Université de Montréal

Apprentissage d'atlas cellulaires par la méthode de Factorized embeddings

par

Assya Trofimov

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Informatique

February 10, 2022

Université de Montréal

Faculté des arts et des sciences

Cette thèse intitulée

Apprentissage d'atlas cellulaires par la méthode de Factorized embeddings

présentée par

Assya Trofimov

a été évaluée par un jury composé des personnes suivantes :

François Major
(président-rapporteur)

Sébastien Lemieux
(directeur de recherche)

Claude Perreault
(codirecteur)

Miklós Csűrös
(membre du jury)

Mathieu Blanchette
(examinateur externe)

(représentant du doyen de la FESP)

Résumé

Le corps humain contient plus de 3.72×10^{13} cellules qui se distinguent par leur morphologie, fonction et état. Leur catalogage en atlas cellulaires c'est entamé il y a plus de 150 ans, avec l'invention des colorants cellulaires en microscopie. Notre connaissance des types cellulaires et leur phénotypes moléculaires nous permet de connaître et prédire leurs fonctions et patrons d'interactions. Ces connaissances sont à la base de la capacité à poser des diagnostics, créer des médicaments et même faire pousser des organes en biologie synthétique. Surprenamment, notre connaissance est loin d'être complète et c'est pourquoi la caractérisation systématique des cellules et l'assemblage des connaissances en atlas cellulaires est nécessaire. Le développement du séquençage à haut débit a révolutionné la biologie des systèmes et ce type de données est parfait pour la construction d'atlas cellulaires entièrement basés sur les données. Un tel atlas cellulaire contiendra une représentation des cellules par des vecteurs de nombres, où chaque vecteur encode le profil moléculaire capturant des informations biologiques de chaque cellule. Chaque expérience de séquençage d'ARN (RNA-Seq) produit des dizaines de milliers de mesures extrêmement riches en information dont l'analyse demeure non-triviale. Des algorithmes de réduction de dimensionnalité, entre autres, permettent d'extraire des données des patrons importants et encoder les échantillons dans des espaces plus interprétables. De cette manière, les cellules similaires sont groupés sur la base d'une multitude de mesures qu'offre le RNA-Seq. Nous avons donc créé un modèle, le *Factorized Embedding* (FE), qui permet d'organiser les données de séquençage d'ARN de la sorte. Le modèle apprend simultanément deux espaces d'encodage: un pour les échantillons et l'autre pour les gènes. Nous avons observé qu'une fois entraîné, que ce modèle groupe les échantillons sur la base de leur similarité d'expression génique et permet l'interpolation dans l'espace d'encodage et donc une certaine interprétabilité de l'espace d'encodage. Du côté de l'encodage des gènes, nous avons remarqué que les gènes se regroupaient selon leurs patrons de co-expression ainsi que selon des similarité de fonctions, trouvées via des ontologies de gènes (Gene Ontology, GO). Nous avons ensuite exploré les propriétés d'une modification du modèle FE, baptisée le Transcriptome Latent (TLT, de l'anglais *The Latent Transcriptome*), où l'encodage des gènes est remplacé par une fonction d'encodage de k-mers provenant de

données brutes de RNA-Seq. Cette modification du modèle capture dans son espace d'encodage des séquence à la fois de l'information sur la similarité et l'abondance des séquences ADN. L'espace d'encodage a ainsi permis de détecter des anormalités génomiques tels les translocations, ainsi que des mutations spécifiques au patient, rendant cet espace de représentation utile autant pour la visualisation que pour l'analyse de données. Finalement, la dernière itération explorée dans cette thèse, du modèle FE, baptisée cette fois-ci le *TCRome*, encode des séquences TCR (récepteurs de cellules T) plutôt que des k-mers, venant du séquençage de répertoires immuns (TCR-Seq). Une irrégularité dans la performance du modèle a mené à une analyse des séquences plus approfondie et à la détection de deux sous-types de TCR. Nous avons analysé les répertoires TCR de plus de 1000 individus et rapportons que le répertoire TCR est composé de deux types de TCR ontogéniquement et fonctionellement distincts. Nous avons découvert des patrons distincts dans les abondances de l'un ou l'autre type, changeant en fonction du sexe, l'âge et dans le cadre de maladies telles chez les sujets portant des mutations dans le gène AIRE et dans le cadre de la maladie du greffon contre l'hôte (GVHD). Ces résultats pointent vers la nécessité d'utiliser des données de séquençage multi-modales pour la construction d'atlas cellulaires, c'est à dire en plus des séquence TCR, des données sur l'expression génique ainsi que des caractérisation moléculaires seront probablement utiles, mais leur intégration sera non-triviale. Le modèle FE (et ses modifications) est un bon candidat pour ce type d'encodage, vu sa flexibilité d'architecture et sa résilience aux données manquantes.

Mots clés: séquençage à haut débit, apprentissage automatique, réseaux de neurones artificiels, séquençage de TCR, atlas cellulaires

Abstract

The human body contains over 3.72×10^{13} cells, that distinguish themselves by their morphology, function and state. Their cataloguing into cell atlases has started over 150 years ago, with the invention of cellular stains for microscopy. Our knowledge of cell types and molecular phenotypes allows is to better know and predict their functions and interaction patterns. This knowledge is at the basis of the ability to diagnose disease, create drugs and even grow organs in synthetic biology. Surprisingly, our knowledge is far from complete and this is why a systematic characterization of cells and the assembly of cell atlases is important. The development of high throughput sequencing has revolutionized systems biology and this type of data is perfect for the construction of entirely data-driven cell atlases. Such an atlas will contain a representation of cells by vectors of numbers, where each vector encodes a molecular profile, capturing biological data about each cell. Each sequencing experiment yields tens of thousands of measurements, extremely rich in information, but their analysis remains non-trivial. Dimensionnality reduction algorithms allow to extract from the data important patterns and encode samples into interpretable spaces. This way, similar cells are grouped on the basis of a multitude of measurements that comes from high throughput sequencing. We have created a model, the *Factorized Embedding* (FE), that allows to organize RNA sequencing (RNA-Seq) data in such a way. The FE model learns simultaneously two encoding spaces: one for samples and one for genes. We have found that the model groups samples on the basis of similar gene expression and allows for smooth interpolation in the encoding space and thus some manner of interpretability. As for the gene encoding space, we observed that gene coordinates were grouped according to co-expression patterns as well as similarity in function, found via gene ontology (GO). We then explored a modification of the FE model, names The Latent Transcriptome (TLT), where the gene encoding function is replaced by a function encoding k-mers, calculated from raw RNA-Seq data. This modification of the model captured in the k-mer encoding space both sequence similarity and sequence abundance. The encoding space allowed for the detection of genomic abnormalities such as translocations, as well as patient-specific mutations, making the encoding space useful for both visualisation and data analysis. Finally, the last iteration of the FE model that we explored, called *TCRome*, encodes amino-acid TCR sequences rather than k-mers.

An irregularity in the model's performance led us to discover two TCR subtypes, entirely based on their sequence. We have thus analyzed TCR repertoires of over 1000 individuals and report that the TCR repertoire is composed of two ontogenically and functionally distinct types. We have discovered distinct pattens in the abundances of each of the sub-types, changing with age, sex and in the context of some diseases such as in individuals carrying a mutated AIRE gene and in graft versus host disease (GVHD). Collectively, these results point towards the necessity to use multi-modal sequencing data for the construction of cell atlases, namely gene expression data, TCR sequencing data and possibly various molecular characterizations. The integration of all this data will however be non-trivial. The FE model (and its modifications) is a good candidate for this type of data organisation, namely because of its flexibility in architecture and resilience to missing data.

Keywords: high throughput sequencing, machine learning, artificial neural network, TCR sequencing, cell atlas

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et des abréviations

ARN, Acide Ribonucléique

RNA-Seq, Séquençage d'ARN, de l'anglais *RNA Sequencing*

ANN, Réseau de Neurones Artificiels, de l'anglais *Artificial Neural Network*

CNN, Réseau de Neurones Artificiels à Convolution, de l'anglais *Convolutional Neural Network*

RNN, Réseau de Neurones Artificiels Récurrent, de l'anglais *Recurrent Neural Network*

TCR, Récepteur de cellule T, de l'anglais *T Cell Repector*

MHC, Complexe Majeur d'Histocompatibilité, de l'anglais *Major Histocompatibility Complex*

AIRE, Régulateur autoimmun d'expression, de l'anglais *Auto-Immune Regulator of Expression*

GVHD, Maladie de greffon contre l'hôte, de l'anglais *Graft Versus Host Disease*

"Science makes people reach selflessly for truth and objectivity; it teaches people to accept reality, with wonder and admiration, not to mention the deep awe and joy that the natural order of things brings to the true scientist." - Lise Meitner

Remerciements

Comme le proverbe japonais dit: mieux que mille an d'études diligentes est une journée avec un excellent professeur. Mes chers directeurs, Claude et Sébastien, je vous suis reconnaissante pour votre patience et votre soutien à travers toutes ces années.

Comme l'a dit Pasteur: "La chance sourit aux esprits préparés" Claude, tu m'as appris l'importance de la chance en science, mais tu m'as également enseigné à être préparée! J'étais à un Midi-Pizza: tu parlais des diables de Tasmanie et c'est à ce moment là que j'ai eu l'étincelle pour l'immunologie et ses formidables questions. Tu m'as engagée pour un stage d'été avant même que j'aie reçu ma lettre d'acceptation pour le programme de baccalauréat; tu m'as donné la chance de faire partie de cette aventure extraordinaire qu'est la recherche dans ton laboratoire; tu m'as dirigée sans pression, me laissant le choix et la direction du projet et des expériences, tu m'as enseignée à creuser et suivre mon instinct. Pour tout ça je te dis un gros merci! Je suis le chercheur que je suis aujourd'hui grâce à toi.

Sébastien, qui n'a pas eu peur d'engager une étudiante sachant à peine programmer, qui m'a poussée à garder un regard critique, et à questionner l'information, peu importe d'où elle provient, à m'inscrire à des cours difficiles, à persévérer, à suivre des projets au bout - un gros merci! Ta bibliothèque infinie de livres pour toutes occasion de vie (qui changent de place sans arrêt) m'a été bien utile :0) Merci pour ta patience, ton support et tes idées!

Aux membres de la plateforme de bioinformatique: Patrick, JP, Jonathan, Geneviève, Eric - merci de m'avoir accueillie à bras ouverts, m'avoir tenu compagnie au dîner (nos dîners me manquent). Merci également pour tout ce que vous m'avez enseigné - j'en sors gagnante! Vous êtes une équipe de feu! Un gros merci aux professeurs de l'IRIC et du DIRO, qui n'étaient pas mes directeurs de recherche mais qui ont prêté oreille sympathisante et m'ont conseillé: François Major, Yoshua Bengio, Vincent Archambault et Brian Wilhelm.

Aux membres présents et passés des laboratoires Perreault et Lemieux, merci pour votre écoute, votre accueil et votre patience. Aux "rockstar" de la science JD, Céline, Eralda: je vous admire, merci pour toutes vos bonnes idées et pour les discussions. Caro (Labelle) - merci pour ton aide et ton support moral. MaVi, merci pour ta bonne humeur et tes "quelle conne! pendeja! " - tu me fais sourire même quand je suis de mauvaise humeur. À Caro (Côté), Marie-Pierre, Sylvie et Krystel: merci pour votre support et votre soutien quand je

commençais ma carrière au laboratoire - je me suis sentie chez moi au laboratoire velocius quam asparagi conquantur!

Merci à ma famille: Maman, Papa, Papa, Faye et Lavy, pour m'avoir soutenu à travers mes hauts et mes bas, d'avoir offert une oreille sympathique et des distractions, au moment où j'en le plus avais besoin. Merci à mes amis: Tristan, Virgile, Meta, Flo, Dima, Sarah, Devon, Morgan, pour votre bonne humeur et les fous rires qu'on a partagé. Je suis reconnaissante à Phil, mon compagnon de vie. Merci pour ta patience et ton optimisme dans les moments difficiles - tu as célébré avec moi les bourses et les articles, tu m'as rassurée aux grands moments de stress. Je suis honorée de t'avoir dans ma vie.

Finalement, je remercie mon comité de thèse d'avoir accepté d'évaluer ma thèse: Dr. Mathieu Blanchette, Dr. François Major et Dr Miklós Csűrös!

Merci aux membres du Département d'informatique et Recherche Opérationelle, surtout à Céline Bégin, pour ta bonne humeur et tes propos rassurants - merci de m'avoir aidée à naviguer le côté administratif de mon cheminement de doc! Finalement, merci également aux organismes subventionnaires de m'avoir soutenu financièrement durant mes études: Le Fonds de Recherche de Santé du Québec et l'Institut de Recherche en Santé du Canada, Génome Québec et Génome Canada.

Introduction

0.1. Mise en contexte

Le corps humain contient 3.72×10^{13} cellules divisées en près de 210 types cellulaires différents (Bianconi et al., 2013). Le type cellulaire est typiquement défini basé sur la fonction de la cellule dans corps (ex.: globule rouge, hépatocyte, neurone etc.). Ceci est basé sur le fait que des phénotypes moléculaires distincts dans les cellules pointent vers des fonctions distinctes. Ainsi, notre connaissance des types cellulaires et leur phénotypes moléculaires nous permet de connaître et prédire leurs fonctions et patrons d'interactions. La maladie, elle, est définie comme la perturbation des fonctions et interactions normales entre les cellules. Par exemple, une cellule tumorale se multiplie rapidement: son fonctionnement normal est perturbé. Il va donc de soi qu'afin de mieux comprendre le vivant, une caractérisation systématique des cellules et l'assemblage de ces connaissances dans un atlas est importante.

0.2. Les atlas cellulaires d'hier, d'aujourd'hui et demain

La caractérisation des divers types cellulaires du corps humain s'est entamée suite au développement de colorants cellulaires en microscopie. L'hematoxyline, développée en 1873, demeure à ce jour l'un de colorants les plus utilisés en histologie (en combinaison avec un autre colorant, l'éosine) (Alturkistani et al., 2015). Au dix-neuvième siècle, le développement de nouveaux colorants pour la coloration différentielle des cellules a pris de l'ampleur et a même mené à la création du premier composé de chimiothérapie: le traitement pour la syphilis, développée en 1913 par Paul Ehrlich, lauréat d'un prix Nobel (Ehrlich, 1913). Les premiers atlas cellulaires contenaient donc des images de microscopie, définissant les types cellulaires par microscopie et coloration différentielle. Les connaissances en histologie ont entre autre permis à une meilleure caractérisation des environnement tumoraux et ont catalysé le développement de nouveaux tests diagnostiques et traitements (Dapson et Horobin, 2009). Par exemple, le dépistage du cancer du col de l'utérus, fait via un Pap-test, utilise une combinaison de 4 colorants, dont l'hematoxyline (Dapson et Horobin, 2009).

Cependant, les premiers atlas cellulaires caractérisaient les cellules utilisaient des colorants choisis par commodité ou découverts par hasard (Regev et al., 2017). La catégorisation

des cellules était principalement basée sur des critères différents: la morphologie, l'expression d'un marqueur de surface, ou la sécrétion de molécules. Nos connaissances actuelles sur les cellules, issus de ces anciens atlas cellulaires, sont donc trouées et incomplètes. Un atlas cellulaire plus précis, issu d'une caractérisation systématique et complète des cellules, permettra de mieux caractériser les états cellulaires, d'améliorer les tests diagnostiques, le développement de médicaments et l'ingénierie cellulaire, pour ne donner que quelques exemples. Aujourd'hui un test sanguin caractérise et dénombre les types cellulaires généraux; avec une meilleure connaissance de biomarqueurs indicatifs de sous-types cellulaires permettra de mieux diagnostiquer et donc d'augmenter l'efficacité des tests sanguins à diagnostiquer les maladies autoimmunes, les cancers ou infections avant même l'apparition des symptômes (Rozenblatt-Rosen et al., 2017). Les médicaments développés pourront avoir moins de toxicité non-intentionnée et la régénération de cellules, tissus et organes pourra être plus précise en médecine regénérative (Regev et al., 2017). Par exemple, en 2018 Montoro et collègues ont découvert que les voies respiratoires contenaient sept types cellulaires plutôt que six (Montoro et al., 2018). Ce nouveau type cellulaire est crucial dans la meilleure compréhension et au développement d'un traitement pour la fibrose kystique, une maladie génétique affectant 1 dans 3000 canadiens et canadiennes.

0.3. L'insaisissable définition du type cellulaire

Récemment, Regev et collègues ont suggéré que la définition de type cellulaire demeure vague (Regev et al., 2017). En effet, tandis que plusieurs sont d'accord sur la catégorisation des cellules par organe d'origine, les définitions plus précises sont souvent sujet à débats. Par exemple, la différence entre un type cellulaire et un état cellulaire est floue, vu que les cellules changent d'état souvent et changent même à l'occasion de type (Wagner et al., 2016), des exemples notables seraient le rythme circadien, le vieillissement et le cycle cellulaire. Le terme *type cellulaire* est donc considéré comme désuet par plusieurs, étant un vestige de la période où étudier les cellules était seulement possible en microscopie (Clevers, 2017). En effet, avec l'avenue de méthodes telles la cytométrie en flux et le séquençage à haut débit, de nouvelles façons d'étudier les cellules sont devenues possible.

Depuis, plusieurs études avancent que les types cellulaires existeraient dans un continuum de caractéristiques plutôt que des groupes discrets (Wagner et al., 2019; Xia et al., 2019; Bendall et al., 2014). Ici, le consensus est qu'il est temps pour un changement de paradigme dans la façon de voir les types cellulaires et plusieurs ont suggéré que la définition de types cellulaires devrait être entièrement appuyée sur les données (Clevers, 2017). Cette idée a mené à l'établissement de divers consortium, tel le *Human Cell Atlas* (Regev et al., 2017) ainsi que le Human Protein Atlas (Uhlen et al., 2015) pour n'en nommer que quelques-uns, pour avancer la création d'atlas cellulaires "nouveau genre".

0.4. L'atlas cellulaire - nouveau format

Il y a eu des changements proposés dans le format qu'un atlas cellulaire idéal devrait avoir, par rapport aux atlas d'hier. L'une des idées novatrices vient de Wagner et collègues, qui affirment que les cellules dans un atlas cellulaire devraient être représentées par des vecteurs de nombres (Wagner et al., 2016), où chaque vecteur encode le profil moléculaire capturant des informations biologiques de chaque cellule. Des exemples de profils moléculaires suggérés par les auteurs étaient i) l'appartenance de la cellule à des types discrets pré-déterminés, ii) l'état de la cellule à travers des phénotypes continus, tels l'inflammation et la différentiation, iii) la vacillation temporelle, telle la division cellulaire, l'activation et, finalement, iv) l'emplacement physique dans le corps et par rapport aux autres cellules (Wagner et al., 2016). Faisant écho à cette étude, plusieurs autres équipes se sont prononcées sur leur définition d'un type cellulaire et l'organisation des cellules dans les atlas cellulaires "nouveau genre" (Clevers, 2017). Certains ont suggéré que l'atlas cellulaire humain devrait être entièrement encodé dans un espace multidimensionnel s'appuyant sur les données, où les cellules similaires devraient être voisines et un état anormal ou malade pourrait être détecté par des différences dans les populations de cellules, des procédés cellulaires aberrants ou même des interactions cellulaires aberrantes (Ponting, 2019). En effet, Ponting suggère que cet atlas cellulaire pourrait établir des métriques de progression de maladies par le moyen de l'observation de décalages dans l'espace multidimensionnel de l'atlas et dans les distributions de populations cellulaires (Ponting, 2019). Plus qu'un simple catalogue d'éléments, l'atlas cellulaire devra également encoder des relations entre les cellules (Regev et al., 2017). Xia et collègues ont plutôt suggéré que l'atlas soit construit à la manière d'un tableau périodique, et à l'image du tableau périodique des éléments de Mendeleïev, ce dernier devrait être en mesure de prédire la présence de nouveaux sous-types cellulaires qui n'ont pas encore été caractérisés (Xia et al., 2019).

0.5. Hypothèse et objectifs

J'ai choisi pour objectif principal de ma thèse la création d'un atlas cellulaire basé entièrement sur des données de séquençage à haut débit.

Plus précisément, j'ai :

- créé un algorithme d'apprentissage automatique permettant à la fois l'encodage des cellules ainsi que des profils moléculaires (expression génique),
- étendu cet algorithme à des données de séquence d'ARN, et
- adapté cet algorithme à des données de profilage de répertoires adaptatif immun.

L'algorithme proposé apprend un espace d'encodage, qui sert d'atlas cellulaire et intègre les propriétés suivantes:

- Encode chaque échantillon dans un espace multidimensionnel, en s'appuyant entièrement sur les données de séquençage ARN, tel que proposé par (Wagner et al., 2016; Ponting, 2019).
- Permet l'observation de décalages entre les échantillons, sur la base de mesures d'expression de gènes individuels, similaire à la suggestion de (Ponting, 2019).
- Offre un espace dense, c'est à dire qui permet l'interpolation entre les échantillons, tel que suggéré par (Xia et al., 2019).

0.6. Organisation de la thèse

Le Chapitre 1 offre une revue de la méthodologie en transcriptomique et dans le Chapitre 2 le lecteur trouvera un entrée en matière pour l'apprentissage automatique. Le modèle *factorized embeddings* (FE) original et sa modification pour traiter des données de séquences ARN sont décrits dans deux articles contenus dans les Chapitres 3 et 4. Une brève introduction à l'immunologie adaptative et aux données de séquençage de récepteurs immuns peut être consultée au Chapitre 5, afin d'avoir tout le matériel nécessaire pour la lecture de l'article d'application du modèle FE à des données de séquençage de récepteurs immuns, au Chapitre 6.

Chapitre 1

Les données de séquençage: le jeu de données idéal pour l'atlas cellulaire

La création d'un atlas cellulaire moderne repose principalement sur les données de séquençages des cellules d'intérêt. Nous verrons comment s'articule en pratique le séquençage et les limites/difficultés qui y sont associées. Ce chapitre offre une revue des méthodes et idées déjà existantes pour la création d'atlas ainsi que quelques informations pratiques sur le séquençage haut débit de l'acide ribonucléique (ARN).

1.1. Le séquençage pour la construction d'atlas cellulaires

Dans la quête pour la compréhension de la haute diversité cellulaire dans le corps humain a mené à l'élaboration de projets tels le Human Protein Atlas (l'atlas humain de protéines) (Uhlen et al., 2015; Sud et al., 2017; Thul et al., 2017), où des équipes ont quantifié systématiquement la localisation cellulaire de plus de 12000 protéines, à travers 13 organelles majeures et dans 32 tissus et organes humains. La construction de cet atlas a requis un effort formidable, car pour chaque protéine et chaque cellule examinée, une expérience d'immunofluorescence a été faite. Ce travail a créé un outil utile; en effet, plusieurs maladies, telles le cancer, sont souvent caractérisées par des mauvaises localisations de protéines (Hegde et Zavodszky, 2019; Hung et Link, 2011; Laurila et Vihinen, 2009), et un tel atlas offre une comparaison commune et donc des notions sur les différences entre la maladie et la santé.

D'autres équipes se sont plutôt concentrées sur le séquençage d'ARN pour quantifier systématiquement les mêmes biomarqueurs. Récemment, un atlas cellulaire de souris a été publié où les auteurs ont développé une méthode rapide et peu coûteuse pour le séquençage à cellule unique (scRNASeq) nommée Microwell-seq, (Han et al., 2018). Dans cette étude, avec l'algorithme de réduction de dimensionnalité t-SNE (voir Section 2.2.2) les auteurs ont rapporté plus de 800 types cellulaires différents groupés en 98 partitions (Han et al., 2018).

Des travaux similaires ont été publiées avec la technologie scRNASeq ne se concentrant que dans des cadres spécifiques, tels le cancer du sein (Wagner et al., 2019), du rein (Park et al., 2018) et du pancréas (Muraro et al., 2016). Ceci indique qu'avec le temps, des jeux de données d'envergure seront disponibles pour leur intégration dans un atlas cellulaire commun. Tous ces atlas susmentionnés ont tous en commun l'inconvénient qu'ils sont plutôt des catalogues de résultats de quantifications. En effet, il leur manque un lien entre les protéines et gènes, une certaine structure sous-jacente qui lierait les biomarqueurs et permettrait de comparer directement les réseaux d'expression génique ou protéique entre les populations et types cellulaires. Xia et collègues ont suggéré une piste de solution: utiliser l'expression de facteurs de transcriptions afin de définir les états cellulaires variés (Xia et al., 2019). Leur raisonnement est que toutes les cellules partageraient une structure latente nommée un complexe régulateur essentiel de transcription génique (core regulatory complex, CoRC), qui est composé de facteurs de transcriptions qui sont exprimés dans les cellules et qui définissent le profil d'expression génique (Arendt, 2008). Les cellules pourront donc être définies en fonction de l'activation différentielle du CoRC et cet atlas cellulaire pourrait établir des liens entre les différentes cellules, traitant plutôt sur la base de patrons d'activation d'expression génique, plutôt que sur les différences d'expression de gènes uniques. Cette idée est supportée par le travail de Pierson et collègues dans leur études où ils ont observé que des facteurs de transcriptions tissu-spécifiques sont ce qui connecte ultimement l'expression génique au tissu (Pierson et al., 2015). Ils ont trouvé des gènes co-exprimés sont conservés entre les tissus et ont conclu que ces "programmes" de transcription sont catalysés par une activité de facteurs de transcriptions similaires. Ces deux travaux offrent une solution raisonnable pour repousser les limites des atlas cellulaires actuels limités à être des catalogues de quantifications.

1.2. Le jeu de données idéal pour l'atlas cellulaire humain

Naturellement, le jeu de données idéal est donc un transcriptome à cellule unique exhaustif de toutes les cellules du corps humain. Ce jeu de données n'existe pas encore mais la technologie de séquençage à cellule unique (scRNASeq) a été perfectionnée et plusieurs jeux de données ont été accumulés pour plusieurs tissus humains (Zheng et al., 2017). Regev et collègues ont adressé cette problématique de couverture du corps humain en jeux de données de séquençage dans leur white paper sur le consortium du Human Cell Atlas (l'atlas de la cellule humain) et l'un de leurs plans est la pratique de l'acquisition de données non-biaisées (Regev et al., 2017). L'une des sources de données déjà disponible est les consortiums de données massives dont j'offre un aperçu dans la sous-section suivante.

1.2.1. Les consortium de données pour la médecine personnalisée

De nos jours plusieurs traitements incluent des données génétiques afin de rendre compte des fines variations qui distinguent les êtres humains. Le terme *médecine personnalisée* a été inventé pour décrire ce type de traitement très spécifique au patient (Lu et al., 2014). Des consortium d'envergure tels The Cancer Genome Atlas (TCGA) et Genotype-Tissue Expression (GTEx) ont été créés afin d'étudier les cellules saines et malades en utilisant divers méthodes telles le séquençage génomique, transcriptomique ainsi que des mesures d'abondance de protéines (Weinstein et al., 2013; Lonsdale et al., 2013). À titre de référence, depuis ses débuts en 2006, TCGA contient maintenant au dessus de 85000 cas de cancer, la plupart ayant des données de séquençage multi-plateforme (ARN, mutations, quantification de protéines etc.). L'autre base de données d'envergure, GTEx contient au dessus de 15000 échantillons de tissus sains. L'objectif principal du consortium GTEx était d'explorer les bases de la maladie via des analyses d'expression de locul de caractères quantitatifs (expression of quantitative trait loci, eQTL), le consortium a également ajouté des données de transcriptomique à sa base de données (Lonsdale et al., 2013). Récemment, avec l'avènement du séquençage à cellule unique (single-cell RNA-Sequencing, scRNAseq), une définition encore plus précise des cellules: cette nouvelle méthode de séquençage permet de quantifier l'expression génique sur une base cellulaire et ajouter de la profondeur à des analyses cellulaires intra-tissu (Wagner et al., 2016; Zhang et Zhang, 2019) ainsi que des trajectoires de développement tissulaire (Kernfeld et al., 2018). Dans cette thèse, les données transcriptomiques issues des consortiums TCGA et GTEx ont été utilisées pour développer et évaluer les algorithmes pour la construction d'atlas dans les Chapitres 3 et 4.

1.3. Que peut-on faire avec des données brutes de séquençage?

Dans cette section de la thèse je survolerai les étapes de générales du traitement et analyse des données de séquençage d'ARN (RNA-Seq). Toute analyse débute avec du matériel génétique purifié à partir des cellules qui est séquencé. Les données de séquençage brutes sortant du séquenceur sont sous la forme de courtes séquences d'ADN séquencé (*read*) constituées d'un alphabet à 4 lettres, A,C,T et G, symbolisant les 4 bases nucléés: adénine, cytosine, thymine et guanine. À chaque base nucléée séquencée est associée une mesure de qualité de séquençage, encodée en un seul caractère ASCII (phred score) (Ewing et Green, 1998). Ce score sera utilisé par les logiciels d'alignement et quantification afin de déterminer la confiance de chaque nucléotide lors de l'assignation des correspondances. Le format décrit ci-dessus est le format du fichier FASTQ, le format le plus utilisé pour les données brutes de RNA-Seq.

1.3.1. L'analyse de séquences brutes en transcriptomique

Les méthodes d'analyse basées sur les k-mers se sont établies dans la dernière décennie dans divers domaines telles la quantification d'expression génique (Audemard et al., 2018), la détection de mutations (Nordström et al., 2013), l'estimation d'expression de transcrits Patro et al. (2014) et même la détection de familles de séquences répétées (Price et al., 2005) ou (pour les expériences de séquençage d'ADN) la résistance aux agents antimicrobiens (Drouin et al., 2016).

1.4. Qu'est-ce qu'un k-mer?

Pour une séquence donnée, un k-mer est n'importe quelle sous-séquence de nucléotides. Ainsi, pour une séquence S de longueur l, les k-mers de longueur k sont l'ensemble des sous-séquences de la position 1 à la position $l - k + 1$:

$$
\begin{array}{ll}
S & \text{AAATCGTAT} \qquad |\, S \,| = l \\
S_{1:4} & \text{AAAT} \\
S_{2:5} & \quad \text{AATC} \\
S_{3:6} & \qquad \text{ATCG} \\
S_{4:7} & \qquad\quad \text{TCGT} \\
S_{5:8} & \qquad\qquad \text{CGTA} \\
S_{6:9} & \qquad\qquad\quad \text{GTAT}
\end{array}
$$

Fig. 1.1. Énumération de tous les k-mers de longueur 4 pour la séquence exemple

Le k-mer est analogue au n-gramme utilisé principalement dans les applications de linguistique et traitement de la langue naturelle et théorie de l'information (Brown et al., 1992).

1.4.1. Le choix de la taille du k-mer

Le choix de la taille du k-mer k pour l'analyse est un exercice non-trivial, qui se base sur les trois principes suivants:

- la longueur du génome analysé
- la longueur de la séquence de discrimination
- la taille de la base de données de k-mers

Pour illustrer le problème du choix de la taille de k-mer, prenons l'exemple où deux individus ou échantillons sont comparés sur la base de leur profil de quantification de k-mers. Considérons premièrement l'exemple extrême où la longueur des k-mers choisie est de 1. Ceci n'est qu'une simple quantification du nombre de chaque nucléotide A,C,G,T dans le génome (Figure 1.2). Malgré que pour certaines applications la quantité de C/G dans la séquence est importante (Hurst et Merchant, 2001), dans notre exemple cette longueur de k-mer est peu

informative pour la distinction à l'intérieur d'une même espèce d'individus, où les génomes ne différent entre eux que par quelques bases. Si on considère l'autre extrême, c'est-à-dire un k-mer de longueur g où g correspond à la longueur du génome entier (6.4 milliard de bases pour l'humain), la quantification d'un g-mer sera également peu informative. En effet, chaque individu aura son propre g-mer (longueur g variant avec l'individu) et donc ne peut prendre les valeurs 0 ou 1, correspondant à l'identité de l'individu (Figure 1.2).

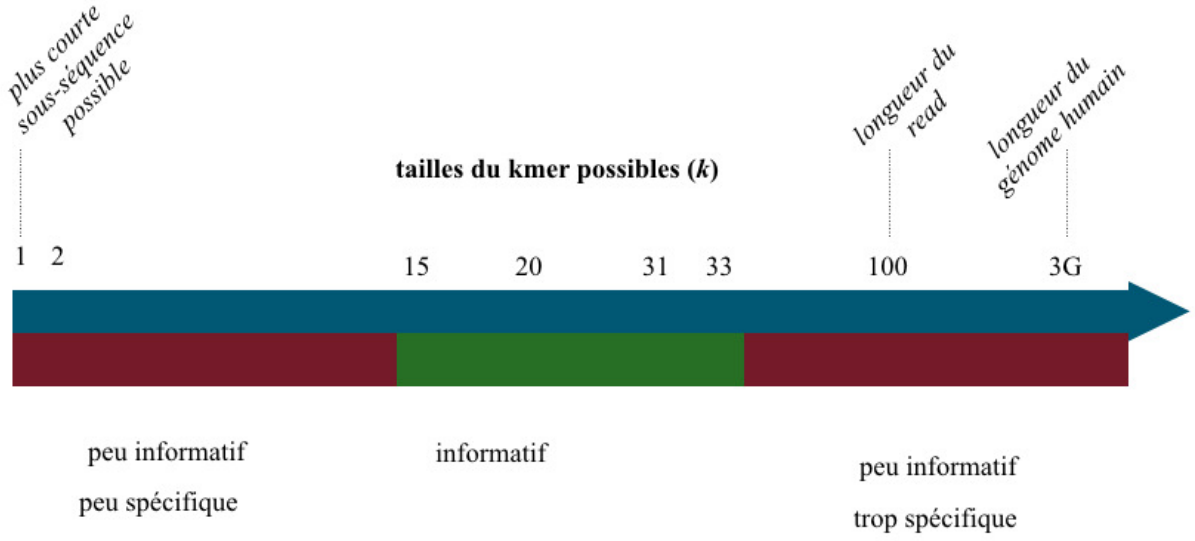


Fig. 1.2. Contenu en information du k-mer en fonction de la taille du k-mer

Dans la réalité, le contexte des applications dicte souvent le choix de la longueur du k-mer. Par exemple, dans les tâches de comparaison de génomes bactériens pour la détection de résistance aux antibiotiques, le génome entier est séquencé et donc les k-mers sont rarement quantifiés: les analyses se concentrent plutôt sur leur présence ou absence. Dans ces cas là, il est important de sélectionner une taille de k-mer qui garantit l'absence d'ambiguïté sur la région de correspondance du k-mer. Dans le cadre du RNA-Seq, les reads ont typiquement une longueur de 50 à 200 nucléotides. Un k-mer de longueur du read offrirait probablement une spécificité maximale sur la région de provenance du génome, vu que la fragmentation des séquences est aléatoire et donc il y a très peu de chances d'avoir exactement le même read.

Cependant, le séquençage à haut-débit est sujet aux erreurs, à un taux moyen de 0.1% par paire de base séquencée (Fox et al., 2014). Ceci signifie que dans une expérience de RNA-Seq à profondeur de séquençage de 30 millions de reads et à des reads de longueur 100, si chaque read contient au plus un seule erreur de séquençage, il y aura près de 3 millions de reads contenant au moins 1 erreur de séquençage. Chaque erreur de séquençage génère k k-mers contenant l'erreur. Et donc plus la longueur du k-mer augmente, plus le nombre

de k-mers totaux augmente, (borné par la longueur maximale du read) et ceci pèse dans la balance du nombre de k-mers qu'il est nécessaire de stocker. Naturellement, le choix de longueur du k-mer est également optimisé en fonction de l'espace de stockage accepté pour les données. Finalement, comme nous examinons un exemple où la comparaison de profils de quantification de k-mers est désirable, il faut s'assurer que la longueur du k-mer choisi garantisse que les deux génomes aient au moins une certaine quantité de k-mers en commun.

Le choix de la taille du k-mer est une délicate opération de compromis entre la taille de la base de données et la spécificité du k-mer à la position, ceci inhérent à la tâche d'analyse. Un exemple de taille du k-mer déterminée par la tâche, le logiciel Kover justifie de cette manière l'utilisation d'une longueur de k-mer de 31 (Drouin et al., 2016). En effet, Drouin et collègues s'attaquent au problème de résistance aux antibiotiques dans les bactéries. Dans leur cas, il faut que le k-mer soit assez spécifique pour éviter les ambiguïtés et potentiellement manquer des gènes de résistance. Par ailleurs, le choix de 31-mer est également utilisé par plusieurs autres outils et aurait rapport avec les architectures d'ordinateur modernes (Chikhi et Medvedev, 2014). En effet, certains expliquent qu'un 31-mer est le plus grand k-mer de longueur impaire pouvant être traduit en un entier 64-bit et comme les ordinateurs d'aujourd'hui manipulent les entiers 64-bit aisément, ceci explique son choix (Rizk et al., 2013). Chikhi et collègues ont même créé un outil nommé *k-merGENIE*, qui évalue de manière automatique la meilleure longueur de k-mer pour chaque génome (Chikhi et Medvedev, 2014). Dans cet article, ils examinent la meilleure taille de k-mer pour une application différente: l'assemblage de génome. Le choix de la taille du k-mer dans le contexte d'assemblage de génome est un problème bien décrit et il existe toute une littérature à ce sujet (Chikhi et Medvedev, 2014), cependant cette thèse ne traîte pas de ce sujet. La sous-section suivante offre un survol des méthodes de quantifications de k-mers qui puisent de ce domaine.

1.5. Les outils de quantification de k-mers

Tandis qu'une approche naïve à la quantification de k-mers est facile à implémenter, elle demeure grandement dépendante sur la longueur des séquences à quantifier et surtout du nombre de séquences. En effet, avec un fichier FASTQ contenant 30 millions de reads de 100 nucléotides de long, il y aura au plus 29 999 969 31-mers (probablement moins) qui devront être ajoutés un à la suite de l'autre dans une structure de données choisie.

Plusieurs approches parallélisées de plus en plus rapides ont été développées à travers les années. Par exemple, le logiciel *jellyfish* publié en 2011 utilise une table de hachage pour stocker les k-mers quantifiés en plus de paralléliser le décompte et donc qui permet de raccourcir grandement le temps de traitement des échantillons. *DSK*, en compétition avec *jellyfish*, utilise moins d'espace mémoire mais prend plus de temps de computation et donc est une bonne option pour le décompte de k-mers sur des machines à basse mémoire vive (Rizk

et al., 2013). En revanche, *BFCounter* utilise plutôt une structure de données nommée un filtre de Bloomau lieu de la table de hachage, s'appuyant sur l'idée que les k-mers uniques sont souvent peu informatifs, puisqu'ils sont probablement issus d'erreurs de séquençage (Melsted et Pritchard, 2011). Un autre logiciel, Reindeer, a vu le jour plus récemment, s'attaque plutôt au problème du temps de recherche de k-mers dans l'index (Marchet et al., 2020). En effet, dans cet article l'approche de l'indexation des k-mers se fait à travers plusieurs échantillons en même temps, ainsi raccourcissant le temps de recherche lors des analyses subséquentes, dans un contexte de comparaison de profils de quantification (Marchet et al., 2020). Pour plus de détails et les différences concrètes entre les méthodes de quantifications de k-mers, j'invite le lecteur intéressé à consulter le travail de Manekar et collègues, qui ont effectué une comparaison (benchmark) directe des méthodes de quantifications de k-mers les plus populaires et montrent entre autres les compromis entre les éléments principaux discutés ici: le temps de computation, l'espace de stockage, la mémoire utilisée (Manekar et Sathe, 2018).

1.6. Méthodes de quantification d'expression génique

Dans cette section du chapitre, il sera question des pipelines d'analyses générales des données transcriptomiques. Un logiciel d'alignement utilise un génome de référence pour déterminer les positions dans le génome d'où proviennent les reads. Pour des séquences ambiguës (par exemple, ayant plus d'une position possible), chaque logiciel d'alignement utilise sa propre stratégie. Par exemple, certains logiciels préfèrent jeter les séquences ambiguës, d'autres tranchent en se basant sur le nombre de nucléotides correspondant à l'annotation (Dobin et al., 2013; Weese et al., 2009; Kim et al., 2013). Ensuite, un logiciel de quantification est utilisé afin de quantifier le nombre de reads sur chaque gène dans l'annotation. Le logiciel de quantification utilise un fichier de référence, qui contient les coordonnées des gènes dans le génome et permet de quantifier l'abondance relative de chaque gène dans l'échantillon de séquençage.

Une autre option de traitement des données RNA-Seq est d'utiliser un logiciel de pseudo-alignement (Figure 1.3 B), tel Kallisto ou Sailfish (Bray et al., 2016; Patro et al., 2014). À la différence des logiciels d'alignement et quantifications, ce type de logiciel n'utilise qu'un fichier d'annotation et calcule directement l'abondance de chaque gène contenu dans le fichier d'annotation. Kallisto, par exemple, brise chaque read et chaque gène dans l'annotation en courtes sous-séquences chevauchantes (voir section 1.4). Ensuite, il compare les ensembles de k-mers entre le read et ceux de chaque séquence dans l'annotation et ainsi détermine les correspondances. Une fois les correspondances assignées, le logiciel calcule l'abondance relative des gènes basé sur le nombre de correspondances avec les reads.

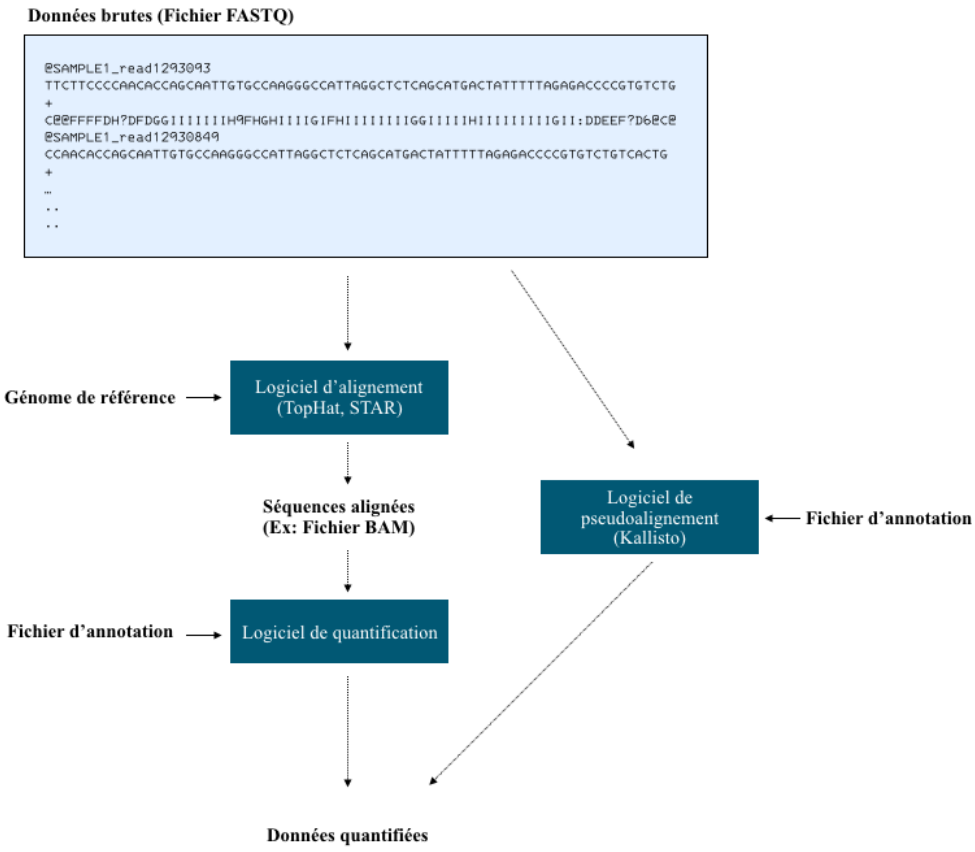


Fig. 1.3. Trajectoires de traitement des données RNA-Seq

Finalement, d'autres options de traitement de données existent: par exemple, RSEM ne nécessite pas un génome de référence et procède plutôt à un assemblage de novo, utile quand le génome de référence n'existe pas (Li et Dewey, 2011).

Peu importe le logiciel ou les étapes de quantification, une fois cette dernière faite, les données sont stockées sous la forme d'un tableau, où chaque rangée correspond à un échantillon et chaque colonne correspond à un gène ou transcrit quantifié. Chaque cellule dans le tableau store donc la valeur d'expression du gène dans l'échantillon correspondant. Le premier article présenté au Chapitre 3 de cette thèse travaille directement sur ce format de données et propose une méthode de construction d'atlas cellulaires à partir de données RNA-Seq quantifiées.

1.6.1. Les fichiers de référence

Les deux approches décrites ci-haut nécessitent un génome ou une banque de gènes de référence et donc ne permettent d'identifier que les séquences présentes dans l'annotation. En effet, typiquement un fichier d'annotation ne contient que les gènes codant pour des protéines (leur séquence ou leur coordonnées dans le génome de référence). Il est clair que lorsque le

protocole d'isolation d'ARN contient une étape de sélection pour les queues polyA, on peut s'attendre que les séquences isolées soient en majorité les séquences qui suivent les étapes de maturation d'ARN messager standard, c'est-à-dire les gènes codant pour les protéines (Zhao et al., 2014). Cependant, ne se concentrer que sur les quelques 20000 gènes codant pour des protéines offre une vision quelque peu biaisée de la biologie. En effet, une multitude d'information est contenue ailleurs, dans les séquences non-codantes, les intros, les microARN (miRNA) et autres régions. La plupart de ces protocoles de quantification d'expression géniques ne se concentraient jadis que sur la partie codante du transcriptome, qui constitue moins que 2% du génome(Huser et al., 2014). En effet, jusqu'à tout récemment, les 98% restants étaient considérés comme de "L'ADN-poubelle" ou "L'ADN-camelote" (Palazzo et Gregory, 2014). Plus précisément, c'est dans la dernière décennie seulement que plusieurs études ont découvert que ces régions étaient non-seulement transcriptionellement actives, mais aussi cruciales pour des procédés cellulaires variés tels la formation de placenta (Chuong, 2018), la maintenance d'état de cellules souches (Gautam et al., 2017) ainsi que la génération tumorale (tumorigenesis) (Kassiotis, 2014). Ces résultats témoignent du besoin de délaisser quelque peu les méthodes basées sur les références et plutôt se concentrer sur des méthodes basées sur les séquences elles-mêmes. Le deuxième article présenté dans au Chapitre 4 de cette thèse adapte le modèle proposé au Chapitre 3 à des quantifications de k-mers issus de données brutes RNA-Seq (non-alignées).

Chapitre 2

L'apprentissage automatique

Les algorithmes d'apprentissage sont des modèles statistiques qui, appliqués à des données, permettent de répondre à des questions quantitativement. Dépendant du type de questions, ils peuvent prédire, résumer ou trouver des patrons intéressants.

Il existe deux grandes catégories d'algorithmes d'apprentissage. Le type de données ou de question détermine quelle catégorie est choisie.

Les prochaines sections offrent un court résumé des principes utiles pour la compréhension de la présente thèse. Je n'aborderai que les thèmes généraux touchés par les articles présentées et au lecteur cherchant plus de détails, je suggère les ouvrages suivants: *Pattern Recognition and Machine Learning* par C.M. Bishop et *Deep learning* par I. Goodfellow, Y. Bengio et A. Courville (Bishop, 2006; Goodfellow et al., 2016).

2.1. L'apprentissage supervisé

L'apprentissage supervisé est un type d'algorithme d'apprentissage utilisé lorsque les exemples d'un jeu de données ont des étiquettes correspondantes. Très souvent, le but sera de prédire l'étiquette associée à de nouveaux exemples. Un exemple classique appliqué à la biologie pourrait être la tâche de prédiction d'un état (malade/sain, cancer/normal, etc) sur la base de l'expression génique. Dans cet exemple, les caractéristiques sont les valeur d'expression de tous les gènes, tandis que les étiquettes sont les états. L'exemple précédent est un exemple de modèle de *classification*, c'est à dire le modèle prédit la classe d'appartenance d'un exemple. Il existe également des modèles de *régression*, où l'étiquette prédite prend une valeur réelle. Un exemple classique appliqué à la biologie pourrait être un tâche de prédiction du coefficient de viabilité, suite au traitement des cellules d'un certain composé chimique, basé sur la concentration du produit chimique. Dans cet exemple, la concentration du composé chimique est la caractéristique, tandis que le coefficient de viabilité est l'étiquette à prédire. L'apprentissage supervisé nécessite pour chaque exemple des étiquettes associées. Formellement, chaque exemple est un vecteur $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_m)$ où chaque x_i représente

une caractéristique de l'exemple et y représente son étiquette. Un algorithme d'apprentissage supervisé va donc ajuster son modèle afin d'apprendre une fonction f qui lui permettrait de prédire précisément l'étiquette y n'utilisant que les données $\mathbf{x}$. Le modèle est ajusté via ses paramètres θ. Ainsi, la prédiction de l'étiquette y à partir des données $\mathbf{x}$ se fait de la manière suivante:

$$f_\theta(\mathbf{x}) = y.$$

Souvent, la fonction idéale f est seulement approximable et donc en réalité le modèle n'apprendra qu'une approximation de cette fonction: $\hat{f}$. Le modèle va donc ajuster ses paramètres, en d'autres termes *apprendre*, en comparant sa prédiction $\hat{y}$ à la vraie étiquette y. Les sous-sections suivantes introduiront quelques algorithmes d'apprentissage supervisés communs en bioinformatique utilisés dans cette thèse.

2.1.1. Régression logistique

Le modèle de régression logistique est un modèle de classification linéaire. Dans le cadre d'une classification binaire (problème de classification à deux classes), le modèle de régression logistique calcule la probabilité que l'exemple $\mathbf{x}$ soit de classe c_1, étant donné les données $\mathbf{x}$:

$$p(\hat{y} = c_1 \mid \mathbf{x}).$$

Ainsi, le modèle émet sa prédiction de la manière suivante:

$$\sigma(W^T\mathbf{x} + \mathbf{b}) = \hat{y},$$

où σ représente la fonction logistique sigmoïde et les paramètres du modèle sont une matrice de poids W, où chaque poids individuel correspond à l'une des valeurs d'entrée dans le vecteur de l'exemple ainsi qu'un vecteur de biais b.

2.1.2. Perceptron multi-couches

Le perceptron multi-couches (MLP, de l'anglais *Multi-Layer Perceptron*)est un modèle d'apprentissage supervisé permettant une classification ou une régression. Similaire au modèle de régression logistique (Section 2.1.1), ce modèle, lui aussi, apprend une matrice de poids W lui permettant de transformer les données d'entrée. Cependant, à la différence de la régression logistique, ce dernier possède des "couches cachées", c'est-à-dire qu'il y a plus d'une étape de transformation entre les données et la cible et donc plus d'une matrice W. Pour un MLP avec un nombre de couches n, il y aura $n + 1$ matrices de poids. Ainsi, pour un modèle à 1 couche cachée, deux matrices W_1 et W_2 seront apprises. De plus, une *fonction d'activation* ϕ transforme les données de manière non-linéaire à chaque couche cachée. La fonction logistique sigmoïde est un exemple de *fonction d'activation* non-linéaire. Les

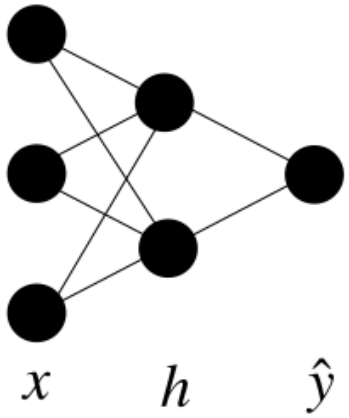


Fig. 2.1. Perceptron multi-couches à 1 couche cachée

transformations entre l'entrée et la sortie du modèle iront donc comme suit:

$$W_1^T x = h$$

$$\phi(h) = \tilde{h}$$

$$\tilde{h}^T W_2 = \hat{y}$$

Chaque unité de la couche cachée s'appelle *neurone* et ceci réfère au nombre de connections entre chaque caractéristique d'entrée et la couche cachée (Figure 2.1). En général, un modèle augmente de capacité soit en augmentant le nombre de couches cachées ou en rendant les couches cachées plus larges. Par exemple, le modèle de régression logistique est très limité car il n'a pas de couche cachée et donc est incapable d'apprendre des relations non-linéaires entre les caractéristiques. Ceci n'est pas du tout le cas pour le MLP, qui peut être de profondeur (et largeur) variable. En fait, les récents succès de l'apprentissage profond (*Deep Learning*) réfèrent spécifiquement aux modèles à plusieurs couches cachées, d'où le nom de modèle profond.

2.1.3. Les modèles traitant les données séquentielles

Des réseaux de neurones artificiels (ANN, de l'anglais *Artificial Neural Network*) ayant des architectures spécialement conçues pour l'analyse de séquence sont utilisées: le réseau à convolution (CNN, de l'anglais *Convolutional Neural Network*) et le réseau à récurrences (RNN, de l'anglais *Recurrent Neural Network*).

Introduits dans les années 80, les CNN se basent sur l'idée que les items dans une séquence ont une certaine relation avec leur voisinage (Fukushima, 1980; LeCun et al., 1999).

Ces réseaux s'inspirent librement de la structure du cortex visuel des mammifères, la découverte de laquelle a mérité un prix Nobel à l'équipe de Hubel et Wiesel (Hubel et Wiesel, 1962). Ces ANN traitent une partie de la séquence à la fois et pour ce faire remplacent la matrice de poids standard entièrement connectée W d'un MLP par une matrice de plus petite taille, qui correspond à la taille de la sous-séquence traitée. Ensuite une opération de *pooling* (mise en commun) est effectuée, telle le calcul de la moyenne sur un sous-ensemble voisins de noeuds (LeCun et al., 1999). En d'autres mots, les données sont traités par des fenêtres ne se concentrant que sur une sous-séquence à la fois. Ceci permet la détection de caractéristiques se répétant dans la séquence et cette opération est invariante aux translations, en d'autres mots, un patron peut être détecté n'importe où dans la séquence. Une application en génomique notable de ce type de modèles est DeepSEA, où le CNN détecte l'effet de variations génétiques non-codantes sur la liaison de facteurs de transcription, l'accessibilité de l'ADN et les marqueurs d'histones avec précision au nucléotide près (Zhou et Troyanskaya, 2015).

Les RNN, eux, sont une famille d'ANN qui modélisent une séquence temporelle à l'intérieur de la séquence (Rumelhart et al., 1986). À la différence du MLP, le RNN partage le vecteur de poids W à travers tous les items dans la séquence, plutôt qu'apprendre un vecteur de poids W différent pour chaque item. Ceci permet d'apprendre des patrons récurrents. Des modifications au modèle, tels les Mémoires à Long-Court Terme (LSTM, de l'anglais *Long Short Term Memory*) ont été proposés plus tard qui permettent une meilleure persistance des patrons dans la mémoire, et surtout utilisés pour de longues séquences (Hochreiter et Schmidhuber, 1997). Une application des RNN en transcriptomique notable est la détection de sites de liaison de facteurs de transcription (Shen et al., 2018). Vu que les facteurs de transcription ont des séquences de préférence, les auteurs arrivent à détecter les sites de liaison par la détection de ces patrons spécifiques dans les séquences d'ADN avec les RNN. Plus de détails sur les diverses architectures et leur bien fondé peut être trouvé dans l'ouvrage (Goodfellow et al., 2016).

2.1.4. La fonction de coût

Dépendant du type de modèle d'apprentissage, une *fonction de coût* sera choisie. La fonction de coût est une fonction qui permet d'évaluer à quel point un modèle a bien appris, en comparant ses prédictions aux étiquettes. Le modèle cherche à minimiser ses erreurs et donc de réduire cette fonction en trouvant des bons paramètres. Par exemple, le modèle de régression logistique et le MLP de classification binaire sont entraînés par descente de gradient et une erreur de coût C utilisée est la log-vraissemblance négative (*Negative Log Likelihood*, aussi appelée entropie croisée ou *cross-entropy*). Cette dernière est calculée de la

manière suivante:

$$C(y,\hat{y}) = -y \ln \hat{y} - (1-y) \ln(1-\hat{y}).$$

Ainsi, pour un jeu de données X, le calcul de l'erreur prend la forme:

$$\text{NLL} = \frac{1}{N} \Sigma_{i=1}^{N} (-y_i \ln \hat{y}_i - (1-y_i) \ln(1-\hat{y}_i)).$$

Vu que c'est une classification binaire, les valeurs de $y \in \{0,1\}$. L'erreur sera maximale quand le modèle prédit une valeur à l'opposé de l'étiquette attendue. Par exemple, si $y = 1$ et le modèle prédit une valeur $\hat{y} = 0.01$, l'erreur calculée sera de $ln(0.01)$. Tandis que si plus la valeur prédite est proche de la valeur attendue, l'erreur sera moindre. C'est cette erreur calculée qui est utilisée par la descente de gradient afin d'ajuster les paramètres W et b du modèle et obtenir une meilleure prédiction. Dans un contexte de régression, c'est à dire où le modèle prédit une valeur plutôt qu'une classe, la fonction de coût la plus communément utilisée est l'erreur quadratique moyenne (MSE, de l'anglais *Mean Squared Error*).

2.1.5. L'optimisation

Plusieurs modèles sont entraînés en utilisant l'algorithme de rétropropagation. Le modèle est entraîné par rétro-propagation en deux phases subséquentes décrites ci-dessous. C'est un processus itératif, où le modèle émet tout d'abord une prédiction, puis ajuste ses paramètres (poids des matrices W et b pour un MLP) proportionnellement à l'erreur calculée entre la prédiction et la valeur de l'étiquette. Afin de mieux suivre le processus, j'utiliserai la lettre t pour désigner une étape d'optimisation spécifique et $t-1$ et $t+1$ sont respectivement l'étape d'avant et celle d'après. Tout d'abord, le modèle émet une prédiction $\hat{y}$. L'erreur entre la valeur prédite $\hat{y}$ et la valeur attendue y est calculée et pour chaque matrice, un ajustement ΔE est calculé par différentiation afin de mieux prédire l'ajustement nécessaire à la matrice actuelle W_t. La matrice de poids W_t est actualisée à chaque étape en modifiant celle de l'étape précédente (W_{t-1}):

$$W_t = W_{t-1} - \mu \Delta E(W_{t-1}),$$

où $\mu > 0$ est le pas d'apprentissage (*learning rate*), une valeur scalaire qui contrôle la taille de l'effet de l'ajustement ΔE calculée. Plus d'information peut être trouvée dans l'ouvrage (Bishop, 2006) dans le chapitre 5.3 à ce sujet.

