Université de Montréal

Étude des signatures géniques dans un contexte d'expériences de RNA- Seq

par
Assya Trofimov

Département de biochimie et médecine moléculaire
Faculté de Médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en bio-informatique

Août, 2017

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Étude des signatures géniques dans un contexte d'expériences de RNA- Seq

présenté par:

Assya Trofimov

Mémoire accepté le: . . . . . .

RÉSUMÉ

Le principal intérêt des expériences de séquençage d'ARN (RNA-Seq) est qu'elles constituent une vue d'ensemble sur les procédés géniques intrinsèques de la cellule. L'état malade différe de l'état sain de par son usage génique et de nombreux efforts ont été canalisés dans les dernières années en bioinformatique, pour affiner ces signatures géniques, notamment dans la classification de leucémies et le typage de cancers du sein. Tous ces modèles voient, cependant, leur performance détériorée par un grand nombre de dimensions d'entrée et la plupart des auteurs choisissent d'imposer un seuil d'exclusion de gènes. J'ai voulu déterminer la nature d'une signature génique et sa taille optimale, en nombre de gènes. Pour déterminer la taille d'une signature génique j'ai appliqué des algorithmes de co-partitionnements à un sous-ensemble de données transcriptomiques afin d'en extraire la signature génique. Mes résultats indiquent que la signature génique ne peut être extraite en entier et l'utilisation de seuils d'exclusions de gènes est le principal problème. J'ai exploré une méthode d'extraction de la signature génique avec un réseau de neurones artificiels (ANN) en calculant le plus petit ajustement en expression génique nécessaire pour passer d'un phénotype à un autre. La signature génique extraite indique que presque la totalité des gènes sont affectés pour un phénotype donné. Conséquemment, il est inapproprié de considérer des méthodes avec seuil d'exclusion de gènes et je propose que les signatures géniques sont des phénomènes omnigéniques. Afin de pallier à l'inconvénient dû à la nécessité d'inclure tous les gènes dans l'analyse, j'ai élaboré une méthode d'apprentissage machine par ANN qui gère simultanément deux espaces : l'espace des gènes et l'espace des échantillons. Les coordonnées des gènes et des échantillons dans leur espaces respectifs sont arrangés de manière à ce qu'ils prédisent l'expression génique. Ma contribution est donc un modèle qui apprend de manière simultanée les interactions entre les gènes et les interactions entre les échantillons. Ma méthode permet également d'inclure dans l'analyse de jeux de données partiellement manquantes, faisant le lien vers l'intégration de données et l'analyses d'échantillons de séquençage de cellule unique (scRNA-Seq).

Mots clés : Apprentissage machine, réduction de dimensionnalité, transcriptome, RNA-Seq.

ABSTRACT

The main appeal of RNA sequencing experiments is that they offer a general view of all cell's intrinsic genetic processes. Diseased state differs from healthy by it's gene usage and many efforts have been channeled in bioinformatics these last few years to purify these gene signatures, in particular in the classification of leukemia and breast cancer subtyping. However, these models see their performance hindered by a large size of input dimensions and most authors chose to impose a threshold of gene exclusion. I wanted to determine what is a gene signature and how many genes it truly contains. To determine it's size, I applied co-clustering algorithms to a subset of transcriptomic data, to extract it's gene signature. My results indicate that the gene signature cannot be extracted entirely and the use of exclusion thresholds is the main problem. I then explored a gene signature extraction method using an artificial neural net (ANN), by calculating the smallest adjustment in gene expression necessary to go from one phenotypic class to another. The extracted gene signature indicated that almost all genes are affected for the given phenotype. Consequently, it seems inappropriate to consider threshold-based methods and I, therefore, propose that gene signatures are omnigenic phenomena. To level the disadvantage of having to include all genes in gene expression analyses, I designed a ANN method that simultaneously manages two spaces: the gene and the sample space. The coordinates for genes and samples in their respective space are arranged to predict the gene expression. My contribution is a model that learns simultaneously about genes and samples. My method allows the analysis of datasets with missing data, making the integration of heterogenous data integration as well as the analysis of single-cell RNA-Seq experiments.

Keywords: machine learning, artificial neural networks, dimensionnality reduction

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

LISTE DES SIGLES

ALL	leucémie aigue lymphoïde
AML	leucémie aigue myeloïde
ANN	Réseau de neurones artificiels
ADN	acide désoxyribonucléique
ARN	acide ribonucléique
FAB	French American British
GTEx	Genotype-tissue Expression Project
GO terms	Gene Ontology terms
IP	immunoprotéasome
KEGG	Kyoto Encyclopedia of Genes and Genomes
MSE	Erreur des moyennes carrées
NMF	Non-negative Matrix Factorization
PCA	Analyse des Composantes Principales
RNA-Seq	séquençage ARN
TCGA	The Cancer Genome Atlas
t-SNE	t-Stochastic Neighborhood Embedding
VAE	Auto-encodeur variationel

À Raissa Moiseevna Brum, scientifique, pédiatre, poétesse, et arrière-grand-mère. Tu y étais jadis et j'y suis maintenant.

REMERCIEMENTS

Tout ceci a été possible grâce à quelques personnes exceptionelles.

J'aimerais remercier mon ('highly logical') directeur, Claude Perreault, pour son support, son optimisme et son esprit avant-gardiste. Vous m'avez appris "To Boldly Go Where No Man Has Gone Before".

Merci à mon co-directeur, Sébastien Lemieux, pour sa pensée "out of the box" et son audace en recherche. Tu ne cesses de m'apprendre des nouvelles choses sur le métier de chercheur et l'informatique. L'escalade aussi.

Je voudrais exprimer ma plus profonde gratitude à Philippe Brouillard, pour avoir choisi sans hésitation de prendre part à mon aventure, à m'offrir ton support et ton coeur, à ensoleiller mes journées et partager tes projets, tes hobbies et ta vie avec moi.

J'aimerais remercier ma famille, pour leurs encouragements, leur amour inconditionel et leur foi en mes capacités. Merci Lavy, pour ton honnêté sans faille, à mes parents, qui m'ont appris à naviguer la vie adulte.

J'aimerais également remercier tous les membres des laboratoires Perreault et Lemieux, présents ou partis, pour leur bonne humeur, leur écoute, et tous ces diners en bonne compagnie.

Merci Pat, d'avoir toujours été là pour ton support quand je me chicanais avec mon ordinateur. Merci JP pour tes magouilles de programmation utiles, merci Eric pour avoir toujours questionné ce que je fais et m'offrir ton opinion franche. Merci à Geneviève, pour partager ton stash de chocolat en moment de panique (!). Merci à Jonathan pour tes hochements de tête silencieux et ton écoute active.

Finalement je suis reconnaissante aux gouvernements Canadien et Québécois, pour m'avoir attribué divers bourses tout au long de ma carrière et avoir tant contribué à ma formation.

CHAPITRE 1

INTRODUCTION

1.1 Mise en contexte

Les maladies complexes, telles le cancer, affectent les niveaux d'expression génique dans les cellules. Ayant un métabolisme différent de la cellule saine, la cellule tumorale altère la production de plusieures protéines, ce qui engendre une différence dans les niveaux d'expression génique observable et quantifiable. Comprendre le lien entre la variation de l'expression génique et la maladie est un des buts centraux de la transcriptomique. Ceci ne fait que souligner l'importance d'acquérir une méthode permettant d'étudier cette variation de l'expression génique, la signature génique (*gene signature*). De nombreux efforts ont donc été canalisés pour développer des modèles prédictifs, pouvant extraire ces signatures géniques. Typiquement, les signatures géniques (ici se limitant à un nombre prédéterminé de gènes) sont utilisées à des fins diagnostiques ou d'inférence du traitement approprié, tels la classification de leucémie [37] ou le typage de cancers du sein [9, 25, 112].

D'autres, plutôt que de se concentrer sur une poignée de gènes, ont préféré grouper les gènes en fonctions cellulaires, et plutôt examiner la différence de procédés cellulaires, tout en laissant le nombre de gènes impliqués libre [87, 95]. Dans le début des années 2000, il y a eu maints efforts de cataloguer les gènes et les lier à des fonctions cellulaires de plus haut niveau ; par exemple, le Kyoto Encyclopedia of Genes and Genomes (KEGG, [76]) et les Gene Ontology terms (GO terms, [1]). Des signatures géniques basées sur des fonctions cellulaires plutôt que directement sur les gènes ont été élaborées, cependant, malgré ces efforts, la reproductibilité des signatures géniques extraites aujourd'hui reste basse [10].

1.2 La signature génique

La signature génique est définie comme une altération d'expression génique simple (un gène) ou combinée (plusieurs gènes), spécifique à un phénotype observé [15]. La première signature génique en leucémie était celle détaillant la différence entre la leucémie myéloide aiguë et la leucémie lymphoide aiguë, par Golub et collègues [37]. Dans cet article, les auteurs ont utilisé 38 profils transcriptomiques afin de bâtir une signature génique permettant de distinguer entre une leucémie aiguë myeloïde (AML) et lymphoide (ALL). Ils ont trié les 6817 gènes quantifiés en ordre de corrélation avec le phénotype, et ils ont ensuite choisi 50 gènes les plus corrélés [37]. Ce choix n'est pas expliqué, surtout que les auteurs ont rapporté avoir trouvé au moins 1100 gènes dits importants pour la distinction entre l'AML et l'ALL [37]. Comme les choix des gènes n'est pas du tout justifié, un scénario où un autre groupe de gènes est choisi est donc envisageable. Ceci amène donc une précision obligatoire de la définition de la signature génique. En effet, bien qu'un petit nombre de gènes soient suffisant pour identifier un phénotype [9, 37, 112], ces gènes peuvent s'avérer non-suffisants dans une cohorte différente [10].

1.3 Hypothèse et objectifs

Notre hypothèse est que les maladies complexes sont en fait des phénomènes omnigéniques, c'est-à-dire, qu'il y a une répercussion sur l'ensemble des gènes. Basé sur cette hypothèse, j'ai répondu aux trois questions suivantes dans mon travail :

1. Qu'est-ce qu'une signature génique ?
2. Comment extraire une signature génique ?
3. S'il y a lieu, quelle est la taille (en nombres de gènes) d'une signature génique ?

Plus précisément, je compte i) déterminer ce qu'est une signature génique dans le contexte de maladie et dans le contexte d'expression génique différentielle (Section 3) et ii) explorer des méthodes d'extraction de la signature génique et d'analyses subséquentes de celle-ci (Section 4).

1.4 Organisation du mémoire

Pour faciliter la compréhension du contenu du mémoire, le chapitre 2 constient une revue de la méthodologie pertinente en transcriptomique (Section 2.1) et apprentissage machine (Section 2.3).

Le chapitre 3 contient un article qui présente l'utilisation de divers algorithmes de partitionnement et de réduction de dimensionnalité dans une étude de la leucémie aigue myeloïde (AML) et a été publié dans la revue *Scientific Reports*. Dans cet article nous avons tenté, avec mes co-auteurs, d'extraire une signature génique dans un sous-type d'AML, afin d'expliquer la vulnérabilité de celle-ci aux inhibiteurs d'immunoprotéasome. Cet article est donc l'étude d'une signature génique, utilisant les méthodologies standards.

J'explore une méthode plus précise pour extraire les signatures géniques absolues dans l'article contenu dans la section 4.1, à l'aide d'un réseau de neurones. Ce court article a été soumis à un workshop à la conférence *Neural Information Processing Systems (NIPS) 2016* et a été sélectionné par le comité de lecture pour une présentation par affiche. Cet article tente de répondre à la question sur la taille de la signature génique : combien de gènes sont impliqués ?

Finalement, la section 4.6 est dédiée au troisième article, qui propose une méthode de réduction de dimensionnalité comme méthode d'intégration de données biologiques, tout en extrayant simultanément de l'information sur les gènes et les échantillons à l'étude. Ce court article a été soumis à un workshop à la conférence *International Conference for Machine Learning (ICML) 2017* et a été sélectionné par le comité de lecture pour une présentation orale et par affiche et a également obtenu un prix de voyage.

Le chapitre 5 contient la discussion combinant tous les résultats et observations faites au long des trois articles ainsi que des projets de travaux futurs que je compte entreprendre dans les prochainnes années.

CHAPITRE 2

REVUE DE LA LITTÉRATURE MÉTHODOLOGIQUE

2.1 Le traitement des données en RNA-Seq

Pour observer l'ensemble des procédés d'une cellule, le séquençage d'ARN (RNA-Seq) est une méthode de choix. Cette méthode consiste en la purification de l'ARN d'une cellule, suivi de son séquençage. Il y a plusieurs étapes de pré-traitement post-séquençage effectuées lors de l'analyse RNA-Seq. Les séquences brutes provenant du séquenceur sont alignées sur un génome de référence, afin d'identifier leur gène ou région génique de provenance [77]. Ceci ne peut être, malheureusement, fait avec des méthodes exactes et les outils d'alignement de séquences (aligneurs) utilisent des heuristiques pour estimer rapidement l'alignement le plus probable des séquences [18, 54, 58, 77]. Il existe cependant des séquences redondantes ou très spécifiques à l'individu, ce qui rend la tâche des algorithmes d'alignement plus difficile. Des différences dans l'alignement sont rapportées dépendant de l'aligneur utilisé et la façon que ce dernier gère les cas spéciaux [86]. Puis, les séquences sont quantifiées, afin d'obtenir un niveau d'expression relatif pour chaque gène. Ici, tout dépend de la façon dont les gènes sont annotés ; en effet, les positions de départ et de fin du gène sont utilisés, et donc une annotation différente mène à une quantification différente [77, 86]. Une fois la quantification faite, les expressions géniques des échantillons analysés sont comparées entre elles. Il faut noter qu'une grande partie de l'information est mise de côté ici, notamment l'information sur les variants (mutations, réarrangements, épissage alternatif). De plus, si la séquence à aligner ne fait pas partie des gènes ou isoformes contenus dans l'annotation utilisée, cette séquence ne sera pas comptabilisée. Ceci constitue l'approche classique du traitement des données RNA-Seq et demeure aveugle aux séquences non-annotées, mutations et variants de gènes [18, 86]. D'autres méthodes plus sophistiquées existent mais ne seront pas discutées dans le cadre de ce mémoire. Un exemple d'analyse transcriptomique utilisant cette méthode standard est contenue dans le chapitre 3.

2.2 Les représentations clairsemées en bioinformatique

La malédiction de la dimensionnalité (également appelée "fléau de la dimension") est un phénomène qui s'observe dans les jeux de données où le nombre de caractéristiques pour chaque exemple est très grand [6]. Les données provenant d'expériences RNA-Seq sont, par leur taille, ce type de données. En effet, chaque échantillon produit entre 10^4 et 10^5 valeurs d'expressions et donc le nombre de caractéristiques par échantillon surpasse largement le nombre total d'échantillons.

Une solution potentielle au problème de haute dimensionnalité sont les représentations dites clairsemées (*sparse*). Les données sont donc réduites à un nombre plus restreint de signaux principaux, un nombre souvent choisi d'avance. Cette "quête de la sparsité" en biologie a d'ailleurs une origine historique. En effet, avant que le séquençage à haut débit ne soit populaire, la plupart des analyses de mécanismes génétiques ne donnaient que quelques candidats. Une analyse sémantique de 28 articles pris au hasard dans 3 grands journaux immunologiques a révélé que 23 des 28 études suivent un modèle strict de "Facteur X interagit avec facteur Y dans le mécanisme Z" [8]. Cette mécanique d'analyse ne garantit, cependant, qu'un recouvrement partiel de la signature génique expliquant les effets observés.

