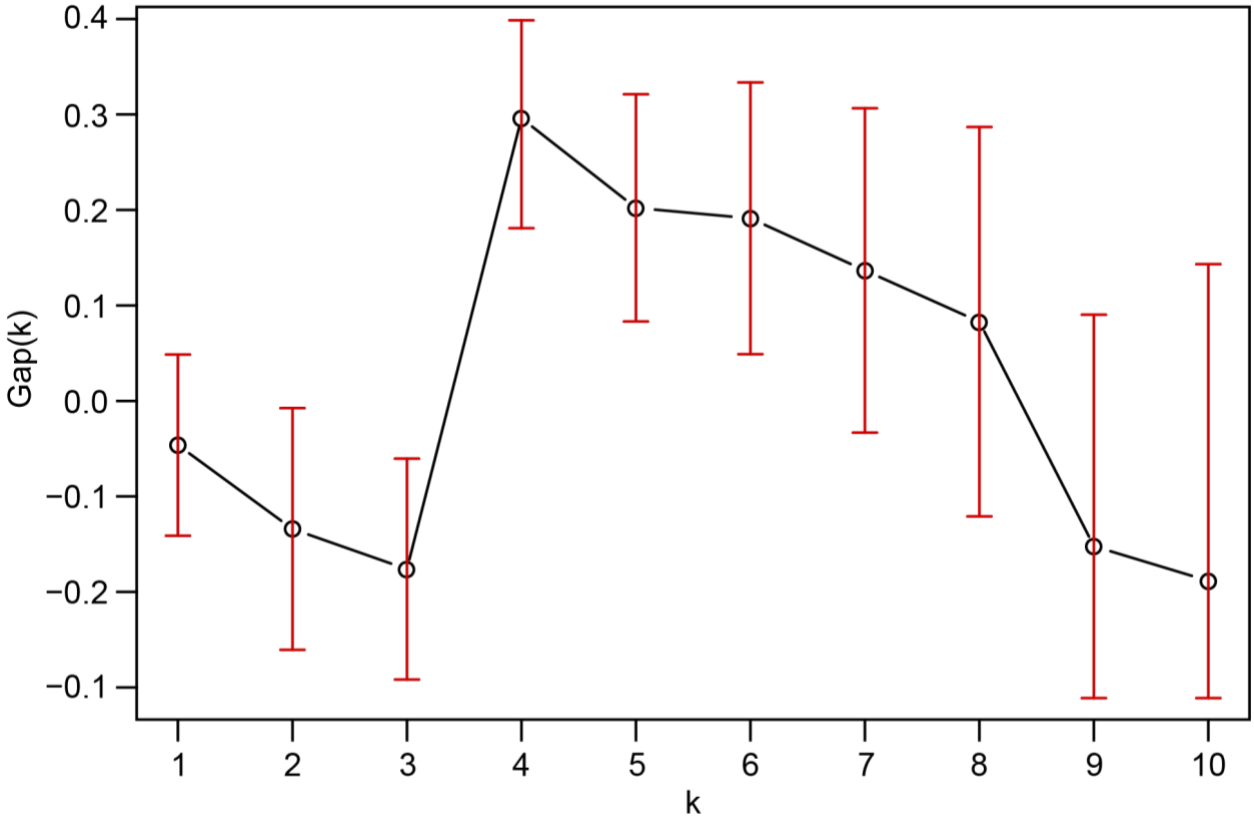
Tables

Table A2.1. Univariate and multivariate analysis of markers of HIV-1 acquisition among vaccinees.

	Univ. OR (95% CI)	Univ. *p*	Multiv. OR (95% CI)	Multiv. *p*
IgA antibodies binding to Env	1.54 [1.06, 2.26]	**0.0237**	1.22 [0.733, 2.02]	0.432
IgG antibodies binding to V1/V2	0.703 [0.446, 1.07]	0.112	1.24 [0.670, 2.29]	0.485
DQB1*06	1.12 [0.408, 2.78]	0.819	0.505 [0.0297, 2.75]	0.517
Interaction IgA:DQB1*06	---	---	9.77 [1.95, 112]	**0.0216**
DPB1*13	0.696 [0.270, 1.64]	0.426	0.453 [0.0975, 1.53]]	0.243
Interaction IgG:DPB1*13	---	---	0.132 [0.0216, 0.537]	**0.0115**
PFS	0.620 [0.390, 0.944]	**0.0322**	0.477 [0.252, 0.844]	**0.0153**
INTERFERON_GAMMA_RESPONSE	0.883 [0.803, 0.967]	**0.00837**	0.974 [0.857, 1.10]	0.677
MTORC1_SIGNALING	1.33 [1.15, 1.57]	**0.000336**	1.25 [1.01, 1.56]	**0.0496**
TNFA_SIGNALING_VIA_NFKB	1.24 [1.10, 1.42]	**0.000871**	1.08 [0.919, 1.28]	0.368

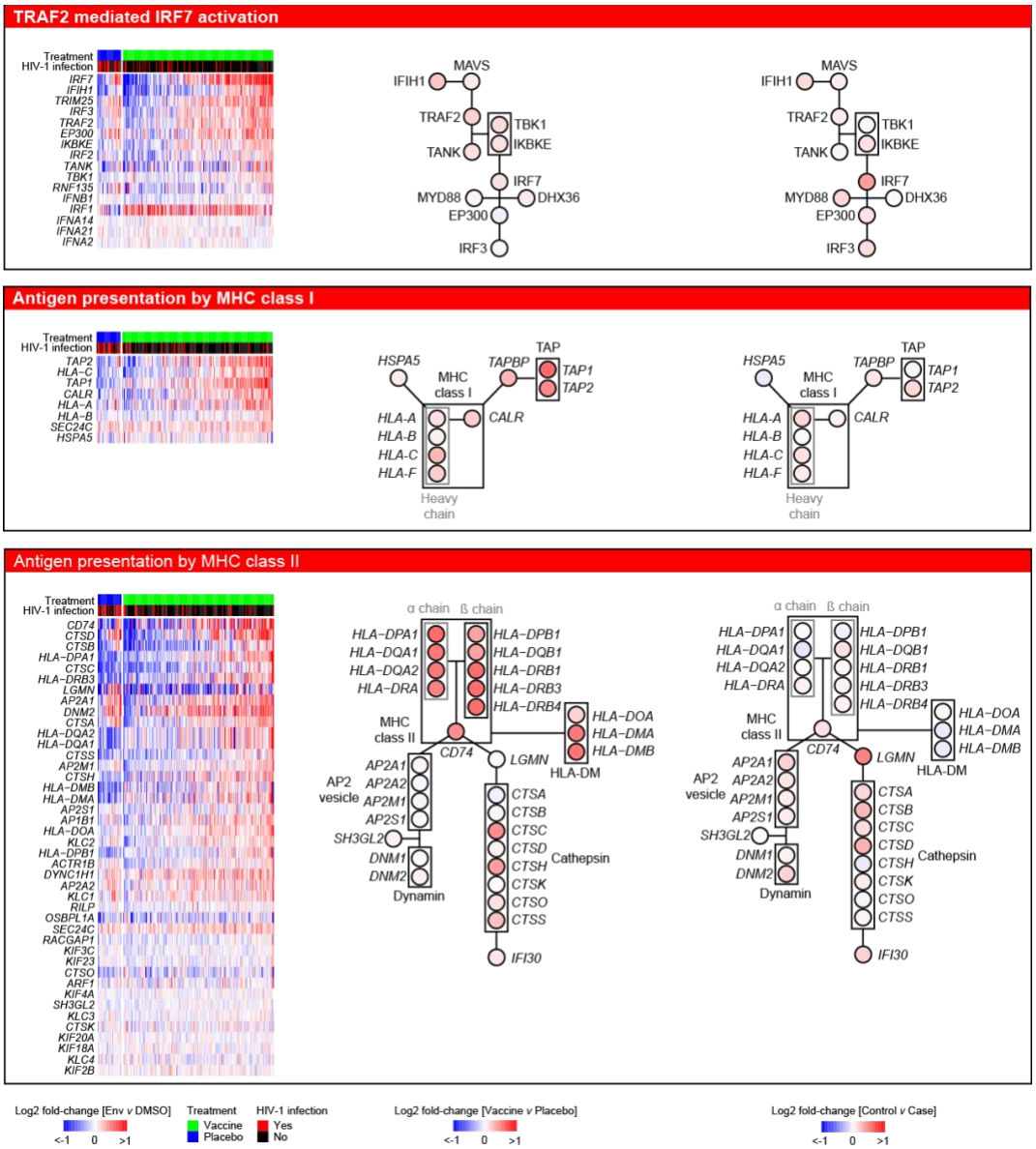
For each variable, the odds ratio (OR) and its 95% confidence interval (CI) is reported per one standard deviation increase. The p-value of a z-test testing that the OR is different from 1 is reported in the table. P-values inferior or equal 0.05 are indicated in bold. All univariate (univ.) and multivariate (multiv.) logistic regression models were adjusted for gender and behavior risk of the participants.

Supplemental Data



Supplementary Figure A2.1. Gap statistic revealed four clusters of correlated genesets

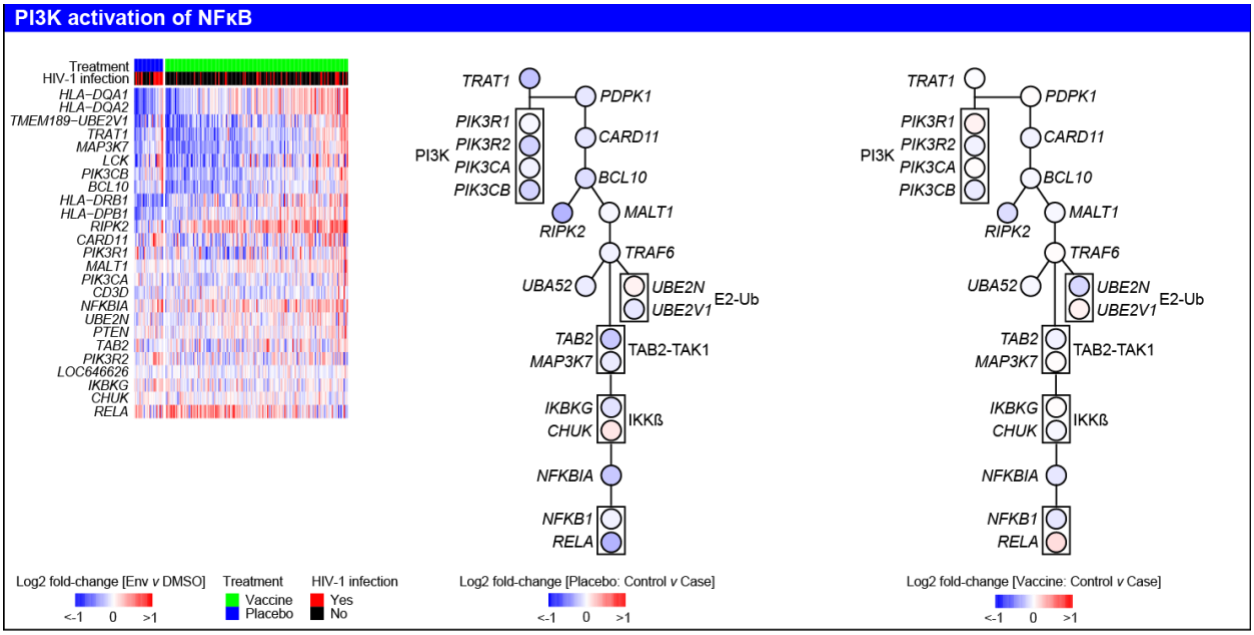
Scatter plot showing the gap statistic estimation of the optimal number of cluster of genes among the 11 pathways induced by the RV144 vaccine in the transcriptomic pilot cohort. The x-axis corresponds to the different number of clusters tested and the y--axis corresponds to the gap statistic (Gap) and its 95% confidence interval calculated over 100 bootstrap iterations. The gap statistic corresponds to the between- clusters variance divided by the intra-cluster variance (the greater the gap statistic, the better the fit). The optimal number of clusters was obtained following the rule described in Tibshirani *et al.*[1] defined as the smallest k such that Gap(k) ≥ Gap(k+1) – sd(k+1). Consequently, four clusters were identified as the optimal number of clusters of pathways.



Supplementary Figure A2.2. IRF7 activation, MHC class I/II are induced by the vaccine and associated with low risk of HIV-1 acquisition

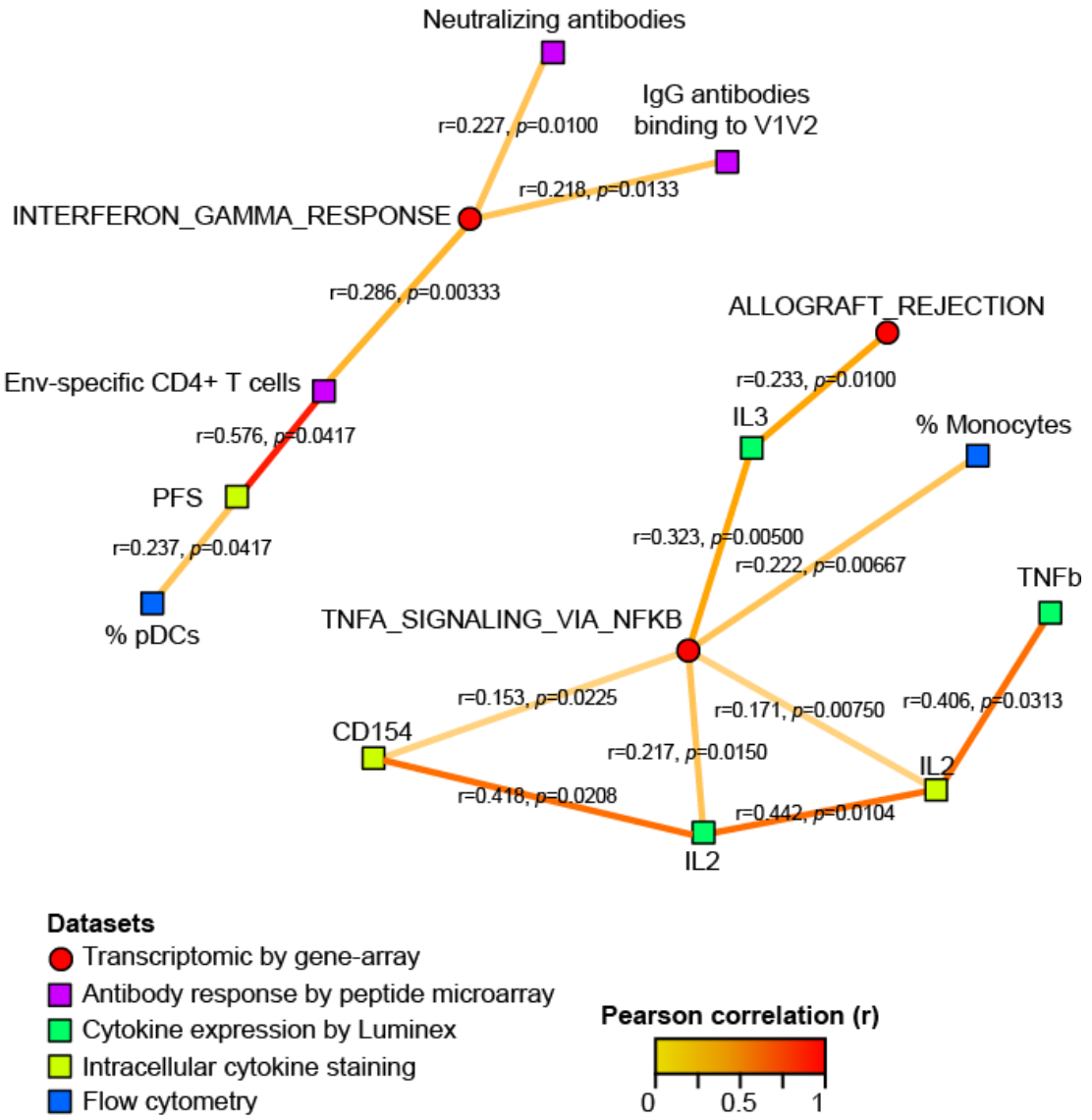
GSEA was used to identify canonical pathways (MSigDB v.5.1, class C2P) enriched in genes induced by the RV144 vaccine and up-regulated in participants that remained HIV-1 negative at last follow-up (Control) compared to participants that acquired HIV-1 despite being vaccinated (Case). Three canonical pathways, TRAF2 activation of IRF7 (upper panel), Antigen presentation by MHC class I (middle panel) and Antigen presentation by MHC class II (lower panel) are presented. For each panel, a heatmap showing the expression of the leading edges genes induced by the vaccine or induced in controls versus cases is presented on the left. A network based on interactions (edges) described in the literature (Reactome, Literome) with each node

(genes) colored by the fold-change between vaccine and placebo (middle of the panel) and by the fold-change between controls versus cases (right side) is given for each canonical pathway.



Supplementary Figure A2.3. NF-κB activation is associated with HIV-1 acquisition in placebo and vaccinees

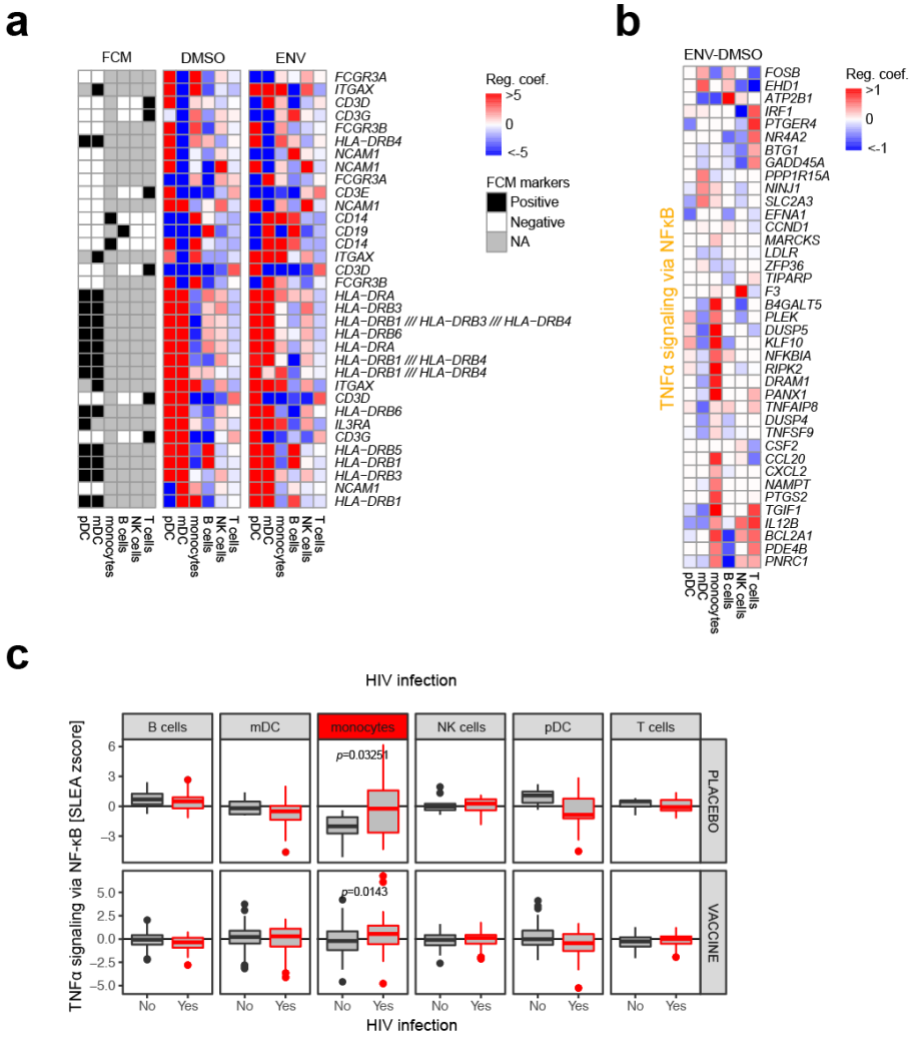
GSEA was used to identify canonical pathways (MSigDB v.5.1, class C2P) enriched in genes down-regulated in participants that remained HIV-1 negative at last follow-up (Control) compared to participants that acquired HIV-1 (Case), for both placebo recipient and RV144 vaccinees. One canonical pathway, the NF-κB activation is shown in the figure. A heatmap presenting the expression of the leading edges genes induced in cases versus controls is displayed on the left side. A network based on interactions (edges) described in the literature (Reactome, Literome) with each node (genes) of the network colored by the fold-change between controls versus cases in placebo recipient (middle of the panel) and vaccinees (right side) is given for the canonical pathway.



Supplementary Figure A2.4. Integrative analysis of the transcriptomic, antibody response, and cytokine expression

The RV144 vaccine modulated four pathways (Hallmark pathways identified using the Gene Set Enrichment Analysis2). A projection-based integrative analysis3 was performed to determine associations between those four pathways and previously investigated readouts. Each dot corresponds to a readout; edges connect two readouts significantly positively correlated (Pearson t-test: $p \leq 0.05$). Each edge is labeled with the Pearson correlation and its corresponding p-value. Among the four pathways modulated by the RV144 vaccine, three pathways (red circle) were significantly correlated to immune subset frequencies, antibody or cytokine expression (squares). IFNγ signaling pathway associated with a low risk of HIV-1 acquisition of RV144 vaccinees (Interferons pathways) were significantly correlated to levels of IgG against V1/V2, a previously identified marker of good response to the RV144 vaccine. TNFα signaling pathway associated

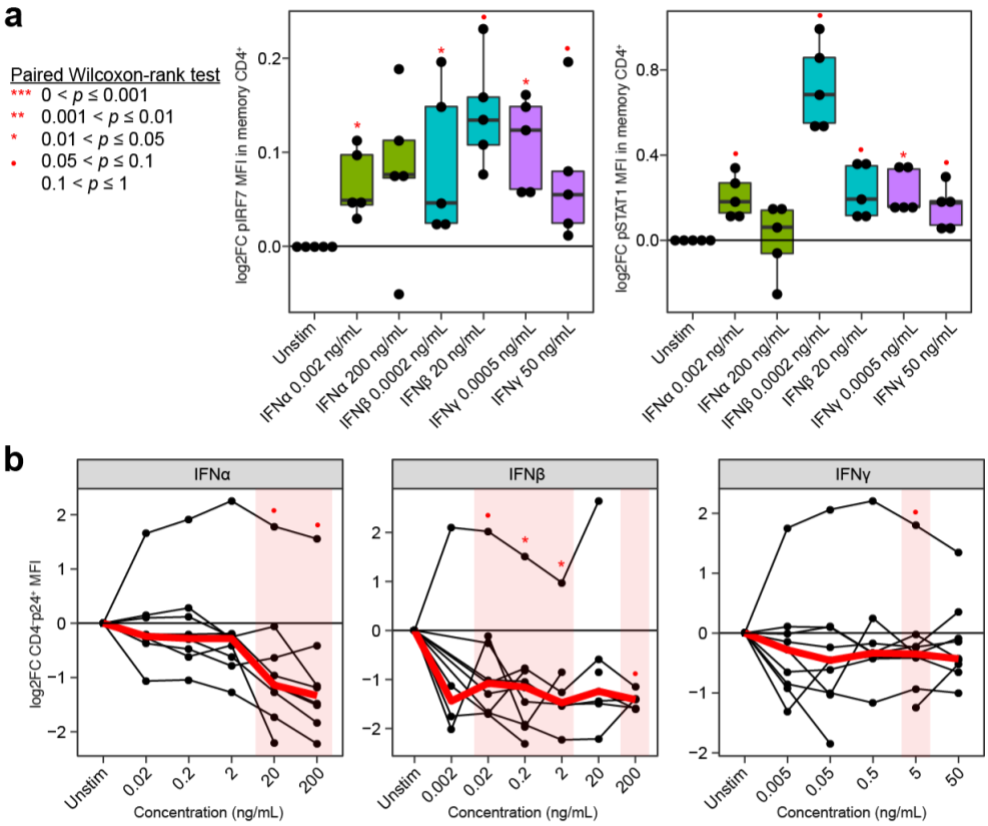
with a higher risk of HIV-1 acquisition was significantly positively correlated with inflammatory cytokines IL2+IL3 and the frequency of monocytes (inflammatory effector cells).



Supplementary Figure A2.5. Deconvolution of PBMC expression

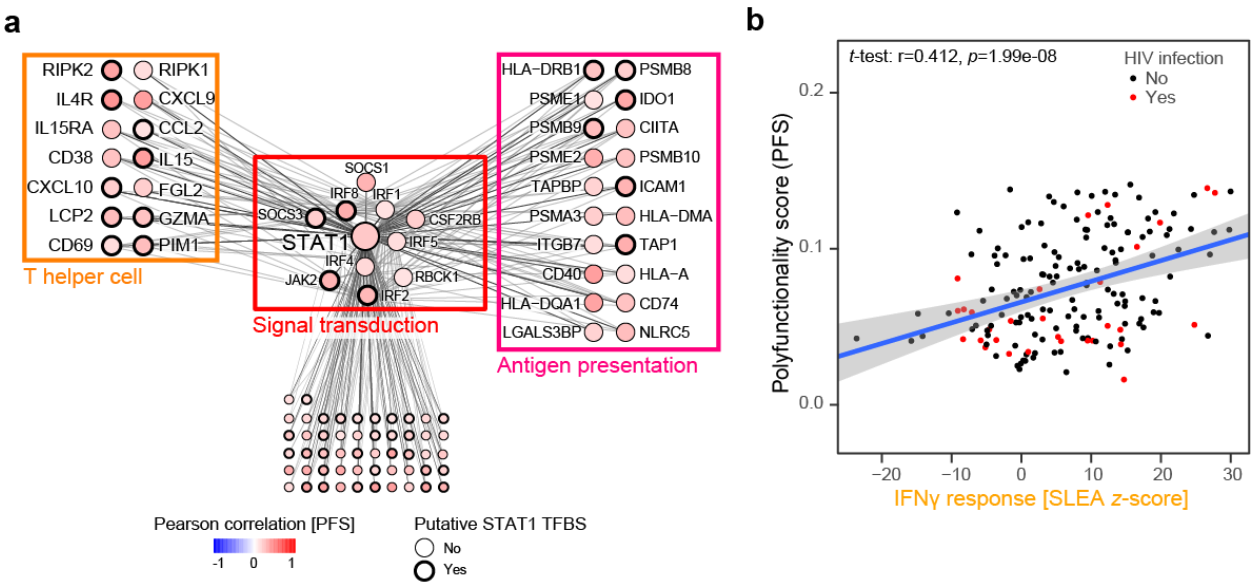
a Heatmap presenting the expression of genes coding for cells surface markers in the deconvoluted matrix. **b** Heatmap showing the expression of the 40 NF-κB target genes in the six deconvoluted subsets. For each gene a linear model was fit and deconvolution was applied. The regression coefficient (i.e. the cell-specific expression for each gene) is plotted using a blue-white-red color gradient. The monocytes subsets expressed the highest levels of the NF-κB target genes compared to the other subsets (Wilcoxon rank-sum test: p=1.64e-06). **c** Boxplots of NF-κB target genes in the six subsets stratified by HIV-1 infection. A Wilcoxon rank-sum test was performed to assess the difference between participants that acquire HIV-1 and those that remained HIV-1 negative. On the boxplot, the lower whisker, the lower hinge, the middle hinge,

the upper hinge and the upper whisker correspond to the interquartile (IQR) from the 1st quartile, the 1st quartile, the median, the 3rd quartile and the IQR from the 3rd quartile, respectively.