2.1.6. Le surapprentissage

L'arrêt de l'optimisation est basé sur un certain critère d'arrêt prédéfini. Par exemple, l'optimisation peut s'arrêter si l'erreur calculée tombe à zéro ou bien si elle descend en dessous d'un certain seuil. D'autres critères d'arrêt populaires relèvent de la notion de généralisation. Dans le cadre de tout modèle, on assume que les données d'entraînement sont issues d'une certaine distribution D inconnue. Le modèle f entraîné tente donc d'approximer cette

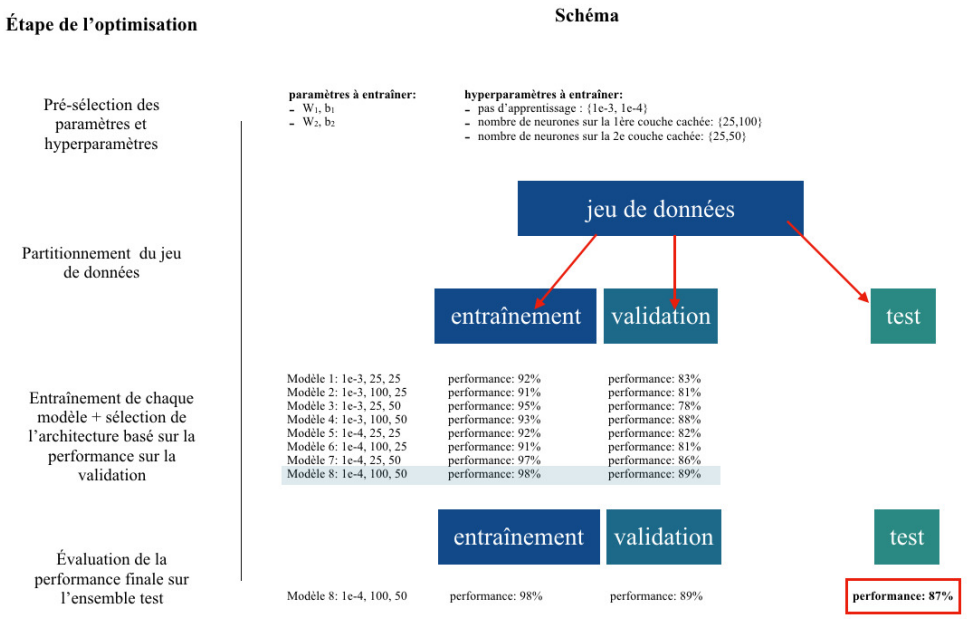


Fig. 2.2. Schéma des étapes de la recherche d'hyperparamètres

distribution D. Cependant, comme les données ne sont qu'un échantillon tiré de la distribution D inconnue, le seul moyen de vérifier que le modèle f approxime bien cette distribution est en tirant d'autres échantillons de la même distribution et en évaluant le modèle sur ces nouvelles données. On dit d'un modèle qui performe bien sur ces nouvelles données qu'il *généralise* bien. En réalité, il est souvent impossible d'obtenir plus d'échantillons à partir de la même distribution, c'est notamment très souvent le cas en bioinformatique. On effectue donc un partitionnement de données en groupes qu'on nomme le groupe d'entraînement et le groupe de validation. Le modèle n'est entraîné que sur le groupe de données d'entraînement pour ensuite être validé sur le groupe de validation. Ceci peut être utilisé pour un critère d'arrêt de l'optimisation, par exemple l'optimisation arrête au moment où le modèle arrête de bien performer sur le groupe de données de validation. Lorsqu'un modèle de généralise pas (faible performance sur les données de validation) bien mais a une bonne performance sur les données d'entraînement est appelé *surapprentissage*. Le surapprentissage n'est pas souhaitable et il existe plusieurs manières de le mitiger, telles les méthodes de régularisation (Chicco, 2017). J'invite le lecteur intéressé à consulter le Chapitre 7 du *Deep Learning Book* (Goodfellow et al., 2016).

2.1.7. Les hyperparamètres

Outre les paramètres du modèle (tels les matrices de poids W dans les ANN), il y a d'autres paramètres externes qui persistent tout au long de l'entraînement du modèle mais dont le choix peut influencer la performance du modèle et ainsi nuire à la généralisation. Ces

paramètres sont appelés *hyperparamètres*; des exemple sont le *learning rate*, le nombre de neurones dans chaque couche cachée du ANN, etc. Typiquement les données sont partitionnées en trois groupes: le groupe d'entraînement, de validation et de test. Chaque modèle est entraîné sur les données d'entraînement et validé sur les données de validation (Figure 2.2). Ensuite, la performance de chaque modèle sur les données de validation est comparée et les hyperparamètres sont choisis sur cette base. La performance finale du modèle est calculée sur les données test, jamais vues auparavant par le modèle (Figure 2.2). Ces étapes sont une recherche d'hyperparamètres standard qui explore toutes les possibilités. Typiquement, les recherches d'hyperparamètres se font soit par recherche aléatoire, recherche dirigée en grille (*grid search*) (Bergstra et Bengio, 2012) ou encore bayésienne (Snoek et al., 2012; Klein, 2017). La variation dans l'entraînement due à la sélection des hyperparamètres influe grandement sur la performance finale du modèle (Bouthillier et al., 2021). Une recherche d'hyperparamètres est donc de mise afin de trouver les meilleures valeurs et garantir une meilleur généralisation.

2.1.8. Classifieur de forêt d'arbres décisionnels

Un autre modèle d'apprentissage supervisé qui n'entre pas exactement dans le même cadre conceptuel que les réseaux de neurones (il est dit non-paramétrique) est l'arbre de décision. L'arbre décisionnel est un modèle qui construit un arbre de décision à partir des caractéristiques des données, afin de mieux classifier les données. Ce modèle est non-linéaire et utilise un critère d'impureté, tel l'entropie de Gini Gini (1936), pour partitionner l'espace des données. Itérativement, le modèle sélectionne une caractéristique d'entrée et sélectionne une valeur seuil qui minimise l'index de Gini. Par exemple, dans la Figure 2.3 on peut voir différentes positions possibles pour le seuil pour les caractéristiques x_1 et x_2 (Figure 2.3 A et B). L'index de Gini peut prendre une valeur entre 0 et 1, où 0 signifie que le noeud est pur, c'est à dire qu'il n'y a pas de mélange de classes dans le noeud.

Le modèle construit ainsi un arbre de décision binaire (Figure 2.3 D), où les feuilles sont des classes prédites et les noeuds intermédiaires correspondent à des seuils de classification. Ainsi, pour un nouvel exemple, l'arbre de décision est parcouru de la racine aux feuilles, en choisissant le chemin correspondant sur chaque noeud parcouru et une fois la feuille atteinte, celle-ci prédit la classe du noeud.

La forêt d'arbres décisionnels est un modèle d'ensemble, c'est à dire que plusieurs arbres décisionnels seront construits et la classe finale sera décidée par un système de vote parmi les arbres construits. Comme tous les modèles d'ensemble, la forêt d'arbres décisionnels permet d'obtenir une classification robuste aux variations dues au bruit dans le modèle et à l'initialisation (Chen et Ishwaran, 2012; Qi, 2012). De plus, la forêt d'arbres décisionnels peut également être un modèle de régression, c'est à dire qu'elle peut prédire une valeur au

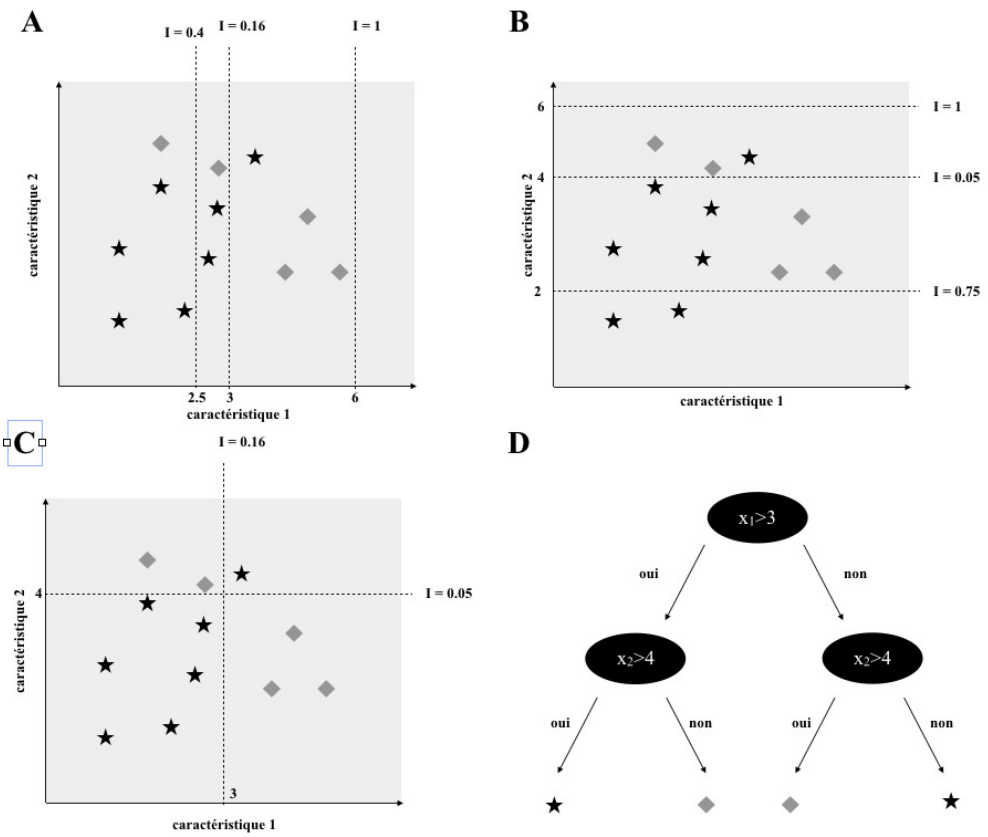


Fig. 2.3. Calcul d'un arbre décisionnel à deux caractéristiques

lieu d'une classe. J'invite le lecteur intéressé par cette modification à consulter (Chen et Ishwaran, 2012) pour plus de détails. Ce modèle a été utilisé dans le chapitre 6 pour prédire de quelle couche proviennent les TCR.

2.2. L'apprentissage non-supervisé

L'autre grande catégorie d'apprentissage est l'apprentissage non-supervisé. Ici, il n'y a pas d'étiquettes associées aux données. L'algorithme ajuste un modèle tentant de trouver une certaine structure sous-jacente dans les données. Deux méthodes pour trouver des structures sous-jacente dans des données sont le partitionnement et la réduction de dimensionnalité.

2.2.1. Le partitionnement

Les algorithmes de partitionnement (*clustering*) permettent de grouper les exemples dans des groupes définis sur la base d'une mesure séléctionnée. Par exemple, le partitionnement *k-moyennes* (*k-means clustering*) requiert à l'utilisateur de choisir d'avance combien de groupes (k) sont présents dans les données pour ensuite sur la base d'une mesure de distance inter-échantillon (par exemple la distance euclidienne) les grouper en k groupes. Le

partitionnement hiérarchique (*hierarchical clustering*), quant à lui, a une approche ascendante (bottom-up), où des groupes de plus en plus gros sont formés en groupant des paires d'échantillons et de groupes itérativement, toujours sur la base d'une mesure de distance. L'agglutination des groupes avec les autres peut se faire sur la base de plusieurs mesures. Ici je détaillerai brièvement les plus communément utilisées en bioinformatique. La fonction de partition UPGMA (*Unweighted Pair Group Method with Arithmetic mean*) agglomère ensemble deux entités basé sur la distance entre leur moyennes respectives (Sokal et Michener, 1958). À la différence, l'algorithme *single-linkage* agglomère les entités basé sur une distance minimale de deux points uniques (Gower et Ross, 1969). Ces deux algorithmes d'agglomération ont des effets différents sur la structure obtenue du dendrogramme hiérarchique. Une fois le dendrogramme obtenu, il est "coupé" soit basé sur la hauteur ou en un certain nombre de partitions choisies. Le double-partitionnement hiérarchique est souvent utilisé en analyse transcriptomique, où il est désirable d'obtenir à la fois un partitionnement des échantillons et des gènes.

2.2.2. La réduction de dimensionnalité

Les algorithmes de réduction de dimensionnalité, quant à eux, permettent de réduire de manière intelligente la dimensionnalité des données. L'algorithme apprendra donc une fonction f qui réduira un jeu de données avec N exemples de M dimensions en un jeu de données en d dimensions, où $d < M$. Par exemple, dans un jeu de données de RNA-Seq chaque exemple est un vecteur de dimension M, où M est le nombre de gènes dont l'expression a été mesurée, ce qui est souvent entre 20000 et 56000. Les algorithmes de réduction de dimensionnalité permettent de réduire chaque exemple à un vecteur plus petit, par exemple de deux dimensions ($d = 2$), dans ce cas-ci permettant la visualisation.

De nombreux algorithmes de réduction de dimensionnalité existent, les plus communs en transcriptomique étant l'analyse en composantes principales (*principal components analysis, PCA*), l'encodage stochastique t-distribué de voisins (*t-distributed stochastic neighbor embedding, t-SNE*) et l'approximation et projection uniforme de la variété géométrique (*Uniform Manifold Approximation and Projection, UMAP*) (Van Der Maaten et Hinton, 2008; McInnes, 2008).

L'algorithme PCA place chaque exemple des données d'entrée dans un nouveau système de coordonnées, où chaque dimension rend compte de la variance dans les données, les dimensions triées en ordre décroissant de variance. Pour des fins de visualisation, souvent que les deux premières dimensions sont utilisées. Cet algorithme est souvent la porte d'entrée pour la plupart des analyses de données transcriptomiques. Il est rapide à calculer et des

implémentations existent dans la plupart des languages communément utilisés en bioinformatique. Son défaut et qu'il est linéaire et donc si les données ont une structure sous-jacente non-linéaire, cet algorithme ne pourra pas la trouver.

Les algorithmes t-SNE et UMAP sont tous les deux des réductions de dimensionnalité non-linéaires. L'algorithme t-SNE optimise un encodage des données dans un espace à basse dimensionnalité, tentant de préserver des distance entre des paires d'échantillons (Van Der Maaten et Hinton, 2008). Pour sauver du temps, t-SNE n'optimise pas toutes les paires de distances et donc ne préserve que des relations dites *locales*. L'algorithme UMAP, quant à lui, procède en deux étapes (McInnes, 2008). La première étape consiste à construire un graphe de voisinage, basé sur l'hyperparamètre v, où chaque exemple est un noeud et une arrête les connecte si ils sont assez proches voisins (défini via l'hyperparamètre de voisinage, où, chaque exemple n'aura au maximum que v voisins et donc un maximum de v connections dans le graphe). Ensuite, la seconde étape consiste à encoder les exemples dans un espace de d dimensions tout en préservant la structure du graphe, ce qui est fait par descente de gradient. Les deux algorithmes sont optimisés de manière stochastique et l'encodage des données appris varie d'une exécution de l'algorithme à l'autre. De plus, les résultats des deux algorithmes dépendent de leur hyperparamètres respectifs et donc en variant ces derniers de très différentes représentations peuvent être obtenues (détails sur les hyperparamètres dans la section 2.1.6 ci-dessus). Une récente publication a même fait le lien entre les deux algorithmes, suggérant qu'ils sont équivalents à certains hyperparamètres (Böhm et al., 2020). En analyse transcriptomique, les algorithmes les plus utilisés pour la réduction de dimensionnalité des vecteurs d'expression génique sont les algorithmes UMAP, t-SNE et PCA. L'encodage des entités biologiques se fait donc principalement utilisant ces trois algorithmes ainsi que ceux qui seront détaillées dans les prochaines sections. Ces algorithmes sont donc de bons points de départ pour l'encodage de cellules basé sur leur transcriptome en en espace numérique commun.

2.3. Inférence des bases de connaissances

"Nous nous noyons dans l'information mais avons faim de connaissances" a écrit Naisbitt dans son livre *Megatrends*. L'utilisation de grandes quantités de données pour faire des prédictions intelligentes est l'un des buts centraux de l'intelligence artificielle. La compilation des données massives en bases de données (DB, de l'anglais *database*) est la pierre d'assise de plusieurs approches en intelligence artificielle. Dans le domaine de la bioinformatique, plusieurs DB existent pour des domaines variés de la biologie, tels la structure de protéines (Berman et al., 2000), l'immunologie(Vita et al., 2015, 2019a) , la génomique (Clark et al., 2016) et la génétique (Landrum et al., 2014). Certes ces DB offrent un espace de stockage d'informations mais la capacité de faire des inférences et recommandations leur manque et

on ne peut donc pas les qualifier d'atlas et selon Huser et collègues, ceci serait en partie dû au manque de données au niveau patient (Huser et al., 2014). L'organisation des données en bases de connaissances, les KB (de l'anglais *Knowledge Bases*) a permis l'élaboration des systèmes experts, des algorithmes aidant aux décisions, basés sur les données (Levesque et Lakemeyer, 2001). En traitement de la langue naturelle, Bordes et collègues ont suggéré une manière d'encoder des données de KB dans des encodages structurés (structured embeddings), où chaque entité serait représentée par un vecteur de nombres réels de d dimensions (Bordes et al.). Dans cet article, les auteurs ont entraîné un modèle qui apprend à prédire des triplets de la forme ("entité"-"relation"-"entité"). Ce type de modèle serait idéal pour la création de nouvelles connaissances, vu qu'il supporte la recherche directement dans l'espace de vecteurs, lorsqu'un des éléments du triplet est manquant. L'algorithme présenté dans le chapitre 3 de cette thèse se base sur l'idée proposée par Bordes et collègues (Bordes et al.), où deux types d'entités sont encodés: l'individu et le gène. La relation entre les deux reste fixe: c'est l'expression du gène dans l'individu.

2.3.1. L'encodage d'entités biologiques

Une contribution substantielle au domaine du traitement de la langue naturelle a été faite sous la forme de vecteurs pré-entraînés de représentation de mots (Mikolov et al., 2013; Pennington et al., 2014). Ces deux modèles utilisent un très large corpus de texte (Google Books ou similaire) afin de s'entraîner basé sur des co-occurences de mots ou séquences. Il a été observé que les vecteurs de représentations entraînés de cette manière capturent des relations entre les mots et offrent la possibilité d'effectuer de l'arithmétique vectorielle directement dans l'espace de représentation. Un exemple frappant observé par Mikolov et collègues est la similarité entre le vecteur liant *Rome* et *Italie* et les vecteurs liant les capitales à leur pays d'origine respectifs (Mikolov et al., 2013). Ceci les a menés à conclure que la représentation des mots capturait un sens, cohérent avec *l'hypothèse distributionnelle* en linguistique, qui avance que les mots ayant un sens similaire, se retrouveront dans des contextes similaires (Harris, 1954). Les représentations apprises sont donc qualifiées de représentations *distribuées*.

Inspirés par ces résultats, des modèles tels *BioVec* (Asgari et Mofrad, 2015), *dna2vec* (Ng, 2017), *seq2vec* (Kimothi et al., 2016) ainsi que *gene2vec* (Du et al., 2019) ont été conçus. Asgari et collègues ont adapté l'algorithme décrit par (Mikolov et al., 2013) à des tri-grammes de protéines et ont rapporté que l'espace d'encodage ainsi entraîné capture des propriétés biochimiques des acides aminés, tels l'hydrophobicité, l'encombrement stérique et la polarité (Asgari et Mofrad, 2015). Kimothi et collègues ont ensuite amélioré *BioVec* en testant leur version d'une extension de *word2vec* (Mikolov et al., 2013) basé sur la séquence ADN sur une tâche de classification de familles de protéines et ont rapporté une performance quasi

parfaite n'utilisant que des k-mers de taille 3 (Kimothi et al., 2016). De la même manière, Ng et collègues ont entraîné un espace d'encodage de k-mers d'ADN et ont montré que la similarité cosinus entre des paires de k-mers est proportionelle à leur score de similarité Needleman-Wunsch, c'est à dire que le modèle gardait la distance entre les k-mers dans l'espace d'encodage proportionelle à leur similarité de séquence (Ng, 2017).

Du et collègues se sont plutôt concentrés à entraîner des encodages de gènes *gene2vec*, basé sur des données de co-expression génique (Du et al., 2019). Ils rapportent que les représentations distribuées apprises par leur modèle sont cohérentes avec des interactions de gènes et capturent même une certaine notion du type de gène (codant pour une protéine, lncRNA etc.). Par ailleurs, Choy et collègues ont entraîné un réseau de neurones artificiels peu profond pour représenter à la fois des gènes et des patients dans un espace à haute dimension Choy et al. (2019). Ils ont construit leur modèle de sorte à ce que le produit scalaire entre le vecteur d'un gène et le vecteur d'un patient soit prédictif de l'expression génique de ce gène dans ce patient. Leurs résultats montrent qu'ils sont en mesure de partitionner les cancers basé sur leur expression génique en méta-groupes qui peuvent ensuite être utilisés pour prédire les patients répondant à de la thérapie d'inhibiteurs de point de contrôle immuns (immune checkpoint therapy) Choy et al. (2019).

Chapitre 3

Factorized embeddings learns rich and biologically meaningful embedding spaces using factorized tensor decomposition

Assya Trofimov[1,2,3], Joseph Paul Cohen[2,3], Yoshua Bengio[2,3], Claude Perreault[1,4,5], Sébastien Lemieux[1,6]

([1]) Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([2]) Department of Computer Science and Research Operations, Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([3]) Montreal Institute for Learning Algorithms (Mila), Montreal, Quebec H2S 3H1, Canada
([4]) Department of Medicine, Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([5]) Maisonneuve-Rosemont Hospital, Montreal, Quebec H1T 2M4, Canada
([6]) Department of Biochemistry at University of Montreal, Université de Montréal, Montreal, Quebec H3C 3J7, Canada

Cet article a été publié dans la revue *Bioinformatics* de Oxford Journals en juilet 2020. Les résultats de l'article ont également été sélectionnés pour une présentation orale à la conférence Intelligent Systems in Molecular Biology (ISMB) 2020.

3.1. Mise en contexte

S'inspirant du travail de Bordes et al., qui a suggéré d'encoder des relations entre des entités (Bordes et al.), nous nous sommes demandé s'il est possible de construire un modèle qui apprend deux espaces d'encodage d'entités, où l'une serait les échantillons et l'autre serait les gènes. Nous avons donc choisi de travailler sur des données de séquençage d'ARN (RNA-Seq) issus de deux consortiums: *Genotype-Tissue Expression* (GTEx) et *The Cancer Genome Atlas* (TCGA), contenant des échantillons RNA-Seq issus respectivement de tissus sains et d'échantillons tumoraux. Cet article présente la toute première itération du modèle *Factorized Embedding*.

Nos résultats montrent que les espaces d'encodage appris par le modèle étaient riches en information: i) l'encodage des patients groupait les échantillons par tissu d'origine et l'espace d'encodage était organisé basé sur l'expression de divers gènes, ii) l'espace d'encodage des gènes regroupait les gènes co-exprimés et participant à des fonctions similaires, ainsi que des groupes de gènes spécifiquement exprimés dans certains tissus. Nous avons démontré qu'il est également possible d'effectuer de l'interpolation entre les points dans l'espace d'encodage, ce qui sépare le modèle FE des autres réductions de dimensionnalité communément utilisées en biologie computationelle. De plus, nous avons comparé les espaces d'encodage d'échantillons appris par FE à trois autres algorithmes de réduction de dimensionnalité sur un total de 32 tâches de classification et régression. Nos résultats montrent que FE se retrouve la plupart du temps en première position de performance pour presque toutes les tâches testées. Ces résultats nous ont permis de conclure que l'architecture de FE est une architecture candidate pour la construction d'atlas cellulaires multidimensionnels basés sur les données de transcriptomique.

3.2. Contributions

Assya Trofimov: A mené le projet, conçu et effectué les expériences, préparé toutes les figures et a écrit l'article.
Joseph Paul Cohen: A effectué des expériences et a participé à l'écriture de l'article.
Yoshua Bengio: A conçu les expériences, participé à des discussions et a participé à l'écriture de l'article.
Claude Perreault et Sébastien Lemieux: Ont conçu et dirigé les expériences, ont

participé aux discussions et analyzé les données, et ont écrit l'article.

3.3. Résumé en français

Motivation: Les récents développements en séquençage ont révolutionné notre compréhension des activités cellulaires internes et la façon de traiter les maladies. Cependant, une seule expérience de séquençage d'ARN (RNA-Seq) mesure des dizaines de milliers de paramètres simultanément. Malgré la richesse en information des résultats, l'analyse de donnée est non-triviale. Les méthodes de réduction de dimensionnalité aident avec cette tâche, en extrayant des patrons à partir des données, en les compressant dans des représentations de vecteurs compactes.

Résultats: Nous présentons le modèle *Factorized Embeddings* (FE), un algorithme auto-supervisé d'apprentissage profond qui apprend simultanément, via une factorization de tenseurs, des espaces de représentation des gènes et des échantillons. Nous avons testé notre modèle sur des données RNA-Seq de deux consortiums et avons observé que l'encodage des échantillons capture à la fois de l'information sur des patrons d'expression géniques globaux et sur la base de gènes individuels. De plus, nous avons observé que l'espace d'encodage des gènes était organisé de sorte à ce que les gènes tissu-spécifiques et hautement correlés, ainsi que les gènes participant aux mêmes fonctions (GO terms) soient regroupés dans l'espace. Finalement, nous avons comparé la représentation vectorielle des échantillons apprise par le modèle FE à d'autres modèles similaires sur 49 tâches de régression. Nous rapportons que les représentations apprises par FE se classent premiers ou deuxièmes en performance pour toutes les tâches, surpassant des fois par une marge considérable, d'autres représentations.

Disponibilité: Un exemple jouet sous la forme d'un Jupyter Notebook ainsi que le code et les encodages pré-entraînés sont disponibles au: `https://github.com/TrofimovAssya/FactorizedEmbeddings`.

Information supplémentaire: Disponible à l'Annexe A .

3.4. Abstract

Motivation: The recent development of sequencing technologies revolutionised our understanding of the inner workings of the cell as well as the way disease is treated. A single RNA sequencing (RNA-Seq) experiment, however, measures tens of thousands of parameters simultaneously. While the results are information rich, data analysis provides a challenge. Dimensionality reduction methods help with this task by extracting patterns from the data by compressing it into compact vector representations.

Results: We present the factorized embeddings (FE) model, a self-supervised deep learning

algorithm that learns simultaneously, by tensor factorization, gene and sample representation spaces. We ran the model on RNA-Seq data from two large-scale cohorts and observed that the sample representation captures information on single-gene and global gene expression patterns. Moreover, we found that the gene representation space was organized such that tissue-specific genes, highly correlated genes as well as genes participating in the same GO terms were grouped. Finally, we compared the vector representation of samples learned by the FE model to other similar models on 49 regression tasks. We report that the FE-trained representations rank first or second in all of the tasks, surpassing, sometimes by a considerable margin, other representations.

Availability: A toy example in the form of a Jupyter Notebook as well as the code and trained embeddings for this project can be found at: `https://github.com/TrofimovAssya/FactorizedEmbeddings`.

Supplementary information: Supplementary data are available at Appendix A.

3.5. Introduction

RNA sequencing data offers a snapshot into all the cellular processes at a specific time. Since the development of high-throughput sequencing, a multitude of other types of -omics experiments have appeared. RNA-Seq remains nonetheless the most accessible functional characterization of a biological sample. It is sufficiently mature to be applied in a clinical context and large-scale datasets of several thousands samples are readily available. In practice, once aligned and quantified, each RNA-Seq experiment yields (for a human sample) a vector of 20K to 60K gene expression values, depending on the gene annotation selected. Most analyses involving transcriptomic data, however, apply some kind of filtering on the genes by either selecting some of them, grouping them by function, or most of the time applying some type of dimensionnality reduction (Gibbons et Roth, 2002; Kim et Kim, 2018; Gönen, 2009).

Dimensionality reduction algorithms popular in bioinformatics analyses are principal components analysis (PCA), t-Stochastic Neighborhood Embeddings (t-SNE) (Van Der Maaten et Hinton, 2008) and Uniform Manifold Approximation and Projection (UMAP) (McInnes, 2008). They all compress and encode (embed) the data into a new vector representation. While this is often done on samples, it is seldom done on genes. We found that gene representations are mainly computed in the context of clustering and factorization of genes into meta-genes (Lemieux et al., 2017; Brunet et al., 2004), with the ultimate goal of looking for molecular patterns in the gene expression data.

Some teams argue for a similarity between gene expression and bag-of-word representations of text corpora (Ng, 2017; Asgari et Mofrad, 2015), linking their work to that of

(Mikolov et al., 2013; Pennington et al., 2014), who introduced distributed vector representations of words to the field of natural language processing. Upon training their models on word co-occurences in context, they found that their representation of words captured some semantic relationships. This result is consistent with the *distributional hypothesis* in linguistics, where words with similar meaning will be found in similar contexts (Harris, 1954).

Inspired by their seminal work, Du and colleagues have used gene co-expression data to train gene embeddings they called *gene2vec*, and reported that their embeddings extract information of both gene type (protein coding, lncRNA etc.) as well as tissue specificity (Du et al., 2019). Recently, Schreiber and colleagues have published Avocado, a deep neural network tensor factorization tool specialized in epigenomics data, that learns a representation of the human genome, allowing for imputation of epigenomics data and other related tasks (Schreiber et al., 2019). Lastly, similar work by Choy and colleagues showcased a shallow artificial neural network (ANN) to represent genes and samples in high-dimensional embedding spaces, while simultaneously extracting information about genes and samples (Choy et al., 2019). Choy and colleagues reported that they were able to cluster cancers according to gene expression into meta-groups that may then be used for predicting immune checkpoint therapy responders (Choy et al., 2019).

In principle, the model proposed by (Choy et al., 2019) yields an attractive for data-mining double representation of genes and samples. In this paper, we extend the idea behind simultaneously training a distributed representation for genes and samples, presented in (Choy et al., 2019) to the notion of factorized embeddings, an artificial neural network that learns independent embedding spaces to represent factors of RNA-Seq data. We present the general framework of the factorized embeddings model and compare it, when possible, to the models from (Choy et al., 2019) and (Du et al., 2019), as well as other standard dimensionality reduction algorithms such as t-SNE, UMAP and PCA. We show that factorized embeddings capture biologically meaningful information and are reusable in auxiliary tasks that involve predicting some biological features of the dataset.

3.6. Approach

We term factorized embeddings (FE) the idea behind training a distributed representation for genes or samples by factorized tensor decomposition. In transcriptomics, to describe a single gene expression value, we refer to "gene Y in sample X"; we propose to treat both samples and genes as factors that contribute to characterizing the data. Learning a factorized embeddings of the data would be learning an embedding space for each of the factors that contribute to the gene expression value variation.

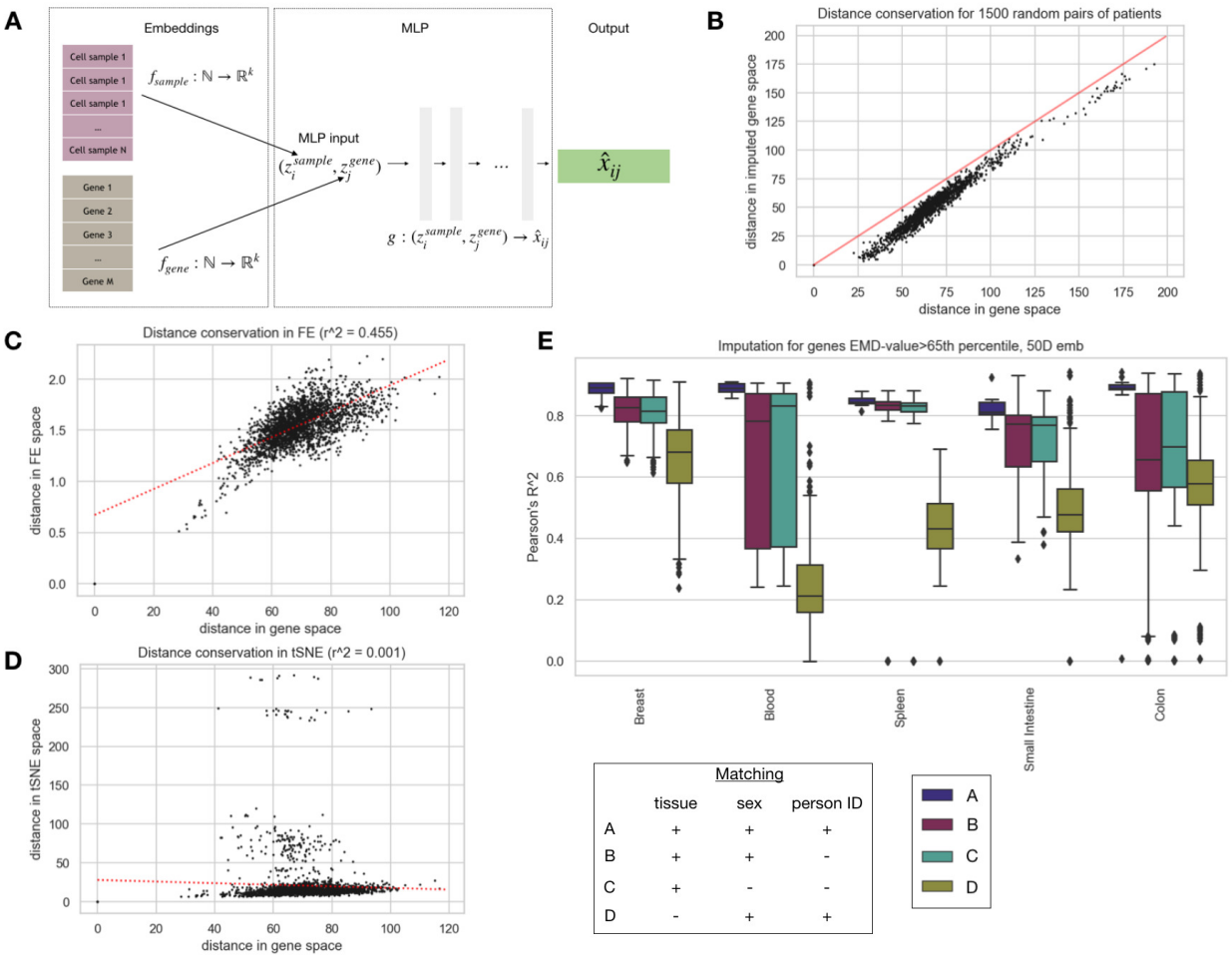


Fig. 3.1. The factorized embeddings model reconstructs data with high accuracy and preserves sample pair-wise distances

A) Schema of the factorized embeddings model. B) Pairwise euclidean distance preservation between 1 500 random pairs of samples in original feature space (gene expression) and reconstructed space C and D) The FE-trained representation preserves more accurately than tSNE pair-wise distances between samples in the embedding space. E) The FE model allows for precise imputation of transcriptomes on a patient-level.

3.6.1. The factorized embeddings model

Each gene expression measurement is a single positive real number. Each single measurement x_{ij} is minimally characterized by a sample i it came from and a gene j it is measuring. Both the sample and the gene are the minimal descriptors for this single gene expression measurement. There may also be other additional descriptors that may influence the gene expression measurement, such as patient, batch, sex, race, etc. The data are cast into a narrow format, where each single measurement x_{ij} for sample i and gene j is described by a series of descriptors.

For clarity, the dataset was built using the two minimal descriptors: The data X is an $N \times M$ array, where there are N samples each of M gene expression measurements.

$$X = [x_1, x_2, ..., x_N]$$

So, an RNA-Seq sample i is represented by a vector of M real values.

$$x_i = [x_{i,1}, x_{i,2}, ..., x_{i,M}]$$

For each sample i and for each gene j in X, an entry in the dataset is created. X is transformed into the following two vectors, where the first contains indices tuples and the second the measured gene expressions $x_{i,j}$:

$$\begin{bmatrix} (1,1) \\ (1,2) \\ \vdots \\ (i,j) \\ \vdots \\ (N,M) \end{bmatrix} \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{i,j} \\ \vdots \\ x_{N,M} \end{bmatrix}$$

Here, each entry in the table on the right is a gene expression measurement for sample i for all genes 1 to M. The table on the left is the descriptor table, where the identity of the sample i as well as the corresponding gene j are recorded. The descriptor table is the input for the neural network, where each doublet of values is an example, while the table on the right contains the targets.

We built a neural network where each of the inputs is embedded in a low-dimensional space of size k. For each field in the inputs, a function f maps a descriptor (for example a sample index) into a k-dimensional space $f : \mathbb{N} \to \mathbb{R}^k$. These spaces are referred to in the text as embedding or representation spaces. All k-dimensional coordinates in embeddings space are concatenated and serve as input for a multi-layer perceptron (MLP). The embedding of an input pair of descriptors (i,j) (e.g. sample i and gene j) is done with two functions $f_{sample} : i \to z_{sample}$ and $f_{genes} : j \to z_{gene}$ resulting in the following concatenated input for the MLP: $[z_{sample}, z_{gene}]$, where Z_{sample} and Z_{gene} are the k-dimensional embedding spaces in which are represented, respectively, the samples 1 through N and the genes 1 through M.

The concatenated embedding coordinates are then fed through a series of fully-connected layers (collectively referred to as g) and the final output layer is a single linear neuron, predicting the gene expression $\hat{x}_{ij}$ for the corresponding descriptors sample i and gene j.

$$g(z_{sample}, z_{gene}) = \hat{x}_{ij}$$

All parameters ($\{\theta, \theta_s, \theta_g\}$) for each layer as well as the embedding functions ($\{f_{\theta_s}, f_{\theta_g}\}$) are optimized together by gradient descent with a mean squared error loss (MSE):

$$L = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} (g_\theta(f_{\theta_s}(i), f_{\theta_g}(j)) - x_{ij})^2$$

This model functions with the assumptions that both the samples and features are independent and identically distributed (IID) amongst themselves (Murphy, 2012). While this may technically not be the case, we found most gene expression analyses work under the assumption of IID genes and IID samples (Maciejewski, 2014). Moreover, the FE model, by design, attempts to preserve all associations in the data between samples and features. We found that this may be in some cases a limitation, especially when the underlying data structure is not linear in the feature space (Figure A.3). We tested the limits of the FE model on a series of toy datasets (Figures A.3,A.4,A.5,A.6) and have made these experiments available in the supplementary section (See Annexe A).

3.7. Methods

3.7.1. RNA-Seq Data

RNA sequencing (RNA-Seq) data for Genotype-Tissue Expression (*GTEx*) and The Cancer Genome Atlas (*TCGA*) cohorts was downloaded from the Xena Browser (Goldman et al., 2019). Xena browser offers a platform of re-aligned and quantified data using the same pipeline to allow for comparison across datasets and across experiments. For each dataset, the RNA-Seq reads were pseudo-aligned and quantified with Kallisto (Bray et al., 2016). Gene expression is represented by transcript per million (TPM) counts. Values were log-transformed with $\log_{10}(\text{TPM} + 1)$.

3.7.2. Tissue-specificity measures

To group genes by tissue-specificity, two methods were used. The first, the *Tau* index τ as described by Yanai and colleagues (Yanai et al., 2005), is a measurement of how tissue-specific a gene expression is, comparing to the other samples. Briefly, for each tissue type c in C, the average expression is calculated for all genes and divided by the maximal value. Then, for each gene j, the τ_j is calculated with:

$$\tau_j = \frac{\sum_{c=1}^{C} (1 - x_{c,j})}{C - 1}$$

The *Tau* index yields a single value between 0 and 1 for each gene. Yanai and colleagues categorize *Tau* index values between 0 and 0.3 as housekeeping genes and values above 0.8 as tissue-specific genes (Yanai et al., 2005).

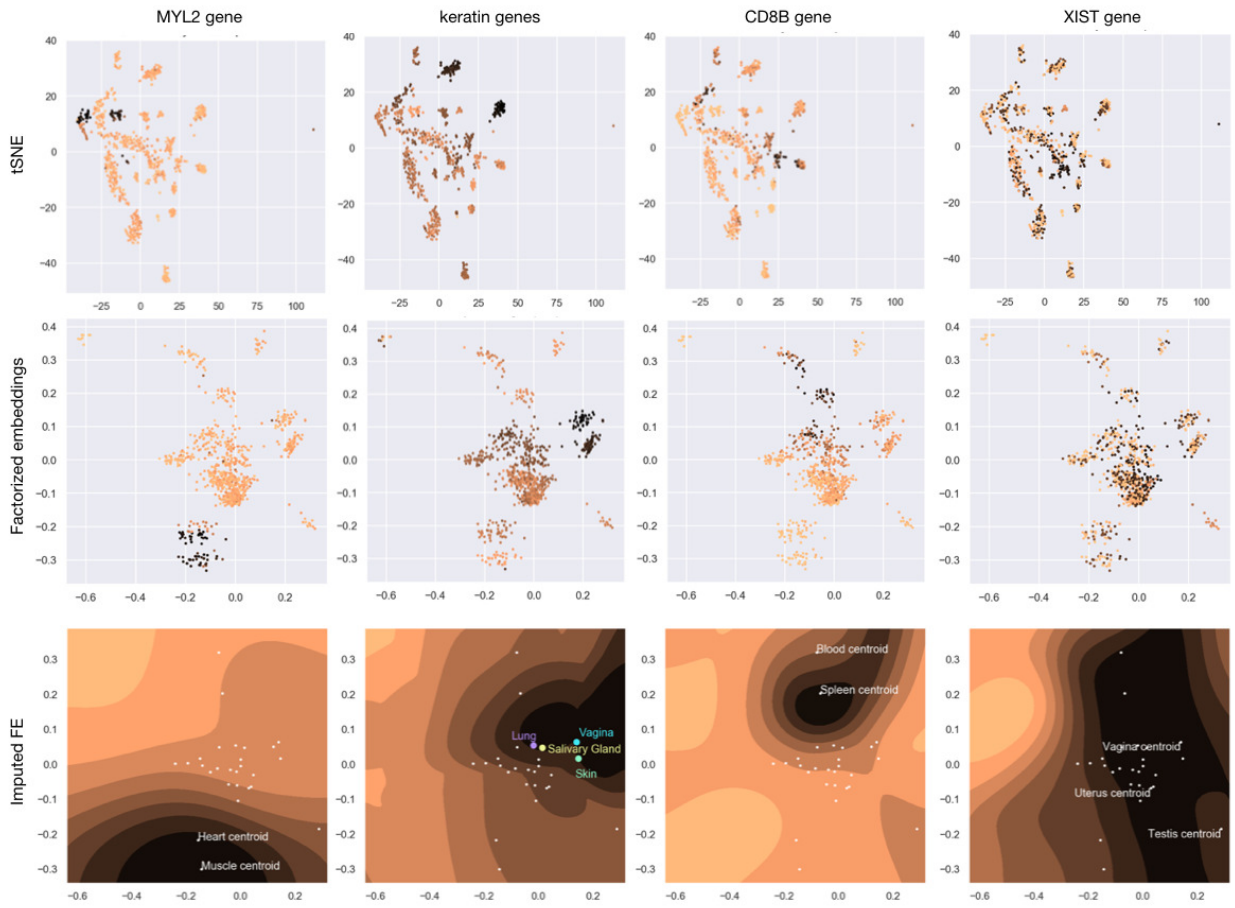


Fig. 3.2. The FE-trained sample embeddings are consistent with individual gene expression levels.

Top and middle) 2-dimensional tSNE and FE of the *GTEx* cohort coloured by the expression level of the 4 chosen reporter genes. **Bottom)** We generated a 2D grid over the embedding space and for every points on that grid, we generated using the trained FE model a new prototype sample. We coloured the space by the predicted expression of each reporter gene for those prototypes

The second one, the tissue-type Earth Mover's distance (EMD) is the Wasserstein distance for each tissue type done as following. Given a gene expression matrix of N samples by M genes, and for a vector tissues C, where each sample i has a tissue $c \in C$. For each gene $j \in M$ and each tissue $c \in C$, we calculate the tissue-specific EMD as the Wasserstein distance between the gene expression of gene j for all samples that belong to class c and the ones of other classes:

Thus, for every tissue type, we obtain a measure of information content for each gene, to distinguish between this tissue type and the others.

3.7.3. Replicating the results of similar models

We attempted to replicate the results published by (Choy et al., 2019) in order to use their model for comparison to the FE model. However, despite significant efforts, we were unable to reproduce embeddings of the quality they presented using the approach as it is described in their publication. We therefore could only compare our FE model on the basis of the published embedding weights available on the author's Github. Consequently, for these comparisons, we trained our FE model on the data processed the same way as described in their publication. Through this work, we maintain comparisons to the Choy model when possible.

For the *gene2vec* model published by Du and colleagues, since it only creates representations of genes, we compared in (section 3.7.8.1) their reported gene representation to the one learned by the FE model.

3.7.4. Benchmarks

For each task, we trained a k-Nearest Neighbors (kNN) regressor model. Similar to the k-Nearest Neighbors (kNN) classifier, the regressor model finds k neighbours for a new point and outputs the average value for those points (instead of the majority class). The kNN was trained on 80% of the data and tested on 20%; this was repeated 100 times for each task. The performance was measured as the Pearson correlation coefficient between the predicted and the actual value. The choice for the correlation was to promote proportional values instead of exact values.

3.7.5. Statistical tests

To compare performances for classifiers trained on various embeddings, we performed an ANOVA test followed by Tukey post-hoc testing. Differences in performance were considered statistically significant when the corrected p-value was lower than 0.05.

3.7.6. Model training

The entire factorized embeddings (FE) model, which includes both embedding spaces and subsequent fully-connected layers (see section 3.6.1), is trained with a mean squared error (MSE) loss function and tanh activation function, Adam optimizer, a $L2$ regularisation of 10^{-5} and a learning-rate of 10^{-3}. The model is implemented in PyTorch (Paszke et al., 2019) and we release the code online (https://github.com/TrofimovAssya/FactorizedEmbeddings). We chose an embedding size of 50D for both genes and samples and a multi-layer perceptron (MLP) of 5 layers, respectively of sizes 250, 150, 100, 50 and 10 (see Figure A.1,A.2). As

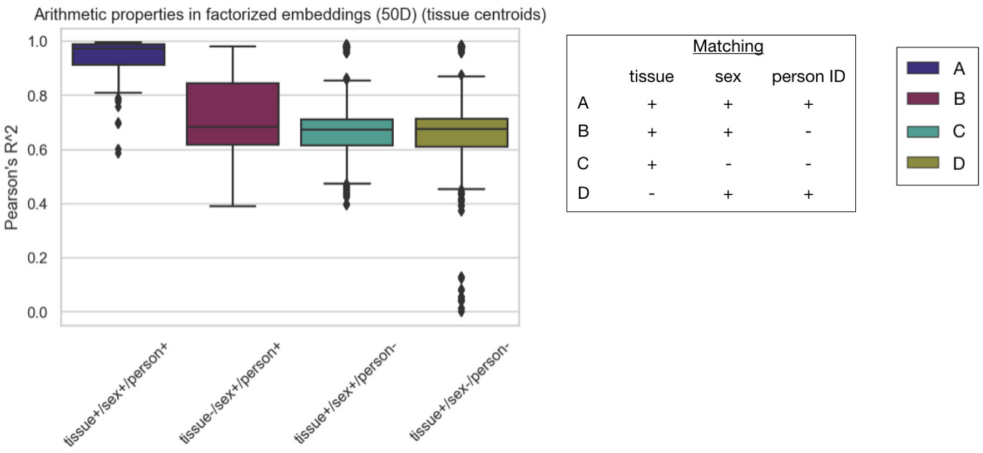


Fig. 3.3. Vector arithmetic properties are conserved in the patient space. For each patient, we compare the prototype transcriptome obtained by vector arithmetics (see text) to the ground truth (A), other tissues for the same person (B) as well as other samples of the same tissue (C,D)

a reference, for a dataset of 1 000 samples each of 60k transcript expression measurements, the model trains 500 epochs in 72 hours on an NVIDIA GTX 1080 Ti.

3.7.7. Reconstruction accuracy of the model

As a sanity check, we first compared pairwise distances between 1500 random sample pairs in original and reconstructed space (Figure 3.1B). This was done to probe if once the data passes through the model, it preserves the proportions between samples. We found that the FE model reconstructs with high accuracy the data (Figure 3.1B).

Moreover, unlike the Locally Linear Embeddings (LLE) (Roweis et Saul, 2000) algorithm or to some extent t-Stochastic Neighborhood embeddings (t-SNE), the factorized embeddings (FE) model is not guaranteed to preserve sample-sample distances in the embedding space. We however report that it is the case for FE, where pairwise distances in original feature space for a random subset of pairs of samples are preserved in embedding space (Figure 3.1C), incidentally better than tSNE (Figure 3.1D).

We finally probed whether the gene expression reconstruction was performing well on hard to predict genes. We randomly selected 20 individuals for every tissue type in the GTEx cohort and compared the reconstructed by the FE model transcriptome as follows (Figure 3.1E):

- (A) itself (*tissue+/sex+/person+*)
- (B) other samples of that tissue type, with matching sex (*tissue+/sex+/person-*)
- (C) other samples of that tissue type, with opposite sex (*tissue+/sex+/person-*)
- (D) other tissues for that individual (*tissue-/sex+/person+*)

We found that the reconstruction of the transcriptomes was always closer to the individual than to other matching tissue samples (Figure 3.1E), with a Pearson's correlation R^2 close to 1. As expected, other tissues for that individual were always less similar to the patient than other examples of the same tissue. We also found that depending on the tissue type, the other samples of the same tissue were at varying degrees of similarity, some closer as seen for *Spleen* and *Breast* and some quite far, such as *Blood* and *Small Intestine* (Figure 3.1E). We hypothesize that since each sample represents a bulk tissue, some tissues might have a higher heterogeneity, with multiple cell lineages represented in each cell subset (Wagner et al., 2016; Regev et al., 2017). We limit the comparison in reconstruction to genes with a tissue-specific Earth-Mover's distance over the 65^{th} percentile (see Methods), in other words, genes with high tissue-specificity. While the 65th percentile was selected arbitrarily, the importance to use percentiles is caused by the fact that the EMD distribution for genes varies between tissue types and so the cutoff value might vary from tissue to tissue. We do this because we expect the large proportion of non-tissue specific genes (housekeeping genes for example) to be well reconstructed and we want to probe the reconstruction of challenging genes.

Taken together, these results suggest that the FE model reconstructs with good accuracy each individual sample and preserves the sample pairwise distances in the embedding space.

3.7.8. Factorized embeddings captures biologically meaningful information on both samples and genes

We found that similarly to t-SNE, the FE model trained on the GTEx dataset groups samples by tissue types and the FE model trained on the TCGA dataset groups samples by cancer types. For t-SNE, grouping of samples in embedding space depends on sample pair-wise distances, since this is what t-SNE optimizes locally in the representation (Van Der Maaten et Hinton, 2008). This characteristic is however not guaranteed in FE, although we found that it is conserved (Figure 3.1B).

For the results presented in figure 3.2, we chose four "reporter" genes or groups of genes (displayed for each column) for the following characteristics: i) *MYL2*, coding for myosin 2 was chosen since it is expressed only by a small number of tissues (heart and muscle), ii) *CD8B*, a marker of T cells was chosen because it is expressed in small quantities by many tissues and in high proportion in blood and spleen, iii) *XIST*, a sex-specific gene, expressed only in tissues belonging to female individuals and iv) *keratin*, a group of proteins expressed widely in epithelial tissues. For each of these reporter genes, we wanted to observe how the level of their expression across samples was represented in the embedding space. We found that the FE model learns a space where the predicted gene expression for genes that are common (keratin) or rare (*MYL2*) across tissues, regardless of their level of expression

(*CD8B*) is smooth (Figure 3.2, lower row). In contrast, we do not find the same smoothness with the t-SNE representation of the sample space.

Finally, we observed that the preservation of the smoothness of the space over the gene expression of *XIST* did not seem to be important nor for t-SNE, nor for the FE model, which suggests some sort of selection of genes by importance for the reconstruction. We found that this is due at least in part to the fact that the *XIST* gene and other sex-specific genes have a lower gene expression profile and constitute a smaller gene group than tissue-specific genes. Further details that lead to this conclusion can be found in the supplementary material (Figure A.8). We conclude here that no matter the gene expression pattern, be it restricted to only some tissues, FE orders samples in embedding space according to individual gene expression.

Moreover, to verify if the embeddings space is dense, we created a 2500 point grid over the sample embedding space and for each coordinate we generated a new transcriptome, by running it through the FE model. The sample embedding space was then coloured by the imputed gene expression and we report that the FE sample embedding space is dense and allows for interpolation between samples (Figure 3.2 bottom).

While this visual comparison is possible when the embedding space is 2-dimensional, we wanted to evaluate this property of interpolation with the 50D embeddings. Inspired by the vector arithmetics in embedding space described in (Mikolov et al., 2013), we performed a similar experiment. We trained an FE model on the GTEx cohort, where multiple tissues are available for the same individual donor. Taking one specific donor-tissue combination, we subtracted the centroid coordinates for that tissue and added the centroid coordinates for a new tissue. The obtained embedding coordinate was then run through the FE model, to generate a new prototype transcriptome. This prototype transcriptome was compared to the actual transcriptome for that donor-tissue combination (ground truth), as well as other donors with either matching tissue type or sex, similar to (see legend in Figure 3.3). We report that the prototype transcriptome imputed by such vector arithmetics is closest to the ground truth than any other transcriptome (Figure 3.3, verified by an ANOVA test followed by a post-hoc Tukey test, corrected p-value <0.01). Moreover, we observe that the reconstructed transcriptome seems to be somewhat closer to other transcriptomes for that donor, leading us to conclude that the FE model encodes donor-specific information.

Yanai and colleagues have suggested that it would be possible to infer ancestral tissue profiles using comparisons between gene expression profiles (Yanai et al., 2005). At that time they had little tissues available and found three major groups among tissues: i) *Bone Marrow*, *Spleen*, *Thymus* and *Lung* were grouped together and shared a common ancestor on a higher level with ii) the *Pancreas*, *Prostate*, *Kidney* and *Liver* group. This meta-group in turn shared a common ancestral gene expression with the third group, consisting of *Heart* and *Muscle* (Yanai et al., 2005). We isolated from both t-SNE, PCA and FE representations

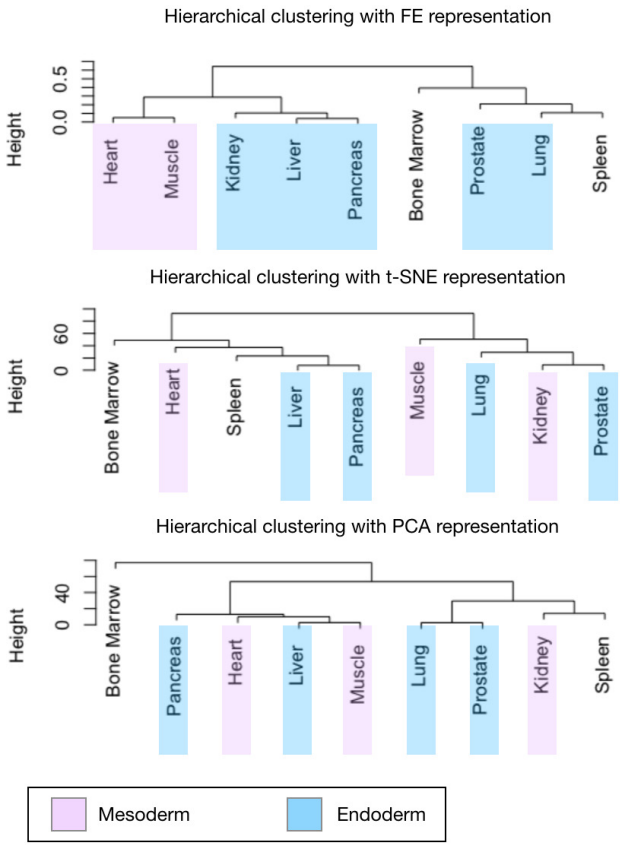


Fig. 3.4. Factorized embeddings learns general gene expression patterns. For FE as well as t-SNE and PCA embeddings of the *GTEx* cohort we performed hierarchical clustering over various tissues originating of either the mesoderm or the endoderm. We coloured each tissue as the germ layer of origin.

the tissues in question and performed a hierarchical clustering with complete-linkage on the embedding coordinates of the centroids for each tissue group. The FE model was found to retain roughly the same hierarchy as described in (Yanai et al., 2005), while t-SNE and PCA did not (Figure 3.4). From here we conclude that the FE model retains global gene expression patterns. It is however possible that this hierarchy is not retained in t-SNE because of the nature of the t-SNE algorithm, since it only optimizes for conservation of local dependencies between points (Moon et al., 2019).

Finally, to compare the learned sample representations to those reported by Choy and colleagues, we trained two versions of the FE model: i)one only on protein-coding genes, to fit what was described by Choy and colleagues and ii) one using all the dataset. We then evaluated these embeddings and compared them to others obtained by t-SNE, PCA and UMAP as well as two instances of the FE model, trained on the two datasets.

For the easiest task - tumor type classification - we found that the weights for embeddings of TCGA cohort published by Choy and colleagues did not perform as well as all the other representations (Figure 3.5).

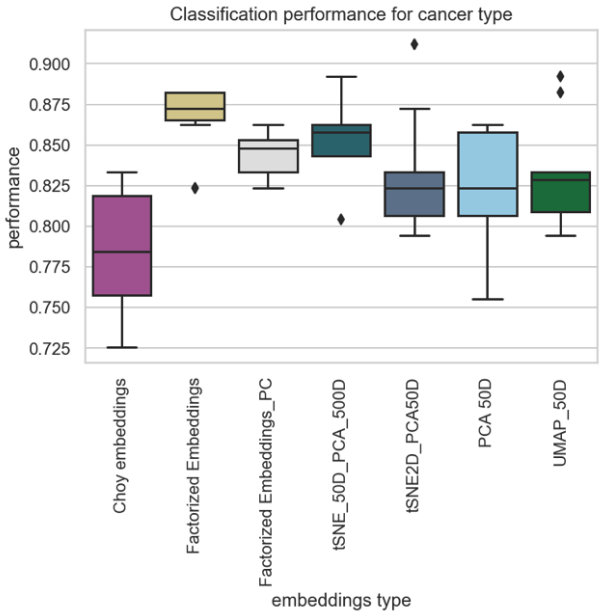


Fig. 3.5. The FE-trained embeddings outperform all other representation in the prediction of cancer type task.
We compared the representations of the FE model trained on all the data (FE) as well as an FE model trained on protein-coding genes only (FE_PC), to the downloaded weights for the Choy et al. model, a 50D tSNE, a 2D tSNE, a 50D UMAP and a 50D PCA. Each box and whiskers plot represents the performance of a 5-nearest neighbor classifier model tested on 20% of withheld data, reshuffled 100 times.

Taken together, these results suggest that the FE model captures biologically relevant information in the sample representation.

3.7.8.1. The nature of gene embeddings. We then concentrated on the gene embeddings, to probe what kind of information might be captured by the FE model for individual genes. While sample embeddings offers a multitude of labels and categories, gene embeddings do not have this type of extensive categorisation.

We compared side-by-side the gene embeddings obtained for the FE model to the weights provided by (Choy et al., 2019) as well as (Du et al., 2019). Both Du and colleagues and Choy and colleagues chose to train their models on a limited set of genes, mainly protein coding, and some microRNAs (respectively 24447 and 20531 genes). For the side-by-side comparison, we focused on protein-coding genes only, to mimic what was done by the other teams (Choy et al., 2019; Du et al., 2019).