2.3 L'apprentissage machine en biologie

L'apprentissage machine est une technique statistique et computationnelle qui élabore un modèle général expliquant les données. L'apprentissage machine est aujourd'hui exploité afin d'automatiser plusieurs tâches laborieuses souvent encore faites par les humains ("text mining", traitement de langue naturelle, classification d'images etc.). L'application de l'apprentissage machine aux données provenant d'expériences biologiques apporte un élément de complication non prévu dans l'apprentissage machine classique. En effet, dans les applications d'apprentissage machine classique, les prédictions peuvent être aisément vérifiées par l'humain. Dans certaines applications bioinformatiques, l'humain est incapable d'analyser directement ces données et émettre une prédiction ; c'est pourquoi les modèles statistiques sont construits. Ce type de problème est

également courant entre autres dans le domaine de la prédiction de phénomènes météorologiques et financiers. Il est donc attendu que ces modèles fournissent plus d'information que ce que l'humain peut extraire et souvent pour une question dont la vraie réponse est inconnue. Ceci ajoute une dimension de complexité aux problèmes d'apprentissage machine en biologie, vu que la vérification humaine directe dans certains cas n'est pas possible.

2.4 L'apprentissage

Les algorithmes d'apprentissage se divisent en deux grandes catégories : i) l'apprentissage supervisé et ii) l'apprentissage non-supervisé. L'algorithme d'apprentissage bâtit et ajuste un modèle statistique, ajusté aux données, et la distinction entre les types d'apprentissages se fait surtout sur le type de modèle qui en résulte et le type d'informations dont l'algorithme dispose. Les sections suivantes offrent plus de détails sur les deux catégories d'apprentissage ainsi que des exemples d'applications.

2.5 Apprentissage supervisé et entrainement de modèle

L'apprentissage supervisé est utilisé lorsque les échantillons sont associés à des valeurs (ou étiquettes) connues. Souvent, les données d'entrée sont dénotés par un ensemble d'exemples ayant la forme

$$(x_1, x_2, ..., x_d, y)$$

où x_i sont des caractéristiques d'entrée et y est une étiquette ou une valeur à prédire. La nature de la valeur à prédire est discrète (contexte de classification) ou réelle (contexte de régression). Le modèle idéal peut être vu comme une certaine fonction f qui permet de passer de données X à une prédiction Y.

$$f : X \to Y$$

Cependant, comme la fonction f idéale est inconnue, la fonction $f(X)$ qui sera élaborée par le modèle, appelons la dorénavant $\hat{f}(X)$, ne sera qu'une approximation de cette vraie fonction $f(X)$.

Pour bien choisir la fonction $\hat{f}$, la prédiction $\hat{Y}$ émise par le modèle est comparée à la vraie valeur Y. Cette comparaison se fait via une fonction de coût pré-déterminée. Par exemple, si Y est un nombre réel, une fonction de coût L appropriée pourrait être une erreur quadratique : $L = (Y - \hat{Y})^2$. L'algorithme effectue ensuite des modifications à son modèle statistique, basé sur cette erreur. Une fois le modèle entrainé, sa capacité à généraliser est évaluée à l'aide d'un jeu de données externe X_{test}. Une basse erreur de prédiction sur le jeu de données X couplée à une haute erreur sur le jeu X_{test} est un indice de sur-apprentissage (faible généralisation), c'est-à-dire d'une trop haute spécialisation du modèle pour les données X.

L'apprentissage consiste donc à l'élaboration d'une fonction $\hat{f} : X \rightarrow Y$ qui i) performerait bien pour la tâche en cours selon la fonction de perte choisie et ii) généraliserait à un nouveau jeu de données X_{test} puisé de la même population.

2.6 Apprentissage non-supervisé

L'apprentissage non-supervisé diffère de l'apprentissage supervisé par le fait qu'il n'y a pas d'étiquettes associées aux échantillons. Ces algorithmes ont pour but de découvrir une structure sous-jacente directement à partir des données d'entrée. Par exemple, l'algorithme de partitionnement en k-moyennes (k-means clustering) [61], utilise une mesure de distance entre les échantillons pour inférer l'appartenance à un sous-groupe. Deux sous-familles d'algorithmes d'apprentissage non-supervisé populaires en bioinformatique sont les algorithmes de partitionnement (*clustering*) et les réductions de dimensionnalité. Dans les prochaines sous-sections, une liste non-exhaustive d'algorithmes d'apprentissage non-supervisé seront introduits, se tenant aux algorithmes les plus utilisés en bioinformatique.

2.6.1 Algorithmes de partitionnement

Il est fréquent, lors d'expériences de séquençage, de suspecter d'avoir une certaine hétérogénéité dans la population à l'étude. Par exemple, en séquençant l'ARN provenant d'une cohorte de patients ayant le cancer, il serait intéressant de voir s'il y a des sous-groupes correspondant à des traits observés en clinique : par exemple, la survie ou non du patient, ou encore un sous-type de la maladie. Grouper les exemples est donc une solution intéressante. Les algorithmes de partitionnement ont tous une mécanique similaire. Ils regroupent les exemples similaires en partitions (*clusters*) utilisant une mesure de distance appropriée (choisie par l'utilisateur), souvent la distance euclidienne [75]. Les algorithmes de partitionnement sont très populaires en bioinformatique et sont utilisés dans plusieurs contextes, dont les analyses transcriptomiques et métabolomiques [45, 93, 104].

2.6.1.1 Le partitionnement de k-moyennes

Introduit pour la première fois en 1967, le partitionnement de k-moyennes agrège les échantillons selon une mesure de distance dans un nombre de groupes k prédéterminé par l'utilisateur. Tout d'abord, k points sont placés aléatoirement dans l'espace des données. Ces points constituent des centroïdes futurs des k groupes. Ensuite, itérativement, pour chaque échantillon, une distance aux centroïdes est calculée et chaque échantillon reçoit l'étiquette de groupe correspondant au centroïde le plus proche. Finalement, la position des centroïdes est actualisée en mesurant le point milieu pour chaque groupe, selon les échantillons contenus dans le groupe en question [61]. Cet algorithme est utilisé en analyses transcriptomiques, quand le nombre de groupes à obtenir est connu ou suspecté [81].

2.6.1.2 Le partitionnement hiérarchique

Le partitionnement hiérarchique diffère principalement du partitionnement k-moyennes, par le fait qu'il n'y a pas la notion de groupes prédetérminés. En effet, le partitionnement hiérarchique calcule plutôt une distance directe entre tous les points. Dans la première

itération de l'algorithme, les points appertiennement chacun à son propre groupe (*cluster*) et à chaque itération, des paires de *clusters* sont joints. Le calcul prend fin lorsque tous les points se retrouvent dans le même *cluster*.

Afin de joindre deux *clusters* A et B, qui sont des ensembles de points $A = \{a_1, ..., a_m\}$ et $B = \{b_1, ..., b_n\}$, une mesure de distance est introduite. Dans la variante de l'algorithme de partitionnement hiérarchique décrit ici (UPGMA [92]), une distance inter-*clusters* est définie par la moyenne entre les distances de chaque point à chaque point de l'autre cluster (équation 2.1).

$$dist(A,B) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} ||a_i - b_j||_2^2 \tag{2.1}$$

Ces liens construit entre les clusters peuvent être représentés sous forme de dendrogramme (exemple dans la section 3.3.6). Afin de séparer les points en *clusters*, c'est à l'utilisateur de déterminer où il compte couper l'arbre, créant ainsi un certain nombre de sous-arbres, qui sont les *clusters* [46].

2.6.1.3 Le co-partitionnement

Le terme "co-partitionnement" suggère un algorithme où plusieurs partitionnements ont lieu simultanément et que ces multiples partitionnements sont liés. Le partitionnement de Potts, développé pour des problèmes dans le domaine de la physique [113], permet ce type de manipulation.

Le co-partitionnement se fait en deux étapes. La première étape consiste à construire deux réseaux relationels sur une base quelconque (distance, corrélation etc.). Ces deux réseaux sont ensuite alignés par une autre mesure. Par exemple, une équipe a implémenté ce type d'algorithme pour trouver des gènes orthologues entre deux espèces [115]. Donc les réseaux de gènes sont construits pour chaque espèce et ensuite les orthologues connus sont joints entre les deux réseaux. Le partitionnement se fait par recuit simulé ou "simulated annealing", où à chaque étape, l'énergie du système ; la température (ou entropie) du système est progressivement baissée, via le paramétre de température, permettant le partitionnement [113, 115].

Ceci fait référence à un modèle de vibration moléculaire, où les mouvement des atomes des molécules est plus rapide à plus haute température. Lorsque la température est baissée, les atomes perdent de la mobilité et adoptent une conformation stable. De façon similaire, le réseau relationnel trouve sa conformation stable lorsque la température du système baisse.

2.6.2 Les algorithmes de réduction de dimensionnalité

Un autre type d'apprentissage non-supervisé est la réduction de dimensionnalité. Cette grande famille de méthodes regroupe plus d'une dizaine d'algorithmes qui ont pour but de réduire, moyennant quelques pertes, le nombre de dimensions des données d'entrée. À la différence des algorithmes de partitionnement, ces modèles permettent d'extraire les traits principaux ou les plus importants (selon une mesure spécifique) d'un jeu de données. C'est-à-dire, l'algorithme crée une fonction f, où les données d'entrée x, ayant un nombre de dimensions d sont réduites (ou encodées) à un espace plus restreint de taille d', soit $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$, où $d' < d$. Bien que plusieurs algorithmes ont été développés, uniquement les plus utilisés en biologie seront détaillés. Un type de réseau de neurones permettant la réduction de dimensionnalité sera également détaillé.

2.6.2.1 L'analyse en composantes principales

Outil standard de l'analyse de données, cet algorithme crée un nouveau système de coordonnées pour les données d'entrée. Plus précisément, l'algorithme définit des nouveaux axes, permettant de séparer le plus possible les données. L'avantage principal de la méthode est qu'elle est rapide à calculer, plusieurs implémentations existent dans une multitude de langages de programmation. Les composantes sont ordonnancées par importance de variance. Souvent, à des fins de visualisation, que les deux premières dimensions sont utilisées. Le principal désavantage est que cette méthode a une contrainte d'orthogonalité sur les vecteurs et donc elle est très restrictive sur le type d'encodage qu'elle peut offrir [90, 103]. L'analyse de composantes principales est appliquée à des transcriptomes humains dans les chapitres 3 et 4 et sa performance est comparée à celle

des autres algorithmes de réduction de dimensionnalité.

2.6.2.2 La factorisation non-négative de matrices

La factorisation non-négative de matrices (NMF) est une méthode de décomposition par parties, où des données d'entrées (ici expressions géniques et échantillons) sont réduites à un plus petit nombre de méta-gènes [30]. Ceci correspond à la factorisation d'une matrice A de taille $N \times M$ en sous-matrices W et H de taille $N \times k$ et $k \times M$, où $k << M, N$. Ces deux matrices ont une contrainte de non-négativité et le nombre k correspond à un nombre (choisi) de meta-gènes [12]. Les deux sous-matrices sont initialisées à des nombres aléatoires et ensuite sont mises à jour itérativement, en maximisant la probabilité de générer A à partir de W et H [12]. Cette méthode a été employée également pour grouper des composés chimiques sur la base de leur protéines cibles [114] ainsi que dans la recherche d'orthologues [99]. Vu l'intérêt accru pour cette méthode, une équipe a également créé un outil d'analyse, bioNMF, permettant l'application de l'algorithme sans avoir à l'implémenter [78]. Malgré le grand intérêt pour la méthode, elle est tout de même très dépendante du k choisi. En effet, vu la variabilité des résultats pour une même expérience, Brunet et collègues ont même développé une stratégie de sélection de modèle, afin de garantir la meilleure performance [12]. Dans la section 4.6, l'algorithme de NMF est comparé à plusieurs autres algorithmes de réduction de dimensionnalité.

2.6.2.3 t-SNE : embedding stochastique du voisinage

Le *t-stochastic neighborhood embedding*, encodage de voisinage stochastique à distribution de t en français, est une méthode non-linéaire de réduction de dimensionnalité [102]. Cette méthode modélise des données de haute dimensionnalité dans un système de coordonnées de 2 ou 3 dimensions. Pour ce faire, une probablitié conditionelle est calculée pour chaque paire de points, c'est-à-dire, des points similaires ont une haute probabilité conditionelle et des points dissimilaires en ont une basse. Ensuite, le méme type de distribution est construit en basse dimension (2 ou 3) [102]. La divergence Kullback-Leibler est minimisée entre ces deux distributions et l'encodage en basse dimension est

rapporté [102, 103]. La section 4.6 contient une application de l'algorithme de t-SNE à des transcriptomes de tissus humains.

2.6.2.4 Réseaux de neurones artificiels

S'inspirant des réseaux neuronaux réels, l'approche connexioniste se base sur l'hypothèse que des propriétés nouvelles pourraient émerger d'un empilement d'entités (ici neurones). En effet, les réseaux de neurones artificiels (ANN) comportent plusieurs unités computationelles, qui sont la plupart du temps organisés en couches. Chaque couche peut être représentée par une fonction linéaire $h = Wx + b$, prennant en entrée un vecteur de valeurs x, le multipliant par une matrice de poids W et lui additionnant un vecteur de constantes (biais) b [105]. Une fonction non-linéaire, appellée également *fonction d'activation*, est habituellement appliquée à ces valeurs transformées, où la tangente hyperbolique et la sigmoide seraient des exemples classiques. Ces couches de transformations linéaires et de non-linéarités peuvent être empilées en grand nombre, d'où l'appellation "apprentissage profond". La capacité à apprendre des représentations plus complexes augmente avec le nombre de couches et cette profondeur rend obsolète la pré-sélection des caractéristiques (*feature engineering*) des données fournies au réseaux [38]. Les poids et biais du réseau total constituent ses paramètres et ces derniers sont optimisés durant l'entrainement, souvent par descente de gradient.

L'auto-encodeur est une architecture de réseau de neurones artificiels, qui tente de reconstruire les données d'entrée x, suite à un encodage (sur une couche de neurones cachée h). Il est typiquement pénalisé sur une distance entre les données d'entrée et les données reconstruites [105]. L'auto-encodeur a souvent un nombre plus petit de neurones artificiels sur la couche cachée que la taille des données d'entrée, d'où son appartenance à la classe des algorithmes de réduction de dimensionnalité. Cette architecture a également plusieurs variations, tels les auto-encodeurs débruiteurs (denoising autoencoder) et les auto-encodeurs variationnels (variational autoencoder, VAE).

L'auto-encodeur variationnel a pour particularité d'être à la fois un modèle de réduction de dimensionnalité et un modèle générateur (*generative model*). Dans ce modèle, l'encodeur apprend pour chaque point les paramètres d'une gaussienne de dimension-

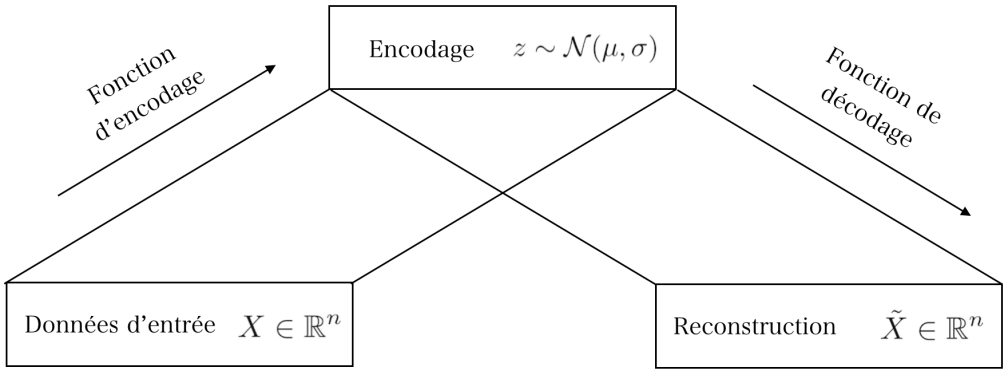


Figure 2.1 : Auto-encodeur Variationnel

nalité pré-determinée, souvent plus petite que la dimensionnalité des données d'entrée [26]. La reconstruction se fait à partir des points dans l'espace d'encodage (espace latent) et le réseau est pénalisé sur une erreur de reconstruction ainsi qu'une probabilité conditionelle du point z sachant les données d'entrée X (Figure 2.1). La section 4.1 contient un exemple d'utilisation d'un auto-encodeur variationnel pour extraire un encodage d'échantillons.