Supplementary Figure A2.6. Treatment of lymphocytes with interferons result in the activation of IRF7 and render them refractory to infection

a Boxplot of the ratio of the phosphorylated IRF7 and phosphorylated STAT1 after interferon stimulation of memory CD4+ compared to unstimulated memory CD4+. The ex vivo experiments were performed on cells from five healthy donors. The fold-change in median of fluorescence intensity (MFI) between interferon stimulated samples and the unstimulated condition is presented as a function of the concentration of interferon used. On the boxplot, the lower whisker, the lower hinge, the middle hinge, the upper hinge and the upper whisker correspond to the interquartile (IQR) from the 1st quartile, the 1st quartile, the median, the 3rd quartile and the IQR from the 3rd quartile, respectively. **b** Lines plot showing the ratio of CD4—p24+ MFI after interferon stimulation over the unstimulated levels as a function of interferon concentration. The red line indicates the median levels of CD4—p24+ MFI across 10 healthy donors. a-b A paired Wilcoxon rank-sum test was used to assess the statistical significance of the fold-change (***: p ≤ 0.001, **: 0.001 < p ≤ 0.01, *: 0.01 < p ≤ 0.05, •: 0.05 < p ≤ 0.1).



Supplementary Figure A2.7. Mechanisms associated with an increased CD4+ T cell response **a** Network showing the genes implicated in IFNγ signaling. Nodes correspond to genes; the color of a node is proportional to the Pearson correlation with Polyfunctionality score (PFS). Edges are inferred by GeneMANIA and correspond to physical interactions, colocalization or co-expression. **b** Scatter plot presenting the association of IFNγ responsive genes as a function of the levels of PFS. The average expression of the list of genes calculated using the SLEA z-score method is presented on the x-axis. A linear regression model (blue line), and its 95% confidence interval (grey zone), was fit between SLEA z-score and PFS. A Pearson correlation and a t-test were performed to assess the significance of the correlation between the transcriptomic data and CD4+ T cell response.

Supplementary Table A2.1. Clinical characteristics of the RV144 cohort and transcriptomic pilot study

	Full cohort	Transcriptomic pilot study	Fisher exact test: *p*
n	16,402	50	
Treatment			2.33e-05
vaccine	8,202 (50.0%)	40 (80%)	
placebo	8,200 (50.0%)	10 (20%)	
HIV infection			0.765
acquired	132 (0.805%)	0 (0%)	
negative	15,823 (96.5%)	50 (100%)	
N/A	447 (2.73%)	0 (0%)	
Gender			0.110
male	10,068 (61.4%)	25 (50%)	
female	6,334 (38.6%)	25 (50%)	
Age group			0.166

≤ 20 yr	4,546 (27.7%)	16 (32%)	
21-25 yr	7,344 (44.8%)	16 (32%)	
≥ 26 yr	4,512 (27.5%)	18 (36%)	
Province			0.575
Chon Buri	8,219 (50.1%)	23 (46%)	
Rayong	8,183 (49.9%)	27 (54%)	
Risk group			0.884
low	7,789 (47.5%)	22 (44%)	
medium	4,664 (28.4%)	15 (30%)	
high	3,949 (24.1%)	13 (26%)	
Per-protocol			2.25e-06
yes	12,542 (76.5%)	50 (100%)	
no	3,860 (23.5%)	0 (0%)	

Supplementary Table A2.2. Clinical characteristics of the placebo and vaccinees of the transcriptomic pilot study

	Vaccine	Placebo	Fisher exact test: *p*
n	40	10	
HIV infection			
controls	40 (100%)	10 (100%)	
Gender			1
male	20 (50%)	5 (50%)	
female	20 (50%)	5 (50%)	
Age group			0.907
≤ 20 yr	12 (30.0%)	4 (40%)	
21-25 yr	13 (32.5%)	3 (30%)	
≥ 26 yr	15 (37.5%)	3 (30%)	
Province			0.0850
Chon Buri	21 (52.5%)	2 (20%)	
Rayong	19 (47.5%)	8 (80%)	
Risk group			0.189
low	19 (47.5%)	3 (30%)	
medium	13 (32.5%)	2 (20%)	
high	8 (20.0%)	5 (50%)	
Per-protocol			
yes	40 (100%)	10 (100%)	

Risk group: Risk of HIV-1 infection based on participants behavior; Per protocol: completed the full immunization course (4/4 immunizations).

Supplementary Table A2.3. Clinical characteristics of the RV144 cohort and transcriptomic case/control study

	Full cohort	Transcriptomic case/control study	Fisher exact test: *p*
n	16,402	213	
Treatment			< 2.2e-16

vaccine	8,202 (50.0%)	183 (85.9%)	
placebo	8,200 (50.0%)	30 (14.1%)	
HIV infection			< 2.2e-16
acquired	132 (0.805%)	48 (22.5%)	
negative	15,823 (96.5%)	165 (77.5%)	
N/A	447 (2.73%)		
Gender			0.671
male	10,068 (61.4%)	134 (62.9%)	
female	6,334 (38.6%)	79 (37.1%)	
Age group			0.385
≤ 20 yr	4,546 (27.7%)	52 (24.4%)	
21-25 yr	7,344 (44.8%)	105 (49.3%)	
≥ 26 yr	4,512 (27.5%)	56 (26.3%)	
Province			0.581
Chon Buri	8,219 (50.1%)	111 (52.1%)	
Rayong	8,183 (49.9%)	102 (47.9%)	
Risk group			0.746
low	7,789 (47.5%)	98 (46.0%)	
medium	4,664 (28.4%)	59 (27.7%)	
high	3,949 (24.1%)	56 (26.3%)	
Per-protocol			2.83E-06
yes	12,542 (76.5%)	190 (89.2%)	
no	3,860 (23.5%)	23 (10.8%)	

Risk group: Risk of HIV-1 infection based on participants behavior; Per protocol: completed the full immunization course (4/4 immunizations).

Supplementary Table A2.4. Clinical characteristics of the placebo and vaccines of the transcriptomic case/control study

	Vaccine	Placebo	Fisher's exact test: *p*
n	183	30	
HIV infection			
acquired	31 (16.9%)	17 (56.7%)	1.05e-05
negative	152 (83.1%)	13 (43.4%)	
Gender			
male	117 (63.9%)	17 (56.7%)	
female	66 (36.1%)	13 (43.3%)	
Age group			
≤ 20 yr	44 (24.1%)	8 (26.7%)	0.966
21-25 yr	91 (49.7%)	14 (46.6%)	
≥ 26 yr	48 (26.2%)	8 (26.7%)	
Province			0.114
Chon Buri	91 (49.7%)	20 (66.7%)	
Rayong	92 (50.3%)	10 (33.3%)	
Risk group			

low	84 (45.9%)	14 (46.7%)	0.651
medium	49 (26.8%)	10 (33.3%)	
high	50 (27.3%)	6 (20.0%)	
Per-protocol			0.0508
yes	160 (87.4%)	30 (100%)	
no	23 (12.6%)	0 (0%)	

Risk group: Risk of HIV-1 infection based on participants behavior; Per protocol: completed the full immunization course (4/4 immunizations).

Supplementary Table A2.5. Associated of the four representative pathways with HIV acquisition

Geneset	Placebo NES	Placebo q-value	Vaccine NES	Vaccine q-value
HALLMARK_TNFA_SIGNALING_VIA_NFKB	-2.17	**0**	-2.21	**0**
HALLMARK_MTORC1_SIGNALING	-1.85	**0.000860**	-1.70	**0.00893**
HALLMARK_INTERFERON_GAMMA_RESPONSE	-1.42	0.0645	1.61	**0.0281**
HALLMARK_ALLOGRAFT_REJECTION	-0.893	0.785	-1.14	0.320

Two pathways are significantly associated with HIV acquisition (GSEA: FDR ≤ 5%) while one pathway is significantly associated with protection from HIV acquisition.

Supplementary Table A2.6. Univariate and multivariate analysis of IgG response and IFN signaling among vaccinees

	Univ.OR (95% CI)	Univ. *p*	Multiv. OR (95% CI)	Multiv. *p*
IgG antibodies binding to V1/V2	0.703 [0.446, 1.07]	0.112	1.22 [0.728, 2.03]	0.443
DPB1*13	0.696 [0.270, 1.64]	0.426	0.406 [0.0890, 1.25]	0.165
Interaction IgG:DPB1*13	---	---	0.111 [0.0187, 0.427]	**0.00472**
INTERFERON_GAMMA_RESPONSE	0.883 [0.803, 0.967]	**0.00837**	0.898 [0.811, 1.01]	0.0542

For each variable, the odds ratio (OR) of HIV-1 acquisition and its 95% confidence interval (CI) is reported per one standard deviation increase. The p-value of a z-test, testing that the OR is different from 1 is reported in the table. P-values inferior or equal 0.05 are indicated in bold. All univariate (univ.) and multivariate (multiv.) logistic regression models were adjusted for gender and behavior risk of the participants.

Article #3 : A crowdsourced analysis to identify ab initio molecular signatures predictive of susceptibility to viral infection

Mise en contexte

Dans le second article de cette thèse, nous n'avons pas réussi à identifier une signature transcriptionnelle qui procure une prédiction indépendante des marqueurs sérologiques qui prédisent la réponse au vaccin RV144. Dans cet article, nous contribuons un support mécanistique à cette observation en démontrant que les voies de signalisation associées à la réponse aux vaccins sont corrélées aux marqueurs sérologiques. Or une autre explication est que 1) l'approche d'apprentissage machine utilisée ne permet pas d'identifier de telles signatures ou 2) qu'il n'existe pas de signature prédictive pour ce vaccin.

Dans les deux premiers articles de cette thèse nous avons utilisé différentes approches d'apprentissage machine (1er article : classificateur naïf bayésien et 2e article : régression logistique) pour identifier des biomarqueurs de la réponse aux vaccins. Dans le 3e article nous avons comparé plusieurs méthodes d'apprentissage machine pour identifier la meilleure approche permettant de prédire la réponse du système immunitaire à une infection virale.

Dans le 3e article, nous avons utilisé la réponse aux virus respiratoires pour tester différentes méthodes d'apprentissage machine. En effet, la réponse aux virus varie considérablement d'un individu à l'autre et il n'existe actuellement aucun prédicteur moléculaire connu pouvant être mesuré durant les premiers stades de l'infection. Une compétition incluant plusieurs sites académiques a été organisée pour déterminer si les facteurs moléculaires avant ou après exposition à des virus ciblant les voies respiratoires pourraient prédire les réponses à ces infections virales. Une vingtaine d'équipes ont participé à cette compétition. Toutes les équipes ont reçu les mêmes profils d'expression transcriptionnels du sang collecté avant l'exposition à l'un des quatre virus respiratoires (H1N1, H3N2, rhinovirus et RSV), ainsi que jusqu'à 24 h après l'exposition. Chaque équipe était libre d'utiliser la méthode d'apprentissage machine de leurs choix. Une analyse communautaire a ensuite été effectuée afin de déterminer (1) si les modèles post-exposition sont de meilleurs prédicteurs des symptômes de l'infection virale que les modèles préexposition (2) si les gènes identifiés par les différents modèles prédictifs chevauchent entre les modèles (3) quelle approche d'apprentissage machine permet de construire les prédicteurs le plus robuste et précis.

Le classificateur ayant remporté la tâche de prédire la sévérité des symptômes après une infection virale est le fruit de mon travail. L'introduction a été rédigée par l'un des organisateurs de la compétition. L'ensemble des figures ont été le résultat de mon effort individuel et générées

en collaboration avec d'autres bio-informaticiens (co-premiers auteurs de l'article). La première version des sections résultats, discussion et matériel et méthodes a été rédigée par moi, et révisée en collaboration avec les autres co-premiers auteurs et le dernier auteur du papier.

A crowdsourced analysis to identify ab initio molecular signatures predictive of susceptibility to viral infection.

Slim Fourati[1,†], Aarthi Talla[1,†], Mehrad Mahmoudian[2,3,†], Joshua G Burkhart[4,5,†], Riku Klén[2,†], Ricardo Henao[6,7], Thomas Yu[8], Zafer Aydın[9], Ka Yee Yeung[10], Mehmet Eren Ahsen[11], Reem Almugbel[10], Samad Jahandideh[12], Xiao Liang[10], Torbjörn E M Nordling[13], Motoki Shiga[14], Ana Stanescu[11,15], Robert Vogel[11,16], Respiratory Viral DREAM Challenge Consortium, Gaurav Pandey[11], Christopher Chiu[17], Micah T McClain[6,18,19], Christopher W Woods[6,18,19], Geoffrey S Ginsburg[6.19], Laura L Elo[2], Ephraim L Tsalik[6,19,20], Lara M Mangravite[21], Solveig K Sieberts[22]

Affiliations

[1]Department of Pathology, School of Medicine, Case Western Reserve University, Cleveland, OH, 44106, USA.

[2]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, FI-20520, Turku, Finland.

[3]Department of Future Technologies, University of Turku, FI-20014 Turku, Finland.

[4]Department of Medical Informatics and Clinical Epidemiology, School of Medicine, Oregon Health & Science University, Portland, OR, 97239, USA.

[5]Laboratory of Evolutionary Genetics, Institute of Ecology and Evolution, University of Oregon, Eugene, OR, 97403, USA.

6Duke Center for Applied Genomics and Precision Medicine, Duke University School of Medicine, Durham, NC, 27710, USA.

[7]Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708, USA.

[8]Sage Bionetworks, Seattle, WA, 98121, USA.

[9]Department of Computer Engineering, Abdullah Gul University, Kayseri, 38080, Turkey.

[10]School of Engineering and Technology, University of Washington Tacoma, Tacoma, WA, 98402, USA.

[11]Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA.

[12]Origent Data Sciences, Inc., Vienna, VA, 22182, USA.

[13]Department of Mechanical Engineering, National Cheng Kung University, Tainan, 70101, Taiwan.

[14]Department of Electrical, Electronic and Computer Engineering, Faculty of Engineering, Gifu University, Gifu, 501-1193, Japan.

[15]Department of Computer Science, University of West Georgia, Carrolton, GA, 30116, USA.
[16]IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598, USA.
[17]Section of Infectious Diseases and Immunity, Imperial College London, London, W12 0NN, UK.
[18]Medical Service, Durham VA Health Care System, Durham, NC, 27705, USA.
[19]Department of Medicine, Duke University School of Medicine, Durham, NC, 27710, USA.
[20]Emergency Medicine Service, Durham VA Health Care System, Durham, NC, 27705, USA.
[21]Sage Bionetworks, Seattle, WA, 98121, USA. lara.mangravite@sagebase.org.
[22]Sage Bionetworks, Seattle, WA, 98121, USA. solly.sieberts@sagebase.org
[†]These authors contributed equally

This work was originally published in Nature Communications:
Nat Commun. 2018 Oct 24;9(1):4418. doi: 10.1038/s41467-018-06735-8. PubMed PMID: 30356117; PubMed Central PMCID: PMC6200745.

Abstract

Respiratory viruses are highly infectious; however, the variation of physiologic responses to viral exposure is poorly understood. Most studies examining molecular predictors of response focus on late stage predictors, typically near the time of peak symptoms. To determine whether pre- or early post-exposure factors could predict response, we conducted a community-based analysis to identify predictors of resilience or susceptibility to several respiratory viruses (H1N1, H3N2, Rhinovirus, and RSV) using peripheral blood gene expression profiles collected from healthy subjects prior to viral exposure, as well as up to 24 hours following exposure. This analysis revealed that it is possible to construct models predictive of symptomatic response using profiles even prior to viral exposure. Analysis of predictive gene features revealed little overlap among models; however, in aggregate, these genes were enriched for common pathways. Heme Metabolism, the most significantly enriched pathway, was associated with higher risk of developing symptoms following viral exposure. This study is the first to demonstrate that pre-exposure molecular predictors can be identified and improve our understanding of mechanisms of response to respiratory viruses.

Introduction

Acute respiratory viral infections are among the most common reasons for outpatient clinical encounters *(1)*. Symptoms of viral infection may range from mild (e.g. sneezing, runny nose) to life-threatening (dehydration, seizures, death), though many individuals exposed to respiratory viruses remain entirely asymptomatic *(2)*. Variability in individuals' responses to exposure has been observed both in natural infections *(3)* and controlled human viral exposure studies. Specifically, some individuals remained asymptomatic despite exposure to respiratory viruses, including human rhinovirus (HRV) *(4–6)*, respiratory syncytial virus (RSV) *(4–6)*, influenza H3N2 *(4–9)* and influenza H1N1 *(4, 5, 9)*. Factors responsible for mediating response to respiratory viral exposure are poorly understood. These individual responses are likely influenced by multiple processes, including the host genetics *(10)*, the basal state of the host upon exposure *(11)*, and the dynamics of host immune response in the early hours immediately following exposure and throughout the infection *(12)*. Many of these processes occur in the peripheral blood through activation and recruitment of circulating immune cells *(13)*. However, it remains unknown whether host factors conferring resilience or susceptibility to symptomatic infectious disease can be detected in peripheral blood before infection or whether they are only apparent in response to pathogen exposure.

In order to identify such gene expression markers of resilience and susceptibility to acute respiratory viral infection, we utilized gene expression data from seven human viral exposure experiments *(6, 7, 9)*. These exposure studies have shown that global gene expression patterns measured in peripheral blood around the time of symptom onset (as early as 36 hours after viral exposure) are highly correlated with symptomatic manifestations of illness *(6, 9)*. However, these later-stage observations do not necessarily reflect the spectrum of early time point immune processes that might predict eventual infection. Since transcriptomic signals are weak at these early time points, the detection of early predictors of viral response has not yet been possible in any individual study. By combining data collected across these seven studies and leveraging the community to implement state-of-the-art analytical algorithms, the *Respiratory Viral DREAM Challenge* (www.synapse.org/ViralChallenge) aimed to develop early predictors of resilience or susceptibility to symptomatic manifestation based on expression profiles that were collected prior to and at early time points following viral exposure.

Results

Human Viral Exposure Experiments

In order to determine whether viral susceptibility could be predicted prior to viral exposure, we collated 7 human viral exposure experiments: one RSV, two influenza H1N1, two influenza H3N2 and two human rhinovirus studies, in which a combined total of 148 healthy volunteers were exposed to virus (**Fig. A3.1B-C**) or sham (n=7) *(6, 7, 9)*. Subjects were excluded if pre-existing neutralizing antibodies were detected, except for the RSV study in which neutralizing antibodies were not an exclusion criteria. Each subject in the study was followed for up to 12 days after exposure and serially sampled for peripheral blood gene expression by Affymetrix Human U133A 2.0 GeneChips. Throughout the trial, subjects self-reported clinical symptom scores across 8-10 symptoms (**Fig. A3.S1**); these data were used to stratify subjects as either symptomatic or asymptomatic and to quantify symptom severity. Additionally, nasopharyngeal swabs measured viral shedding; these data were used to stratify subjects as either shedders or non-shedders (Fig. 1D). Clinical symptoms were summarized based on a modified Jackson score *(14)* and viral shedding was determined to be present if two or more measurable titers or one elevated titer was observed within 24 hours following viral exposure *(15)*. Viral shedding and clinical symptoms were provided to *Respiratory Viral DREAM Challenge* teams only for the training data set (**Fig. A3.1B**). An additional, but not previously available, human exposure experiment to the RSV virus (n = 21) was used as an independent test data set (**Fig. A3.1B-C**). The study design for this data set was similar to those of the 7 original data sets.

Data Analysis Challenge

Using these data, an open data analysis challenge, the *Respiratory Viral DREAM Challenge*, was formulated. Teams were asked to predict viral shedding and clinical symptoms based on peripheral blood gene-expression from up to two timepoints: prior to viral exposure (T_0) or up to 24 hours post viral exposure (T_{24}). Based on gene expression data from the two timepoints, teams were asked to predict at least one of three outcomes: presence of viral shedding (subchallenge 1 (SC1)), presence of symptoms, defined as a modified Jackson score ≥ 6 (subchallenge 2 (SC2)), or symptom severity, defined as the logarithm of the modified Jackson score (subchallenge 3 (SC3)). Teams were asked to submit predictions based on gene-expression and basic demographic (age and gender) data from both timepoints to enable cross-timepoint comparison. The 7 collated data sets served as a training dataset on which teams could build their predictive models. For a subset of subjects (n = 23), phenotypic data were withheld to serve as a leaderboard test set for evaluation with real-time feedback to teams (**Fig. A3.1A**).

Teams were asked to submit at least one leaderboard submission at each timepoint to be evaluated on the leaderboard test set. Performance metrics for these models were returned in real-time, and teams could update their submissions accordingly up to a maximum of 6 combined submissions per subchallenge. At the end of this exercise, teams were asked to provide leave-one-out cross-validation-based predictions on the training set (LOOCVs) and predictor lists for each of their best models.

Each team's best models (one for T_0 and one for T_{24}) per subchallenge were ultimately assessed on the held-out human RSV exposure data set that had not been publicly available, previously (**Fig. A3.1A**). Predictions for the binary outcomes (shedding and symptoms) were assessed using Area Under the Precision-Recall (AUPR) and Receiver Operating Characteristic (AUROC) curves, and ranked using the mean rank of these two measures. The predictions for the continuous outcome (symptom severity) were assessed using Pearson's correlation with the observed values. In each case, permutation-based p-values were used to identify submissions that performed significantly better than those expected at random.

Challenge Results

For presence of symptoms (SC2), 27 models were assessed on the independent test data; 13 models were developed using T_0 predictors, and 14 models using T_{24} predictors. Four of the T_0 models and three of the T_{24} models achieved a nominal p-value of 0.05 for AUPR or AUROC, with the best scoring models at each timepoint achieving similar scores (AUPR(T_0)=0.958,

AUROC(T_0)=0.863, AUPR(T_{24})=0.953, AUROC(T_{24})=0.863). Team *Schrodinger's Cat* was the only team that achieved significance for all measures and timepoints. Despite the few teams achieving statistical significance, the models submitted were overall more predictive than expected at random (enrichment p-values 0.008, 0.002, 0.021, and 0.05 for AUPR(T_0), AUROC(T_0), AUPR(T_{24}), and AUROC(T_{24}), respectively; **Fig. A3.2A**).

For symptom severity (SC3), 23 models were assessed on the independent test data; 11 models were developed using T_0 predictors and 12 models using T_{24} predictors. Four of the T_0 models and two of the T_{24} models achieved a nominal p-value of 0.05 for correlation with the observed log-symptom score, and as above, the best performing models scored similarly at both timepoints (r=0.490 and 0.495 for T_0 and T_{24}, respectively). Teams *cwruPatho* and *Schrodinger's Cat* achieved significant scores at both timepoints. Consistent with SC2, we also saw that the models submitted were overall more predictive than expected at random (enrichment p-values 0.005 and 0.035 for T_0 and T_{24}, respectively; **Fig. A3.2B**). For both SC2 and SC3, enrichment was more pronounced at T_0 compared to T_{24}. Correlation between final scores and leaderboard scores was higher at T_0, suggesting T_{24} predictions may have been subject to a greater degree of overfitting.

For viral shedding (SC1), 30 models were assessed from 16 different teams; 15 models were developed using T_0 predictors and 15 models using T_{24} predictors. No submissions were statistically better than expected by random. In aggregate, these submissions showed no enrichment (enrichment p-values 0.94, 0.95, 0.82, and 0.95, for AUPR(T_0), AUROC(T_0), AUPR(T_{24}), and AUROC(T_{24}), respectively). In contrast, final scores were negatively correlated with leaderboard scores (correlation -0.22, -0.19, -0.65, and -0.54 for AUPR(T_0), AUROC(T_0), AUPR(T_{24}), and AUROC(T_{24}), respectively) suggesting strong overfitting to the training data or a lack of correspondence to viral shedding as assessed in the independent test data set, relative to the training data sets. The negative correlation was strongest at T_{24} (**Fig. A3.S2**). Accordingly, results based on this subchallenge were excluded from further analysis.

Best performing approaches

The two overall best performing teams were *Schrodinger's Cat* and *cwruPatho*. Team Schrodinger's Cat used the provided gene expression profiles before the viral exposure to predict shedding and log symptom scores (binary and continuous outcomes, respectively). For the T_0 models, arithmetic means over measurements prior to exposure were calculated, whereas for the T_{24} models, only the latest measurements before viral exposure were used. Epsilon support vector regression (epsilon-SVR) *(16)* with radial kernel and 10-fold cross-validation were used to develop

the predictive models. Their work demonstrated that predictive models of symptoms following viral exposure can be built using pre-exposure gene-expression.

Team *cwruPatho* constructed models of infection based on pathway modulation, rather than gene expression, to predict infection outcomes. To do so, they used a sample-level enrichment analysis *(17)* approach to summarize the expression of genes implicated in the Hallmark gene sets *(18)* of the Molecular Signature DataBase (MSigDB) *(19)*. They then fitted LASSO regularized regression models, which integrate feature selection with a regression fit *(20)*, on the pathways to predict shedding, presence of symptoms and symptom severity following viral exposure. Their work demonstrated that including multiple genes sharing the same biological function results in more robust prediction than using any single surrogate gene.

Teams *Schrodinger's Cat* and *cwruPatho* used different feature transformation methods and machine learning approaches, suggesting that methods can successfully identify pre- or early post-exposure transcriptomic markers of viral infection susceptibility or resilience. To gauge the range of approaches taken, we extended this comparison to all *Respiratory Viral DREAM Challenge* teams who reported details on the methods they used to develop their submissions. We assessed the range of data preprocessing, feature selection and predictive modeling approaches employed for the submissions, to determine whether any of these methods were associated with prediction accuracy. Details of these three analysis steps (preprocessing, feature selection and predictive modeling) were manually extracted from reports of 24 teams (35 separate reports) who submitted predictions either for the leaderboard test set or the independent test set. To more precisely reflect the conceptual variations across employed methodologies, each of these three analysis tasks was broken down into 4 data preprocessing categories, 7 feature selection categories and 9 predictive modeling categories (**Table A3.S1**). Twenty of 24 (83.3%) teams employed some version of data preprocessing, the task most significantly associated with predictive ability (**Fig. A3.S3A**). Specifically, exclusion of sham-exposed subjects and data normalization associated best with predictive performance (**Fig. A3.3**).

Feature selection and predictive modeling approaches positively associated with predictive ability differed depending on whether the task was classification (presence of symptoms) or regression (symptom severity). Random forest-based predictive models performed slightly better than SVM/SVR methods at predicting symptom status (SC2) (**Fig. A3.S3B**). However, there was no discernible pattern relating feature selection and improved performance in SC2. Feature selection using machine learning approaches such as cross-validation was associated with improved performance in predicting symptom severity (SC3) (**Fig. A3.3**), as were SVM/SVR approaches when compared to linear regression model-based methods (e.g. logistic

regression; **Fig. A3.S3C**). Of note, SVM/SVR approaches were the most popular among the submissions.