We found that both the Choy and the FE models and to some extent the *gene2vec* model grouped genes by overall tissue specificity (Tau index) (Figure 3.6 top row). However, we report that the FE model seems to aggregate genes by the maximal EMD for each tumor type (Figure 3.6 middle row). Finally, for each gene, we selected the tumor type for which

it is the most specific (see Methods). We grouped tumors according to their tissue of origin and found that FE organises genes by tissue-specificity (Figure 3.6 bottom row).

Besides tissue-specificity, gene-gene co-expression (measured by correlation) drives differential gene expression analyses. We found that correlated genes were located close-by in gene embedding space (Figure 3.7). However, we found that proximity in location in the gene embedding space does not necessarily mean correlation in gene expression (Figure 3.7).

Our final way of characterizing genes is their involvement in common cell processes. To probe the recovery of this gene feature, we selected a range of Gene Ontology terms (GO

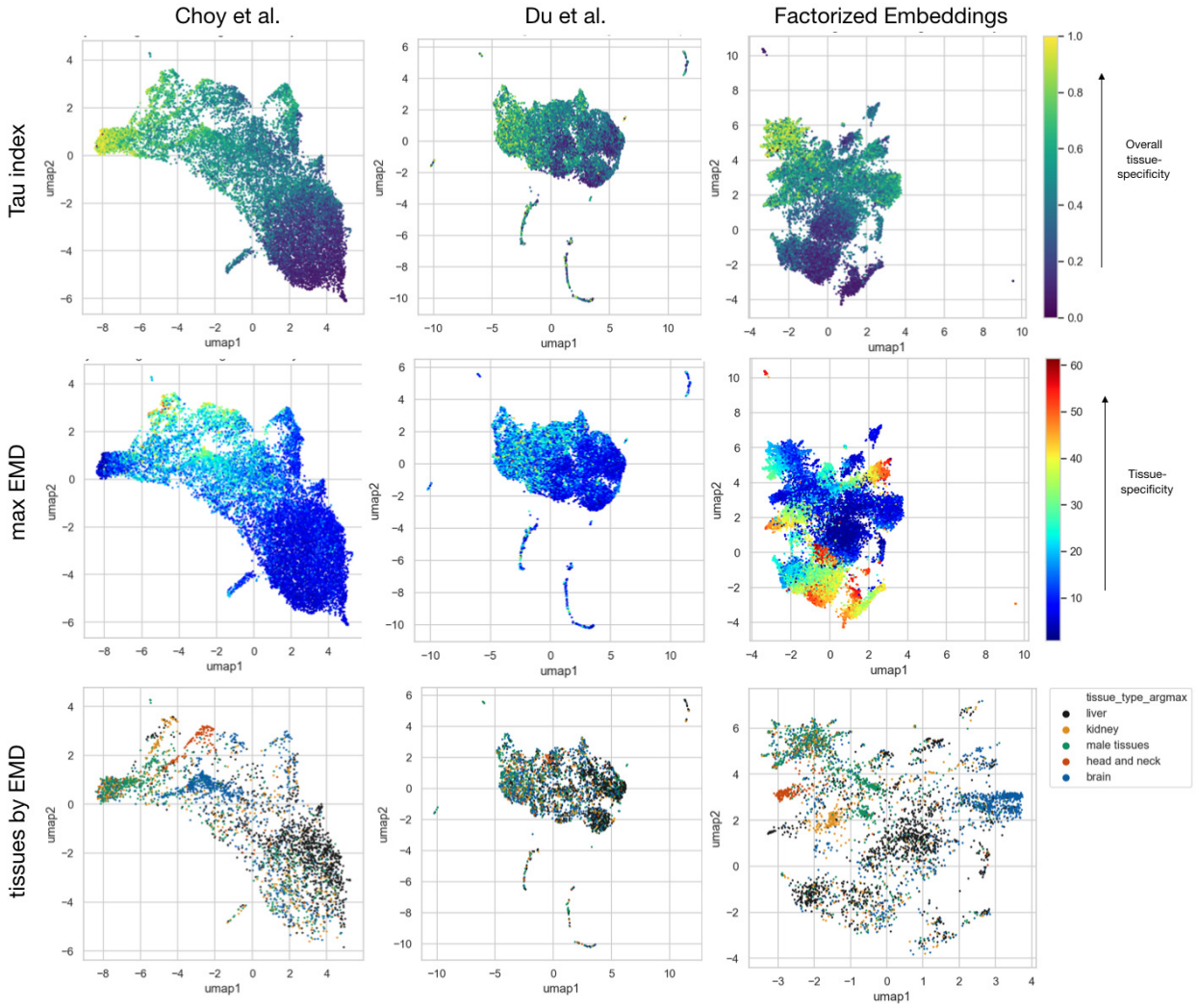


Fig. 3.6. Gene embeddings trained with FE group tissue-specific genes. For the gene representations obtained for the Choy, Du and FE models, we calculated for each gene the Tau index of tissue specificity and a tissue-specific EMD (Methods). Each point represents a gene and they are coloured by either Tau index or maximal EMD over all tissues. The bottom row shows genes coloured by the tissue, to which they are specific, obtained by taking the maximal argument over tissue specificities.

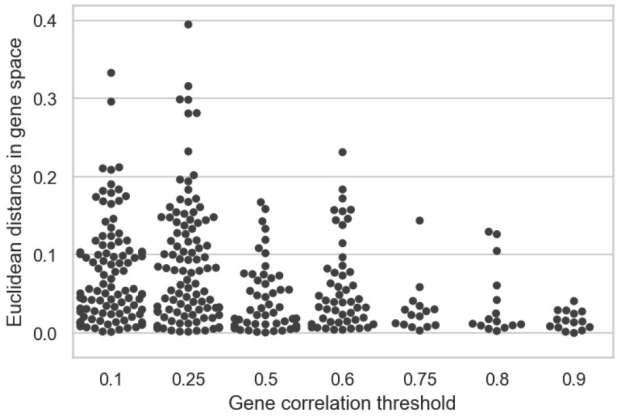


Fig. 3.7. Factorized embeddings groups correlated genes together in embedding space

. For randomly selected pairs of genes at various correlation intervals, we measure the pairwise euclidean distance.

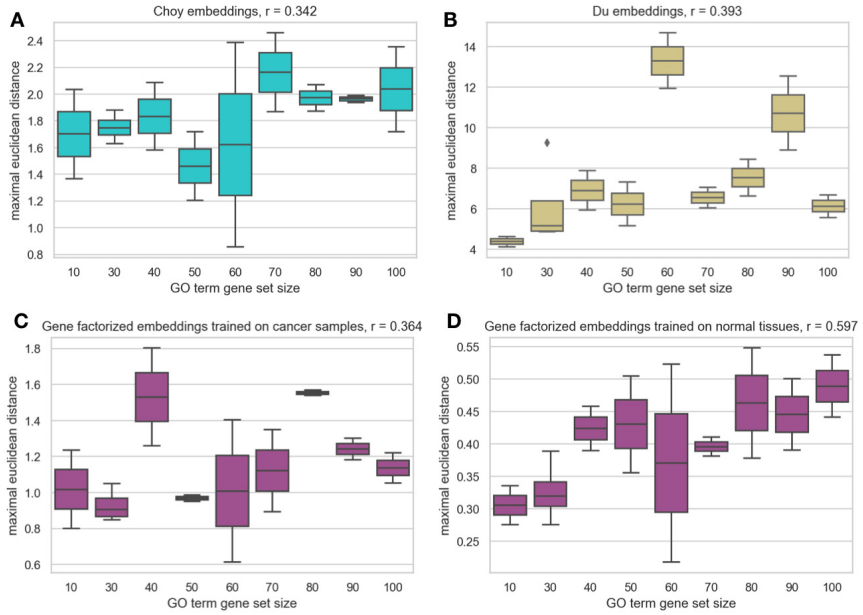


Fig. 3.8. The FE-trained on *GTEx* gene representations capture GO term participation.

We selected GO terms with increasing gene set size (and decreasing specificity) and calculated for each gene set the maximal intra-set distance in the various embedding spaces. We calculate a Pearson correlation coefficient for A) Choy, B) *gene2vec* (Du), C) Factorized Embeddings trained on cancer samples and D) Factorized embeddings trained on healthy tissues.

terms), by gene set size (Ashburner et al., 2000). The rationale is that the smaller the GO term gene set size, the more precise the GO term is. We hypothesize that if the gene embeddings capture gene participation in similar processes, it should group closer together

genes participating in narrow GO terms and vice versa. We found that FE trained on GTEx but not the Choy nor the *gene2vec* embeddings preserve the relationship between GO term size and max euclidean distance (Figure 3.8). This may be at least in part attributed to the fact that the Choy model was trained on the *TCGA* cohort and Du and colleagues trained their model on an amalgam of GEO datasets. Indeed, the change in dataset drastically changes the recovery of the relationship (Figure 3.8C and D). Taken together, our results show that the gene representations learned by the FE model capture both gene tissue-specificity, co-expression patterns as well as cellular processes and gene type (Figure A.9).

3.7.9. Validation of the embeddings on auxiliary task

We believe that the nature of the RNA-Seq data offers a rich glimpse into cell processes that goes beyond just gene expression profiles. For example, mutations in some regions are reported to alter gene expression and therefore will leave an imprint of the transcriptome (Audemard et al., 2018). Moreover, Rappoport and colleagues report that their computational method performs in some cases best when trained on RNA-Seq data, compared to training on various multi-omics datasets (Rappoport et Shamir, 2018). In our work we have shown that the embedding spaces learned by the FE model capture biologically meaningful information - on gene function and gene-gene co-expression, patient specific gene expression as well as tissue-specific gene expression patterns. We wanted to validate the usage of the FE model as a dimensionality reduction method in a series of tasks that involve other types of assays. Conveniently, Thorsson and colleagues used the *TCGA* RNA-Seq data combined with microscopy, copy-number variant, whole genome sequencing and additional RNA-Seq data processing pipelines to characterize the tumors from an immunological point of view (Thorsson et al., 2018). We obtained from their work a total of 49 additional labels for the samples of the *TCGA* cohort. We grouped the labels by the nature of the additional data or algorithm that was required to obtain these labels. These groups are as follows:

- *Cibersort* refers to the prediction of infiltration of various immune cells obtained by the Cibersort algorithm (Newman et al., 2015).
- *Thorsson immune profiles* are the final immune profile categories specified by (Thorsson et al., 2018).
- *Immune repertoire* is a measure of B cell and T cell receptor diversity, requiring a special type of quantification done on bulk RNA-Seq (Bolotin et al., 2015a).
- *Genomic instability* is a group of tasks that includes measures such as incidence of synonymous and non-synonymous mutations, aneuploidy score, homologous recombination defects.

- *Microscopy* refers to tasks that require interpretation of microscopy images of tumor biopsies. Examples of such tasks are quantification of leukocyte fraction and stromal fractions and estimation of intra-tumoral heterogeneity.

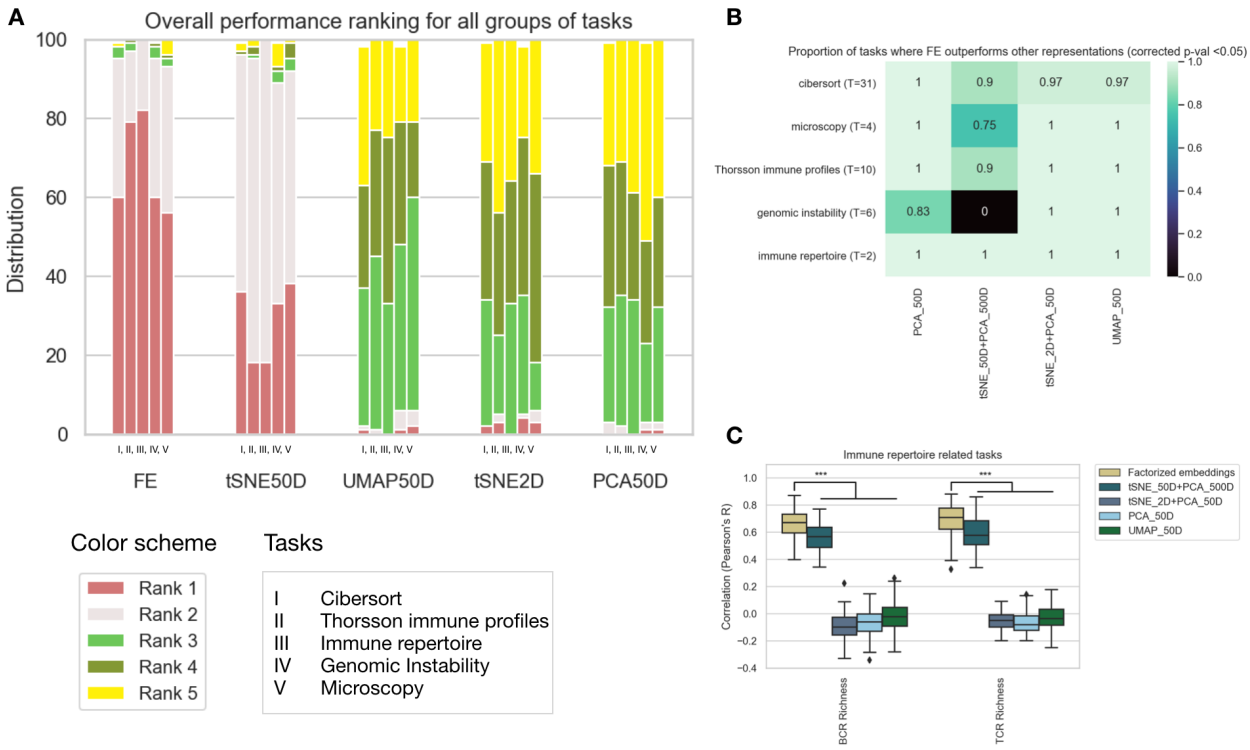


Fig. 3.9. The FE representation of samples of the *TCGA* cohort outperforms in a series of 49 tasks all the other representations.
A) For every task group we ranked the algorithms by performance and separated the tasks into task groups (Methods) B) For every task group, we showed the proportion of tasks where the FE embeddings outperformed each other representation (corrected p-val<0.05, ANOVA followed by post-hoc Tukey test). C) An example of task group where the difference in performance between the FE representation and the others is statistically significant (p-val<0.001)

We hypothesize that the sample embeddings trained by the FE model contain enough signal to perform on each of these 49 regression tasks. We compare the FE sample embeddings to the embeddings obtained by a 50-dimensional t-SNE, UMAP, PCA as well as a 2-dimensional t-SNE. For each task, we trained a kNN regressor model on 80% of the dataset and test on a held out 20%. This process was repeated a total of 100 times (Annexe A.3). Then, we grouped the tasks by the type of data and ranked performance-wise the various embeddings we compared (Figure 3.9 A). We found that the FE model consistently ranked first performance-wise in 60% of the tasks overall and second in almost 40% (Figure 3.9 A). The second-best embedding was the 50-dimensional t-SNE, that ranked first about 30% of

the time and second almost 70% of the time (Figure 3.9 A). There is no clear distinction in the performance of the other three embeddings. We verified if the difference in performance is statistically significant when comparing FE to other embeddings using an ANOVA test, followed by a post-hoc Tukey test. We report that the FE-trained sample embeddings outperformed all three UMAP50D, PCA50D and t-SNE2D almost 100% of the time (Figure 3.9 B). However, while the FE-trained embeddings ranked higher in performance, the difference was not always statistically significant when compared to the 50-dimensional t-SNE-trained embeddings (Figure 3.9 B) - notably, for the genomic instability task, none of the performances was statistically significant (Annexe A.3).

We examined more closely one of these results, the immune repertoire task, where the FE-trained embeddings outperformed all other embeddings (Figure 3.9 B, bottom row and C). We suspect that the high performance of the FE-trained embeddings on this task is due to the link between the tumor type and the immune infiltration. Indeed, it has previously been reported that immune infiltration varies by tumor type, amongst other things (Thorsson et al., 2018; Iglesia et al., 2016). We therefore conclude that at least part of the good performance of the FE-trained embeddings stems from the performance at the easiest task - the tumor-type categorisation (Figure 3.5).

3.8. Discussion and Conclusions

Together with the work of Choy, Shreiber and Du and colleagues, the factorized embeddings model fits into a small family of self-supervised learning algorithms, that perform tensor factorization of the dataset into separate latent spaces. We demonstrated the utility and performance of FE on two large-scale RNA-Seq cohort: *TCGA* and *GTEx*. When compared to the most similar model, published by (Choy et al., 2019), we found that FE captures more biologically meaningful information in the sample and gene embeddings, which is probably due to the fact that the FE model but not the Choy model uses a set of fully connected layers on top of the embeddings layers. Unexpectedly, we found that the FE model preserves gene expression pair-wise distances in embedding space as well as being coherent with both single gene expressions across samples and more broad gene expression patterns. Moreover, we found it possible to perform the same type of vector-space arithmetics in sample embedding space as described in the works of (Mikolov et al., 2013) and (Pennington et al., 2014), transforming one tissue into the other, while preserving the patient specific gene expression profile. This feature is something that non-parametric distance-based approaches, such as t-SNE, do not allow, since there is no way to reconstruct the data from the representation.

Finally, we demonstrated the utility of the encoding samples into a smaller, information-rich representation, by running a total of 49 benchmark tasks, that involve predicting results from other assays on the same samples. We found that FE-trained representations rank

mostly first and sometimes second and outperforms all the other dimensionality reduction algorithms on all 49 benchmarks. We found that in a small amount of cases, it is not statistically different from the performance of a 50-dimensional t-SNE. We would like to point out one of the possible limitations behind the performance of the benchmark experiments is that most of these labels are closely related to the tumor type and therefore some of the outstanding performance of the FE-trained embeddings may be attributable to this. Also, we were unable to identify gene categories standing out within the gene embedding space. This likely reflects the fact that most genes are multi-functional and involved in several processes across different tissue types and cancers. It is also unfortunate that very little categories are available for genes besides GO annotation, which are mostly derived from the study of healthy tissues. One possible extension would be to add some notion of gene function and/or localisation, that is available in the literature.

An appealing feature of the FE model is that it can, with little adjustments, be adapted to train on large-scale multi-omics datasets by introducing supplementary embedding spaces and jointly optimizing as many functions $g()$ (see Approach and Methods) as there are sources of observations, which could very well be additional clinical information on patients. Embeddings representing spaces shared by multiple data sources would be constrained to integrate these sources. This extension would naturally take advantage of the fact that our approach is not affected by missing data. Importantly, it would not require that datasets be complete, where all modalities would be measured for all samples. Furthermore, exploiting this later feature would support the use of the FE model for missing data imputation. More challenging would be the extension of the FE model to less "categorical" spaces, a direction we have previously explored, in a limited context, by adapting the FE model to the work with transcript sequences instead of relying on a predefined transcriptome annotation (Trofimov et al., 2018). Non trivial implementation issues, resulting in poor scalability of the proposed model, have so far limited its development.

We believe that the FE model is a highly customizable architecture that provides a strong foundation to develop omics-based predictors based on integrated data sources, resilient to missing data and provide similar benefits to other dimensionality reduction techniques in extracting patterns from omics data.

Acknowledgements

We would like to thank Theodore J. Perkins and Mathieu Blanchette for their helpful comments and ideas on the experiments and preliminary results. We also thanks Caroline Labelle, Jeremie Zumer, Alex Fedorov, Tristan Sylvain and Philippe Brouillard for their feedback on the manuscript.

Funding

This work was supported by CIFAR, Calcul Quebec, and Oncopole. We acknowledge the support of IVADO and the Canada First Research Excellence Fund (Apogée/CFREF). A.T's work has been supported by the Frederick Banting and Charles Best Canada Graduate Scholarships Doctoral Award of the Canadian Institute for Health Research (CIHR).

Chapitre 4

The Latent Transcriptome

Assya Trofimov[1,2,3], Francis Dutil[2,3], Claude Perreault[1,4,5], Sébastien Lemieux[1,6], Yoshua Bengio[2,3], Joseph Paul Cohen[2,3]

([1]) Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([2]) Department of Computer Science and Research Operations, Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([3]) Montreal Institute for Learning Algorithms (Mila), Montreal, Quebec H2S 3H1, Canada
([4]) Department of Medicine, Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([5]) Maisonneuve-Rosemont Hospital, Montreal, Quebec H1T 2M4, Canada
([6]) Department of Biochemistry at University of Montreal, Université de Montréal, Montreal, Quebec H3C 3J7, Canada

Cet article a été publié sur arxiv en 2018

4.1. Mise en contexte

Le modèle *Factorized Embedding* (FE) a été validé comme une architecture candidate pour la construction d'atlas cellulaires basés sur les données de séquençage ARN (RNA-Seq) (voir Chapitre 3). Nous avons voulu étendre le modèle à un modèle pouvant capturer à la fois des différences d'expression génique et des modifications dans la séquence même du gène. En effet, dans certains cancers, tels la Leucémie Myéloïde Aiguë (AML, de l'anglais *Acute Myeloid Leukemia*), l'encodage des échantillons dans un atlas cellulaire doit inclure également des modifications génomiques, telles les mutations, inversions, répétitions en tandem et translocations chromosomiques, qui sont souvent utilisées à des fins diagnostiques et pour orienter le choix du traitement. Malheureusement, simplement analyser les données quantifiées de RNA-Seq ne capture qu'en partie ces informations (Audemard et al., 2018) et donc nous avons cherché à modifier le modèle FE existant afin de trouver une solution accommodant ce type de jeu de données. Pour ce faire, nous avons remplacé l'espace d'encodage des gènes par un espace d'encodage de k-mers. Nous nous attendions à ce que le modèle encode les k-mers selon leur similarité en abondance ainsi que leur similarité de séquence, cependant, nous ne savions pas dans quelle mesure le modèle allait prioriser l'une ou l'autre des caractéristiques. Cette itération du modèle FE a été baptisée le *Transcriptome Latent* (TLT, de l'anglais *The Latent Transcriptome*). Nous avons observé que l'espace d'encodage de k-mers appris par le modèle TLT était tout d'abord sectionné selon l'abondance des k-mers, puis les k-mers étaient groupés par similarité de séquence. Dans l'espace appris, les k-mers chevauchant, faisant partie de la même séquence, se regroupaient en partitions (*clusters*). Cette particularité du modèle a permis également de capturer une modification génomique: un translocation chromosomique créant un gène de fusion. En effet, les k-mers chevauchant la translocation entre deux chromosomes formaient dans l'espace d'encodage une forme de pont, à mi-chemin entre les clusters correspondant aux exons des deux gènes participant à la fusion. L'ensemble de ces résultats a permis de conclure que le modèle TLT était une itération intéressante du modèle FE pour les atlas cellulaires où l'encodage des échantillons nécessite d'inclure des données de la séquence elle-même.

4.2. Contributions

Assya Trofimov: A mené le projet, conçu et effectué les expériences, analysé les données, préparé toutes les figures et a écrit l'article.
Francis Dutil: A effectué des expériences, a participé aux discussions et à l'analyse de données et a participé à l'écriture de l'article.

Claude Perreault, Sébastien Lemieux, Yoshua Bengio, Joseph Paul Cohen: Ont conçu et dirigé les expériences, ont participé aux discussions et analysé les données, et ont écrit l'article.

4.3. Texte de l'article

4.3.1. Résumé

4.3.2. Introduction

The fundamental issue we would like to address in this paper is the need for a flexible representation of RNA-Seq experimental data. Standard RNA-Seq analysis pipelines discard rich information for a more canonical result (Zielezinski et al., 2017; Valbuena et al., 1978). This information may be crucial, since diseases such as cancer are known for their high mutational burden, multiple rearrangements and unconventional genomes, which do not fit the assumptions of the standard RNA-Seq pipeline, that uses hand-crafted features as a basis for analysis. To address this, we learn a latent space that captures gene-like structures from raw RNA-Seq data. We find that our proposed model successfully recovers information on gene sequence similarity, mutations, and chromosomal rearrangements.

The transcriptome is a subset of all possible sequences of the genome that are used by the cell at any given moment and constitutes less than 2% of all genomic sequences (if we consider only one cell type). Of this transcriptome, only a small amount is captured in standard RNA-Seq analysis pipelines, mainly transcripts that encode proteins (total of 20-60k sequences). The goal of these pipelines is to count the relative abundance of each transcript in the cell.

The raw data actually contains much more information than just gene abundance, namely patient-specific mutations and chromosomal rearrangements. RNA-Seq experiments yield very rich data, that can be informative both in terms of sequence abundance as well as sequence composition. Reducing this rich data to only the detection of annotated genes (which are "hand-crafted" features of the sequence) is not appropriate for analysis. Indeed, although the simple story relating each gene to a protein is correct to first approximation, there are important phenomena such as gene homology, patient-specific mutations, translocations and other genomic alterations that are excluded from the analysis, despite their presence in the data.

In this work, we consider the problem of including the rich patient-specific sequence information from RNA-Seq data via a continuous representation that will account for both gene expression as well as mutations and chromosomal rearrangements. We propose a model

which learns gene-like representations from the raw patient-specific sequence RNA-Seq data. We study how this model handles situations that are standard in cancer genomics but considered edge case in standard pipelines.

4.3.3. Related work

4.3.3.1. The standard RNA-Seq analysis pipeline. Once RNA is extracted from cells in the lab, it is processed by a sequencer. Individual RNA sequences are fragmented into short 100-200 base pair sequences (each of which is called a *read*) before entering the sequencer and then processed in bulk. A sequencing experiment produces a vast number (billions) of short character sequences (reads, R), each character (A,C,T,G) representing each of the four nucleic acids (Figure 4.1A). A good analogy of the way RNA-Seq experiments are done would be to compare the output of a sequencer to the output of a shredder. To deal with a shredder-generated output, a reference text would be helpful. This reference text (approximately like the text of the shredded document) is used to look-up the shredded sequences to determine their regions of origin in the text. This works well so long as the true underlying text and the reference are fairly close and that their differences are local. Unfortunately, this assumption breaks in the case of cancer-induced mutations.

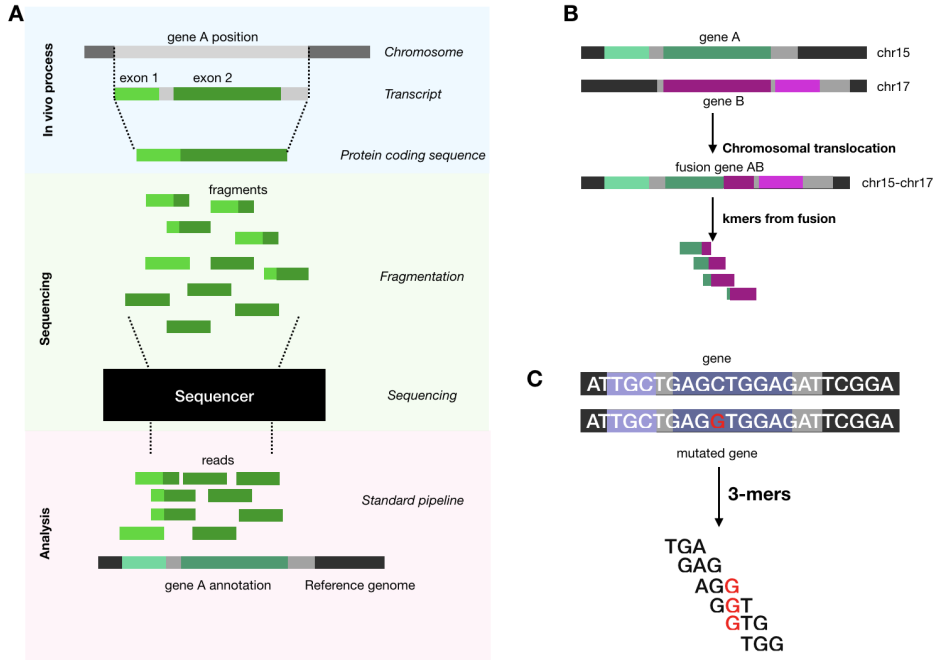


Fig. 4.1. Overview of pipelines and k-mers
A) Standard RNA-Seq analysis pipeline, including alignment of sequencing reads. B)Translocation between two chromosomes and creation of a fusion gene AB. k-mers overlapping the fusion will partially match both original sequences. C) Every patient-specific mutation creates at least k new k-mers. Here new k-mers are shown for k=3

A reference genome exists for most of the widely studied species, see Figure 4.1 for the standard processing pipeline. In this reference genome, gene locations and exact sequence are annotated according to the current revision. The reference genome is renewed every five years or so, to take into account new discoveries in genomics. Zielezinski and colleagues report that sequence alignment methods fail in the edge cases of high sequence similarity (homology) (Zielezinski et al., 2017). Moreover, sequence alignment to the genome is an NP-hard problem (Chatzou et al., 2016), that is sped up using heuristics and yielding the "most probable" alignment, which adds a level of uncertainty to results.

In the event of a chromosomal translocation (Figure 4.1B), a common occurrence in cancer, reads that cover both parts will only be matched to the un-fused sequences based on a sequence similarity score (typically $< 50\%$ match) (Zielezinski et al., 2017). Therefore, chromosomal rearrangements such as translocations do not directly match the reference genome. For this specific case, we deem the reference genome inappropriate. Indeed, reference based methods yield a relative abundance measurement of genes, which are by definition, hand crafted features.

4.3.3.2. Merging RNA-Seq experiments with additional genomic data. Cancer cells have often unconventional genomes, many showing chromosomal rearrangements, mutations, copy-number variations (CNV) and repeated regions that have a clinical interest (Weinstein et al., 2013). The correct identification of these rearrangements is a non trivial challenge for reference genome-based pipelines, since these modifications are not included in the reference genome.

Many recent advances in the field of cancer research have become possible either by combining standard pipeline-derived RNA-Seq data with other sequence-specific data, such as SNP arrays, miRNA-Seq and even whole genome sequencing (Koboldt et al., 2012; Gerstung et al., 2015; Hu et al., 2016; Weinstein et al., 2013; Gao et al., 2013; Liu et al., 2017).

Other teams preferred to develop reference-free methods for variant calling. One such reference-free method, *km*, stores n-grams (also known as k-mers in computational biology) coming directly from reads into a De Bruijn graph-like structure (Audemard et al., 2018). They argue that since only a small part of the genome is expressed, variant detection can be limited to the transcriptome. Their tool, *km*, uses only the abundance of these k-mers in patients to detect genomic abnormalities. While this method does not depend on any type of reference genome, it still shares the same problem as the ones that combine RNA-Seq with other experiments; all these methods require a predefined sequence for analysis. In other words, to find an abnormality in the cancer samples, one must know in advance the exact abnormal sequence to look for.

4.3.3.3. Representation learning for biological sequences. After the success of distributed representations in NLP (Mikolov et al., 2013) some teams have attempted to create distributed representations for biological sequences. Asgari et Mofrad (2015) adapted Word2Vec to create BioVec, distributed representations for biological sequences, based solely on sequence similarity. They report that their representation captures biochemical properties of proteins such as mass, volume and charge. Jaeger et al. (2018) has also extended the model to chemical compounds and observe that modeling chemical compounds with vectors yields a better performance in bioactivity prediction tasks. This work focuses on a different aspect of the problem. We consider the idea of using an unsupervised learning approach to directly learn a representation for RNA-Seq data from scratch, without the need for a reference genome and the corresponding definition of genes as clearly delineated and non-overlapping regions in the overall sequence.

4.3.4. Method

We represent the raw data as $\mathcal{R}$, where each read $\mathbf{r} \in \mathcal{R}$ is a sequence of length 100, where $r_j \in \{A,C,G,T\}$. We define a k-mer (ngram) $\mathbf{x}$ of length k as a subsequence for some read $\mathbf{r}$ from positions l to $l+k-1$:

$$\mathbf{x} = \mathbf{r}_{l:l+k-1}$$

Each patient i generates a set $\mathcal{R}_i$ of reads. We extract k-mers of length 24 from all the reads $\mathcal{R}_i$ to generate the k-mer table $\mathbf{X}_i$, one for each patient. Table $\mathbf{X}_i$ contain K_i unique k-mers from reads $\mathcal{R}_i$.

$$\mathbf{X}_i = \{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_{K_i}\}$$

This table as well as the patient's id are used by our model to learn the k-mer representation space.

The method uses the factorized embedding model combined with an RNN. This specific method was previously shown to provide a multi-view embedding of genomics data. We however found that it lacked patient-specific information on the RNA sequence level. Our modification of this model allows for the integration of this sequence information to the factorized embeddings model. Concretely, the model is given a pair of inputs $< \mathbf{x}_{ij}, \mathbb{1}_i >$ and predicts the corresponding count y_{ij}. We represent each nucleotide as a *one-hot* vector. For example, adenine and cytidine, A and C, would be represented by respectively [0,0,0,1] and [0,0,1,0].

For each input pair, a corresponding pair of embeddings $< \mathbf{u}_{ij}, \mathbf{v}_i >$ is then computed. For $\mathbf{v}_i$, a lookup table is used, so each patient has its own embedding coordinates in patient embedding space. For $\mathbf{u}_{ij}$, we use a bidirectional LSTM to compute the embedding

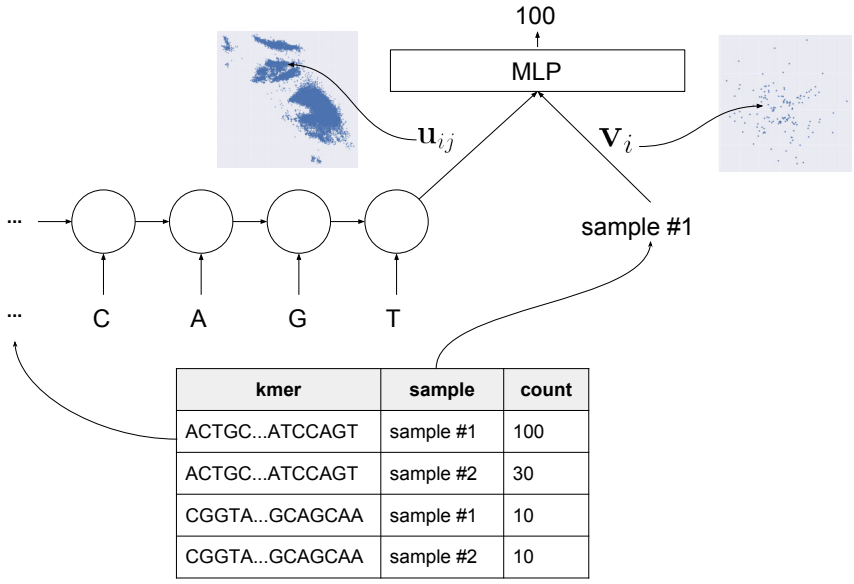


Fig. 4.2. Model overview
The model, where we predict the count of each k-mer based on the their sequence (Using an RNN) and the sample id. We can then visualise the learned embeddings, or use them for some downtream tasks.

coordinates (Hochreiter et Schmidhuber, 1997), where the equations are given by:

$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
\tilde{c}_t &= tanh(W_c x_t + U_c h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot tanh(c_t).
\end{aligned}
$$

The embeddings $\mathbf{u}_{ij}$ are then computed as follow:

$$\mathbf{u}_{ij} = f_{linear}([\mathbf{h}^{\leftarrow}_{rnn}, \mathbf{c}^{\leftarrow}_{rnn}, \mathbf{h}^{\rightarrow}_{rnn}, \mathbf{c}^{\rightarrow}_{rnn}]). \tag{4.3.1}$$

Where $\mathbf{h}^{\leftarrow}_{rnn}$ and $\mathbf{c}^{\leftarrow}_{rnn}$ are the hidden states of the forward RNN, and $\mathbf{h}^{\rightarrow}_{rnn}$ and $\mathbf{c}^{\rightarrow}_{rnn}$ are the hidden states of the backward RNN. A linear function f_{linear} is learned to reduce the dimensionality of the embedding $\mathbf{u}_{ij}$ for visualisation purposes.

The two embeddings are then fed to a MLP to predict the corresponding count:

$$\hat{y}_{ij} = f_{pred}([\mathbf{u}_{ij}, \mathbf{v}_i]). \tag{4.3.2}$$

We use the quadratic loss as a training objective:

$$\mathcal{L} = \sum_{i,j} (\hat{y}_{ij} - y_{ij})^2 \tag{4.3.3}$$

Intuitively, k-mers that come from the same gene would have the same count, since gene expression is counted in terms of transcripts. Overlapping k-mers are also expected to be closer together, since their sequence determines the embedding (Ng, 2017; Kimothi et al., 2016). The model is thus encouraged to group k-mers in embedding space U that generally occurs together across all patients. Similarly, patients that have the same k-mer occurrence profile should be grouped together in the embedding space V. A plan of the model is presented in Figure 4.2.

4.3.5. Experiments

We divide our experiments into three tasks. Each task aims to test the behaviour of our model when presented with DNA sequences that have specific properties. Using all reads from an organism would be optimal for experiments however the scale of the computation is currently intractable (as discussed in the Limitations section). Instead we focus on just specific regions of the genome that contain only a few genes. We determine this region using the a standard alignment pipeline but include the entire region of the gene (not just exons) and also use the aligned data per patient which includes patient specific mutations.

- Task 1: Embedding of genes sequences of high similarity.
- Task 2: Embedding of genes with different sequences
- Task 3: Embedding of genes that participate in a chromosomal translocation.

4.3.5.1. Data. For all our experiments we used aligned, unquantified RNA-Seq data (BAM format files) from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). The dataset contains 149 patients with acute myeloid leukemia (AML), a cancer well known for its multiple chromosomal rearrangements. For all our experiments, we isolated the reads (Figure 4.1) that span specific regions of interest, in order to speed up computation (Table 4.1). Upon extraction of the reads from the BAM files, we obtained a total of 22,008,292 k-mers for all patients, with an average of 125,000 k-mers per patient. We excluded k-mers with occurrence of < 2, as these are most likely sequencing errors. All k-mer occurrences were log-normalised.

4.3.5.2. Experimental details. The Bidirectional LSTM model had 2 layers with 256 hidden units. The size of each embedding (i.e. the output of the bi-RNN and sample id) was of size 2 for visualisation purposes. The prediction model is a two layers MLP of size 150 and 100 units respectively with the ReLU activation function. Each model was trained for 200 epochs with RMSProp with a learning rate of 0.001, $\alpha = 0.99$ and a momentum of 0.

Table 4.1. Regions of interest isolated for all experiments in this paper

gene name	chromosome	start-stop	manuscript section
ZFX	chrX	24148934-24216255	4.3.5.3
ZFY	chrY	2934416-2982508	4.3.5.3
PML	chr15	73994673-74047812	4.3.5.4
RARα	chr17	40309171-40357643	4.3.5.4

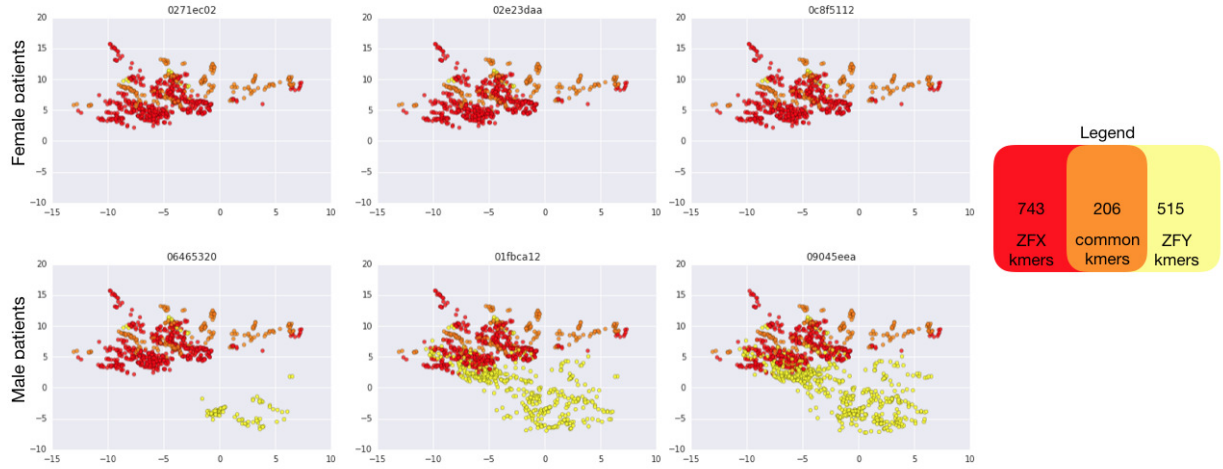


Fig. 4.3. Embedding of homologous genes
A) Embedding space for k-mers in male and female patients. ZFY and ZFX genes share 206 k-mers. Points are coloured according to their matching sequence of origin.

4.3.5.3. Task 1: Representation of genes with highly similar sequences. For the first task, we wanted to test how k-mers from two highly similar sequences would be embedded. The human genome contains many highly similar sequences and it is generally believed that similar sequences can have a similar function in the cell (Pearson, 2013). Although this is not always the case, we chose two genes that share a significant amount of overlap in sequence: ZFY and ZFX genes. It has been reported that they encode proteins of almost identical structure, containing 13 zinc fingers (Palmer et al., 1990; Schneider-Gädicke et al., 1989). While these genes are similar in sequence, they are located on two different chromosomes in the genome. Moreover, ZFY gene is only present in male patients, since it is located on the Y chromosome (Palmer et al., 1990). We argue that this is a case where standard gene annotation separates these two genes into two quantified features and the two reported gene expressions are not related.

We obtained the reference sequences from the reference genome and found that they share 206 k-mers (Figure 4.3A) in the protein coding regions. We expected the model to group the common (all sexes) k-mers in one region and possibly push aside in embedding space the k-mers that were ZFY gene-exclusive.

While our model does not use a reference genome for representation, in order to identify the k-mers in embedding space, we obtained the reference sequences (GRCh38/hg38 Assembly of December 2013) for both genes. We then coloured each k-mer in embedding space according to the gene sequence it belongs to. We also annotated k-mers that match both sequences. As expected, k-mers from female patients were all grouped in the same region, while ZFY-gene exclusive k-mers from male patients were located in another region of the embedding space (Figure 4.3A).

This result supports our claim that when genes share a large part of their sequence, a representation that groups these similar sequences together would contain more information. Although standard pipelines would label these genes as different features and quantify them separately, reads that would fall in the common region would be ambiguous (i.e. matching both regions) and therefore would be either clipped or redistributed according to the mapping software strategy (Kim et al., 2013). We argue that our representation captures both gene expression (via the k-mer counts) as well as the gene sequence similarity.

4.3.5.4. Task 2: Representation of genes with different sequences. For our second task, we tested the embeddings model using two gene regions that are highly different (no homology). This task verifies how the model would arrange in embedding space k-mers that come from different genes. Unlike the previous task, the two gene sequences are very different and the model has to learn sequence similarity between k-mers as well as k-mer abundance variations along the gene. This task is in line with the problems that come from scaling up our method to an entire organism.

The first gene of interest that we chose is the promyelocytic leukemia gene (PML), a tumor suppressor gene, involved in the regulation of p53 response to oncogenic signals. The second gene is the retinoic acid receptor alpha gene (RARα), a gene involved in many core cellular functions such as transcription of clock genes. These genes do not share a significant amount of k-mers and serve as an example to test how the model will arrange k-mers when the sequences are different.

By matching patient k-mers to each of the exons from the reference genome (we used GRCh38/hg38 Assembly of December 2013) of the PML and RARα genes, we notice that the embedding model grouped k-mers in embedding space by matching exon (Figure 4.4A). Exons are subsequences of the transcripts that actually encode the protein (Figure 4.1A). The information recovered by the standard RNA-Seq analysis pipeline corresponds to the exons for these two genes, since they code for both proteins. We conclude that our model recovers correctly information obtainable by the standard RNA-Seq analysis pipeline, when looking at two protein coding regions.

Moreover we observe that there seems to be a discrepancy in the k-mer counts within the gene transcripts. Indeed, for both genes, some exons seem to be present at higher counts

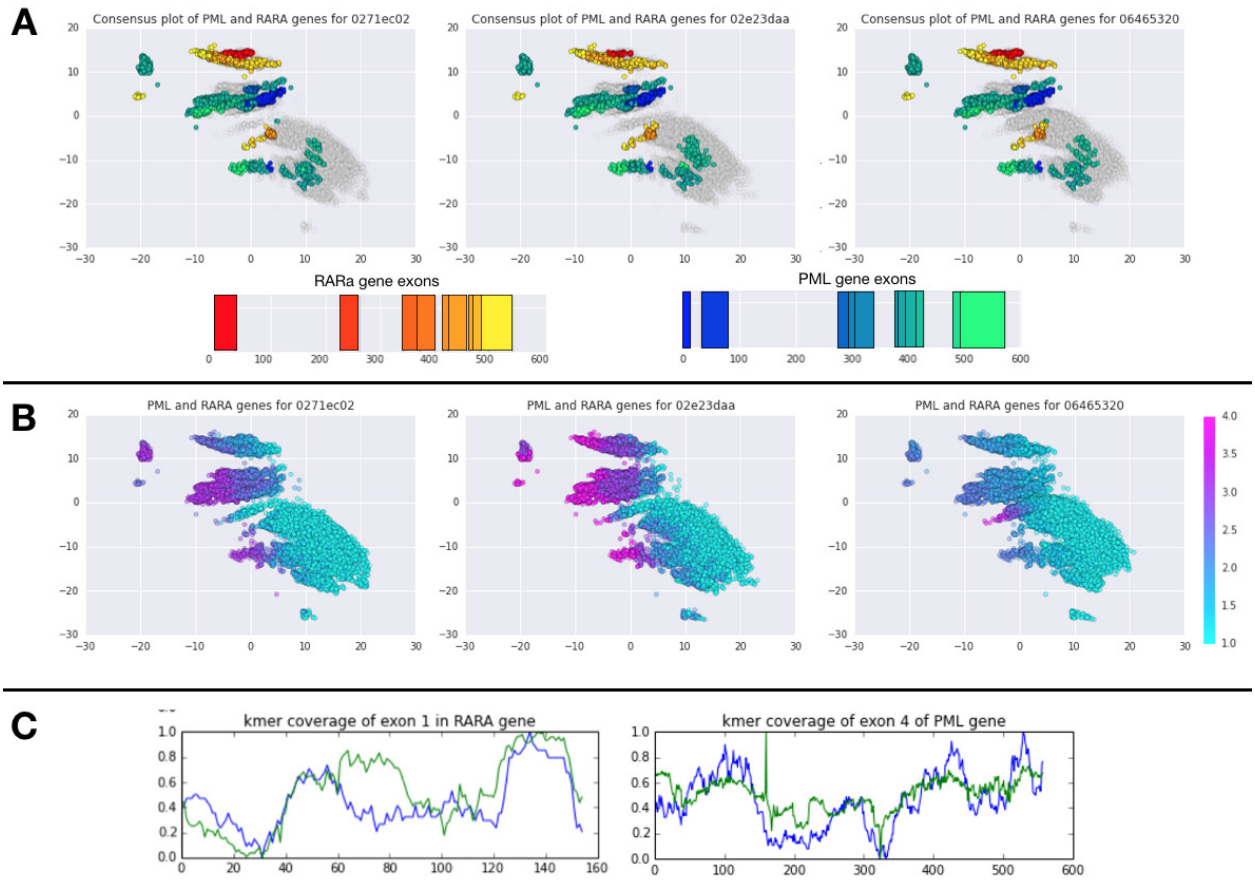


Fig. 4.4. k-mer embedding space

A) k-mer embedding space is shown with k-mers of 3 individuals. Points are coloured according to the corresponding k-mer position in exons for both PML gene (blue), RARa gene (red) and unknown k-mers in grey. B) k-mer embedding space is shown with k-mers of 3 individuals. Points are coloured according to the k-mer counts C) Actual k-mer coverage (blue) of exons and prediction by the model (in green)

(Figure 4.4B). Although the canonical gene expression model states that each transcript is copied in its entirety before being spliced and then translated to a protein (Figure 4.1A), there exists the highly documented phenomenon of sequencing bias (Benjamini et Speed, 2012). This is an entirely experimental bias that is explained by the fact that RNA-Seq is done using an enzyme, polymerase. This enzyme has a bias in terms of G-C content of sequences.

The occurrence of guanine and cytosine (G and C) is measured by counting the G or C nucleotides and then dividing by the total length of the sequence. For example, the sequence $AATTGAGCGA$ would have a G-C content of $(3G + 1C)/10 = 0.4$. We verified the relative exon composition and found that in general, k-mers with a higher count overlap exons with a lower G-C occurrence (no figure shown).

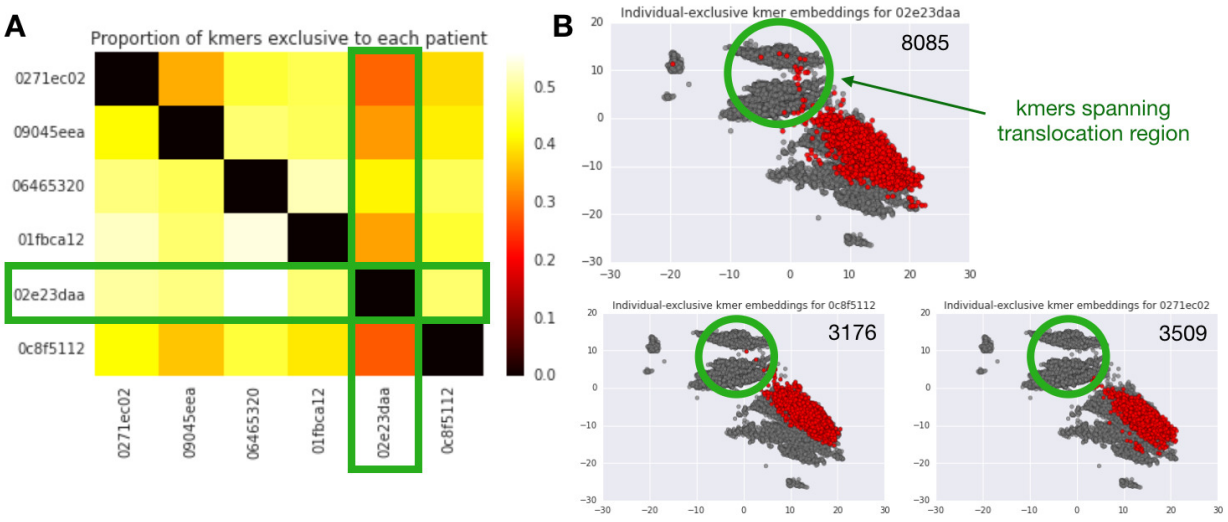


Fig. 4.5. Individual-exclusive k-mers

. A) Heatmap showing the proportion of individual-exclusive k-mers for each pairwise comparisons. B) Patient-exclusive k-mers are highlighted in red. Green circle highlights k-mers that span the translocation.

Moreover, we compared the model's predicted k-mer count and the actual count over various exons from both genes. We found that the predicted k-mer count is proportional to the predicted k-mer count, which confirms that our model captures fine-grained k-mer abundance variations along exons (Figure 4.4C).

4.3.5.5. Task 3: Detection of abnormal genomic structures. The PML and RARα genes are also often involved in a chromosomal translocation (Figure 4.1B), which yields a new fusion gene the PML-RARα in patients that have acute promyelocytic leukemia (de Thé et al., 1991). In our dataset, 10% of the patients are of the acute promyelocytic leukemia subtype and the choice of these two genes serves the dual role for testing the embedding model on two different genes as well as detecting a chromosomic translocation and resulting fusion gene.

Upon further examination of the embedding space for k-mers in task 2, we noticed that a large amount of the k-mers from both genes were not included in the exonic sequences according to the annotation (Figure 4.4B). These sequences are excluded for any of the following reasons: i) the k-mer sequence does not match (exact match) the reference sequence (example in figure 4.1C) ii) the k-mer sequence is not included in the exonic region (3' and 5'-UTR and introns) or iii) the k-mer matches to the t15:17 translocation site in the given patient (example in figure 4.1B). To better address this observation, we performed pairwise comparisons of patient k-mer sets and found that patient 02e23daa has the most different transcriptome, with at most 50% k-mers difference (Figure 4.5A).

We then compared patient-exclusive k-mer sets (Figure 4.5B in red) and found that most patients have between 3-10k exclusive k-mers. We isolated the k-mers from patient 02e23daa and reassembled them into a consensus sequence. We used the software BLAST to perform multiple alignment of this sequence in the reference genome. We report that half of this sequence matches to the PML gene and the other half to the RARα gene, a scenario matching that of a fusion gene, result of a chromosomal translocation (Figure 4.1B). This was confirmed by verifying the clinical data for patient 02e23daa, where the translocation was previously detected and annotated in the clinic (figure not shown). From these results, we conclude that our model captures real genomic abnormalities and allows to extract directly from the k-mer embedding space the abnormal sequence.

4.3.6. Limitations

The main limitation of our model is its scalability. In all the tasks performed in this paper, we heavily restrained the number of k-mers in the dataset. Indeed, without pre-filtering the BAM files, each sample would generate approximately 10-30 billion k-mers (compared to 125,000 per sample in the current dataset). While this number is very high, we suggest that since k-mers are overlapping by definition, it would be possible to sample the k-mers while training and therefore greatly reduce the number of k-mers seen by the model, thus reducing the processing time. Finally, we have not yet explored parallelizing the training onto multiple GPUs, which would greatly reduce the training time.

Finally, while we used the pre-aligned BAM file to filter our regions of interest, our goal is to optimize this model to move away from reference genomes entirely. To do so, we plan on using only two "seed" k-mers and using the total k-mer table, as a means to extract the k-mers of interest. Indeed, exploring all paths supported by k-mers between the two seeds is an (almost) reference-free method to generate k-mer tables, without relying on the alignment product.

4.3.7. Acknowledgements

The results obtained in this publication are based upon data generated by the Leucegene group essentially located at IRIC in Montreal, Canada and supported by Genome Canada and Genome Québec. This data was made possible through human AML specimens provided by the BCLQ, Montreal, Canada. The data is accessible at the following GEO Superseries: GSE67040 GSE48173. This work was supported by CIFAR, Calcul Quebec, and Oncopole. We acknowledge the support of IVADO and the Canada First Research Excellence Fund (Apogée/CFREF). A.T's work has been supported by the Frederick Banting and Charles Best Canada Graduate Scholarships Doctoral Award of the Canadian Institute for Health Research (CIHR).

Chapitre 5

Introduction à l'immunologie adaptative

Le système immunitaire a deux composantes qui varient en spécificité: une composante innée et une adaptative. La composante innée est la première ligne de défense contre les envahisseurs; la réponse se fait à l'échelle de minutes, cependant elle manque de spécificité. En effet, les cellules immunitaires innées sont dotées de *récepteurs de reconnaissance de patrons* (PRR, de l'anglais *pattern-recognition receptors*) et les utilisent pour détecter un grand éventail de patrons moléculaires associés aux pathogènes (PAMPs, de l'anglais *Pathogen-Associated Molecular Patterns*), tels certains acides nucléiques et polysaccharides (Janeway et al. (2001) Section 2-15). Malgré la réponse rapide, ce système ne peut former de mémoire immunitaire. La composante immunitaire adaptative, elle, est dotée d'une haute spécificité antigénique et est caractérisée par l'habileté de former une mémoire immunitaire et de distinguer le Soi immun du non-Soi. La réponse immunitaire adaptative est plus lente mais est spécifique au pathogène et, une fois une mémoire immunitaire formée, le temps de réponse peut être grandement raccourci (Punt et al., 2018). Deux catégories de cellules immunes sont la milice du système immunitaire adaptatif, soit les lymphocytes B et T. Leur nom vient de leur organe de développement, respectivement le thymus (*thymus*, cellule T) et la moelle (*bone marrow*, cellule B), cependant certains affirment que B est pour la bourse de Fabricius (Glick et al., 1956). Le travail présenté dans le chapitre 6 se concentre uniquement sur les cellules T et leur récepteurs et les sections suivantes offrent un survol des thèmes principaux qui y sont abordés. Le lecteur intéressé par plus de détails est invité à consulter les excellents ouvrages *Immunobiologie de Janeway* (Janeway et al., 2001) ainsi que *Immunologie de Kuby* (Punt et al., 2018).

5.1. La tolérance au Soi immun

La tolérance au Soi et l'habileté à reconnaître et éliminer le non-Soi sont les cartes de visite d'un système immunitaire en santé. Pour les cellules T, le Soi immun est présenté par le biais de courts peptides sur les molécules du *Complexe Majeur d'Histocompatibilité* (MHC,

de l'anglais *Major Histocompatibility Complex*). Les cellules T nécessitent d'une interaction avec les molécules MHC via leur *récepteur de cellule T* (TCR, de l'anglais *T cell receptor*) pour leur survie en périphérie (Tanchot et al., 1997). Si lors d'une interactions la cellule T, dite naïve (n'ayant pas été activée auparavant), détecte un peptide du non-Soi, elle initie une cascade d'activation, se différenciant soit en cellule T mémoire ou cellule T effectrice et détruit la cellule présentant le peptide-cible. Mais comment la cellule T différencie le Soi du non-Soi? La prochaine sous-section du présent chapitre offrira quelques pistes à ce sujet.

5.1.1. Le complexe majeur d'histocompatibilité

Les molécules de MHC se distinguent en deux classes de par leur expression dans les cellules ainsi que par les caractéristiques des peptides qu'elles lient. Les molécules de MHC de classe I sont présentes sur toutes les cellules nucléées du corps (à quelques exceptions près) tandis que les molécules de MHC classe II ne sont présentes que sur les cellules présentatrices d'antigènes (cellules dendritiques, cellules B, macrophages) (Janeway et al. (2001) Chapitre 5). Le MHC de l'humain est également appelé HLA (de l'anglais *Human Leukocyte Antigen*). L'humain possède trois régions (locus) principales encodant des molécules MHC classe I: HLA-A, HLA-B et HLA-C. Il y a également trois régions (locus) additionnelles pour les principales molécules MHC classe II: HLA-DP, HLA-DQ et HLA-DR (Punt et al., 2018). La région du MHC est la région génomique la plus polymorphique dans le génome humain, c'est à dire qu'il a plus de variation dans la séquence génique dans cette région qu'ailleurs dans le génome (Robinson et al., 2020). En effet, il existe plus de 23 000 différentes allèles dans la base de données IMGT/HLA (Gonzalez-Galarza et al., 2020; Robinson et al., 2020). Cependant, ces allèles ont une fréquence de recombinaison de près de 0.5%, c'est à dire que les deux allèles d'un individu s'échangent rarement de l'information génique afin de créer de nouvelles séquences. De ce fait, les individus héritent des allèles sous la forme de groupes, en fonction des allèles des parents (Punt et al., 2018). On appelle un groupe d'allèles transmis de manière héréditaire *haplotype*. Chaque être humain possède 12 allèles de MHC classe I et la vraisemblance que deux individus aient des *haplotypes* identiques grandit si les individus sont apparentés. Or, les allèles de l'individu ont une influence directe sur la nature du Soi présenté aux cellules T et c'est ce dont il sera question dans la prochaine section.