L'intérêt principal des auto-encodeurs est que chaque point dans l'espace d'encodage est 'décodable' en dimensions d'entrées et que l'espace d'encodage est directement accessible via la fonction d'encodage.

2.7 L'espace d'encodage et les *embeddings*

La plupart des algorithmes de réduction de dimensionnalité basés sur les réseaux de neurones contenus dans la section 2.6.2.4 permettent l'accès à l'utilisateur à la fonction d'encodage des données d'entrée. Il a été observé que les réseaux de neurones, durant l'apprentissage, sous pression de la fonction de coût, aggrégent les données semblables, en leur donnant des coordonnées similaires dans l'espace latent. Dans les dernières années, plusieurs ont entrepris d'examiner ces espaces d'encodage afin de mieux comprendre les structures sous-jacentes des données à l'étude. En effet, Salakhutdinov et collègue, a encodé 20000 articles de nouvelles utilisant un RBM (Restricted Boltzmann Machine) et rapportent que les articles de nouvelles se groupent selon des thèmes simi-

laires [88].

Les équipes de Mikolov et de Pennington ont quant à eux créé un espace d'encodage où chaque mot à sa propre coordonnée (*embeddings*) et des mots ayant un sens similaire se retrouvent groupés dans l'espace [67, 79]. Une examination plus poussée de l'espace montre que des relations sémantiques entre les mots sont conservées. En effet, les *embeddings* de mots se prêtent à des transformations arithmétiques ; un exemple célèbre le met en évidence : dans l'espace d'encodage, les coordonnées respectives pour 'Roi' + 'femme' - 'homme' donne les coordonnées pour le mot 'Reine' [79]. De plus, la distance relative entre les coordonnées pour 'Paris' et 'France' sont presque les mêmes que ceux pour 'Rome' et 'Italie' [67]. Cette propriété de l'espace d'encodage des réseaux de neurones les rend d'autant plus intéressantes pour l'analyse de transcriptomes.

Dans la section 4.6, des *embeddings* pour des gènes et des patients sont crées et les relations entre les coordonnées sont examinées de plus près.

CHAPITRE 3

DANS LES CANCERS HUMAINS, L'EXPRESSION DES GÈNES DE L'IMMUNOPROTÉASOME EST RÉGULÉE PAR DES FACTEURS INTRINSÈQUES ET EXTRINSÈQUES DE LA CELLULE

Contribution de l'étudiant

Ma contribution à l'article c'est fait dans la planification, conception et réalisation des analyses bioinformatiques.

Je me suis occupée de télécharger et traiter les données de RNA-Seq et les données cliniques, ainsi que d'adapter un module d'implémentation de l'algorithme de *co-clustering* à nos données. J'ai également implémenté un algorithme effectuant des test de Fisher multiples, afin d'effectuer l'enrichissement par GO terms du résultat du *co-clustering*.

J'ai effectué la majorité des analyses statistiques contenues dans cet article, sauf les études de survie. J'ai également pris part à l'interprétation des données et à rédaction du manuscrit et la création des figures.

Expression of immunoproteasome genes is regulated by cell-intrinsic and -extrinsic factors in human cancers

Alexandre Rouette[1,2,6], Assya Trofimov[1,2,3,6], David Haberl[1], Geneviève Boucher[1], Vincent-Philippe Lavallée[1,2,4,5], Giovanni D'Angelo[4,5], Josée Hébert[1,2,4,5], Guy Sauvageau[1,2,4,5], Sébastien Lemieux[1,3] and Claude Perreault[1,2,4*]

[1]Institute for Research in Immunology and Cancer; [2]Department of Medicine and [3]Department of Computer Science and Operations Research, Université de Montréal; [4]Division of Hematology-Oncology and [5]Quebec Leukemia Cell Bank, Maisonneuve-Rosemont Hospital, Montreal, Quebec, Canada;

[6]These authors contributed equally to this work

[*]CORRESPONDENCE : Claude Perreault (claude.perreault@umontreal.ca), Institute for Research in Immunology and Cancer, P.O. Box 6128, Station Centre-Ville, Montreal, QC, Canada, H3C 3J7. Phone : 514-343-6126; Fax : 514-343-5839

Article publié dans Scientific Reports, le 23 septembre 2016 [87]

Le texte de l'article a été modifié de sa version originale afin d'accomoder les exigences de présentation du mémoire. Les références de cet articles sont incluses dans les références totales du mémoire et les numéros de figures portent le numéro de chapitre.

3.1 Abstract

Based on transcriptomic analyses of thousands of samples from The Cancer Genome Atlas, we report that expression of constitutive proteasome (CP) genes (PSMB5, PSMB6, PSMB7) and immunoproteasome (IP) genes (PSMB8, PSMB9, PSMB10) is increased in most cancer types. In breast cancer, expression of IP genes was determined by the abundance of tumor infiltrating lymphocytes and high expression of IP genes was associated with longer survival. In contrast, IP upregulation in acute myeloid leukemia (AML) was a cell-intrinsic feature that was not associated with longer survival. Expression of IP genes in AML was IFN-independent, correlated with the methylation status of IP genes, and was particularly high in AML with an M5 phenotype and/or MLL rearrangement. Notably, PSMB8 inhibition led to accumulation of polyubiquitinated proteins and cell death in IP^{high} but not IP^{low} AML cells. Co-clustering analysis revealed that genes correlated with IP subunits in non-M5 AMLs were primarily implicated in immune processes. However, in M5 AML, IP genes were primarily co-regulated with genes involved in cell metabolism and proliferation, mitochondrial activity and stress responses. We conclude that M5 AML cells can upregulate IP genes in a cell-intrinsic manner in order to resist cell stress.

3.2 Introduction

All eukaryotes express constitutive proteasomes (CPs) that possess three catalytic subunits (PSMB5, PSMB6 and PSMB7). In addition to CPs, vertebrates also express immunoproteasomes (IPs), in which the catalytic β-subunits are replaced by IFN-γ-inducible homologues : PSMB8 for PSMB5, PSMB9 for PSMB6 and PSMB10 for PSMB7[3]. The first non-redundant role ascribed to IPs was their enhanced ability to generate MHC I-associated peptides[23]. However, recent work has revealed that IPs can be expressed by non-immune cell[49, 110] and that differential cleavage of transcription factors by CPs and IPs has pleiotropic effects on cell function[22]. Indeed, CPs and IPs differentially modulate the abundance of transcription factors that regulate signaling pathways with prominent roles in cell differentiation, inflammation and neoplastic transformation (e.g., NF-kB, IFNs, STATs and Wnt)[22].

In cancer cells, genomic instability and oncogene addiction cause proteotoxic and oxidative stress[60]. Indeed, aneuploidy and variations in transcript levels produce imbalances in the stoichiometry of protein complexes and thereby lead to accumulation of misfolded proteins and formation of aggregates (proteotoxic stress)[74, 89, 106]. Moreover, oncogenic signaling and dysregulation of mitochondrial function generate reactive oxygen species which damage DNA and proteins (oxidative stress). Proteasomes are key players in stress response since they degrade damaged (misfolded or oxidized) proteins[31, 40, 65]. Accordingly, cancer cells are presumed to be unduly dependent on proteasomal function[108]. Besides, tumors are commonly infiltrated by IFN-γ-producing lymphocytes specific for neo-antigens[71], and IFN-γ directly upregulates IP genes1. Hence, several factors could influence the abundance of proteasomes in neoplastic cells.

The goal of our work was therefore to determine whether CPs and IPs were differentially expressed in normal vs. neoplastic human cells and whether the two types of proteasomes played non-redundant roles in cancer cells. Here we report that overexpression of proteasomes is present in a wide variety of cancer types. Differential expression of CP genes had no impact on survival. However, IP upregulation in breast cancer sho-

wed a strong correlation with the abundance of interferon-producing tumor infiltrating lymphocytes and was associated with a good prognosis. In contrast, IP upregulation in AML was a cell-intrinsic feature that was not associated with improved survival. IP expression was particularly high in AML with an M5 phenotype according to the French-American-British (FAB) classification or in AML with an MLL rearrangement. IP expression in AML correlated with the methylation status of IP genes, and specific IP inhibition led to accumulation of polyubiquitinated proteins and cell death in IP^{high} but not IP^{low} AML cells. We conclude that expression of IP genes in human cancers is regulated by cancer cell-extrinsic (IFN-γ) and -intrinsic (cell stress) factors. Furthermore, our work identifies a functional vulnerability in IP^{high} AML cells because of an undue sensitivity to treatment with an IP-specific inhibitor.

3.3 Results

3.3.1 Genes encoding proteasome catalytic subunits are overexpressed in several cancer types.

In order to evaluate the expression of proteasome catalytic subunits in cancer, we first downloaded RNA-Seq data from TCGA, along with clinical metadata, from the Cancer Genomics Hub (see Methods). The initial analysis covered primary samples from thirteen tumor types from eleven different tissues, with normal tissue controls available for eight cancer types (Figure 3.1). We analyzed the expression of the three CP- and the three IP-specific catalytic subunits. For the eight cancer types with available normal tissue controls, we found that a mean of five (out of six) proteasome catalytic subunits were slightly, but significantly, overexpressed in cancer samples (range 3-6) relative to normal tissue (Figure 3.1). We conclude that proteasome upregulation is a general feature of cancer tissues.

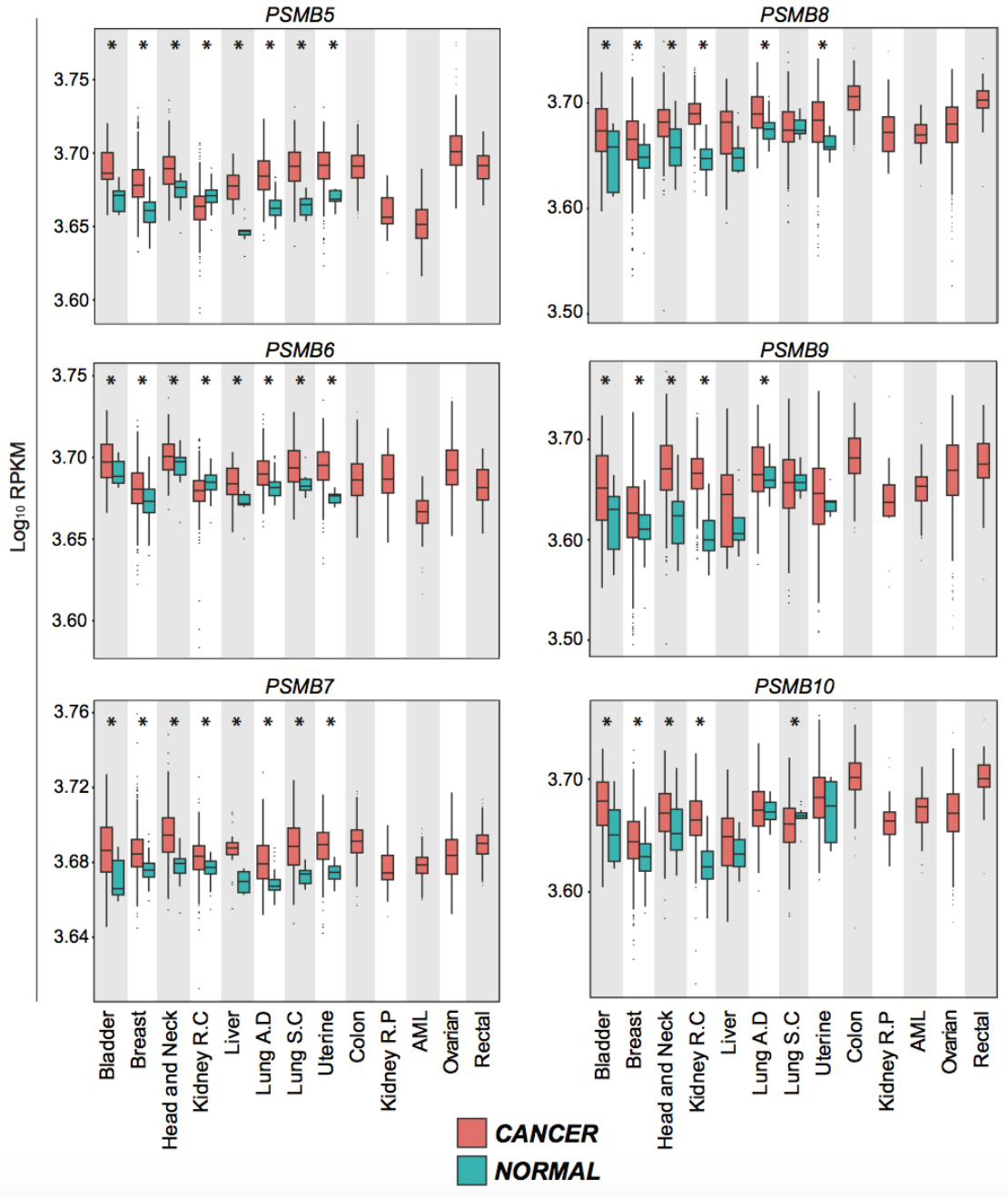


Figure 3.1 : Genes encoding proteasome catalytic subunits are overexpressed in several cancer types.

3.3.2 High expression of IP genes is associated with improved survival in breast cancer

We then sought to determine whether expression of CP- or IP-encoding genes correlated with survival in patients with various cancer types. For each patient in the TCGA cancer cohorts, expression of CP- or IP-encoding genes was transformed in z-score and summed. Based on this score, patient cohorts were separated in two or three groups of similar size (see Methods). This allowed us to evaluate the survival of patients with low or high expression of proteasome genes in their tumor sample. For most cancer types, expression of CP and IP genes showed no correlation with survival (Supplementary Fig. S1, Annexe I). However, IP gene expression did correlate with survival in breast cancer, as IP^{high} status was associated with a decreased risk of death (hazard ratio = 0.53 for 2 groups - Figure 3.2a and Table 3.I). Indeed, survival at ten years was 61.9%$\pm$ 11.7% for patients whose IP gene expression ranked in the top third of the cohort (IP^{high}) relative to 36.1%$\pm$8.0% for those in the bottom third (IP^{low}) (Figure 3.2 a). Furthermore, expression of individual IP genes PSMB8 and PSMB10 was associated with a decreased risk of death (Supplementary Table S1, Annexe I). However, expression of CP genes did not correlate with survival in breast cancer : i) high global expression of CP genes was not associated with better prognosis when the cohort was separated in two or three groups (Figure 3.2a), and ii) no individual CP gene was associated with prolonged survival (Supplementary Table S1, Annexe I).