We also sought to compare cross-timepoint predictions to determine the stability of predictions by timepoint. Significant correlation was observed between predictions using T_0 and T_{24} gene expression for symptomatic classification (SC2) (Leaderboard: ρ=0.608, p=1.04e-61; Independent test set: ρ=0.451, p=2.05e-25). Interestingly, we observed that approximately 25% of subjects were difficult to predict based on T_0 gene expression profile (inherently difficult; **Fig. A3.S4**); similarly, approximately 25% of subjects were correctly predicted by the majority of teams (inherently easy; **Fig. A3.S4**). Inherently difficult subjects were also misclassified when T_{24} gene expression data was used for the predictions. Inherently easy subjects were also consistently easy to classify using T_{24} gene expression data. This suggests *ab initio* characteristics allow some subjects to be more susceptible or resilient to symptomatic disease and that, within 24 hours, those characteristics are not substantially altered in post-exposure peripheral blood expression profiles.

Biological Interpretation of Predictors

In addition to predictions, each team was asked to submit lists of gene expression features used in their predictive models. Twenty-four teams submitted predictive models with AUROC > 0.5 for SC2 or r > 0 for SC3 (leaderboard test set) for either T_0 or T_{24}, among which 6 teams submitted separate models for each virus and reported virus-specific predictors. The remaining 18 reported models independent of the virus, submitting a single model for all viruses. With the exception of the list from *cwruPatho*, which used pathway information in the selection of features, pathway analysis of individual predictor lists showed no enrichment of pathways from MSigDB *(19)*, possibly due to the tendency of most feature selection algorithms to choose one or few features from within correlated sets.

We then assessed whether models showing predictive ability (leaderboard test set AUROC > 0.5 for SC2 or r > 0 for SC3) tended to pick the same gene features, or whether the different gene sets may provide complementary information. Within each subchallenge and timepoint, significance of the overlap among predictor lists was calculated for every combination of two or more predictor lists across teams. All two-way, three-way, four-way, etc. overlaps were considered. This analysis revealed that there was no gene shared among all teams for any timepoint or subchallenge (**Fig. A3.4A**).

Despite the paucity of overlap among predictor lists, we sought to identify whether genes used in the predictive models were part of the same biological processes or pathways. In other

words, we examined if different teams might have chosen different surrogate genes to represent the same pathway. To test this hypothesis, we performed pathway enrichment analysis of the union of predictors across predictor lists within timepoint and subchallenge. We observed significant enrichments in each case (**Fig. A3.4B**), suggesting that predictive gene features are indeed complementary across models. More pathways were enriched among predictors from T_{24} models (SC2=17 pathways and SC3=20 pathways) than from T_0 models (SC2=15 pathways and SC3=17 pathways). At T_0, genes involved in the metabolism of heme and erythroblast differentiation (HEME METABOLISM), genes specifically up-regulated by KRAS activation (KRAS_SIGNALING_UP), genes defining an inflammatory response (INFLAMMATORY RESPONSE) and genes mediating cell death by activation of caspases (APOPTOSIS) were associated with presence of symptoms in both SC2 and SC3 (**Fig. A3.4B**). At T_{24}, along with HEME METABOLISM, the expression of several inflammatory response pathways like KRAS SIGNALING, INFLAMMATORY RESPONSE, genes up-regulated in response to the gamma cytokine IFNg (INTERFERON GAMMA RESPONSE), genes upregulated by IL6 via STAT3 (IL6 JAK STAT3 SIGNALING), genes regulated by NFkB in response to TNF (TNFA SIGNALING VIA NFKB) and genes encoding components of the complement system (COMPLEMENT) were associated with symptoms in both SC2 and SC3 (**Fig. A3.4B**). Additionally, there was a significant overlap in genes across timepoints and subchallenges in each of these enriched pathways (Fisher's exact test p-value ≤ 0.05).

A meta-analysis across subchallenges (SC2 and SC3) and timepoints (T_0 and T_{24}) was performed in order to identify the most significant pathways associated with outcome. HEME METABOLISM was the most significantly associated with developing symptoms (susceptibility), while OXIDATIVE PHOSPHORYLATION and MYC TARGETS were the most significantly associated with a lack of symptoms (resilience) (**Fig. A3.S5**). This indicates that heme, known to generate inflammatory mediators through the activation of selective inflammatory pathways *(21)* is the best predictor of becoming symptomatic both pre- and early post-exposure to respiratory viruses. Genes in HEME METABOLISM associated with symptoms include genes coding for the hemoglobin subunits (HBB, HBD, HBQ1 and HBZ), the heme binding protein (HEBP1) and genes coding for enzymes important for the synthesis of heme (ALAS2, FECH, HMBS, UROD). It also includes glycophorins, which are the major erythrocyte membrane proteins (GYPA, GYPB, GYPC and GYPE), which are known receptors for the influenza virus (**Fig. A3.4C**) *(22, 23)*. Genes essential for erythroid maturation and differentiation (NEF2, TAL1, EPOR and GATA1), including the transcription factor GATA1 and its targets, the hemoglobin subunit genes HBB and HBG1/2,

were also part of HEME METABOLISM associated with an increase in symptom frequencies and severity.

Discussion

Using an open data analysis challenge framework, this study showed that models based on transcriptomic profiles, even prior to viral exposure, were predictive of infectious symptoms and symptom severity, which has not been previously demonstrated. The best scoring individual models for predicting symptoms and log-symptom score, though statistically significant, fall short of practical significance. However, these outcomes suggest that there is potential to develop models and ultimately, clinically relevant tests, based on the knowledge gained from these results. However, this would necessitate further efforts to generate more data or identify different biomarker assays which more accurately assess the mechanisms observed in the transcriptomic models. Additionally, since these studies focused on healthy adults, further data generation should extend to a wider range of subjects with respect to age and health status, as well as tracking and modeling these co-factors.

A generally useful exercise in crowdsourcing-based challenges is to construct ensembles from the submissions to assimilate the knowledge contained in them, and boost the overall predictive power of the challenge *(24)*. This exercise has yielded useful results in earlier benchmark studies *(25, 26)* and the *DREAM Rheumatoid Arthritis Challenge (27)*. However, the ensembles constructed for the *Respiratory Viral DREAM Challenge* did not perform better than the respective best performers among all the individual submissions for the various subchallenges and time points. We attribute this shortcoming partly to the relatively small training set (118 subjects), which may incline the ensemble methods to overfit these data, and the assumption of class-conditioned independence of the submissions inherent in SUMMA may not have been appropriate in this challenge *(28)*. The relative homogeneity, or lack of diversity, among the submissions for the various subchallenges and timepoints may have been another potential factor behind the diminished performance of the ensembles *(29)*.

The relative homogeneity of submissions and observation that the same subjects are misclassified by almost all participating teams suggests there may be a plateau in predictive ability when using gene expression to predict the presence of symptoms or symptom severity. It is possible that an integrative analysis supplementing or replacing the gene expression data with post-transcriptional (such as metabolomic or proteomic) data could further improve accuracy. For example, metabolomic data have been used to differentiate patients with influenza H1N1 from others with bacterial pneumonia or non-infectious conditions as well as differentiate influenza

survivors from non-survivors *(30)*. With respect to proteomics, Burke *et al.* used four of the viral exposure studies described here to derive and validate a proteomic signature from nasal lavage samples which distinguish, with high accuracy, symptomatic from asymptomatic subjects at time of maximal symptoms *(31)*. Cytokines are a special class of proteins that has been investigated in a variety of infectious disease conditions. Of particular relevance, cytokine profiling has been performed for one of the influenza H3N2 studies used in this Challenge. In that work, McClain *et al.* demonstrated that several cytokines were upregulated early after viral exposure (within 24 hours in some cases) and differentiated symptomatic from asymptomatic cases *(32)*. Baseline differences in cytokine expression were not observed, however, suggesting that cytokine expression is useful for predicting response to viral exposure but not baseline susceptibility. To our knowledge, no study has identified baseline metabolomic or proteomic predictors of resilience or susceptibility to respiratory viral infection. In addition, the combination of these data with transcriptomic predictors has not yet been investigated and may yield robust predictors of susceptibility or resistance to infection.

Our analyses revealed a significant concordance between predictions at T_0 and T_{24} (**Fig. A3.S4**), as well as a significant overlap between predictors at each of these timepoints. Given the stability of predictions and predictors between T_0 and T_{24}, it appears that the pre-exposure biological mechanisms conferring susceptibility or resilience to respiratory viral infection may be observable up to one day post-exposure. We also observed significant overlap between gene signatures at both T_0 and T_{24} and late stage signatures of viral infection, reported in the literature, and derived from gene-expression 48 hours or later after viral exposure *(5–9, 15, 33–38)*. The overlap between the predictors identified in this study and the later stage signatures was more significant at T_{24} than T_0, suggesting that pre-exposure signatures of susceptibility differ somewhat from post-exposure signatures of active infection, and T_{24} predictors may reflect some aspects of both. The T_0 gene signatures may encompass novel insight into *ab initio* factors that confer resilience or susceptibility.

Pathway enrichment analysis in our study revealed that the most significantly enriched pathway associated with symptomatic infection was HEME METABOLISM, known to have a direct role in immunity through activation of innate immune receptors on macrophages and neutrophils *(21)*. Of note, genes part of HEME METABOLISM were also enriched among late stage signatures of viral infection (ex. Hemoglobin gene HBZ and the iron containing glycoprotein ACP5 in *(33)*). Iron (obtained from heme) homeostasis is an important aspect of human health and disease. Viruses require an iron-rich host to survive and grow, and iron accumulation in macrophages has been shown to favor replication and colonization of several viruses (e.g. HIV-

1, HCV) and other pathogenic microorganisms *(39)*. Furthermore, iron-replete cells have been shown to be better hosts for viral proliferation *(39)*. Increased iron loading in macrophages positively correlates with mortality *(39)* and it has been shown that viral infection can cause iron overload which could further exacerbate disease. Additionally, previous evidence suggests counteracting iron accumulation may limit infection *(21, 39)*. Studies have shown that limiting iron availability to infected cells (by the use of iron chelators) curbed the growth of several infectious viruses and ameliorated disease *(21, 39–41)*. This important role of iron in the susceptibility and response to infection may be the mechanism by which HEME METABOLISM genes conferred susceptibility to respiratory viral infection. As such, it represents an important biological pathway potentially offering a means by which an individual's susceptibility or response to infection can be optimized. Such a relationship should be investigated in future studies of infection susceptibility. In addition, Heme-oxygenase (HMOX1), a heme-degrading enzyme that antagonizes heme induced inflammation and is essential for the clearance of heme from circulation *(42)*, was among the predictors from the T_0 models. Interestingly, the expression of this gene at baseline was associated with lack of symptoms (for both SC2 and SC3), in concordance with its reported antiviral role during influenza infection *(43, 44)*. Augmentation of HMOX1 expression by gene transfer had provided cellular resistance against heme toxicity *(45)*. Hence enhancing HMOX1 activity could be an alternative to antagonize heme induced effects and thereby controlling infection and inflammation.

In addition to HEME METABOLISM, pro-inflammatory pathways such as INFLAMMATORY RESPONSE, KRAS SIGNALING, and APOPTOSIS were also associated with susceptibility to viral infection in our study, while homeostatic pathways, such as OXIDATIVE PHOSPHORYLATION and MYC TARGETS, were associated with resilience, both prior to and post-viral exposure (**Fig. A3.4**). Enrichment of these pathways among T_{24} predictors was more significant than among the T_0 predictors, suggesting these mechanisms are not only emblematic of baseline system health, but also response to viral invasion. Additional pathways enriched among T_{24} predictors include INTERFERON GAMMA RESPONSE and COMPLEMENT, which are involved in innate and acquired immunity. Several genes among T_0 and T_{24} predictors overlapped with genes positively associated with flu vaccination response *(46)*. Among them, *FCER1G* and *STAB1*, members of the inflammatory response pathway positively associated with symptoms in this study and were elevated prior to vaccination in young adults who showed good response to vaccination *(46)* (Fisher exact test: $p=0.0338$ for T_0 and $p=0.000673$ for T_{24}). This suggest that individuals predicted at a higher risk of presenting symptoms following influenza exposure may also be the most likely to benefit from vaccination.

The *Respiratory Viral DREAM Challenge* is to date the largest and most comprehensive analysis of early stage prediction of viral susceptibility. The open data analysis challenge framework is useful for comparing approaches and identifying the most scientifically or clinically relevant model or method in an unbiased fashion *(24)*. In this case, we observed few commonalities among the best performing models of symptomatic susceptibility to respiratory viral exposure. Indeed, the overall best performing teams in the challenge used different machine learning techniques to build their models. Interestingly, data preprocessing was the analysis task most significantly associated with model accuracy, suggesting what has often been speculated, that adequate attention to data processing prior to predictive modeling is a crucial first step *(47)*.

The open data challenge framework is also useful in arriving at consensus regarding research outcomes that may guide future efforts within a field *(24)*. Through this challenge, we have identified *ab initio* transcriptomic signatures predictive of response to viral exposure, which has provided valuable insight into the biological mechanisms conferring susceptibility to infection. This insight was not evident from any individual model, but became apparent with the meta-analysis of the individual signatures. While development of a diagnostic test of baseline susceptibility in not yet feasible based on these findings, they suggest potential for development in this area.

Methods

Training Data

Training data came from seven related viral exposure trials, representing four different respiratory viruses. The datasets are *DEE1 RSV*, *DEE2 H3N2*, *DEE3 H1N1*, *DEE4X H1N1*, *DEE5 H3N2*, *Rhinovirus Duke*, and *Rhinovirus UVA (6, 7, 9)*. In each of these human viral exposure trials, healthy volunteers were followed for seven to nine days following controlled nasal exposure to the specified respiratory virus. Subjects enrolled into these viral exposure experiments had to meet several inclusion and exclusion criteria. Among them was an evaluation of pre-existing neutralizing antibodies to the viral strain. In the case of influenza H3N2 and influenza H1N1, all subjects were screened for such antibodies. Any subject with pre-existing antibodies to the viral strain was excluded. For the rhinovirus studies, subjects with a serum neutralizing antibody titer to RV39 > 1:4 at pre-screening were excluded. For the RSV study, subjects were pre-screened for neutralizing antibodies, although the presence of such antibodies was not an exclusion criterion.

Symptom data and nasal lavage samples were collected from each subject on a repeated basis over the course of 7-9 days. Viral infection was quantified by measuring release of viral particles

from nasal passages ("viral shedding"), as assessed from nasal lavage samples via qualitative viral culture and/or quantitative influenza RT-PCR. Symptom data were collected through self-report on a repeated basis. Symptoms were quantified using a modified Jackson score *(14)*, which assessed the severity of eight upper respiratory symptoms (runny nose, cough, headache, malaise, myalgia, sneeze, sore throat, and stuffy nose) rated 0-4, with 4 being most severe. Scores were integrated daily over 5-day windows.

Blood was collected and gene expression of peripheral blood was performed 1 day (24 to 30 hours) prior to exposure, immediately prior to exposure, and at regular intervals following exposure. These peripheral blood samples were gene expression profiled on the Affy Human Genome U133A 2.0 array.

All subjects exposed to influenza (H1N1 or H3N2) received oseltamivir 5 days post-exposure. However, 14 (of 21) subjects in the DEE5 H3N2 cohort received early treatment (24 hours post-exposure) regardless of symptoms or shedding. *Rhinovirus Duke* additionally included 7 volunteers who were exposed to sham rather than active virus.

All subjects provided written consents, and each of the seven trials was reviewed and approved by the appropriate governing IRB.

RSV Test Data

Healthy non-smoking adults aged 18-45 were eligible for inclusion after screening to exclude underlying immunodeficiencies. A total of 21 subjects (10 female) were inoculated with 10^4 plaque-forming units of RSV A Memphis 37 (RSV M37) by intranasal drops and quarantined from 1 day before inoculation to the 12th day after. Peripheral blood samples were taken immediately before inoculation and regularly for the next 7 days and profiled on the Affy Human Genome U133A 2.0 array. Subjects were discharged after study day 12, provided no or mild respiratory symptoms and a negative RSV antigen respiratory secretions test. Shedding was determined by polymerase chain reaction (PCR) in nasal lavage and defined as detectable virus for ≥2 days between Day +2 and Day +10 to avoid false-positives from the viral inoculum and to align case definitions with the other 7 studies. Subjects filled a diary of upper respiratory tract symptoms from Day −1 to Day +12, which was summarized using a modified Jackson score. All subjects returned for further nasal and blood sampling on Day +28 for safety purposes. All subjects provided written informed consent and the study was approved by the UK National Research Ethics Service (London-Fulham Research Ethics Committee ref. 11/LO/1826).

Gene Expression Normalization

Both raw (CEL files) and normalized versions of the gene-expression data were made available to teams in the Challenge. Both versions contained only profiles that pass QC metrics including those for RNA Degradation, scale factors, percent genes present, β-actin 3' to 5' ratio and GAPDH 3' to 5' ratio in the *Affy* Bioconductor package. Normalization via RMA was performed on all expression data across all timepoints for the training and leaderboard data sets. The RSV data were later normalized together with the training and leaderboard data, and teams were free to further QC and normalize the data in the way they deemed appropriate.

Analysis Challenge Design

The training data studies were split into training and leaderboard sets, where the leaderboard subjects were chosen randomly from 3 of the trials: *DEE4X H1N1*, *DEE5 H3N2*, and *Rhinovirus Duke*, which were not publicly available at the time of challenge launch. Outcome data for the leaderboard set were not provided to the teams, but instead, teams were able to test predictions in these individuals using the leaderboard, with a maximum of 6 submissions per subchallenge, the purpose of which was to allow teams to optimize their models prior to assessment on the independent test data. Of these, at least one submission was required to use only data prior to viral exposure and at least one using data up to 24 hours post-exposure.

For the training data, teams had access to clinical and demographic variables: age, sex, whether the subject received early oseltamivir treatment (*DEE5 H3N2* only) and whether the subject received sham exposure rather than virus (*Rhinovirus Duke* only), as well as gene expression data for the entire time-course of the studies. They also received data for the three outcomes used in the data analysis challenge:

- Subchallenge 1: SHEDDING_SC1, a binary variable indicating presence of virus in nasal swab following exposure
- Subchallenge 2: SYMPTOMATIC_SC2, a binary variable indicating post-exposure maximum 5-day integrated symptom score >= 6
- Subchallenge 3: LOGSYMPTSCORE_SC3, a continuous variable indicating the log of the maximum 5-day integrated symptom score+1

as well as the granular symptom data by day and symptom category. For the leaderboard test data, they were supplied with the clinical and demographic variables and gene expression data up to 24 hours post-exposure.

Final assessment of optimized models was performed in the RSV Test Data (*i.e.* independent test set), and outcomes for these subjects were withheld from teams. In order to assure that predictions were limited to data from the appropriate time window, the gene-expression data were

released in two phases corresponding to data prior to viral exposure, and data up to 24 hours post exposure. Teams were also supplied with age and sex information for these subjects.

The Challenge was launched and training data were released May 15th, 2016 for participants to use to begin analyzing the data and building their models. In total 38 teams registered for the challenge and 37 participated (Supplementary Table S5). The leaderboards opened approximately 2 months later, and were open for approximately 3 months (July to September) to allow participants to optimize their models with feedback from the scores on the leaderboard data. At the close of this round on September 30th, participating teams were also required to submit code, methodological write-ups, predictor lists, and LOOCVs and doing so qualified participants to be included as authors (either Consortium or by-line) on this manuscript. Participating teams could opt to evaluate their optimized models in the independent test data, which occurred January to February 2017. At the close of the challenge, participating teams were invited to collaborate with the Challenge Organizers to analyze the results. Prior to the launch of the challenge, substantial effort was put forth by the Challenge organizers to collate and vet the data, to determine the feasibility of the Challenge and define the Challenge objectives. For further details on the organizational efforts required to prepare for a challenge, see Saez-Rodriguez et al. (2016) (*24*).

Submission Scoring

Team predictions were compared to true values using AUPR and AUROC for subchallenges 1 and 2, and Pearson correlation for subchallenge 3. For each submission, a p-value, estimating the probability of observing the score under the null hypothesis that the predicted labels are random, was computed by 10,000 permutations of the predictions relative to the true values.

We also had access to leaderboard predictions from 10,000 models build on data with randomly permuted labels for 3 teams for SC2 and 2 teams for SC3. This second test estimates the probability of observing the score under the null hypothesis that the independent variables does not contain information about the target variable within the model structure used in the predictor. Comparisons between permutation p-values and scores from models built on the permuted data showed that the latter approach to p-value computation was slightly more conservative (data not shown), and presumably more robust to overfitting the training data. Albeit theoretically preferable, the computational demands of this approach makes it infeasible for most challenges.

Heterogeneity of the Predictions

T_0 and T_{24} predictions for each outcome and team were collected to assess whether they were correlated to each other. Three teams provided predictions as binary values while 12 teams provided predictions as continuous values on different scales. In order to compare binary and continuous predictions, we first transformed them into ranks (with ties given the same average rank) and then ordered subjects increasingly by their mean rank across outcomes (*mean-rank*). The lower the mean-rank, the more likely a subject was predicted by the teams as not showing shedding/symptoms, whereas a higher mean-rank means a subject was predicted by most of the teams as showing shedding/symptoms. Distribution of the mean-rank (**Fig. A3.S4**) revealed three groups of subjects: (1) ~25% of subjects correctly predicted by most of the teams (i.e. inherently easy), (2) ~25% of subjects incorrectly predicted by most of the teams (i.e. inherently difficult) and (3) ~50% of subjects who were predicted differently by the teams.

Ensemble Prediction

We constructed a variety of ensembles from the teams' submissions to the various subchallenges as a part of the collaborative phase of the *Respiratory Viral DREAM Challenge*. To enable a comparative analysis between individual and ensemble models in the collaborative phase, the teams were requested to submit leave-one-out cross-validation (LOOCV)-derived predictions on the training examples using the same methods used to generate leaderboard and/or test set predictions in the competitive phase. The LOOCV setup, which doesn't involve random subsetting of the training data, was chosen to avoid potential overfitting that can otherwise occur from training and testing on predictions made on the same set of examples *(25)*. We used three types of approaches for learning ensembles, namely stacking and its clustering-based variants *(25)*, Reinforcement Learning-based ensemble selection *(26)* methods, as well as SUMMA, an unsupervised method for the aggregation of predictions *(28)*. Consistent with the process followed by the individual teams, we learnt all the ensembles using the training set LOOCV-derived predictions described above, and used the leaderboard data to select the final models to be evaluated on the test data.

Combined Gene Sets

Statistical significance of the overlap among predictor lists was calculated using the multi-set intersection probability method implemented in the SuperExactTest R package *(48)*. A first set of analysis was performed with teams whose leaderboard AUROC > 0.5. A second set of analysis aimed at identifying genes that overlap virus-specific, subchallenge-specific and timepoint-specific predictive models, was restricted to teams that provided virus-specific (*Nautilus*, *aydin*,

SSN_Dream_Team, *Txsolo*, *cwruPatho* and *Aganita*), subchallenge-specific (*aydin, SSN_Dream_Team, cwruPatho, jhou*) and timepoint-specific predictors (*aydin, SSN_Dream_Team, cwruPatho, Espoir, jdn, jhou, burkhajo*) and participated in the leaderboard phase of the challenge, respectively. For both analyses, overlapping predictors associated with p-values less than or equal to 0.05 were considered significant.

Pathway enrichment analysis

To assess pathway enrichment among predictors of infection, we considered predictors from teams with leaderboard AUROC > 0.5 (SC2) or Pearson correlation, r > 0 (SC3). Affymetrix Human U133A 2.0 GeneChip probe identifiers were mapped to gene symbols. We removed probes matching multiple genes, and when multiple probes matched a single gene, we retained the probe with the maximum median intensity across subjects.

For the list of predictors of presence of symptoms (SC2), we calculated the log2 fold-change of features (symptomatic(1)/asymptomatic(0)) at T_0 and T_{24}, and for prediction of the symptom scores (SC3), we calculated the Spearman's correlation coefficient of the features, at T_0 and T_{24}, with the outcome. Pathway enrichment was then performed on the union of all predictors (across the teams) that were associated with presence/increase severity of symptoms (SC2: log2 fold-change > 0 and SC3: Spearman's correlation > 0), as well as, for the union of all predictors (across teams) that were associated with lack of symptoms/lower symptoms severity (SC2: log2 fold-change < 0 and SC3: Spearman's correlation < 0), separately by timepoint and subchallenge. We used the Hallmark gene sets (version 6.0) *(18)* of the Molecular Signature DataBase (MSigDB) *(19)* for the enrichment, and calculated the significance of enrichment using Fisher's exact test. The resulting p-values were corrected for multiple comparisons using the Benjamini and Hochberg algorithm. Only significantly enriched pathways (corrected p-value < 0.05) were reported. Meta-analyses across subchallenges and timepoints were performed using the maxP test statistic *(49)*.

References

1. G. C. Lee, K. R. Reveles, R. T. Attridge, K. A. Lawson, I. A. Mansi, J. S. Lewis, C. R. Frei, Outpatient antibiotic prescribing in the United States: 2000 to 2010, BMC Med. 12 (2014).

2. C. L. Byington, K. Ampofo, C. Stockmann, F. R. Adler, A. Herbener, T. Miller, X. Sheng, A. J. Blaschke, R. Crisp, A. T. Pavia, Community Surveillance of Respiratory Viruses Among Families in the Utah Better Identification of Germs-Longitudinal Viral Epidemiology (BIG-LoVE) Study, Clin. Infect. Dis. (2015).

3. K. K. W. To, J. Zhou, J. F. W. Chan, K. Y. Yuen, Host genes and influenza pathogenesis in humans: An emerging paradigm, Curr. Opin. Virol. 14, 7–15 (2015).
4. L. Carin, A. Hero, J. Lucas, D. Dunson, M. Chen, R. Heñao, A. 2-Puig, A. Zaas, C. W. Woods, G. S. Ginsburg, High-Dimensional Longitudinal Genomic Data: An analysis used for monitoring viral infections, IEEE Signal Process. Mag. 29, 108–123 (2012).
5. M. Chen, D. Carlson, A. Zaas, C. W. Woods, G. S. Ginsburg, A. Hero, J. Lucas, L. Carin, Detection of viruses via statistical gene expression analysis, IEEE Trans. Biomed. Eng. 58, 468–479 (2011).
6. A. K. Zaas, M. Chen, J. Varkey, T. Veldman, A. O. Hero, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, N. C. Øien, B. Nicholson, S. Kingsmore, L. Carin, C. W. Woods, G. S. Ginsburg, Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans, Cell Host Microbe 6, 207–217 (2009).
7. Y. Huang, A. K. Zaas, A. Rao, N. Dobigeon, P. J. Woolf, T. Veldman, N. C. Øien, M. T. McClain, J. B. Varkey, B. Nicholson, L. Carin, S. Kingsmore, C. W. Woods, G. S. Ginsburg, A. O. Hero, Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection, PLoS Genet. 7 (2011).
8. M. T. McClain, B. P. Nicholson, L. P. Park, T. Y. Liu, A. O. Hero, E. L. Tsalik, A. K. Zaas, T. Veldman, L. L. Hudson, R. Lambkin-Williams, A. Gilbert, T. Burke, M. Nichols, G. S. Ginsburg, C. W. Woods, A genomic signature of influenza infection shows potential for presymptomatic detection, guiding early therapy, and monitoring clinical responses, Open Forum Infect. Dis. 3 (2016).
9. C. W. Woods, M. T. McClain, M. Chen, A. K. Zaas, B. P. Nicholson, J. Varkey, T. Veldman, S. F. Kingsmore, Y. Huang, R. Lambkin-Williams, A. G. Gilbert, A. O. Hero, E. Ramsburg, S. Glickman, J. E. Lucas, L. Carin, G. S. Ginsburg, A Host Transcriptional Signature for Presymptomatic Detection of Infection in Humans Exposed to Influenza H1N1 or H3N2, PLoS One 8 (2013).
10. A. R. Everitt, S. Clare, T. Pertel, S. P. John, R. S. Wash, S. E. Smith, C. R. Chin, E. M. Feeley, J. S. Sims, D. J. Adams, H. M. Wise, L. Kane, D. Goulding, P. Digard, V. Anttila, J. K. Baillie, T. S. Walsh, D. A. Hume, A. Palotie, Y. Xue, V. Colonna, C. Tyler-Smith, J. Dunning, S. B. Gordon, R. L. Smyth, P. J. Openshaw, G. Dougan, A. L. Brass, P. Kellam, IFITM3 restricts the morbidity and mortality associated with influenza, Nature 484, 519–523 (2012).
11. M. Pichon, B. Lina, L. Josset, Impact of the Respiratory Microbiome on Host Responses to Respiratory Viral Infection, Vaccines 5, 40 (2017).