5.1.2. L'immunopeptidome

Chaque allèle MHC diffère quant aux peptides qu'elle peut lier. Les peptides liés par les molécules MHC et présents en surface cellulaire chez un individu sont appelés collectivement *l'immunopeptidome*.

Ce dernier est très plastique, c'est-à-dire qu'il change face à des infections, à des stimuli externes et avec l'âge, reflétant les changements en contenu protéique cellulaire (de Verteuil

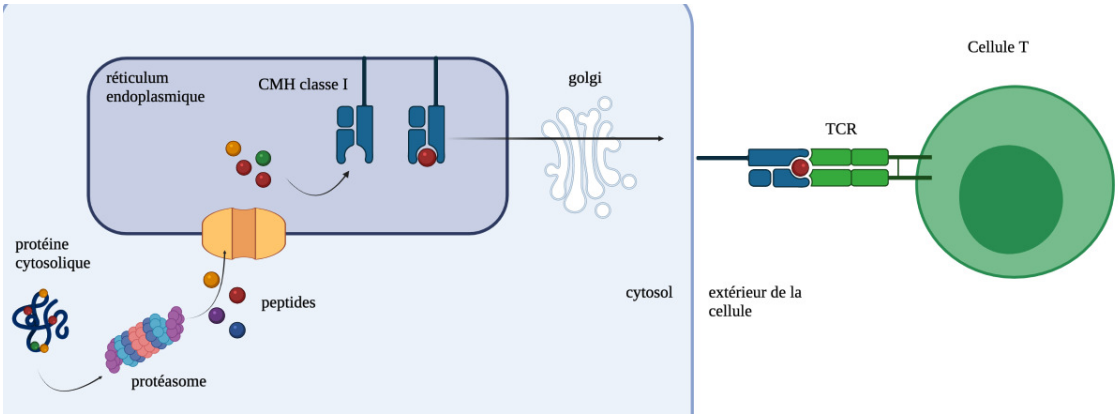


Fig. 5.1. Schéma de la dégradation des peptides et leur présentation en surface cellulaire par les molécules du MHC classe I

et al., 2010; Granados et al., 2009). La voie cytosolique de présentation peptidique spécifique aux molécules MHC classe I se fait en une série d'étapes recrutant plusieurs complexes protéiques et le schéma en Figure 5.1 en montre les étapes principales. Les protéines cytosoliques sont ciblées pour la dégradation par le protéasome. Le protéasome est un complexe protéique de forme cylindrique dont la fonction principale est la dégradation de protéines en courts peptides. Le protéasome utilise le transporteur associé au traitement d'antigène (TAP, de l'anglais *transporter associated with antigen processing*) pour acheminer les peptides à l'intérieur du réticulum endoplasmique. Ici, les peptides se lient aux molécules MHC et sont exportés par la voie d'excrétion normale vers la surface cellulaire, par l'appareil de Golgi (Figure 5.1). Le protéasome a le potentiel de dégrader toutes les protéines, cependant seulement une fraction des peptides se lient aux molécules MHC. En effet, les molécules MHC différentes ont des affinités de liaison différentes et donc le contenu de l'immunopeptidome sera dépendant des allèles d'un individu.

5.2. Les cellules T

Dans la section précédente nous avons vu la nature du Soi pour les cellules T et la prochaine section aura pour thème la maturation des cellules T ainsi que leur éducation à différentier le Soi du non-Soi.

5.2.1. La maturation des TCR

Le thymus, est un organe immunitaire spécialisé, localisé au dessus du coeur et il constitue le lieu de la maturation des cellules T. Les précurseurs de cellules T voyagent du cortex vers la medulla du thymus, tout en passant à travers des étapes clé de la maturation. Durant la maturation, la cellule T crée son TCR par un processus appelé recombinaison V(D)J,

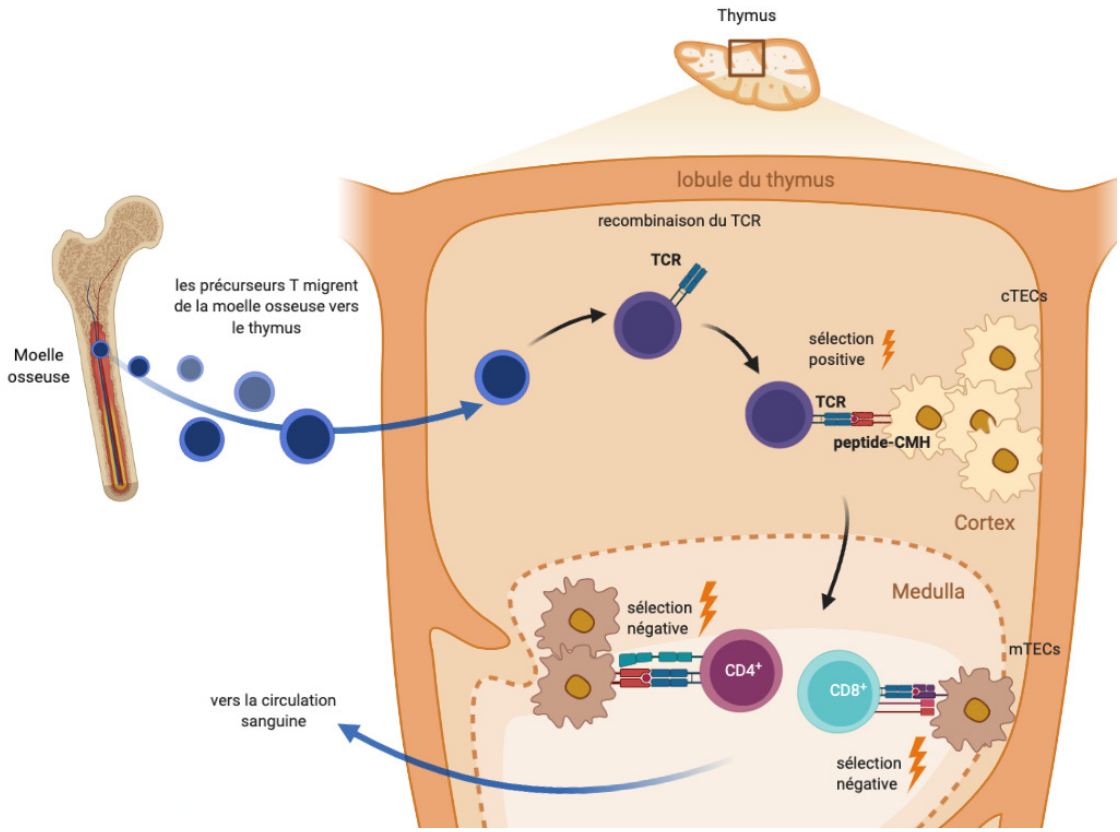


Fig. 5.2. Schéma des étapes de la maturation des cellules T dans le thymus

sélectionne une des deux molécules co-stimulatoires (CD8 ou CD4) et passe un processus de sélection rigoureux avant d'être exportée en périphérie (Figure 5.2).

Le processus de sélection s'assure que les cellules T migrant en périphérie ne sont pas auto-réactives et reconnaissent convenablement le non-Soi. Il est hautement sélectif, quelque 97% des cellules T ne survivront pas le processus de sélection (Egerton et al., 1990; Thomas-Vaslin et al., 2008; Merkenschlager et al., 1997). Le processus complet se divise en deux grandes étapes: i) *la sélection positive* et ii) *la sélection négative*. et est supporté par les cellules épithéliales thymiques (TEC). Le processus de sélection positive élimine les cellules incapables de reconnaître les molécules MHC. En effet, comme les cellules T ont besoin d'une constante stimulation par des molécules MHC pour leur survie, les cellules ne pouvant pas effectuer une interaction adéquate sont tuées par sous-stimulation (négligence).

L'étape de sélection négative, elle, élimine les cellules reconnaissant avec trop de force des peptides du Soi. Ce processus est supporté par les cellules épithéliales médullaires thymiques (mTECs, de l'anglais *medullary thymic epithelial cells*). Ces cellules présentent sur leur surface cellulaire sur leurs molécules MHC des peptides du Soi. En fait, ces cellules expriment une grande quantité de protéines trouvées dans divers tissus en périphérie avec le seul but d'en présenter les peptides en surface cellulaire pour la sélection négative. Ce processus

catalysé par le facteur de transcription AIRE (de l'anglais *Auto-immune regulator*), s'appelle *promiscuous gene expression*, ou expression génique de promiscuité. Les cellules T survivant la sélection négative sont donc tolérantes au Soi immun et vont circuler dans le corps et interagir avec les complexes peptide-MHC dans les tissus du corps. Un processus de sélection fautif mène à plusieurs maladies auto-immunes, dont le APS-1 (de l'anglais *Autoimmune polyendocrine syndrome type 1*), où des mutations dans le gène encodant AIRE mène à son activité fautive et donc l'export en périphérie de cellules T auto-réactives (Bruserud et al., 2016). Finalement, l'ensemble des récepteurs T d'un individu est appelé *répertoire TCR* ou *répertoire T*. Dans les prochaines sections je vais me pencher sur la diversité des répertoires TCR ainsi que les méthodes computationelles qui sont utilisées pour leur analyse.

5.2.2. La diversité des TCR

Chaque TCR est un hétérodimère, composé d'une chaine α et d'une chaine β (ou dans certains cas γ et δ) qui se recombinent indépendamment dans les cellules T, durant leur maturation dans le thymus (Section 5.2.1). La grande majorité des TCR trouvés dans le répertoire humain sont des TCR $\alpha\beta$ et donc je ne vais me concentrer que sur ces derniers (Glusman et al., 2001). Chaque TCR est constitué de régions (boucles) CDR1, CDR2 et CDR3, dont la région CDR3 est la région variable, responsable pour l'interaction avec le peptide (De Simone et al., 2018). Cette variabilité est due au fait que la boucle CDR3 chevauche les jonctions entre les gènes V, D et J suite à leur recombinaison. En effet, la recombinaison de chaque chaine du TCR se fait via un processus stochastique appelé *recombinaison V(D)J*, où un gène V, un gène D et un gène J sont sélectionnés et recombinés ensemble (Figure 5.3). Une séquence constante (C) fait office d'ancrage au TCR en surface cellulaire (Figure 5.3). La Figure 5.3 contient une représentation visuelle des étapes de recombinaison pour la chaîne β; la recombinaison de la chaîne α se fait de manière similaire (avec les gènes Vα et Jα) mais le gène D est omis. Dans le génome humain, il y a 67 gènes V, 2 gènes D et 14 gènes J pour la chaîne β (Punt et al., 2018), ce qui mène à 1876 combinaisons de CDR3β possibles. De la diversité additionnelle dans les séquences CDR3 est créée avec de l'addition de nucléotides aléatoires par la protéine TdT (de l'anglais *Terminal deoxynucleotidyl transferase*). Le nombre théorique de séquences possibles est donc de 10^{15} (Mayer et al., 2019) à 10^{61} (de Greef et al., 2020) mais en pratique, chaque personne compte près de 4^{11} récepteurs différents dans son répertoire (Jenkins et al., 2010b). Comme la sélection des TCR dépend directement des peptides présentés par les molécules MHC (voir section 5.2.1 et ces derniers changent selon l'haplotype HLA de l'individu, les répertoires TCR sont hautement spécifiques à l'individu. Finalement, il est à noter que la probabilité de sélection des gènes V, D et J n'est pas uniforme et des équipes ont même créé un modèle

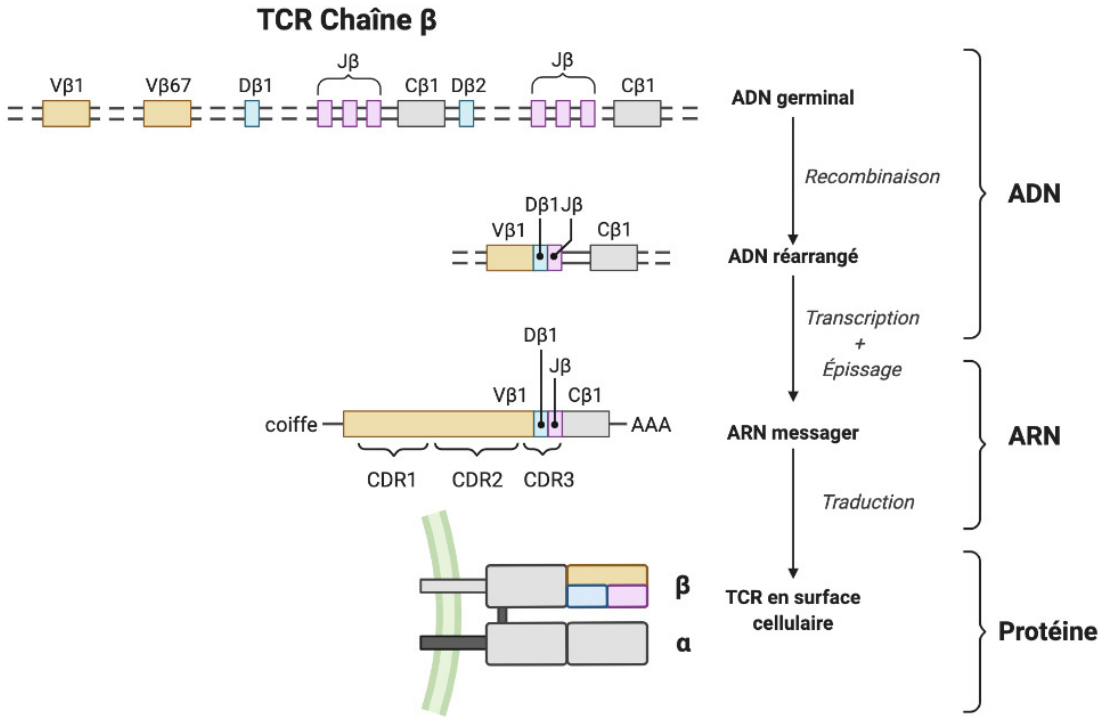


Fig. 5.3. Schéma des étapes de la recombinaison V(D)J de la chaîne β du TCR

probabiliste décrivant ces observations (Marcou et al., 2018). La Section 5.3 offre plus de détail sur les méthodes computationnelles abordant la diversité des répertoires TCR.

Les TCR individuels ainsi que les répertoires sont étudiés dans divers contextes de recherche, tels les maladies infectieuses, les maladies auto-immunes, ainsi que le cancer. En effet, des bases de données telles la VDJdb (VDJ database) ainsi que le McPAS-TCR (base de données sélectionnée manuellement de TCR associés aux pathologies, de l'anglais *Manually curated pathology associated TCR databate*) et IEDB (base de données d'épitopes immuns, de l'anglais *Immune Epitope Database*) répertorient une grande quantité de TCR ainsi que des peptides associés (Tickotsky et al., 2017; Vita et al., 2019b). En plus de la composition en acides aminés de chaines CDR3β spécifiques, la distribution de population de divers TCR est étudiée et les répertoires individuels sont comparés entre eux. Vu que lors de son activation, une cellule T prolifère, une plus grande quantité d'un certain TCR peut être un indice d'une infection récente (Emerson et al., 2017). Dans la prochaine section il sera question des méthodes computationnelles développées pour l'analyse de récepteurs immuns. Il est à noter que le consensus général est que la chaine β du TCR est responsable de l'interaction avec le peptide et donc la plupart des séquençages à haut débit de TCR se font uniquement pour la chaîne β, vu que cette dernière offre le plus d'information quant à la nature du peptide reconnu par le TCR.

5.3. Méthodes computationnelles pour récepteurs immuns

Le développement en 2009 du TCR-Seq a mené à une révolution de l'immunologie computationelle et de l'immunologie des sytèmes. Le Rep-Seq (de l'anglais *Repertoire sequencing*) réfère à la fois au séquençage de répertoires TCR et BCR (récepteurs de cellules B) et j'utiliserai le terme TCR-Seq dans le reste de la thèse. Ce qui différentie le TCR-Seq du RNA-Seq est l'amplification de la région CDR3 suite à l'isolation de l'ARN utilisant des séquences d'apprêt (*primer*) spécifiques (De Simone et al., 2018; Valkiers et al., 2022).

Toute expérience de TCR-Seq est cependant grandement sous-échantillonnée. En effet, avec 1 mL de sang, il est possible de séquencer entre 250000 et 1000000 de récepteurs immuns différents (Mora et Walczak, 2019). Vu que le corps humain contient près de 3×10^{11} cellules dont 6×10^9 se trouvent dans le sang (Mora et Walczak, 2019), chaque échantillon ne montre qu'une partie du vrai répertoire TCR d'un individu. Les analyses de séquences ainsi que de répertoires TCR doivent donc tenir compte de ces particularités des données et de leur provenance.

Selon Miho et collègues, les principales analyses faites sur les répertoires immuns, qui incluent à la fois les répertoires de récepteurs de cellules T et B, sont: i) les analyses de diversité, ii) les analyses d'architecture, iii) les analyses d'évolution et iv) les analyses de convergence (Miho et al., 2018).

Dans les analyses de diversité s'insèrent les analyses de séquence, ainsi que les analyses de composition des répertoires T. Les analyses d'architecture, elles, se concentrent plutôt à grouper les séquences basé sur des caractéristiques communes. La catégorie d'analyse d'évolution est très spécifique aux analyses des BCR et traite du phénomène de maturation par affinité. Finalement, la catégorie d'analyses de convergence consiste en des travaux qui comparent plusieurs des répertoires TCR pour y trouver des biais de convergence, notamment dans le cadre de la reconnaissance de peptides spécifiques. Plusieurs modèles et méthodes d'analyse développées depuis peuvent appartenir à plus d'une catégorie, surtout que les catégories ii) et iv) se rejoignent quelque peu en domaine. Dans les prochaines sections je survolerai quelques travaux pertinents ayant proposé des solutions aux différentes problématiques décrites ci-dessus.

5.3.1. Les analyses de diversité

Les analyses de diversité en biologie touchent souvent à des analyses d'un nombre d'espèces dans une certaine population ainsi que la distribution des individus dans ces espèces. Il existe notamment plus d'une soixantaine de mesures de diversité (Tucker et al., 2017) et la plupart calculent un index tenant compte de la richesse (nombre d'espèces différentes) et de la régularité (nombres de représentants de chaque espèce).

Les mesures de diversité sont également importantes dans l'étude des répertoires TCR et les mesures les plus communes en analyses de TCR sont l'entropie de Shannon, le coefficient de Gini et la diversité de Simpson inverse (Tucker et al., 2017; Chiffelle et al., 2020).

La mesure d'entropie $H(X)$ est définie comme suit: soit une variable aléatoire discrète X qui peut prendre l'une des valeurs (espèces) contenues dans $\{x_1,x_2,x_3,...,x_n\}$. À chaque espèce, une certaine probabilité $p(x_i)$ est associée. L'entropie de Shannon se calcule donc de la manière suivante (Shannon, 1948):

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log(p(x_i)).$$

Dans le cadre d'un répertoire TCR, on compte chaque séquence TCR distincte comme une des valeurs x_i et la probabilité $p(x_i)$ est sa clonalité dans le répertoire. Les répertoires varient en terme de distribution clonale et en terme de nombre de TCR distincts.

Les valeurs que peuvent prendre l'entropie de Shannon pour des récepteurs varient donc entre 0 et un certain nombre, proportionnel au nombre de catégories. Une autre mesure fréquemment utilisée est l'index de Simpson inverse. Ce dernier représente la probabilité suite à deux échantillonnages que la variable discrète aléatoire X prenne la même valeur catégorique x_i. Concrètement, les probabilités $p(x_i)$ sont mises au carré et sommées:

$$S(X) = \sum_{i=1}^{n} p(x_i)^2.$$

Dans la littérature, l'index de Simpson est souvent transformé en son inverse: $1 - S(X)$. Finalement, le coefficient de Gini $G(X)$ vient du calcul de la distribution de la richesse de Max Lorenz (Gini, 1936). Les catégories x_i sont triées en ordre croissant et le calcul se fait:

$$G(X) = \frac{\sum_{i=1}^{n}(2i - n - 1)x_i}{n\sum_{i=1}^{n} x_i}.$$

Le coefficient de Gini, quant à lui prend des valeurs entre 0 et 1 et à la différence de l'index de Simpson et l'entropie de Shannon n'est pas dépendant du nombre de catégories. Dans le contexte d'un répertoire TCR, un coefficient de Gini de 0 signifie une distribution égale de chaque TCR et 1 signifie un TCR dominant qui occupe la totalité de répertoire.

Dans la catégorie d'analyse de diversité, il est question de la nature de la séquence ainsi que de la diversité des séquences dans le répertoire TCR. Les équipes des laboratoires de Alexandra Walczak et Thierry Mora ont publié au courant des dernières années plusieurs outils effectuant ce type d'analyses. On y trouve l'outil *IGoR*, qui modélise un scénario de recombinaison V(D)J (détails à la section 5.2.2) pour une séquence de nucléotides codant pour un TCR donné (Marcou et al., 2018). Ce modèle calcule un alignement de chaque séquence contre la banque de données des séquences correspondant aux gènes V, D et J dans le génome de référence. L'outil *OLGA*, très similaire à *IGoR*, quant à lui, permet de

calculer la probabilité de recombinaison à partir d'une séquence d'acides aminés (Sethna et al., 2019). Outre la création de ces outils, les auteurs ont pu montrer un certain degré de chevauchement entre les répertoires d'individus, en soulignant un certain biais général pour certaines combinaisons de recombinaison (Elhanati et al., 2018). En effet, avec assez de séquences pour un seul individu, les auteurs ont entraîné un modèle personnalisé calculant des biais de recombinaison différents pour chaque individu. Leur conclusion était la présence d'un partage de séquences, surtout basé sur une convergence de recombinaison de séquence ainsi que la redondance du code génétique (Elhanati et al., 2018). Une autre équipe a montré que tout ce qu'apprend le modèle OLGA (Sethna et al., 2019) peut également être approximé par un auto-encodeur variationnel (VAE, de l'anglais *Variational Auto-Encoder*), un type d'hybride entre le réseau de neurones artificiel (ANN) et le modèle graphique (Davidsen et al., 2019). Dans cet article, les auteurs ont conclu que les règles de recombinaison peuvent être apprises par un ANN avec assez de données et sans les lui fournir explicitement.

5.3.2. Analyse d'architecture de répertoires

Du côté de l'analyse d'architecture de répertoires, l'idée est plutôt de développer des mesures pour grouper des TCR basé sur une certaine caractéristique similaire. Par exemple, l'outil ALICE (de l'anglais *Antigen-specific Lymphocyte Identification by Clustering of Expanded sequences*) utilise un seuil de dissimilarité d'un acide aminé pour créer des graphes de voisinage et pour partitionner les séquences en *clusters* (Pogorelyy et al., 2019). Dans les résultats rapportés par les auteurs, ALICE arrive à identifier des TCR spécifiquement expansés suite à une infection virale (Pogorelyy et Shugay, 2019).

Outre la distance Levenshtein, qui correspond au nombre de lettres non-correspondantes entre deux séquences, exploitée par ALICE, d'autres mesures de similarité ont été explorées, notamment des distances basées sur les propriétés physicochimiques des acides aminés. En effet, la distance de Levenshtein assume que remplacer un acide aminé par un autre aurait le même effet, peu importe l'acide aminé. Or, les acides aminés différent entre eux sur la base de propriétés physicochimiques, tels la polarité, la charge et l'encombrement stérique. BLOSUM62 et Grantham sont deux mesures de distance rendant compte de ces propriétés (Henikoff et Henikoff, 1992; Grantham, 1974). Plus formellement, chaque séquence $\mathbf{S}$ est constituée d'entiers symbolisant les divers acides aminés (l'ordre est préétabli d'avance, par exemple via une fonction prenant en entrée une lettre symbolisant un acide aminé et produisant l'entier correspondant). La distance entre une paire de séquences d'acides-aminés $\mathbf{S}_1$ et $\mathbf{S}_2$ est calculée de la manière suivante:

$$D(\mathbf{S}_1,\mathbf{S}_2) = \sum_{i=1}^{n} C[\mathbf{S}_1^i,\mathbf{S}_2^i],$$

où **S** correspond aux vecteurs de lettres que sont les séquences et C est une matrice de 20 par 20 contenant les valeurs de distances pour chaque paires d'acides aminés, dont les coordonnées (ex.: $C[1:3]$) correspondent aux acides aminés encodés par les entiers (par exemple *alanine* et *acide aspartique*). J'effectue ici un abus de notation afin de simplifier, Cette matrice C change selon le type de mesure utilisée, par exemple, la distance Levenshtein score pour chaque non-correspondance une distance de 1 et 0 pour une correspondance (*match*). Pour BLOSUM, qui est un acronyme de *Blocks Substitution Matrix*, la matrice C contient de l'information de substitution évolutionnelle d'acides aminés (Henikoff et Henikoff, 1992). Grantham, quant à elle, est une mesure de dissimilarité entre les acides aminés qui tient plutôt compte de la polarité, l'encombrement stérique et de la charge, en plus du taux de substitution (Grantham, 1974).

Dash et collègues ont publié une méthode nommée TCRdist, se basant sur la matrice de substitution BLOSUM62 pour calculer les distances entre les paires de TCR (Dash et al., 2017; Mayer-Blackwell et al., 2021). Ce calcul est à la base de l'outil CONGA, un modèle qui construit simultanément deux graphes: l'un utilisant TCRdist pour grouper les TCR basé sur leur similarité de séquence et l'autre, utilisant des données scRNA-Seq pairés au TCR-Seq pour grouper les cellules ayant une expression génique similaire (Schattgen et al., 2021). Cette méthode se base sur l'idée que deux cellules ayant des TCR très similaires auront possiblement des expressions géniques similaires, car elles auraient des réponses similaires à des envahisseurs et donc seront recrutés au site d'infection de la même manière. Les auteurs ont ainsi découvert des associations entre les séquences TCR et l'expression génique des cellules T et décrit une nouvelle sous-population de cellules T (Schattgen et al., 2021).

Une étape de granularité d'analyse supérieure consiste à comparer directement des répertoires. Pogorelyy et al ont élaboré une méthode comparant la réponse immunitaire suite à une vaccination (Pogorelyy et Shugay, 2019). Dans cette publication, les auteurs ont effectué la vaccination de 8 individus sains et ont séquencé des échantillons pris à des temps suite à la vaccination. Afin de détecter les TCR ayant répondu au vaccin, les auteurs font la modélisation suivante: Chaque échantillon est séquencé en duplicat qui sert à modéliser la présence et l'expansion d'un TCR côte à côte. En effet, les auteurs motivent ce choix par l'incertitude autour de l'absence d'une séquence TCR et le sous-échantillonage du TCR-Seq par rapport aux répertoires complets. Ils utilisent un modèle de Poisson pour modéliser la présence "de base" d'une séquence dans l'échantillon et utilisent la *loi de puissance* (*Power Law*) pour modéliser l'expansion des TCR à travers plusieurs *time points* d'analyse. De cette manière, les auteurs ont extrait un nombre de TCR de confiance ayant répondu au vaccin de la fièvre jaune à travers 5 individus analysés.

5.3.3. Les analyses de convergence

Nous avons vu précédemment que deux TCRs aux séquences similaires peuvent être le produit de convergence de recombinaison, due au biais dans le processus de recombinaison V(D)J (Elhanati et al., 2018). Malgré tout, une certaine redondance dans la séquence du TCR a été également observée pour des TCR réagissant au même peptide (Glanville et al., 2017; Venturi et al., 2008; Qi et al., 2016). En d'autres mots, les TCR reconnaissant le même peptide ont souvent des séquences TCR similaires, même parfois à travers plus d'une espèce (Madi et al., 2014). C'est entre autre la base des analyses d'architecture et le développement de diverses métriques pour grouper les séquences une à une.

Une autre stratégie de comparaison de séquences utilisée par la méthode GLIPH, se base plutôt sur des similarités de k-mers provenant des séquences TCR à travers plusieurs répertoires immuns. Chaque TCR est brisé en courts k-mers (des 4-mers) et les similarités d'interactions avec les peptides sont évaluées sur la base du k-mer, plutôt que la séquence TCR entière (Glanville et al., 2017). Utilisant cette méthode, Glainville et al. concluent que les TCR ayant une séquence similaire et partageant des sous-séquences dans leur TCR répondraient aux mêmes peptides du non-Soi, pointant vers une possibilité de regroupement de séquences TCR contenant le patron à plusieurs places dans la séquence.

Un autre modèle ayant pour but de trouver des séquences "intéressantes" utilise la notion statistique de transport optimal. Olson et collègues utilisent tout d'abord TCRdist pour calculer une distance entre toutes les paires de TCR entre deux répertoires (Olson, 2020). Les auteurs hypothétisent que les répertoires contiennent les mêmes éléments et ont une composition homogène, cependant ces éléments peuvent quelque peu différer en séquence. Ils ont donc utilisé la méthode de transport optimal pour cibler les séquences similaires entre les répertoires individuels et de cette manière détectent les réponses immunitaires au vaccin contre la fièvre jaune (Olson, 2020).

Finalement, utilisant la méthode TCRdist, Mayer-Blackwell et al. ont trouvé des clusters de TCR à la fois spécifiques aux divers HLA et répondant aux peptides provenant du virus SARS-CoV2 (Mayer-Blackwell et al., 2021). Ces applications sont trois exemples concrets d'analyses de convergence, comparant des répertoires afin de trouver des similarités.

5.4. Application du modèle TLT aux données TCRSeq

Dans le contexte d'un atlas cellulaire, les méthodes d'analyse de répertoires immuns décrites dans la section 5.3 ont tous en commun le fait qu'ils ne permettent pas d'apprendre une représentation de l'individu. En effet, même la méthode ANN décrite dans (Davidsen et al., 2019) traite un répertoire à la fois, sans apprendre de représentation pour les individus. C'est donc cette particularité qui rend les modèles de type Factorized Embeddings et ses dérivés attrayants pour l'intégration de multiples répertoires immuns dans un atlas cellulaire

de répertoires. De plus, étant donné que des 10^6 cellules T séquençées, chaque échantillon contiendra entre 10^4 et 10^6 séquences TCR différentes, ceci constitue un jeu de données attrayant pour tester une version du modèle TLT (Chapitre 4) sur un jeu de données plus petit et donc moins sujet aux problèmes de traitement de données. Les données TCR-Seq sont un jeu de données offrant une diversité de séquences pour chaque individu tout en restant de plus petite dimension que le jeu de données RNA-Seq utilisé dans le chapitre 4. J'ai donc choisi d'adapter le modèle de TLT (Chapitre 4) à des séquences TCR et je référerai à cette itération du modèle *TCRome* dans le reste de la thèse.

5.4.1. Description des données

Le jeu de données de N individus de TCR-Seq contient N groupes de M séquences TCR chaque obtenues par séquençage. Chaque échantillon i est donc un ensemble de séquences TCR S_i, où chaque séquence $t \in S_i$. La séquence t est de longueur variable (entre 8 et 27 résidus) et est constituée d'un alphabet de 20 lettres, correspondant aux 20 acides aminés.

5.4.2. Description du modèle

Tout comme pour le modèle TLT, le modèle TCRome apprend simultanément deux fonctions d'encodage f_{TCR} et $f_{patient}$. La fonction $f_{patient}$ prend en entrée un nombre entier, correspondant à l'index de l'individu et effectue un encodage de cet entier dans un espace à k dimensions:

$$f_{patient} : i \rightarrow z_i^{patient},$$

où $z_i^{patient} \in \mathbb{R}^k$. De la même manière, la fonction f_{TCR} encode chaque TCR S_j dans un espace à k dimensions:

$$f_{TCR} : S_j \rightarrow z_i^{TCR}.$$

Ces deux fonctions f_{TCR} et $f_{patient}$ sont dites fonctions d'*embeddings*. Les deux ensembles de coordonnées en k dimensions sont transmises à un perceptron multi-couches (MLP, voir section 2), qui tente de prédire la probabilité que ce patient ait dans son répertoire ce TCR.

Quelques différences avec le modèle TLT sont à noter:

- Les séquences sont composées d'acides aminés et sont de longueur variable.
- C'est un modèle de classification à deux classes, où les classes 0 et 1 symbolisent respectivement l'absence et la présence du TCR dans le répertoire donné.

Pour le premier point de différence, j'ai choisi d'utiliser un réseau à convolution (CNN, voir section 2) et rembourrer (*padding*) les séquences pour qu'elles aient la même longueur. Chaque acide aminé est pré-encodé en utilisant l'encodage *one-hot* ou l'encodage Grantham (Grantham, 1974). Pour le deuxième point de différence, j'ai choisi la fonction de coût de log négatif vraissemblance (NLL, de l'anglais *Negative Log Likelihood*, voir Section 2.1.1), typique pour les classifications binaires.

Comme chaque répertoire individuel ne contient que les TCR observés, j'ai bonifié chaque répertoire d'un nombre de TCR égal non-observés dans cet individus, mais observés ailleurs dans la cohorte. Ces séquences ajoutés artificiellement sont tous de la classe correspondant à "absent" (ou 0) et sont là afin que le modèle puisse apprendre à classifier les séquences.

Vu que chaque cellule T n'a qu'une seule séquence TCR unique, on peut imaginer que l'espace d'encodage des séquences TCR sera en quelque sorte un atlas de séquences ou de cellules, tandis que l'espace d'encodage des patients pourrait être utilisée pour des analyses de plus grosse granularité, au niveau du patient.

5.4.3. Résultats préliminaires

J'ai entraîné un modèle TCRome sur les individus de la cohorte Thome (Thome et al., 2016), qui contient 120 répertoires TCR provenant d'individus âgés entre 2 et 99 ans. Chaque séquence d'acides aminés a été encodée en one-hot ou bien utilisant la matrice de dis-similarité Grantham (Grantham, 1974).

Chacun des modèles utilisait un espace d'encodage en 10 dimensions et pour des fins de visualisation, les espaces d'encodage des TCR a été réduit à deux dimensions utilisant l'algorithme UMAP (voir Section 2.2.2). J'ai tout d'abord visuellement comparé les deux espaces d'encodage des TCR, suivant la méthode d'encodage des acides aminés. Vu que la séquence TCR contient une partie de la séquence du gène V ainsi que le gène J, je m'attendais à ce que les TCR ayant les mêmes gènes J et/ou V soient regroupés dans l'espace d'encodage. C'était effectivement le cas, comme on peut voir dans la Figure 5.4. Dans cette figure, chaque point est un TCR distinct et ils sont coloriés soit par le gène J utilisé ou le gène V et la colonne de droite correspond à un encodage one-hot des acides aminés tandis que la colonne de gauche correspond à un encodage de Grantham (Figure 5.4). La conclusion principale est que l'encodage Grantham semble mieux capturer les gènes J mais pas le gène V, comparé à l'encodage one-hot. En effet, comme Grantham contient de l'information supplémentaire sur la similarité des acides aminés, j'ai également comparé les TCR ayant été associés à des virus de la base de données VDJdb, dans l'espoir de retrouver des groupement basés sur des pathogènes connus.

J'ai tout d'abord sélectionné les peptides annotés dans VDJdb pour le virus de fièvre jaune (YFV) ainsi que le virus de Dengue type 1 (DENV1). Le regroupement basé sur la distance Grantham semblait mieux capturer des similarités de séquences dues à une réactivité virale commune (Figure 5.5). C'était également le cas pour deux peptides spécifiques provenant respectivement de la protéine NS3 du virus de l'hépatite C (HCV) ainsi que de la protéine Gag du virus d'immunodéficience humaine de type 1 (VIH-1) (Figure 5.6). Dans les deux cas, encoder les TCR basé sur une matrice de similarité tenant compte des similarités des acides aminés semble capturer non seulement une similarité de séquence large (gène J) mais

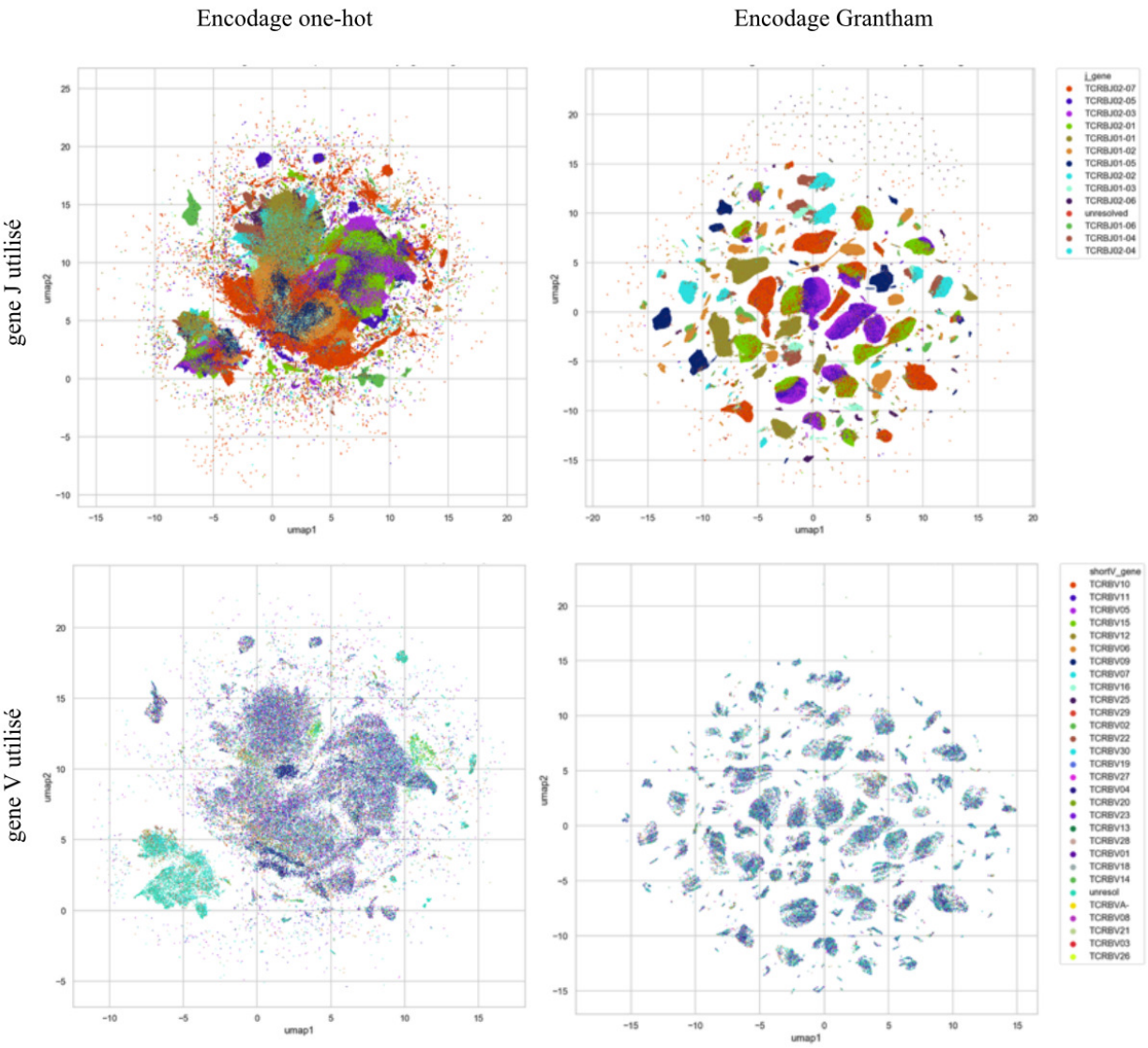


Fig. 5.4. Encodage des TCR appris par le modèle coloriés par identité du gène J (A-B) ou gène V (C-D)

également une similarité de séquence plus aiguë (réactivité commune aux peptides). Pour valider la performance du classifieur à prédire la présence ou non d'un TCR, je l'ai évalué sur des répertoires TCR qu'il n'a jamais vu durant son entraînement. Pour ce faire, utilisant les prédictions faites par le modèle lorsque présenté avec des séquences nouvelles, j'ai calculé une courbe ROC (de l'anglais *Receiver Operation Curve*) qui mesure le taux de vrai et faux positifs. Brièvement, l'aire sous cette courbe (AUC, de l'anglais *Area Under the Curve*) est la mesure utilisée pour évaluer la performance d'un classifieur binaire. Des valeurs avoisinant 0.5 montrent une performance aléatoire, tandis qu'une aire sous la courbe de 1 témoigne d'une performance parfaite (sans erreur). Dans mes résultats, j'ai remarqué assez rapidement qu'il y avait un biais lorsque l'ensemble de séquence d'évaluation était séléctionné basé sur

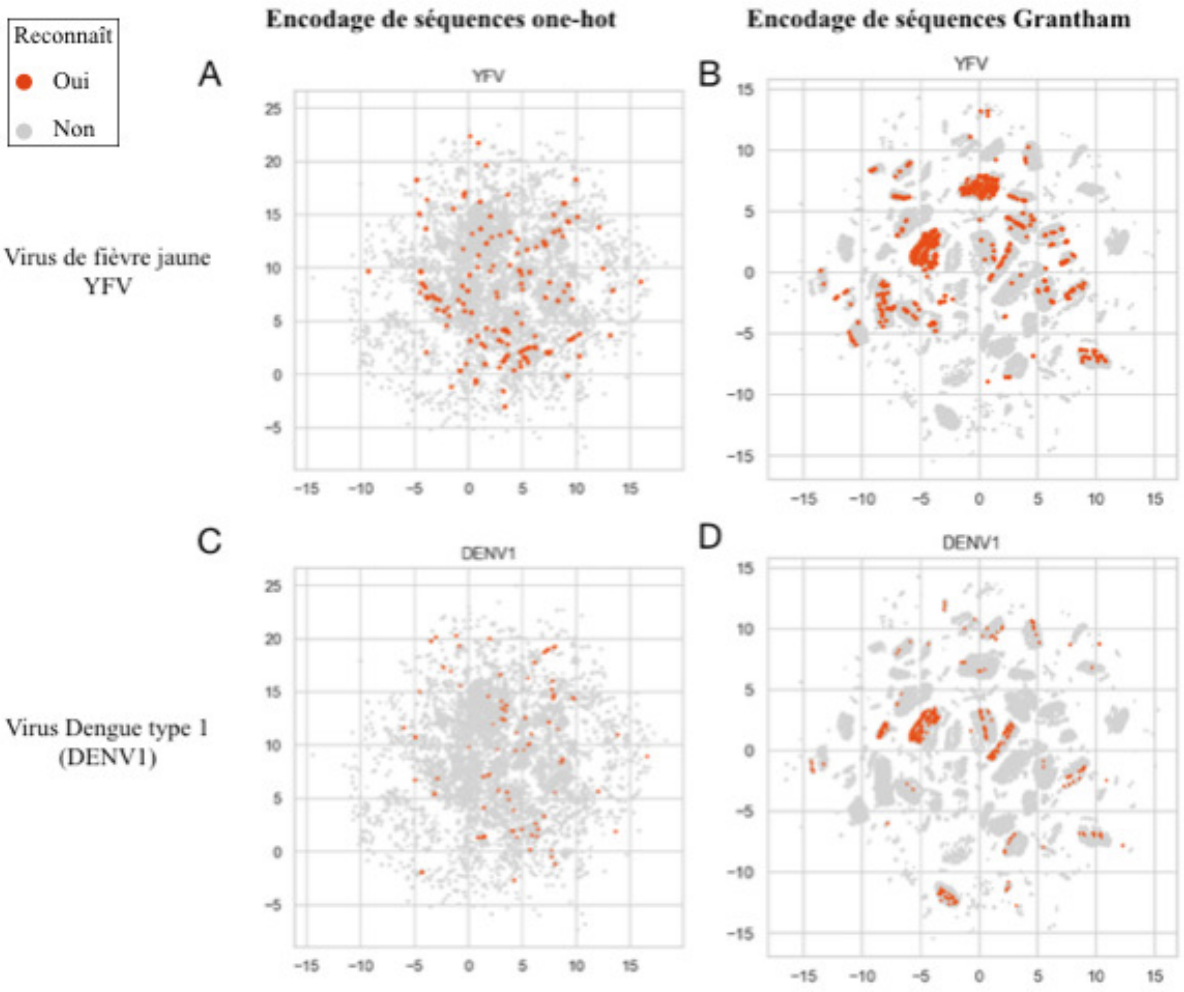


Fig. 5.5. Groupement des TCR répondants à des peptides issus de deux virus

la fréquence du TCR dans le répertoire (Figure 5.7A). En effet, le classifieur performait de manière excellente lorsqu'il était évalué sur les séquences les plus fréquentes dans le répertoire (Figure 5.7 B-C), tandis que sa performance était aléatoire pour les séquences rares (Figure 5.7 D-E).

Ceci suggère qu'il y a une grande redondance entre les séquences les plus fréquentes à travers plusieurs individus, c'est-à-dire que ces séquences "nouvelles" ne le sont pas du tout, car elles sont partagées par plusieurs individus. Les résultats encourageants suggèrent que le modèle TLT est une architecture intéressante pour l'analyse de répertoires immuns mais des particularités du jeu de données semblent interférer avec son fonctionnement. J'ai donc effectué une analyse approfondie des similarités de séquences et de la nature du partage de séquences TCR entre des individus non-apparentés dans ce jeu de données et quatre autres. Les résultats que j'ai obtenus ont mené à l'élaboration de l'article présenté au chapitre suivant.

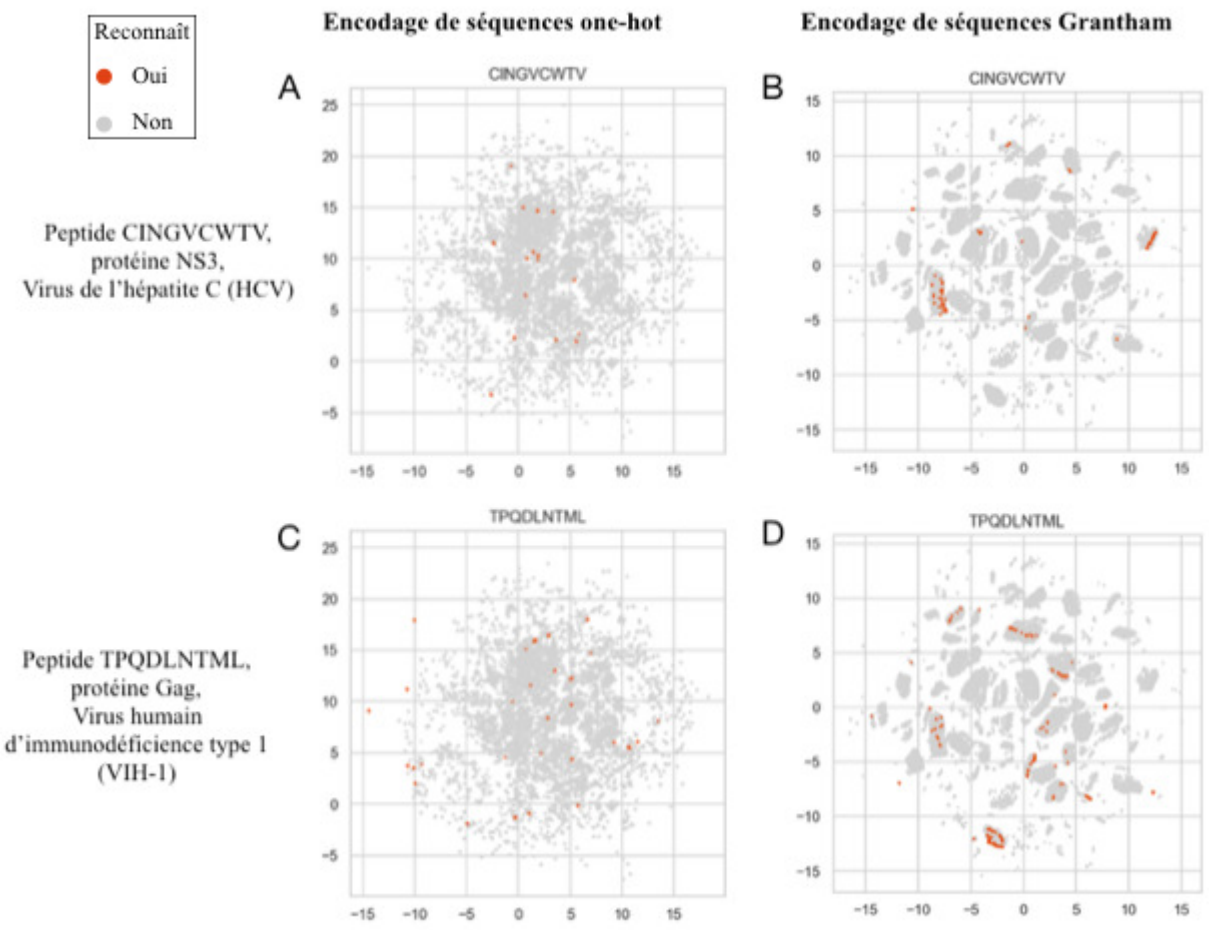


Fig. 5.6. Groupement des TCR répondants à des peptides issus de deux peptides provenant de virus

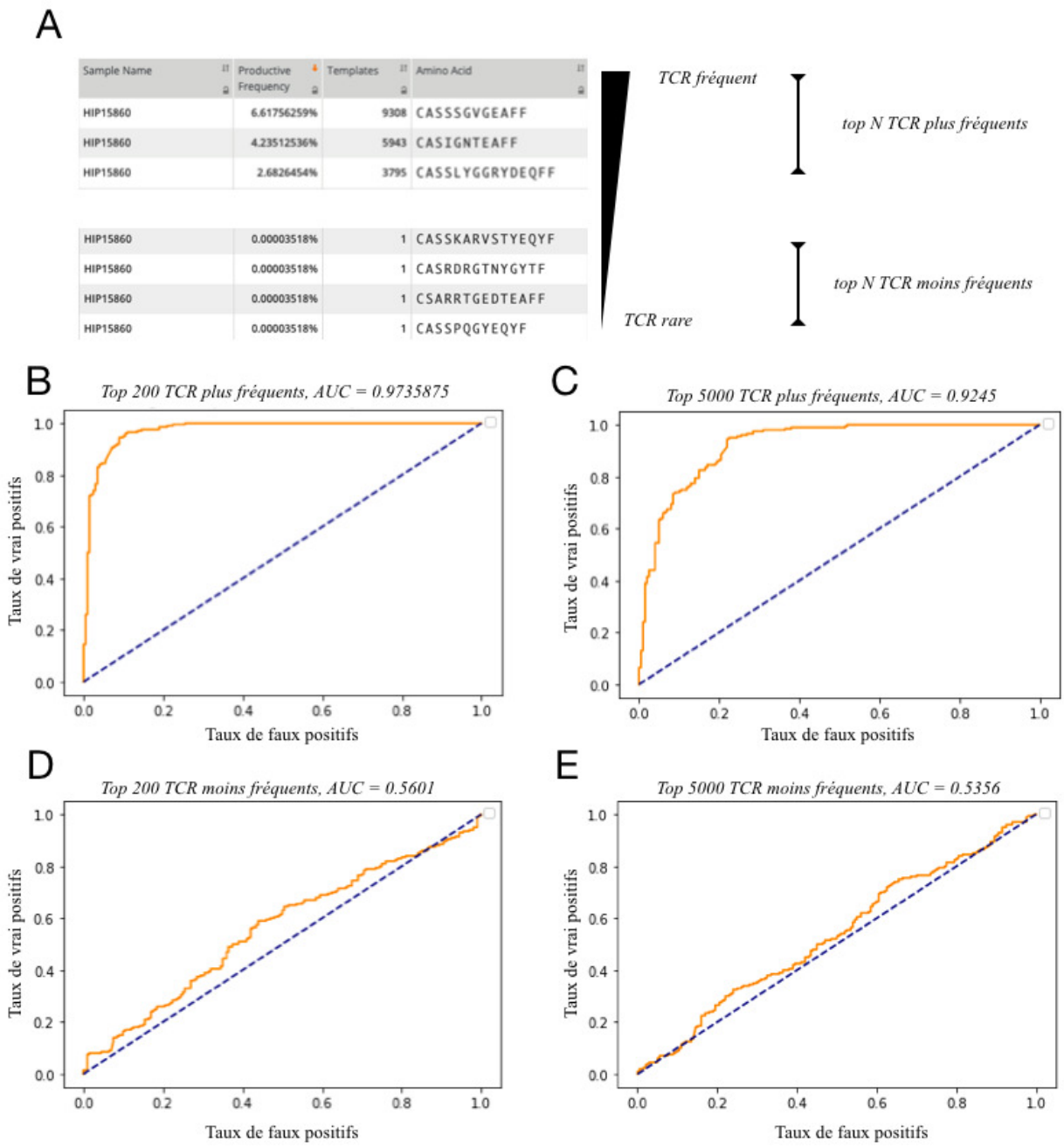


Fig. 5.7. Évaluation de la performance du modèle TCRome sur des séquences que le modèle n'a jamais vues

Chapitre 6

Two types of human TCR differentially regulate reactivity to self and non-self antigens

Assya Trofimov*[1,2,3], Philippe Brouillard[2,3], Jean-David Larouche[1,4], Jonathan Séguin[1], Jean-Philippe Laverdure[1], Ann Brasey[5], Gregory Ehx[1,6], Lambert Busque[5], Silvy Lachance[4,5,8], Sébastien Lemieux[1,2,7,8], Claude Perreault[1,4,5,8,9]

([1]) Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([2]) Department of Computer Science and Research Operations, Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([3]) Montreal Institute for Learning Algorithms (Mila), Montreal, Quebec H2S 3H1, Canada
([4]) Department of Medicine, Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([5]) Maisonneuve-Rosemont Hospital, Montreal, Quebec H1T 2M4, Canada
([6]) Currently Interdisciplinary Cluster for Applied Geno-Proteomics (GIGA-I3), University of Liege, Liege 4000, Belgium
([7]) Department of Biochemistry at University of Montreal, Université de Montréal, Montreal, Quebec H3C 3J7, Canada
([8]) Senior author
([9]) Lead contact

Cet article a été soumis à la revue *Immunity* de Cell Press en janvier 2022

6.1. Mise en contexte

Dans l'article présenté au Chapitre 4 nous avons heurté une embûche de taille quant à la taille du jeu de données. En effet, pour les expériences présentées aux Chapitre 4 nous avons dû nous limiter aux k-mers dans certaines régions seulement, par économie de temps d'entraînement du modèle. Voulant tester le modèle sur un jeu de données de plus grande envergure et sans se limiter à des régions génomiques, nous avons donc cherché un jeu de données plus petit. Par petit, nous désignons un jeu de données nécessitant l'encodage d'un ensemble de séquences pour chaque échantillon, mais avec un nombre de séquences par échantillon plus petit. Nous avons donc choisi d'adapter le modèle TLT (Chapitre 4) à des séquences de récepteurs de cellules T (voir Chapitre 5), des cellules effectrices du système immunitaire adaptatif. Les particularités du modèle et sa performance ont été discutés dans la Section 5.4. L'article présenté dans ce chapitre détaille les deux sous-types de TCR que nous avons découvert suite à l'analyse plus approfondie des données TCR. Dans cet article, nous avons observé des différences inter-individuelles quant aux proportions de quantités clonales des deux sous-types de TCR, variant selon l'âge, le sexe et même dans le cas de certaines maladies médiées par les cellules T (résumé graphique à la Figure 6.1). Outre la découverte et caractérisation des deux sous-types de TCR, ces résultats nous ont démontré la nécessité d'inclure dans l'analyse plusieurs modalités de données.

6.2. Contributions

J'ai mené le projet, conçu et effectué les expériences principales, interprété les résultats, participé aux discussions et j'ai écrit le manuscrit.

Les détails des contributions des auteurs sont les suivantes:
Assya Trofimov, Claude Perreault et Sébastien Lemieux ont conçu l'étude.
Assya Trofimov a effectué les analyses bioinformatiques et l'interprétation des résultats principales.
Jonathan Séguin, Jean-Philippe Laverdure, Ann Brasey, Lambert Busque, Silvy Lachance, Sébastien Lemieux ont effectué le séquençage RNA-Seq et ont participé à l'analyse de résultats.
Philippe Brouillard, Jean-David Larouche et Greg Ehx ont contribué aux analyses bioinformatiques.
Philippe Brouillard, Jean-David Larouche, Greg Ehx, Sébastien Lemieux et

Claude Perreault ont contribué à l'analyse et l'interprétation des données et résultats. **Assya Trofimov et Claude Perreault** ont écrit le manuscrit et tous les auteurs ont édité et approuvé le manuscrit final.

6.3. Résumé français

Basé sur des analyses de séquences TCR provenant de plus de 1000 individus, nous rapportons que le répertoire TCR est composé de deux types de TCR ontogéniquement et fonctionellement distincts. Leur production est régulée par de la variation de production thymique et l'activité de la terminal deoxunucleotidyl transferase (TDT). Les TCR néonataux dérivent de progéniteurs TDT-négatifs et persistent à travers la vie d'un individu, sont hautement partagés entre les sujets et sont polyréactifs envers les antigènes microbiaux et du soi. Ainsi, >50% des TCR du sang de cordon répondent au SARS-CoV2 et d'autres pathogènes communs. Les TCR TDT-dépendants présentent des caractéristiques structurales distinctes et sont moins partagés entre les individus. La production de TCR TDT-dépendants est maximale pendant l'enfance, quand la production thymique et l'activité de TDT sont à leur sommet. Ces TCR sont plus abondants dans les sujets portant des mutations AIRE et semblent jouer un rôle dominant dans la maladie du greffon contre l'hôte (GVHD). Les facteurs influançant à la baisse la productivité thymique (l'âge, le sexe masculin) ont un impact négatif sur la diversité des TCR. Les mâles compensent pour le faible diversité de répertoire TCR par une hyperexpansion de certains clonotypes TCR.

Mots-clés: viellissement, autoimmunité, région de détermination complémentaire 3, maladie du greffon contre l'hôte, infection, régression logistique, SARS-CoV2, dimorphisme sexuel, répertoire de cellules T, recombinaison V(D)J

6.4. Abstract

Based on analyses of TCR sequences from over 1,000 individuals, we report that the TCR repertoire is composed of two ontogenically and functionally distinct types of TCRs. Their production is regulated by variations in thymic output and terminal deoxynucleotidyl transferase (TDT) activity. Neonatal TCRs derived from TDT-negative progenitors persist throughout life, are highly shared among subjects, and are polyreactive to self and microbial antigens. Thus, >50% of cord blood TCRs are responsive to SARS-CoV2 and other common pathogens. TDT-dependent TCRs present distinct structural features and are less shared among subjects. TDT-dependent TCRs are produced in maximal numbers during infancy when thymic output and TDT activity reach a summit, are more abundant in subjects with AIRE mutations, and seem to play a dominant role in graft-versus-host

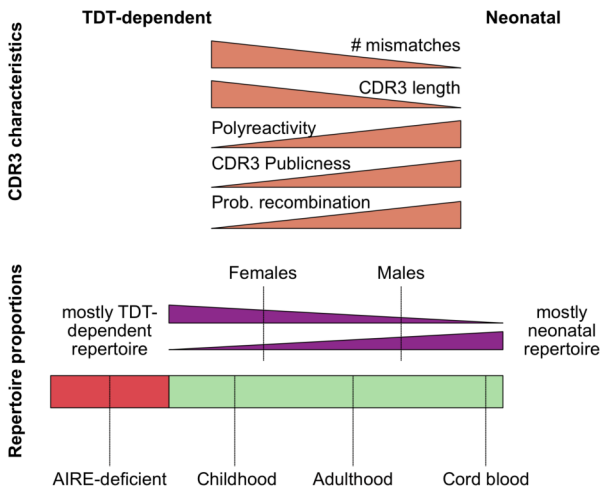


Fig. 6.1. Résumé graphique accompagnant l'article

disease. Factors decreasing thymic output (age, male sex) negatively impact TCR diversity. Males compensate for their lower repertoire diversity via hyperexpansion of selected TCR clonotypes.