Breast Cancer			
Category	# of groups	Hazard Ratio [95% Confidence Interval]	*p-value*
CP^{high} vs CP^{low}	2	1.30 [0.83-2.02]	0.245
CP^{high} vs CP^{low}	3	1.02 [0.78-1.33]	0.960
IP^{high} vs IP^{low}	2	0.53 [0.35-0.82]	0.003
IP^{high} vs IP^{low}	3	0.74 [0.56-0.97]	0.074
AML			
Category	# of groups	Hazard Ratio [95% Confidence Interval]	*p-value*
CP^{high} vs CP^{low}	2	1.35 [0.91 - 1.99]	0.133
CP^{high} vs CP^{low}	3	1.28 [0.79 - 2.08]	0.404
IP^{high} vs IP^{low}	2	1.42 [0.96 - 2.11]	0.076
IP^{high} vs IP^{low}	3	1.54 [0.95 - 2.51]	0.099

Tableau 3.I : Correlation entre le risque de décès et l'expression du protéasome

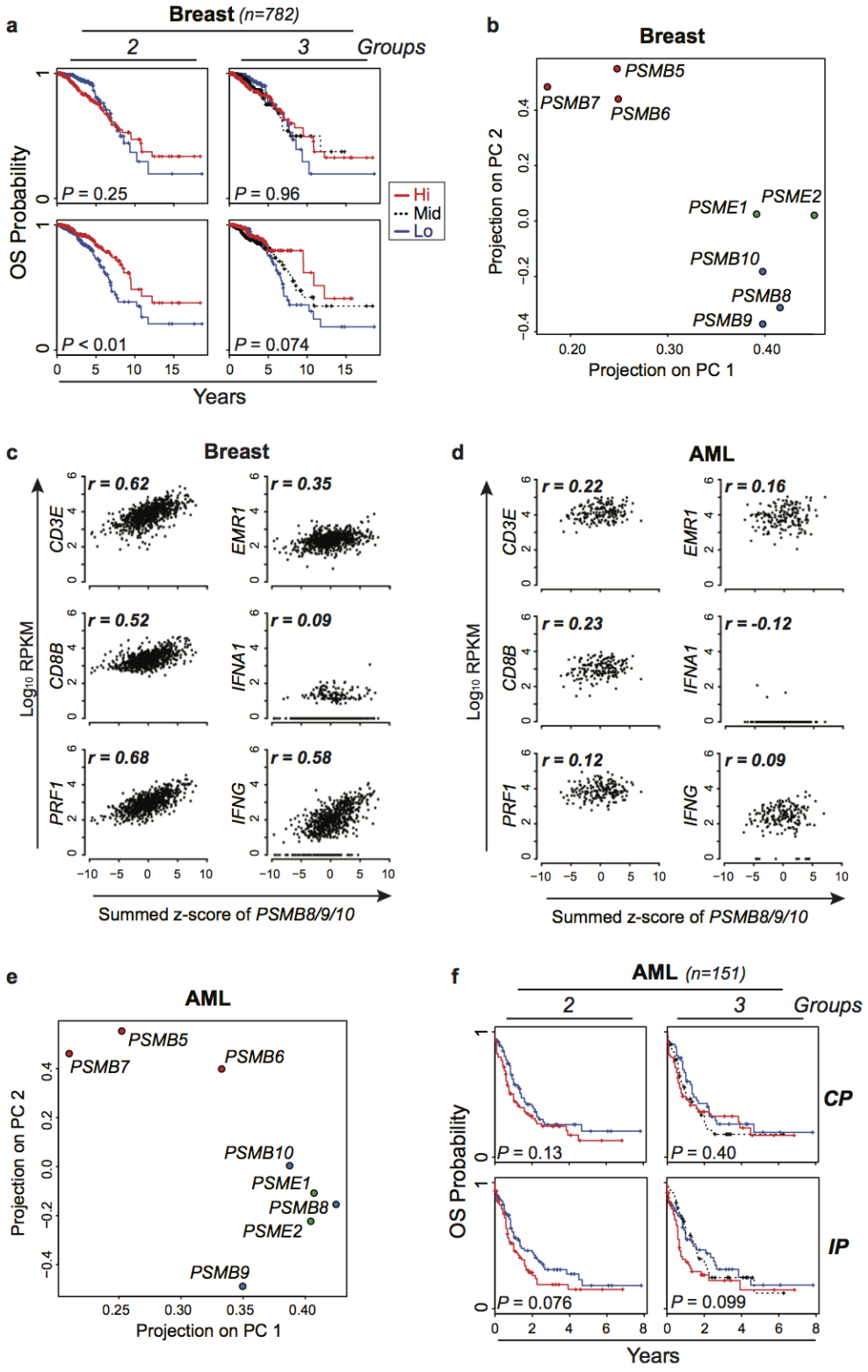


Figure 3.2 : Expression of IP subunits is cell-autonomous in AML.

3.3.3 IP subunits are co-expressed in breast cancer samples.

In normal cells, assembly of IPs is cooperative : the three catalytic subunits (PSMB8, 9 and 10) interact with each other to favor their common incorporation in homogenous IPs[41]. However, intermediate proteasomes, containing CP and IP subunits, can be assembled and display some unusual proteolytic cleavage preferences[42]. To assess whether CP and IP catalytic subunits were co-expressed in breast cancer samples, we performed a principal component analysis (PCA) on gene expression data for CP- and IP-encoding genes and regulatory subunits PA28α and PA28β (encoded by PSME1 and PSME2). PCA enriches for differences and variations by finding a rotation of the input data matrix that maximises the data variations in the first few dimensions. We found that, IP catalytic subunits clustered together with PSME1 and PSME2, apart from the CP subunits (Figure 3.2b). These results suggest that, like what is found in normal cells[110], expression of IP subunits occurs in a coordinated manner in breast cancer cells.

3.3.4 IP expression is cell-autonomous in AML but not in breast cancer.

Expression of IPs can be upregulated by cell autonomous signaling or via paracrine secretion of IFN-γ by surrounding NK cells and CD8 T lymphocytes[43]. This is particularly relevant in tumors where CD8+ tumor-infiltrating lymphocytes (TILs) secrete copious amounts of IFN-γ[66, 111]. We therefore asked whether IP expression correlated with the abundance of transcripts reflecting infiltration by CD8 TILs (CD3E, CD8, PRF1), macrophages (EMR1) and IFN secretion (IFNA1, IFNG). Based on RNA-Seq data from TCGA, expression of IP genes showed a strong correlation with expression of IFNG and T-cell genes in breast cancer (Figure 3.2c). Since infiltration by CD8 TILs is associated with a good prognosis in many cancer types (including breast cancer)[34], we surmise that the IP^{high} status in breast cancer is a marker of TIL infiltration and thereby correlates with prolonged survival (Figure 3.2a). IP expression also correlated with the expression of CD3E, CD8 and PRF1 in colon cancer, another form of solid tumor infiltrated by TILs[70] (data not shown).

Because infiltration by TILs has not been reported in hematologic malignancies such

as AML, we then studied how IP expression was regulated in AML. We found that expression of IP subunits was coordinated in AML, but showed no significant correlation with infiltration by TILs nor with abundance of IFN transcripts (Figure 3.2d-e). Furthermore, we observed a trend toward better prognosis in patients with IP^{low} relative to IP^{high} AML (Figure 3.2f, Table 3.I and Supplementary Table S1, Annexe I), but this trend did not reach statistical significance ($p = 0.07$ taking into consideration all AML subtypes except for acute promyelocytic leukemia - see Methods). Studies on additional patient cohorts will therefore be necessary in order to determine whether IP expression is a prognostic marker in AML. Nonetheless, these data reveal a clear dichotomy in IP regulation between breast cancer and AML, and beg the question : what is the nature of the cell autonomous (TIL-independent) signals that regulate IP levels in AML ?

3.3.5 IP subunits are highly expressed in myeloid and lymphoid cancer cell lines.

We reasoned that if IP upregulation is cell-intrinsic in AML but TIL-dependent in breast cancer, we should detect higher IP levels in AML cell lines than in breast cancer cell lines (since cell lines contain no TILs). The transcriptional portrait of 675 human cancer cell lines was recently reported[52]. When we interrogated this resource, we found that, as predicted, IP genes were expressed at higher levels in myeloid leukemias ($n = 21$) than in breast cancer cell lines ($n = 70$) (Supplementary Fig. S2, Annexe I). When we analyzed other types of hematopoietic cancer cell lines, we found high expression of IP genes in lymphoid leukemia and lymphoma cell lines (Supplementary Fig. S2, Annexe I). Two points can be made from these data. First, they support the concept that high expression of IP genes is an intrinsic feature of AML but not breast cancer. Second, they suggest that overexpression of IP genes may be found not only in AML but also in other types of hematolymphoid malignancies.

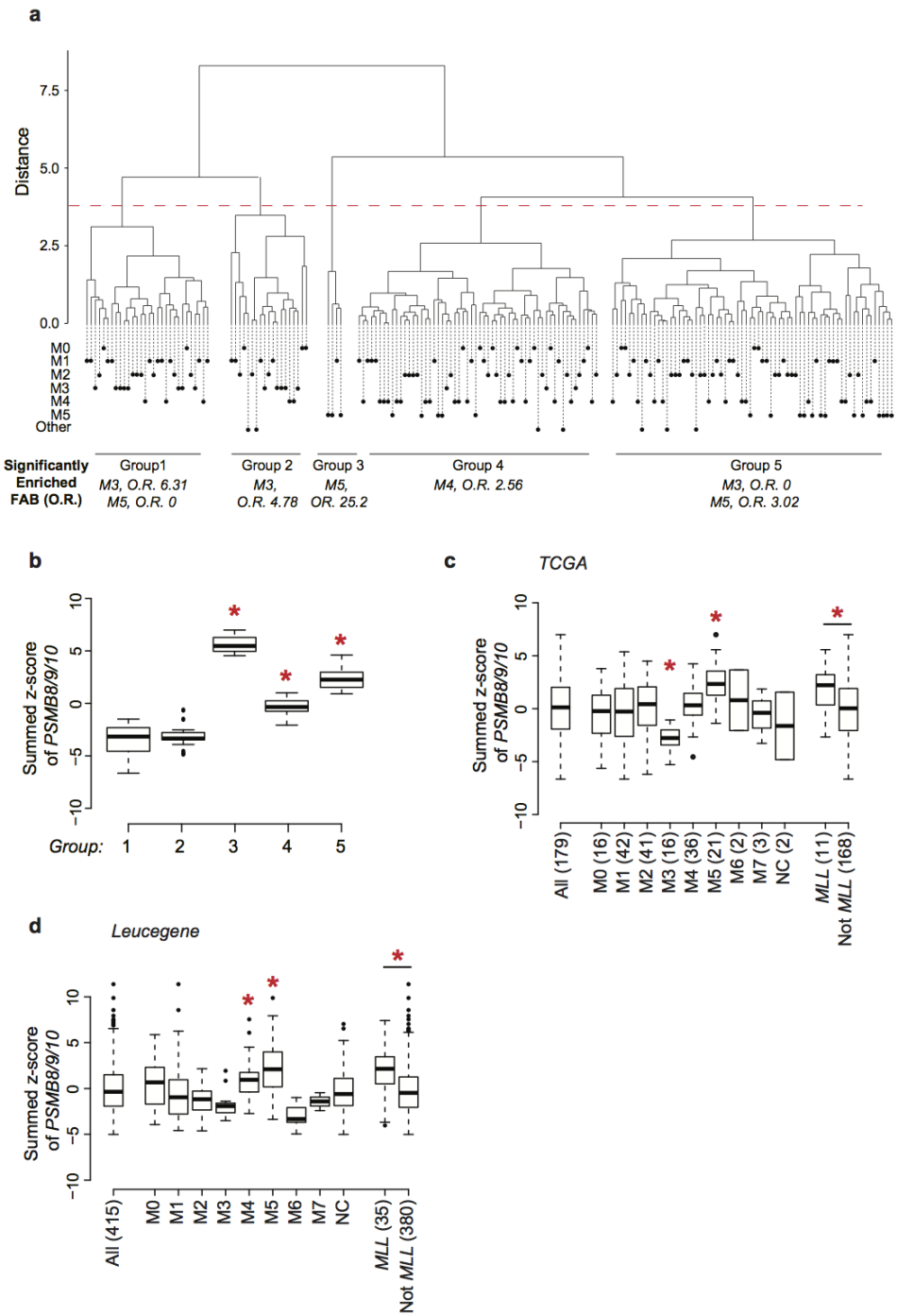


Figure 3.3 : IP expression is upregulated in AML with an M5 phenotype or MLL rearrangements

3.3.6 IP expression is upregulated in AML with an M5 phenotype or MLL rearrangement.

AML is a complex and heterogeneous disease, which can be divided into distinct classes based on cytogenetic and molecular profiles[27]. Hence, to gain insights into IP regulation and function in primary AML samples, we performed a hierarchical clustering analysis of all TCGA AML samples based solely on the expression of IP-encoding genes (see Methods). This analysis led to the identification of five clusters of AML patients (Figure 3.3a). Then, enrichment analyses were performed on each cluster for known cytogenetic markers, major translocations or morphologic subtypes (FAB classification) (see Methods). Only enrichment in FAB categories are shown since they yielded significant enrichments : patients with M3 AMLs were found to be enriched in clusters 1 and 2, which express low levels of IP, while M5 AMLs were enriched in clusters 3 and 5, which express high levels of IP (Figure 3.3a-b). Furthermore, targeted classification of samples according to FAB subtypes confirmed that M5 AMLs expressed higher levels of IP than other FAB classes whereas M3 AMLs expressed lower IP levels (Figure 3.3c). We also confirmed the overexpression of IP genes in M5 AMLs using the Leucegene cohort (415 samples), an independent cohort of AML samples with RNA-Seq data (Figure 3.3d)[55, 56]. In AML, MLL rearrangements are frequently associated to the M4 and M5 morphologic subtypes, while M3 AML are caused by promyelocytic leukemia/retinoic acid receptor-α (PML-RARA) oncoproteins[28, 73]. Accordingly, we found superior IP gene expression in AMLs with MLL fusions in both the TCGA and Leucegene cohorts (Figure 3.3c,d). Overall, these data show that IP expression is particularly high in AML with an M5 phenotype and/or MLL rearrangement.

3.3.7 IP expression is regulated by DNA methylation.

Next, we sought to investigate the mechanisms responsible for the differential expression of IP subunits in AML, using M5 vs. M3 AMLs as a model for high and low IP-expressing cancer types. First, we analyzed whole-exome and whole-genome sequencing data available from TCGA for AML samples[13]. No recurrent mutation was