12. A. Iwasaki, P. S. Pillai, Innate immunity to influenza virus infectionNat. Rev. Immunol. 14, 315–328 (2014).
13. J. Heidema, J. W. a Rossen, M. V Lukens, M. S. Ketel, E. Scheltens, M. E. G. Kranendonk, W. W. C. van Maren, A. M. van Loon, H. G. Otten, J. L. L. Kimpen, G. M. van Bleek, Dynamics of human respiratory virus-specific CD8+ T cell responses in blood and airways during episodes of common cold., J. Immunol. 181, 5551–9 (2008).
14. F. Carrat, E. Vergu, N. M. Ferguson, M. Lemaitre, S. Cauchemez, S. Leach, A. J. Valleron, Time lines of infection and disease in human influenza: A review of volunteer challenge studiesAm. J. Epidemiol. 167, 775–785 (2008).
15. T. Y. Liu, T. Burke, L. P. Park, C. W. Woods, A. K. Zaas, G. S. Ginsburg, A. O. Hero, An individualized predictor of health and disease using paired reference and target samples, BMC Bioinformatics 17 (2016).
16. C. Chang, C. Lin, LIBSVM : A Library for Support Vector Machines, ACM Trans. Intell. Syst. Technol. 2, 1–39 (2013).
17. N. Lopez-Bigas, S. De, S. A. Teichmann, Functional protein divergence in the evolution of Homo sapiens, Genome Biol. 9 (2008).
18. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, P. Tamayo, The Molecular Signatures Database Hallmark Gene Set Collection, Cell Syst. 1, 417–425 (2015).
19. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A 102, 15545–15550 (2005).
20. R. Tibshirani, Regression Shrinkage and Selection via the Lasso, J. R. Stat. Soc. Ser. B 58, 267–288 (1996).
21. F. F. Dutra, M. T. Bozza, Heme on innate immunity and inflammationFront. Pharmacol. 5, 115 (2014).
22. K. Ohyama, S. Yamauchi, T. Endo, S. Ohkuma, Presence of influenza virus-reactive glycophorins other than glycophorin A in human erythrocyte membranes, Biochem. Biophys. Res. Commun. 178, 79–84 (1991).
23. K. Ohyama, T. Endo, S. Ohkuma, T. Yamakawa, Isolation and influenza virus receptor activity of glycophorins B, C and D from human erythrocyte membranes, Biochim Biophys Acta 1148, 133–138 (1993).

24. J. Saez-Rodriguez, J. C. Costello, S. H. Friend, M. R. Kellen, L. Mangravite, P. Meyer, T. Norman, G. Stolovitzky, Crowdsourcing biomedical research: Leveraging communities as innovation enginesNat. Rev. Genet. 17, 470–486 (2016).
25. S. Whalen, O. P. Pandey, G. Pandey, Predicting protein function and other biomedical characteristics with heterogeneous ensembles, Methods 93, 92–102 (2016).
26. A. N. A. STANESCU, G. PANDEY, in Biocomputing 2017, pp. 288–299.
27. S. K. Sieberts, F. Zhu, J. García-García, E. Stahl, A. Pratap, G. Pandey, D. Pappas, D. Aguilar, B. Anton, J. Bonet, R. Eksi, O. Fornés, E. Guney, H. Li, M. A. Marín, B. Panwar, J. Planas-Iglesias, D. Poglayen, J. Cui, A. O. Falcao, C. Suver, B. Hoff, V. S. K. Balagurusamy, D. Dillenberger, E. C. Neto, T. Norman, T. Aittokallio, M. Ammad-Ud-Din, C. A. Azencott, V. Bellón, V. Boeva, K. Bunte, H. Chheda, L. Cheng, J. Corander, M. Dumontier, A. Goldenberg, P. Gopalacharyulu, M. Hajiloo, D. Hidru, A. Jaiswal, S. Kaski, B. Khalfaoui, S. A. Khan, E. R. Kramer, P. Marttinen, A. M. Mezlini, B. Molparia, M. Pirinen, J. Saarela, M. Samwald, V. Stoven, H. Tang, J. Tang, A. Torkamani, J. P. Vert, B. Wang, T. Wang, K. Wennerberg, N. E. Wineinger, G. Xiao, Y. Xie, R. Yeung, X. Zhan, C. Zhao, J. Greenberg, J. Kremer, K. Michaud, A. Barton, M. Coenen, X. Mariette, C. Miceli, N. Shadick, M. Weinblatt, N. De Vries, P. P. Tak, D. Gerlag, T. W. J. Huizinga, F. Kurreeman, C. F. Allaart, S. Louis Bridges, L. Criswell, L. Moreland, L. Klareskog, S. Saevarsdottir, L. Padyukov, P. K. Gregersen, S. Friend, R. Plenge, G. Stolovitzky, B. Oliva, Y. Guan, L. M. Mangravite, Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis, Nat. Commun. 7 (2016).
28. M. E. Ahsen, R. Vogel, G. Stolovitzky, Unsupervised Evaluation and Weighted Aggregation of Ranked Predictions, (2018) (available at http://arxiv.org/abs/1802.04684).
29. L. I. Kuncheva, C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. 51, 181–207 (2003).
30. M. M. Banoei, H. J. Vogel, A. M. Weljie, A. Kumar, S. Yende, D. C. Angus, B. W. Winston, J. Batt, J. Hutchison, A. Fox-Robichard, P. Liaw, J. F. Cailhier, E. Charbonney, Plasma metabolomics for the diagnosis and prognosis of H1N1 influenza pneumonia, Crit. Care 21 (2017).
31. T. W. Burke, R. Henao, E. Soderblom, E. L. Tsalik, J. W. Thompson, M. T. McClain, M. Nichols, B. P. Nicholson, T. Veldman, J. E. Lucas, M. A. Moseley, R. B. Turner, R. Lambkin-Williams, A. O. Hero, C. W. Woods, G. S. Ginsburg, Nasopharyngeal Protein Biomarkers of Acute Respiratory Virus Infection, EBioMedicine 17, 172–181 (2017).
32. M. T. Mcclain, R. Henao, J. Williams, B. Nicholson, T. Veldman, L. Hudson, E. L. Tsalik, R. Lambkin-Williams, A. Gilbert, A. Mann, G. S. Ginsburg, C. W. Woods, Differential evolution of

peripheral cytokine levels in symptomatic and asymptomatic responses to experimental influenza virus challenge, Clin. Exp. Immunol. 183, 441–451 (2015).
33. M. Chen, A. Zaas, C. Woods, G. S. Ginsburg, J. Lucas, D. Dunson, L. Carin, Predicting viral infection from high-dimensional biomarker trajectories, J. Am. Stat. Assoc. 106, 1259–1279 (2011).
34. A. K. Zaas, T. Burke, M. Chen, M. McClain, B. Nicholson, T. Veldman, E. L. Tsalik, V. Fowler, E. P. Rivers, R. Otero, S. F. Kingsmore, D. Voora, J. Lucas, A. O. Hero, L. Carin, C. W. Woods, G. S. Ginsburg, A host-based RT-PCR gene expression signature to identify acute respiratory viral infection, Sci. Transl. Med. 5 (2013), doi:10.1126/scitranslmed.3006280.
35. D. Proud, R. B. Turner, B. Winther, S. Wiehler, J. P. Tiesman, T. D. Reichling, K. D. Juhlin, A. W. Fulmer, B. Y. Ho, A. A. Walanski, C. L. Poore, H. Mizoguchi, L. Jump, M. L. Moore, C. K. Zukowski, J. W. Clymer, Gene expression profiles during in vivo human rhinovirus infection insights into the host response, Am. J. Respir. Crit. Care Med. 178, 962–968 (2008).
36. B. Chen, M. Chen, J. Paisley, A. Zaas, C. Woods, G. S. Ginsburg, A. Hero, J. Lucas, D. Dunson, L. Carin, Bayesian inference of the number of factors in gene-expression analysis: Application to human virus challenge studies, BMC Bioinformatics 11 (2010).
37. J. Muller, E. Parizotto, R. Antrobus, J. Francis, C. Bunce, A. Stranks, M. Nichols, M. McClain, A. V. S. Hill, A. Ramasamy, S. C. Gilbert, Development of an objective gene expression panel as an alternative to self-reported symptom scores in human influenza challenge trials, J. Transl. Med. 15 (2017).
38. E. E. Davenport, R. D. Antrobus, P. J. Lillie, S. Gilbert, J. C. Knight, Transcriptomic profiling facilitates classification of response to influenza challenge, J. Mol. Med. 93, 105–114 (2015).
39. H. Drakesmith, A. Prentice, Viral infection and iron metabolism, Nat. Rev. Microbiol. 6, 541–552 (2008).
40. E. D. Weinberg, Roles of metallic ions in host-parasite interactions, Bacteriol. Rev. 30, 136–151 (1966).
41. E. D. Weinberg, Iron and Susceptibility to Infectious Disease, Science 184, 952–956 (1974).
42. F. A. D. T. G. Wagener, A. Eggert, O. C. Boerman, W. J. G. Oyen, A. Verhofstad, N. G. Abraham, G. Adema, Y. Van Kooyk, T. De Witte, C. G. Figdor, Heme is a potent inducer of inflammation in mice and is counteracted by heme oxygenase, Blood 98, 1802–1811 (2001).
43. N. W. Cummins, E. A. Weaver, S. M. May, A. J. Croatt, O. Foreman, R. B. Kennedy, G. A. Poland, M. A. Barry, K. A. Nath, A. D. Badley, Heme oxygenase-1 regulates the immune response to influenza virus infection and vaccination in aged mice., FASEB J. 26, 2911–8 (2012).

44. X. Qi, H. Zhang, T. Xue, B. Yang, M. Deng, J. Wang, Down-regulation of cellular protein heme oxygenase-1 inhibits proliferation of avian influenza virus H9N2 in chicken oviduct epithelial cells, J. Gen. Virol. 99, 36–43 (2018).
45. N. G. Abraham, Y. Lavrovsky, M. L. Schwartzman, R. A. Stoltz, R. D. Levere, M. E. Gerritsen, S. Shibahara, A. Kappas, Transfection of the human heme oxygenase gene into rabbit coronary microvessel endothelial cells: protective effect against heme and hemoglobin toxicity., Proc. Natl. Acad. Sci. U. S. A. 92, 6798–802 (1995).
46. H.-C. S. P. Team, H.-I. Consortium, Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses, Sci. Immunol. 2 (2017).
47. E. Bilal, J. Dutkowski, J. Guinney, I. S. Jang, B. A. Logsdon, G. Pandey, B. A. Sauerwine, Y. Shimoni, H. K. Moen Vollan, B. H. Mecham, O. M. Rueda, J. Tost, C. Curtis, M. J. Alvarez, V. N. Kristensen, S. Aparicio, A. L. Børresen-Dale, C. Caldas, A. Califano, S. H. Friend, T. Ideker, E. E. Schadt, G. A. Stolovitzky, A. A. Margolin, Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling, PLoS Comput. Biol. 9 (2013).
48. M. Wang, Y. Zhao, B. Zhang, Efficient Test and Visualization of Multi-Set Intersections, Sci. Rep. 5, 16923 (2015).
49. B. Wilkinson, A statistical consideration in psychological research, Psychol. Bull. 48, 156–158 (1951).
50. D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, Q. Morris, The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function, Nucleic Acids Res. 38 (2010).
51. H. Akaike, Information theory and an extensión of the maximum likelihood principle, Int. Symp. Inf. theory , 267–281 (1973).
52. D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y.

Xie, C. Young, J. Zinman, V. E. Johnson, Redefine statistical significance, Nat. Hum. Behav. (2017).

Acknowledgements

This work was supported by Defense Advanced Research Projects Agency and the Army Research Office through Grant W911NF-15-1-0107. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. JGB was supported by a training grant from the National Institutes of Health, USA (NIH grant 4T15LM007088-25).GP and AS's work was supported by NIH grant # R01GM114434 and an IBM faculty award to GP. TEMN was supported by the Ministry of Science and Technology of Taiwan grant 105-2218-E-006-016-MY2. KYY was supported by NIH grants U54 HL127624 and R01GM126019. MS was supported by Grants-in-Aid for Scientific Research JP16H02866 from the Japan Society for the Promotion of Science.

We wish to thank Rafick P. Sekaly (Case Western Reserve University) for his critical feedback during the writing process.

Author Contributions

RH, CC, MTM, CWW, GSG, and ELT devised and performed the viral exposure experiments. RH, GSG, ELT, LM and SKS designed and ran the data analysis challenge. SF, AT, MM, JGB, RK, ZA, KYY, RA, SJ, XL, TEMN, MS, LLE, and The Respiratory Viral DREAM Challenge Consortium members participated in the Challenge and SF, AT, MM, JGB, RK, RH, ZA, KYY, MEA, RA, SJ, XL, TEMN, MS, AS, RV, GP, LLE, SKS analyzed the data.

Data Availability

Data are available through GEO GSE73072. Challenge results and methods and code for individual models are available through www.synapse.org/ViralChallenge.

Materials & Correspondence

Correspondence should be addressed to Dr. Sieberts: solly.sieberts@sagebase.org.

Figures

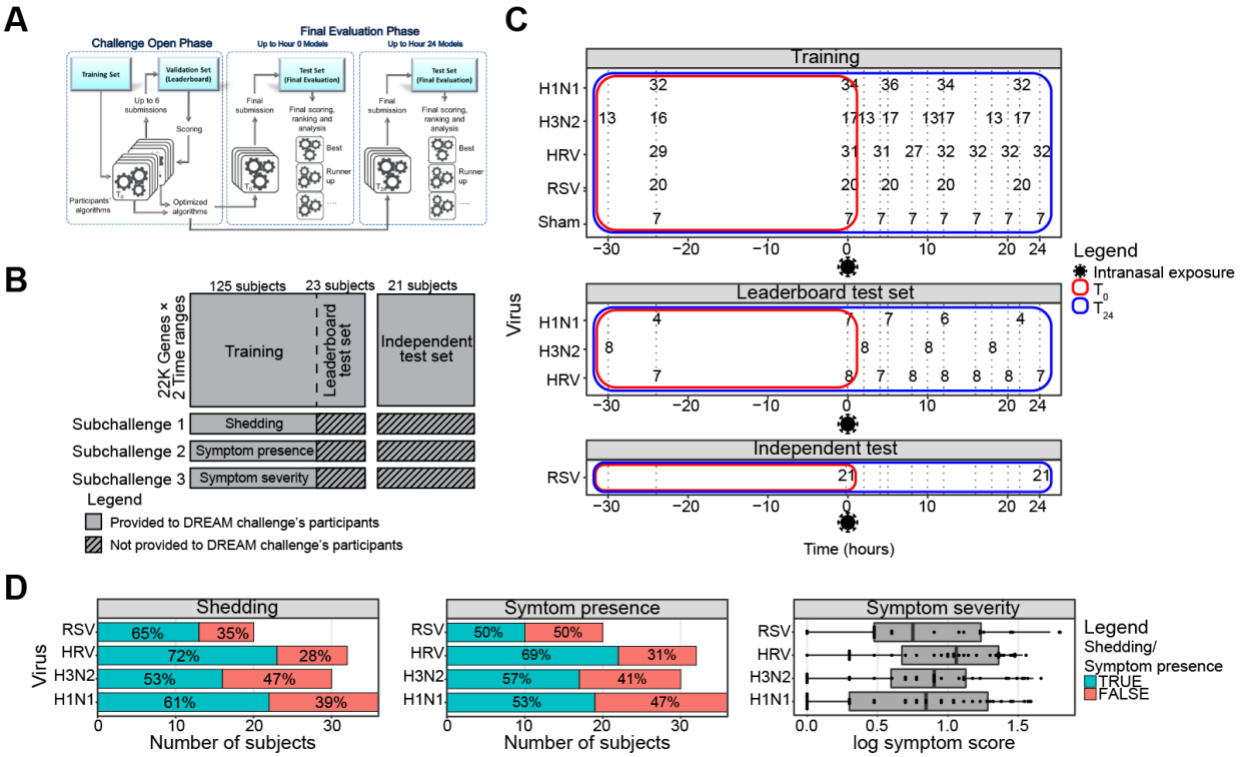


Figure A3.1. Respiratory Viral DREAM Challenge

(**A**) Schematic representation of the *Respiratory Viral DREAM Challenge* workflow. Participants used feedback from evaluation on the leaderboard test set to optimize their T_0 and T_{24} models, and submitted a single model, per timepoint, for final evaluation on the Independent Test Set. (**B**) Schematic representing the data provided to participants. 125 subjects were provided as training data, 23 subjects were provided as a leaderboard test set, and 21 subjects from an independent data set were used for final evaluation. (**C**) Challenge data come from seven viral exposure trials with sham or one of 4 different respiratory viruses (H1N1, H3N2, Rhinovirus, and RSV). In each of these trials, healthy volunteers were followed for seven to nine days following controlled nasal exposure to one respiratory virus. Blood was collected and gene expression of peripheral blood was performed 1 day (24 to 30 hours) prior to exposure, immediately prior to exposure and at regular intervals following exposure. Data were split into a training, leaderboard, and independent test set. Outcome data for the leaderboard and independent test set were not provided to the teams, but instead teams were asked to predict them based on gene-expression pre-exposure (T_0) or up to 24 hours post-exposure (T_{24}). (**D**) Symptom data and nasal lavage samples were collected from each subject on a repeated basis over the course of 7-9 days. Viral infection was quantified by measuring release of viral particles from viral culture or by qRT-PCR ("viral shedding"). Symptomatic data were collected through self-report on a repeated basis. Symptoms

were quantified using a modified Jackson score, which assessed the severity of 8 upper respiratory symptoms (runny nose, cough, headache, malaise, myalgia, sneeze, sore throat and stuffy nose).

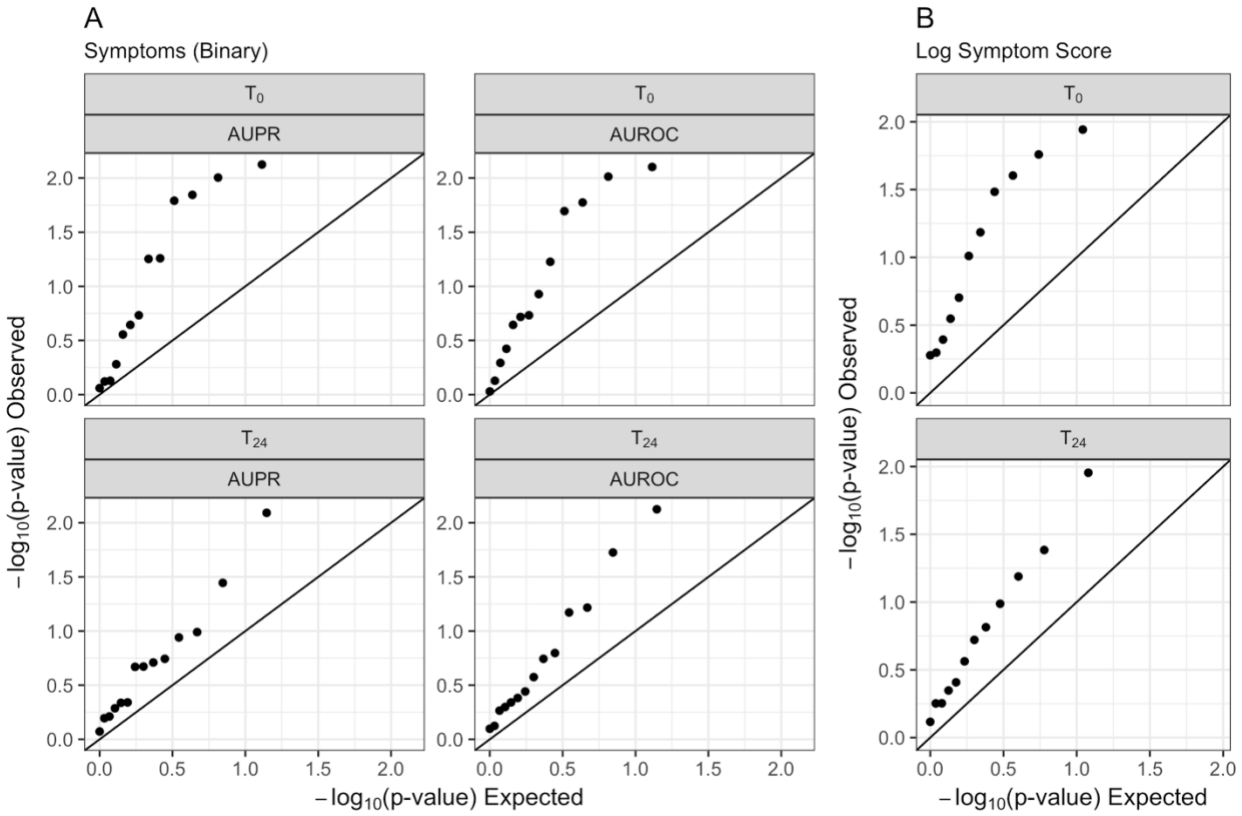


Figure A3.2. Models predict presence of symptoms and symptom severity better than expected at random

Observed -$\log_{10}$(p-value) versus the null expectation for submitted predictions for predicting (**A**) presence of symptoms (SC2) and (**B**) log symptom score (SC3). For both subchallenges significant enrichment of p-values (enrichment p-value 0.008, 0.002, 0.021, and 0.05 for AUPR(T_0), AUROC(T_0), AUPR(T_{24}), and AUROC(T_{24}), respectively, for presence of symptoms, and enrichment p-value 0.005 and 0.035 for T_0 and T_{24}, respectively, for log symptom score) across submissions demonstrates that pre- and early post-exposure transcriptomic data can predict susceptibility to respiratory viruses.

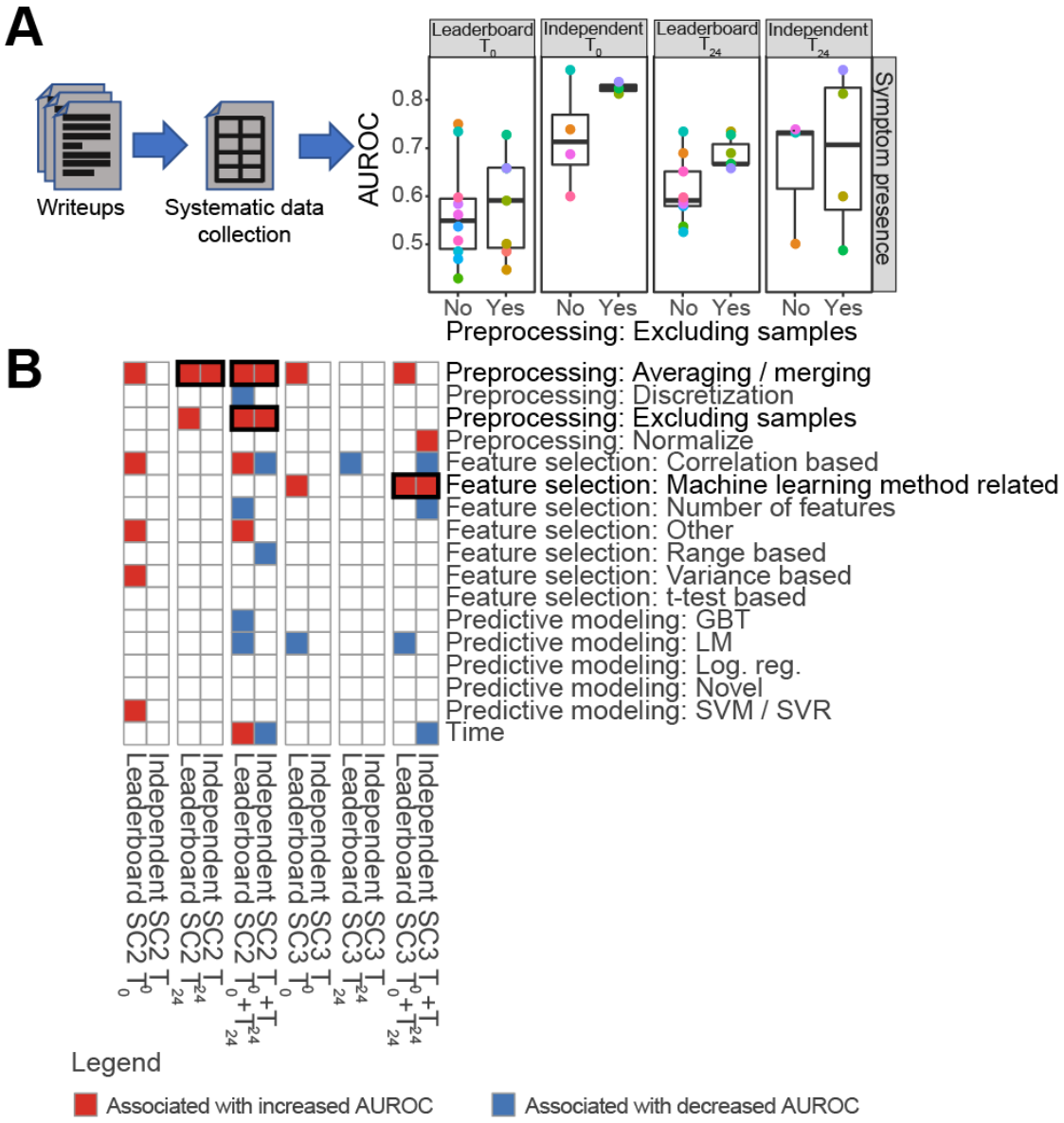


Figure A3.3. Adequate preprocessing leads to more accurate predictors of symptoms presence and severity

(**A**) Schematic representation of the analysis of the participating teams' writeups to identify methodological steps associated with more accurate prediction of symptoms. First, the writeups were manually inspected to identify the preprocessing, feature selection and predictive modeling method used by each team. Second, the methods were regrouped into general categories across teams. Third, each general method was assessed for its association with predictive model accuracies on the leaderboard test set and the independent test set. (**B**) Heatmap showing the association of each general method with prediction ability (i.e. AUROC for subchallenge 2 (prediction of symptom presence; SC2) and Pearson's correlation coefficient for subchallenge 3

(prediction of symptom severity; SC3)). For each general method, a Wilcoxon rank-sum test was used to assess the association between using the method (coded as a binary variable) and prediction ability.