Keywords: aging, autoimmunity, complementarity-determining region 3, graft-versus-host disease, infection, logistic regression, SARS-CoV2, sexual dimorphism, T cell repertoire, V(D)J recombination

6.5. Introduction

Jawed vertebrates absolutely need a diversified TCR repertoire because classic $\alpha\beta$ T cells must respond with exquisite specificity to an enormous diversity of ligands (Mittelbrunn et Kroemer, 2021). TCR diversity is generated by somatic recombination of V(D)J gene segments and is further increased postnatally by nucleotide insertion mediated by terminal deoxynucleotidyl transferase (TDT). Notably, neonatal thymocytes, which derive from fetal hematopoietic stem cells, do not express TDT (Rudd, 2020). TDT expression reaches maximal expression in humans between 10 and 40 months (mo) of age, then decreases progressively during adolescence and adulthood (Deibel et al., 1983; Pahwa et al., 1981). Recent estimates of the potential number of TCRs produced by V(D)J recombination range from 10^{15} (Mayer et al., 2019) to 10^{61} (de Greef et al., 2020), which vastly outnumbers the number of distinct TCRs present in a human body. Indeed, the adult human body contains approximately 4×10^{11} T cells (Jenkins et al., 2010a) composed of about 10^{10} TCR clonotypes of various sizes (de Greef et al., 2020; Lythe et al., 2016). Initially, T cell repertoires have been presumed to be almost entirely private, and the occurrence of the same TCR in two unrelated individuals was attributed to coincidence. However, with the development of

high-throughput TCR sequencing and state-of-the-art analytical algorithms, it became clear that interindividual sharing of TCR clonotypes was more common than expected (Pogorelyy et al., 2017; Sethna et al., 2019; Soto et al., 2020). Furthermore, some public clones were found to persist through an individual's life (Chu et al., 2019; Pogorelyy et al., 2017). Still, the extent of interindividual sharing of TCR clonotypes is not precisely known, with estimates ranging from 1% (Johnson et al., 2021) to 3-8% (Soto et al., 2020). Which factors influence TCR diversity? At face value, the reduced thymic output associated with aging and male sex (Clave et al., 2018) should impinge on TCR diversity. However, in contrast to mice, humans can compensate for a reduction of thymic output via minimal adjustments in homeostatic T cell proliferation (Goronzy et Weyand, 2019). Hence, the relation between thymic output and TCR diversity may not be linear. Nonetheless, analyses of TCR sequences in large cohorts have revealed a negative impact of aging on TCR diversity, while the effect of sex remains questionable (DeWitt et al., 2018; Krishna et al., 2020). Furthermore, there is an agreement that HLA polymorphism (i.e., heterozygosity for divergent alleles) positively correlates with TCR diversity and that some pathogens (e.g., CMV) can influence the composition of the TCR repertoire (DeWitt et al., 2018; Krishna et al., 2020). In this study, we analyzed the physical properties of CDR3 beta sequences in seven cohorts of individuals and their implication in immune responses against pathogens, autoantigens, and alloantigens. We found stark differences between male and female TCR repertoires, where males maintain a lower diversity but high clonality (i.e., highly abundant TCR clonotypes). In comparison, female repertoires have a higher diversity of CDR3 at low clonal frequencies. A salient finding was the identification of two non-redundant CDR3 repertoire layers based on physical characteristics, including length, number of insertions, and V/J gene usage. The neonatal layer constitutes the entire TCR repertoire of cord blood, while the TDT-dependent layer appears later in life. Unexpectedly, the cord blood TCR repertoire contains mainly public and polyreactive CDR3s that are responsive to common pathogens.

6.6. Results

6.6.1. Physical characteristics of public and superpublic CDR3s

We defined as public a CDR3aa (CDR3 amino acid sequence) seen in at least two individuals, while a superpublic CDR3aa is present in at least half of the subjects. For our first experiment, we used the Britanova cohort (Figure 6.2A), consisting of 79 healthy volunteers aged from 0-100 (see Methods). We found 2,862,268 public and 15,088 superpublic CDR3aa, of which 21 were ubiquitous (present in all samples) (Figure 6.2A). To define the physical properties of public and superpublic CDR3aa, we first analyzed their V and J gene usage by grouping the CDR3aa sequences by the annotated V or J gene identity. As expected

(De Simone et al., 2018), while each unique CDR3aa sequence was encoded by mostly 1 or 2 J genes, many V genes contributed to the same CDR3aa sequence. At the population level, we observed an average of 26 different V genes per CDR3aa sequence (Figure 6.2B,C). For both public and superpublic CDR3aas, sequences encoded by a higher diversity of J genes were also encoded by numerous V genes (Figure 6.2D,E). In single individuals, up to eight different V genes could contribute to the same CDR3aa (Figure 6.2G). Finally, as previously reported (Gil et al., 2020; Madi et al., 2014; Venturi et al., 2008), we confirmed a positive correlation between the extent of CDR3aa sharing and the number of different nucleotide sequences encoding each CDR3aa (Figure 6.2F). This positive correlation points towards a convergent recombination trend for public and superpublic CDR3aa (Quigley et al., 2010). We then used the software IgBLAST to obtain the number of mismatched (i.e., not germline) nucleotides in each CDR3aa sequence in the Britanova cohort (see Methods). We found that sequences shared by more individuals were also sequences with fewer mismatches (Figure 6.2H). This is consistent with the idea that non-templated nucleotide addition is a random process, and therefore each nucleotide mismatch lowers the likelihood of sequence sharing (Marcou et al., 2018; Sethna et al., 2019). Using the OLGA software (see Methods), we calculated the probability of a CDR3 nucleotide sequence being generated during V(D)J recombination for individual CDR3aa sequences. We confirmed a positive correlation between sequence publicness and recombination probability (Figure 6.2I). Finally, public sequences were shorter than private ones (Figure 6.2J), presumably because non-templated nucleotide addition lengthens the sequence (Marcou et al., 2018; Sethna et al., 2019).

6.6.2. CDR3aa sharing patterns change with age

Analysis of the Britanova cohort revealed a tight correlation between the extent of CDR3aa sharing among subjects and the frequency of the corresponding clonotypes in individual subjects. Superpublic CDR3aa were coded by high-frequency TCR clonotypes, and the most superpublic CDR3aa were found at higher-than-expected cumulative frequencies (Figure 6.3A). We calculated pairwise repertoire overlap distance based on the Jaccard index (see Methods). Using this distance, we performed hierarchical clustering (Figure B.1A,B) and found that individuals clustered by age and repertoire diversity (Figure 6.3A), especially when looking at superpublic CDR3 sharing patterns. Indeed, upon splitting the dendrogram of superpublic CDR3s into four clusters (Figure 6.3B), we found that individuals in the different clusters had significantly different repertoire sizes and age distribution (Figure 6.3C,D). Clusters #2 and #4 showed maximum divergence: individuals in cluster #2 had an average of 0.6×10^6 different CDR3 sequences and a mean age of 40, against only 0.1×10^6 sequences and a mean age of 93 in cluster #4 (Figure 6.3C,D). We wondered whether variations in TDT

activity with age (Deibel et al., 1983; Pahwa et al., 1981) could impact on repertoire sharing. When we aligned each CDR3 to the germline from the reference genome and counted the number of mismatches (see Methods), we found that indeed, cord blood CDR3s (TDT-negative) contained fewer mismatches than samples from other age groups (Figure 6.3E). Finally, when we grouped CDR3aa by descending order of frequency (see Methods), we found that the most frequent CDR3aa displayed fewer mismatches than those with lower frequency, most distinctively in cord blood (Figure 6.3F). Further clone size analyses showed that as individuals age, they accumulate in their repertoires more high-frequency CDR3aa, which have a high recombination frequency (Figure 6.3G). We used a two-step strategy to evaluate the relationship between clonality and the probability of recombination at different ages. We fitted a Gaussian mixture model for each age group, and then we calculated the log-likelihood of data from other age groups under this model (see Methods). We found that models fitted on repertoires of younger individuals did not fit with data from older individuals. However, since models fitted on older individuals had a similar likelihood for all age groups, we concluded that older repertoires retain characteristics of younger repertoires and outgrow them with time (Figure 6.3H). What distinguishes older repertoires from younger ones is a large quantity of high-frequency (presumably hyperexpanded) CDR3aa with a high recombination probability (Figure 6.3G,H). For individual samples in the Britanova cohort, the proportion of public CDR3aa was maximum in cord blood, dropped abruptly in children, and increased progressively with age after that (Figure 6.3I). As a result, the proportion of public CDR3aa in subjects $\geq$ 65 years of age was similar to that in cord blood. The progressive increase in the fraction of public CDR3aa from childhood to old age was even more conspicuous when considering the clonality of each CDR3aa (see Methods, the section on CDR3 sharing): almost 70% of repertoires in individuals $\geq$ 65 years of age were composed of public CDR3aa (Figure 6.3J). Though cord blood and samples from subjects $\geq$ 65 years of age contained similar proportions of public CDR3aa (Figure 6.3I), their clonality was very different (Figure 6.3J). Cord blood cells had a more uniformly distributed repertoire of public CDR3aa, without the hyperexpanded clones present in subjects $\geq$ 65 (Figure 6.3I-J). We validated our observation in two additional cohorts. In the Emerson cohort, containing TCR-Seq data from 666 healthy individuals (Emerson et al., 2017), we could split individuals by CMV status. We found that an age-related public fraction skew can be observed in CMV+ and CMV- subjects (Figure B.2A-D). The Thome cohort is smaller but contains TCR-Seq data from deceased donors' spleen and lymph nodes rather than blood (Thome et al., 2016). T cells were sorted by naive or effector memory phenotype in this study; we, therefore, analyzed those categories separately. We found the same trend of sharing by age group for the naive T cells (Figure B.2E-F) but not for the effector memory T (TEM) cells in secondary lymphoid organs (Figure B.2G-H). The latter divergence warrants further investigation but must be considered preliminary because it is based on analyses of a small cohort of deceased

donors. These results indicate that as individuals age, their repertoire becomes preferentially populated by clones with high recombination frequencies. A high recombination frequency is likely instrumental in the abundance of highly public clones. Another possible explanation could be a preferential expansion of these T cells due to homeostatic proliferation (Murray et al., 2003) or immune activation. To test the latter hypothesis, we used the ERGO software (Springer et al., 2020) to predict the recognition of a large set of HLA-associated peptides recognized by public and private CDR3aa. Our dataset included 25,270 human peptides (Pearson et al., 2016) and 20,961 viral-derived peptides (Vita et al., 2019b). We found that peptides recognized by shorter and more shared CDR3aa had higher TCR binding scores than peptides recognized by longer and less shared CDR3aa (Figure B.3A-C). For further validation, we used single-cell TCR sequencing analyses of T cells responding to 50 different HLA-associated peptides (see Methods). We labeled CDR3aa that recognized more than one peptide as polyreactive. The highly polyreactive CDR3aa had an average length ranging from 10 to 17 amino acids (Figure B.3D), had fewer mismatches (Figure B.3E), and a higher recombination probability (Figure B.3F). One caveat of these data is the limited number of peptides tested. Nevertheless, they highlight polyreactivity as another characteristic of shared CDR3aa and suggest that polyreactivity contributes to the high clonality of public CDR3aa in older individuals.

6.6.3. The impact of sex on the TCR repertoire

Aside from age, male sex is the factor with the most negative impact on thymic output (Clave et al., 2018). Therefore, we analyzed the potential influence of sex on CDR3 repertoire diversity and publicness by grouping individuals into broader age groups to maintain adequate comparison numbers between categories (Figure B.3J). Overall, we found that males had fewer CDR3aa in their repertoires than females: this was the case for public (Figure 6.4A) and superpublic CDR3aa (Figure 6.4B). We then analyzed a set of almost 160,000 high-confidence SARS-CoV2-specific TCRs (Nolan et al., 2020). Since the Britanova cohort was sequenced in 2014, all individuals were unexposed to SARS-CoV2. Notably, we found a high number of SARS-CoV2-specific CDR3aa in these repertoires (Figure 6.4C). Again, males had fewer SARS-CoV2-specific CDR3aa in their repertoires than females (Figure 6.4C). When we examined repertoire diversity using Shannon entropy, we found that repertoires of females were significantly more diverse than those of males (Figure 6.4D-F). Accordingly, small-size CDR3aa clonotypes represented 70% of the repertoire in females and 50% in males (Figure 6.4 G,H). In contrast, hyperexpanded CDR3 clonotypes constituted 30% of repertoire in males and 10% in females. Differences between males and females were present in all age groups and always reached statistical significance in subjects aged 2-45 but not in other groups (Figure 6.4A-F). These results highlight a prominent sexual dimorphism

in the TCR repertoire and suggest that it results from differences in thymic output. Female repertoires are more diverse, and males compensate for their lower repertoire diversity via hyperexpansion of selected TCR clonotypes.

6.6.4. Sharing of disease-specific CDR3s in different age groups

Prompted by the results of our analyses of SARS-CoV2-specific CDR3s, we downloaded and explored the McPAS database, a manually curated catalog of pathology-associated TCR sequences (Tickotsky et al., 2017). We found minimal overlap (0.1-3%) between TCRs in two McPAS categories: microbial pathogens and autoimmune diseases (Figure B.4A). To gain further insight into disease-related CDR3s, we took CDR3aa listed in the McPAS microbial pathogens dataset and analyzed their frequency in subjects from the Britanova cohort (Figure 6.5A). The hierarchical clustering dendrogram was separated into three clusters for individuals (I1 to I3) and five clusters for CDR3aa (C1 to C5). Age had a dramatic influence on both dimensions of this orthogonal clustering. Among clusters for individuals, cluster I2 was composed solely of cord blood samples, whereas individuals in clusters I1 and I3 had a mean age of 82 and 26 years of age, respectively (Figure 6.5B). The CDR3aa-based clustering adopted the following pattern: i) CDR3aa in cluster C1 were present almost exclusively in cord blood, ii) those in cluster C2 were present in few individuals without any clear pattern, and iii) CDR3aa in clusters C4 and C5 were present in young individuals (cord blood and <45 y.o.) (Figure 6.5A). Cluster C3 was remarkable in that it contained the most highly shared CDR3aa; they were found at high frequency in cord blood and lower frequency in almost all other individuals. CDR3aa in cluster C3 were shorter and displayed a greater recombination frequency than CDR3aa in the four other clusters (Figure 6.5C,D). Observations on microbial pathogens-related CDR3aas were replicated in autoimmune disease-associated CDR3aa (Figure B.4B-E). First, cord blood (cluster I1 in Figure B.4B) contained more autoimmunity-associated CDR3aa. Second, the most highly shared CDR3aa (cluster C1 in Figure B.4B) were shorter and displayed a greater recombination frequency than CDR3aa in the four other clusters.

The key finding was that almost all disease-related CDR3aas were found in cord blood. Indeed, 18 to 75% of CDR3aa in individual cord blood samples (Britanova cohort) were responsive to SARS-CoV2 [i.e., present in high-confidence SARS-CoV2-specific TCRs (Nolan et al., 2020)] (Figure 6.5E). A significant proportion of cord blood CDR3aa was also responsive to other pathogens and autoantigens (Figure 6.5E). The large size of the SARS-CoV2-specific TCR dataset explains why more cord blood CDR3aa appeared responsive to SARS-CoV2 than other pathogens and autoantigens. In some cord blood samples, the summed proportions of CDR3aa responsive to autoantigens, SARS-CoV2, and other pathogens were superior to 100% (Figure 6.5E). This is most likely justified by the polyreactivity of public

TCRs (Figure B.3). We conclude that all individuals have many disease-reacting clones in their repertoires before birth. Are disease-related CDR3s present in older subjects? To address this question, we calculated the percentage of disease-related CDR3aa present in top N CDR3aa from individuals of various age groups (Figure 6.5F, S5A-C). Two points can be made from this analysis. First, in cord blood, disease-related CDR3aa are enriched in high-frequency clonotypes. Second, the remarkable representation of disease-related CDR3aa in the "pre-immune" repertoire of cord blood is lost in older individuals.

Our data support the notion that TCRs generated during fetal life can persist (or be continuously generated) for decades in adults (Pogorelyy et al., 2017). More importantly, they show that most of these TCRs participate in a wide variety of immune responses in adult life. Globally, our data presented so far suggest the existence of two types of CDR3: the superpublic ones, shared by many individuals and present before birth, and the private repertoire, dependent on TDT modifications. For the remainder of the study, we will refer to these two types of TCRs as neonatal and TDT-dependent.

6.6.5. Negative selection targets TDT-dependent TCRs

Irrespective of their TCR type, neonatal or TDT-dependent, T cells are subjected to intrathymic positive and negative selection. AIRE mutation selectively and consistently perturbs negative selection and thereby causes autoimmunity (Liston et al., 2003). We, therefore, analyzed the CDR3s of the Sng cohort, which contains subjects with AIRE mutations and healthy controls (Sng et al., 2019). We found that AIRE-mutated CDR3aa repertoires had a lower public fraction than healthy repertoires (Figure 6.6A), with a lower recombination frequency (Figure 6.6B), a higher number of mismatches per CDR3aa (Figure 6.6C) for both regulatory and conventional T cell compartments, and longer CDR3aas (Figure 6.6D). These results point towards enrichment in TDT-dependent CDR3s in AIRE-mutated repertoires. They also suggest that thymocytes with TDT-dependent TCRs are prime subjects of negative selection.

6.6.6. Effect of the TCR repertoire on graft-versus-host disease

To further evaluate the potential impact of the two types of TCRs, we reasoned that the best strategy would be to use a model in which the readout depends exclusively on T cells. Acute graft-versus-host disease (aGVHD) following allogeneic hematopoietic cell transplantation (AHCT) represents such a model. Indeed, donor T cells, particularly the CD4+ subset, are necessary and sufficient for the occurrence of aGVHD (Ni et al., 2017; Socié et Blazar, 2009). They initiate aGVHD via recognition of host alloantigens (Martin et al., 2017; Vincent et al., 2011). Therefore, we analyzed TCRs in purified CD4+ T cells from 73 AHCT donors. Donors and recipients were HLA-matched siblings. The T cells were

obtained from the peripheral blood of donors on the day of transplantation and submitted to RNA sequencing. To extract CDR3 sequences from RNA sequencing reads, we used the MIXCR software (Bolotin et al., 2015b). We classified donors as aGVHD+ or aGVHD-, depending on whether their recipient presented or not severe aGVHD (see Methods). Notably, aGVHD+ donors had lower CDR3 diversity than aGVHD- grafts (Figure 6.7A). We used a treemap to display both diversity and clone size in two representative donors. Treemaps offer a visual representation of diversity at a glance, and we used these plots to compare two representative examples of aGVHD- and aGVHD+ donor repertoires. In the aGVHD+ donor, three hyperexpanded clones occupied almost $\frac{1}{3}$ of the repertoire (Figure 6.7B), while the aGVHD- donor did not have this skew (Figure 6.7C). The CDR3aa in aGVHD+ grafts were longer (Figure 6.7D), had a lower recombination frequency, and more numerous mismatches than CDR3aa in aGVHD donors (Figure B.5D-E). We then split the cohort by the median or quartiles and generated Kaplan-Meier curves to assess the impact of CDR3 features on the occurrence of aGVHD (Figure 6.7E-H, Figure B.6). Overall, grafts containing a higher proportion of CDR3 with neonatal features caused less aGVHD. These features were: CDR3 length in amino acids (Figure 6.7E), percentage overlap with cord blood samples (Figure 6.7F), recombination frequency (Figure 6.7G), and Simpson diversity index (Figure 6.7H). Finally, we used Cox proportional hazards (CoxPH) models to evaluate more accurately the impact of clinical and CDR3 features on the risk of aGVHD. For the clinical characteristics model, the sole significant correlation was a higher rate of aGVHD in male recipients of female grafts (Figure 6.7I). These results are concordant with previous reports (Kim et al., 2016). For the CDR3 model, we found that a high number of neonatal CDR3 and a high average recombination frequency decreased the risk of aGVHD (Figure 6.7J). Other characteristics and clinical traits such as donor age and CMV status had no significant impact (Figure 6.7I,J). Collectively, these results strongly suggest that donors with a higher proportion of neonatal TCRs cause less aGVHD and that aGVHD is initiated primarily by TDT-dependent TCRs.

6.6.7. A stratified model of the TCR repertoire

Our final goal was to evaluate the importance of discrete features in defining neonatal and TDT-dependent TCRs. Our reasoning was based on two assumptions. First, we assumed that cord blood samples contained exclusively neonatal TCRs while all other age groups contained a mix of neonatal and TDT-dependent TCRs. Second, since thymic output and TDT activity reach their zenith during childhood, we postulated that children would generate the greatest diversity of TDT-dependent TCRs. Therefore, to get a pure and diversified population of TDT-dependent CDR3s, we selected CDR3s present in children but not in cord blood. We then confirmed that, compared to neonatal CDR3s, the TDT-dependent CDR3s

were longer (Figure 6.8A) had more mismatches and a lower recombination probability (Figure 6.8B-C). Notably, they also displayed a different V and J gene usage (Figure 6.8D-E). On this dataset, we trained a logistic regression model and random forest to verify if the nonlinearity of the model could have an impact on the performance. Using all the five features (recombination frequency, # mismatches, CDR3 length, V gene, J gene), we performed an ablation study by obtaining all possible combinations of presence/absence, totaling 31 combinations of features (Figure 6.8F). We trained the two models on the dataset for each combination and evaluated their performance on a held-out CDR3 repertoire of each type (the entire individual's repertoire). The performance of each model on the held-out data is represented as a single column, where black squares symbolize the absence and white squares the presence of a feature, and the performance squares are colored by the percentage of accuracy of classification (Figure 6.8F). The CDR3 length was crucial to the model; without the CDR3 length, the model's performance was close to the baseline of 60%, which is the proportion of neonatal CDR3s in the dataset. Adding the length improves classification accuracy by about 10% for all conditions. Numbers of mismatches and V/J gene usage had a more modest effect on the performance, with an accuracy gain of about 5% each. Moreover, V and J gene usage was non-redundant and having both yielded better performance than only having one or the other for both models. The inclusion of the recombination frequency did not impact the performance, most likely because it is largely redundant with CDR3 length (Figure B.7A). Finally, to validate the order of importance of the features, we fit a linear regression model on the presence/absence of features (see Methods). This allowed for the direct comparison of the relative importance of the features based on the coefficients assigned to each feature (Figure 6.8G). We found the following importance hierarchy: length (# mismatches, VJ gene usage) probability of recombination (Figure 6.8G). Consistency between the logistic regression and the random forest models suggests that these features robustly discriminate between neonatal and TDT-dependent TCRs. We then used the trained model to classify each CDR3 in the cohort and found that, as expected, there is a considerable dip in the proportion of neonatal TCRs after birth (Figure B.7B). Then, from infancy to adulthood, there is a progressive increase in the proportion of neonatal clonotypes (Figure B.7B), probably because of their polyreactivity (Figure B.3). Afterward, the proportion of neonatal CDR3 remains relatively stable with a slight trend downwards with advancing age (Figure B.7B).

6.7. Discussion

This report analyzed the amino acid sequence of over 100 million TCR CDR3 beta chains from over 1,000 subjects. Focusing on CDR3 amino acid sequences instead of nucleotides and V/J gene usage allowed us to uncover more numerous public and superpublic CDR3aa

than anticipated. We found stark differences between male and female repertoires, as well as age-specific and disease-specific repertoire features. Age and sex are associated with important differences in immune responses to pathogens and self-antigens (Brodin et Davis, 2017; Liston et al., 2016). Thymic involution is instrumental in decreasing immunocompetence with age and represents a major public health issue, as illustrated by the COVID pandemic (Mittelbrunn et Kroemer, 2021; Palmer et al., 2018; Yousefzadeh et al., 2021). Aside from age, male sex is the factor with the most negative impact on thymic output (Palmer et al., 2018). We report that both aging and male sex are associated with decreased TCR diversity and hyperexpansion of public clonotypes. Female TCR repertoires are more diverse, and males compensate for their lower repertoire diversity via hyperexpansion of selected TCR clonotypes. These data argue for a strong mechanistic link between thymic output and TCR diversity. Analyses of cord blood samples were particularly instructive. In the absence of TDT, TCRs produced before birth have short CDR3s, few mismatches (relative to germline sequences), and a biased V/J gene usage. These neonatal TCRs persist (or are continuously replenished) throughout life, are highly shared among subjects, and are polyreactive to self and microbial HLA-associated peptides. Three factors likely contribute to the large clone size and extensive sharing of neonatal TCRs over a lifetime. First, they have a high recombination frequency; in other words, they are easy to assemble during V(D)J recombination. Second, their high reactivity to self-antigens (Figure B.2A) should theoretically favor their positive selection in the thymus and their homeostatic proliferation in the periphery (Ernst et al., 1999; Hogquist et Jameson, 2014). Third, our analyses of subjects with AIRE mutations revealed that neonatal TCRs were less affected by negative selection in the thymus than TDT-dependent TCRs. Thus, neonatal TCRs may integrate all the "Goldilocks" conditions for intrathymic selection and survival in the periphery. Notably, polyreactivity to self-antigens could also favor the commitment of thymocytes bearing neonatal TCRs toward either the regulatory or alternative T cell lineages (Sood et al., 2021; Vrisekoop et al., 2014). This possibility should be explored in future studies. Remarkably, in individual cord blood samples, 10-80% of CDR3s were reactive to SARS-CoV2, other pathogens, or autoimmune diseases (Figure 6.5E). This means that humans are born with a TCR repertoire that can have a lifelong influence on their response to pathogens and the risk of autoimmunity. From an evolutionary perspective, the size of human populations has been limited by the rate of infant mortality. Hence, it would seem convenient to be born with a polyreactive T cell repertoire responsive to common pathogens. In contrast to neonatal TCRs, TDT-dependent TCRs are longer, less shared, contain more mismatches, and display a different V/J gene profile. Their production is maximal during infancy, when thymic output and TDT activity reach a summit, and slowly decreases after that. We found that TDT-dependent TCRs were more abundant in subjects with AIRE mutations. This suggests that negative selection preferentially eliminates TDT-dependent TCRs. The ultimate role of TDT remains

unclear. By ultimate role, we mean the evolutionary selected biological advantage conferred by TDT. In mice, deletion of TDT does not increase susceptibility to pathogens or the incidence of autoimmunity but decreases the breadth of anti-viral responses (Haeryfar et al., 2008; Kedzierska et al., 2008). However, for the immune system, evolutionary convergence towards a higher diversity is thought to be a protection mechanism to get ahead of the arms race with pathogens (Liston et al., 2021). Therefore, a plausible hypothesis is that the presence of TDT-dependent TCRs confers an additional, more "private" layer of security against the emergence of antigen-loss variants. aGVHD is a harbinger of chronic GVHD and has remained the nemesis of patients and physicians during the entire history of AHCT, partly because its occurrence is unpredictable. Our aGVHD cohort was composed of HLA-matched siblings. In this situation, aGVHD is caused by donor T cells that react against host minor histocompatibility antigens (Vincent et al., 2011; Warren et al., 2012). On the other hand, while histoincompatibility is necessary, it is insufficient to elicit fatal GVHD. Indeed, in patients that received AHCT from donors presenting multiple disparities for minor histocompatibility antigens, only 73% developed aGVHD (Martin, 1991). It has been hypothesized that some AHCT donors might be stronger alloresponders than others (Baron et al., 2007). In our cohort of 73 donor-recipient pairs, the occurrence of severe aGVHD was strongly associated with a low proportion of neonatal TCRs in the donor repertoire. Such a protective effect of neonatal TCRs would explain reports that AHCT with cord blood rather than adult hematopoietic cells may be associated with a lower risk of GVHD (Cohen et al., 2020). If our observation is validated in further studies, it will justify the preferential selection of AHCT donors with a high proportion of neonatal TCRs in their peripheral blood. Together, our data support an emerging model in which the T cell repertoire is composed of two strata with differential reactivity to self and non-self antigens: public neonatal TCRs and private TDT-dependent TCRs. This model is remarkably coherent with insightful theoretical predictions by Vrisekoop and colleagues who labeled the two strata the "somatic" repertoire and the "ur"-repertoire (Vrisekoop et al., 2014). Our model is also consistent with functional studies demonstrating that neonatal T cells can no longer be considered immature versions of adult cells. On the contrary, they are highly functional and respond rapidly to antigenic challenges (Davenport et al., 2020; Rudd, 2020).

6.8. Acknowledgments

This study was supported by grant FDN-148400 from the Canadian Institutes of Health Research (to C.P.). A.T. was supported by a studentship from the Canadian Institutes of Health Research, and J.D.L. by a studentship from the Fonds de Recherche Québec – Santé.

6.9. Author Contributions

A.T., C.P. and S.Le. designed the study. A.T. performed the main bioinformatic analyses and result interpretation. J.S., J.-P.L., S.La., L.B., S.Le. and A.B. performed RNA Sequencing experiments. P.B., J.-D.L. and G.E. contributed to bioinformatic analyses. A.T., P.B., J.-D.L., G.E., S.Le. and C.P. contributed to the analysis and interpretation of data and results. A.T. and C.P. wrote the manuscript and all authors edited and approved the final manuscript.

6.10. Declaration of Interests

The authors declare no competing interests.

6.11. Methods

6.11.1. TCR sequencing datasets

We downloaded TCR sequences and additional data from four non-overlapping cohorts: (Britanova et al., 2016; Emerson et al., 2017; Sng et al., 2019; Thome et al., 2016). A total of 980 human subjects were included in these cohorts, with 403 females and 519 males; the sex of 47 subjects was unknown.

6.11.2. Disease-specific CDR3 sets

A set of 160,000 unique SARS-CoV2-specific CDR3 was obtained from the ImmuneCODE™ database (Nolan et al., 2020). From the McPAS CDR3 datasets, we downloaded the McPAS database on 2021-08-12 (Tickotsky et al., 2017). We included in our study CDR3beta amino acid sequences of human origin found in the two top categories of diseases: Pathogens and Autoimmune.

6.11.3. Combined TCR-Seqsingle-cell RNA-Seq data from antigen-specific CD8+ T cells

We obtained from the 10x Genomics data repository V(D)J sequence information generated by 10x Genomics CellRanger for 150,000 CD8+ T cells isolated from four healthy donors. Data was downloaded from website (`https://www.10xgenomics.com/welcome?closeUrl=%2Fresources%2Fdatasets&lastTouchOfferName=CD8%2B%20T%20cells%20of%20Healthy%20Donor%201&lastTouchOfferType=Dataset&redirectUrl=%2Fresources%2Fdatasets%2Fcd-8-plus-t-cells-of-healthy-donor-1-1-standard-3-0-2`). We did the following data pre-processing. We filtered out barcodes associated with non-unique cells or did not have a resolved CDR3 amino acid or nucleotide sequence. As previously reported

(Schattgen et al., 2021), we found substantial non-specific binding in donor 1 and excluded this sample.

6.11.4. RNA-Seq dataset

The GVHD cohort included 73 healthy sibling-matched donors. Written informed consent was obtained from all patients or their legal guardians before sample collection or hematopoietic stem cell transplantation. For each sample, peripheral blood mononuclear cells (PBMC) were collected, and CD4 T cells were isolated by immunomagnetic positive selection (EasySep Human CD4 positive selection kit; Stemcell Technologies). RNA was extracted from purified CD4 T cells by Trizol-Column (PureLink RNA Mini kit; Thermo Fisher Scientific). RNA was quantified by U.V. spectrophotometry (Tecan Infinite M1000), and quality was verified by Bioanalyzer (Nano RNA Chip; Agilent). Whole transcriptome libraries were prepared with the Ion Torrent Total RNA-Seq Kit v2 (Thermo Fisher Scientific) from 200ng total poly-A enriched RNA (Dynabead mRNA direct Micro Kit; Ambion). Sequencing was done on an Ion P1 chip using the Thermo Fisher Ion Proton System to a minimum of 30M reads.

6.11.5. Isolating CDR3 from bulk RNA-Seq in silico

From each RNA-Seq donor sample from the GVHD cohort, we isolated CDR3 contigs using the MIXCR software (Bolotin et al., 2015a). Since the Ion Proton sequencing system generates variable-length reads, we allowed for partial alignments and performed two passes of contig assembly. To rescue as many CDR3s as possible, for incomplete TCR CDR3s, we allowed for extension via the V/J genes, since it has been shown to introduce limited errors, because of the very conserved nature of TCRs on both ends (pattern CASS——-EF) (Bolotin et al., 2015a).

6.11.6. Peptide sets and TCR-binding prediction

We downloaded 25,270 human-derived HLA-associated peptides from (Pearson et al., 2016) on 2020-06-07 and 20,961 viral-derived peptides from the Immune Epitope Database (Vita et al., 2019a) on 2020-11-23. Using the ERGO model (Springer et al., 2020) commit version 5c6fc37, cloned from the GitHub repository (https://github.com/louzounlab/ERGO), we applied the long short-term memory model pre-trained on McPAS on selected CDR3. ERGO outputs a probability of TCR recognition score between 0 and 1, where 1 means recognized.

6.11.7. CDR3 sharing and repertoire overlaps

We calculated sharing of individual CDR3 sequences based solely on the amino acid sequence without matching V/J genes and nucleotide sequences. This approach was selected to assess sharing of the final protein product of CDR3s found in the body rather than to look at specific mRNA features. Thus, for public fraction calculations based on unique CDR3aa sequences, we calculated the number of public sequences in an individual's repertoire and divided it by the total number of sequences in the repertoire (results of Figure 6.3I). To assess the clonality of public CDR3aa, we summed the clonal frequencies attributed to individual public sequences (results of Figure 6.3J). We use the Jaccard distance $d_{(J(A,B))}$ as a measure of dissimilarity between two CDR3 sets A and B:

$$d_{J(A,B)} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Here a value of 0 means exact overlap of the CDR3 sets, and 1 means no overlap.

6.11.8. Diversity measurements

We used the inverse Simpson clonality and Shannon entropy as diversity measurements. Inverse Simpson diversity index is calculated by weighted arithmetic mean of each clone abundance (Simpson, 1949), and we implemented this as a function in python. Shannon entropy (Shannon, 1948) is calculated using the scipy.stats python package.

6.11.9. Recombination probability prediction

CDR3aa recombination probability was predicted using the OLGA software (Sethna et al., 2019). We downloaded the command-line tool commit version 4e0bc36 from the repository (https://github.com/statbiophys/OLGA) and selected humanTRB as the alignment database for the predictions.

6.11.10. Number of mismatches

We used IgBLAST (Ye et al., 2013) to align each nucleotide sequence to the annotated germline V, D, and J sequences to calculate the number of mismatches. We downloaded the package from the repository (https://github.com/ncbi/igblast) commit version dfb98f8. We used the human database and specified TCRs. We obtained the aligned sequence for each result and counted the number of mismatches between the aligned sequence and the germline.

6.11.11. Gaussian mixture model

We used the gaussian mixture model from the python GaussianMixture function from the sklearn library to fit each Gaussian Mixture Model. We selected a 2 component Gaussian Mixture Model with a diagonal covariance type. We grouped individuals of the Britanova cohort by age group. For each age group, we fitted a separate Gaussian Mixture Model on the recombination frequency and clonality quantifications. Then, we evaluated the fit of this model in other age groups. This fit was calculated as a log-likelihood of fit for the data to the pre-trained model. We then reported the average log-likelihood for each age-group - model pair in the heatmap in Figure 6.3.

6.11.12. Hierarchical clustering

We used hierarchical clustering with an unweighted pair group method with arithmetic mean agglomerative function for all hierarchical clustering experiments in this study. Visual assessment was used to split each dendrogram into clusters manually. We used the clustermap function from the seaborn python library to plot heatmaps and associated dendrograms.

6.11.13. Expected cumulative frequency

For each CDR3aa sequence, we calculated the sharing percentage and grouped sequences according to the CDR3aa sharing bins. Then, we calculated for each CDR3aa the cumulative repertoire frequency by summing frequencies of all CDR3aa in each bin across all individual repertoires. The average repertoire frequency for each bin was $4.12 \times 10^{-6} \pm 1.2 \times 10^{-6}$. To draw the expected cumulative frequency line, we multiplied this overall frequency by the median number of individuals of each sharing bin. We reported this value as the mean expected cumulative frequency (red dotted line on the plot).

6.11.14. Overlaps by top N most frequent CDR3

We ranked CDR3 in descending order of clonal frequency, and for a growing N, we selected the top N most frequent CDR3 in each repertoire. Then, we calculated the percentage overlap with the disease-specific CDR3 set. Individual percentages were grouped by age group, and for each age group, the standard deviation within the age group is shown on each line plot.

6.11.15. Treemap

We used the treemap function from the squarify python library. The package was given the CDR3 clonal frequencies and a random color palette.

6.11.16. Survival model

We used the survival package in R for plotting the Kaplan-Meier plots and the lifelines packages in python for the CoxPH model. For each Kaplan-Meier plot, we split the group by median as well as 25% and 75% quantiles to attempt to find the best group separation for each CDR3 characteristic. All plots can be seen in Figure B.6 with associated statistical testing. For the CoxPH models, we used the lifelines python library with the option for right-censored data. On each plot, we reported the log (hazard ratios) as well as the bottom and top 95% confidence intervals.

6.11.17. Classification and regression models

The logistic regression and random forest models from the sklearn python library were used to classify neonatal and TDT-dependent CDR3s in Figure 6.8. We used the default parameters for each model: respectively, an L2 penalty with a regularization strength of 1 and the L-BFGS solver for the logistic regression and 100 estimators, a Gini impurity criterion, no max depth, and a minimum number of samples of 2 for the split for the random forest classifier. For the ablation study, each model received the selected combination of features and learned to classify CDR3 into two classes: neonatal or TDT-dependent. During the dataset preparation, two repertories of each type (cord blood and child) were held out. These repertoires comprised the test set of new data. The performance of each iteration of the model given the combination of input features was reported in Figure 6.8, with performances color-coded for visual comparison.

A linear regression model was trained to determine the relative importance of the features. We used the LinearRegression function from the sklearn python library with the following default features: intercept fitting was allowed, and negative coefficients were allowed. This model received as input a binary vector of the presence/absence of the features and learned to predict the performance of either of the two models (logistic regression or random forest) obtained previously for each feature combination. The coefficients attributed to each binary feature presence/absence were used to compare relative importance. Coefficients close to zero meant there was little weight attributed to the model and vice versa. The ranking of features was obtained by ordering the absolute values of the coefficients in descending order.

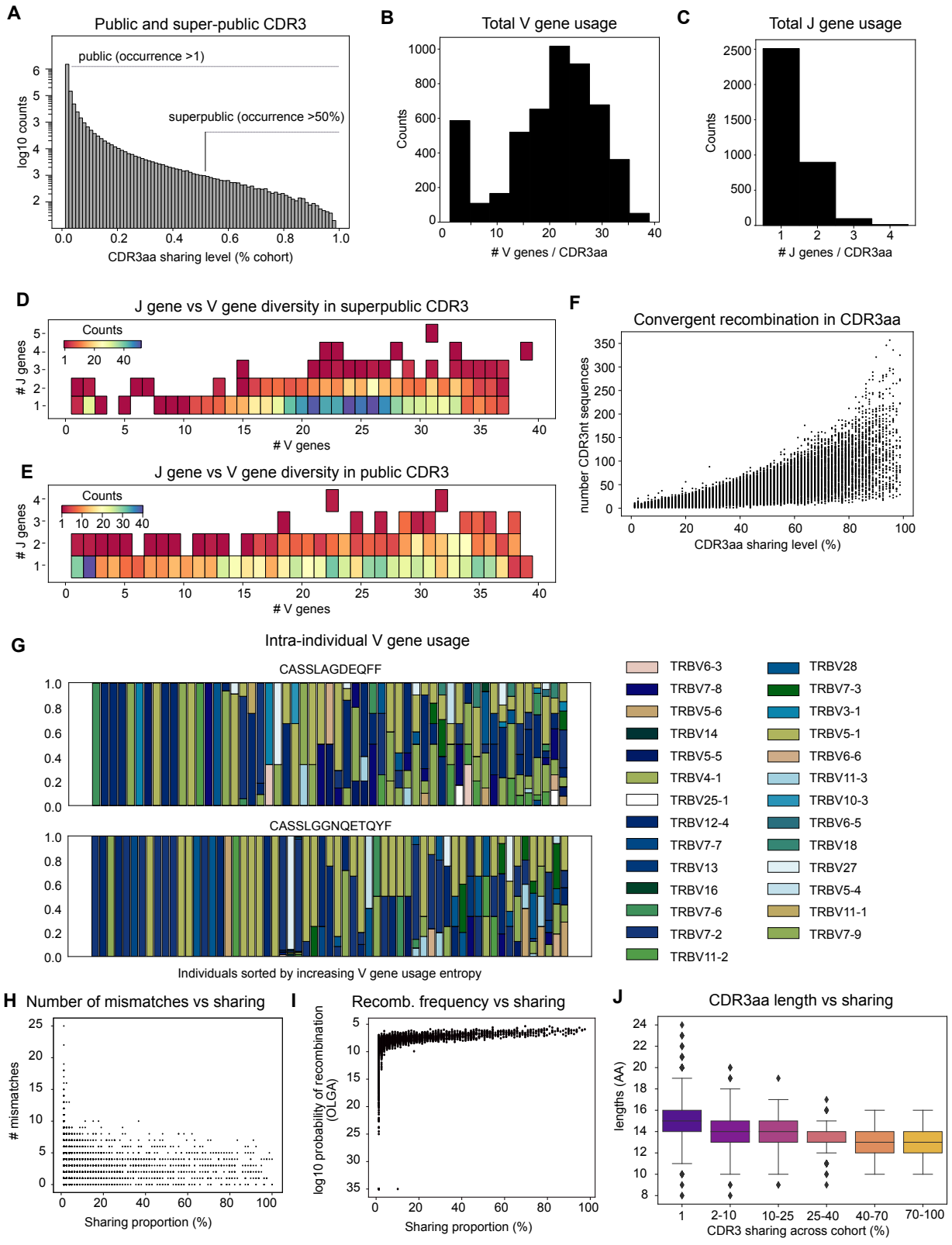


Fig. 6.2. The physical characteristics of public CDR3s.

Fig. 6.2. (A) Public CDR3s are defined as seen in at least two people in the cohort, while superpublic CDR3s are seen in at least half of the cohort. Number of unique (B) V and (C) J genes encoding individual public CDR3aa. Relationship between the number of unique V and J genes shown by 2D histogram in (D) superpublic and (E) public CDR3aa. (F) Scatterplot showing convergent recombination of CDR3 nucleotide sequences in public CDR3aa: highly shared CDR3aa are coded by multiple synonymous nucleotides sequences. (G) Intra-individual V gene usage diversity for two superpublic CDR3aa sequences, sorted by intra-individual entropy: CASSLAGDEQFF and CASSLGGNQETQYF. (H) CDR3aa sharing and number of mismatches to the annotated germline. (I) Relation between the predicted recombination frequency and CDR3aa cohort sharing percentage. (J) CDR3aa length binned by CDR3aa cohort sharing percentage.

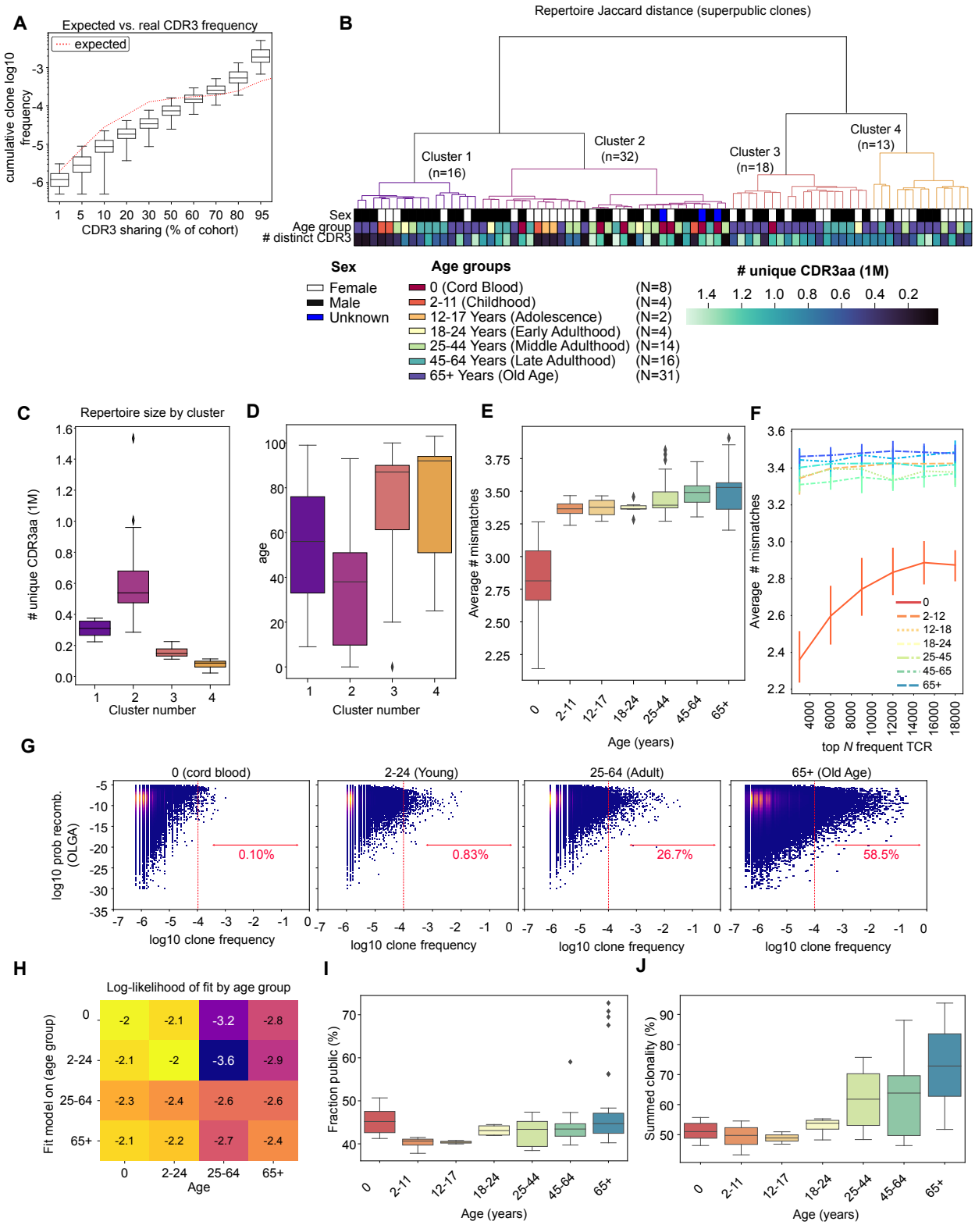


Fig. 6.3. CDR3 sharing between individuals as a function of age

Fig. 6.3. (A) Cumulative log10 frequency of CDR3 binned by cohort sharing percentage. The red dotted line indicates the expected frequency given the number of individuals in each sharing bin. (B) Hierarchical clustering of individual repertoires based on pairwise Jaccard distance. Dendrogram leaves representing individuals are colored by sex, age group, and the number of distinct CDR3aa found in individual repertoires. (C and D) Boxplots show the distribution of the number of distinct public CDR3aa found in individuals and the age of individuals in each of the four clusters from Figure 2B. (E) The median number of mismatches to the germline found in CDR3aa of individuals by age group. (F) The average number of mismatches to the germline found in CDR3aa of individuals grouped by CDR3aa frequency in each repertoire. Colors represent various age groups. (G) Aging correlates with an accumulation of high-frequency clonotypes with a high recombination frequency (as determined by OLGA score). (H) Mean log-likelihood of fit for CDR3aa found in age groups in abscissa for models trained on age groups in ordinate. Each row represents a model trained on the age group, and each column represents the test CDR3aa; each cell contains the mean log-likelihood of fit for a Gaussian mixture model (see Methods). (I) The proportion of public CDR3aa in different age groups. (J) Summed clonality of public CDR3aa in various age groups.

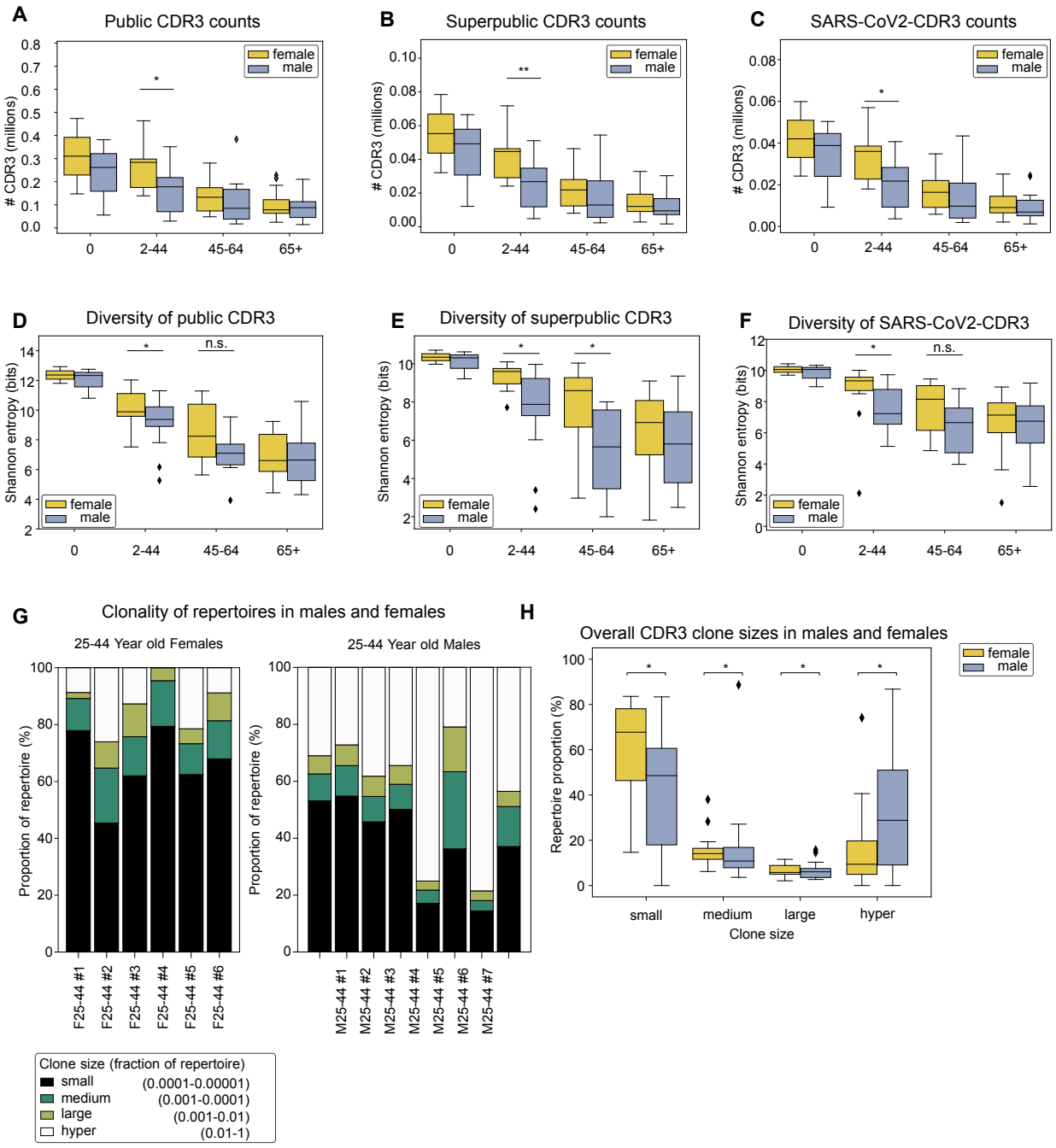


Fig. 6.4. CDR3 sharing between individuals as a function of sex

Fig. 6.4. Absolute numbers of (A) public, (B) superpublic, and (C) SARS-CoV2-specific CDR3aa in male and female individuals by broad age groups. Difference statistically significant ($p<0.05$) for young individuals (ages 2-44). Shannon entropy for (D) public, (E) superpublic, and (F) SARS-CoV2-specific CDR3aa in males and females. Statistically significant differences ($p<0.05$, Mann-Whitney-Wilcoxon) for subjects aged 2-44 and 45-65. (G) Clonality of CDR3aa in repertoires of adult males and females, binned by clone size. (H) Boxplot showing the distribution of overall clone sizes of CDR3aa in males and females of all age groups.

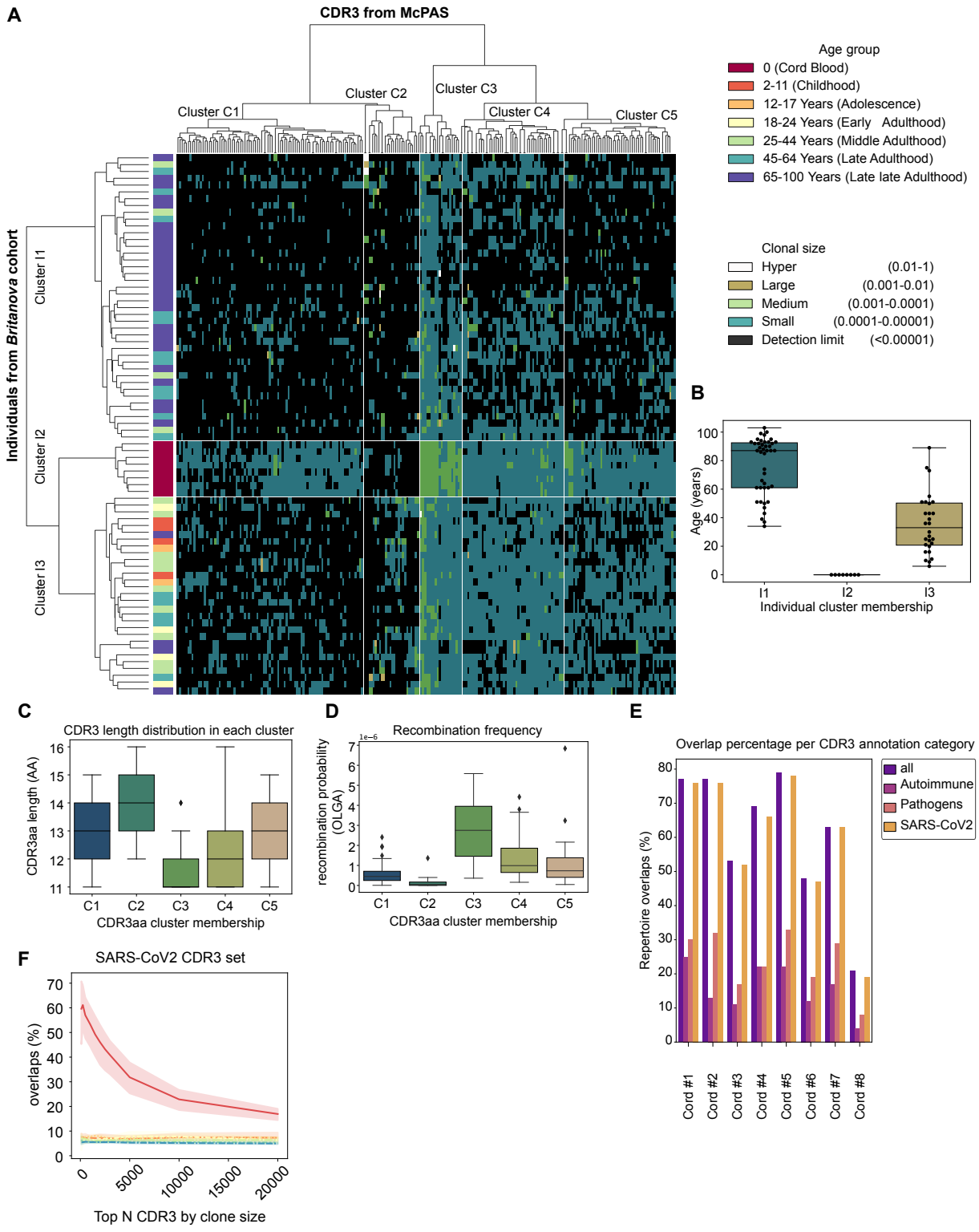


Fig. 6.5. Cord blood samples contain pathology annotated CDR3s.

Fig. 6.5. (A) Heatmap shows, for subjects of the Britanova cohort, the frequency of CDR3aa listed in the McPAS microbial pathogens dataset (Tickotsky et al., 2017). Rows represent individuals, columns unique CDR3aa, and cell color indicates CDR3aa clone size. Row dendrogram leaves are colored by age group. (B) Age distribution for individuals in three individual (Y-axis) clusters from (A). (C) Boxplot showing CDR3 lengths for CDR3 in the five X-axis clusters from (A). (D) Boxplot showing predicted recombination frequency for CDR3 in five X-axis clusters from (B). (E) Barplots show the percentage CDR3aa responsive to autoantigens (Tickotsky et al., 2017), SARS-CoV2 (Nolan et al., 2020), or other pathogens (Tickotsky et al., 2017) in individual cord blood samples from the Britanova cohort. (F) Line plots showing the percentage of SARS-CoV2-specific CDR3aa among the top N most frequent CDR3aa. Line colors and types correspond to age groups as in panel A.