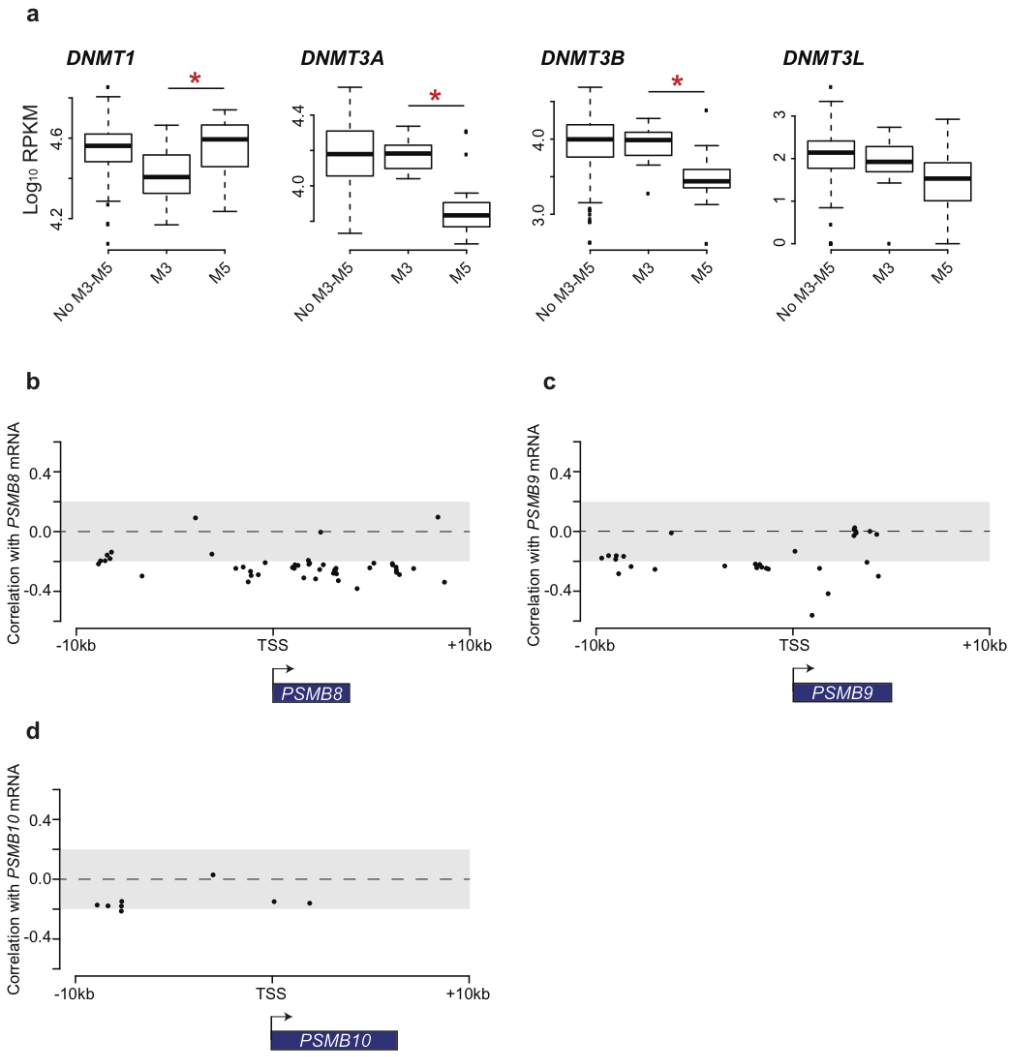


Figure 3.4 : DNA methylation in primary AML samples

present in the coding sequences of IP genes, or in upstream or downstream regulatory sequences (±9 000 kb; data not shown), leaving epigenetics as a possible mechanism for dysregulation of IP expression in AML. Methylation of DNA is a stable epigenetic modification leading to transcriptional repression [5]. Interestingly, enzymes that control DNA methylation (DNMT1 and DNMT3A/B) were differentially expressed between M3 and M5 AMLs (Figure 3.4a). Furthermore, the intensity of DNA methylation on several cytosines located in the coding regions of PSMB8 and PSMB9 was inversely correlated to their gene expression (Figure 3.4b-d). In order to investigate the role of DNA methylation in IP expression, we studied two cell lines : NB4 (M3, PML-RARA+) and THP1 (M5, MLL-AF9 rearrangement). In accordance with our observations in primary AML samples, NB4 cells expressed low levels of IP genes, relative to the THP1 cells (Figure 3.5a-b). Treatment with 5-azacytidine, an analog of cytidine inhibiting DNA methyltransferases[17], reduced levels of DNMT1 and DNMT3A in NB4 cells after 24 and 48 hours (Figure 3.5c). Notably, 5-azacytidine treatment increased the expression of PSMB8 and PSMB9, both at the mRNA (Figure 3.5d) and protein level (Figure 3.5e-f). We conclude that differential methylation of IP genes or their promoters can explain their contrasting expression levels in M3 vs. M5 AML. Nonetheless, further studies will be useful to determine the precise mechanisms responsible for the upregulation of IP genes by 5-azacytidine.

3.3.8 IP expression correlates with distinct functional networks in M5 vs. non-M5 AML.

Gene co-expression can yield systems-level insights into the function of genes and networks[107]. In order to investigate the role of IPs in AML subsets we split TCGA AMLs into two classes : M5 and non-M5. We then created correlation networks for each class and performed co-clustering using OrthoClust for both networks (Figure 3.6a). We then retrieved the genes correlated to IP genes within each cluster and performed GO term enrichment on each gene set. Enriched GO terms were grouped into general categories and their distributions were compared (Figure 3.6b and Supplementary Table S2, Annexe I). We found that genes correlated with IP subunits in non-M5 AMLs were

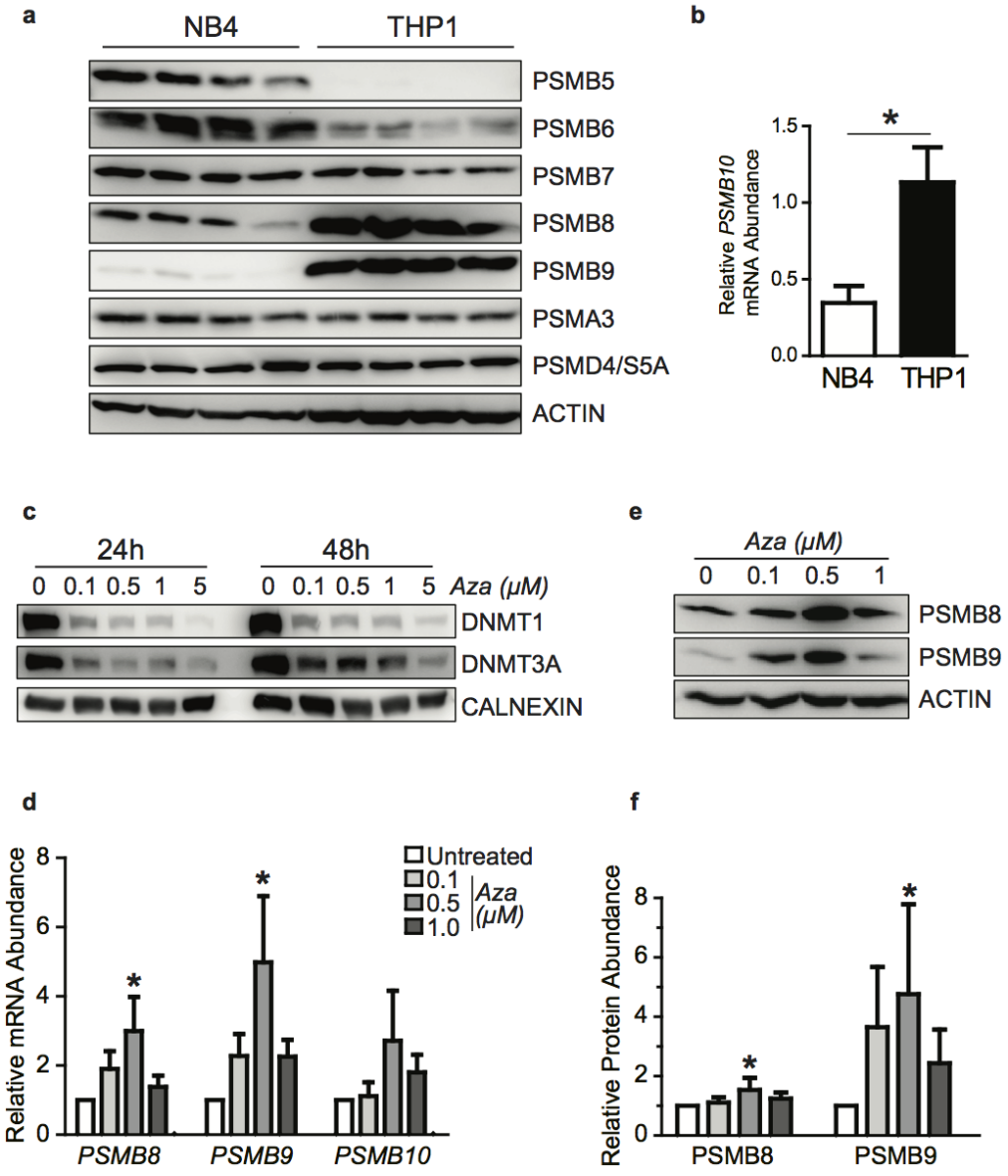


Figure 3.5 : 5-azacytidine treatment increases levels of PSMB8 and PSMB9 in NB4 cells

primarily implicated in immune processes. In stark contrast, genes correlated with IP subunits in M5 AMLs were involved in metabolic and cell cycle processes, but not in immune processes. Moreover, genes correlated with PSMB8 and PSMB9 in M5 AMLs were enriched in processes linked to mitochondrial activity and stress responses, respectively. We conclude that while IP genes are mainly instrumental in immune processes in non-M5 AMLs, they are primarily connected to cell metabolism, proliferation and mitochondrial activity in M5 AMLs.

3.3.9 THP1 cells are addicted to IPs.

If IP overexpression is linked to vital cell processes specifically in M5 AMLs, then M5 AMLs should be overly sensitive to IP inhibition. To test this hypothesis, we treated THP1 cells (M5, MLL-AF9 rearrangement) and NB4 cells (M3, PML-RARA+) with non-selective proteasome inhibitors (MG132 and Bortezomib) and with the PSMB8-specific inhibitor ONX-0914[72]. THP1 and NB4 cells showed high and low IP expression, respectively (Figure 3.5a-b). MG132 and Bortezomib increased the amounts of polyubiquitinated proteins and decreased the viability of both NB4 and THP1 cells (Figure 3.7a-d). However, the two cell lines showed divergent responses to ONX-0914. Indeed, ONX-0914 caused a massive accumulation of polyubiquitinated proteins and decreased the viability of THP1 cells, but had no effect on NB4 cells (Figure 3.7a-d). These data show that, at least for the THP1 cell line, IP overexpression in AML cells correlates with susceptibility to a selective IP inhibitor.

3.3.10 IP expression correlates with sensitivity to non-selective proteasome inhibitors.

While inhibition of proteasome activity is effective for treatment of several cancer types[108], there is limited knowledge about mechanisms of resistance to proteasome inhibitors[47]. To directly evaluate whether IP expression might regulate resistance to proteasome inhibition, we downloaded data from the Genomics of Drug Sensitivity in Cancer database, a public resource of drug responsiveness and gene expression data

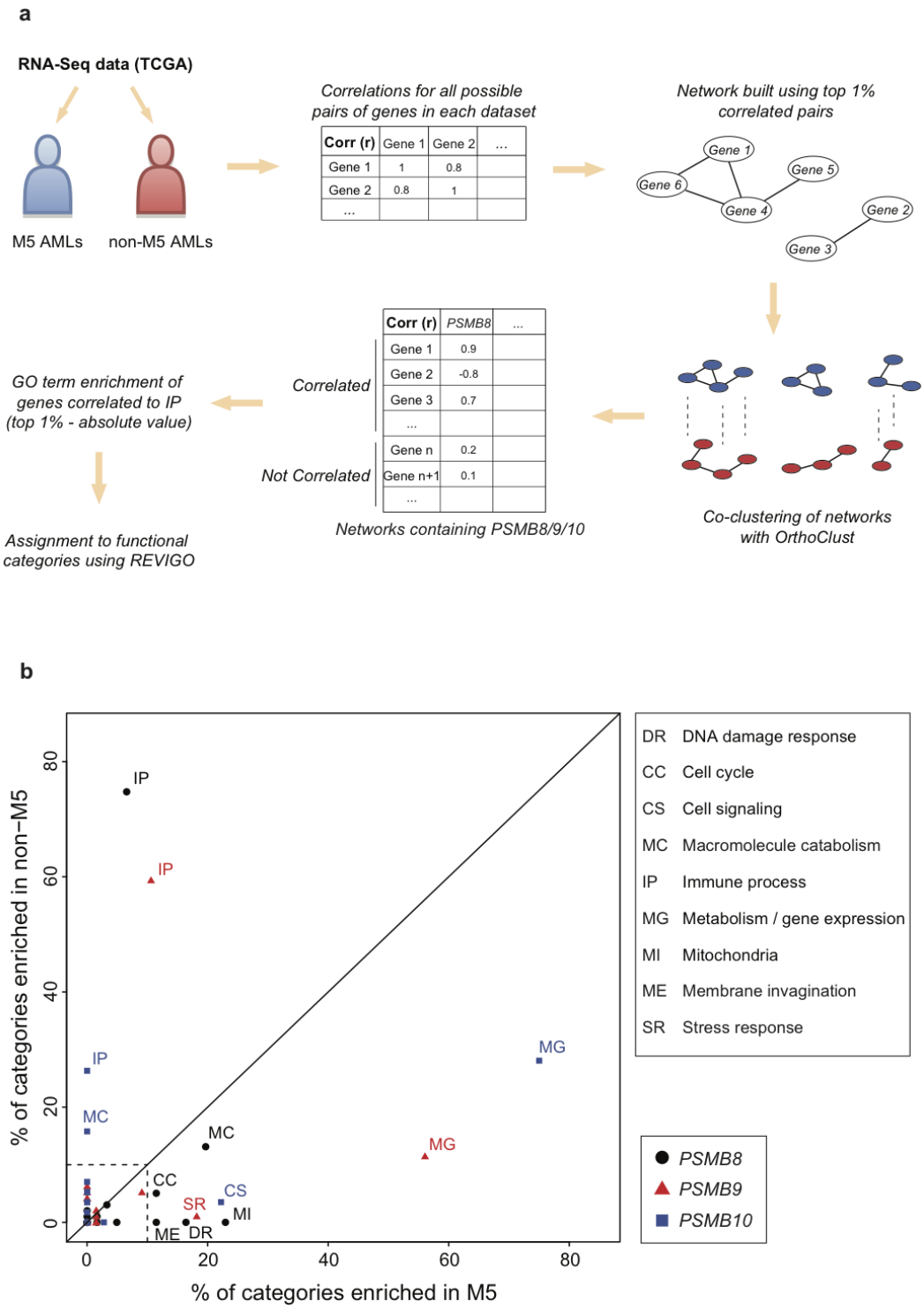


Figure 3.6 : IP expression correlates with distinct functional networks in M5 vs. non-M5 AMLs

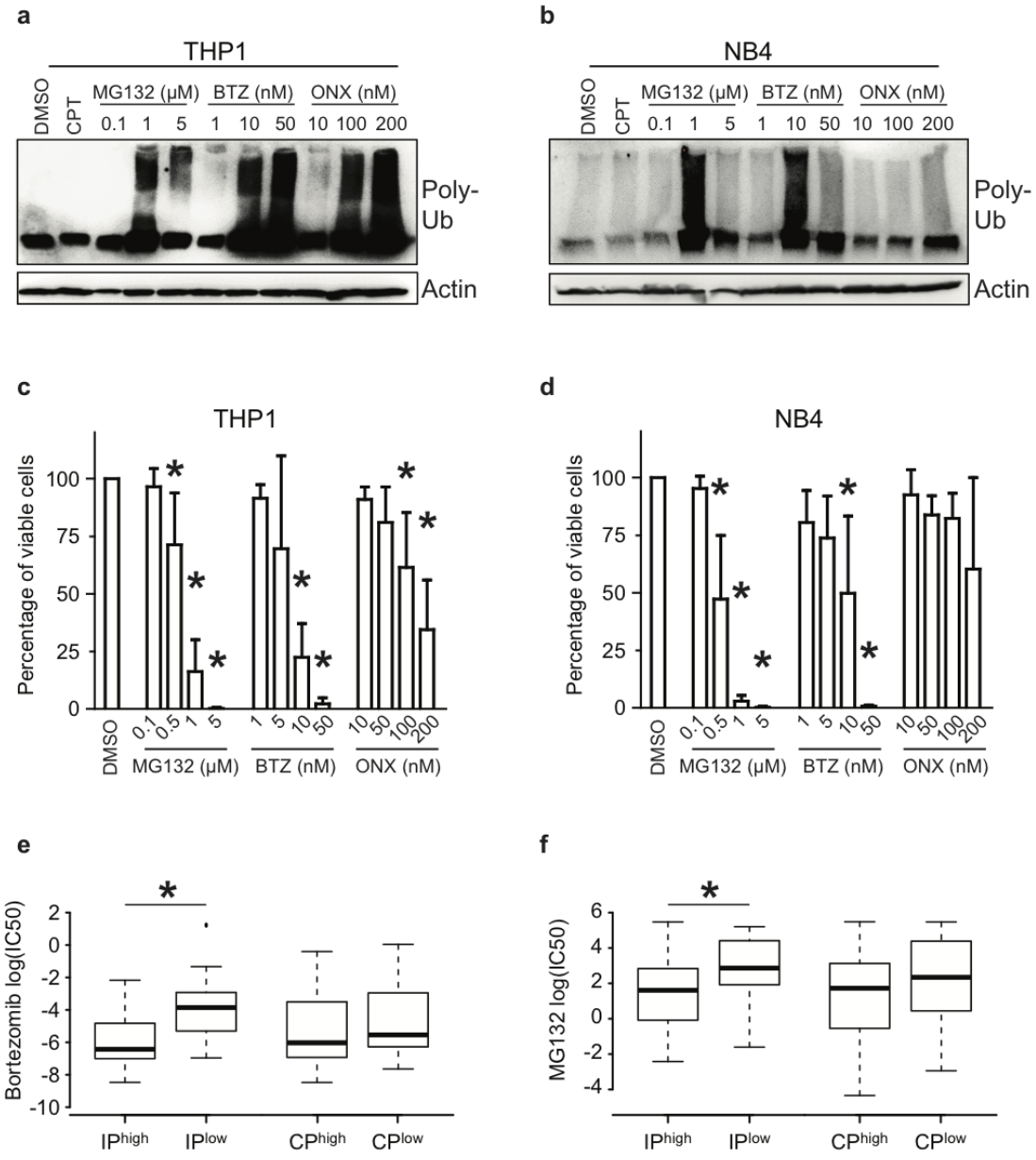


Figure 3.7 : THP1 cells are addicted to IPs

collected from a large panel of human cancer cell lines[35]. We ranked the 309 cell lines with available pharmacogenomic data according to their expression of IP or CP (irrespective of their cell lineage). Then top 10% and bottom 10% groups were used to compare their sensitivity to Bortezomib and MG132. Interestingly, IP^{high} cells were more sensitive to Bortezomib or MG132 treatment than IP^{low} cells, while no differences were observed between CP^{high} and CP^{low} cells (Figure 3.7e-f). These analyses show that high IP expression is a marker of sensitivity to non-selective proteasome inhibitors.