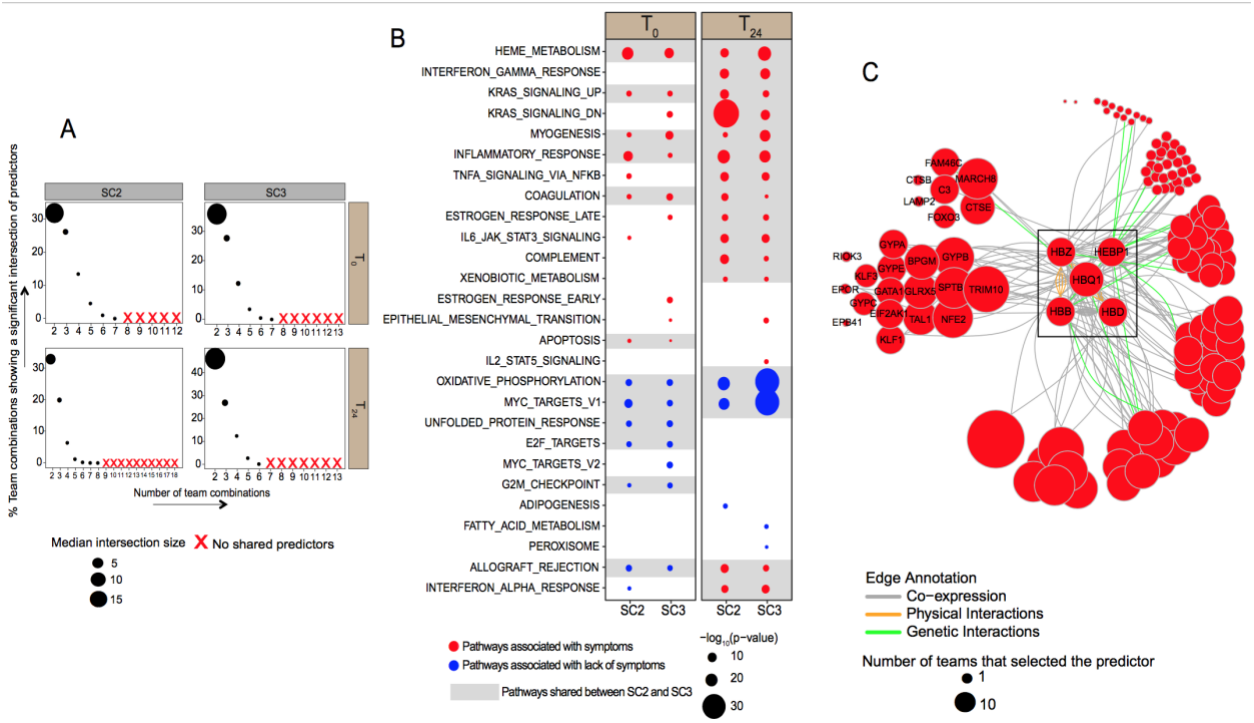
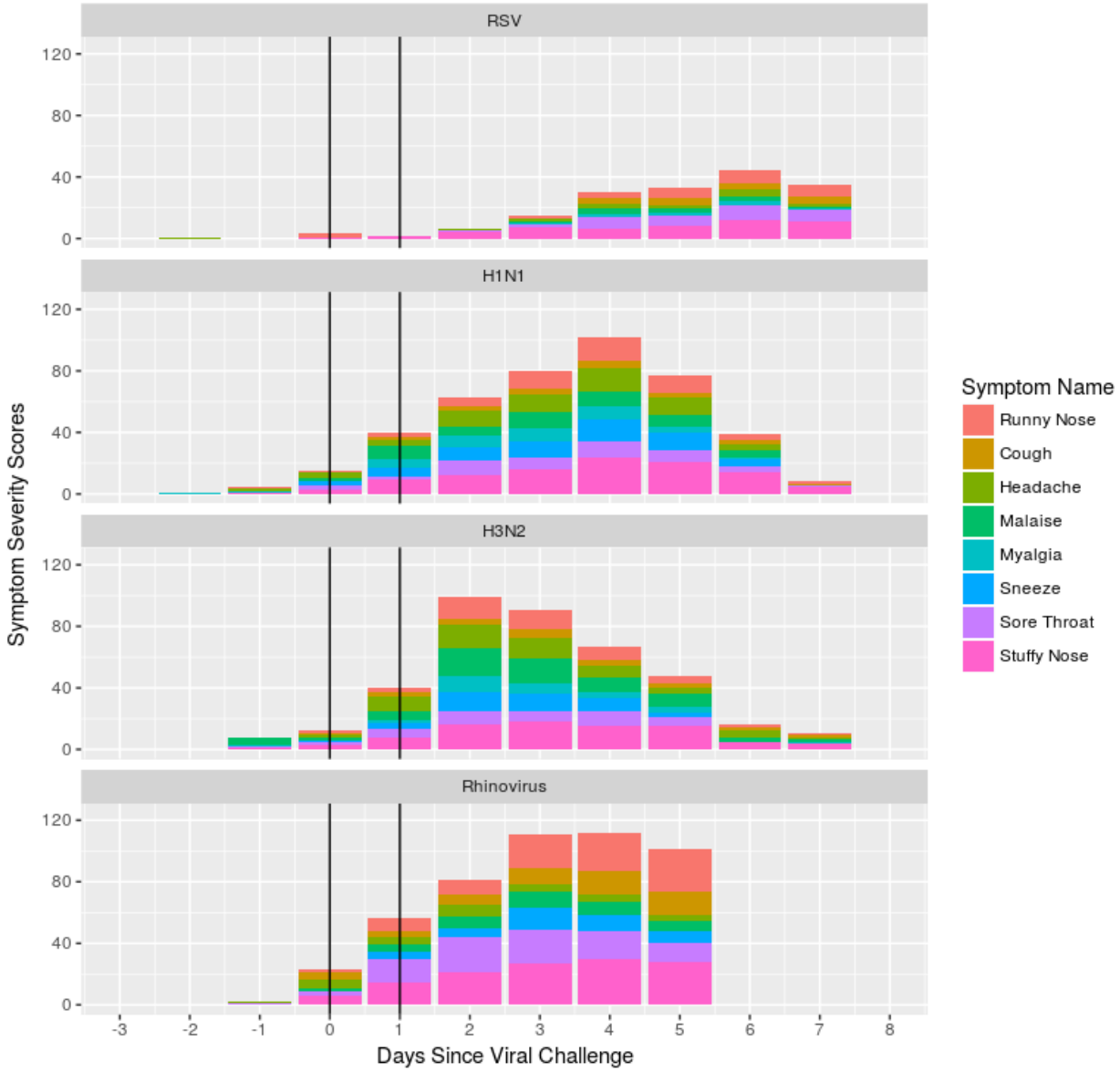


Figure A3.4. Overlap and pathway enrichment among predictors of symptoms

(**A**) Percent of team combinations showing statistically significant intersections of predictors at T_0 and T_{24}. Only teams whose with AUROC ≥ 0.5 or r ≥ 0 for subchallenge 2 and 3, respectively, were used for this analysis. The x-axis indicates the number of teams included in the combination. For example, the value 2 corresponds to pairwise overlaps, 3 corresponds to 3-way overlaps, etc. The y-axis indicates the percentage of team combinations with a statistically significant (p-value < 0.05) predictor intersection. Point size indicates median intersection size of predictors among team combinations with significant predictor intersection; 'X' indicates no significant predictor intersection. (**B**) Pathway enrichment among predictors of infection for each subchallenge (SC2 and SC3) at T_0 and T_{24}. The x-axis indicates subchallenge and each grid indicates timepoint. The y-axis indicates pathways enriched among predictors with a Benjamini-Hochberg corrected p-value < 0.05. Point size represents the fisher's exact test enrichment $-\log_{10}$(p-value). Point colors indicate whether the pathway was associated with symptoms (red) or lack thereof (blue). Pathways shared between both SC2 and SC3 at each timepoint are highlighted in grey. Pathways are ordered by the decreasing maxP test statistic as determined in Fig S5 (**C**) GeneMANIA network of the union of predictors involved in the Heme metabolism pathway across time points

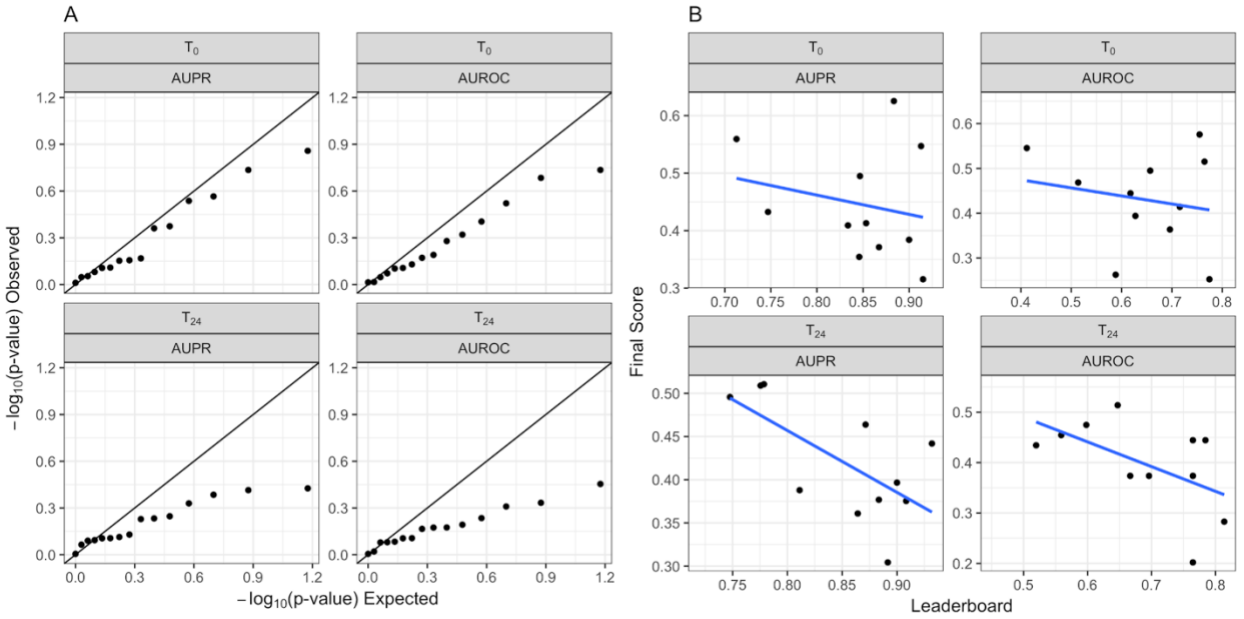
(T_0 and T_{24}) and subchallenges (SC2 and SC3). Edges are inferred by GeneMANIA *(50)* corresponding to co-expression (purple), physical interactions (orange) and genetic interactions (green) among genes. Node size corresponds to the number of teams that selected the predictor.

Supplemental Data



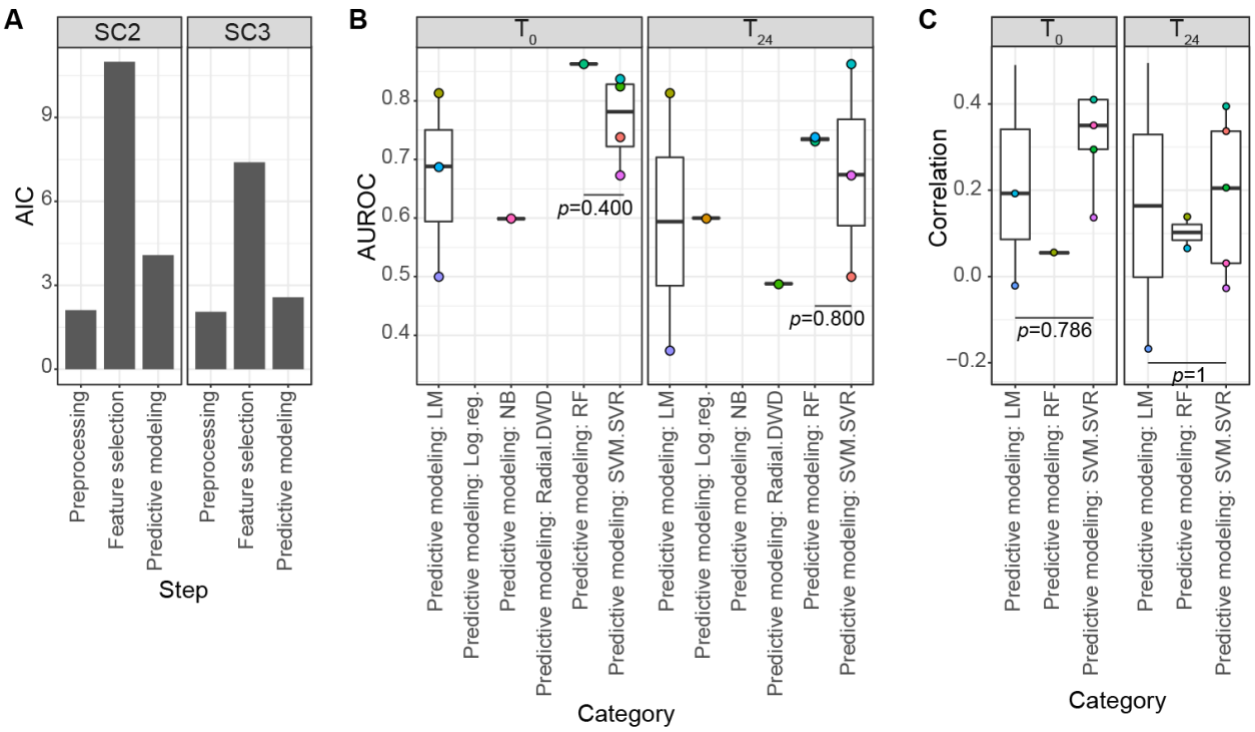
Supplementary Figure A3.1. Total aggregated symptom load by virus (RSV, H1N1, H3N2, Rhinovirus)

While self-reported symptom distributions differ across the different viruses, in each case peak symptoms occur at least one day after the latest time point examined in this study (24 hours post-viral exposure).



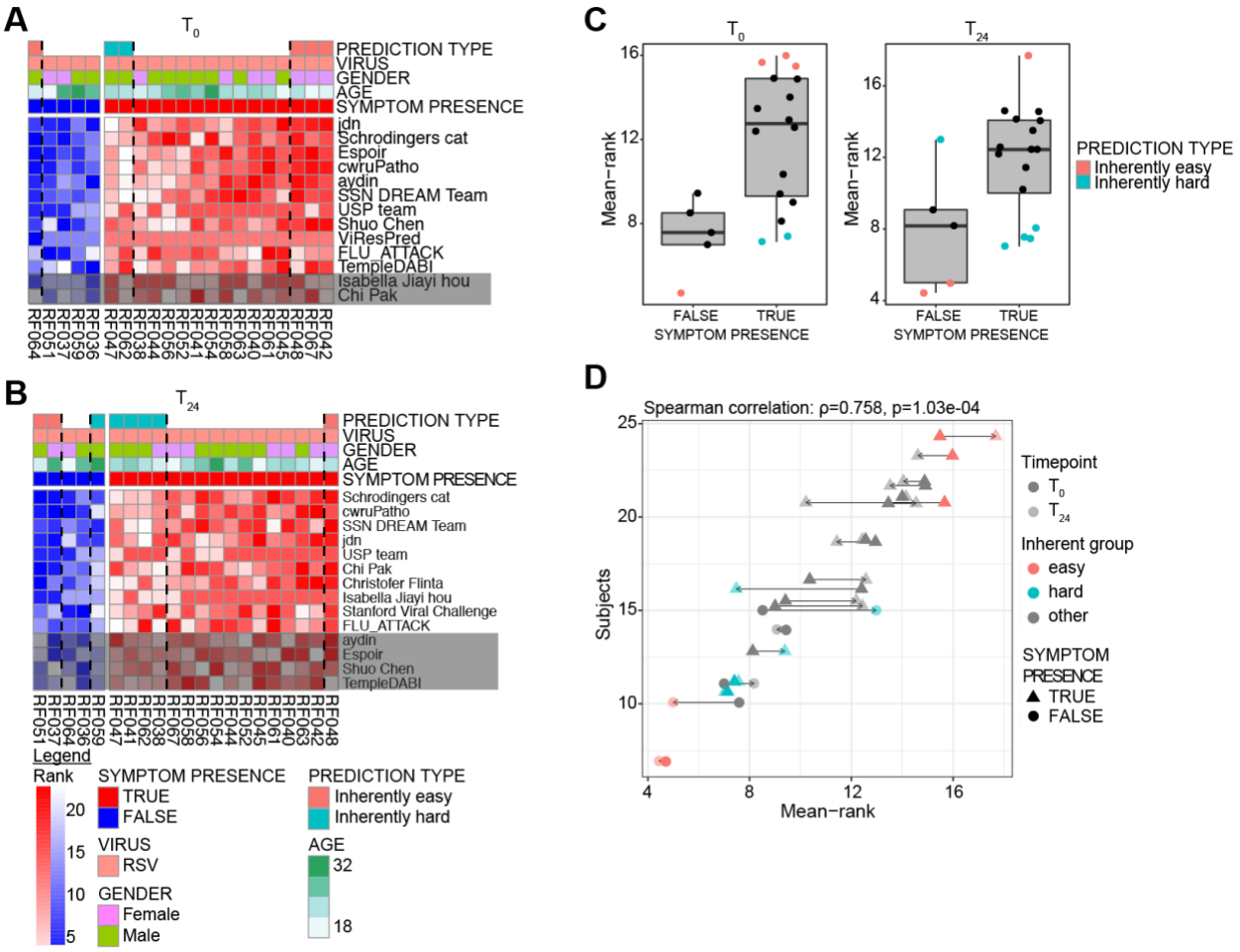
Supplementary Figure A3.2. Models show inability to predict viral shedding

(**A**) Observed -$\log_{10}$(p-value) versus the null expectation for submitted predictions for classifying viral shedding (SC1) demonstrates a lack of enrichment (enrichment p-values 0.94, 0.95, 0.82 and 0.95, for AUPR(T_0), AUROC(T_0), AUPR(T_{24}) and AUROC(T_{24}), respectively). (**B**) Correlations between scores from the leaderboard test set and independent test set for SC1 are negative (correlations -0.22, -0.19, -0.65, and -0.54 for AUPR(T_0), AUROC(T_0), AUPR(T_{24}) and AUROC(T_{24}), respectively), suggesting overfitting of the training and leaderboard data.



Supplementary Figure A3.3. Preprocessing and predictive modeling approaches leading to better predictive ability

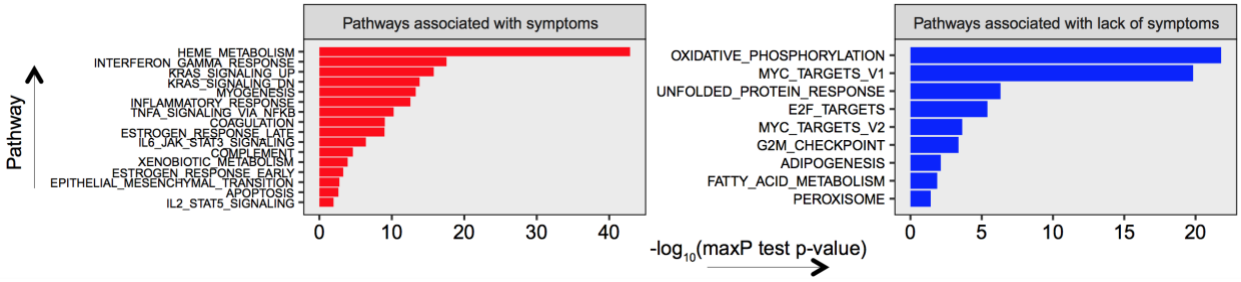
(**A**) Akaike information criterion (AIC) *(51)*, an estimate of the relative information loss under a predictive model (the smaller the AIC values the better is the trade-off between the goodness of fit and the simplicity of a model) for models representing each of the three major steps in predictive model building. Analysis was performed separately for SC2 and SC3. (**B**) Area under a ROC curve (AUROC) as function of predictive modeling method used to build predictive models of presence of symptoms (SC2). A Wilcoxon rank-sum test was used to assess the variation of prediction ability across the methods. LM: linear least square regression model; Log. reg.: logistic regression; NB: naive bayes; Radial DWD: Radial distance Weighted Discrimination; RF: random forest; SVM: support vector machine; SVR: support vector regression. (**C**) Pearson correlation (Correlation) as function of predictive modeling approaches used to build predictive models of symptoms severity (SC3). A Wilcoxon rank-sum test was used to assess the variation of prediction ability across the methods. On the boxplots (**B-C**), the lower whisker, the lower hinge, the mid hinge, the upper hinge and the upper whisker correspond to the $-1.5 \times$ the interquartile (IQR) from the 1st quartile, the 1st quartile, the median, the 3rd quartile and $1.5 \times$ IQR from the 3rd quartile of the AUROC/Pearson Correlation, respectively



Supplementary Figure A3.4. Subjects inherently difficult to predict both at T_0 and T_{24}

Heatmaps of the predictions of the symptomatic score classifiers for the Independent test set. Predictions were either binary outcome (for 3 teams at T_0 and 2 teams at T_{24}) or continuous probability (for 10 teams at T_0 and 12 teams at T_{24}). Predictions were transformed to ranks in order to be comparable across teams. (**A**) T_0 subjects and (**B**) T_{24} subjects were ordered left to right from the subject predicted by the teams as asymptomatic (*i.e.* symptom presence = FALSE) to the subject predicted by the teams as symptomatic (*i.e.* symptom presence = TRUE). Teams with AUROC < 0.5 (highlighted in grey) were not used for the ordering of the subjects. (**C**) Prediction type groups were identified by investigating the distribution of the mean-rank. Inherently hard subjects that presented symptoms were defined as having mean-rank below the median of the mean-rank of all asymptomatic subjects. Similarly, inherently hard subjects that did not presented symptoms had mean-rank above the median of symptomatic subjects. Inherently easy asymptomatic subjects were defined by having mean-rank strictly below the 1st quartile of asymptomatic subjects while inherently easy symptomatic subjects had mean-rank had mean-rank strictly above the 3rd quartile of the mean-ranks of symptomatic subjects. On the boxplot,

the lower whisker, the lower hinge, the mid hinge, the upper hinge and the upper whisker correspond to the $-1.5 \times$ the interquartile (IQR) from the 1st quartile, the 1st quartile, the median, the 3rd quartile and $1.5 \times$ IQR from the 3rd quartile of the mean-rank respectively (**D**) Scatter plot of the average prediction by the teams (x-axis) for each subject (y-axis) by timepoint. Lines connect subjects from the same subjects. Subjects are colored by their prediction type group. Spearman's correlation coefficient and *t*-test were used to evaluate the correlation between T_0 predictions and T_{24} predictions.



Supplementary Figure A3.5. Heme metabolism best predicts symptoms across time points and subchallenges

Pathways associated with symptoms and lack of symptoms across time points (T_0 and T_{24}) and subchallenges (SC2 and SC3). The pathways that were enriched at each timepoint for each subchallenge at an adjusted p-value < 0.05 was considered. The statistical significance of each pathway was calculated across time points and subchallenges using the maxP test statistic. The x-axis represents the -$\log_{10}$(maxP test p-value) value and the y-axis corresponds to the pathways associated with symptoms (in red) and pathways associated with lack of symptoms (in blue) ordered by the decreasing value of -$\log_{10}$(maxP test p-value).

Supplementary Table A3.1. Methods used by the teams for the predictions of viral shedding and symptoms

			Number of teams reported (n=24)		
Category	Step	Description / Criteria	SC1 (%)	SC2 (%)	SC3 (%)
Excluding subjects	Preprocessing	Exclusion of subjects based on some criteria (SHAM, missing values, etc.)	7 (29)	8 (33)	6 (25)
Normalize	Preprocessing	Use of any specific normalization on the data	14 (58)	12 (50)	11 (46)
Averaging / merging	Preprocessing	Merging of multiple features (or time points) together to generate new features	9 (38)	9 (38)	7 (29)
Discretization	Preprocessing	Division of a continuous attribute into n distinct bins where each bin contains N instances	2 (8)	2 (8)	2 (8)
Machine learning method related	Feature selection	Use of any machine learning-based approach to do the feature selection	13 (54)	12 (50)	11 (46)
Variance based	Feature selection	Filtering out a set of features based on their variance	4 (17)	3 (12)	3 (12)
Correlation based	Feature selection	Filtering out a set of features based on correlation	3 (12)	3 (12)	2 (8)
T-test based	Feature selection	Feature selection based on an approach similar to t-test	3 (12)	3 (12)	2 (8)
Range based	Feature selection	Feature selection based on value range (Defining a cut-off etc.)	3 (12)	4 (17)	3 (12)
Number of features	Feature selection	Number of features used in predictive models	2 (8)	2 (8)	2 (8)
Other	Feature selection	Any other feature that is not explained by the terms above, such as DISR, Feature hashing, etc.	3 (12)	3 (12)	2 (8)
LM	Predictive modeling	Linear model of any form (including "Generalized Linear Model")	6 (25)	5 (21)	5 (21)
Log. reg.	Predictive modeling	Logistic regression	1 (4)	1 (4)	0 (0)
RF	Predictive modeling	Random forest	2 (8)	3 (12)	3 (12)
SVM / SVR	Predictive modeling	Support Vector Machine	9 (38)	6 (25)	7 (29)
NB	Predictive modeling	Naive Bayes	0 (0)	1 (4)	0 (0)
Guass. proc. reg.	Predictive modeling	Gaussian process regression	1 (4)	1 (4)	1 (4)
GBT	Predictive modeling	Gradient Boosted Trees	1 (4)	1 (4)	1 (4)
Radial DWD	Predictive modeling	Radial Distance Weighted Discrimination	0 (0)	1 (4)	0 (0)
Novel	Predictive modeling	Methods that are unconventional and are developed by the group that used them, such as ROSETTA, LIFT, ROAD, etc.	3 (12)	2 (8)	2 (8)

Twenty-four of the thirty four teams participating in the Challenge provided writeups describing the method they used to build their predictive models, which were then classified into methodological category for three processing steps: preprocessing, feature selection, and predictive modeling. The numbers (and proportion) of teams that using each methods are indicated in the table, by subchallenge.

Discussion

1. Utilité clinique des biomarqueurs prédictifs de la réponse aux vaccins

Les trois articles présentés dans cette thèse nous ont permis d'exploiter des données large échelle pour identifier des biomarqueurs de la réponse aux vaccins (1er et 2d article) ou de la sévérité de symptômes d'une infection virale (3e article).

Notre analyse des prédicteurs de la réponse à la vaccination contre l'HepB nous permis d'identifier des différences entre l'âge chronologique et l'âge biologique. Nous avons généré un score composite intégrant les changements transcriptionnels associés à l'âge, appelé BioAge, et nous avons démontré que ce score était associé négativement à la réponse au vaccin contre l'HepB Twinrix. Le BioAge s'est avéré être un meilleur prédicteur de la réponse du vaccin contre l'HepB que l'âge chronologique (AUC du BioAge = 67%; AUC de l'âge=60%). Depuis la publication de ce 1er article, d'autres groupes de recherche ont entamé une validation indépendante de l'association du BioAge avec la réponse immunitaire spécifique à d'autres vaccins contre l'HepB (Engerix-B) et également avec la réponse à d'autres vaccins où l'âge chronologique est associé avec une réponse plus faible aux vaccins (ex. vaccin contre les virus Influenza). Ces exercices permettront de valider le classificateur et également de déterminer si le BioAge peut être utilisé comme prédicteurs de la réponse pour d'autres vaccins.