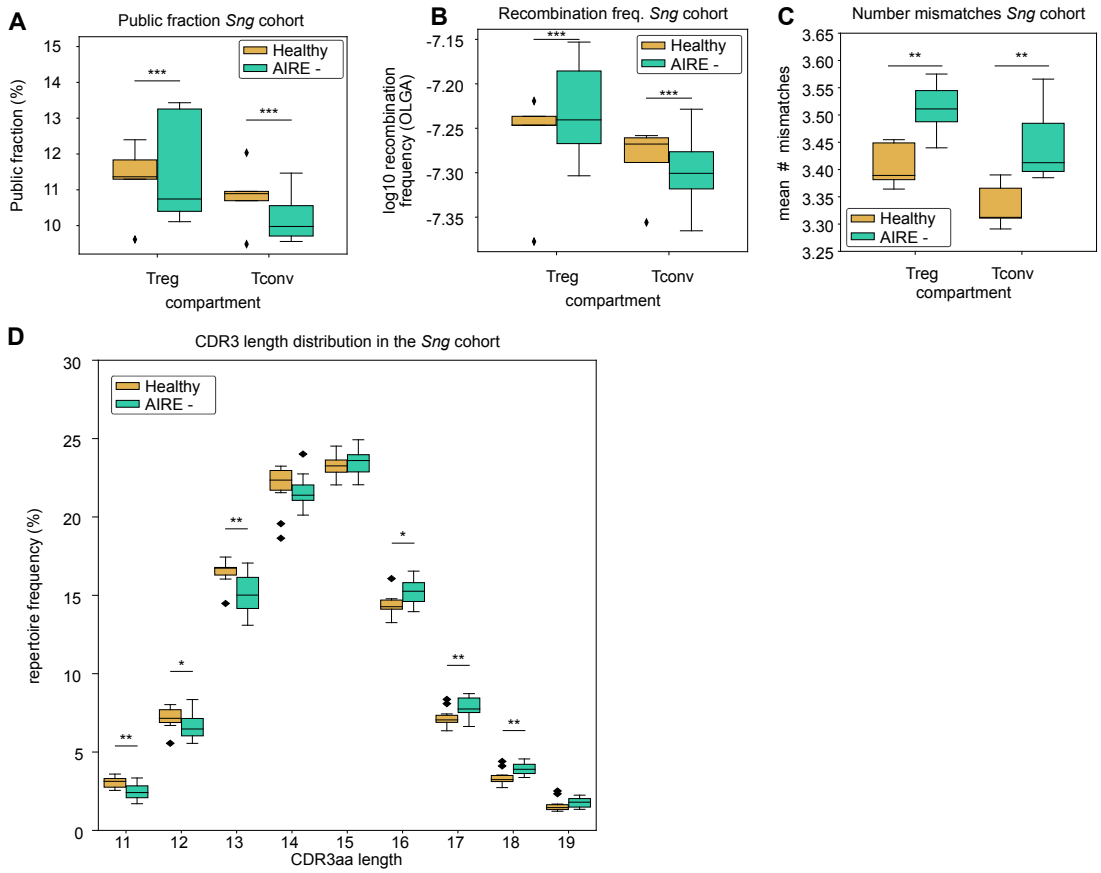


Fig. 6.6. CDR3aa profile in subjects with AIRE mutations

Fig. 6.6. Boxplots showing, for the regulatory (Treg) and conventional T cell (Tconv) compartments, (A) the public fraction, (B) the recombination frequency, and (C) the number of mismatches in subjects with AIRE mutations vs. controls. (D) Boxplots show the CDR3aa length distribution in the Sng cohort. (*p<0.05, ** p< 0.01, ***p<0.001, Mann-Whitney-Wilcoxon).

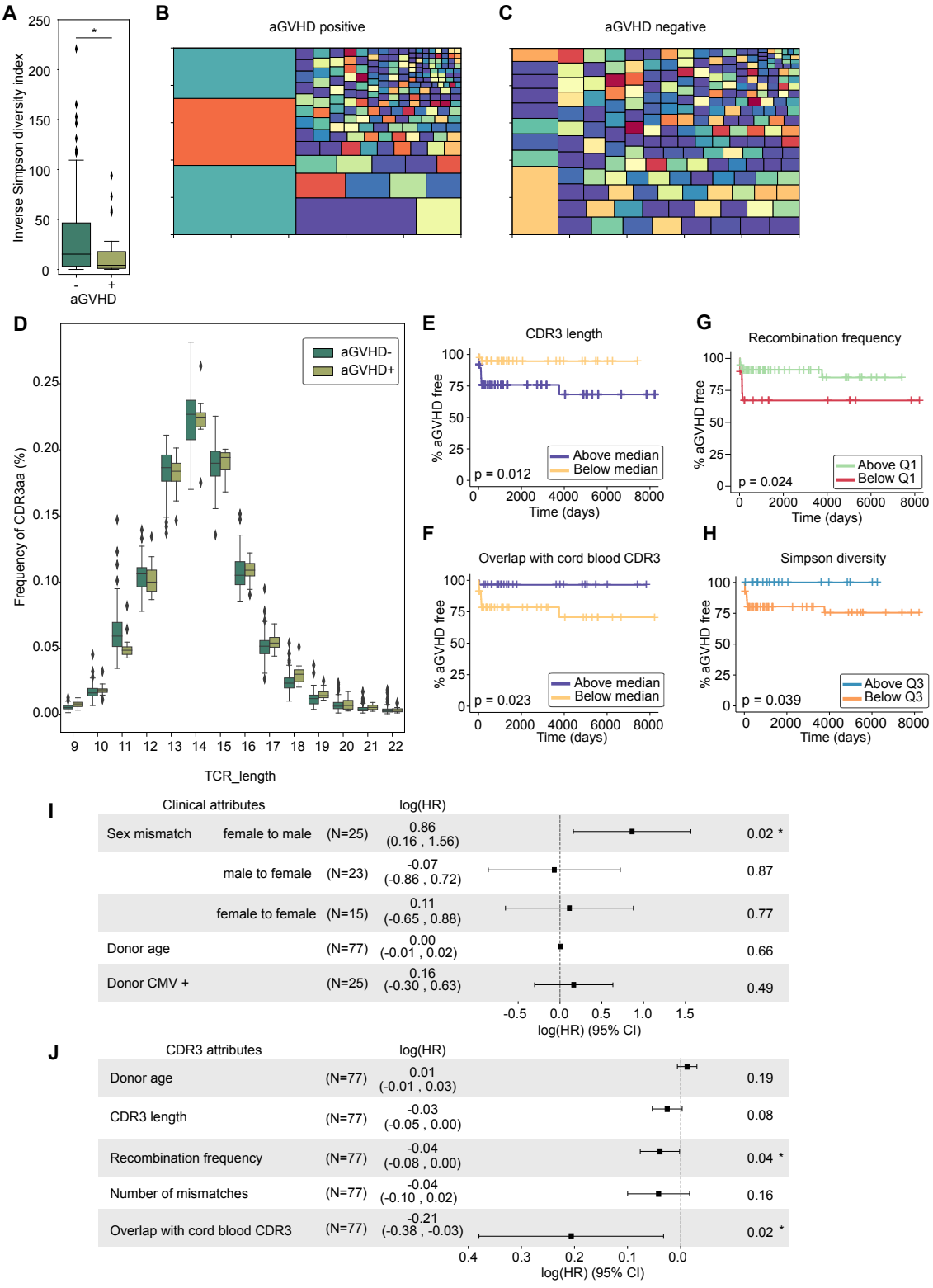


Fig. 6.7. CDR3aa in CD4 T cells from aGVHD+ and aGVHD- AHCT donors

Fig. 6.7. (A) Inverse Simpson diversity index of CDR3 repertoires found in aGVHD+ and aGVHD- grafts. Treemaps showing CDR3aa diversity and clone sizes for two representative donors (B) aGVHD+ and (C) aGVHD-, colors selected at random for better visual distinction. (D) CDR3aa length distribution in aGVHD+ and aGVHD- donors. Kaplan-Meier curves representing aGVHD onset for grafts split into two groups by median or quantiles according to (E) CDR3 length, (F) overlap with cord blood CDR3aa, (G) recombination frequency, and (H) Simpson diversity index. CoxPH models calculating hazard ratios for (I) clinical characteristics and (J) CDR3 repertoire of the donor. (*$p<0.05$, **$p<0.01$, ***$p<0.001$).

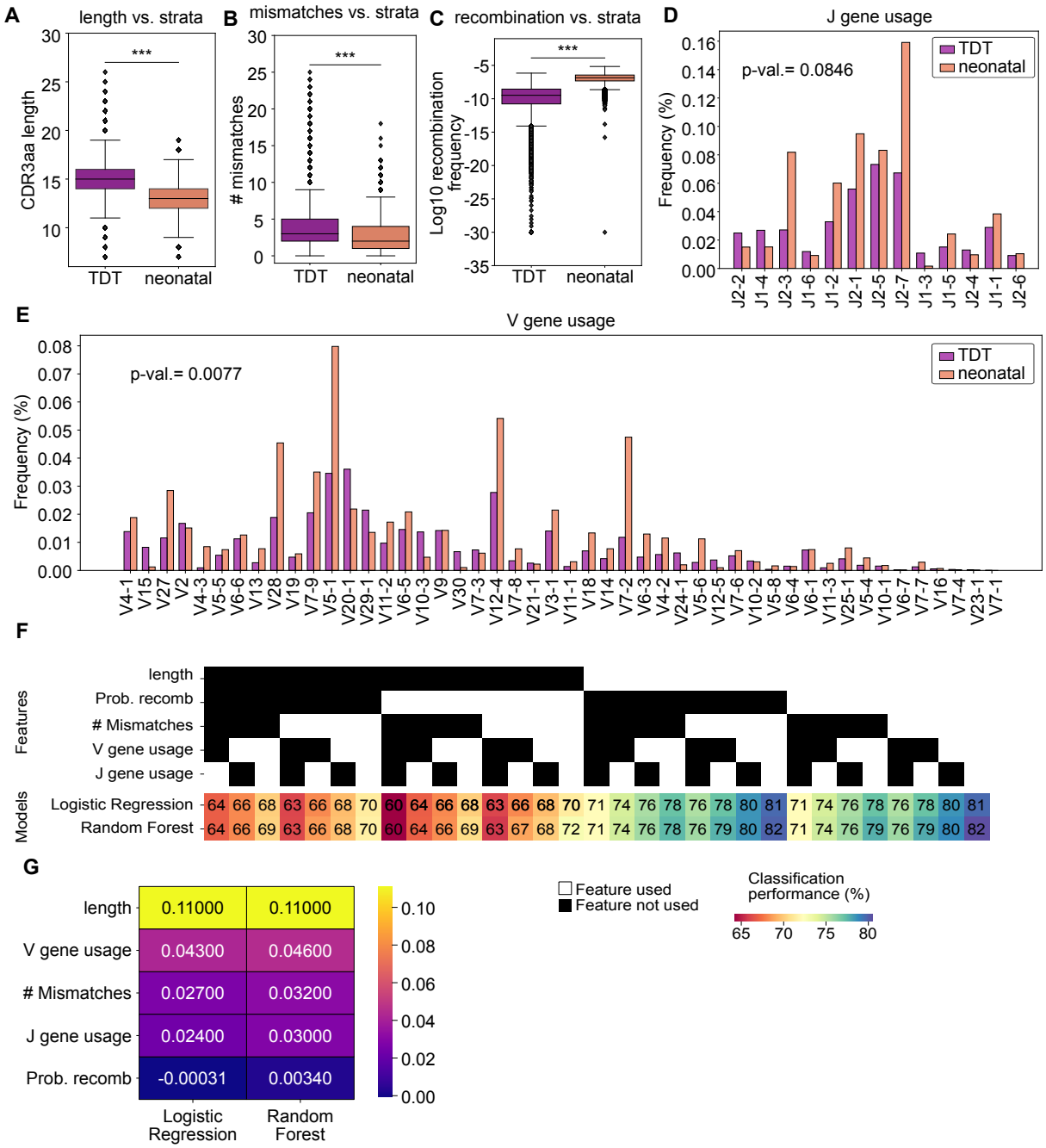


Fig. 6.8. Features of neonatal vs. TDT-dependent TCRs

Fig. 6.8. Boxplots depict (A) the CDR3aa length, (B) the median number of mismatches to the germline, and (C) the median log10 recombination frequency of the TDT-dependent and neonatal strata. (D) J gene and (E) V gene usage frequencies for CDR3aa in TDT-dependent and neonatal strata. (F) Feature ablation study showing classification accuracy on held-out data for each feature combination. Black/ white squares signal exclusion/inclusion of features in the dataset, and the color scheme shows classification performance. (G) Coefficients of the linear model fitted on feature ablation study (see Methods). (*$p<0.05$, **$p<0.01$, ***$p<0.001$, Mann-Whitney-Wilcoxon).

Chapitre 7

Conclusion

Dans cette thèse j'ai développé un modèle basé sur les réseaux de neurones artificiels (ANN), le *Factorized Embeddings* (FE), permettant de construire des atlas cellulaires entièrement basés sur des données de séquençage à haut débit. Pour ce faire, le modèle FE apprend deux espaces d'encodage l'un pour les cellules ou échantillons, et l'autre pour les gènes ou séquences. La nouveauté de l'approche est dans le fait qu'un espace des gènes est également appris, à la différence de la plupart des algorithmes de réduction de dimensionnalité utilisés pour construire des atlas cellulaires (Section 2.2.2). Cette particularité rend le modèle résilient aux données manquantes et permet d'explorer à la fois l'espace d'encodage des patients et celui des gènes.

Dans la première itération du modèle, j'ai observé que le modèle FE apprenait un espace d'encodage des échantillons rendant compte du tissu d'origine de l'échantillon cellulaire le tout basé sur l'expression génique. L'espace d'encodage, ordonné selon les similarités d'expressions de gènes individuels, permet l'interpolation régulière entre les coordonnées des échantillons. Ainsi pour toute coordonnée dans l'espace il est possible de prédire l'expression de tous les gènes du transcriptome, ce qui permet d'étudier l'espace d'encodage appris. Du côté de l'espace d'encodage des gènes, les distances entre des paires de gènes reflétaient leurs niveaux de co-expressions, tandis que l'espace appris reflétait la variance d'expression par groupes de tissus (spécificité au tissu). Ce modèle a permis de montrer qu'il est possible d'apprendre simultanément plusieurs espaces d'encodage pour des données de transcriptomique. De plus, il a établi une base solide validée pour les deux prochaines itérations du modèle.

Dans la deuxième itération du modèle, *The Latent Transcriptome* (TLT), j'ai remplacé la fonction d'encodage de gènes, qui est très dépendante de la méthode de quantification, par une fonction d'encodage de séquences d'ADN. Le passage d'un espace d'encodage de valeurs catégoriques vers des valeurs continues avait pour but d'ajouter au modèle existant la possibilité de faire des liens entre les séquences basé sur leur composition en nucléotides

et la présence de patrons dans les séquences. De plus, ce modèle était développé pour traiter un jeu de données, où des modifications dans la séquence même du gène influence des caractéristiques. Le jeu de données que j'ai choisi, Leucegène (Macrae et al., 2013), un jeu de données de RNA-Seq d'échantillons de leucémie myéloïde aiguë, où des mutations, inversions et translocations de gènes sont importantes pour la classification des échantillons et le modèle TLT devait permettre de détecter ce type de modification génomiques via l'espace d'encodage des k-mers. Pour ce faire, des tables d'abondance de k-mers ont été quantifiées à partir des données brutes de transcriptome, contournant ainsi les procédures d'alignement et quantifications géniques (Section 1.6). Cette itération était limitée par une complexité (O) élevée et donc appliquer le modèle à tous les k-mers du transcriptome se révéla à être un problème non-trivial. Cette itération m'a tout de même permis d'explorer le nouveau modèle dans un cas restreint. En effet, je me suis limitée à quelques régions génomiques ayant des propriétés intéressantes: i) deux gènes quasi-identiques mais dont un n'est présent que dans les personnes de sexe masculin dans la cohorte et ii) deux gènes participant à une translocation dans près de 10% de la cohorte. Je m'attendais à ce que le modèle apprenne un espace d'encodage pour les k-mers basé sur leur séquence et que les k-mers aux séquences similaires soient groupés dans l'espace d'encodage. Je m'attendais également que le modèle conserve des propriétés de la première itération du modèle *Factorized Embeddings*, dans le sens où les k-mers dont l'abondance était similaire dans le transcriptome seraient groupés, tel que je l'ai observé pour les gènes. Cependant, je ne savais pas laquelle des deux propriétés allaient être prioritaire pour le modèle. Le résultat observé était que le modèle optimisait les deux propriétés en même temps; en effet, les k-mers étaient ordonnés dans l'espace en fonction de leur abondance mais également en fonction de leur similarité, permettant ainsi de grouper les k-mers contigus selon leurs exons d'appartenance. Cette particularité a permis de détecter facilement les séquences participant dans la translocation entre les deux gènes, vu qu'ils se retrouvaient à mi-chemin entre les exons respectifs dans l'espace d'encodage. Malgré le problème d'extensibilité (*scalability*) au jeu de données entier, le modèle a tout de même montré des caractéristiques intéressantes et donc j'ai tenté de trouver un jeu de données plus petit pour la troisième itération du modèle.

Finalement, en tentant d'adapter le modèle TLT à un jeu de données de séquences plus petit que le transcriptome, j'ai construit la troisième itération du modèle: *TCRome*. Cette itération est adaptée aux données de séquençage TCR (TCR-Seq) (Section 5.3). Ainsi, le modèle prédit la présence ou non d'un TCR dans le répertoire d'un individu. En examinant l'espace d'encodage des séquences TCR, j'ai observé que comme avec le modèle TLT, les TCR aux séquences similaires étaient groupés ensemble. Cependant, ce modèle ne généralisait pas du tout aux séquences nouvelles. De plus, j'ai observé des patrons intrigants de partage des TCR entre les individus. L'analyse de ces patrons de partage ont mené à la découverte de deux types de récepteurs, présents à diverses proportions selon le sexe et l'âge de l'individu.

Les proportions étaient corrélées à des évènements auto-immuns tels la maladie du greffon contre l'hôte (GVHD) et des mutations dans le gènes AIRE.

Collectivement, ces travaux démontrent que l'architecture du modèle FE se prête bien aux exigences d'un atlas cellulaire. En effet, chaque échantillon est encodé dans espace multidimensionnel, tout en s'appuyant sur les données. Les échantillons voisins sont similaires, selon leur expression génique ainsi que leur appartenance à un tissu. Grâce à la possibilité d'interpoler directement dans l'espace d'échantillon, des décalages entre les personnes ont pu être mesurées, sur la base d'expression de gènes spécifiques. Cette propriété peut se prêter à un contexte différent, par exemple pour la mesure de décalages entre plusieurs échantillons d'une maladie, dont on veut mesurer la progression, tel que suggéré par (Ponting, 2019). Ceci permet de conclure que l'architecture FE est une solution potentielle pour la construction d'atlas cellulaires. Cependant, tout travail a des limitations; les trois itérations du modèle Factorized Embedding sont toutes imparfaites et dans la prochaine section je détaillerai quelques solutions possibles.

7.1. Limitations

Comme tout réseau de neurones artificiel (ANN), le modèle FE a ses limitations. Je vais tout d'abord détailler les limitations générales, dont souffrent les ANN, suivi des limitations propres au modèle FE et ses dérivés.

7.1.1. Limitations générale des ANN: le jeu de données y est pour beaucoup

Tout d'abord, la performance de tout ANN dépend énormément des propriétés du jeu de données et le modèle FE ne fait pas exception. La richesse des données est nécessaire pour garantir un bon apprentissage et dans le cas du modèle FE, dans un jeu de données pauvre, par exemple ne contenant que quelques échantillons ou même que quelques tissus différents, le résultat va en souffrir. Dans un contexte de pauvreté du jeu de données, l'espace d'encodage appris le sera également, ne reflétant que les relations entre les gènes dans ce jeu de données spécifique. Effectivement, les gènes appartenant à la même famille ou ayant une fonction similaire ne seront pas groupés s'ils ne sont pas exprimés dans le jeu de données analysé. Ceci signifie que la performance du modèle FE dépend largement de la richesse du jeu de données.

Toujours en terme de richesse du jeu de données, le jeu de données idéal pour la construction d'atlas cellulaires devra fort probablement contenir plus d'une modalité de quantification. En effet, dans le travail présenté au Chapitre 6, je n'ai effectué que l'analyse des séquences TCR. Or, outre se distinguer par leur séquence TCR, les cellules expriment en surface cellulaire des marqueurs permettant d'identifier entre autres leur sous-type et leur état

d'activation. Dans le cadre de construction d'atlas cellulaires de répertoires immuns, plus d'un type de quantification pourrait être nécessaire, notamment pour grouper les cellules selon d'autres caractéristiques communes, autres que leurs séquences TCR. Une solution possible serait de construire un atlas cellulaire utilisant des données pairées, similaire à ce qu'ont fait (Schattgen et al., 2021). En effet, des transcriptomes à cellule unique donneraient un aperçu de l'expression génique, incluant les marqueurs de surface. Un atlas cellulaire incluant ces deux modalités de quantification bénéficierait grandement, quant à l'organisation des cellules dans l'espace d'encodage.

Pour les modèles TLT et TCRome, aucune information supplémentaire n'a été utilisée, mis à part la séquence. Or, les résultats du Chapitre 6 témoignent de la présence de deux types de séquences mais qui ne sont pas basés uniquement sur un patron dans la séquence de protéines. Effectivement, les deux types de TCR diffèrent sur la base de la longueur, la présence d'insertions et l'utilisation du gène V et J (Chapitre 6). Mis à part la longueur, aucune de ces propriétés de la séquence n'ont pu être analysées par le modèle TCRome. Le travail du Chapitre 6 a mis en valeur que des propriétés de séquences doivent être étudiées au delà de la séquence elle-même. Le modèle TCRome est effectivement agnostique aux notions de recombinaison et donc ne possède pas l'information nécessaire pour quantifier les insertions ni l'utilisation du gène V et J, outre leur similarité de séquence. Ceci signifie que des modifications au modèle original sont nécessaire pour "capturer" ce type de caractéristiques et découvrir les deux sous-types de séquences. Il n'est pas exclus que ce même genre de caractéristiques inaccessibles peuvent être à l'origine d'un nouveau sous-type cellulaire et l'élaboration de modèles profitant au maximum de plusieurs modalités d'information sont nécessaires. Du côté de l'architecture, par exemple, un modèle recevant le gène V et J et le nombre d'insertions, en plus de la séquence pourrait potentiellement intégrer ces informations. Un modèle hybride modélisant chaque séquence de TCR comme un scénario de recombinaison, tel celui proposé par Marcou et al. (Marcou et al., 2018) mais conservant l'encodage des échantillons pourrait également être une solution. Cependant, tout ceci assume bien entendu qu'on connaît d'avance ces caractéristiques. Comme tout modèle ANN, l'utilisation d'information complémentaire est toujours à son avantage.

Finalement, le modèle FE est très dépendant de l'annotation des gènes. En effet, vu que les fichiers d'annotation diffèrent sur la base des gènes qui y sont inclus (Section 1.6), il arrive que deux jeux de données RNA-Seq quantifiés ne puissent pas être fusionnés (Wang et al., 2018; Zhang et al., 2020; Keilwagen et al., 2018). Je parle ici du cas où le modèle est entraîné sur deux cohortes qui ont subi des quantifications différentes. Comme avec la plupart des modèles ANN, FE est sensible aux effets de lot (*batch effect*) et donc deux annotations contenant des gènes différents, ou même basés sur les gènes/transcrits ne peuvent être combinés. Ceci signifie que pour une fusion adéquate des données il faut refaire la quantification ensemble de toutes les données, ce qui peut être problématique dans certains

cas. Des solutions ont cependant été développées pour minimiser les effets de lot et ceux-ci peuvent être intégrés au modèle si le besoin se manifeste (Zhang et al., 2020; Leek, 2014; Risso et al., 2014).

7.1.2. Les limitations propres au modèle FE et ses dérivés

Les principales limitations du modèle FE sont liées à l'architecture et je détaillerai leur nature ci-dessous.

Un des problèmes principaux avec le modèle FE est qu'il apprend un encodage pour les patients et donc dans un contexte de validation et/ou prédiction, il nécessite le ré-entrainement du modèle pour émettre des prédictions pour un nouveau patient. Effectivement, on ne peut simplement utiliser le modèle pour émettre des prédictions, lorsqu'un nouveau patient est présenté, son encodage nécessite d'être ré-entraîné. Si le nouveau patient diffère trop des patients utilisés pour entraîner le réseau, (hors distribution), le modèle n'émettra probablement pas une bonne prédiction. J'ai observé ceci à petite échelle avec deux patients dans le jeu de données de GTEx, où le modèle mettait de côté tous les échantillons provenant de ces patients. J'ai ensuite découvert que ceci était dû à un patron d'expression génique particulier qui correspondait à un problème technique, confirmé par d'autres équipes, où les échantillons biologiques se sont dégradés avant le séquençage (Gallego Romero et al., 2014). La limitation principale en terme de prédiction est donc qu'il est nécessaire de ré-entraîner le modèle pour prédire l'expression génique d'un nouveau patient.

Ensuite, lorsque j'ai voulu entraîner le modèle TLT sur le transcriptome complet transformé en table d'abondance de k-mers, j'ai heurté un problème, pas du tout observé pour le modèle FE. Chacun des 149 patients de la cohorte Leucegene (Macrae et al., 2013) avait sa propre table de 1×10^{10} à 3×10^{10} de k-mers et donc pour une itération d'entraînement, le modèle devait traiter entre 1.5×10^{12} et 4.5×10^{12} k-mers. Le modèle TLT était impossible à entraîner dans un temps raisonnable, vu la taille du jeu de données. Plusieurs changements étaient possibles:

- changer la taille du k-mer
- réduire le jeu de données d'entraînement et limiter le nombre de k-mers

Un k-mer de plus petite taille mène à la réduction exponentielle du jeu de données (Figure 1.2). Cependant, un k-mer court sera ambigu dans sa région génomique de provenance. Ainsi pour des applications où il est nécessaire de connaître avec précision la région génomique d'origine du k-mer, par exemple dans des expériences de recherche de cible pour l'immunothérapie du cancer (Laumont et al., 2018), réduire la taille du k-mer peut être désavantageuse. Aussi, dans le contexte de détection d'altérations génomiques, telles les insertions, délétions, translocations et répétitions en tandem (*tandem repeats*), si la longueur

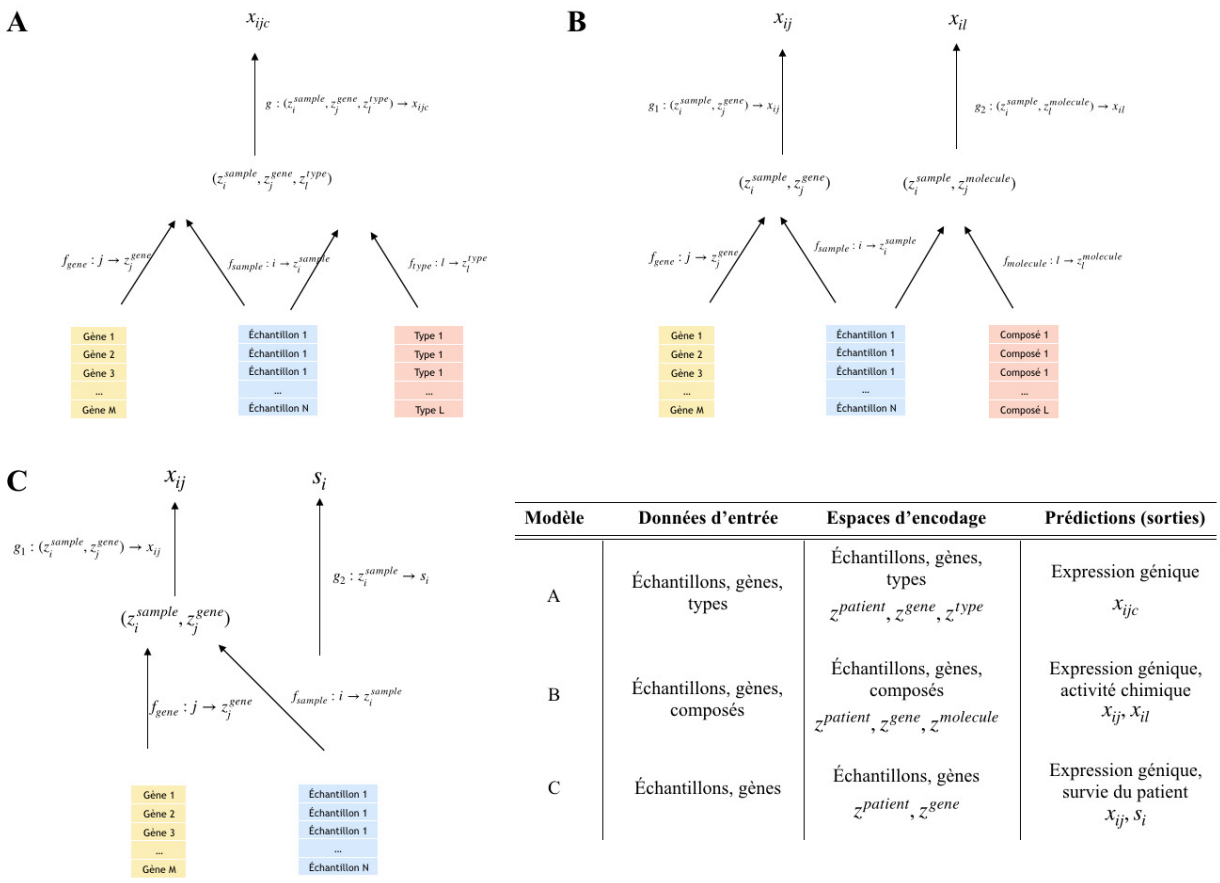
du k-mer est plus petite que l'altération, sa détection sera problématique. L'autre possibilité serait de réduire le jeu de données d'entraînement en réduisant le nombre de k-mers, profitant du fait que les k-mers se chevauchent et donc sont redondants en information. Par exemple, l'utilisation d'une méthode de hashage tel *min-hash* (Ondov et al., 2016). Cette option rejoint presque le concept derrière le développement des puces à ADN (*microarrays*). En effet, la puce à ADN quantifie l'expression génique en mesurant l'abondance d'une région spécifique dans un gène (les sondes). On peut donc imaginer la sélection de quelques k-mers en tant que "sondes" et donc la réduction massive du jeu de données. Ceci enlève par contre quelques propriétés de l'espace d'encodage des k-mers, notamment dans le cadre de détection de translocations ou modifications géniques, vu que désormais le jeu de données ne sera limité qu'à quelques k-mers.

7.2. Travaux futurs

Le modèle FE est une architecture adaptable aux particularités du jeu de données et offre une base pour le développement de modèles utilisant et intégrant des données multi-omiques. Je détaillerai dans les prochaines section quelques-unes des avenues qui sont les plus prometteuses.

7.2.1. Factorized Embeddings pour les données manquantes

L'un des points forts de l'architecture FE est que les données sont lues par le modèle par paire (gene, échantillon) et donc le modèle peut tout de même apprendre dans des contextes où les données sont partielles ou manquantes. Par exemple, dans le contexte de séquençage à cellule unique le phénomène de *drop-out* cause des problèmes pour l'analyse de données. Ce phénomène est dû principalement au manque de profondeur de séquençage dont la cause est la petite quantité de matériel génétique présent dans une seule cellule, le modèle FE pourrait tout de même bien fonctionner. En effet, près de 80% des gènes ne sont pas quantifiés dans les séquençages à cellule unique et donc les stratégies d'analyse implémentent typiquement une étape d'imputation d'expression génique, avant les algorithmes de réduction de dimensionnalité (Lopez et al., 2018; Van Dijk et al., 2018). Le séquençage à cellule unique offre une avenue de quantification très alléchante pour la construction d'atlas cellulaires, vu la richesse des données obtenues, capturant la variabilité inter-cellulaire dans un même tissu. De par sa versatilité, le modèle FE pourrait être adapté pour l'analyse de ce type de séquençage.



Modèle	Données d'entrée	Espaces d'encodage	Prédictions (sorties)
A	Échantillons, gènes, types	Échantillons, gènes, types $z^{patient}, z^{gene}, z^{type}$	Expression génique x_{ijc}
B	Échantillons, gènes, composés	Échantillons, gènes, composés $z^{patient}, z^{gene}, z^{molecule}$	Expression génique, activité chimique x_{ij}, x_{il}
C	Échantillons, gènes	Échantillons, gènes $z^{patient}, z^{gene}$	Expression génique, survie du patient x_{ij}, s_i

Fig. 7.1. Exemples d'architectures possibles pour l'intégration de données multi-omiques dans le modèle Factorized Embeddings

7.2.2. Factorized Embeddings pour l'intégration de données multi-omiques

L'autre avantage de l'architecture du modèle FE est que le nombre de données d'entrées et sorties peut varier. En effet, FE peut être modifié pour émettre deux ou plus de prédictions et/ou apprendre plus de deux espaces. Ceci signifie qu'il est tout à fait concevable d'avoir un modèle FE modifié apprenant et intégrant des données multi-omiques.

Par exemple, le modèle pourrait avoir une seule sortie mais avoir trois entrées au modèle (Figure 7.1A). Un exemple exploitant ce type d'itération du modèle serait d'ajouter des données catégoriques (par exemple des données cliniques) spécifiques au patient - type de maladie par exemple. Ceci offrirait la possibilité d'apprendre un espace de patients et gènes en fonction des catégories et donc d'émettre des prédictions sur l'expression génique du patient, s'il appartenait à une autre classe (un autre type de maladie dont le patient ne fait pas partie).

Dans un autre exemple, le modèle pourrait apprendre 3 espaces d'encodage, pour les gènes, les patients et les composés chimiques. Dans cet exemple, le nombre de sorties (prédictions) du modèle est de 2, où le modèle prédit à la fois l'expression génique dans un échantillon cellulaire, ainsi que l'activité chimique d'un composé dans le même échantillon (Figure 7.1B). Cette itération du modèle exploiterait également sa résilience aux données manquantes (Section 7.2.1) et permettrait d'émettre des prédictions à la fois sur l'expression génique ainsi que des activités chimiques manquantes dans le jeu de données. Finalement, il est possible de modifier le modèle FE pour ajouter des modules d'ANN performant des fonctions spécifiques. Par exemple, le modèle *DeepSurv* de Katzman et collègues, est un ANN implémentant une généralisation du modèle de survie de Cox (CoxPH, modèle à risque proportionnel de Cox, de l'anglais *Cox Proportionnal Hasards*) (Katzman et al., 2016). Une itération de l'architecture de FE pourrait être construite, où le modèle *DeepSurv* pourrait être un module contenu dans le modèle FE, et qui utiliserait les coordonnées d'encodage de patients en tant que caractéristiques d'entrée afin de prédire la survie des patients. Ce modèle apprendrait donc deux espaces d'encodage: celui des gènes et celui des échantillons mais tentera de simultanément prédire l'expression génique ainsi que la survie d'un patient (Figure 7.1C). Tous ces exemples montrent la polyvalence du modèle FE (Figure 7.1D), dû principalement au fait que c'est un réseau de neurones.

7.2.3. Le modèle TCRome pour apprendre des individus

Le modèle TCRome pourrait être amélioré afin de palier aux problèmes décrits dans la Section 7.1.2. En effet, en remplaçant la fonction d'encodage du TCR par un modèle probabiliste modélisant les caractéristiques de recombinaison, tel que celui décrit dans (Sethna et al., 2020; Marcou et al., 2018), ceci permettrait d'ajouter de la précision dans le groupement et donc récupérer les particularités des TCR décrites au Chapitre 6. De plus, une nouvelle technologie de séquençage récemment développée permet de séquencer à la fois l'expression génique, le TCR et même de mesurer la réactivité de la cellule T à des peptides spécifiques (Zemmour et al., 2018; Valkiers et al., 2022). Schattgen et al. ont déjà exploré cette idée, montrant que cette méthode permettait de regrouper les cellules T non seulement en fonction de leur séquence TCR mais également de leur expression génique (Schattgen et al., 2021). Ce type de jeu de données sera idéal pour la création d'atlas cellulaires de répertoires immuns, car il regrouperait une énorme quantité d'information, tant du point de vue séquence du TCR, que d'expression génique et réactivité du TCR à des peptides, et donc permettra une meilleure analyse et compréhension de l'organisation des répertoires de cellules T.

Finalement, un moyen de palier au problème de la nécessité de ré-entraînement du modèle avec l'inclusion d'un nouveau patient, serait de remplacer l'encodage du patient par un

encodage basé sur un ensemble de caractéristiques communes à plusieurs patients. Par exemple, dans l'optique de l'apprentissage des atlas de cellules T dans divers individus, l'index symbolisant l'individu pourrait être remplacé par un encodage des divers allèles MHC (Section 5) que cet individu présente. Ce nouvel atlas apprendrait donc un espace d'encodage des molécules MHC et chaque individu serait représenté par l'ensemble de ces molécules encodées. Ceci permettrait donc à un modèle entraîné d'émettre des prédictions sur des patients qu'il n'a jamais vus, basé sur les molécules MHC de l'individu.

7.3. Le mot de la fin

Les avancées en séquençage à haut débit et autres méthodes "omique" ont transformé la biologie des systèmes. Ces nouvelles technologies approfondissent notre compréhension des cellules humaines, de leurs états moléculaires dans toute leur diversité. Le travail sur la méthode computationelle présentée dans cette thèse ne fait qu'effleurer la surface du potentiel des réseaux de neurones artificiels dans la construction d'atlas cellulaires entièrement basés sur les données. La mise en place de l'initiative du Human Cell Atlas (Lindeboom et al., 2021; Rozenblatt-Rosen et al., 2017; Regev et al., 2017) en 2016 a déjà mené à plus de 400 publications, ce qui témoigne de l'intérêt de ce projet. Regroupant plus de 2200 personnes à travers 75 pays différents, l'initiative c'est donné pour mission l'acquisition de données exhaustive et ouverte. Les domaines d'impact suite à la construction de cet atlas seront grands: une meilleure compréhension des maladies, le développement du médicament avec moins de toxicité, des meilleurs résultats en biologie synthétique. Tout cela met en évidence la nécessité de développer des méthodes computationnelles, permettant d'intégrer l'ensemble des données et de les rendre interprétables.

Références bibliographiques

H. A. Alturkistani, F. M. Tashkandi, et Z. M. Mohammedsaleh. Histological Stains: A Literature Review and Case Study. *Glob. J. Health Sci.*, 8(3):72–79, 2015.

D. Arendt. The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.*, 9(11):868–882, 2008.

E. Asgari et M. R. K. Mofrad. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS One*, 10(11):141287, 2015.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, et G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

E. O. Audemard, P. Gendron, V.-P. Lavallée, J. Hébert, G. Sauvageau, et S. Lemieux. Targeted variant detection in leukemia using unaligned RNA-Seq reads. *bioRxiv*, page 295808, 2018.

C. Baron, R. Somogyi, L. D. Greller, V. Rineau, P. Wilkinson, C. R. Cho, M. J. Cameron, D. J. Kelvin, P. Chagnon, D.-C. Roy, L. Busque, R.-P. Sékaly, et C. Perreault. Prediction of graft-versus-host disease in humans by donor gene-expression profiling. *PLoS Med.*, 4 (1):e23, 2007.

S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, et D. Pe'er. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, 157(3):714–725, 2014.

Y. Benjamini et T. P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, 40(10):e72, 2012.

J. Bergstra et Y. Bengio. Random search for Hyper-Parameter optimization. *J. Mach. Learn. Res.*, 13(10):281–305, 2012.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, et P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, et S. Canaider. An estimation

of the number of cells in the human body. *Ann. Hum. Biol.*, 40(6):463–471, 2013.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Aug. 2006.

J. N. Böhm, P. Berens, et D. Kobak. Attraction-Repulsion spectrum in neighbor embeddings. July 2020.

D. A. Bolotin, S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva, et D. M. Chudakov. MiXCR: Software for comprehensive adaptive immunity profiling, 2015a.

D. A. Bolotin, S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva, et D. M. Chudakov. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*, 12(5):380–381, May 2015b.

A. Bordes, J. Weston, et Y. Bengio. Learning Structured Embeddings of Knowledge Bases. Technical report.

X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Sepah, E. Raff, K. Madan, V. Voleti, S. E. Kahou, V. Michalski, D. Serdyuk, T. Arbel, C. Pal, G. Varoquaux, et P. Vincent. Accounting for variance in machine learning benchmarks. Mar. 2021.

N. L. Bray, H. Pimentel, P. Melsted, et L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, 2016.

O. V. Britanova, M. Shugay, E. M. Merzlyak, D. B. Staroverov, E. V. Putintseva, M. A. Turchaninova, I. Z. Mamedov, M. V. Pogorelyy, D. A. Bolotin, M. Izraelson, A. N. Davydov, E. S. Egorov, S. A. Kasatskaya, D. V. Rebrikov, S. Lukyanov, et D. M. Chudakov. Dynamics of individual T cell repertoires: From cord blood to centenarians. *J. Immunol.*, 196(12):5005–5013, June 2016.

P. Brodin et M. M. Davis. Human immune system variation. *Nat. Rev. Immunol.*, 17(1): 21–29, Jan. 2017.

P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, et R. L. Mercer. Class-Based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–480, 1992.

J.-P. Brunet, P. Tamayo, T. R. Golub, et J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.*, 101(12):4164–4169, Mar. 2004.

Ø. Bruserud, B. E. Oftedal, A. B. Wolff, et E. S. Husebye. AIRE-mutations and autoimmune disease. *Curr. Opin. Immunol.*, 43:8–15, Dec. 2016.

M. Chatzou, C. Magis, J.-M. Chang, C. Kemena, G. Bussotti, I. Erb, et C. Notredame. Multiple sequence alignment modeling: methods and applications. *Brief. Bioinform.*, 17 (6):1009–1023, 2016.

X. Chen et H. Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6): 323–329, June 2012.

D. Chicco. Ten quick tips for machine learning in computational biology. *BioData Min.*, 10: 35, Dec. 2017.

J. Chiffelle, R. Genolet, M. A. Perez, G. Coukos, V. Zoete, et A. Harari. T-cell repertoire analysis and metrics of diversity and clonality. *Curr. Opin. Biotechnol.*, 65:284–295, Oct. 2020.

R. Chikhi et P. Medvedev. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37, Jan. 2014.

C. T. Choy, C. H. Wong, et S. L. Chan. Embedding of Genes Using Cancer Gene Expression Data: Biological Relevance and Potential Application on Biomarker Discovery. *Front. Genet.*, 9:682, 2019.

N. D. Chu, H. S. Bi, R. O. Emerson, A. M. Sherwood, M. E. Birnbaum, H. S. Robins, et E. J. Alm. Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors. *BMC Immunol.*, 20(1):19, June 2019.

E. B. Chuong. The placenta goes viral: Retroviruses control gene expression in pregnancy. *PLoS Biol.*, 16(10):e3000028, 2018.

K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, et E. W. Sayers. GenBank. *Nucleic Acids Res.*, 44(D1):67–72, 2016.

E. Clave, I. L. Araujo, C. Alanio, E. Patin, J. Bergstedt, A. Urrutia, S. Lopez-Lastra, Y. Li, B. Charbit, C. R. MacPherson, M. Hasan, B. L. Melo-Lima, C. Douay, N. Saut, M. Germain, D.-A. Trégouët, P.-E. Morange, M. Fontes, D. Duffy, J. P. Di Santo, L. Quintana-Murci, M. L. Albert, A. Toubert, et Milieu Intérieur Consortium. Human thymopoiesis is influenced by a common genetic variant within the TCRA-TCRD locus. *Sci. Transl. Med.*, 10(457), Sept. 2018.

H. Clevers. What is your conceptual definition of "cell type" in the context of a mature organism? *Cell Syst*, 4(3):255–259, Mar. 2017.

S. Cohen, J. Roy, S. Lachance, J.-S. Delisle, A. Marinier, L. Busque, D.-C. Roy, F. Barabé, I. Ahmad, N. Bambace, L. Bernard, T. Kiss, P. Bouchard, P. Caudrelier, S. Landais, F. Larochelle, J. Chagraoui, B. Lehnertz, S. Corneau, E. Tomellini, J. J. A. van Kampen, J. J. Cornelissen, M. Dumont-Lagacé, M. Tanguay, Q. Li, S. Lemieux, P. W. Zandstra, et G. Sauvageau. Hematopoietic stem cell transplantation using single UM171-expanded cord blood: a single-arm, phase 1-2 safety and feasibility study. *Lancet Haematol*, 7(2): e134–e145, Feb. 2020.

R. W. Dapson et R. W. Horobin. Dyes from a twenty-first century perspective. *Biotech. Histochem.*, 84(4):135–137, Aug. 2009.

P. Dash, A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. O. Nguyen, K. Kedzierska, N. L. La Gruta, P. Bradley, et P. G. Thomas. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661):89–93, June 2017.

M. P. Davenport, N. L. Smith, et B. D. Rudd. Building a T cell compartment: how immune cell development shapes function. *Nat. Rev. Immunol.*, 20(8):499–506, June 2020.

K. Davidsen, B. J. Olson, W. S. DeWitt, 3rd, J. Feng, E. Harkins, P. Bradley, et F. A. Matsen, 4th. Deep generative models for T cell receptor protein sequences. *Elife*, 8, Sept. 2019.

P. C. de Greef, T. Oakes, B. Gerritsen, M. Ismail, J. M. Heather, R. Hermsen, B. Chain, et R. J. de Boer. The naive t-cell receptor repertoire has an extremely broad distribution of clone sizes. *Elife*, 9, Mar. 2020.

M. De Simone, G. Rossetti, et M. Pagani. Single cell T cell receptor sequencing: Techniques and future challenges. *Front. Immunol.*, 9:1638, July 2018.

H. de Thé, C. Lavau, A. Marchio, C. Chomienne, L. Degos, et A. Dejean. The PML-RAR alpha fusion mRNA generated by the t(15;17) translocation in acute promyelocytic leukemia encodes a functionally altered RAR. *Cell*, 66(4):675–684, 1991.

D. de Verteuil, T. L. Muratore-Schroeder, D. P. Granados, M.-H. Fortier, M.-P. Hardy, A. Bramoullé, E. Caron, K. Vincent, S. Mader, S. Lemieux, P. Thibault, et C. Perreault. Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules. *Mol. Cell. Proteomics*, 9(9):2034–2047, 2010.

M. R. Deibel, Jr, L. K. Riley, M. S. Coleman, M. L. Cibull, S. A. Fuller, et E. Todd. Expression of terminal deoxynucleotidyl transferase in human thymus during ontogeny and development. *J. Immunol.*, 131(1):195–200, July 1983.

H. Deng et G. Runger. Feature Selection via Regularized Trees. Technical report.

W. S. DeWitt, 3rd, A. Smith, G. Schoch, J. A. Hansen, F. A. Matsen, 4th, et P. Bradley. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife*, 7, Aug. 2018.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, et T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan. 2013.

A. Drouin, S. Giguère, M. Déraspe, M. Marchand, M. Tyers, V. G. Loo, A.-M. Bourgault, F. Laviolette, et J. Corbeil. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, 17(1):754, Sept. 2016.

J. Du, P. Jia, Y. Dai, C. Tao, Z. Zhao, et D. Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(S1):82, 2019.

M. Egerton, R. Scollay, et K. Shortman. Kinetics of mature t-cell development in the thymus. *Proc. Natl. Acad. Sci. U. S. A.*, 87(7):2579–2582, Apr. 1990.

P. Ehrlich. Address in Pathology, ON CHEMIOTHERAPY: Delivered before the Seventeenth International Congress of Medicine. *Br. Med. J.*, 2(2746):353–359, 1913.

Y. Elhanati, Z. Sethna, C. G. Callan, Jr, T. Mora, et A. M. Walczak. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol. Rev.*, 284(1):167–179, July 2018.

R. O. Emerson, W. S. DeWitt, M. Vignali, J. Gravley, J. K. Hu, E. J. Osborne, C. Desmarais, M. Klinger, C. S. Carlson, J. A. Hansen, M. Rieder, et H. S. Robins. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.*, 49(5):659–665, May 2017.

B. Ernst, D. S. Lee, J. M. Chang, J. Sprent, et C. D. Surh. The peptide ligands mediating positive selection in the thymus control T cell survival and homeostatic proliferation in the periphery. *Immunity*, 11(2):173–181, Aug. 1999.

B. Ewing et P. Green. Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Res.*, 8(3):186–194, Mar. 1998.

E. J. Fox, K. S. Reid-Bayliss, M. J. Emond, et L. A. Loeb. Accuracy of next generation sequencing platforms. *Next Gener Seq Appl*, 1, 2014.

K. Fukushima. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, 36(4):193–202, 1980.

I. Gallego Romero, A. A. Pai, J. Tung, et Y. Gilad. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.*, 12:42, May 2014.

J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, et N. Schultz. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.*, 6(269):l1–pl1, 2013.

P. Gautam, T. Yu, et Y.-H. Loh. Regulation of ERVs in pluripotent stem cells and reprogramming. *Curr. Opin. Genet. Dev.*, 46:194–201, 2017.

M. Gerstung, A. Pellagatti, L. Malcovati, A. Giagounidis, M. G. D. Porta, M. Jädersten, H. Dolatshad, A. Verma, N. C. P. Cross, P. Vyas, S. Killick, E. Hellström-Lindberg, M. Cazzola, E. Papaemmanuil, P. J. Campbell, et J. Boultwood. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.*, 6(1):5901, 2015.

F. D. Gibbons et F. P. Roth. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.*, 12(10):1574–1581, Oct. 2002.

A. Gil, L. Kamga, R. Chirravuri-Venkata, N. Aslan, F. Clark, D. Ghersi, K. Luzuriaga, et L. K. Selin. Epstein-Barr virus Epitope-Major histocompatibility complex interaction combined with convergent recombination drives selection of diverse T cell receptor α and β repertoires. *MBio*, 11(2), Mar. 2020.

C. Gini. On the measure of concentration with espacial reference to income and wealth. *Cowles Commission*, 2(3), 1936.

J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. L. Arlehamn, A. Sette, S. D. Boyd, T. J. Scriba, O. M. Martinez, et M. M. Davis. Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547(7661):94–98, July 2017.

B. Glick, T. S. Chang, et R. G. Jaap. The bursa of fabricius and antibody production. *Poult. Sci.*, 35(1):224–225, Jan. 1956.

G. Glusman, L. Rowen, I. Lee, C. Boysen, J. C. Roach, A. F. Smit, K. Wang, B. F. Koop, et L. Hood. Comparative genomics of the human and mouse T cell receptor loci. *Immunity*, 15(3):337–349, Sept. 2001.

M. Goldman, B. Craft, M. Hastie, K. Repečka, A. Kamath, F. McDade, D. Rogers, A. N. Brooks, J. Zhu, et D. Haussler. The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv*, page 326470, 2019.

M. Gönen. Statistical aspects of gene signatures and molecular targets. *Gastrointest. Cancer Res.*, 3(2 Suppl):19–21, 2009.

F. F. Gonzalez-Galarza, A. McCabe, E. J. M. D. Santos, J. Jones, L. Takeshita, N. D. Ortega-Rivera, G. M. D. Cid-Pavon, K. Ramsbottom, G. Ghattaoraya, A. Alfirevic, D. Middleton, et A. R. Jones. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.*, 48(D1): D783–D788, Jan. 2020.

I. Goodfellow, Y. Bengio, et A. Courville. *Deep Learning*. MIT Press, Nov. 2016.

J. J. Goronzy et C. M. Weyand. Mechanisms underlying T cell ageing. *Nat. Rev. Immunol.*, 19(9):573–583, Sept. 2019.

J. C. Gower et G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 18(1):54–64, 1969.

D. P. Granados, P.-L. Tanguay, M.-P. Hardy, E. Caron, D. de Verteuil, S. Meloche, et C. Perreault. ER stress affects processing of MHC class I-associated peptides. *BMC Immunol.*, 10(1):10, 2009.

R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185 (4154):862–864, Sept. 1974.

S. M. M. Haeryfar, H. D. Hickman, K. R. Irvine, D. C. Tscharke, J. R. Bennink, et J. W. Yewdell. Terminal deoxynucleotidyl transferase establishes and broadens antiviral CD8+ T cell immunodominance hierarchies. *J. Immunol.*, 181(1):649–659, July 2008.

X. Han, R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye, D. Huang, Y. Xu, W. Huang, M. Jiang, X. Jiang, J. Mao, Y. Chen, C. Lu, J. Xie, Q. Fang, Y. Wang, R. Yue, T. Li, H. Huang, S. H. Orkin, G.-C. Yuan, M. Chen, et G. Guo. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 172(5):1091–1107, 2018.

Z. S. Harris. Distributional Structure. *Distributional Structure, WORD*, 10(3):146–162, 1954.

R. S. Hegde et E. Zavodszky. Recognition and Degradation of Mislocalized Proteins in Health and Disease. *Cold Spring Harb. Perspect. Biol.*, page a033902, 2019.

S. Henikoff et J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, 89(22):10915–10919, Nov. 1992.

S. Hochreiter et J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

K. A. Hogquist et S. C. Jameson. The self-obsession of T cells: how TCR signaling thresholds affect fate 'decisions' and effector function. *Nat. Immunol.*, 15(9):815–823, Sept. 2014.

Y. Hu, T. Hase, H. P. Li, S. Prabhakar, H. Kitano, S. K. Ng, S. Ghosh, et L. J. K. Wee. A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. *BMC Genomics*, 17(Suppl 13):1025, 2016.

D. H. Hubel et T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160:106–154, Jan. 1962.

M.-C. Hung et W. Link. Protein localization in disease and therapy. *J. Cell Sci.*, 124(20): 3381–3392, 2011.

L. D. Hurst et A. R. Merchant. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc. Biol. Sci.*, 268(1466): 493–497, Mar. 2001.

V. Huser, M. Sincan, et J. J. Cimino. Developing genomic knowledge bases and databases to support clinical management: current perspectives. *Pharmgenomics. Pers. Med.*, 7: 275–283, 2014.

M. D. Iglesia, J. S. Parker, K. A. Hoadley, J. S. Serody, C. M. Perou, et B. G. Vincent. Genomic Analysis of Immune Cell Infiltrates Across 11 Tumor Types. *J. Natl. Cancer Inst.*, 108(11):djw144, 2016.

S. Jaeger, S. Fulle, et S. Turk. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.*, 58(1):27–35, 2018.

C. A. Janeway, Jr, P. Travers, M. Walport, et M. J. Shlomchik. *Immunobiology*. Garland Science, 2001.

M. K. Jenkins, H. H. Chu, J. B. McLachlan, et J. J. Moon. On the composition of the preimmune repertoire of T cells specific for peptide-major histocompatibility complex ligands. *Annu. Rev. Immunol.*, 28:275–294, 2010a.

M. K. Jenkins, H. Hamlet Chu, J. B. McLachlan, et J. J. Moon. On the composition of the preimmune repertoire of T cells specific for Peptide–Major histocompatibility complex ligands. *Annual review of*, Mar. 2010b.

S. A. Johnson, S. L. Seale, R. M. Gittelman, J. A. Rytlewski, H. S. Robins, et P. A. Fields. Impact of HLA type, age and chronic viral infection on peripheral t-cell receptor sharing between unrelated individuals. *PLoS One*, 16(8):e0249484, Aug. 2021.

G. Kassiotis. Endogenous retroviruses and the development of cancer. *J. Immunol.*, 192(4): 1343–1349, 2014.

J. Katzman, U. Shaham, J. Bates, A. Cloninger, T. Jiang, et Y. Kluger. DeepSurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. June 2016.

K. Kedzierska, P. G. Thomas, V. Venturi, M. P. Davenport, P. C. Doherty, S. J. Turner, et N. L. La Gruta. Terminal deoxynucleotidyltransferase is required for the establishment of private virus-specific CD8+ TCR repertoires and facilitates optimal CTL responses. *J. Immunol.*, 181(4):2556–2562, Aug. 2008.

J. Keilwagen, F. Hartung, M. Paulini, S. O. Twardziok, et J. Grau. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*, 19(1):189, May 2018.

E. M. Kernfeld, R. M. J. Genga, K. Neherin, M. E. Magaletta, et P. Xu. A Single-Cell Transcriptomic Atlas of Thymus Organogenesis Resolves Cell Types and Developmental Maturation. 2018.

D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, et S. L. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, Apr. 2013.

H. Kim et Y. M. Kim. Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types. *Sci. Rep.*, 8(1), 2018.

H. T. Kim, M.-J. Zhang, A. E. Woolfrey, A. St Martin, J. Chen, W. Saber, M.-A. Perales, P. Armand, et M. Eapen. Donor and recipient sex in allogeneic stem cell transplantation: what really matters. *Haematologica*, 101(10):1260–1266, Oct. 2016.

D. Kimothi, A. Soni, P. Biyani, et J. M. Hogan. Distributed Representations for Biological Sequence Analysis. 2016.

A. Klein. RoBO : A flexible and robust bayesian optimization framework in python. *undefined*, 2017.

D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, et R. K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22(3):568–576, 2012.

C. Krishna, D. Chowell, M. Gönen, Y. Elhanati, et T. A. Chan. Genetic and environmental determinants of human TCR repertoire diversity. *Immun. Ageing*, 17:26, Sept. 2020.

M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, et D. R. Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, 42(Database issue):980–985, 2014.

C. M. Laumont, K. Vincent, L. Hesnard, É. Audemard, É. Bonneil, J.-P. Laverdure, P. Gendron, M. Courcelles, M.-P. Hardy, C. Côté, C. Durette, C. St-Pierre, M. Benhammadi, J. Lanoix, S. Vobecky, E. Haddad, S. Lemieux, P. Thibault, et C. Perreault. Noncoding

regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.*, 10 (470), Dec. 2018.

K. Laurila et M. Vihinen. Prediction of disease-related mutations affecting protein localization. *BMC Genomics*, 10(1):122, 2009.

Y. LeCun, P. Haffner, L. Bottou, et Y. Bengio. Object recognition with Gradient-Based learning. In D. A. Forsyth, J. L. Mundy, V. di Gesú, et R. Cipolla, editors, *Shape, Contour and Grouping in Computer Vision*, pages 319–345. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

J. T. Leek. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.*, 42(21), Dec. 2014.

S. Lemieux, T. Sargeant, D. Laperrière, H. Ismail, G. Boucher, M. Rozendaal, V. P. Lavallée, D. Ashton-Beaucage, B. Wilhelm, J. Hébert, D. J. Hilton, S. Mader, et G. Sauvageau. MiSTIC, an integrated platform for the analysis of heterogeneity in large tumour transcriptome datasets. *Nucleic Acids Res.*, 45(13), 2017.

H. J. Levesque et G. Lakemeyer. *The Logic of Knowledge Bases*. MIT Press, Feb. 2001.

B. Li et C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, Aug. 2011.

R. G. H. Lindeboom, A. Regev, et S. A. Teichmann. Towards a human cell atlas: Taking notes from the past. *Trends Genet.*, 37(7):625–630, July 2021.

A. Liston, S. Lesage, J. Wilson, L. Peltonen, et C. C. Goodnow. Aire regulates negative selection of organ-specific T cells. *Nat. Immunol.*, 4(4):350–354, Apr. 2003.

A. Liston, E. J. Carr, et M. A. Linterman. Shaping variation in the human immune system. *Trends Immunol.*, 37(10):637–646, Oct. 2016.

A. Liston, S. Humblet-Baron, D. Duffy, et A. Goris. Human immune diversity: from evolution to modernity. *Nat. Immunol.*, 22(12):1479–1489, Dec. 2021.

K. Liu, L. He, Z. Liu, J. Xu, Y. Liu, Q. Kuang, Z. Wen, et M. Li. Mutation status coupled with RNA-sequencing data can efficiently identify important non-significantly mutated genes serving as diagnostic biomarkers of endometrial cancer. *BMC Bioinformatics*, 18 (Suppl 14):472, 2017.

J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand,

M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, et H. F. Moore. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, 45(6):580–585, 2013.

R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, et N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15(12):1053–1058, 2018.