3.4 Discussion

The present work shows that expression of CP and IP genes is increased in most cancer types. Regulation of IP genes is of particular interest because it is regulated by both cell-intrinsic and -extrinsic factors in different types of cancer. We found that in breast cancer, upregulation of IP genes is a cancer cell-extrinsic process correlating with the presence of IFN-γ-secreting tumor-infiltrating lymphocytes. Hence, in accordance with the fact that lymphocyte infiltrates and IFN-γ secretion in solid tumors are favorable prognostic markers[48, 70, 109], high expression of IP genes correlated with improved survival in patients with breast cancer. In contrast, in AML, upregulation of IP genes did not correlate with improved survival. Furthermore, levels of IP transcripts in AML were found to be IFN-independent and cell-intrinsic features associated with hypomethylation of IP genes. In AML, the three CP genes (PSMB5, 6 and 7) are co-expressed, as are the three IP genes (PSMB8, 9 and 10), but the CP and the IP trios are regulated independently (Figure 3.2d). In line with this, the total amounts of proteasomes were similar in NB4 and THP1 cells (cf the non-catalytic subunits PSMA3 and PSMD4), but NB4 cells express mainly PSMB5, 6 and 7 CP units whereas THP1 contain mainly PSMB8, 9 and 10 IP units (Figure 3.5a-b). This is consistent with evidence that IP gene upregulation leads to replacement of CPs with IPs rather than to the addition of IPs to CPs[50]. Replacement of CPs with IPs can have far reaching consequences. Indeed, these two types of proteasomes show differences in kinetics of substrate processing and in cleavage preferences that can lead to differential expression of thousands of genes[22, 69, 101].

What can explain the cell-autonomous upregulation of IP genes in some cancers, and particularly in AML ? One IFN-γ-independent factor was shown to preferentially induce transcription of IPs over CPs : oxidative stress. Indeed, IP upregulation has been reported in cells affected by several degenerative diseases linked to oxidative stress including amyotrophic lateral sclerosis (neurons), Duchenne muscular dystrophy (myocytes) and macular degeneration (retinal cells)[7, 14, 20, 29]. Notably, relative to normal hematopoietic cells, AML cells have an increased reliance on oxidative phosphorylation and a higher mitochondrial mass, suffer from dysregulated mitochondrial biogenesis and metabolism, and are more susceptible to oxidative stress[53, 94]. Consistent with this, we found that in M5 AML, the IP network was enriched in genes involved in metabolic processes, mitochondrial function and stress responses (Figure 3.6). We therefore speculate that IP upregulation in AML cells may be driven by oxidative stress. In fact, from an evolutionary perspective, dealing with oxidative and other forms of cell stress may be the most conserved role of IPs. PSMB8 and PSMB9 orthologues have been found in invertebrates (who have no adaptive immune system), including the most basal branch of Metazoans-the placozoan Trichoplax adhaerens[97]. In invertebrates, the role of PSMB8 and PSMB9 orthologues is to help cells dealing with oxidative and proteotoxic stress. The notion that IPs are important for response to proteotoxic and oxidative stress is consistent with their expression in non-immune cells and their implication in functional processes such as cell differentiation and self-renewal[2, 22, 110]. Furthermore, we noted that cells expressing high levels of IP (but not CPs) were unduly sensitive to both selective IP inhibitors and unselective proteasome inhibition (Figure 3.7). This supports the notion that cell-autonomous IP upregulation is driven by proteotoxic and oxidative stress in cancer cells.

Irrespective of the mechanisms causing IP upregulation in M5 AML cells, their susceptibility to a selective IP inhibitor (Figure 3.7) identifies a functional vulnerability that warrants further studies. Unselective proteasome inhibitors have anti-myeloma and anti-AML activity13. However, since proteasomal activity is required for normal cell function and survival, unselective proteasome inhibitors cause substantial side effects and their therapeutic window is relatively narrow13. In contrast, transient inhibition of

IPs has no effect on normal cells[4]. Hence, we propose that recently discovered IP-specific inhibitors[21, 68] could have substantial efficacy for treatment of IP^{high} AMLs (mostly AML with an M5 phenotype and/or MLL rearrangements), and perhaps other types of IP^{high} cancers, particularly those of hematopoietic origin (Supplementary Fig. S2, Annexe I). Systematic analyses using various IP-specific inhibitors on a large panel of hematopoietic cancer cell lines will be required in order to evaluate the potential clinical relevance of IP-specific inhibitors. The correlation that we found between IP gene methylation and expression could also be relevant to AML treatment regimens containing hypomethylating agents, which are being used particularly often in elderly subjects. Indeed, expression of DNMT3A is decreased in M5 AMLs (Figure 3.4a), and mutations affecting this gene are known to be common and to have a negative impact on prognosis in AML[116]. Furthermore, we noted that treatment of the NB4 cell line with the hypomethylating drug 5-azacytidine increased IP expression (Figure 3.5d-f). Hence, since our work suggests that IPs help AML cells to withstand oxidative stress, it might be advantageous to include IP-specific inhibitors to regimens containing hypomethylating agents.The present work shows that expression of CP and IP genes is increased in most cancer types. Regulation of IP genes is of particular interest because it is regulated by both cell-intrinsic and -extrinsic factors in different types of cancer. We found that in breast cancer, upregulation of IP genes is a cancer cell-extrinsic process correlating with the presence of IFN-γ-secreting tumor-infiltrating lymphocytes. Hence, in accordance with the fact that lymphocyte infiltrates and IFN-γ secretion in solid tumors are favorable prognostic markers[48, 70, 109], high expression of IP genes correlated with improved survival in patients with breast cancer. In contrast, in AML, upregulation of IP genes did not correlate with improved survival. Furthermore, levels of IP transcripts in AML were found to be IFN-independent and cell-intrinsic features associated with hypomethylation of IP genes. In AML, the three CP genes (PSMB5, 6 and 7) are co-expressed, as are the three IP genes (PSMB8, 9 and 10), but the CP and the IP trios are regulated independently (Figure 3.2d). In line with this, the total amounts of proteasomes were similar in NB4 and THP1 cells (cf the non-catalytic subunits PSMA3 and PSMD4), but NB4 cells express mainly PSMB5, 6 and 7 CP units whereas THP1 contain mainly PSMB8, 9 and

10 IP units (Figure 3.5a-b). This is consistent with evidence that IP gene upregulation leads to replacement of CPs with IPs rather than to the addition of IPs to CPs37. Replacement of CPs with IPs can have far reaching consequences. Indeed, these two types of proteasomes show differences in kinetics of substrate processing and in cleavage preferences that can lead to differential expression of thousands of genes[22, 69, 101].

What can explain the cell-autonomous upregulation of IP genes in some cancers, and particularly in AML ? One IFN-γ-independent factor was shown to preferentially induce transcription of IPs over CPs : oxidative stress. Indeed, IP upregulation has been reported in cells affected by several degenerative diseases linked to oxidative stress including amyotrophic lateral sclerosis (neurons), Duchenne muscular dystrophy (myocytes) and macular degeneration (retinal cells)[7, 14, 20, 29]. Notably, relative to normal hematopoietic cells, AML cells have an increased reliance on oxidative phosphorylation and a higher mitochondrial mass, suffer from dysregulated mitochondrial biogenesis and metabolism, and are more susceptible to oxidative stress[53, 94]. Consistent with this, we found that in M5 AML, the IP network was enriched in genes involved in metabolic processes, mitochondrial function and stress responses (Figure 3.6). We therefore speculate that IP upregulation in AML cells may be driven by oxidative stress. In fact, from an evolutionary perspective, dealing with oxidative and other forms of cell stress may be the most conserved role of IPs. PSMB8 and PSMB9 orthologues have been found in invertebrates (who have no adaptive immune system), including the most basal branch of MetazoansâĂŤthe placozoan Trichoplax adhaerens[97]. In invertebrates, the role of PSMB8 and PSMB9 orthologues is to help cells dealing with oxidative and proteotoxic stress. The notion that IPs are important for response to proteotoxic and oxidative stress is consistent with their expression in non-immune cells and their implication in functional processes such as cell differentiation and self-renewal[2, 22, 110]. Furthermore, we noted that cells expressing high levels of IP (but not CPs) were unduly sensitive to both selective IP inhibitors and unselective proteasome inhibition (Figure 3.7). This supports the notion that cell-autonomous IP upregulation is driven by proteotoxic and oxidative stress in cancer cells.

3.5 Methods

3.5.1 Gene expression data.

For normal and cancer tissues, RNA sequencing (RNA-Seq) datasets were downloaded from The Cancer Genome Atlas (TCGA) Hub (https ://tcga-data.nci.nih.gov/tcga/) : BLCA3.1.1, BRCA3.1.2, COAD3.3.4, HNSC3.1.4, KIRC3.1.1, KIRP3.1.1, LAML3.1.7, LIHC3.1.0, LUAD3.1.2, LUSC3.1.3, OV3.1.5, READ3.1.2, UCEC3.2.7). RPKM values from each patient were log-transformed (log10 [1000*RPKM + 1]) for normal/cancer comparison. For all other analyses, log-transformed expression of individual proteasome catalytic subunits (PSMB5, PSMB6 and PSMB7 for CP, and PSMB8, PSMB9 and PSMB10 for IP) was transformed to z-score and summed to get global CP or IP z-scores. We defined z-scores as [gene X expression for a given sample - mean gene X expression of all samples] / standard deviation for gene X expression values in all samples]. Pearson's correlation coefficients were calculated between global IP z-scores and log-transformed RPKM expression of indicated genes. For analysis human cancer cell lines, RNA-Seq datasets were obtained from the ArrayExpress database under accession number E-MTAB-2706 (http ://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2706/) 22. Mean expression of IP genes (Gene IDs for PSMB8 : 5696, PSMB9 : 5698 and PSMB10 : 5699) was calculated for specific cancer cell lineages according to Klijn et al. 22. The IC50 values for Bortezomib and MG132 as well as gene expression data across 309 cancer cell lines were obtained from http ://www.cancerrxgene.org/ 34. Cell lines were ranked according to their mean expression of IP or CP genes, then top 10% ("high") and bottom 10% ("low") groups were isolated and IC50 values of Bortezomib and MG132 were compared between both groups.

3.5.2 Kaplan-Meier curves and survival analyses.

For all cancer types, clinical datasets were downloaded from the TCGA Data Portal Hub and relevant information was extracted from clinicalpatientXXX.txt files. Columns labeled "vitalstatus" were used for patient status (expired or living) and columns labeled "daystodeath" and "daystolastfollowup" were used for survival analysis. Analyses

were conducted using the R package *survival*, available at the CRAN official website (https ://cran.r-project.org/web/packages/survival/index.html). Samples were divided in two or three equal groups based on global CP or IP z-scores. Cox proportional hazards models were used to estimate hazard ratios (high group/low group) and 95% confidence intervals. The log-rank test was used to calculate p-values corrected for three or more comparisons. Since they were not treated with cytarabine-based protocols protocols (like other AML sub-groups), patients with acute promyelocytic leukemia (n=16) were not included in analyses of patient survival (Figure 3.2 and Table 3.I).

3.5.3 Analysis of DNA methylation

For AML, DNA methylation datasets were downloaded from the TCGA Data Portal Hub and methylation intensity (Beta-value) was retrieved for all sites (n = 485,577). We kept CpG sites for which Beta-value was correlated > 0.2 or < -0.2 with the log-transformed RPKM expression of either PSMB8, PSMB9 or PSMB10 (Pearson's correlation) in all AML samples. CpG sites present <10 kb upstream or downstream of the transcription start site of PSMB8, PSMB9 or PSMB10 were plotted against their level of correlation with expression of PSMB8, PSMB9 or PSMB10.

3.5.4 Cell culture

THP-1 and NB4 cell lines were obtained from Dr. Brian Wilhelm (IRIC, Université de Montréal, Canada) and maintained in RPMI-1640 media supplemented with 10% fetal bovine serum and 100 U/mL penicillin-streptomycin (Thermo Fisher Scientific, Waltham, MA). For cell viability assays, THP1 and NB4 cells were seeded at 5 x 104 cells in 100 Îijl in 96-well plates and treated for 72 hours with MG132 (EMD millipore, Etobicoke, Canada), Bortezomib (New England Biolabs, Whitby, Canada) or ONX-0914 (Cayman Chemicals, Ann Harbor, MI). CellTiter-Glo Luminescent Viability reagent (Promega, Madison, WI) was added to the wells according to manufacturer's instructions and signal was read using a plate reader. For inhibition of DNA methylation assays, NB4 cells were seeded at 5 x 105 cells in 1 mL in 24-well plates and treated 48

or 72 hours with 5-azacytidine (Sigma-Aldrich, St-Louis, MO).

3.5.5 RT-qPCRanalyses

Total RNA was extracted from cells with TRIzol RNA reagent and retro-transcribed with the High-Capacity cDNA Reverse Transcription kit (Thermo Fisher Scientific). Quantitative PCRs were performed using Taqman technology with ViiA^{TM} 7 Real-Time PCR system (Thermo Fisher Scientific) and results were analysed with the ViiA^{TM} 7 software. Primer sequences were designed with the Universal Probe Library system (Roche Life Sciences, Madison, WI) as follows : PSMB8 (Fwd : accccgcgtgacactact, Rev : gggactggaagaattctgtgg, Probe 17), PSMB9 (Fwd : accaaccggggacttacc, Rev : tcaaacactcggttcaccac, Probe 86), PSMB10 (Fwd : ggttccagccgaacatga, Rev : gcccaggtcacccaagat, Probe 31).

3.5.6 Western blot analyses

Cells were lysed in RIPA buffer (50 mmol/L Tris-HCl pH 7.4, 1% Nonidet P-40, 0.25% Na-deoxycholate, 150 mmol/L NaCl, 1 mmol/L EDTA) containing complete protease inhibitor mixture (Roche Life Sciences), 1 mmol/L Na3VO4 pH 9, 5 mmol/L NaF and 10mM NEM (Sigma-Aldrich). Samples were resolved by SDS-PAGE and immunoblotted with the following antibodies : anti-β1/PSMB6, anti-β5/PSMB5, anti-LMP2/PSMB10, anti-LMP7/PSMB8 (Abcam, Cambridge, UK) ; anti-β2/PSMB7, anti-poly-Ub (FK1 clone) (Enzo Life Sciences,Farmingdale, NY) ; anti-PSMA3, anti-S5A/PSMD4 (New England Biolabs) ; anti-β-actin (AC-15) (Sigma-Aldrich). After incubation with anti-mouse (BD Bioscience, Franklin Lakes, NJ) or anti-rabbit (New England Biolabs) horseradish peroxidase-conjugated secondary antibodies, chemiluminescent signal was detected using SuperSignalâĎć West Femto Detection Kit (Thermo Fisher Scientific) and an ImageQuant LAS-4000 imaging system (Fujifilm, GE Healthcare, Baie d'Urfe, QC, Canada).