Typiquement les vaccins contre l'HepB sont administrés en trois doses avec un intervalle d'un mois séparant la 1re dose de la seconde, et 5 mois séparant la 2de dose de la 3e dose (189). Le pourcentage de participants ayant répondu après avoir reçu deux doses du vaccin contre l'HepB est aux alentours de 50% alors qu'après 3 doses le taux de réponse est au-delà de 90% (190). Ceci suggère que les personnes avec un BioAge élevé ayant répondu faiblement aux deux premières doses du vaccin ont de fortes chances de produire des Acs anti-HepB à un niveau protecteur et pourront être considérés comme des répondants au vaccin dès qu'ils auront reçu la troisième dose du vaccin contre l'HepB.

Tous les participants utilisés dans le 1er article de cette thèse sont âgés de 25 ou plus. L'association du BioAge avec la réponse aux vaccins contre l'HepB chez les enfants est inconnue. Au Canada, les récipients majoritaires du vaccin contre l'HepB sont les nouveaux nés/jeunes enfants (191). Les enfants reçoivent les trois doses standard des vaccins contre l'HepB, données dans un intervalle de 6 mois. Il est concevable que le BioAge puisse également être capable de distinguer les enfants ne nécessitant que les deux premières doses du vaccin

des enfants nécessitant une troisième dose. De plus amples recherches cliniques sont nécessaires pour tester cette hypothèse.

Dans le 1er article de cette thèse, nous avons réussi à identifier des prédicteurs de la réponse au vaccin contre l'HepB uniquement un utilisant des données transcriptionnelles, protéiques et cellulaire mesurées prévaccination. En effet, les données obtenues après la vaccination (une semaine après la première dose du vaccin) n'ont produit aucun classificateur significatif (comparer à une séparation aléatoire des participants) de la réponse au vaccin de l'HepB. Plusieurs approches d'apprentissage machine incluant une approche semi-supervisée basée sur les travaux de Pulendran et collab. utilisant des modules de gènes corrégulés dans le sang(192) et des approches supervisées alternatives (régression linéaire, machine à vecteurs de support) n'ont pas réussi à identifier un classificateur pour le vaccin contre l'HepB basé sur les données transcriptionnelles après vaccination. De plus, le fait que deux groupes de bio-informaticiens sont arrivés à la même conclusion de façon indépendante (mon étude et les travaux parallèles de Dr Cristescu, le 2e auteur du 1er article) suggère qu'il est difficile d'obtenir un classificateur basé sur les données post-vaccination sur cette cohorte. De plus, nous avons appliqué la même stratégie sur un jeu indépendant (ImmPort(193) : SDY690) et nous avons observé que les données prévaccination ont une performance comparable que celle observée sur la cohorte du 1er article de cette thèse (**Annexe 1**). Ceci suggère que pour certains vaccins, tels que celui contre l'HepB, la variance observée prévaccination entre individus est plus utile pour prédire la réponse au vaccin que la variance entre individus juste après la vaccination.
Une conclusion identique est rapportée dans le 3e article de cette thèse. Dans ce 3e article, les données transcriptionnelles collectées dans les 24 heures suivant l'infection des individus par un virus respiratoire n'ont pas permis de construire des classificateurs plus robustes ou plus précis que ceux construits à partir des données transcriptionnelles générées sur des cellules de sang périphérique obtenues avant l'infection des volontaires.
Les travaux présentés dans cette thèse ont instigué l'utilisation de données large échelle pour l'identificateur de marqueurs mesurés prévaccination et prédictifs de la réponse aux vaccins. En effet, avant la publication du 1er article de cette thèse, uniquement deux articles avaient décrit des signatures transcriptionnelles capables de prédire la réponse aux vaccins (141, 142). Ces deux études ont eu recours à une méthode de validation croisée pour évaluer la précision de leurs classificateurs et n'ont pas inclus de cohorte test. Dans le 1er article de cette thèse, nous avons utilisé deux approches indépendantes pour identifier des classificateurs de la réponse au vaccin contre l'HepB (une méthode semi-supervisée basée sur l'âge et une méthode supervisée). Les

deux classificateurs présentés dans le 1er article démontrent un chevauchement des gènes utilisés et des voies de signalisation représentées. Ces deux classificateurs ont été testés en utilisant un groupe de participants, séparé au début de l'étude, pour déterminer la prédiction ces classificateurs. L'apport méthodologique et la découverte de marqueurs prédictifs de la réponse au vaccin contre l'HepB font de ce 1er article un travail pionnier cité plus de 75 fois dans la littérature scientifique.

Toutefois, l'importance des données prévaccination n'est pas généralisable à toutes les perturbations du système immunitaire (vaccin ou infection). En effet, dans le 1er article présenté dans cette thèse, tous les participants dans l'étude ont reçu le vaccin contre l'HepB, mais également des vaccins contre la diphtérie, le tétanos et le choléra. Pour ces trois autres vaccins les données prévaccination n'ont pas produit des classificateurs significatifs (i.e. permettant une prédiction meilleure qu'une partition aléatoire des participants en forts et faibles répondants) de la réponse à ces vaccins. Les participants inclus dans cette étude n'ont pas reçu au préalable le vaccin contre l'HepB alors que plusieurs d'entre eux ont reçu dans le passé les vaccins contre la diphtérie, le tétanos et le choléra. Le même résultat (c.-à-d. l'inhabilité de produire des classificateurs significatifs prévaccination pour les vaccins contre la diphtérie, le tétanos et le choléra) est obtenu en excluant les participants présentant des titres d'Acs prévaccination contre la diphtérie, le tétanos et le choléra. Il est possible que d'autres facteurs prévaccination génétique ou environnementaux que l'on n'a pas mesuré dans cette étude soient responsable de l'hétérogénéité des profils d'expression que l'on observe prévaccination respectivement dans les de participants répondants et non-répondants aux vaccins contre la diphtérie, le tétanos et le choléra. Cependant, les données collectées 7 jours après la vaccination semblent plus à même de prédire la réponse à ces trois vaccins (ces résultats vont faire partie d'un article couramment en préparation) (**Annexe 2**). Ceci suggère qu'il n'est pas toujours possible d'identifier des marqueurs prévaccination prédictifs de la réponse à un vaccin, et que les marqueurs identifiés dans le 1er article ne sont pas universellement prédictifs de la réponse aux vaccins. Des études supplémentaires ont été effectuer pour déterminer si de tels prédicteurs pourraient être identifiés et ont démontré qu'il est essentiel d'effectuer une méta-analyse incluant plusieurs (13) vaccins distincts et une large cohorte pour chacun des vaccins pour pouvoir identifier de tels biomarqueurs (194).

Le BioAge et le classificateur naïf bayésien détaillés dans le 1er article et les deux meilleurs classificateurs décrits dans le 3e article (machine à vecteurs de support et le modèle de régression linéaire clairsemée) pour prédire, respectivement, la réponse aux vaccins et la sévérité des

symptômes d'une infection virale possèdent des précisions similaires aux autres classificateurs décrits dans la littérature scientifique (cf. Table 2 de l'introduction de cette thèse). Ces précisions ne sont toutefois pas assez bonnes pour que ces classificateurs puissent être considérés des outils cliniques viables, car ils font en moyenne 20% à 30% d'erreurs. Dû au faible coût des vaccins et le fait que leurs effets secondaires néfastes sont rares, il est possible qu'un classificateur robuste avec une précision quasi parfaite pourrait être envisagé comme viable cliniquement (195, 196).

Une autre utilisation clinique des biomarqueurs issus de l'analyse de données large échelle est d'identifier des corrélats de protection pouvant être utilisée comme substitut de la protection contre une infection virale. Ces corrélats de protection mesurés juste après la vaccination peuvent prédire une protection contre une infection virale pouvant subvenir quelques jours à plusieurs semaines/mois après la vaccination. De tels corrélats peuvent notamment être utilisés pour accélérer la recherche clinique, en évitant d'effectuer un long et coûteux suivit des participants dans un essai clinique pour un vaccin. Toutefois il est nécessaire d'être prudent sur la validité de ces marqueurs. Un exemple de corrélat de protection identifié en utilisant des données large échelle est les niveaux des Acs IgG reconnaissant la boucle V1/V2 du VIH associés à la réponse du vaccin RV144 (103). Ce corrélat a été utilisé comme critère pour évaluer le succès des essais cliniques de phase 2 et comme aval aux essais de phase 3 du vaccin ALVAC-HIV+protéine en Afrique du Sud (essai clinique HVTN702 testant un vaccin similaire au vaccin RV144) (101). Toutefois, cet essai clinique de phase 3 a échoué à démontrer l'efficacité du vaccin ALVAC-HIV+protéine à protéger contre l'infection par le VIH. Ceci met en doute la validité du marqueur IgG reconnaissant la boucle V1/V2 comme un corrélat robuste de la réponse du vaccin RV144. Alternativement, d'autres marqueurs cliniques et facteurs expérimentaux comme le sous-type de VIH circulant (clade B versus clade A/E), la formulation du vaccin ou l'ethnicité (Asiatique pour RV144 et Africaine pour HVTN702) pourrait également être responsable des résultats contradictoires de ces deux essais cliniques même si les titres d'Acs ne semblent pas être affectés par ces facteurs. Ceci suggère que les Acs ne sont pas des bons corrélats de protection et justifie notre approche focalisée sur l'utilisation de données à large échelle pour définir des corrélats de protection prédictifs ayant une précision accrue.

Une importante assomption et implication de la recherche de signatures prévaccination prédictives de la réponse aux vaccins et qui identifient des voies de signalisation ou des cellules du système immunitaire associées à ces réponses est qu'une intervention médicamenteuse

prévaccination pourrait potentiellement améliorer les résultats de la vaccination. Il existe des preuves (préliminaires ou indirectes) supportant cette hypothèse (157, 197). Si l'on peut moduler les facteurs associés à une réponse sous optimale au vaccin par le biais d'une intervention ciblant les cellules ou les voies de signalisation identifiées cela pourrait permettre l'implémentation de stratégies de vaccination qui peuvent conduire à une réponse immunitaire plus efficace, moins associée à des effets secondaires néfastes et qui permettrait une protection accrue. De telles approches permettraient d'augmenter l'efficacité des vaccins. Cela pourrait également permettre d'implémenter des stratégies permettant l'administration d'un nombre réduit de doses de vaccin (idéalement une seule dose), diminuant par le fait le coût des vaccins et le risque d'engendrer des effets secondaires néfastes (198). De plus, de nombreuses interventions autorisées (ex. médicaments, adjuvants, produits biologiques) connues pour avoir des fonctions immunomodulatrices pourraient être potentiellement réutilisées pour modifier l'état de base du système immunitaire de manière ciblée. Grâce aux technologies large échelle et la modélisation informatique; il est possible de déterminer, par exemple, les modulateurs immunitaires prévaccination qui pourraient être administré, de les personnaliser pour un individu et un type de vaccin.

Dans le 1er article, nous avons démontré qu'un faible BioAge (jeune adulte) était caractérisé par une expression accrue des gènes impliqués dans l'activation des cellules T et B, tandis qu'un BioAge élevé (personne âgée) était caractérisé par la forte expression de gènes pro-inflammatoires et d'interféron de type II. L'association d'une inflammation prévaccination avec une faible réponse aux vaccins a aussi été démontrée pour d'autres vaccins comme les vaccins contre l'Influenza (139). Bloquer l'inflammation prévaccination par l'utilisation de faible concentration de l'immunosuppresseur silorimus (rapamycine) a résulté en une réponse humorale plus forte suivant la vaccination contre l'influenza suggérant que les marqueurs prévaccination prédictifs peuvent être utilisé pour personnaliser le régiment de vaccination d'un individu (197).

En résumé, les biomarqueurs identifiés dans les trois articles présentés dans cette thèse n'ont pas une précision assez élevée pour être considérés comme des outils cliniques à part entière. Toutefois les informations biologiques qui en découlent peuvent potentiellement être utilisées en clinique pour personnaliser la formulation des vaccins reçus par une personne.

2. Apports mécanistiques des données large échelle

Les trois articles présentés dans cette thèse illustrent comment des données large échelle peuvent être utilisées pour émettre des hypothèses sur les mécanismes régissant la réponse aux

vaccins. Dans ces trois articles, une analyse bio-informatique d'enrichissement des voies de signalisation a permis d'identifier les voies de signalisation affectée par une perturbation d'un système immunitaire (à la suite de la vaccination dans les articles 1 et 2 ou une infection virale dans l'article 3). Une limite de ces méthodes d'enrichissement de terme biologique est qu'ils dépendent de la qualité des bases de données publiées de voies de signalisation et ne sont pas capables d'identifier de nouveaux éléments impliqués dans une voie de signalisation. Pour pallier ce problème, nous avons utilisé des méthodes de corrélation et entrepris la construction de réseaux en utilisant une approche d'intégration multiomique pour (1) déterminer si la fonction de voies de signalisations spécifiques peut être associées à des modes distincts d'expression (ARN messager, protéine) et (2) identifier de nouveaux éléments impliqués dans une voie de signalisation.

Dans le 1er article de cette thèse, les titres d'Acs spécifiques de l'Ag de surface du virus de l'HepB sont les corrélats établis de l'efficacité du vaccin contre l'HepB, ce qui implique que la présence d'une sous-population spécifique de lymphocytes B avant la vaccination qui permettrait le développement d'une réponse optimale au vaccin. Les résultats de l'analyse intégrative multiomique présentés dans le 1er article confirment cette hypothèse. Premièrement, les résultats de la cytométrie en flux démontrent que le pourcentage d'une sous-population de lymphocytes B mémoire (% IgG+ cellules B mémoire) est un prédicteur univarié significatif de la réponse au vaccin contre l'HepB, bien que le nombre total absolu de lymphocytes B ne fût pas un prédicteur significatif de la réponse au vaccin contre l'HepB. L'analyse des données transcriptionnelles démontre que plusieurs marqueurs des lymphocytes B, y compris des composants du complexe BCR (IGH, CD79A, CD79B et CD19), ont un niveau plus élevé chez les répondeurs au vaccin contre l'HepB. Des résultats similaires ont été rapportés par Tsang JS. *et collab.* où un phénotype spécifique des lymphocytes B, à savoir le pourcentage de lymphocytes B mémoire CD38+CD27+, était le principal marqueur d'un modèle qui prédisait la réponse humorale au vaccin contre le virus Influenza H1N1 avant la vaccination (cf. Table 2 de l'introduction de cette thèse) (151).

Dans le 2e article de cette thèse, l'analyse transcriptionnelle du sang des sujets vaccinés avec ALVAC-HIV+gp120 (i.e. vaccin RV144) stimulés avec des peptides dérivés de l'enveloppe virale du VIH a été caractérisée par l'expression élevée des gènes associés à la présentation de l'Ag, la maturation du complexe majeur d'histocompatibilité de classe II et des gènes dotés de fonctions antivirales. Ces gènes et ces réseaux transcriptionnels étaient induits uniquement chez les participants qui sont restés séronégatifs pour le VIH-1 lors de leur dernier suivi (3 ans après

la vaccination) par rapport aux participants qui ont été infectés par le VIH-1. L'induction de ces voies n'a pas été observée chez les receveurs de placebo ou en l'absence de stimulation avec des peptides Env indiquant que ces voies de signalisations ont été induites spécifiquement en réponse à l'administration du vaccin. L'analyse intégrative multiomique a démontré que l'activité du facteur de transcription IRF7, un régulateur clef d'une réponse immunitaire innée antivirale induit par les interférons, était liée à la fréquence de lymphocytes T capable de produire les cytokines effectrices de la réponse immunitaire. Pour valider les résultats de l'analyse in silico des données multiomiques, nous avons effectué des expériences *ex vivo* où des lymphocytes T ont été stimulé par des interférons et infectés in vitro par le VIH. Ces expériences ont montré que l'activité transcriptionelle d'IRF7 était associée à un faible risque chez les vaccinés et de plus était exprimé par les lymphocytes T. L'expression d'IRF7 par les lymphocytes T peut rendre ces cellules réfractaires à l'infection par le VIH-1. Ces résultats fournissent un mécanisme par lequel les cellules T CD4+ de sujets immunisés avec le vaccin RV144 vont monter une réponse à l'interféron de type II spécifique au VIH rendant ainsi les lymphocytes résistants à l'infection par le VIH-1. Ce résultat n'est toutefois qu'un seul parmi d'autres hypothèses pouvant expliquer le rôle d'IRF7 dans la réponse au vaccin RV144; une autre explication est qu'IRF7 est aussi un marqueur de l'infection de monocytes/macrophages par le vaccin RV144 montrant que le vaccin infecte ses cellules cibles chez les sujets vaccinés (199). Le rôle d'IRF7 dans la présentation d'Ag n'a pas été extensivement étudié (IRF7 est un régulateur en amont du transporteur TAP1 qui permet les antigènes peptiques d'être présentés par le CMH de classe I (200) ainsi qu'en amont du régulateur de l'expression des molécules CMH de classe II CIITA (201)); cette fonction d'IRF7 mérite d'être le sujet d'expériences additionnelles.

D'une part, dans le 2^{e} article de cette thèse, nous montrons que les niveaux d'expression du gène IRF7 et son activité transcriptionnelle, mesurés deux semaines après la vaccination, sont associés à un plus faible risque d'infection chez les personnes vaccinées avec le vaccin RV144. D'autre part, dans le 1er article de cette thèse nous avons montré que les niveaux d'expression du gène IRF7 et l'activation de la voie transcriptionnelle en aval d'IRF7, mesurés prévaccination, sont associés à une faible réponse humorale au vaccin contre l'HepB. Le facteur de transcription IRF7 est un exemple de marqueur pouvant être impliqué dans le mécanisme de la réponse aux vaccins, mais pouvant avoir un impact distinct sur l'expression génique dans les cellules du système immunitaire inné. Les fonctions positives et négatives des interférons sont l'objet d'importante recherches (202). De plus amples recherches sont nécessaires pour déterminer si l'âge des participants (personnes âgées de plus de 65 ans dans le 1er article; adultes de moins

de 30 ans dans le 2e article), le temps écoulé après la vaccination étudiée (prévaccination dans le 1er article; deux semaines après la vaccination dans le 2e article), le vaccin étudié (vaccin contre l'HepB dans le 1er article; vaccin contre le VIH dans le 2e article) quel facteur dicte si l'activité d'IRF7 est bénéfique ou néfaste à la réponse aux vaccins

Dans le 1er article présenté dans cette thèse, l'expression de HMOX1, l'hémoxydase qui catalyse la dégradation de l'hème, est induite chez les personnes ayant répondu faiblement au vaccin contre l'HepB alors que dans le 3e article de cette thèse, les niveaux d'expression de HMOX1 sont associés à une absence de symptômes après une infection par un virus respiratoire en concordance avec son activité antivirale reportée dans la littérature scientifique. En effet, dans des expériences effectuées chez la souris, l'augmentation de l'expression de HMOX1 procure une résistance cellulaire contre la toxicité de l'hème et permet de contrôler l'infection et l'inflammation induites par une infection virale (203). Le fait que certains marqueurs tels que HMOX1 sont associés à la sévérité d'une infection virale, mais aussi à la réponse aux vaccins est une observation qui pourrait être généralisable. Dans le 3e article de cette thèse, nous avons essayé d'évaluer l'intersection entre les marqueurs associés à une infection virale naturelle et les marqueurs de la réponse aux vaccins (simulant une infection virale), et avons constaté que plusieurs des marqueurs associés à une plus grave sévérité des symptômes sont également de marqueurs de la réponse au vaccin, suggérant que les individus à risque de développer des symptômes sévères sont également ceux le plus à même de bénéficier de la vaccination. Il est important de noter que cette intersection n'était pas statistiquement significative, néanmoins reste intéressante et mérite d'être évaluée sur d'autres cohortes et pour d'autres infections virales.

3. Limite des données larges échelles

L'analyse des données de large échelle nécessite l'utilisation d'approches descendantes (*top-down*) pour identifier les mécanismes qui contrôlent la réponse au vaccin. Les approches descendantes utilisées dans les trois articles présentés dans cette thèse incluent l'analyse d'enrichissement de vois de signalisation, les méthodes de déconvolution pour identifier les cellules immunitaires en jeu durant la réponse au vaccin et les méthodes d'intégration multiomique pour établir des liens corrélatifs entre les différentes données transcriptionnelles et protéiques générées par l'analyse des échantillons cellulaires et plasmiques analysés dans cette thèse. Les réseaux identifiés par l'utilisation de ces méthodes ont permis d'établir des liens entre plusieurs éléments (ARN messager, protéines ou cellules) associés à la réponse aux vaccins. Les deux principales limites de ce type d'analyse sont que (1) les éléments identifier peuvent être

des marqueurs de la réponse aux vaccins sans toutefois être impliqués dans le mécanisme protectif induit par les vaccins et (2) les liens de connectivité entre ces éléments sont basés sur des corrélations et non sur des liens causes/effets.
Pour pallier ces limites, nous avons utilisé des méthodes d'extraction de connaissances de la littérature scientifique (une discipline de l'apprentissage machine utilisant les concepts du traitement automatique des langues) pour identifier des données externes d'invalidation génique supportant l'implication directe de l'élément dans la réponse aux vaccins (204). L'extraction de connaissances nous a permis d'identifier des jeux de données *in vitro* où certains des biomarqueurs de la réponse au vaccin RV144 sont surexprimés et ceci affecte le cycle de réplication du VIH suggérant qu'ils puissent être impliqués dans les mécanismes de protection induits par le vaccin contre le VIH (ex. XCL1, un des gènes cible de IRF7, est capable de bloquer l'attachement du VIH à une cellule cible et ainsi pourrait bloquer l'infection). Les méthodes d'extraction de connaissances nous ont aussi permis de déterminer l'impact positif ou négatif de la régulation entre chaque paire d'éléments et les liens causaux entre ces éléments (ex. IL6 est une cible transcriptionelle de NFκB (205); ces liens causaux putatives peuvent être validés par des expériences de mutagenèse génique telles que CRISPR-KO).

L'approche descendante (i.e. avec en apex les études du transcriptome global et les réduisant à un nombre restreint de gènes utilisés comme biomarqueurs) adoptée dans les trois articles de cette thèse est un moyen d'identifier les mécanismes de réponse aux vaccins. Une approche ascendante reposant sur des connaissances *a priori* de la réponse aux vaccins et où le système immunitaire est modélisé en tant que réseaux booléens, bayésien ou entant qu'équations différentielles aurait pu être tentée dans les articles de cette thèse. Toutefois, il existe aujourd'hui un fossé entre l'utilisation des données large échelle et l'utilisation de modèles *in silico* de la réponse immunitaire. En effet, la majorité de ces modèles nécessitent des données mesurées sur un nombre restreint d'éléments transcriptionnels, protéiques ou cellulaires échantillonnés à plusieurs reprises après la perturbation du système immunitaire (par la vaccination ou par une infection virale). Or, la plupart des jeux de données utilisés dans cette thèse n'ont été mesurés qu'à un nombre restreint de périodes après la perturbation (1er article : prévaccination et 7 jours après la vaccination; 2^{e} article : 2 semaines après la vaccination et 3^{e} article : avant l'infection et toutes les deux heures jusqu'à 24 heures après l'infection). L'utilisation de données collectées à plusieurs intervalles après la perturbation permettrait l'utilisation de ces modèles *in silico* et améliorerait notre capacité d'identifier des relations causales entre biomarqueurs (où un biomarqueur n'étant induit que brièvement après la perturbation à de faibles chances d'être

impliqué dans les mécanismes de la réponse à une infection pouvant survenir plusieurs jours/mois après la vaccination).

L'utilisation des méthodes d'apprentissage machine avec comme entrée des jeux de données large échelle ne permet pas toujours de générer des biomarqueurs robustes de la réponse immunitaire aux vaccins. Le 3e article de cette thèse prouve ce point; une vingtaine d'équipes ont reçu les mêmes jeux de données transcriptionnelles, ont toutes utilisé diverses méthodes d'apprentissage machine et ont développé des classificateurs ayant différents niveaux de précision suggérant que le choix des méthodes d'apprentissage a un effet sur la précision du classificateur. Un des principaux points à retenir du 3e article de cette thèse est que les deux classificateurs les plus précis de la sévérité des symptômes ont soit utilisé l'ensemble des gènes du génome pour établir leurs prédictions ou ont encore consolidé les gènes impliqués dans une même voie de signalisation et ensuite ont utilisé ces voies de signalisation comme variables initiales. Une même observation peut être faite dans le 1er article de cette thèse ou le BioAge, basé sur l'expression de plus de 2000 gènes séparés en 20 modules de gènes corrélés entre eux, possède une précision plus importante que le classificateur naïf bayésien basé sur l'expression de 15 gènes. En d'autres mots, la redondance dans les prédicteurs utilisés dans ces classificateurs semblerait améliorer notre capacité à prédire la réponse aux vaccins ou à une infection virale.
Un exercice consistant à construire un classificateur basé sur l'ensemble des classificateurs soumis et utilisant le vote majoritaire comme prédiction (i.e. un participant est classifié comme allant développer des symptômes sévères si plus de 50% des classificateurs l'ont prédit à risque), n'a pas permis d'augmenter le pouvoir prédictif global du défi, que ce soit en utilisant les classificateurs basés que sur l'expression transcriptionnelle préinfection ou collectée dans les 24 heures suivant l'infection. Nous attribuons cette lacune en partie au faible volume du jeu de données d'entraînement ou encore à l'homogénéité relative des données soumises pour l'analyse bio-informatique.

L'utilisation de méthodes intégratives ne se solde pas toujours en des modèles prédictifs plus précis. En effet, dans le 1er article, les données du protéome plasmatique n'ont pas permis de construire un modèle prédictif même si les données incluaient un nombre important de cytokines et de chimiokines. De plus, l'union de données transcriptionnelles et cellulaires n'a pas permis de générer des modèles prédictifs plus précis que ceux construits sur la base de chaque type de données séparément. De façon similaire, dans le 2e article, l'ajout des données transcriptionnelles

à un modèle basé sur les niveaux d'Acs, les haplotypes, l'âge, le sexe et la pratique sexuelle des participants n'a pas entraîné en un bénéfice en termes de prédiction du risque d'infection des personnes vaccinées avec le vaccin RV144 (suggérant que les mêmes voies de signalisations sont identifiées par ces différentes modalités de données). L'utilisation de données transcriptionnelles longitudinales peut potentiellement permettre d'améliorer ces modèles prédictifs.

L'approche intégrative utilisée dans les 1er et 2e articles de thèse a également des limites. L'approche utilisée nécessite des données complètes sans valeurs manquantes (des méthodes d'imputation ont été utilisées dans ces deux articles) et ne peut pas être appliquée à des données longitudinales. Des approches alternatives pour pallier ces limitations ont été développées (ex. MOFA pour des valeurs manquantes (206) et OmicsLonDA pour des données longitudinales (207)); ces méthodes nécessitent d'être évaluées non seulement sur des données simulées, mais également en utilisant des données expérimentales (ce qui n'est pas le cas à ce jour).