Y.-F. Lu, D. B. Goldstein, M. Angrist, et G. Cavalleri. Personalized medicine and human genetic diversity. *Cold Spring Harb. Perspect. Med.*, 4(9):a008581, 2014.

G. Lythe, R. E. Callard, R. L. Hoare, et C. Molina-París. How many TCR clonotypes does a body maintain? *J. Theor. Biol.*, 389:214–224, Jan. 2016.

H. Maciejewski. Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.*, 15(4):504–518, 2014.

T. Macrae, T. Sargeant, S. Lemieux, J. Hébert, E. Deneault, et G. Sauvageau. RNA-Seq reveals spliceosome and proteasome genes as most consistent transcripts in human cancer cells. *PLoS One*, 8(9):e72884, Sept. 2013.

A. Madi, E. Shifrut, S. Reich-Zeliger, H. Gal, K. Best, W. Ndifon, B. Chain, I. R. Cohen, et N. Friedman. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.*, 24(10): 1603–1612, Oct. 2014.

S. C. Manekar et S. R. Sathe. A benchmark study of k-mer counting methods for high-throughput sequencing. *Gigascience*, 7(12), Dec. 2018.

C. Marchet, Z. Iqbal, D. Gautheret, M. Salson, et R. Chikhi. REINDEER: efficient indexing of k-mer presence and abundance in sequencing datasets. *Bioinformatics*, 36(Suppl_1): i177–i185, July 2020.

Q. Marcou, T. Mora, et A. M. Walczak. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.*, 9(1):561, Feb. 2018.

P. J. Martin. Increased disparity for minor histocompatibility antigens as a potential cause of increased GVHD risk in marrow transplantation from unrelated donors compared with related donors. *Bone Marrow Transplant.*, 8(3):217–223, Sept. 1991.

P. J. Martin, D. M. Levine, B. E. Storer, E. H. Warren, X. Zheng, S. C. Nelson, A. G. Smith, B. K. Mortensen, et J. A. Hansen. Genome-wide minor histocompatibility matching as related to the risk of graft-versus-host disease. *Blood*, 129(6):791–798, Feb. 2017.

A. Mayer, V. Balasubramanian, A. M. Walczak, et T. Mora. How a well-adapting immune system remembers. *Proc. Natl. Acad. Sci. U. S. A.*, 116(18):8815–8823, Apr. 2019.

K. Mayer-Blackwell, S. Schattgen, L. Cohen-Lavi, J. C. Crawford, A. Souquette, J. A. Gaevert, T. Hertz, P. G. Thomas, P. Bradley, et A. Fiore-Gartland. TCR meta-clonotypes for biomarker discovery with tcrdist3: identification of public, HLA-restricted SARS-CoV-2 associated TCR features. Mar. 2021.

C. McInnes. Progress in the evaluation of CDK inhibitors as anti-tumor agents. *Drug Discov. Today*, 13(19-20):875–881, 2008.

P. Melsted et J. K. Pritchard. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics*, 12:333, Aug. 2011.

M. Merkenschlager, D. Graf, M. Lovatt, U. Bommhardt, R. Zamoyska, et A. G. Fisher. How many thymocytes audition for selection? *J. Exp. Med.*, 186(7):1149–1158, Oct. 1997.

E. Miho, A. Yermanos, C. R. Weber, C. T. Berger, S. T. Reddy, et V. Greiff. Computational strategies for dissecting the High-Dimensional complexity of adaptive immune repertoires. *Front. Immunol.*, 9:224, Feb. 2018.

T. Mikolov, K. Chen, G. Corrado, et J. Dean. Efficient estimation of word representations in vector space. Jan. 2013.

M. Mittelbrunn et G. Kroemer. Hallmarks of T cell aging. *Nat. Immunol.*, 22(6):687–698, June 2021.

D. T. Montoro, A. L. Haber, M. Biton, V. Vinarsky, B. Lin, S. E. Birket, F. Yuan, S. Chen, H. M. Leung, J. Villoria, N. Rogel, G. Burgin, A. M. Tsankov, A. Waghray, M. Slyper, J. Waldman, L. Nguyen, D. Dionne, O. Rozenblatt-Rosen, P. R. Tata, H. Mou, M. Shivaraju, H. Bihler, M. Mense, G. J. Tearney, S. M. Rowe, J. F. Engelhardt, A. Regev, et J. Rajagopal. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature*, 560(7718):319–324, Aug. 2018.

K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, et S. Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.*, 37(12):1482–1492, 2019.

T. Mora et A. M. Walczak. How many different clonotypes do immune repertoires contain? *Current Opinion in Systems Biology*, 18:104–110, Dec. 2019.

M. J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti, E. J. P. de Koning, et A. van Oudenaarden. A Single-Cell transcriptome atlas of the human pancreas. *Cell Syst*, 3(4):385–394.e3, Oct. 2016.

K. P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, 2012.

J. M. Murray, G. R. Kaufmann, P. D. Hodgkin, S. R. Lewin, A. D. Kelleher, M. P. Davenport, et J. J. Zaunders. Naive T cells are maintained by thymic output in early ages but by proliferation without phenotypic change after age twenty. *Immunol. Cell Biol.*, 81(6):

487–495, Dec. 2003.

A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, et A. A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, 12(5):453–457, 2015.

P. Ng. dna2vec: Consistent vector representations of variable-length k-mers. 2017.

X. Ni, Q. Song, K. Cassady, R. Deng, H. Jin, M. Zhang, H. Dong, S. Forman, P. J. Martin, Y.-Z. Chen, J. Wang, et D. Zeng. PD-L1 interacts with CD80 to regulate graft-versus-leukemia activity of donor CD8+ T cells. *J. Clin. Invest.*, 127(5):1960–1977, May 2017.

S. Nolan, M. Vignali, M. Klinger, J. N. Dines, I. M. Kaplan, E. Svejnoha, T. Craft, K. Boland, M. Pesesky, R. M. Gittelman, T. M. Snyder, C. J. Gooley, S. Semprini, C. Cerchione, M. Mazza, O. M. Delmonte, K. Dobbs, G. Carreño-Tarragona, S. Barrio, V. Sambri, G. Martinelli, J. D. Goldman, J. R. Heath, L. D. Notarangelo, J. M. Carlson, J. Martinez-Lopez, et H. S. Robins. A large-scale database of t-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Res Sq*, Aug. 2020.

K. J. V. Nordström, M. C. Albani, G. V. James, C. Gutjahr, B. Hartwig, F. Turck, U. Paszkowski, G. Coupland, et K. Schneeberger. Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nat. Biotechnol.*, 31(4):325–330, 2013.

B. J. Olson. *Statistical Methods for Adaptive Immune Receptor Repertoire Analysis and Comparison*. PhD thesis, University of Washington, Ann Arbor, United States, 2020.

B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, et A. M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 17(1):132, 2016.

R. N. Pahwa, M. J. Modak, T. McMorrow, S. Pahwa, G. Fernandes, et R. A. Good. Terminal deoxynucleotidyl transferase (TdT) enzyme in thymus and bone marrow. i. age-associated decline of TdT in humans and mice. *Cell. Immunol.*, 58(1):39–48, Feb. 1981.

A. F. Palazzo et T. R. Gregory. The case for junk DNA. *PLoS Genet.*, 10(5):e1004351, May 2014.

M. S. Palmer, P. Berta, A. H. Sinclair, B. Pym, et P. N. Goodfellow. Comparison of human ZFY and ZFX transcripts. *Proc. Natl. Acad. Sci. U. S. A.*, 87(5):1681–1685, 1990.

S. Palmer, L. Albergante, C. C. Blackburn, et T. J. Newman. Thymic involution and rising disease incidence with age. *Proc. Natl. Acad. Sci. U. S. A.*, 115(8):1883–1888, Feb. 2018.

J. Park, R. Shrestha, C. Qiu, A. Kondo, S. Huang, M. Werth, M. Li, J. Barasch, et K. Suszták. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science*, 360(6390):758–763, 2018.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et Others. Advances in neural information processing systems

32. *Curran Associates, Inc*, pages 8024–8035, 2019.

R. Patro, S. M. Mount, et C. Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, 32(5):462–464, 2014.

H. Pearson, T. Daouda, D. P. Granados, C. Durette, E. Bonneil, M. Courcelles, A. Rodenbrock, J.-P. Laverdure, C. Côté, S. Mader, S. Lemieux, P. Thibault, et C. Perreault. MHC class i-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.*, 126(12):4690–4701, Dec. 2016.

W. R. Pearson. An Introduction to Sequence Similarity ("Homology") Searching. *Curr. Protoc. Bioinformatics*, 42(1):1–3, 2013.

J. Pennington, R. Socher, et C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

E. Pierson, D. Koller, A. Battle, S. Mostafavi, et S. Mostafavi. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Comput. Biol.*, 11(5):e1004220, 2015.

M. V. Pogorelyy et M. Shugay. A framework for annotation of antigen specificities in High-Throughput T-Cell repertoire sequencing studies. *Front. Immunol.*, 10:2159, Sept. 2019.

M. V. Pogorelyy, Y. Elhanati, Q. Marcou, A. L. Sycheva, E. A. Komech, V. I. Nazarov, O. V. Britanova, D. M. Chudakov, I. Z. Mamedov, Y. B. Lebedev, T. Mora, et A. M. Walczak. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Comput. Biol.*, 13(7):e1005572, July 2017.

M. V. Pogorelyy, A. A. Minervina, M. Shugay, D. M. Chudakov, Y. B. Lebedev, T. Mora, et A. M. Walczak. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol.*, 17(6):e3000314, June 2019.

C. P. Ponting. The Human Cell Atlas: making 'cell space' for disease. *Dis. Model. Mech.*, 12(2):dmm037622, 2019.

A. L. Price, N. C. Jones, et P. A. Pevzner. De novo identification of repeat families in large genomes. *Bioinformatics*, 21(Suppl 1):i351–i358, 2005.

J. Punt, S. Stranford, P. Jones, et J. Owen. *Kuby Immunology*. W. H. Freeman, Oct. 2018.

Q. Qi, M. M. Cavanagh, S. Le Saux, L. E. Wagar, S. Mackey, J. Hu, H. Maecker, G. E. Swan, M. M. Davis, C. L. Dekker, L. Tian, C. M. Weyand, et J. J. Goronzy. Defective T memory cell differentiation after varicella zoster vaccination in older individuals. *PLoS Pathog.*, 12(10):e1005892, Oct. 2016.

Y. Qi. Random forest for bioinformatics. In *Ensemble Machine Learning*, pages 307–323. Springer US, Boston, MA, 2012.

M. F. Quigley, H. Y. Greenaway, V. Venturi, R. Lindsay, K. M. Quinn, R. A. Seder, D. C. Douek, M. P. Davenport, et D. A. Price. Convergent recombination shapes the clonotypic

landscape of the naive t-cell repertoire. *Proc. Natl. Acad. Sci. U. S. A.*, 107(45):19414–19419, Nov. 2010.

N. Rappoport et R. Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.*, 46(20):10546–10562, 2018.

A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. Van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, et N. Yosef. The human cell atlas. *Elife*, 6, 2017.

D. Risso, J. Ngai, T. P. Speed, et S. Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, 32(9):896–902, Sept. 2014.

G. Rizk, D. Lavenier, et R. Chikhi. DSK: k-mer counting with very low memory usage. *Bioinformatics*, 29(5):652–653, Mar. 2013.

J. Robinson, D. J. Barker, X. Georgiou, M. A. Cooper, P. Flicek, et S. G. E. Marsh. IPD-IMGT/HLA database. *Nucleic Acids Res.*, 48(D1):D948–D955, Jan. 2020.

S. T. Roweis et L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Technical report, 2000.

O. Rozenblatt-Rosen, M. J. T. Stubbington, A. Regev, et S. A. Teichmann. The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453, Oct. 2017.

B. D. Rudd. Neonatal T cells: A reinterpretation. *Annu. Rev. Immunol.*, 38:229–247, Apr. 2020.

D. E. Rumelhart, G. E. Hinton, et R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct. 1986.

S. A. Schattgen, K. Guion, J. C. Crawford, A. Souquette, A. M. Barrio, M. J. T. Stubbington, P. G. Thomas, et P. Bradley. Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nat. Biotechnol.*, Aug. 2021.

A. Schneider-Gädicke, P. Beer-Romero, L. G. Brown, R. Nussbaum, et D. C. Page. ZFX has a gene structure similar to ZFY, the putative human sex determinant, and escapes X inactivation. *Cell*, 57(7):1247–1258, 1989.

J. Schreiber, T. J. Durham, J. Bilmes, et W. S. Noble. Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *bioRxiv*, page 364976, 2019.

Z. Sethna, Y. Elhanati, C. G. Callan, A. M. Walczak, et T. Mora. OLGA: fast computation of generation probabilities of B- and t-cell receptor amino acid sequences and motifs.

Bioinformatics, 35(17):2974–2981, Sept. 2019.

Z. Sethna, G. Isacchini, T. Dupic, T. Mora, A. M. Walczak, et Y. Elhanati. Population variability in the generation and selection of t-cell repertoires. *PLoS Comput. Biol.*, 16 (12):e1008394, Dec. 2020.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, Oct. 1948.

Z. Shen, W. Bao, et D.-S. Huang. Recurrent neural network for predicting transcription factor binding sites. *Sci. Rep.*, 8(1):15270, Oct. 2018.

E. H. Simpson. Measurement of diversity. *Nature*, 163(4148):688–688, Apr. 1949.

J. Sng, B. Ayoglu, J. W. Chen, J.-N. Schickel, E. M. N. Ferre, S. Glauzy, N. Romberg, M. Hoenig, C. Cunningham-Rundles, P. J. Utz, M. S. Lionakis, et E. Meffre. AIRE expression controls the peripheral selection of autoreactive B cells. *Sci Immunol*, 4(34), Apr. 2019.

J. Snoek, H. Larochelle, et R. P. Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, et K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

G. Socié et B. R. Blazar. Acute graft-versus-host disease: from the bench to the bedside. *Blood*, 114(20):4327–4336, Nov. 2009.

R. R. Sokal et C. D. Michener. *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas, 1958.

A. Sood, M.-È. Lebel, M. Dong, M. Fournier, S. J. Vobecky, É. Haddad, J.-S. Delisle, J. N. Mandl, N. Vrisekoop, et H. J. Melichar. CD5 levels define functionally heterogeneous populations of naïve human CD4+ T cells. *Eur. J. Immunol.*, Mar. 2021.

C. Soto, R. G. Bombardi, M. Kozhevnikov, R. S. Sinkovits, E. C. Chen, A. Branchizio, N. Kose, S. B. Day, M. Pilkinton, M. Gujral, S. Mallal, et J. E. Crowe, Jr. High frequency of shared clonotypes in human T cell receptor repertoires. *Cell Rep.*, 32(2):107882, July 2020.

I. Springer, H. Besser, N. Tickotsky-Moskovitz, S. Dvorkin, et Y. Louzoun. Prediction of specific TCR-Peptide binding from large dictionaries of TCR-Peptide pairs. *Front. Immunol.*, 11:1803, Aug. 2020.

A. Sud, B. Kinnersley, et R. S. Houlston. Genome-wide association studies of cancer: Current insights and future perspectives. *Nat. Rev. Cancer*, 2017.

C. Tanchot, F. A. Lemonnier, B. Pérarnau, A. A. Freitas, et B. Rocha. Differential requirements for survival and proliferation of CD8 naïve or memory T cells. *Science*, 276(5321): 2057–2062, June 1997.

V. Thomas-Vaslin, H. K. Altes, R. J. de Boer, et D. Klatzmann. Comprehensive assessment and mathematical modeling of T cell population dynamics and homeostasis. *J. Immunol.*,

180(4):2240–2250, Feb. 2008.

J. J. C. Thome, K. L. Bickham, Y. Ohmura, M. Kubota, N. Matsuoka, C. Gordon, T. Granot, A. Griesemer, H. Lerner, T. Kato, et D. L. Farber. Early-life compartmentalization of human T cell differentiation and regulatory function in mucosal and lymphoid tissues. *Nat. Med.*, 22(1):72–77, Jan. 2016.

V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, T.-H. Ou Yang, E. Porta-Pardo, G. F. Gao, C. L. Plaisier, J. A. Eddy, E. Ziv, A. C. Culhane, E. O. Paull, I. K. A. Sivakumar, A. J. Gentles, R. Malhotra, F. Farshidfar, A. Colaprico, J. S. Parker, L. E. Mose, N. S. Vo, J. Liu, Y. Liu, J. Rader, V. Dhankani, S. M. Reynolds, R. Bowlby, A. Califano, A. D. Cherniack, D. Anastassiou, D. Bedognetti, Y. Mokrab, A. M. Newman, A. Rao, K. Chen, A. Krasnitz, H. Hu, T. M. Malta, H. Noushmehr, C. S. Pedamallu, S. Bullman, A. I. Ojesina, A. Lamb, W. Zhou, H. Shen, T. K. Choueiri, J. N. Weinstein, J. Guinney, J. Saltz, R. A. Holt, C. S. Rabkin, Cancer Genome Atlas Research Network, A. J. Lazar, J. S. Serody, E. G. Demicco, M. L. Disis, B. G. Vincent, et I. Shmulevich. The immune landscape of cancer. *Immunity*, 48(4):812–830.e14, Apr. 2018.

P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, Å. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. von Feilitzen, K. S. Lilley, M. Uhlén, et E. Lundberg. A subcellular map of the human proteome. *Science*, 356(6340): eaal3321, 2017.

N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, et N. Friedman. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*, 33 (18):2924–2929, Sept. 2017.

A. Trofimov, F. Dutil, C. Perreault, S. Lemieux, Y. Bengio, et J. P. Cohen. Towards the Latent Transcriptome. 2018.

C. M. Tucker, M. W. Cadotte, S. B. Carvalho, T. J. Davies, S. Ferrier, S. A. Fritz, R. Grenyer, M. R. Helmus, L. S. Jin, A. O. Mooers, S. Pavoine, O. Purschke, D. W. Redding, D. F. Rosauer, M. Winter, et F. Mazel. A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biol. Rev. Camb. Philos. Soc.*, 92(2):698–715, May 2017.

M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson,

M. Zwahlen, G. von Heijne, J. Nielsen, et F. Ponten. Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419, 2015.

O. Valbuena, K. B. Marcu, C. M. Croce, K. Huebner, M. Weigert, et R. P. Perry. Chromosomal locations of mouse immunoglobulin genes. *Proc. Natl. Acad. Sci. U. S. A.*, 75(6): 2883–2887, 1978.

S. Valkiers, N. de Vrij, S. Gielis, S. Verbandt, B. Ogunjimi, K. Laukens, et P. Meysman. Recent advances in t-cell receptor repertoire analysis: bridging the gap with multimodal single-cell RNA sequencing. *ImmunoInformatics*, page 100009, Jan. 2022.

L. J. P. Van Der Maaten et G. E. Hinton. Visualizing high-dimensional data using t-sne. *J. Mach. Learn. Res.*, 9(Nov):2579–2605, 2008.

D. Van Dijk, R. Sharma, J. Nainys, G. Wolf, S. Krishnaswamy, D. Pe'er Correspondence, et G. A. Gene. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion In Brief Population Analysis Archetypal Analysis Gene Interactions. 2018.

V. Venturi, H. Y. Chin, T. E. Asher, K. Ladell, P. Scheinberg, E. Bornstein, D. van Bockel, A. D. Kelleher, D. C. Douek, D. A. Price, et M. P. Davenport. TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. *J. Immunol.*, 181(11): 7853–7862, Dec. 2008.

K. Vincent, D.-C. Roy, et C. Perreault. Next-generation leukemia immunotherapy. *Blood*, 118(11):2951–2959, Sept. 2011.

R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, et B. Peters. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.*, 43(Database issue):405–412, 2015.

R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, et B. Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.*, 47(D1):D339–D343, 2019a.

R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, et B. Peters. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.*, 47(D1):D339–D343, Jan. 2019b.

N. Vrisekoop, J. P. Monteiro, J. N. Mandl, et R. N. Germain. Revisiting thymic positive selection and the mature T cell repertoire for antigen. *Immunity*, 41(2):181–190, Aug. 2014.

A. Wagner, A. Regev, et N. Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, 34(11):1145–1160, 2016.

J. Wagner, M. A. Rapsomaniki, S. Chevrier, T. Anzeneder, C. Langwieder, A. Dykgers, M. Rees, A. Ramaswamy, S. Muenst, S. D. Soysal, A. Jacobs, J. Windhager, K. Silina, M. van den Broek, K. J. Dedes, M. Rodríguez Martínez, W. P. Weber, et B. Bodenmiller. A Single-Cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell*, 177 (5):1330–1345.e18, May 2019.

Q. Wang, J. Armenia, C. Zhang, A. V. Penson, E. Reznik, L. Zhang, T. Minet, A. Ochoa, B. E. Gross, C. A. Iacobuzio-Donahue, D. Betel, B. S. Taylor, J. Gao, et N. Schultz. Unifying cancer and normal RNA sequencing data from different sources. *Sci Data*, 5: 180061, Apr. 2018.

E. H. Warren, X. C. Zhang, S. Li, W. Fan, B. E. Storer, J. W. Chien, M. J. Boeckh, L. P. Zhao, P. J. Martin, et J. A. Hansen. Effect of MHC and non-MHC donor/recipient genetic disparity on the outcome of allogeneic HCT. *Blood*, 120(14):2796–2806, Oct. 2012.

D. Weese, A.-K. Emde, T. Rausch, A. Döring, et K. Reinert. RazerS–fast read mapping with sensitivity control. *Genome Res.*, 19(9):1646–1654, Sept. 2009.

J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, et J. M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120, 2013.

B. Xia, Y. Yan, M. Baron, F. Wagner, D. Barkley, M. Chiodin, S. Y. Kim, D. L. Keefe, J. P. Alukal, J. D. Boeke, et I. Yanai. Widespread transcriptional scanning in the testis modulates gene evolution rates. *bioRxiv*, page 282129, 2019.

I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, et O. Shmueli. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–659, 2005.

J. Ye, N. Ma, T. L. Madden, et J. M. Ostell. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.*, 41(Web Server issue):W34–40, July 2013.

M. J. Yousefzadeh, R. R. Flores, Y. Zhu, Z. C. Schmiechen, R. W. Brooks, C. E. Trussoni, Y. Cui, L. Angelini, K.-A. Lee, S. J. McGowan, A. L. Burrack, D. Wang, Q. Dong, A. Lu, T. Sano, R. D. O'Kelly, C. A. McGuckian, J. I. Kato, M. P. Bank, E. A. Wade, S. P. S. Pillai, J. Klug, W. C. Ladiges, C. E. Burd, S. E. Lewis, N. F. LaRusso, N. V. Vo, Y. Wang, E. E. Kelley, J. Huard, I. M. Stromnes, P. D. Robbins, et L. J. Niedernhofer. An aged immune system drives senescence and ageing of solid organs. *Nature*, 594(7861):100–105, June 2021.

D. Zemmour, R. Zilionis, E. Kiner, A. M. Klein, D. Mathis, et C. Benoist. Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nat. Immunol.*, 19(3):291–301, Mar. 2018.

L. Zhang et Z. Zhang. Recharacterizing Tumor-Infiltrating Lymphocytes by Single-Cell RNA Sequencing. *Cancer Immunology Research*, 7(7):1040–1046, 2019.

Y. Zhang, G. Parmigiani, et W. E. Johnson. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform*, 2(3):lqaa078, Sept. 2020.

W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, et C. M. Perou. Comparison of RNA-Seq by poly (a) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*, 15:419, June 2014.

G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, et J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, 2017.

J. Zhou et O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, 12(10):931–934, Oct. 2015.

A. Zielezinski, S. Vinga, J. Almeida, et W. M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.*, 18(1):186, 2017.

Appendix A

Article 1: Matériel supplémentaire

Ce chapitre contient de l'information supplémentaire pour l'article présenté au Chapitre 3. Le texte intégral inclus dans le supplément est en anglais.

Supplementary information accompanying the submission: Factorized embeddings learns rich and biologically meaningful embedding spaces using factorized tensor decomposition

A.1. Table of Content

We have included for easy reference a table of content for the figures.

This table of content does not contain the auxiliary task (Chapter 3) supplementary material. These additional figures can be found in section A.3.

A.2. Supporting figures - part I

A.2.1. Hyper-parameter selection for the factorized embeddings model

We trained five factorized embeddings models on the *TCGA* cohort using embedding spaces of varying size. To assess the performance of each trained embedding space, we test each embedding space using a classification task: prediction of *TCGA* cancer type. We report that the while 2-dimensional embeddings did not perform well, there was no significant difference between spaces of size 25, 50 and 200 (Figure A.1). For the rest of our work, we chose to focus on the 50-dimensional embedding space.

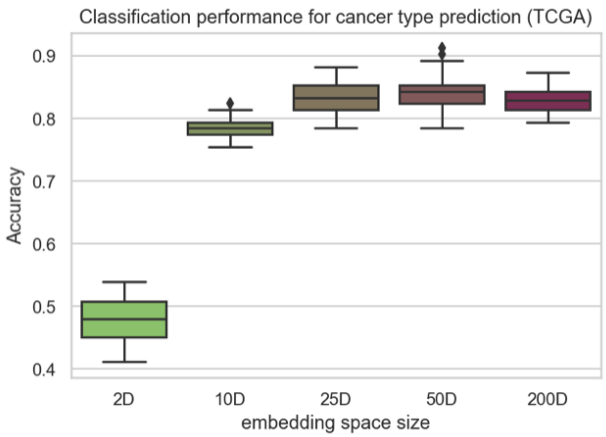


Fig. A.1. Effect of the embedding size on classification performance

Then, keeping the embedding space constant at 50 dimensions, we varied the capacity of the multi-layer perceptron (MLP) ($g()$ function, see Methods) by varying the number of layers. We trained five different FE models on the *TCGA* with different MLP architectures, as described in Figure A.2. To assess the performance of each embedding space thus trained, we tested on the same cancer type prediction classification task. We report that the best performance was obtained by the 5-layer MLP with architecture [250,150,100,50,10], where each number is the number of neurons per layer.

A.2.2. Toy dataset testing: Swiss roll

We then compared the performance of the FE model to tSNE by training both models on two toy datasets popular in manifold learning.

We found that while both FE and tSNE learn a similar representation of the S dataset, this was not the case for the *Swissroll* dataset (Figure A.3). We hypothesize that this may be due to the fact that FE attempts to preserve *all* feature similarities, whereas tSNE preserves

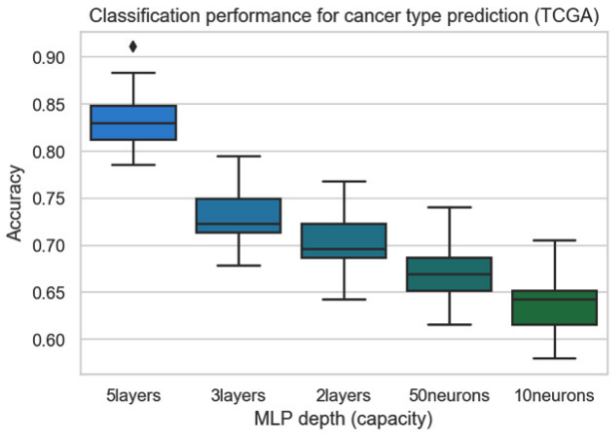


Fig. A.2. Effect of the MLP capacity on classification performance

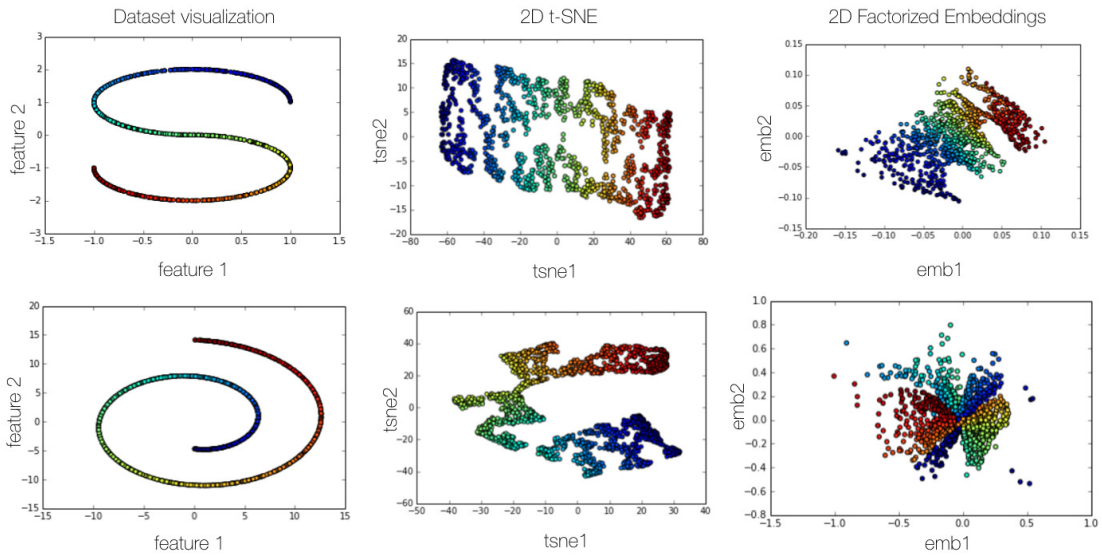


Fig. A.3. Factorized embedding visualisation of the *S* and the *Swissroll* datasets

only local relationships. This can be seen in the way in the FE representation, the dark blue points are located near both the yellow and the red, a constraint that tSNE does not enforce (Figure A.3).

A.2.2.1. Toy dataset testing: MNIST dataset. We then trained the factorized embeddings model on the MNIST dataset (LeCun el. al 1998). Briefly, this dataset consists of 50000 grey-scale images of handwritten digits (0-9) of 28x28 pixels. While it has been shown in the literature that the most appropriate architecture for training models on image tasks are convolutional-neural networks, we chose to perform this experiment because the MNIST dataset is well characterized and can offer some insight into the limitations of the factorized embedding model. We first compared the embeddings obtained from training the factorized embeddings model to those obtained either with a PCA or tSNE (Figure A.4A). We found

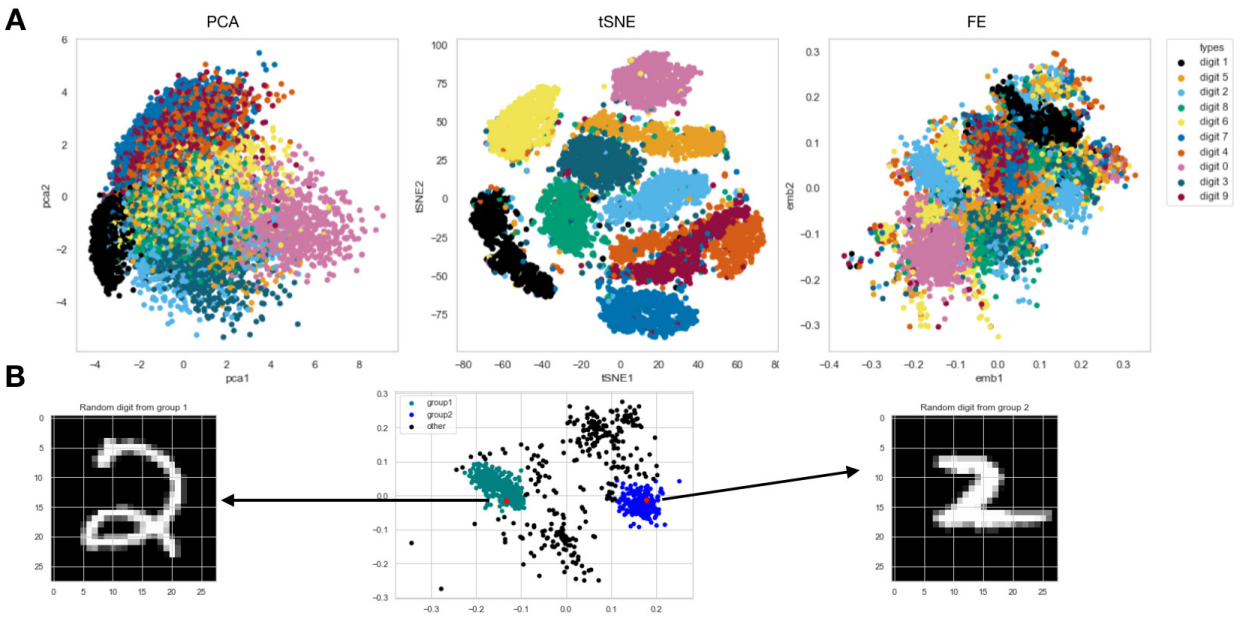


Fig. A.4. Sample embedding space for MNIST
A) Comparison side by size of 2D PCA, tSNE and factorized embeddings for MNIST dataset. Colours represent various digits of the MNIST dataset. B) Analysis of the two largest clusters for the digit 2. Two example images are shown

that while tSNE grouped images by digit type, factorized embeddings did so too but to a lesser extent. Notably, we observed that some digits were split into multiple groups, unlike in the tSNE embeddings. We investigated more closely and found that this may be in part due to a change of style that may be captured by the factorized embeddings model (Figure A.4B). Similar to the results obtained with the Swiss roll dataset, the FE model seemed to attempt to capture all variations.

To validate that the FE model was performing well in terms of reconstruction, we measured the reconstruction accuracy by correlation between either an image and itself or an image (Figure A.5B) and another image of the same digit (Figure A.5A). We found that the FE model seemed to reconstruct with good accuracy digits, other digits of the same type had a poor correlation, confirming that the model reconstructed the specific digit and did not limit itself to predicting average pixel values for the digit class.

To validate if the learned embedding space is dense, we created a 50 by 50 point grid over the embedding space and for each point generated a new digit image. We measured the euclidean distance between this prototype digit and each class average and found that the embeddings space is dense and may allow for interpolation between samples (Figure A.66). We found that while digits 0 and 1 are polar opposites in terms of FE digit space, the digit space for 1 closely resembles that of digit 7, as well as 3 resembles 8 (Figure A.6). We rationalize that this may be due to a shared amount of similarity in the digit writing and hence a high amount of overlap in activated pixels.

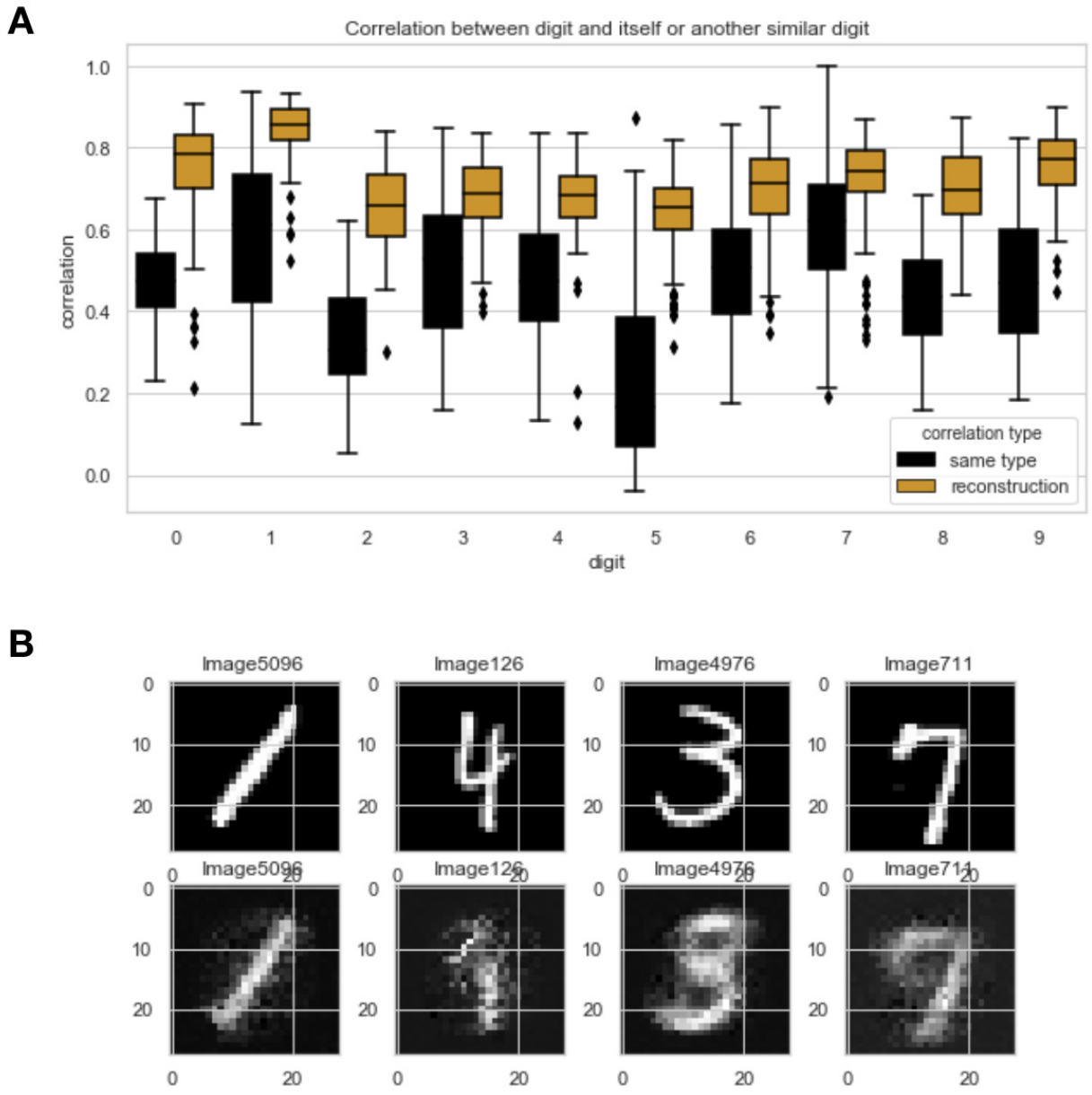


Fig. A.5. Reconstruction and imputation tasks for MNIST
A) Reconstruction accuracy for each pixel measured by correlation between the image and its original. Images are split by digit type B) Reconstruction of some selected digits.

Finally, we looked at the pixel embedding space, and found that as in Supp. Figure 6, the pixel space for similar digits has almost identical patterns (Figure A.7A). Indeed, digits 4 and 9, digits 3, 5, 6 and 8, and digits 1 and 7, have similar pixel activation, made evident by the FE model (Figure A.7A). Moreover, when the edge pixels in the MNIST dataset are always "off" (no digit is written there" and we found that these pixels were grouped together and located around the (0,0) coordinate, suggesting that the feature value plays a role in the final embedding coordinate (Figure A.7B).

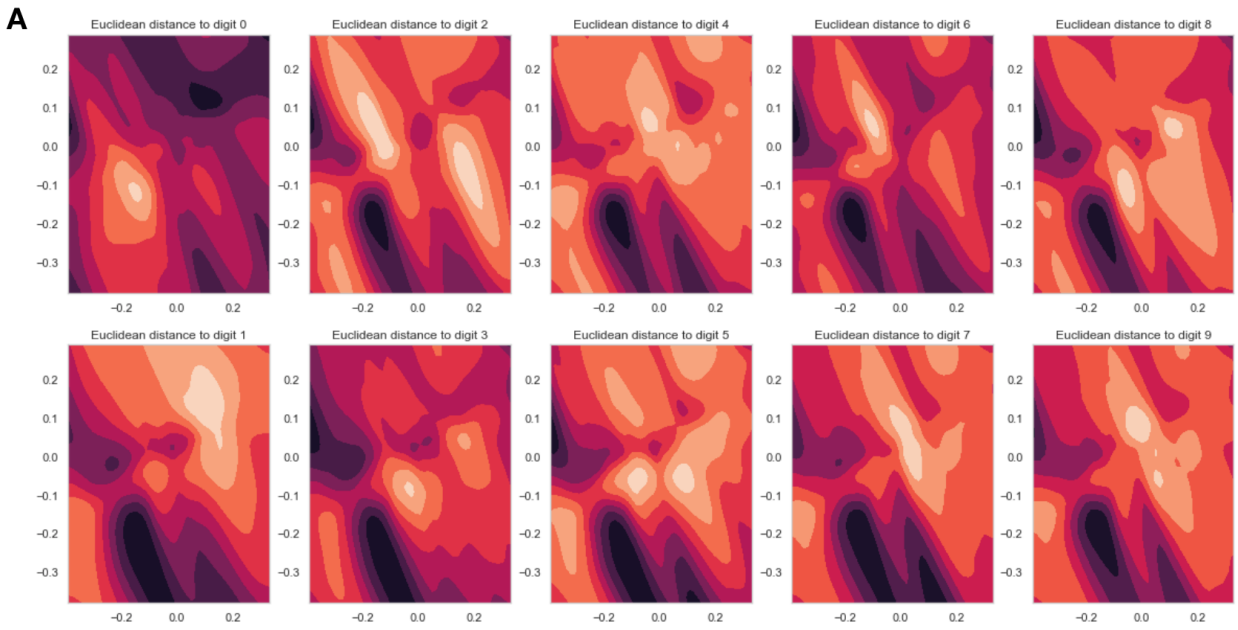


Fig. A.6. Euclidean distance between each generated image and the average image for that class

Taken together, these results on the toy datasets suggest that the FE model attributes some importance to the feature values - all-zeros will most likely not play any role in the training. Moreover, while the FE model is not appropriate for image analysis, it had some mild success in grouping together images of the same class. We conclude that this is probably due to similarity in pixel activation. Finally, we state that the experiments done on the Swissroll and MNIST datasets put forward some limitations of the model: the FE model is not appropriate for datasets that are non-linear in their feature space (as seen for Swissroll) or is made of images.

A.2.3. Limits of the Factorized embeddings model in gene expression datasets

This section is dedicated to some experiments that showcase the limits of the FE model. Following the experiment of Figure 2, showcasing some partiality of the model towards tissue-specific genes over sex-specific genes, we investigated with more details the reason why the XIST gene did not yield the same results as the others. Since the figure in question was generated using 2D embeddings, we verified if varying the size of the embedding may capture more information on sex-specific genes. To do so, we compare the performance of a classifier trained on 4 sizes of embedding spaces in predicting either tissue type or sex (Figure A.8).

We found that varying the size of the embedding did not have an impact on the classification performance for the prediction of sex.

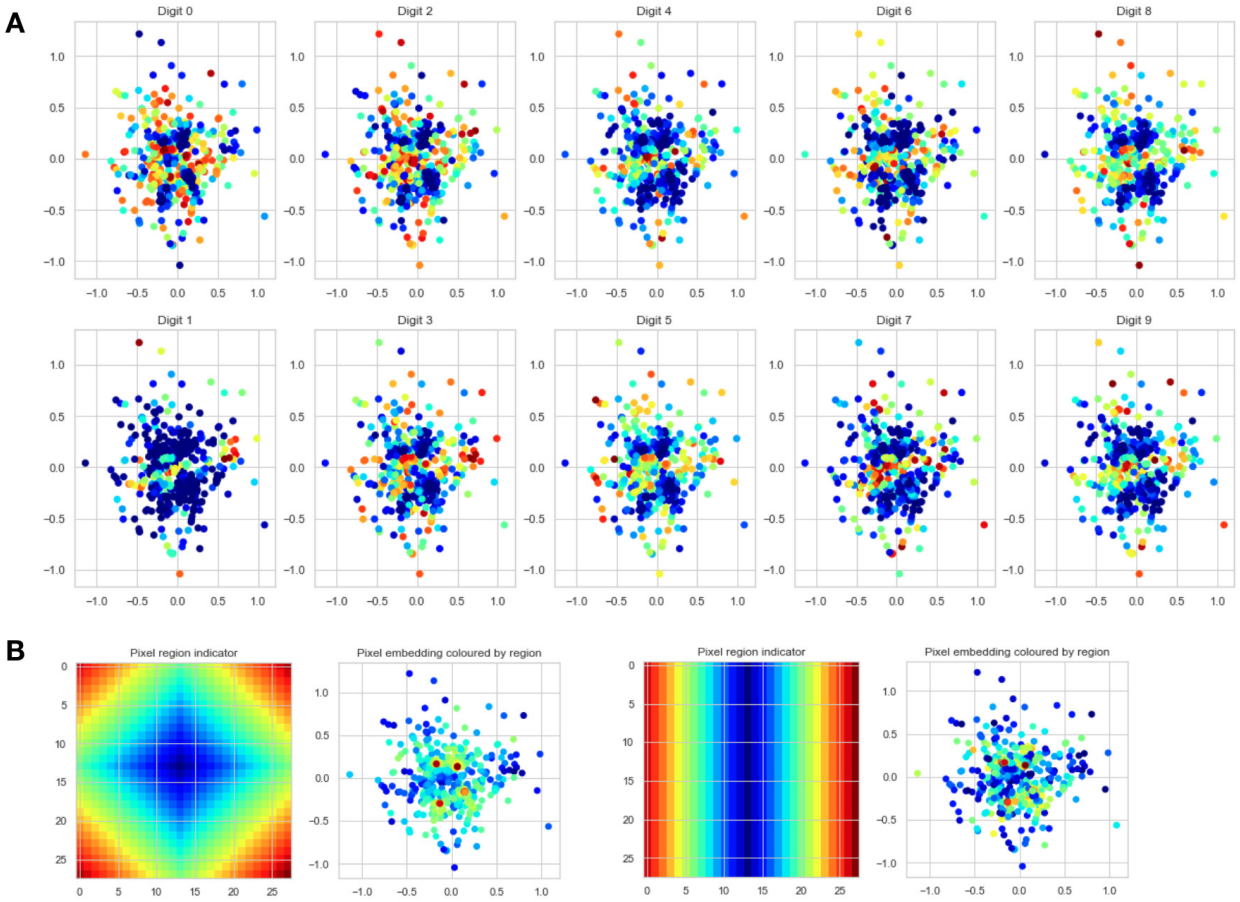


Fig. A.7. Pixel embeddings in MNIST

A) Pixel embeddings coloured by average value for that class B) Pixel embeddings coloured by position as shown on each region indicator plot

To attempt to quantify the signal contained for classification of tissue types and sex, we trained a random forest classifier algorithm on each task and queried the feature usage, as seen in ((Deng et Runger)). Briefly, we train for each task a random forest on all genes and extract the genes that were used for the classification. This is performed 100 times and this yields a relative percentage of usage of each gene for the classification. We use this usage metric as a proxy for the signal measurement. We found that for the tissue-type prediction task 12079 genes were found to be signal-rich, whereas for the sex-type task only 7277 genes were used (Figure A.8B-D). While there was a mild overlap, we found that in general, the reconstruction of sex-specific genes was a lot worse than the reconstruction of tissue-specific genes and genes in both categories (Figure A.8C-D).

Taken together these results show that overall the FE model captured better the tissue-specific genes as opposed to the sex-specific genes. This may be in part due to the larger set size that tissue-specific genes constitute as well as the higher gene expression values tissue-specific genes have, when compared to sex-specific.

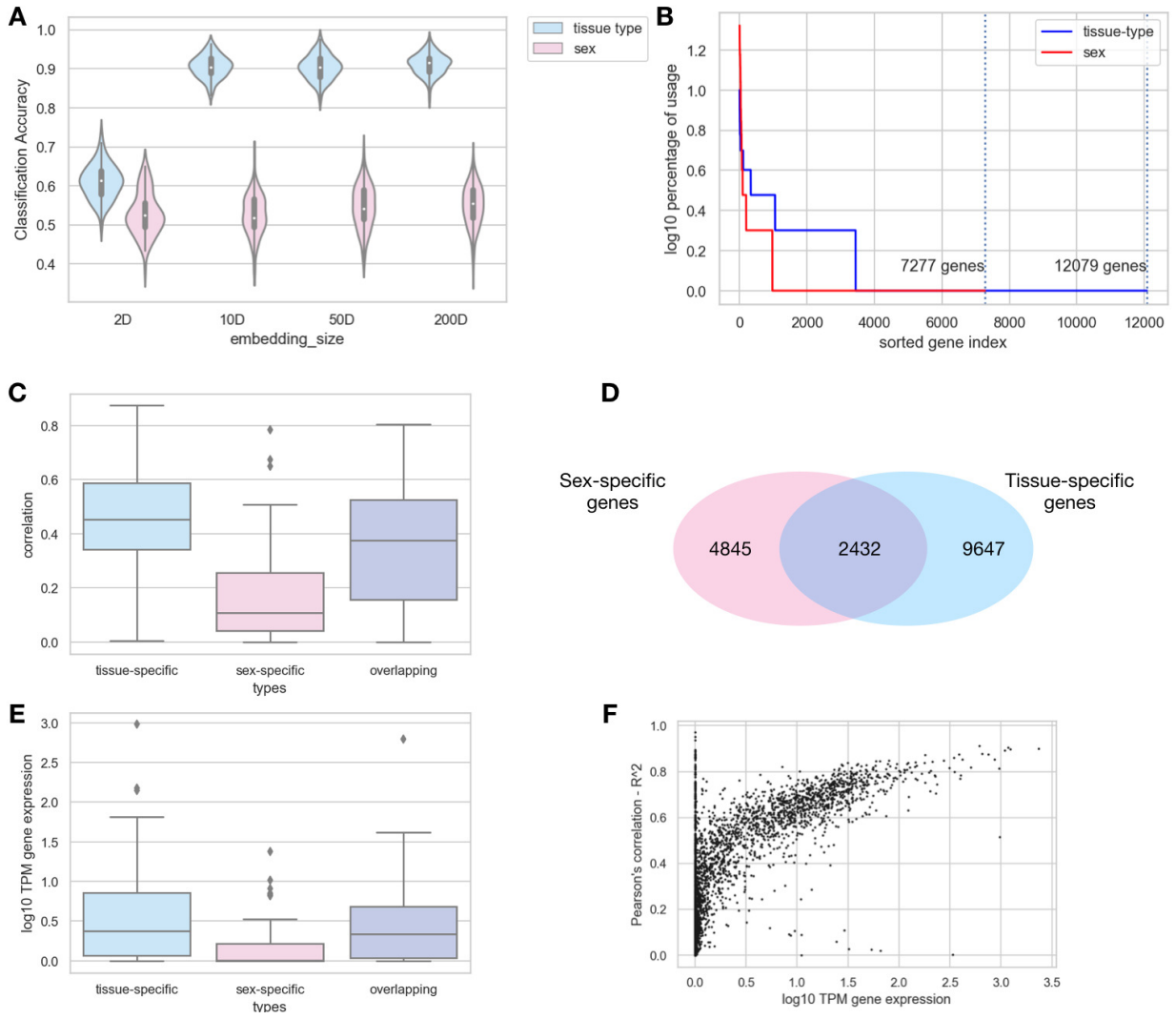


Fig. A.8. Tissue- and sex-specific gene signals in FE embeddings
A) Performance of a classifier trained on embedding spaces of varying size to predict either tissue type of sex B) Percentage of gene usage (log10 scale) over 100 random shuffles in the classification tasks for both tissue-type and sex. Only a total of 12079 genes for tissue-type and 7277 genes for sex prediction were used in at least 1% of all re-shuffles C) Tissue-specific and sex-specific genes sets mildly overlap with 2432 genes in common. D) Average gene expression values for each group E) Reconstruction performance for each group of signal-rich genes measured by correlation.

A.2.4. Gene embeddings supplementary information

In their work, Du and colleagues mention that their gene embeddings recover gene type information ((Du et al., 2019)). We wanted to verify if the FE-trained gene embeddings recovered this information as well. We found that similar to the *gene2vec* model by Du and colleagues, the FE model dedicated part of its gene space to different types of genes, grouping, for instance, protein coding genes together (Figure A.9).

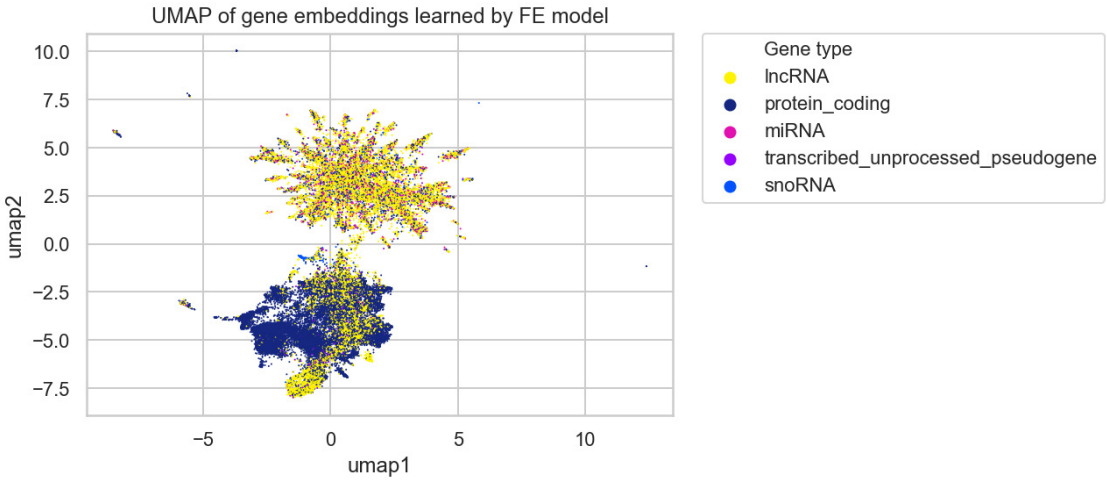


Fig. A.9. Gene embeddings by gene type

A.3. Supplementary information for the auxiliary tasks

For the reader that might be interested in a specific task performance, we plot the performance for each of the 49 auxiliary tasks.

A.3.1. Microscopy task group

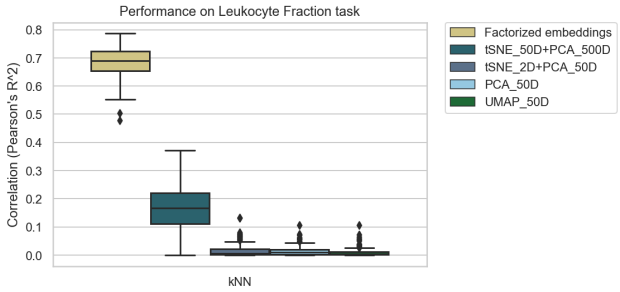


Fig. A.10. Performance on prediction of the leukocyte fraction

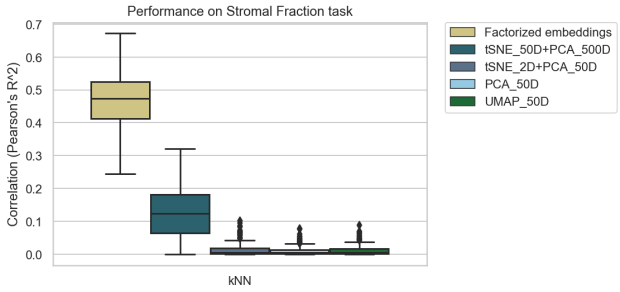


Fig. A.11. Performance on prediction of the stromal fraction

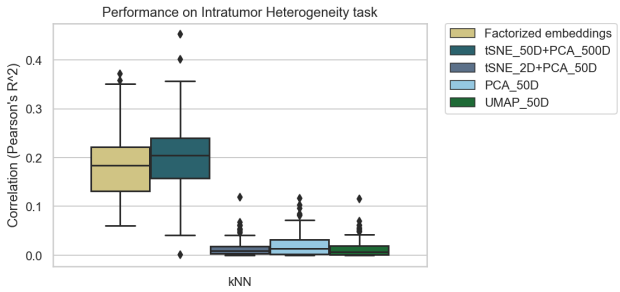


Fig. A.12. Performance on prediction of intra-tumor heterogeneity

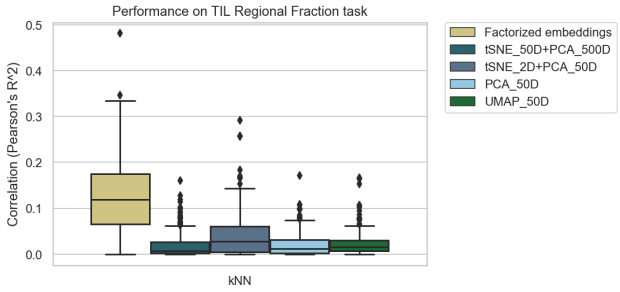


Fig. A.13. Prediction of tumor-infiltrating lymphocyte regional fraction

A.3.2. Cibersort task group

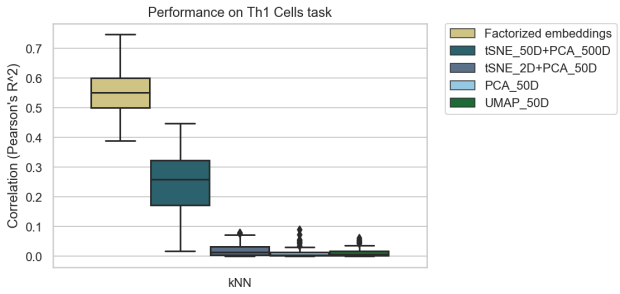


Fig. A.14. Prediction of the abundance of infiltrating Th1 Cells

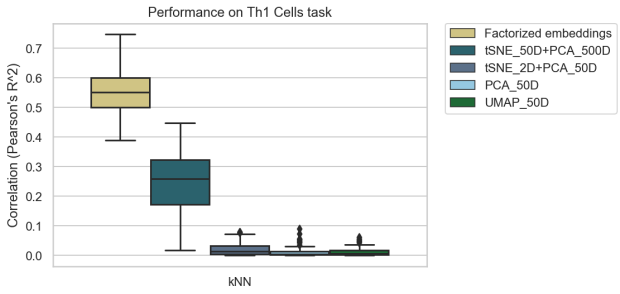


Fig. A.15. Prediction of the abundance of infiltrating Th1 Cells

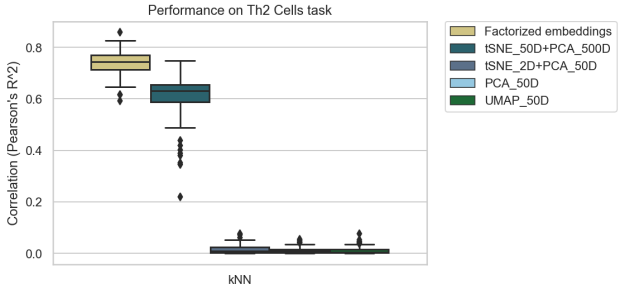


Fig. A.16. Prediction of the abundance of infiltrating Th2 Cells

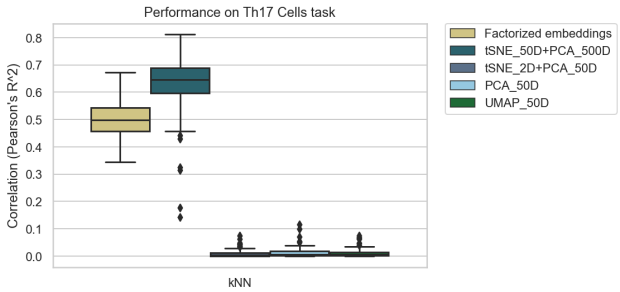


Fig. A.17. Prediction of the abundance of infiltrating Th17 Cells

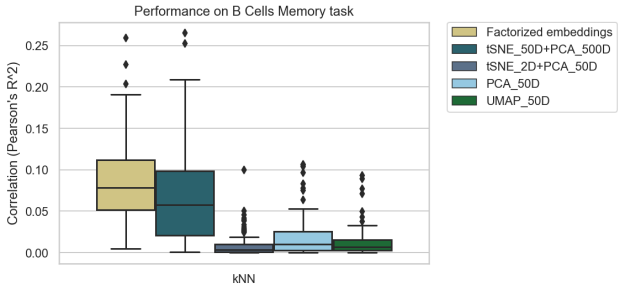


Fig. A.18. Prediction of the abundance of infiltrating Memory B cells

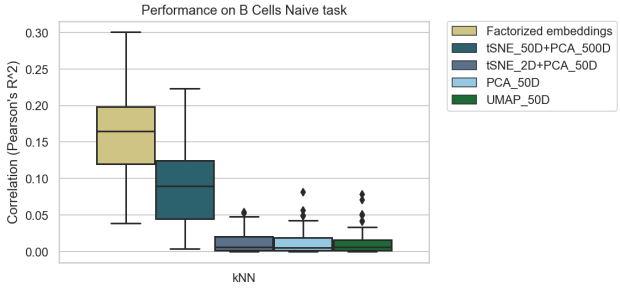


Fig. A.19. Prediction of the abundance of infiltrating Naive B Cells

A.3.2.1. Thorsson immune profiles.