3.5.7 Hierarchical clustering

Using R statistical software, the Euclidean distance was calculated for all TCGA AML samples (n = 179), according to expression of PSMB8, PSMB9, PSMB10, followed by hierarchical clustering.

The clustering results are shown in a dendrogram tree, built using the R package ggdendro (https ://cran.r-project.org/web/packages/ggdendro/index.html). The tree was manually separated into 5 clusters and Fisher's exact tests were performed to determine enrichment in FAB categories within each cluster.

3.5.8 Co-clustering analysis.

TCGA AML samples were split into M5 (n = 21) and non-M5 groups (n = 158). Correlation networks were built within each group based on log-transformed RPKM expression values. Network edges were drawn between two genes only when Pearson's correlation absolute value between them was in the first centile. Networks were aligned using distribution similarity (Mann-Whitney-Wilcoxon test, p-value <0.05) of a given gene in both groups, in order to enrich for gene similarities (same distribution and correlations) between groups during the co-clustering. Both networks and their alignment were fed to the computational framework OrthoClust [115], which performed co-clustering of all genes.

3.5.9 GO term enrichment.

Within clusters containing IP-encoding genes (PSMB8, PSMB9, PSMB10), all genes correlating with individual IP subunits were extracted, according to the top 1% threshold used previously. GO term enrichment was performed on these gene lists using an in-house written tool (https ://github.com/TrofimovAssya/GOrichr) that implements a modified threshold Fisher's exact test (Odds Ratio > 2) to avoid detection of slightly enriched terms in large groups. GO terms that passed these thresholds were grouped under general categories using REVIGO (http ://revigo.irb.hr/)[96] (Supplementary Table S2, Annexe I).

3.5.10 Principal component analysis

The expression of each IP- and CP-encoding genes and of regulatory subunits PSME1 and PSME2 was normalized to z-score and put through a principal component analysis using R Statistical Software for breast cancer and AML samples. Samples were projected in two dimensions using the two components accounting for most of the sample variation.

3.6 Acknowledgements

We thank the personnel of the bioinformatics, genomics and screening core facilities of the Institute for Research in Immunology and Cancer for assistance. We acknowledge the team of the Banque de cellules leucémiques du Québec (BCLQ) who provided and characterized all AML samples included in the Leucegene cohort. The BCLQ is supported by grants from the Cancer Research network of the Fonds de Recherche du Québec-Santé. This work was supported by the Canadian Cancer Society (Grant number 701564). A. Rouette was supported by studentships from the Canadian Institutes of Health Research (CIHR) and the Cole Foundation. A. Trofimov is supported by a CIHR studentship. C. Perreault and G. Sauvageau hold Canada Research Chairs in Immunobiology, and in Molecular Genetics of Stem Cells, respectively. J. Hébert holds a Université de Montréal Research Chair in Leukemia, supported by Industrielle-Alliance.

3.7 Author Contributions

AR, AT, SL and CP designed experiments. AR, AT, DH, GB, VPL and GD performed experiments. AR and AT wrote the first draft of the manuscript. SL, GS, JH and CP analyzed data and edited the final manuscript.

3.8 Additional Information

Supplementary information accompanies this paper at http ://www.nature.com/srep and in Appendix I

Competing financial interests : The authors declare no competing financial interests.

3.9 Detailed Figure Legends

Figure 1. Genes encoding proteasome catalytic subunits are overexpressed in several cancer types. Boxplots of log10 [1000 x RPKM + 1] values for genes encoding proteasome catalytic subunits were drawn for the indicated cancer types. CP genes (on the left) are PSMB5, PSMB6 and PSMB7, whereas IP genes (on the right) are PSMB8, PSMB9 and PSMB10. Red boxplots represent cancer samples and blue boxplots represent normal samples (when data are available). Differences in mean values between groups were determined by two-tailed unpaired Student's t-tests. * indicates $p < 0.05$. R.C : Renal Cell ; R.P : Renal Papillary ; AML : Acute Myeloid Leukemia ; A.D : Adenocarcinoma ; S.C : Squamous Cell.

Figure 2. Expression of IP subunits is cell-autonomous in AML. (a) Kaplan-Meier plots of overall survival (OS) for CPhigh vs. CPlow patients or IPhigh vs. IPlow patients with breast cancer. The log-rank test was used to calculate p-values. (b) Principal component analysis was performed on log10 RPKM values for genes encoding CP, IP and regulatory cap subunits (PSME1 and PSME2) in breast cancer. Plots represent the projections on the first and second principal components (PC). (c-d) For AML and breast cancer samples from TCGA, the summed z-scores of PSMB8/9/10 were plotted against log10 RPKM values of the indicated genes and Pearson's correlation coefficient (r) was calculated. (e) Principal component analysis was performed for AML samples as described above. (f) Kaplan-Meier plots of overall survival (OS) for CPhigh vs. CPlow patients or IPhigh vs. IPlow patients in non-M3 AML samples. The log-rank test was used to calculate p-values.

Figure 3. IP expression is upregulated in AML with an M5 phenotype or MLL rearrangements. (a) AML samples from TCGA were subjected to hierarchical clustering based on the expression of IP-encoding genes and separated into 5 clusters. Leafs of the dendrogram are annotated with circles that represent individual samples and their FAB subgroup. Enrichment in specific FAB classes within groups was determined by

Fisher's exact test and overall enrichment (O.R.) is shown. (b) Summed z-score of PSMB8/9/10 in groups of patients separated by hierarchical clustering in (a). (c,d) Summed z-score of PSMB8/9/10 expression in primary AML samples from (c) TCGA or (d) Leucegene cohorts were analyzed according to FAB classification and presence of MLL rearrangement. Differences between means of FAB groups were determined by one-way analysis of variance (ANOVA) followed by Tukey's post-hoc test, and difference between means as a function of MLL status was determined by Student's t-test (* indicates $p < 0.05$). Numbers in parentheses indicate number of samples.

Figure 4. DNA methylation in primary AML samples. (a) Boxplots of log10 [1000 x RPKM + 1] values for the indicated genes in FAB M3 and M5 AMLs. Differences between means were determined by Student's t-test. * indicates $p < 0.05$. (b-d) Correlation coefficient between mRNA expression of (b) PSMB8, (c) PSMB9 and (d) PSMB10 and intensity of DNA methylation on cytosines located near coding sequences of PSMB8/9/10 for all AML TCGA samples. Arrows indicates direction of transcription. TSS : Transcription Start Site.

Figure 5. 5-azacytidine treatment increases levels of PSMB8 and PSMB9 in NB4 cells. (a) Western blot analysis was performed to evaluate expression of proteasome subunits at the protein level in untreated THP1 and NB4 cells. β-actin serves as loading control. (b) Quantitative PCR analysis was performed on untreated THP1 and NB4 cells to determine total levels of PSMB10 transcripts and data were normalized according to expression of ACTB and TBP (mean $\pm$ SD of four independent experiments). (c-f) NB4 cells were treated with the indicated concentrations of 5-azacytidine for (c) 24, 48 or (e-f) 72 hours followed by western blot analysis, or (d) for 48 hours followed by quantitative PCR analysis (mean $\pm$ SD of three independent experiments). Data were normalized according to expression of ACTB and TBP for quantitative PCR and β-actin or Calnexin served as loading control for western blot. Blots are representative of three independent experiments and were quantified using ImageJ software. Differences in means between groups were determined by one-way analysis of variance (ANOVA) followed by Dunnett's post-hoc test. (* indicates $p < 0.05$). Aza : 5-azacytidine.

Figure 6. IP expression correlates with distinct functional networks in M5 vs.

non- M5 AMLs. (a) Workflow for the analysis of functional networks in M5 vs. non-M5 AMLs. (b) Co-clustering was performed on aligned correlation networks derived from expression of all genes in M5 and non-M5 AMLs. GO term enrichment was performed on genes correlating with IP subunits within their respective cluster. Significantly enriched GO terms were split into general semantically similar process categories using Revigo. The plot represents the % of functional categories for each IP subunit, in M5 and non-M5 AMLs. Categories representing less than 10% are not shown.

Figure 7. THP1 cells are addicted to IPs. (a) THP1 or (b) NB4 cells were treated for 24 hours with DMSO, 5 μM camptothecin (CPT), or with the indicated concentrations of MG132, Bortezomib (BTZ) or ONYX-0914 (ONX). Then, western blot analysis was performed to estimate total levels of polyubiquitinated proteins. Îš-actin served as loading control and blots are representative of three independent experiments. (c) THP1 or (d) NB4 cells were treated for 72 hours with DMSO or with the indicated concentrations of MG132, Bortezomib (BTZ) or ONYX-0914 (ONX), then cell viability was monitored using cell viability luminescence assay (mean $\pm$ SD of four independent experiments). Intensities were normalized as a percentage of viable cells relative to the DMSO control. Differences in means between groups were determined by one-way analysis of variance (ANOVA) followed by Dunnett's post-hoc test. (* indicates $p < 0.05$). (e-f) Cell lines from the Genomics of Drug Sensitivity in Cancer database (n = 309) were ranked according to their mean expression of IP or CP genes. Cell lines that ranked above the 90th percentile and below the 10% percentile with regards to proteasome expression were used to create IPhigh/IPlow and CPhigh/CPlow groups. We compared these groups for their sensitivity to (e) Bortezomib and (f) MG132. Differences in log(IC50) between groups were determined by paired Mann-Whitney test (* indicates $p < 0.05$).

Table 1. Correlation between risk of death and proteasome expression. Patients were divided into two or three equal groups based on CP or IP z-score. Patients with M3 AML (n=16) were not included in the analysis because they were not treated with the same chemotherapy regimen as other AML patients. Cox proportional hazards models were used to estimate hazard ratios (high group/low group) and 95% confidence intervals. The log-rank test was used to calculate p-values.

CHAPITRE 4

EMBEDDINGS : NOUVELLES AVANCÉES MÉTHODOLOGIQUES POUR L'IDENTIFICATION DES SIGNATURES GÉNIQUES

Le chapitre suivant regroupe le contenu de deux courts articles (extended abstracts), faisant état des nouvelles avancées méthodologiques proposées pour l'identification de signatures géniques dans des données de séquençage d'ARN (RNA-Seq). Les textes des articles ont été modifiés de leurs versions originales, afin d'accomoder les exigences de présentation du mémoire. Les références de ces articles sont incluses dans les références totales du mémoire et les numéros de figures portent le numéro de chapitre. Des informations additionelles ont également été incluses afin de rajouter à la compréhension du chapitre.

Contribution de l'étudiant

Ma contribution aux articles c'est fait dans l'implémentation des réseaux de neurones artificiels et la planification, conception et réalisation des analyses bioinformatiques.

Je me suis occupée de télécharger et traiter les données de RNA-Seq et implémenter les réseaux de neurones variés utilisés dans les articles. J'ai également réalisé les expériences et interété les résultats. Finalement, j'ai rédigé les manuscrits et la créé les figures.

Warp : a method for neural network interpretability applied to gene expression profiles

Assya Trofimov[1,2,3], Sébastien Lemieux[1,3] and Claude Perreault[1,2]

[1]Institute for Research in Immunology and Cancer; [2]Department of Medicine and [3]Department of Computer Science and Operations Research, Université de Montréal, Montreal, Quebec, Canada;

Article présenté dans le Machine Learning in Computational Biology Workshop, à la conférence Neural Information Processing Systems (NIPS), le 10 décembre 2016.

4.1 Abstract

We show a proof of principle for warping, a method to interpret the inner working of neural networks in the context of gene expression analysis. Warping is an efficient way to gain insight to the inner workings of neural nets and make them more interpretable. We demonstrate the ability of warping to recover meaningful information for a given class on a sample-specific individual basis. We found warping works well in both linearly and nonlinearly separable datasets. These encouraging results suggest that warping has a potential to be the answer to neural networks interpretability in computational biology.

4.2 Introduction

Application of machine learning is off to a promising start in the field of computational biology. Every year, innovations in the field help bridge a little more the gap between measured biological data and observed phenotype. However, in some cases, algorithms are applied in a "black box" setting, which hinders the proper use of learning algorithms. In an image classification setting, it is expected that, while learning, the neural nets (NN) extract the representation of visual characteristics (shape, color) for a given class [57]. Notably, Simonyan and colleagues [91] show that when giving a pre-trained and frozen neural network an input of noise, they can adjust the input to maximize the class score and this way generate what seems to be the model's captured class notion. Using this type of analysis allows to visualize what the NN detects [98] and what it learned about the characteristics of a given class. Moreover, recent work from Mikolov and colleagues [67] have shown through vector arithmetic in the learned representation space that learning algorithms extract meaningful data representations.

Unlike the field of image analysis, however, machine learning in computational biology calls for a more challenging problem. Image analysis is about training a machine to recognize images, a task at which humans excel. In contrast, machine learning algorithms in computational biology are asked to do the humanly impossible : crunch tremendous amounts of data of abstract construct and extract patterns that can be exploited to predict a phenotype. Some algorithms, such as random forests, are more readily interpretable, which is probably the reason they are so popular in biological data analysis. Indeed, random forests offer easy access to variables and thresholds deemed important for the prediction. In contrast, NN offer a bigger challenge in terms of interpretability and are often overlooked as good candidates for prediction tasks, since users want to avoid a "black box" situation and prefer to fully understand how the analysis is done.

In the following study, we propose a proof-of-principle to the difficult interpretability of NN in the analysis of biological data. For simplicity, we limit ourselves to the specific task of phenotype prediction based on gene expression data. However, we think that our solution has a broader applicability for tasks of different nature. Inspired by the work

of Simonyan [91] and Mikolov [67], our method calculates the change necessary in the input dimension space, to classify the sample as one of the other class, while remaining on the data manifold. Here we define the technique and examine its behavior on a real gene expression dataset.

4.3 Methods

4.3.1 Datasets

Two datasets were selected for the experiments in this study. The first dataset called XOR is a standard visualization and testing dataset that is custom generated. It consists of 2-dimensional data points from two classes (here N=100), arranged in a non-linear fashion. The second dataset, Phenotype, was obtained from The Cancer Genome Atlas (TCGA), a National Institute of Health (NIH) cancer research consortium. We used the normal lung and kidney control RNA-Seq of 100 patients.

4.3.2 Input Dimension Interpolation

We pre-trained a multi-layer perceptron to classify the XOR dataset. Once the trained neural net architecture in place, we froze the layers, making updates to weights and biases impossible (more details in section 2.6.2.4). We then added an additional layer of the same size as the input between the input and the first layer and only allowed updates to biases. We then proceeded to create a "new" dataset, where all members of a specific class were flipped to the other one. The network was then presented with one example and allowed to update the added bias layer and reach convergence. The bias layer was then extracted and the values were examined to determine the smallest necessary adjustment in input dimension space to classify the sample as one of the other class.