4. Limitation des données acquises sur les cellules du sang périphérique

La quasi-totalité des données disponibles pour étudier la réponse aux vaccins est dérivée d'échantillons sanguins. Le sang est un compartiment facile à collecter comparé à d'autres compartiments du corps humain. Des études chez les animaux ont démontré que la contribution de la réponse immunitaire mucosale est aussi importante que la réponse immunitaire sanguine (**Annexe 3** et **Annexe 4**). En effet, en utilisant des techniques multiomiques similaires à celles présentées dans cette thèse et en intégrant les réponses immunitaires dans les tissus muqueux et systémiques dans le sang, nous avant identifier de meilleurs biomarqueurs de la réponse aux vaccins qu'en utilisant uniquement les données du sang. Pour le VIH, le site d'infection est les tissus des muqueuses vaginales ou rectales/anales; la détection d'une réponse mucosale est plus importante qu'une réponse par les lymphocytes du sang périphérique (208). Par exemple, chez les macaques nous avons démontré que les niveaux d'IgG reconnaissant l'enveloppe du VIH dans le tissu rectal sont un meilleur marqueur de la protection conférée par la vaccination que les niveaux de marqueurs mesurés dans le sang (**Annexe 3**).

De façon similaire, les infections par les virus respiratoires peuvent induire des réactions dans les poumons différentes de celles observées dans le sang, l'utilisation des technologies large échelle avec des échantillons des tissus mucosaux peut potentiellement résulter en l'identification de marqueurs diffèrent de ceux dans le sang, plus prédictif, et potentiellement plus à même d'être impliqués dans l'induction de mécanismes responsables de la réponse protectrice induite par les vaccins.

L'utilisation de modèles animaux n'est pas le seul moyen d'étudier la réponse aux vaccins dans les tissues. L'utilisation de nouvelles technologies, telle que l'aspiration des nœuds lymphatiques (209), va permettre de collecter des échantillons des tissus en demeurant minimalement invasif. L'utilisation des technologies permettant de mesurer le transcriptome et le protéome d'une cellule va nous permettre d'utiliser moins de matériel biologique et obtenir plus d'information de ces échantillons. Par exemple, ces technologies à l'échelle de la cellule ont permis d'étudier le rôle des lymphocytes B et leur différentiation en lymphocytes B mémoires à la suite de la vaccination par un vaccin spécifique des virus Influenza (209). De telles études vont grandement améliorer nos connaissances sur les mécanismes de la réponse aux vaccins et permettre d'identifier de meilleurs biomarqueurs de la réponse aux vaccins.

Conclusion

Les vaccins représentent l'un des plus grands succès de la médecine moderne. Toutefois, la plupart des vaccins ont été développés de manière empirique, avec peu ou pas de connaissances sur leurs mécanismes d'action. Les technologies d'analyse moléculaire à large échelle telles que les biopuces à ADN offrent de nouvelles perspectives qui permettraient d'identifier les mécanismes de l'immunité induite par les vaccins. Les travaux présentés dans cette thèse ont permis la mise au point de modèles robustes pour la prédiction de la réponse au vaccin en exploitant les profils d'expression de gènes et de protéines dans le sang.

Les articles de cette thèse démontrent que les profils transcriptionnels obtenus avant la perturbation du système immunitaire soit par un vaccin ou soit par une infection virale peuvent être utilisés pour identifier des biomarqueurs de la réponse immunitaire maximale survenant une ou plusieurs semaine(s) après la perturbation. Dans les cas présentés dans cette thèse, ces biomarqueurs sont plus précis que des marqueurs obtenus après la perturbation suggérant que le statut du système immunitaire avant la perturbation pourrait jouer un rôle prépondérant dans la réponse à un vaccin. Les marqueurs identifiés dans cette thèse n'ont toutefois qu'une précision modérée suggérant que d'autres facteurs (microbiome, métabolome, épigénome) peuvent jouer un rôle important dans la modulation de la réponse aux vaccins.

Les articles de cette thèse montrent également comment l'intégration de plusieurs types de données analytiques procure des informations additionnelles sur les mécanismes de réponse aux vaccins. Ces mécanismes ne peuvent pas être identifiés dans leur intégrité à la suite de l'analyse d'un seul type de données. Ceci suggère que l'étude intégrée de plusieurs modalités analytiques (génétique, transcriptomique, protéique) aurait un apport synergique sur la capacité de décrypter les mécanismes de la réponse aux vaccins.

Références bibliographiques

1. Canada.ca. *Report On Hepatitis B and C Surveillance in Canada*. 2019; Available from: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/report-hepatitis-b-c-canada-2019.html.
2. Canada.ca. *HIV in Canada: 2020 Surveillance highlights*. 2020.
3. Canada.ca. *FluWatch annual report: 2019-2020 influenza season*. 2020; Available from: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/fluwatch/2019-2020/annual-report.html.
4. Seeger, C. and Mason, W.S., *Molecular biology of hepatitis B virus infection.* Virology, 2015. **479-480**, 672-86.
5. Hou, J., Liu, Z., and Gu, F., *Epidemiology and Prevention of Hepatitis B Virus Infection.* Int J Med Sci, 2005. **2**(1), 50-57.
6. Winer, B.Y. and Ploss, A., *Determinants of hepatitis B and delta virus host tropism.* Curr Opin Virol, 2015. **13**, 109-16.
7. Gitlin, N., *Hepatitis B: diagnosis, prevention, and treatment.* Clin Chem, 1997. **43**(8 Pt 2), 1500-6.
8. Hu, W.S. and Hughes, S.H., *HIV-1 reverse transcription.* Cold Spring Harb Perspect Med, 2012. **2**(10).
9. McCutchan, F.E., *Global epidemiology of HIV.* J Med Virol, 2006. **78 Suppl 1**, S7-S12.
10. Campbell, E.M. and Hope, T.J., *Live cell imaging of the HIV-1 life cycle.* Trends Microbiol, 2008. **16**(12), 580-7.
11. Doitsh, G., et al., *Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection.* Nature, 2014. **505**(7484), 509-14.
12. Persaud, D., et al., *A stable latent reservoir for HIV-1 in resting CD4(+) T lymphocytes in infected children.* J Clin Invest, 2000. **105**(7), 995-1003.
13. Vaillant, A.A.J. and Naik, R., *HIV-1 associated opportunistic infections.* StatPearls [Internet], 2019.
14. Chaisson, R.E., et al., *Impact of opportunistic disease on survival in patients with HIV infection.* AIDS, 1998. **12**(1), 29-33.
15. Shaw, G.M. and Hunter, E., *HIV transmission.* Cold Spring Harb Perspect Med, 2012. **2**(11).
16. Chu, C. and Selwyn, P.A., *Diagnosis and initial management of acute HIV infection.* Am Fam Physician, 2010. **81**(10), 1239-44.
17. Makela, M.J., et al., *Viruses and bacteria in the etiology of the common cold.* J Clin Microbiol, 1998. **36**(2), 539-42.
18. Paules, C. and Subbarao, K., *Influenza.* Lancet, 2017. **390**(10095), 697-708.
19. Bouvier, N.M. and Palese, P., *The biology of influenza viruses.* Vaccine, 2008. **26 Suppl 4**, D49-53.
20. Brankston, G., et al., *Transmission of influenza A in human beings.* Lancet Infect Dis, 2007. **7**(4), 257-65.
21. Herold, S., et al., *Influenza virus-induced lung injury: pathogenesis and implications for treatment.* Eur Respir J, 2015. **45**(5), 1463-78.
22. Moghadami, M., *A Narrative Review of Influenza: A Seasonal and Pandemic Disease.* Iran J Med Sci, 2017. **42**(1), 2-13.
23. Collins, P.L., Fearns, R., and Graham, B.S., *Respiratory syncytial virus: virology, reverse genetics, and pathogenesis of disease.* Curr Top Microbiol Immunol, 2013. **372**, 3-38.
24. Kutter, J.S., et al., *Transmission routes of respiratory viruses among humans.* Curr Opin Virol, 2018. **28**, 142-151.

25. Tayyari, F., et al., *Identification of nucleolin as a cellular receptor for human respiratory syncytial virus.* Nat Med, 2011. **17**(9), 1132-5.
26. Krause, C.I., *The ABCs of RSV.* Nurse Pract, 2018. **43**(9), 20-26.
27. C, A.H., Caya, C., and Papenburg, J., *Rapid and simple molecular tests for the detection of respiratory syncytial virus: a review.* Expert Rev Mol Diagn, 2018. **18**(7), 617-629.
28. Bartlett, N., Wark, P., and Knight, D., *Rhinovirus Infections : Rethinking the Impact on Human Health and Disease*. 2019, San Diego, UNITED STATES: Elsevier Science & Technology.
29. Peltola, V., et al., *Rhinovirus transmission within families with children: incidence of symptomatic and asymptomatic infections.* J Infect Dis, 2008. **197**(3), 382-9.
30. Blaas, D. and Fuchs, R., *Mechanism of human rhinovirus infections.* Mol Cell Pediatr, 2016. **3**(1), 21.
31. Greenberg, S.B., *Respiratory consequences of rhinovirus infection.* Arch Intern Med, 2003. **163**(3), 278-84.
32. Jacobs, S.E., et al., *Human rhinoviruses.* Clin Microbiol Rev, 2013. **26**(1), 135-62.
33. Vermaelen, K., *Vaccine Strategies to Improve Anti-cancer Cellular Immune Responses.* Front Immunol, 2019. **10**, 8.
34. Lauring, A.S., Jones, J.O., and Andino, R., *Rationalizing the development of live attenuated virus vaccines.* Nat Biotechnol, 2010. **28**(6), 573-9.
35. Block, S.L., et al., *A randomized, double-blind noninferiority study of quadrivalent live attenuated influenza vaccine in adults.* Vaccine, 2011. **29**(50), 9391-7.
36. Trombetta, C.M., Gianchecchi, E., and Montomoli, E., *Influenza vaccines: Evaluation of the safety profile.* Hum Vaccin Immunother, 2018. **14**(3), 657-670.
37. Dhillon, S., *DTPa-HBV-IPV/Hib Vaccine (Infanrix hexa): A Review of its Use as Primary and Booster Vaccination.* Drugs, 2010. **70**(8), 1021-58.
38. Middaugh, J.P., *Side effects of diphtheria-tetanus toxoid in adults.* Am J Public Health, 1979. **69**(3), 246-9.
39. Clark, T.G. and Cassidy-Hanley, D., *Recombinant subunit vaccines: potentials and constraints.* Dev Biol (Basel), 2005. **121**, 153-63.
40. *Vaccine-Preventable Diseases*, in *Immunology for Pharmacy*, D.K. Flaherty, Editor. 2012, Mosby: Saint Louis. p. 197-213.
41. Leitner, W.W., Ying, H., and Restifo, N.P., *DNA and RNA-based vaccines: principles, progress and prospects.* Vaccine, 1999. **18**(9-10), 765-77.
42. Abdelzaher, H.M., et al., *RNA Vaccines against Infectious Diseases: Vital Progress with Room for Improvement.* Vaccines (Basel), 2021. **9**(11).
43. Fan, Y.J., Chan, K.H., and Hung, I.F., *Safety and Efficacy of COVID-19 Vaccines: A Systematic Review and Meta-Analysis of Different Vaccines at Phase 3.* Vaccines (Basel), 2021. **9**(9).
44. Painter, M.M., et al., *Rapid induction of antigen-specific CD4(+) T cells is associated with coordinated humoral and cellular immunity to SARS-CoV-2 mRNA vaccination.* Immunity, 2021. **54**(9), 2133-2142 e3.
45. Ewer, K.J., et al., *Viral vectors as vaccine platforms: from immunogenicity to impact.* Curr Opin Immunol, 2016. **41**, 47-54.
46. Rerks-Ngarm, S., et al., *Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand.* N Engl J Med, 2009. **361**(23), 2209-20.
47. Li, Y.D., et al., *Coronavirus vaccine development: from SARS and MERS to COVID-19.* J Biomed Sci, 2020. **27**(1), 104.
48. Frederiksen, L.S.F., et al., *The Long Road Toward COVID-19 Herd Immunity: Vaccine Platform Technologies and Mass Immunization Strategies.* Front Immunol, 2020. **11**, 1817.

49. Woo, P.C., et al., *SARS coronavirus spike polypeptide DNA vaccine priming with recombinant spike polypeptide from Escherichia coli as booster induces high titer of neutralizing antibody against SARS coronavirus.* Vaccine, 2005. **23**(42), 4959-68.
50. Shah, R.R., Hassett, K.J., and Brito, L.A., *Overview of Vaccine Adjuvants: Introduction, History, and Current Status.* Methods Mol Biol, 2017. **1494**, 1-13.
51. Zuckerman, J.N., *The importance of injecting vaccines into muscle. Different patients need different needle sizes.* BMJ, 2000. **321**(7271), 1237-8.
52. Zhang, L., Wang, W., and Wang, S., *Effect of vaccine administration modality on immunogenicity and efficacy.* Expert Rev Vaccines, 2015. **14**(11), 1509-23.
53. Jilg, W., Schmidt, M., and Deinhardt, F., *Vaccination against hepatitis B: comparison of three different vaccination schedules.* J Infect Dis, 1989. **160**(5), 766-9.
54. Pellegrini, M., et al., *MF59-adjuvanted versus non-adjuvanted influenza vaccines: integrated analysis from a large safety database.* Vaccine, 2009. **27**(49), 6959-65.
55. Lu, S., *Heterologous prime-boost vaccination.* Curr Opin Immunol, 2009. **21**(3), 346-51.
56. Nichol, K.L., et al., *The effectiveness of vaccination against influenza in healthy, working adults.* N Engl J Med, 1995. **333**(14), 889-93.
57. Miller, E.R., et al., *Deaths following vaccination: What does the evidence show?* Vaccine, 2015. **33**(29), 3288-92.
58. Jackson, S., et al., *Immunogenicity of a two-dose investigational hepatitis B vaccine, HBsAg-1018, using a toll-like receptor 9 agonist adjuvant compared with a licensed hepatitis B vaccine in adults.* Vaccine, 2018. **36**(5), 668-674.
59. Joines, R.W., et al., *A prospective, randomized, comparative US trial of a combination hepatitis A and B vaccine (Twinrix) with corresponding monovalent vaccines (Havrix and Engerix-B) in adults.* Vaccine, 2001. **19**(32), 4710-9.
60. Scheifele, D.W., et al., *Immunogenicity and safety of 3-dose primary vaccination with combined DTPa-HBV-IPV/Hib vaccine in Canadian Aboriginal and non-Aboriginal infants.* Vaccine, 2015. **33**(16), 1897-900.
61. Vesikari, T., et al., *Safety and immunogenicity of a modified process hepatitis B vaccine in healthy infants.* Pediatr Infect Dis J, 2011. **30**(7), e109-13.
62. Jain, V.K., et al., *Time to Change Dosing of Inactivated Quadrivalent Influenza Vaccine in Young Children: Evidence From a Phase III, Randomized, Controlled Trial.* J Pediatric Infect Dis Soc, 2017. **6**(1), 9-19.
63. DiazGranados, C.A., et al., *Efficacy of high-dose versus standard-dose influenza vaccine in older adults.* N Engl J Med, 2014. **371**(7), 635-45.
64. Treanor, J.T., et al., *Immunogenicity and safety of a quadrivalent inactivated influenza vaccine compared with two trivalent inactivated influenza vaccines containing alternate B strains in adults: A phase 3, randomized noninferiority study.* Vaccine, 2017. **35**(15), 1856-1864.
65. van de Witte, S.V., et al., *Trivalent inactivated subunit influenza vaccine Influvac®: 30-year experience of safety and immunogenicity.* Trials in Vaccinology, 2012. **1**, 42-48.
66. Vajo, Z., *The seasonal influenza vaccine Agriflu((R)).* Expert Rev Vaccines, 2011. **10**(11), 1513-7.
67. Jackson, L.A., et al., *Safety, efficacy, and immunogenicity of an inactivated influenza vaccine in healthy adults: a randomized, placebo-controlled trial over two influenza seasons.* BMC Infect Dis, 2010. **10**, 71.
68. Tsai, T.F., *Fluad(R)-MF59(R)-Adjuvanted Influenza Vaccine in Older Adults.* Infect Chemother, 2013. **45**(2), 159-74.
69. Shepherd, B.O., et al., *HIV and SARS-CoV-2: Tracing a Path of Vaccine Research and Development.* Curr HIV/AIDS Rep, 2022. **19**(1), 86-93.
70. Corey, L., et al., *Two Randomized Trials of Neutralizing Antibodies to Prevent HIV-1 Acquisition.* N Engl J Med, 2021. **384**(11), 1003-1014.

71. Alter, M.J., *Epidemiology and prevention of hepatitis B.* Semin Liver Dis, 2003. **23**(1), 39-46.
72. Weinbaum, C., Lyerla, R., and Margolis, H.S., *Prevention and control of infections with hepatitis viruses in correctional settings*. 2003: Massachusetts Medical Society Waltham, MA.
73. Kwon, H. and Lok, A.S., *Hepatitis B therapy.* Nat Rev Gastroenterol Hepatol, 2011. **8**(5), 275-84.
74. Merson, M.H., et al., *The history and challenge of HIV prevention.* Lancet, 2008. **372**(9637), 475-88.
75. Cihlar, T. and Fordyce, M., *Current status and prospects of HIV treatment.* Curr Opin Virol, 2016. **18**, 50-6.
76. Kovacs, J.A. and Masur, H., *Prophylaxis against opportunistic infections in patients with human immunodeficiency virus infection.* N Engl J Med, 2000. **342**(19), 1416-29.
77. Spinner, C.D., et al., *HIV pre-exposure prophylaxis (PrEP): a review of current knowledge of oral systemic HIV PrEP in humans.* Infection, 2016. **44**(2), 151-8.
78. Nicol, M.R., Adams, J.L., and Kashuba, A.D., *HIV PrEP Trials: The Road to Success.* Clin Investig (Lond), 2013. **3**(3).
79. Fonner, V.A., et al., *Effectiveness and safety of oral HIV preexposure prophylaxis for all populations.* AIDS, 2016. **30**(12), 1973-83.
80. Régie de l'assurance maladie du Québec. *Liste des médicaments*. [Format PDF] 2021; Available from: https://www.ramq.gouv.qc.ca/sites/default/files/documents/liste-med-2021-02-03-fr.pdf.
81. Simancas-Racines, D., et al., *Vaccines for the common cold.* Cochrane Database Syst Rev, 2017. **5**, CD002190.
82. Powell, K., *The race to make vaccines for a dangerous respiratory virus.* Nature, 2021. **600**(7889), 379-380.
83. Couch, R.B., *Prevention and treatment of influenza.* N Engl J Med, 2000. **343**(24), 1778-87.
84. Garcia-Sastre, A., *Snatch-and-Grab Inhibitors to Fight the Flu.* Cell, 2019. **177**(6), 1367.
85. Villani, A.C., et al., *Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors.* Science, 2017. **356**(6335).
86. Pulendran, B., *Modulating vaccine responses with dendritic cells and Toll-like receptors.* Immunol Rev, 2004. **199**, 227-50.
87. Hasegawa, H. and Matsumoto, T., *Mechanisms of Tolerance Induction by Dendritic Cells In Vivo.* Front Immunol, 2018. **9**, 350.
88. Wykes, M., et al., *Dendritic cells interact directly with naive B lymphocytes to transfer antigen and initiate class switching in a primary T-dependent response.* J Immunol, 1998. **161**(3), 1313-9.
89. Woodland, D.L. and Kohlmeier, J.E., *Migration, maintenance and recall of memory T cells in peripheral tissues.* Nat Rev Immunol, 2009. **9**(3), 153-61.
90. Trinchieri, G., *Cytokines acting on or secreted by macrophages during intracellular infection (IL-10, IL-12, IFN-gamma).* Curr Opin Immunol, 1997. **9**(1), 17-23.
91. Germic, N., et al., *Regulation of the innate immune system by autophagy: monocytes, macrophages, dendritic cells and antigen presentation.* Cell Death Differ, 2019. **26**(4), 715-727.
92. Esser, M.T., et al., *Memory T cells and vaccines.* Vaccine, 2003. **21**(5-6), 419-30.
93. Vivier, E., et al., *Functions of natural killer cells.* Nat Immunol, 2008. **9**(5), 503-10.
94. Buisman, A.M., et al., *Long-term presence of memory B-cells specific for different vaccine components.* Vaccine, 2009. **28**(1), 179-86.
95. Ma, J., et al., *Cryo-EM structure of S-Trimer, a subunit vaccine candidate for COVID-19.* J Virol, 2021.

96. Kareko, B.W., et al., *Persistence of Neutralizing Antibody Responses Among Yellow Fever Virus 17D Vaccinees Living in a Nonendemic Setting.* J Infect Dis, 2020. **221**(12), 2018-2025.
97. Ellebedy, A.H., et al., *Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination.* Nat Immunol, 2016. **17**(10), 1226-34.
98. Krammer, F., *The human antibody response to influenza A virus infection and vaccination.* Nat Rev Immunol, 2019. **19**(6), 383-397.
99. Clem, A.S., *Fundamentals of vaccine immunology.* J Glob Infect Dis, 2011. **3**(1), 73-8.
100. Duffy, P.E. and Patrick Gorres, J., *Malaria vaccines since 2000: progress, priorities, products.* NPJ Vaccines, 2020. **5**(1), 48.
101. Gray, G.E., et al., *Vaccine Efficacy of ALVAC-HIV and Bivalent Subtype C gp120-MF59 in Adults.* N Engl J Med, 2021. **384**(12), 1089-1100.
102. Kaufmann, S.H.E., *Vaccination Against Tuberculosis: Revamping BCG by Molecular Genetics Guided by Immunology.* Front Immunol, 2020. **11**, 316.
103. Haynes, B.F., et al., *Immune-correlates analysis of an HIV-1 vaccine efficacy trial.* N Engl J Med, 2012. **366**(14), 1275-86.
104. Hannoun, C., Megas, F., and Piercy, J., *Immunogenicity and protective efficacy of influenza vaccination.* Virus Res, 2004. **103**(1-2), 133-8.
105. Beyer, W.E., et al., *Seroprotection rate, mean fold increase, seroconversion rate: which parameter adequately expresses seroresponse to influenza vaccination?* Virus Res, 2004. **103**(1-2), 125-32.
106. CDC. *Seasonal Flu Vaccine Effectiveness Studies*. 2022; Available from: https://www.cdc.gov/flu/vaccines-work/effectiveness-studies.htm.
107. Plotkin, S.A. and Gilbert, P.B., *Nomenclature for immune correlates of protection after vaccination.* Clin Infect Dis, 2012. **54**(11), 1615-7.
108. Yates, N.L., et al., *Vaccine-induced Env V1-V2 IgG3 correlates with lower HIV-1 infection risk and declines soon after vaccination.* Sci Transl Med, 2014. **6**(228), 228ra39.
109. Lau, Y.L., et al., *Response of preterm infants to hepatitis B vaccine.* J Pediatr, 1992. **121**(6), 962-5.
110. Zhang, L., et al., *Effects of hepatitis B immunization on prevention of mother-to-infant transmission of hepatitis B virus and on the immune response of infants towards hepatitis B vaccine.* Vaccine, 2014. **32**(46), 6091-7.
111. Halasa, N.B., et al., *Safety and immunogenicity of trivalent inactivated influenza vaccine in infants.* J Infect Dis, 2008. **197**(10), 1448-54.
112. Ventura, M.T., et al., *Immunosenescence in aging: between immune cells depletion and cytokines up-regulation.* Clin Mol Allergy, 2017. **15**, 21.
113. Weinberger, B., *Adjuvant strategies to improve vaccination of the elderly population.* Curr Opin Pharmacol, 2018. **41**, 34-41.
114. Wolters, B., et al., *Immunogenicity of combined hepatitis A and B vaccine in elderly persons.* Vaccine, 2003. **21**(25-26), 3623-8.
115. Goodwin, K., Viboud, C., and Simonsen, L., *Antibody response to influenza vaccination in the elderly: a quantitative review.* Vaccine, 2006. **24**(8), 1159-69.
116. Edelman, R., et al., *The SENIEUR protocol and the efficacy of hepatitis B vaccination in healthy elderly persons by age, gender, and vaccine route.* Immun Ageing, 2020. **17**, 9.
117. Chambers, C., et al., *Should Sex Be Considered an Effect Modifier in the Evaluation of Influenza Vaccine Effectiveness?* Open Forum Infect Dis, 2018. **5**(9), ofy211.
118. Sakiani, S., Olsen, N.J., and Kovacs, W.J., *Gonadal steroids and humoral immunity.* Nat Rev Endocrinol, 2013. **9**(1), 56-62.
119. Jones, B.G., et al., *Binding of estrogen receptors to switch sites and regulatory elements in the immunoglobulin heavy chain locus of activated B cells suggests a direct influence of estrogen on antibody expression.* Mol Immunol, 2016. **77**, 97-102.

120. Furman, D., et al., *Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination.* Proc Natl Acad Sci U S A, 2014. **111**(2), 869-74.
121. Grosheide, P.M., et al., *Immune response to hepatitis B vaccine in pregnant women receiving post-exposure prophylaxis.* Eur J Obstet Gynecol Reprod Biol, 1993. **50**(1), 53-8.
122. Schlaudecker, E.P., et al., *Declining responsiveness to influenza vaccination with progression of human pregnancy.* Vaccine, 2018. **36**(31), 4734-4741.
123. Wiedmann, M., et al., *Decreased immunogenicity of recombinant hepatitis B vaccine in chronic hepatitis C.* Hepatology, 2000. **31**(1), 230-4.
124. Gross, P.A., et al., *Relation of chronic disease and immune response to influenza vaccine in the elderly.* Vaccine, 1989. **7**(4), 303-8.
125. Saco, T.V., Strauss, A.T., and Ledford, D.K., *Hepatitis B vaccine nonresponders: Possible mechanisms and solutions.* Ann Allergy Asthma Immunol, 2018. **121**(3), 320-327.
126. Beck, C.R., et al., *Influenza vaccination for immunocompromised patients: systematic review and meta-analysis by etiology.* J Infect Dis, 2012. **206**(8), 1250-9.
127. Jacobson, I.M., et al., *Immunogenicity of hepatitis B vaccine in renal transplant recipients.* Transplantation, 1985. **39**(4), 393-5.
128. Kreuter, J., *Liposomes and nanoparticles as vehicles for antibiotics.* Infection, 1991. **19 Suppl 4**, S224-8.
129. Quach, S., et al., *Influenza vaccination coverage across ethnic groups in Canada.* CMAJ, 2012. **184**(15), 1673-81.
130. Erwin-Cohen, R.A., et al., *Human transcriptome response to immunization with live-attenuated Venezuelan equine encephalitis virus vaccine (TC-83): Analysis of whole blood.* Hum Vaccin Immunother, 2017. **13**(1), 169-179.
131. Alper, C.A., et al., *Genetic prediction of nonresponse to hepatitis B vaccine.* N Engl J Med, 1989. **321**(11), 708-12.
132. Wu, T.W., et al., *SNP rs7770370 in HLA-DPB1 loci as a major genetic determinant of response to booster hepatitis B vaccination: results of a genome-wide association study.* J Gastroenterol Hepatol, 2015. **30**(5), 891-9.
133. Wang, C., et al., *HLA and cytokine gene polymorphisms are independently associated with responses to hepatitis B vaccination.* Hepatology, 2004. **39**(4), 978-88.
134. Davila, S., et al., *New genetic associations detected in a host response study to hepatitis B vaccine.* Genes Immun, 2010. **11**(3), 232-8.
135. Chen, J., et al., *Toll-like receptors and cytokines/cytokine receptors polymorphisms associate with non-response to hepatitis B vaccine.* Vaccine, 2011. **29**(4), 706-11.
136. Poland, G.A., Ovsyannikova, I.G., and Jacobson, R.M., *Immunogenetics of seasonal influenza vaccine response.* Vaccine, 2008. **26 Suppl 4**, D35-40.
137. Prentice, H.A., et al., *HLA class II genes modulate vaccine-induced antibody responses to affect HIV-1 acquisition.* Sci Transl Med, 2015. **7**(296), 296ra112.
138. Bartholomeus, E., et al., *Transcriptome profiling in blood before and after hepatitis B vaccination shows significant differences in gene expression between responders and non-responders.* Vaccine, 2018. **36**(42), 6282-6289.
139. HIPC-Chi Signatures Project Team and HIPC-I. Consortium, *Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses.* Sci Immunol, 2017. **2**(14).
140. Weinberger, B., et al., *Impaired Immune Response to Primary but Not to Booster Vaccination Against Hepatitis B in Older Adults.* Front Immunol, 2018. **9**, 1035.
141. Matsumiya, M., et al., *Roles for Treg expansion and HMGB1 signaling through the TLR1-2-6 axis in determining the magnitude of the antigen-specific immune response to MVA85A.* PLoS One, 2013. **8**(7), e67922.