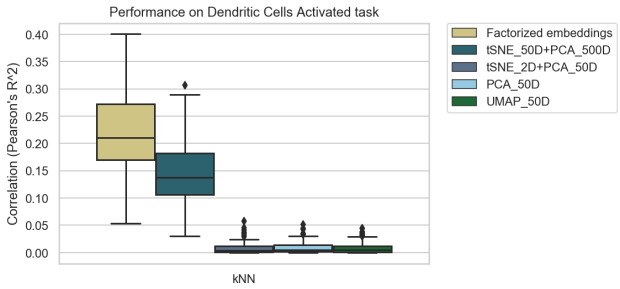


Fig. A.20. Prediction of the abundance of infiltrating activated dendritic cells

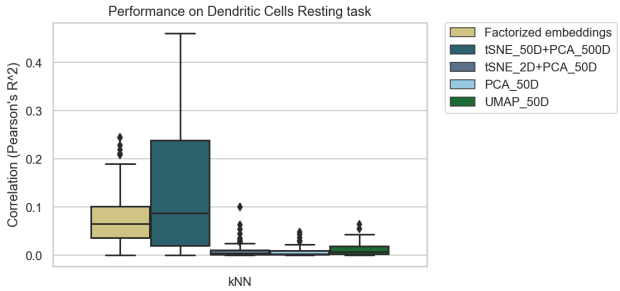


Fig. A.21. Prediction of the abundance of infiltrating resting dendritic cells

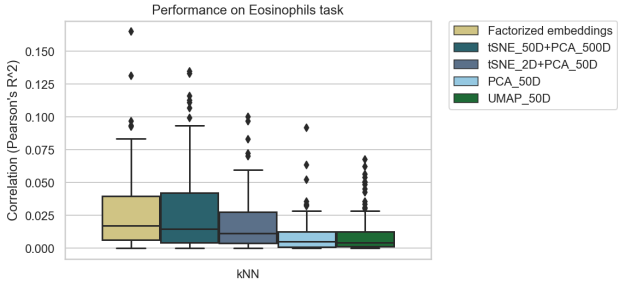


Fig. A.22. Prediction of the abundance of infiltrating eosinophils

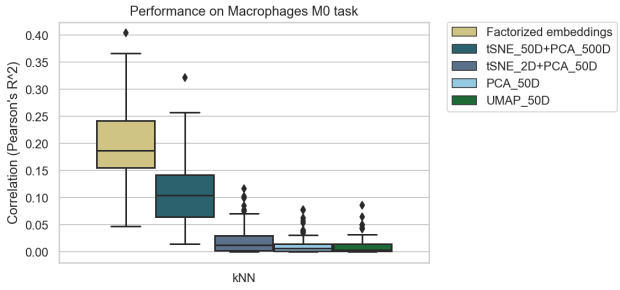


Fig. A.23. Prediction of the abundance of infiltrating M0 macrophages

A.3.3. Genomic instability task group

A.3.4. Immune repertoire task group

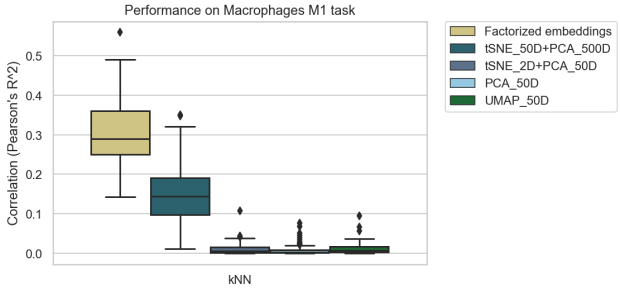


Fig. A.24. Prediction of the abundance of infiltrating M1 macrophages

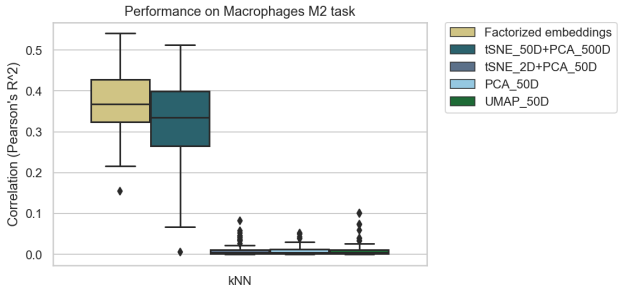


Fig. A.25. Prediction of the abundance of infiltrating M2 macrophages

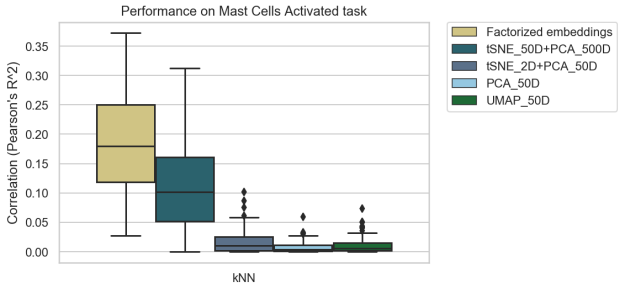


Fig. A.26. Prediction of the abundance of infiltrating activated mast cells

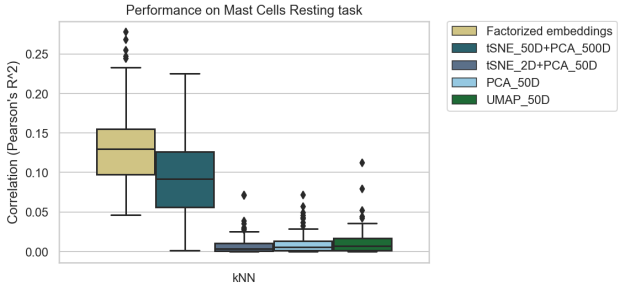


Fig. A.27. Prediction of the abundance of infiltrating resting mast cells

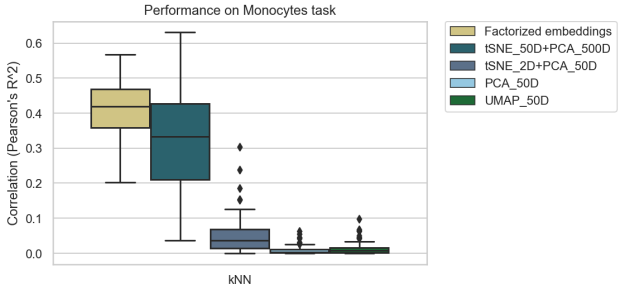


Fig. A.28. Prediction of the abundance of infiltrating monocytes

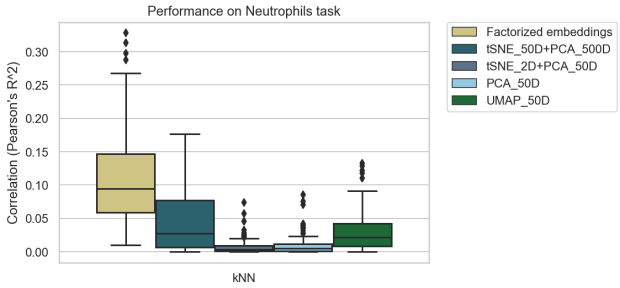


Fig. A.29. Prediction of the abundance of infiltrating neutrophils

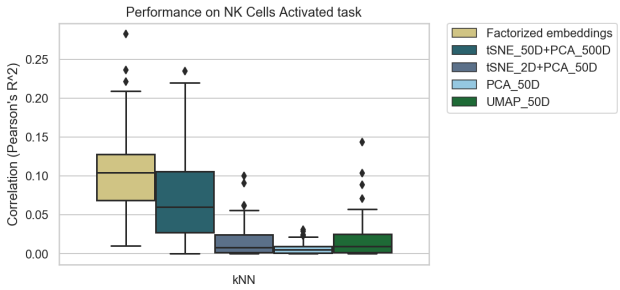


Fig. A.30. Prediction of the abundance of infiltrating activated natural killer (NK) cells

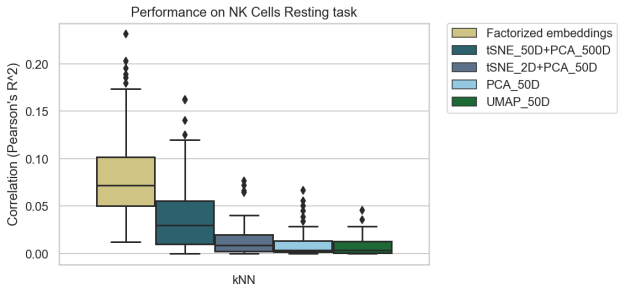


Fig. A.31. Prediction of the abundance of infiltrating resting NK cells

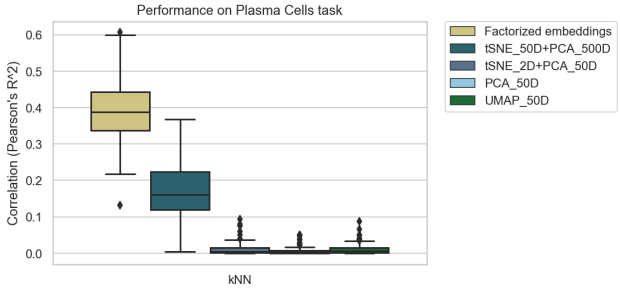


Fig. A.32. Prediction of the abundance of infiltrating plasma cells

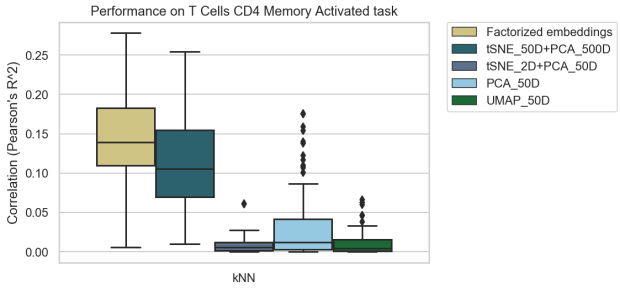


Fig. A.33. Prediction of the abundance of infiltrating activated CD4 memory T cells

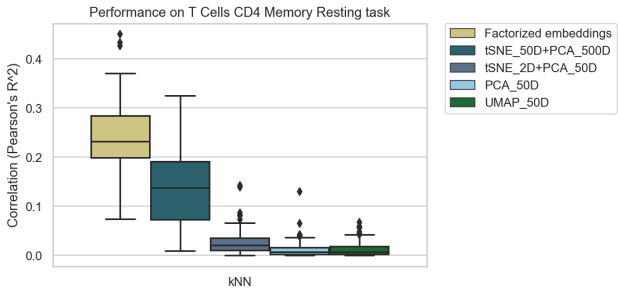


Fig. A.34. Prediction of the abundance of infiltrating resting CD4 memory T cells

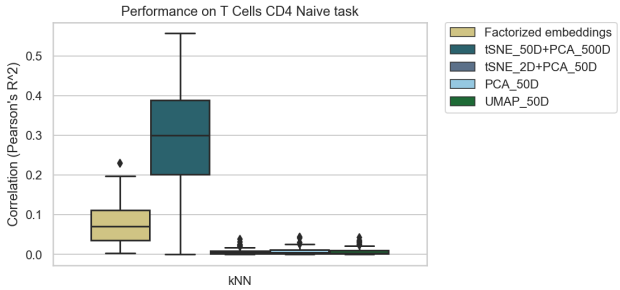


Fig. A.35. Prediction of the abundance of infiltrating naive CD4 T cells

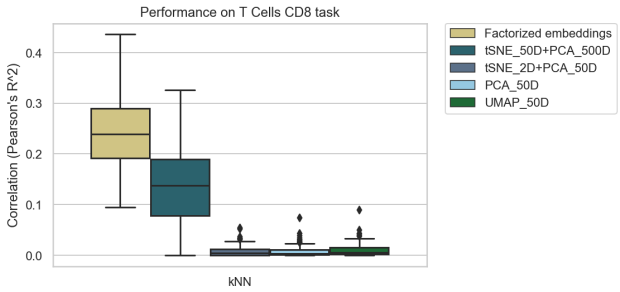


Fig. A.36. Prediction of the abundance of infiltrating resting CD8 T cells

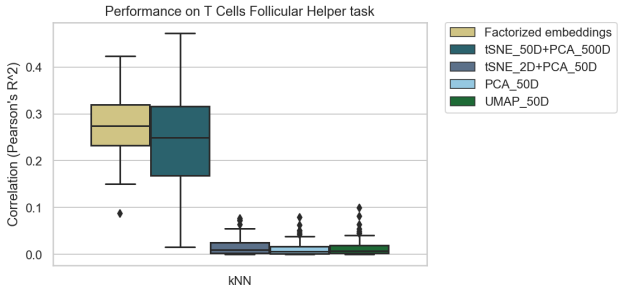


Fig. A.37. Prediction of the abundance of infiltrating follicular helper T cells

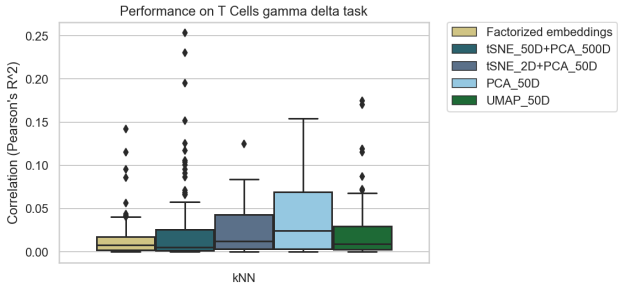


Fig. A.38. Prediction of the abundance of infiltrating gamma delta T cells

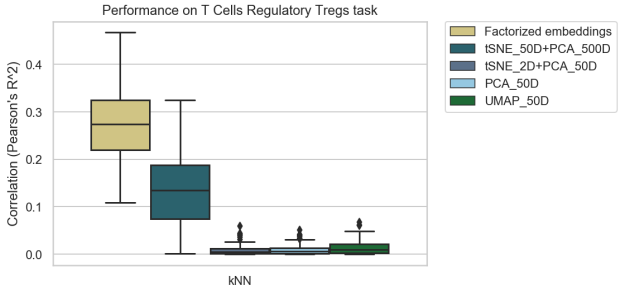


Fig. A.39. Prediction of the abundance of infiltrating regulatory T cells (T_{reg})

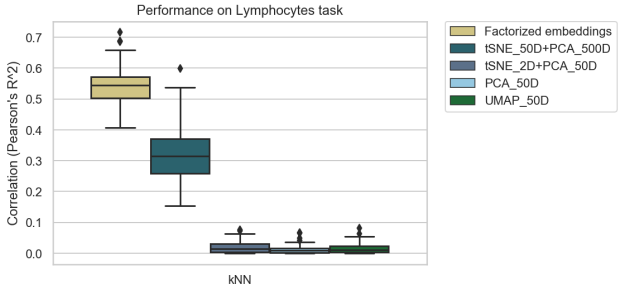


Fig. A.40. Prediction of the abundance of infiltrating lymphocytes

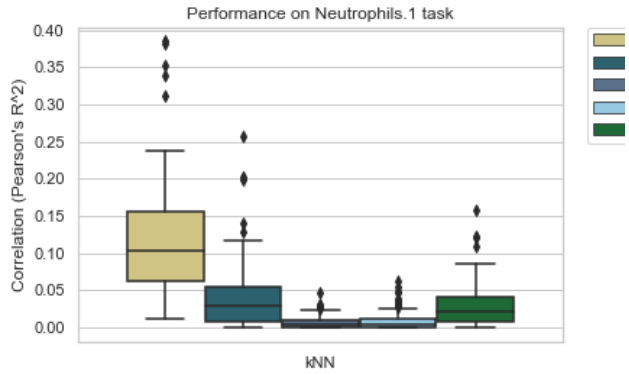


Fig. A.41. Prediction of the abundance of infiltrating neutrophils

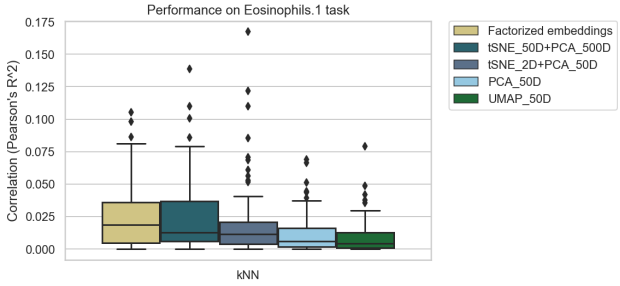


Fig. A.42. Prediction of the abundance of infiltrating eosinophils

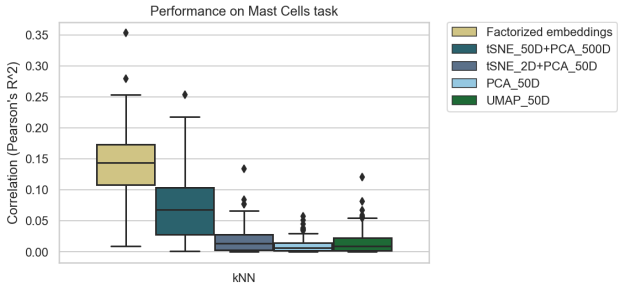


Fig. A.43. Prediction of the abundance of total infiltrating mast cells

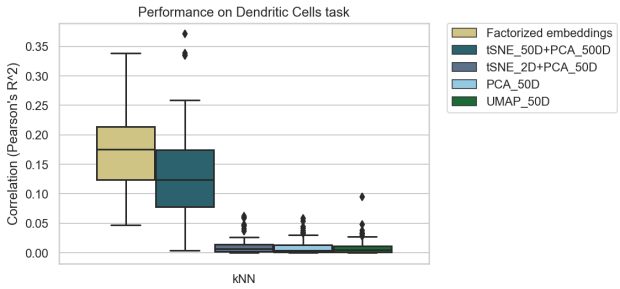


Fig. A.44. Prediction of the abundance of total infiltrating dendritic cells

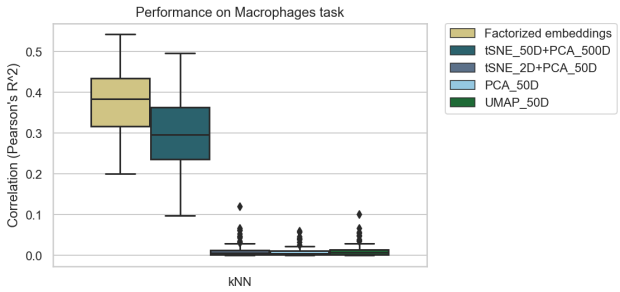


Fig. A.45. Prediction of the abundance of total infiltrating macrophages

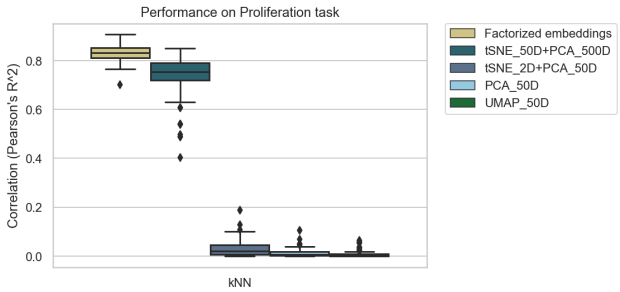


Fig. A.46. Prediction of proliferation immune profile

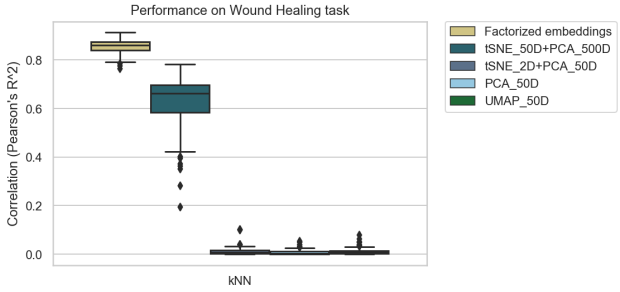


Fig. A.47. Prediction of wound healing immune profile

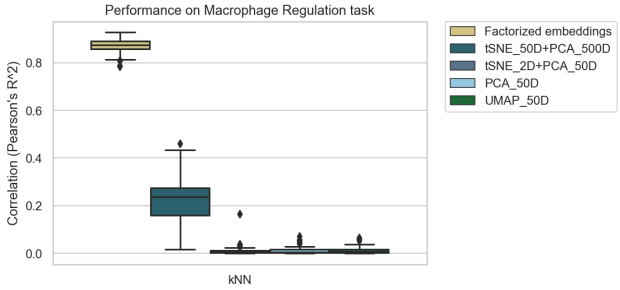


Fig. A.48. Prediction of macrophage regulation immune profile

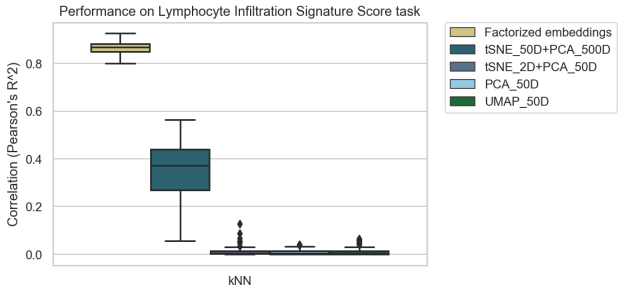


Fig. A.49. Prediction of lymphocyte infiltration signature score

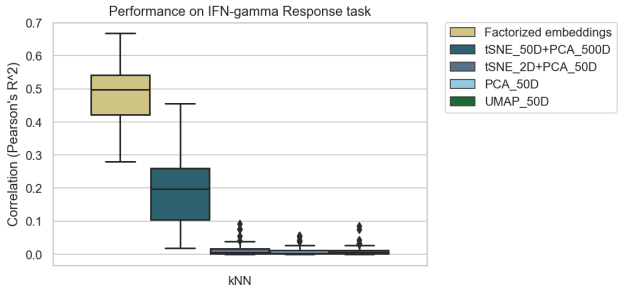


Fig. A.50. Prediction of IFN-γ response immune profile

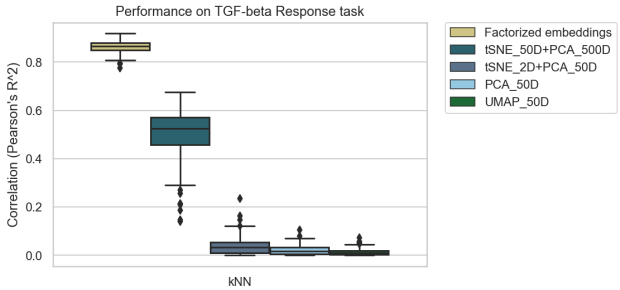


Fig. A.51. Prediction of TGF-β response immune profile

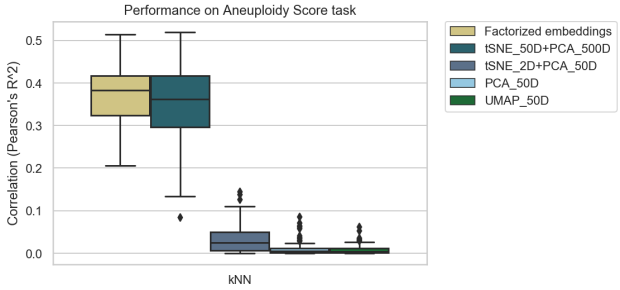


Fig. A.52. Prediction of aneuploidy score

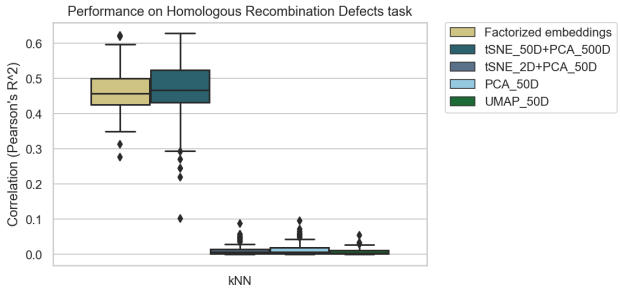


Fig. A.53. Prediction of homologous recombination defects occurence

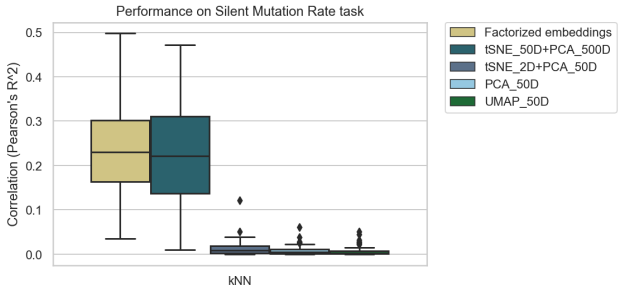


Fig. A.54. Prediction of silent mutation rate

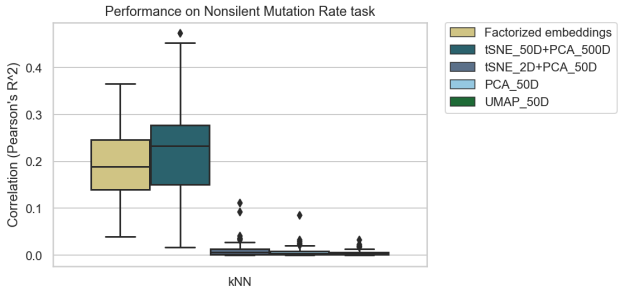


Fig. A.55. Prediction of non-silent mutation rate

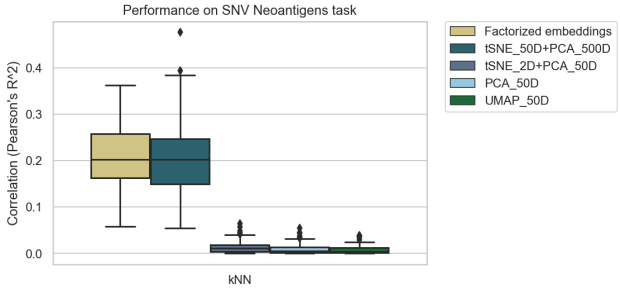


Fig. A.56. Prediction of SNV neoantigen occurence score

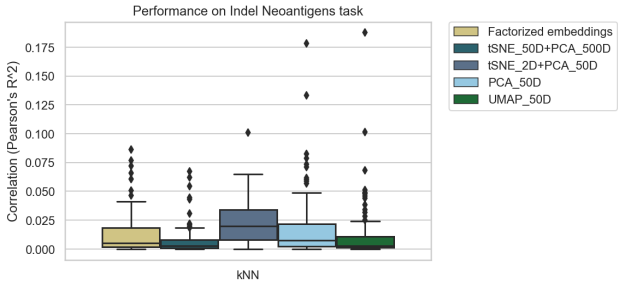


Fig. A.57. Prediction of indel neoantigen occurence score

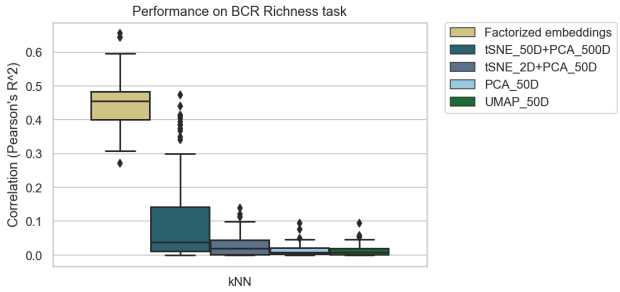


Fig. A.58. Prediction of B cell receptor richness score

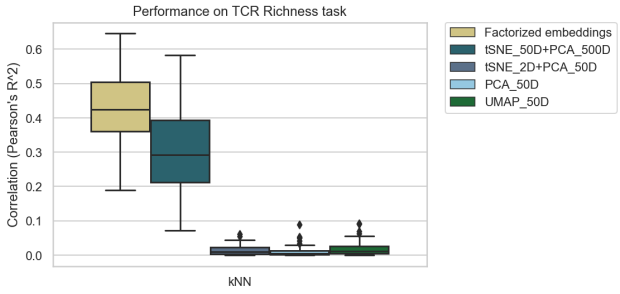


Fig. A.59. Prediction of T cell receptor richness score

Appendix B

Article 3: Matériel supplémentaire

Cette annexe contient les légendes des figures et les figures supplémentaires de l'article présenté au Chapitre 6, telles que présentées lors de la soumission de l'article au journal Immunity.

B.1. Légendes des Figures supplémentaires

Figure B.1: Hierarchical clustering of repertoire sharing (pairwise Jaccard distance) among subjects of the Britanova cohort for (A) public and (B) superpublic CDR3s.

Figure B.2: Boxplots showing (A-C) public fraction percentages and (B-D) summed clonality of public CDR3aas in the Emerson cohort CMV+ and CMV- individuals. Boxplots showing (E-G) public fraction percentages and (F-H) summed clonality of public CDR3aas for naïve and effector-memory T cells in the Thome cohort.

Figure B.3: (A) ERGO scores were used to predict the binding of CDR3aa longer or shorter than 15 amino acids to MHC-associated peptides of viral and human origin. ERGO scores were used to assess the relationship between two features of CDR3aa from the Britanova cohort: publicness and polyreactivity to (B) viral and (C) human peptides. Polyreactivity vs. (D) CDR3aa length, (E) the number of mismatches, and (F) log10 recombination frequency in the 10x Genomics dataset. (G) Distribution of Britanova cohort individuals by sex and with broad age groups depicted in Figure 3.

Figure B.4: (A) Venn diagram showing the overlap between two major CDR3 categories from McPAS: autoimmunity and microbial pathogens. (B) Heatmap shows, for subjects of the Britanova cohort, the frequency of CDR3aa listed in the McPAS autoimmune dataset. Rows represent individuals, columns unique CDR3aa, and cell color indicates CDR3aa clone size. Row dendrogram leaves are colored by age group. (C) Age distribution for subjects in three individual clusters from (A). (D) Boxplot showing CDR3aa length in five clusters from (B). (E) Boxplot showing predicted recombination frequency for CDR3aa in five clusters from (B).

Figure B.5: Line plots showing percentage overlaps with McPAS (A) Pathogen-specific and (B) Autoimmune-specific CDR3 set, with the top N most frequent CDR3aa. Line colors and types correspond to age groups. (C) SARS-CoV2-specific CDR3 overlaps vs. a randomly selected set of CDR3 in cord blood. Boxplots showing (D) the recombination frequency and (E) the number of mismatches for CDR3aa in aGVHD+ vs. aGVHD- donors.

Figure B.6: Kaplan-Meier plots showing splits by all four and individual quartiles for (A-C) CDR3 length, (D-F) the number of mismatches, (G-I) recombination probability, and (J-L) Simpson diversity. This figure complements the results from Figure 6.

Figure B.7: (A) Scatterplot showing the recombination frequency for CDR3s based on the classification probability output of the logistic regression model (neonatal vs. TDT-dependent). Each dot is a CDR3s, and the color scheme represents the degree of sharing through the cohort. (B) The proportion of neonatal CDR3s by age group in the Britanova cohort.

B.2. Figures supplémentaires

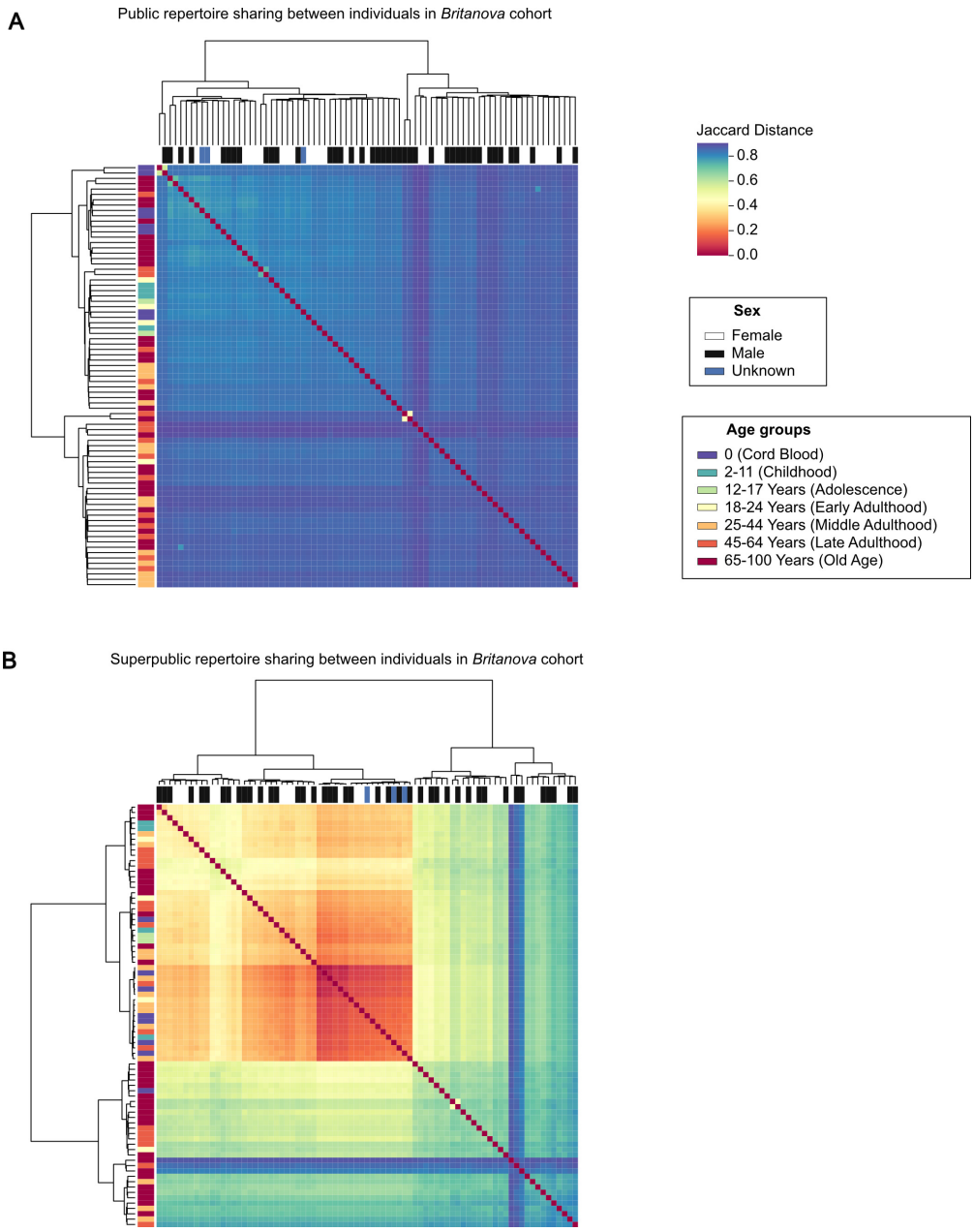


Fig. B.1. Supplementary Figure 1

Fig. B.1. Hierarchical clustering of repertoire sharing (pairwise Jaccard distance) among subjects of the Britanova cohort for (A) public and (B) superpublic CDR3s.

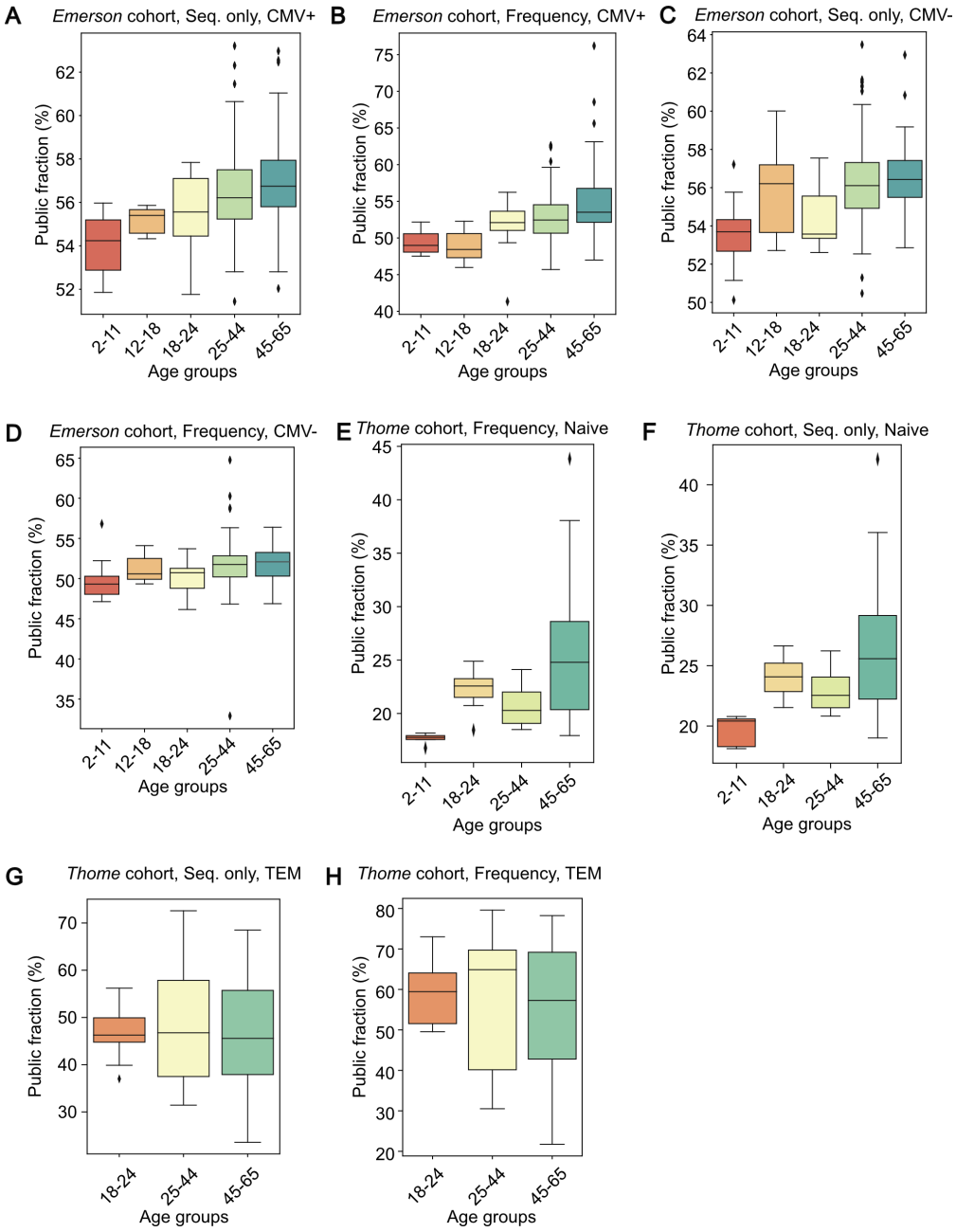
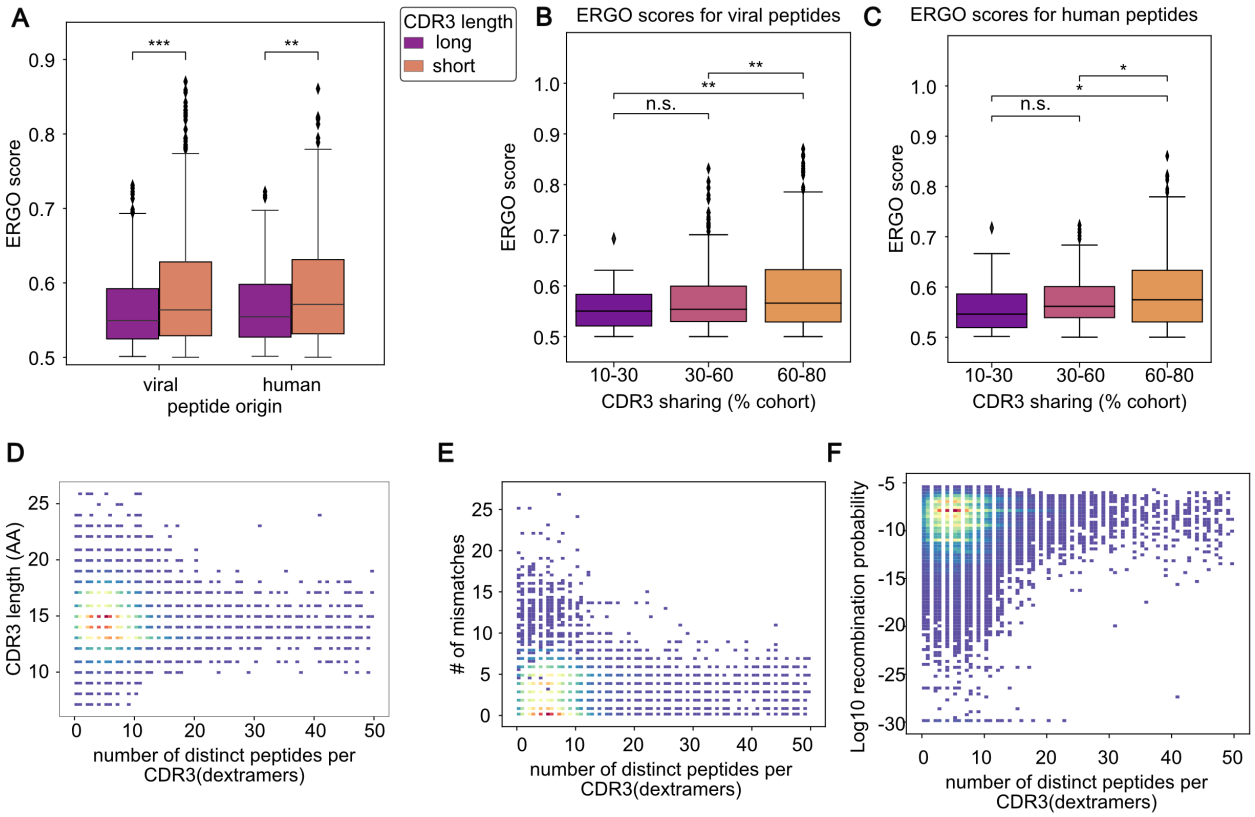


Fig. B.2. Supplementary Figure 2

Fig. B.2. Boxplots showing (A-C) public fraction percentages and (B-D) summed clonality of public CDR3aas in the Emerson cohort CMV+ and CMV- individuals. Boxplots showing (E-G) public fraction percentages and (F-H) summed clonality of public CDR3aas for naïve and effector-memory T cells in the Thome cohort.



G

Britanova cohort broad age groups vs sex

	Male	Female	Unknown
0 (Cord blood)	3	2	3
2-44 (Young)	14	10	0
45-64 (Adult)	6	10	0
65+ (Old Age)	11	20	0

Fig. B.3. Supplementary Figure 3

Fig. B.3. (A) ERGO scores were used to predict the binding of CDR3aa longer or shorter than 15 amino acids to MHC-associated peptides of viral and human origin. ERGO scores were used to assess the relationship between two features of CDR3aa from the Britanova cohort: publicness and polyreactivity to (B) viral and (C) human peptides. Polyreactivity vs. (D) CDR3aa length, (E) the number of mismatches, and (F) log10 recombination frequency in the 10x Genomics dataset. (G) Distribution of Britanova cohort individuals by sex and with broad age groups depicted in Figure 3.

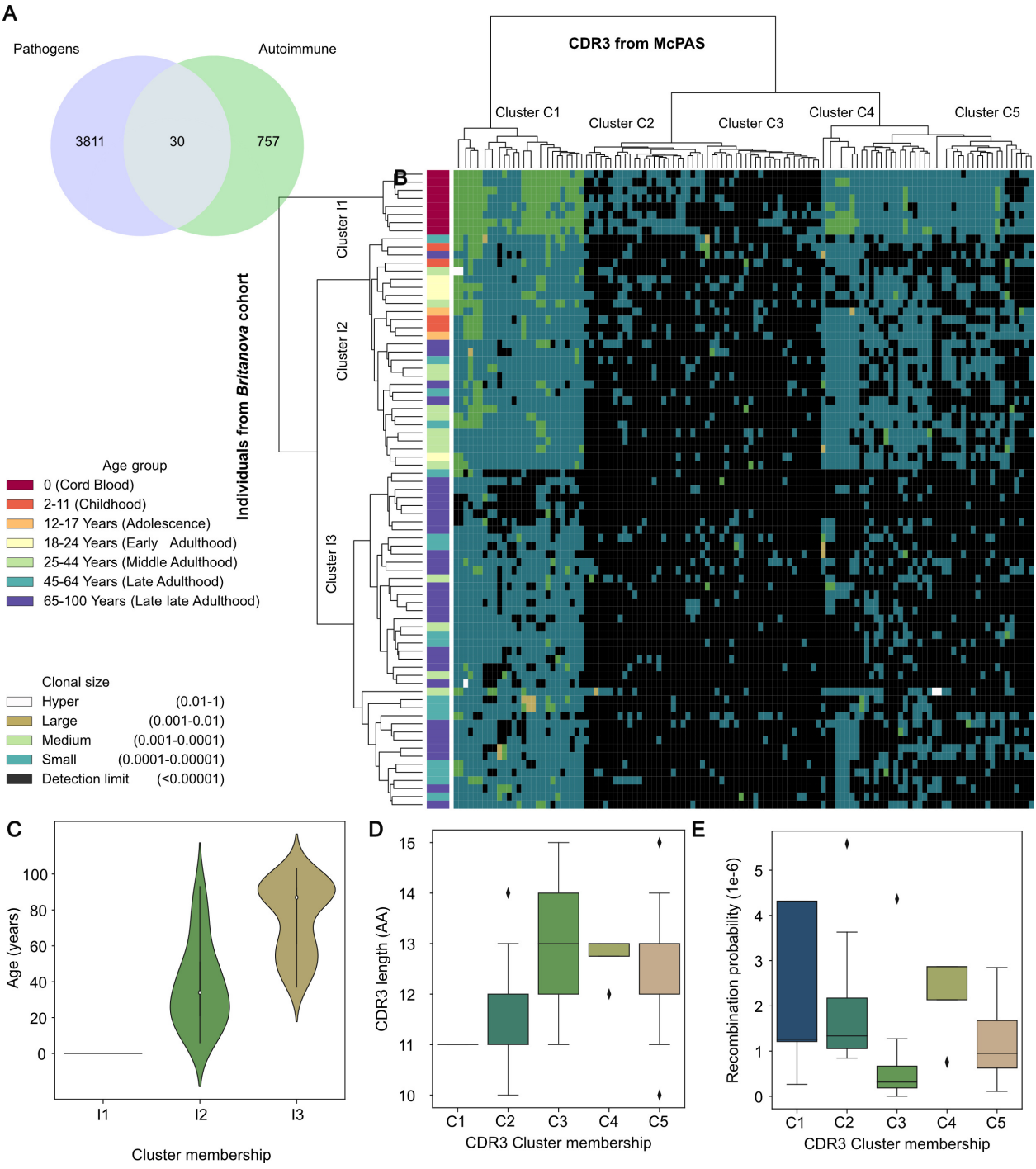


Fig. B.4. Supplementary Figure 4

Fig. B.4. (A) Venn diagram showing the overlap between two major CDR3 categories from McPAS: autoimmunity and microbial pathogens. (B) Heatmap shows, for subjects of the Britanova cohort, the frequency of CDR3aa listed in the McPAS autoimmune dataset. Rows represent individuals, columns unique CDR3aa, and cell color indicates CDR3aa clone size. Row dendrogram leaves are colored by age group. (C) Age distribution for subjects in three individual clusters from (A). (D) Boxplot showing CDR3aa length in five clusters from (B). (E) Boxplot showing predicted recombination frequency for CDR3aa in five clusters from (B).

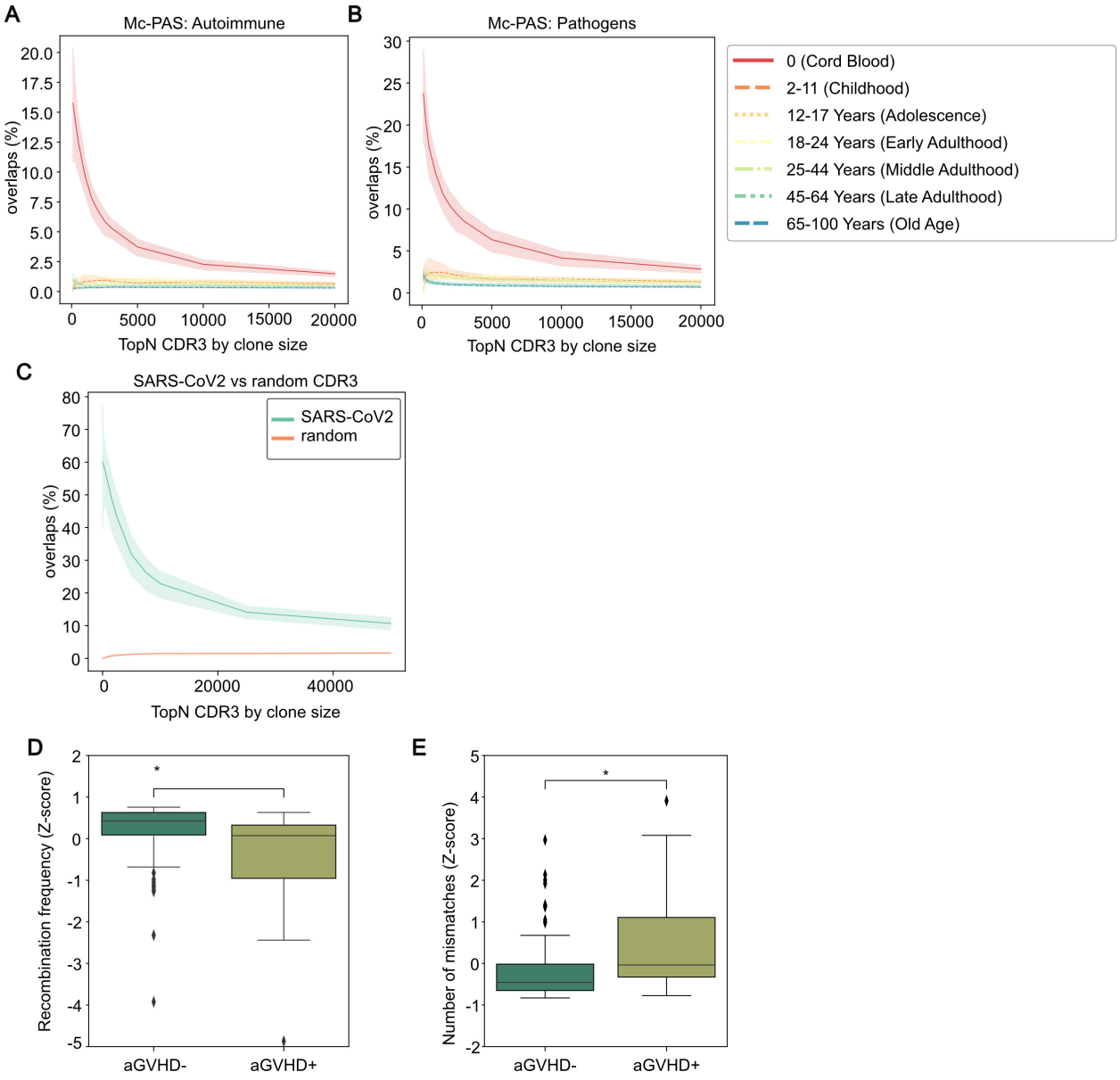


Fig. B.5. Supplementary Figure 5

Fig. B.5. Line plots showing percentage overlaps with McPAS (A) Pathogen-specific and (B) Autoimmune-specific CDR3 set, with the top N most frequent CDR3aa. Line colors and types correspond to age groups. (C) SARS-CoV2-specific CDR3 overlaps vs. a randomly selected set of CDR3 in cord blood. Boxplots showing (D) the recombination frequency and (E) the number of mismatches for CDR3aa in aGVHD+ vs. aGVHD- donors.

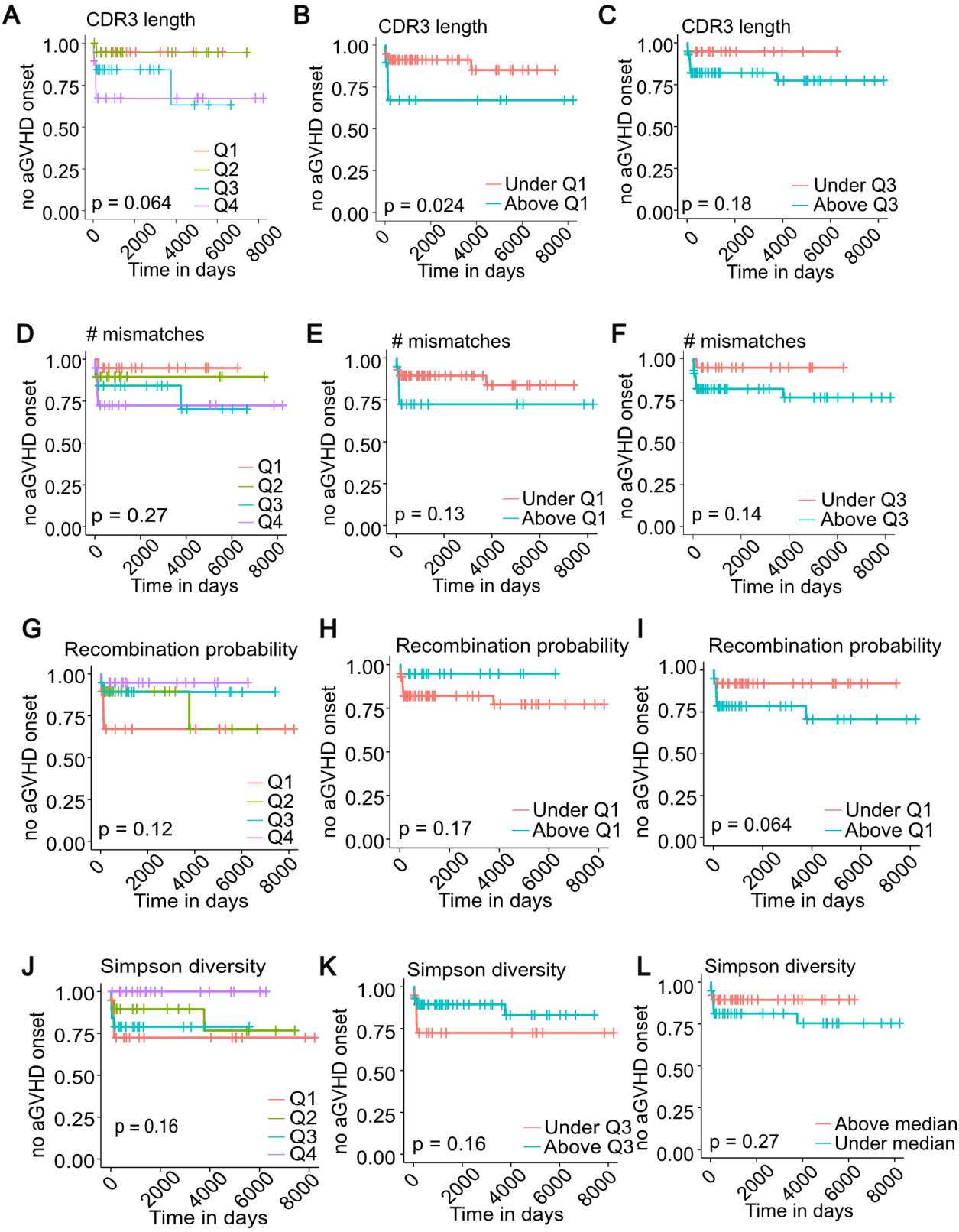


Fig. B.6. Supplementary Figure 6

Fig. B.6. Kaplan-Meier plots showing splits by all four and individual quartiles for (A-C) CDR3 length, (D-F) the number of mismatches, (G-I) recombination probability, and (J-L) Simpson diversity. This figure complements the results from Figure 6.7.

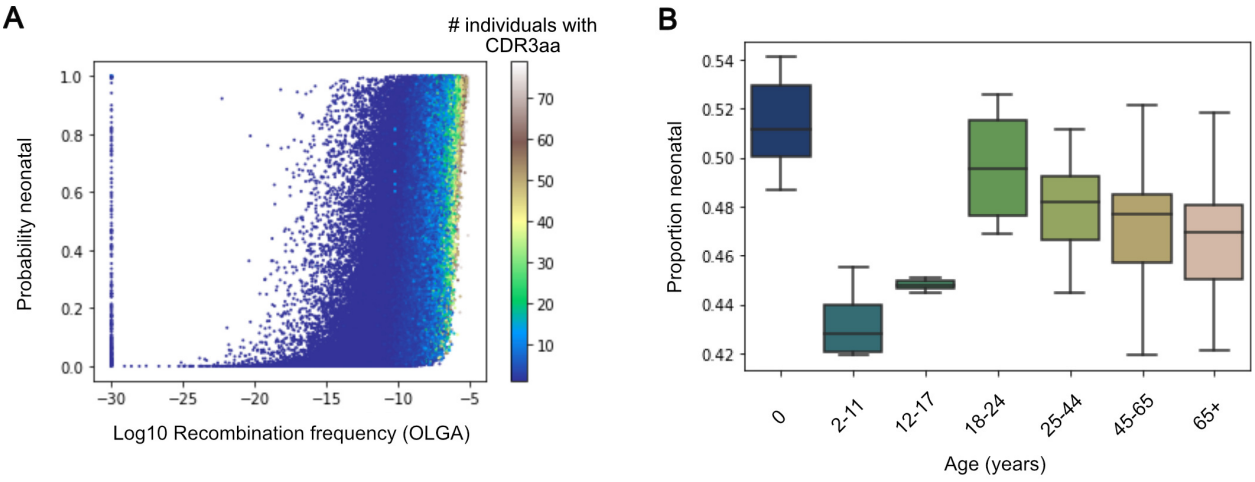


Fig. B.7. Supplementary Figure 7

Fig. B.7. (A) Scatterplot showing the recombination frequency for CDR3s based on the classification probability output of the logistic regression model (neonatal vs. TDT-dependent). Each dot is a CDR3s, and the color scheme represents the degree of sharing through the cohort. (B) The proportion of neonatal CDR3s by age group in the Britanova cohort.