142. Furman, D., et al., *Apoptosis and other immune biomarkers predict influenza vaccine responsiveness.* Mol Syst Biol, 2013. **9**, 659.
143. Kotliarov, Y., et al., *Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus.* Nat Med, 2020. **26**(4), 618-629.
144. Howard, L.M., et al., *AS03-Adjuvanted H5N1 Avian Influenza Vaccine Modulates Early Innate Immune Signatures in Human Peripheral Blood Mononuclear Cells.* J Infect Dis, 2019. **219**(11), 1786-1798.
145. Rechtien, A., et al., *Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV.* Cell Rep, 2017. **20**(9), 2251-2261.
146. Kazmin, D., et al., *Systems analysis of protective immune responses to RTS,S malaria vaccination in humans.* Proc Natl Acad Sci U S A, 2017. **114**(9), 2425-2430.
147. Howard, L.M., et al., *Cell-Based Systems Biology Analysis of Human AS03-Adjuvanted H5N1 Avian Influenza Vaccine Responses: A Phase I Randomized Controlled Trial.* PLoS One, 2017. **12**(1), e0167488.
148. O'Connor, D., et al., *High-dimensional assessment of B-cell responses to quadrivalent meningococcal conjugate and plain polysaccharide vaccine.* Genome Med, 2017. **9**(1), 11.
149. Ovsyannikova, I.G., et al., *Gene signatures associated with adaptive humoral immunity following seasonal influenza A/H1N1 vaccination.* Genes Immun, 2016. **17**(7), 371-379.
150. Nakaya, H.I., et al., *Systems Analysis of Immunity to Influenza Vaccination across Multiple Years and in Diverse Populations Reveals Shared Molecular Signatures.* Immunity, 2015. **43**(6), 1186-98.
151. Tsang, J.S., et al., *Global analyses of human immune variation reveal baseline predictors of postvaccination responses.* Cell, 2014. **157**(2), 499-513.
152. Nakaya, H.I., et al., *Systems biology of vaccination for seasonal influenza in humans.* Nat Immunol, 2011. **12**(8), 786-95.
153. Bucasas, K.L., et al., *Early patterns of gene expression correlate with the humoral immune response to influenza vaccination in humans.* J Infect Dis, 2011. **203**(7), 921-9.
154. Querec, T.D., et al., *Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans.* Nat Immunol, 2009. **10**(1), 116-125.
155. Lin, L., et al., *COMPASS identifies T-cell subsets correlated with clinical outcomes.* Nat Biotechnol, 2015. **33**(6), 610-6.
156. Furman, D., et al., *Cytomegalovirus infection enhances the immune response to influenza.* Sci Transl Med, 2015. **7**(281), 281ra43.
157. Hagan, T., et al., *Antibiotics-Driven Gut Microbiome Perturbation Alters Immunity to Vaccines in Humans.* Cell, 2019. **178**(6), 1313-1328 e13.
158. Li, S., et al., *Metabolic Phenotypes of Response to Vaccination in Humans.* Cell, 2017. **169**(5), 862-877 e17.
159. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258), 561-3.
160. Butte, A., *The use and analysis of microarray data.* Nat Rev Drug Discov, 2002. **1**(12), 951-60.
161. Bammler, T., et al., *Standardizing global gene expression analysis between laboratories and across platforms.* Nat Methods, 2005. **2**(5), 351-6.
162. Pease, A.C., et al., *Light-generated oligonucleotide arrays for rapid DNA sequence analysis.* Proc Natl Acad Sci U S A, 1994. **91**(11), 5022-6.
163. Gentleman, R., et al., *Bioinformatics and computational biology solutions using R and Bioconductor*. 2006: Springer Science & Business Media.
164. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235), 467-70.

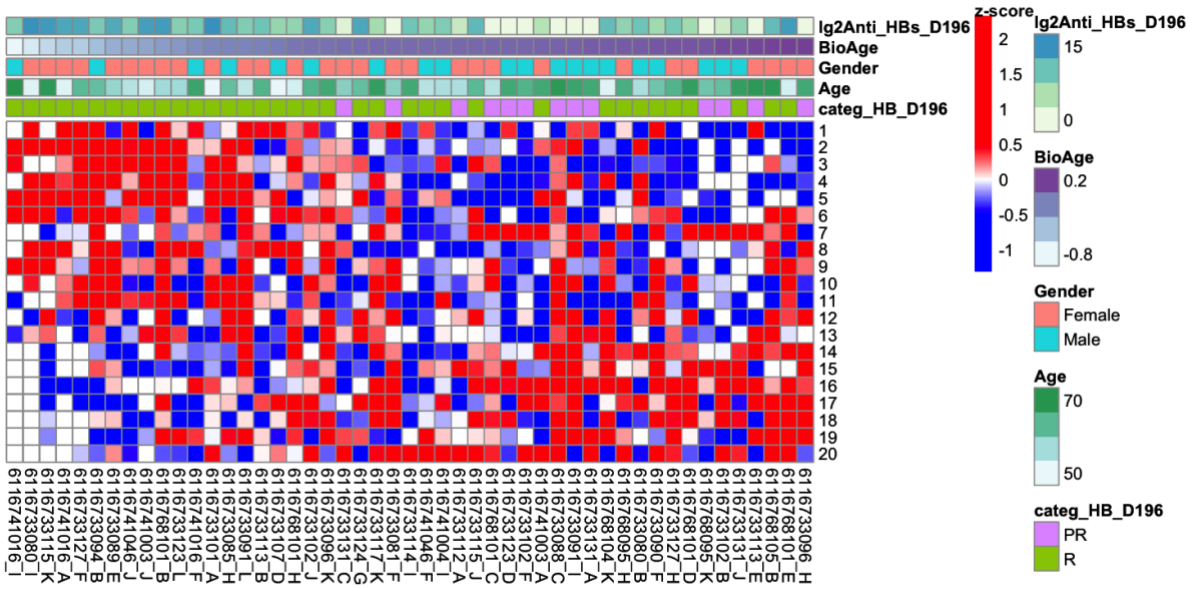
165. Tomaras, G.D., et al., *Vaccine-induced plasma IgA specific for the C1 region of the HIV-1 envelope blocks binding and effector function of IgG.* Proc Natl Acad Sci U S A, 2013. **110**(22), 9019-24.
166. Ackerman, M.E., Barouch, D.H., and Alter, G., *Systems serology for evaluation of HIV vaccine trials.* Immunol Rev, 2017. **275**(1), 262-270.
167. Pepper, S.D., et al., *The utility of MAS5 expression summary and detection call algorithms.* BMC Bioinformatics, 2007. **8**, 273.
168. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2), 249-64.
169. MAQC Consortium, et al., *The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.* Nat Biotechnol, 2006. **24**(9), 1151-61.
170. Khatri, P., Sirota, M., and Butte, A.J., *Ten years of pathway analysis: current approaches and outstanding challenges.* PLoS Comput Biol, 2012. **8**(2), e1002375.
171. Louhimo, R. and Hautaniemi, S., *CNAmet: an R package for integrating copy number, methylation and expression data.* Bioinformatics, 2011. **27**(6), 887-8.
172. Nguyen, H., et al., *PINSPlus: a tool for tumor subtype discovery in integrated genomic data.* Bioinformatics, 2019. **35**(16), 2843-2846.
173. Dimitrakopoulos, C., et al., *Network-based integration of multi-omics data for prioritizing cancer genes.* Bioinformatics, 2018. **34**(14), 2441-2448.
174. Vaske, C.J., et al., *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.* Bioinformatics, 2010. **26**(12), i237-45.
175. Shi, Q., et al., *Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data.* Bioinformatics, 2017. **33**(17), 2706-2714.
176. Rohart, F., et al., *mixOmics: An R package for 'omics feature selection and multiple data integration.* PLoS Comput Biol, 2017. **13**(11), e1005752.
177. Hao, Y., et al., *Integrated analysis of multimodal single-cell data.* Cell, 2021. **184**(13), 3573-3587 e29.
178. Cantini, L., et al., *Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer.* Nat Commun, 2021. **12**(1), 124.
179. Ein-Dor, L., et al., *Outcome signature genes in breast cancer: is there a unique set?* Bioinformatics, 2005. **21**(2), 171-8.
180. Haibe-Kains, B., et al., *A three-gene model to robustly identify breast cancer molecular subtypes.* J Natl Cancer Inst, 2012. **104**(4), 311-25.
181. Dudoit, S., Fridlyand, J., and Speed, T.P., *Comparison of discrimination methods for the classification of tumors using gene expression data.* Journal of the American statistical association, 2002. **97**(457), 77-87.
182. Lee, J.W., et al., *An extensive comparison of recent classification tools applied to microarray data.* Computational Statistics & Data Analysis, 2005. **48**(4), 869-885.
183. Michiels, S., Koscielny, S., and Hill, C., *Prediction of cancer outcome with microarrays: a multiple random validation strategy.* Lancet, 2005. **365**(9458), 488-92.
184. Simon, R., *Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data.* Br J Cancer, 2003. **89**(9), 1599-604.
185. Hastie, T., Tibshirani, R., and Friedman, J., *High-Dimensional Problems: p.. N. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 5th printing*. 2011, New York: Springer.
186. Efron, B., *Estimating the error rate of a prediction rule: improvement on cross-validation.* Journal of the American statistical association, 1983. **78**(382), 316-331.
187. Dupuy, A. and Simon, R.M., *Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting.* J Natl Cancer Inst, 2007. **99**(2), 147-57.

188. Ein-Dor, L., Zuk, O., and Domany, E., *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer.* Proc Natl Acad Sci U S A, 2006. **103**(15), 5923-8.
189. Van der Wielen, M., et al., *Hepatitis A/B vaccination of adults over 40 years old: comparison of three vaccine regimens and effect of influencing factors.* Vaccine, 2006. **24**(26), 5509-15.
190. Hudu, S.A., et al., *Antibody and immune memory persistence post infant hepatitis B vaccination.* Patient Prefer Adherence, 2013. **7**, 981-6.
191. Canada.ca. *Hepatitis B Vaccine: Canadian Immunization Guide*. 2017; Available from: https://www.canada.ca/en/public-health/services/publications/healthy-living/canadian-immunization-guide-part-4-active-vaccines/page-7-hepatitis-b-vaccine.html.
192. Li, S., et al., *Molecular signatures of antibody responses derived from a systems biology study of five human vaccines.* Nat Immunol, 2014. **15**(2), 195-204.
193. Bhattacharya, S., et al., *ImmPort, toward repurposing of open access immunological assay data for translational and clinical research.* Sci Data, 2018. **5**, 180015.
194. Fourati, S., et al., *An innate immune activation state prior to vaccination predicts responsiveness to multiple vaccines.* bioRxiv, 2021, 2021.09.26.461847.
195. de Bruyn, G., et al., *Safety profile of recombinant canarypox HIV vaccines.* Vaccine, 2004. **22**(5-6), 704-13.
196. Leroux-Roels, G., et al., *Prevention of hepatitis B infections: vaccination and its limitations.* Acta Clin Belg, 2001. **56**(4), 209-19.
197. Mannick, J.B., et al., *mTOR inhibition improves immune function in the elderly.* Sci Transl Med, 2014. **6**(268), 268ra179.
198. Hyer, R., et al., *Safety of a two-dose investigational hepatitis B vaccine, HBsAg-1018, using a toll-like receptor 9 agonist adjuvant in adults.* Vaccine, 2018. **36**(19), 2604-2611.
199. Harenberg, A., et al., *Gene profiling analysis of ALVAC infected human monocyte derived dendritic cells.* Vaccine, 2008. **26**(39), 5004-13.
200. Jiang, H., et al., *Interferon-alpha promotes MHC I antigen presentation of islet beta cells through STAT1-IRF7 pathway in type 1 diabetes.* Immunology, 2022.
201. Ohtani, F., et al., *Role of interferon regulatory factor 7 in corneal endothelial cells after HSV-1 infection.* Sci Rep, 2021. **11**(1), 16487.
202. Sandler, N.G., et al., *Type I interferon responses in rhesus macaques prevent SIV infection and slow disease progression.* Nature, 2014. **511**(7511), 601-5.
203. Cummins, N.W., et al., *Heme oxygenase-1 regulates the immune response to influenza virus infection and vaccination in aged mice.* FASEB J, 2012. **26**(7), 2911-8.
204. Mitchell, J.A., et al., *Gene indexing: characterization and analysis of NLM's GeneRIFs.* AMIA Annu Symp Proc, 2003, 460-4.
205. Han, Y., Runge, M.S., and Brasier, A.R., *Angiotensin II induces interleukin-6 transcription in vascular smooth muscle cells through pleiotropic activation of nuclear factor-kappa B transcription factors.* Circ Res, 1999. **84**(6), 695-703.
206. Argelaguet, R., et al., *Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets.* Mol Syst Biol, 2018. **14**(6), e8124.
207. Sailani, M.R., et al., *Deep longitudinal multiomics profiling reveals two biological seasonal patterns in California.* Nat Commun, 2020. **11**(1), 4933.
208. Haase, A.T., *Targeting early infection to prevent HIV-1 mucosal transmission.* Nature, 2010. **464**(7286), 217-23.
209. Turner, J.S., et al., *Human germinal centres engage memory and naive B cells after influenza vaccination.* Nature, 2020. **586**(7827), 127-132.

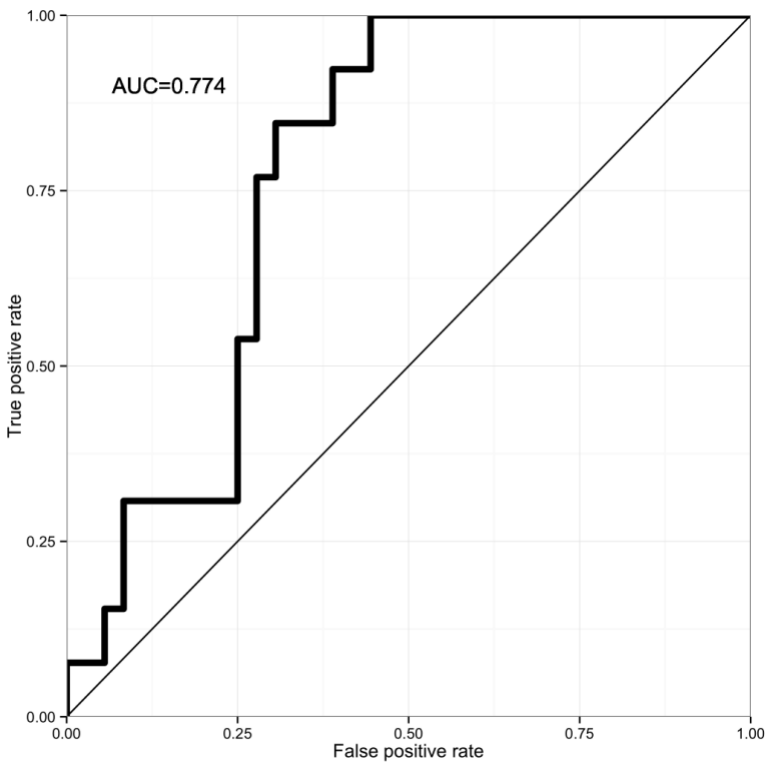
Annexes

1. Annexe 1

A

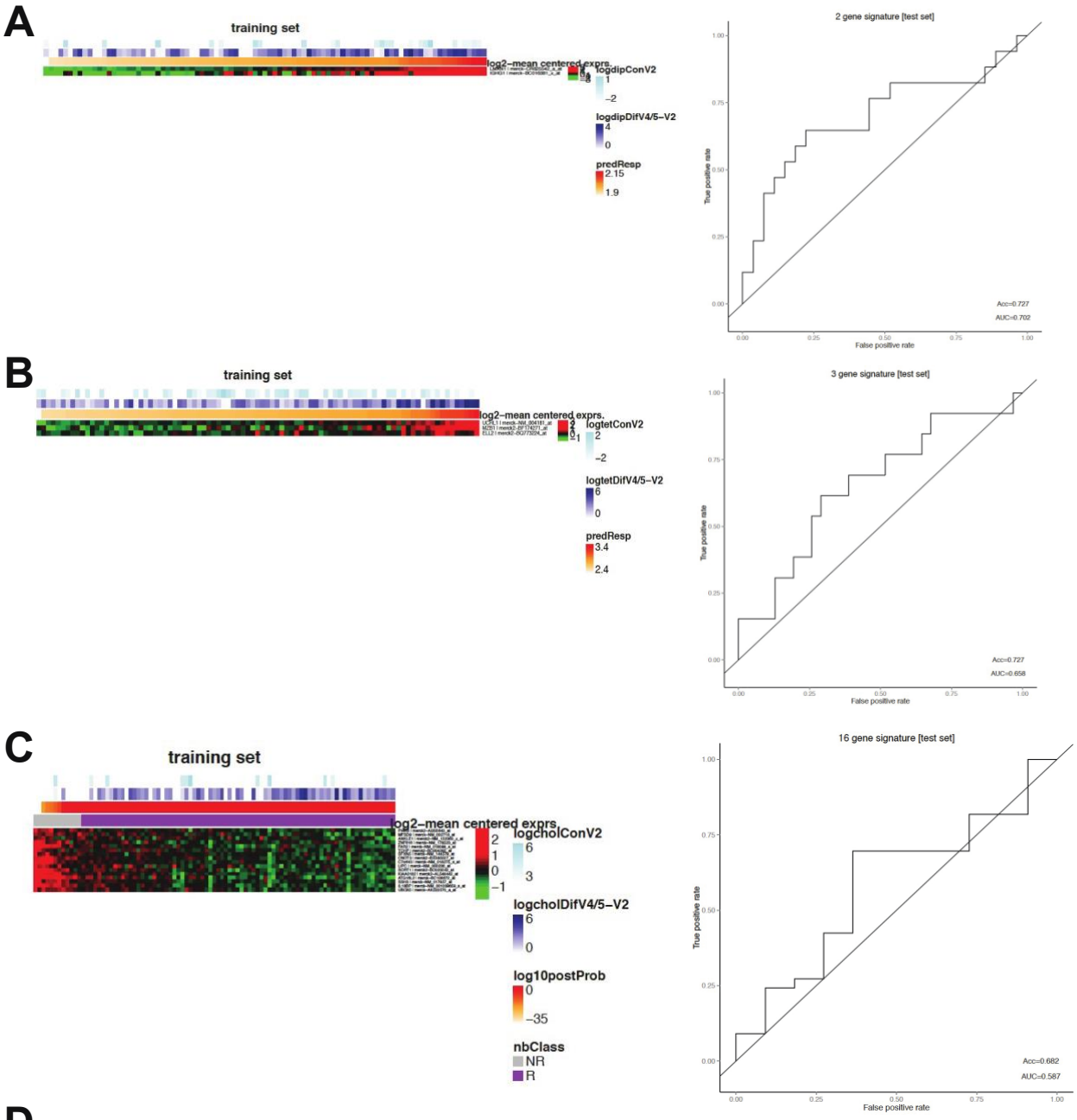


B



La figure ci-dessus résume l'évaluation de la signature BioAge (Article 1) sur un jeu de données indépendant (ImmPort (193) : SDY690). Des personnes âgées (entre 50 et 70 ans), sans précédent de vaccination contre l'HepB, ont été vaccinées avec Engerix-B (vaccin contre l'HepB). Des échantillons sanguins ont été collectés avant la vaccination et leurs transcriptomes ont été analysés par micropuces à ADN de la compagnie Illumina. Les titres d'Acs contre l'Ag de surface de l'HepB ont été mesurés deux semaines après la dernière immunisation avec Engerix-B (Ig2Anti_HBs_D196). La signature BioAge a été appliquée à ces données de la manière décrite dans le 1er article de cette thèse (panneau A). Une analyse par courbe ROC a été utilisée pour estimer la précision du classificateur BioAge, estimée ici a plus de 77% (panneau B).

2. Annexe 2


A
training set
log2−mean centered exprs.
logdipConV2
1
−2
logdipDifV4/5−V2
4
0
predResp
2.15
1.9
2 gene signature [test set]
True positive rate
False positive rate
Acc=0.727
AUC=0.702
B
training set
log2−mean centered exprs.
logtetConV2
2
−2
logtetDifV4/5−V2
6
0
predResp
3.4
2.4
3 gene signature [test set]
True positive rate
False positive rate
Acc=0.727
AUC=0.658
C
training set
log2−mean centered exprs.
2
1
0
−1
logcholConV2
6
3
logcholDifV4/5−V2
6
0
log10postProb
0
−35
nbClass
NR
R
16 gene signature [test set]
True positive rate
False positive rate
Acc=0.682
AUC=0.587
D
10-fold CV [training set]
Accuracy
nb of genes

La figure ci-dessus présente les classificateurs construits sur les jeux de données transcriptionnelles sept jours après la première immunisation avec les vaccins Twinrix (vaccin contre les hépatites B et C), vaccin Diphteria-Tetanus booster et Dukoral (vaccin contre le choléra). La même cohorte de participants et la même stratégie d'apprentissage utilisée dans le premier article de cette thèse ont été utilisées pour construire un classificateur capable de prédire la réponse humorale contre diphtérie (panneau A), tétanos (panneau B) et choléra (panneau C) un mois après la vaccination. Pour chaque classificateur, une analyse par courbe ROC a été utilisée pour estimer la précision du classificateur, estimée à 73% pour la diphtérie (panneau A), 73% pour le tétanos (panneau B) et 68% pour le choléra (panneau C). (Panneau D) Précision estimée par validation-croisée en fonction de nombre de gènes dans le classificateur naïf bayésien pour prédire la réponse au vaccin contre l'HepB en utilisant les données transcriptionnelles 7 jours après la vaccination. Aucun de ces classificateurs possède une précision meilleure qu'une classification aléatoire de la réponse au vaccin (nombre de gènes dans le classificateur égal à zéro).

3. Annexe 3

Il existe deux principales limites à l'étude d'un vaccin chez l'homme. La première limite est la difficulté de distinguer un participant n'ayant pas été en contact avec le virus après vaccination d'un participant ayant été en contact avec le virus, mais protéger par le vaccin. La deuxième limite est la difficulté de collecter des échantillons dans les tissues. Pour remédier à ces problèmes, nous avons utilisé dans l'article annexe ci-joint les macaques rhésus, un modèle animal de l'homme pour les études vaccinales. Grâce à cela, nous avons pu nous assurer que les tous les animaux ont été en contact avec le virus après vaccination et on a pu collecter des échantillons rectaux pour mesurer des marqueurs de la réponse immunitaire dans le site d'infection par le VIH. Entre autres, cette étude a permis de confirmer que les Acs IgG reconnaissant la loupe V1/V2 du VIH sont un corrélat de protection par le vaccin RV144. De plus, nous démontrons que le même Acs dans le rectum est un meilleur corrélat que dans le sang. L'analyse intégrative des données a aussi permis d'identifier des enzymes responsables de catalyser des modifications post-traductionnelles des Acs, étant associées à la réponse au vaccin RV144 chez le rhésus.
Dans cet article, l'intégralité de l'analyse bio-informatique des données transcriptionnelles, de cytométrie de flux et de cytokines a été ma responsabilité. L'analyse intégrative multiomique a été ma responsabilité. Le développement de classificateurs a été principalement mon travail avec l'aide du Dr Kellogs. J'ai généré les figures 4G, 5 et 6, et rédigé la moitié de la partie résultat et de la discussion sous la supervision du dernier auteur.

Vaccari M, Gordon SN, Fourati S, Schifanella L, Liyanage NP, Cameron M, Keele BF, Shen X, Tomaras GD, Billings E, Rao M, Chung AW, Dowell KG, Bailey-Kellogg C, Brown EP, Ackerman ME, Vargas-Inchaustegui DA, Whitney S, Doster MN, Binello N, Pegu P, Montefiori DC, Foulds K, Quinn DS, Donaldson M, Liang F, Loré K, Roederer M, Koup RA, McDermott A, Ma ZM, Miller CJ, Phan TB, Forthal DN, Blackburn M, Caccuri F, Bissa M, Ferrari G, Kalyanaraman V, Ferrari MG, Thompson D, Robert-Guroff M, Ratto-Kim S, Kim JH, Michael NL, Phogat S, Barnett SW, Tartaglia J, Venzon D, Stablein DM, Alter G, Sekaly RP, Franchini G. Adjuvant- dependent innate and adaptive immune signatures of risk of SIVmac251 acquisition. Nat Med. 2016 Jul;22(7):762-70. doi: 10.1038/nm.4105. Epub 2016 May 30.

4. Annexe 4

Nous avons identifié plusieurs biomarqueurs candidats associés à la protection par le vaccin ALVAC-SIV (Annexe 1). Toutefois, nous n'avons pas établi si ces marqueurs sont vaccin spécifique ou sont capables de prédire correctement la protection induite par d'autres vaccins.
Ici, nous avons testé ces marqueurs avec un autre vaccin basé sur le vecteur viral NYVAC (dérivé du virus Vaccinia) comparé au vaccin ALVAC-SIV. Une méta-analyse a ensuite été effectuée pour tester les marqueurs avec de multiples vaccins, à différentes durées après vaccination et différentes routes d'infection par SIV.
Dans l'article ci-joint, l'intégralité de l'analyse bio-informatique des données transcriptionnelles, de cytométrie de flux et de cytokines ont été ma responsabilité. La méta-analyse a été ma responsabilité. J'ai généré les figures 2, 5 et Table 1 et rédigé la moitié de la partie résultat et de la discussion sous la supervision du dernier auteur.

Gorini G, Fourati S, Vaccari M, Rahman MA, Gordon SN, Brown DR, Law L, Chang J, Green R, Barrenäs F, Liyanage NPM, Doster MN, Schifanella L, Bissa M, Silva de Castro I, Washington-Parks R, Galli V, Fuller DH, Santra S, Agy M, Pal R, Palermo RE, Tomaras GD, Shen X, LaBranche CC, Montefiori DC, Venzon DJ, Trinh HV, Rao M, Gale M Jr, Sekaly RP, Franchini G. Engagement of monocytes, NK cells, and CD4+ Th1 cells by ALVAC-SIV vaccination results in a decreased risk of SIVmac251 vaginal acquisition. PLoS Pathog. 2020 Mar;16(3):e1008377. doi: 10.1371/journal.ppat.1008377. eCollection 2020 Mar. PubMed PMID: 32163525; PubMed Central PMCID: PMC7093029.