4.3.3 Variational Autoencoder and latent space vector arithmetic

Our variational autoencoder (VAE) [51] encodes gene expression data into smaller representation space, while maximizing the likelihood of the data X in representation space Z (Gaussian Z). The general cost function for the VAE is :

$$L(\zeta, \theta | x^{(i)}) = -D_{KL}(q_\zeta(z|x^{(i)})||p_\theta(z)) + \mathbb{E}_{q_\zeta(z|x^{(i)})}[log(p_\theta(x^{(i)}|z))]$$

We want to find ζ (the parameters of the encoder $q_\zeta(z|x^{(i)})$) and θ (the parameters of the decoder $p_\theta(x^{(i)}|z)$) given data X. The first term, $-D_{KL}(q_\zeta(z|x^{(i)})||p_\theta(z))$ is the Kullback-Leibler divergence between the probability distribution of z given X, under parameters ζ and the probability distribution of z, under parameters θ. The second term $\mathbb{E}_{q_\zeta(z|x^{(i)})}[log(p_\theta(x^{(i)}|z))]$, is the expectation of having a X given a z. In other words, for parameters ζ and θ, the network maximizes the likelihood of finding a good representation in z space for every X, while minimizing the reconstruction error when decoding z to X. More detailed information can be found in (section 2.6.2.4). We performed latent space vector arithmetic, similar to Mikolov[67], by using the general class centroids and show (Figure 4.2) that the VAE has extracted meaningful representations of the different data classes.

4.4 Results

4.4.1 Intuition on synthetic data

For the purpose of this article, we define the term *class gene signature* as a specific set of genes as well as their expression levels, that are sufficient to identify a phenotype or class. The standard approach to examining *class gene signatures* would be to calculate for each class, a mean gene expression profile and then examine the difference. This technique does not capture class-specific patterns that are not linearly separable. Our first experiment was performed on the XOR dataset (Figure 4.1).

Here, comparing mean values between classes across features will not work, since both classes have the same mean value over each dimension. We then implemented and

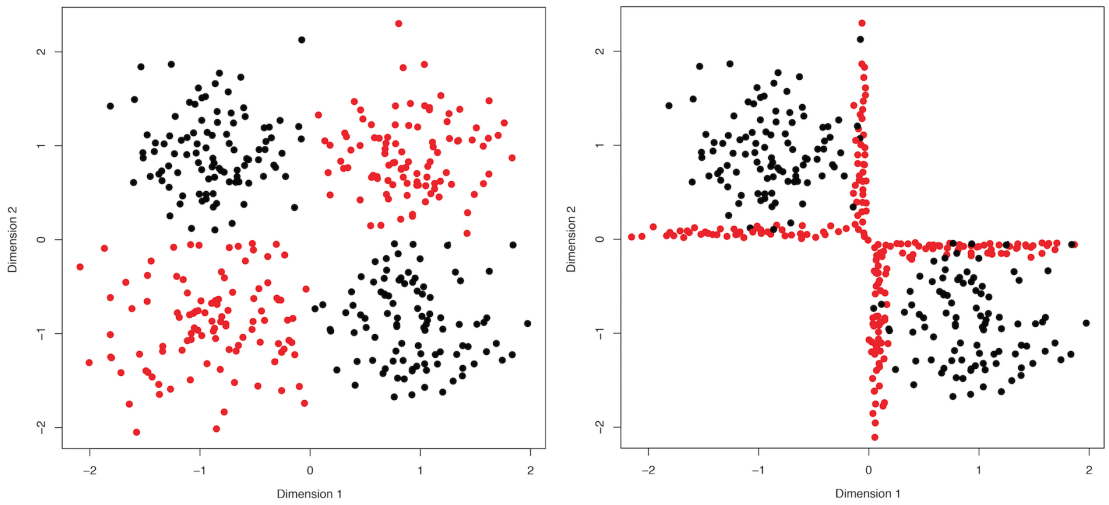


Figure 4.1 : Linear interpolation on XOR dataset

trained an MLP on the dataset and proceeded to perform the warp for all examples of class A (red data points).

When applying the warping vectors to class A, we observe that all data points seem to have migrated inside of the decision boundaries (Figure 4.1). This confirms that our approach functions well with non-linearly separable datasets and highlights the complexity of the output, since bias vectors are calculated on a sample-specific level.

4.4.2 Important flaw in the linear interpolation model

This linear interpolation technique is however flawed, since it assumes the data manifold is continuous in input space. Since this assumption is likely false, this technique will necessarily create adversarial examples [98]. Since the representation learning task is to maximize efficient representation of the data manifold within the representation space, we propose to move the interpolation to the hidden layers, thus guaranteeing that any new point in the space will belong to a plausible example.

4.4.3 Corrected proof of concept with sex and tissue source-site classification

A VAE is used to encode the Phenotype dataset. We observe that the plotted coordinates in representation space seem to cluster by group (Figure 4.2).

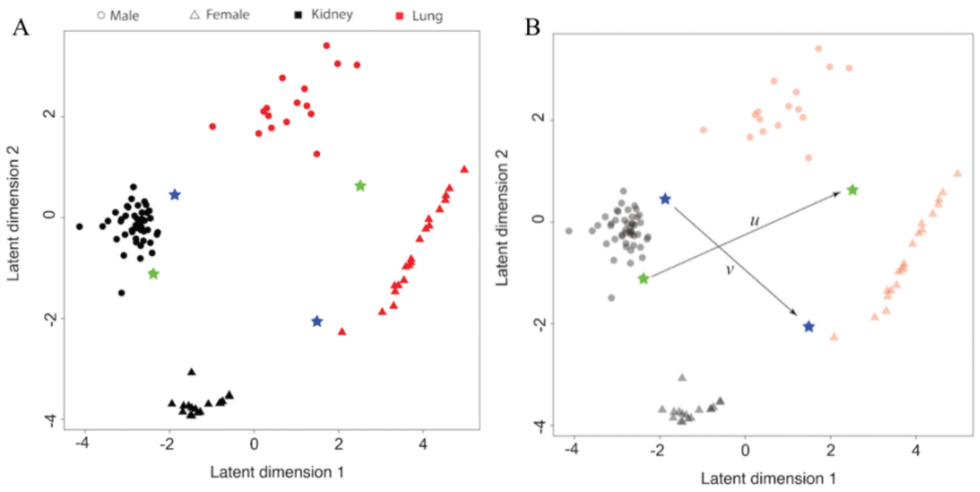


Figure 4.2 : Samples aggregate in encoding space by sex and tissue type

We then performed interpolation in representation space similar to the technique described in Section 4.3.3 (Figure 4.2). For an example point with the class labels male and lung, the coordinates in representation space are [0.672, 2.77]. When added the female centroid and kidney centroid and then subtracted the male and lung centroids, we obtain a new point coordinate, [-1.147, -2.46] (Figure 4.3).

Example point

male, lung	+	*female centroid, kidney centroid*	-	*male centroid, lung centroid*	=	*Nearest Neighbour: Female, Kidney*
[0.672, 2.77]		[-0.925, -3.149]		[-0.634, -1.089]		[-1.147, -2.46]

VAE Decoder

Gene Name	Gene ID	Description	Original Value	Warped Value
SFTPA1	653509	surfactant protein A1	0.996	0.000
SFTPA2	729238	surfactant protein A2	0.997	0.000
AGER	177	advanced glycosylation end-product specific receptor	0.998	0.000
SLC13A3	64849	solute carrier family 13 member 3	0.000	0.393
CYS1	192668	cystin 1	0.000	0.923
RPS4Y1	6192	ribosomal protein S4, Y-linked 1	0.785	0.000
DDX3Y	8653	DEAD-box helicase 3, Y-linked	0.825	0.000
EIF1AY	9086	eukaryotic translation initiation factor 1A, Y-linked	0.884	0.000
TTTY14	83869	testis-specific transcript, Y-linked 14 (non-protein coding)	0.838	0.000
TXLNGY	246126	taxilin gamma pseudogene, Y-linked	0.056	0.000

Figure 4.3 : Example of a linear transformation in encoding space

The nearest neighboring points hold the labels female and kidney. Moreover, the decoder part of the VAE allows the decoding of this new point into input dimensions,

yielding the exact adjustment values for that example point to become one of the other class (top 10 genes according to adjustment magnitude are shown) (Figure 4.3). Indeed, originally, the patient expressed genes of the sufractant family[44], as well as chromosome Y-linked genes. In order to be classified as female and kidney, the network has suggested that these genes be adjusted to an expression value of 0 and simultaneously boost the expression value of the solute carrier family proteins as well as the cystin family, two genes known to be overexpressed in kidney cells[32, 64].

4.5 Conclusion

In this proof-of-principle, our method shows encouraging results as a valid alternative for examining *class gene signatures* post neural network training. The approach offers great flexibility, detecting various types of signatures. Ongoing experiments are being conducted to determine how destructive a warp is to other gene signatures within the samples as well as overcoming reconstruction blurriness caused by the VAE nature of encoding.

Uncovering the gene usage of human tissue cells with joint factorized embeddings

Assya Trofimov[1,2], Joseph Paul Cohen [1], Yoshua Bengio[1], Claude Perreault[2,3] and Sebastien Lemieux[2]

[1]Montreal Institute for Learning Algorithms (MILA) ; [2]Institute for Research in Immunology and Cancer ; [3]Department Medecine, Université de Montréal ; Montreal, Quebec, Canada ;

Correspondence to : Assya Trofimov <assya.trofimov@umontreal.ca>.

Article présenté dans le Workshop for Computational Biology, à la conférence International Conference for Machine Learning (ICML), le 10 août 2017.

4.6 Abstract

We present a factorized embedding method, that learns simultaneously gene and sample embeddings in their respective latent space, in a gene function-dependent manner. Running the model on RNA-Seq data, we observed that tissue samples aggregated spontaneously in latent space by tissue similarity while genes aggregated in latent space by gene function.

Our method recovered most gene-gene association reported in the reference database Genetic Network Annotation Tool (GNAT) [80]. Our approach has the potential to uncover not yet known gene interactions, in a tissue-specific manner.

4.7 Introduction

We propose a deep learning-based dimensionality reduction method (see section 2.6 for more details) that simultaneously learns about samples and genes, while embedding them into their respective latent space based on function. While RNA-seq is a powerful tool to gain insight into a cell's mechanisms, it has many pitfalls, most of which are mostly based on the size of the data. Indeed, an RNA-Seq experiment yields for each studied sample a long vector of gene expression values (10^4 - 10^5). Often analysis is done only on samples to monitor gene expression changes using some type of selection scheme, and to those selected genes are associated functions, through enrichment analysis pipelines. While these methods seem to perform well, they are subjected to stringent cutoffs and are often not reproducible in other datasets[10]. Moreover, rare are the methods that examine the gene changes for all samples.

Our contributions :

- We present a method that aggregates tissue samples by similarity in latent space.
- Our method also simultaneously aggregates genes based on their function.
- Our gene embedding space represents gene function and recovers most of reported co-expressions in two gene annotation databases as well as additional relationships.

4.8 Factorized Embedding

The core of our approach relies on finding good embedding coordinates for gene and tissue samples in order to minimize the prediction error of the network (more details in section 2.6). We create two spaces where gene and tissue samples are represented. The network receives as input pairs of indices ; for a gene expression matrix with N samples by M genes, the network will receive $N \times M$ pairs of indexes and attempt to predict the gene expression for each index pair. The coordinates in gene space and sample space for each of these indices are concatenated and fed through one fully-connected non-linear layer of the neural network. The output is a single neuron, predicting the gene expression value for the input pair and the network is penalized on a Mean Squared Error (MSE) (Figure 4.4). To contrast our method with bi-clustering [62], we factor by concepts we control, namely genes and tissue types. Moreover, we found that the embedding space captures functional relationships between genes as well as tissues, something not guaranteed in bi-clustering.

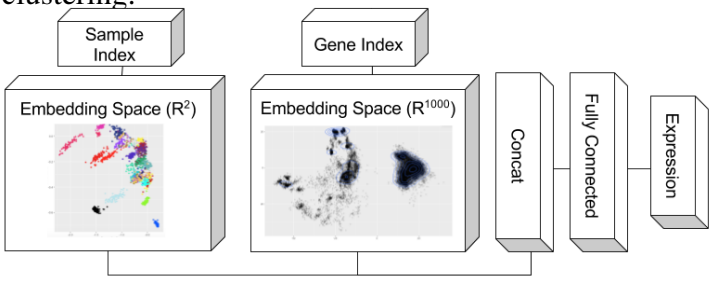


Figure 4.4 : Neural net architecture

We kept the network above the embeddings minimal, in order to maximize the encoding within the embeddings, rather than the parameters of the fully connected layer. We have tested increasing sizes of gene and sample embedding space, but only changing gene embedding space seemed to significantly improve performance. For all experiments, the fully connected layer contains 10 or 25 neurons. Training is done by iterating through every pair of sample and gene indexes and adjusting the embeddings and network parameters to decrease prediction error. We train using a batch size of 10,000 over the entire dataset of 112M pairs, with an RMSprop optimizer, at a learning rate of 10^{-3}. Our code was implemented using Theano [100] and Keras [16] and we release our

source code online.

4.9 Experiments

4.9.1 Data

We use the GTEx dataset, obtained from the Genotype-Tissue expression project (GTEx), a RNA-Seq dataset, that contains 8910 tissue samples, over 30 different tissue types [59]. We removed two tissue types, Brain and Fallopian Tube, mainly for the sample size and ambiguous labeling.

Gene expression values are continuous values, that have been log-transformed for this analysis.

4.9.2 Tissue embeddings

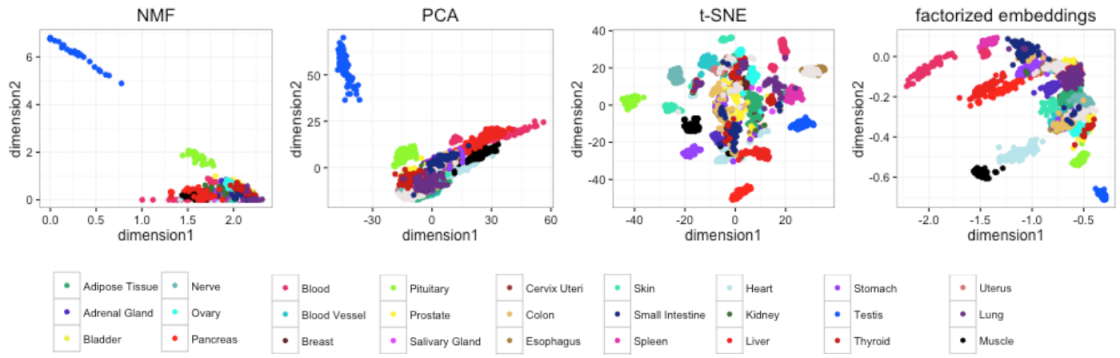


Figure 4.5 : Comparison of embeddings from different popular dimensionality reduction techniques

We first examined the tissue embedding space. We found that our method as well as t-SNE separate the samples into groups (Figure 4.5). We then investigated how these tissue groups relate to each other in the sample embedding space. Four genes or gene groups were chosen as representative examples of genes that should be expressed across multiple tissues; i) keratin genes (KRT) for different epithelial tissues ii) MYH6 for muscle tissues (heart and muscle), iii) CD8B as a marker for immune cells and iv) XIST

as a sex-determining gene (Figure 4.6). Our results suggest that the factorized embeddings, but not t-SNE, segregate space into regions of tissue similarities, according to gene expression values. We also note that expression of sex-related genes is not a major contributor to localization in embedding space.

Algorithms	NMF	PCA	t-SNE	Factorized Embeddings
1-NN	0.31± 0.02	0.38± 0.03	0.62± 0.04	0.62± 0.02
Decision Tree	0.28± 0.03	0.32± 0.04	0.50± 0.05	0.50± 0.05
Gaussian NB	0.22± 0.04	0.22± 0.04	0.28± 0.04	0.29± 0.05
Random Forest	0.28± 0.03	0.31± 0.03	0.47± 0.04	0.47± 0.04
AdaBoost	0.16± 0.04	0.18± 0.04	0.25± 0.06	0.27± 0.06

Tableau 4.I : Semi-supervised classification of embeddings of tissues by different methods methods.

We benchmarked our method against other dimensionality reduction techniques, such as principal components analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) [103] and non-negative matrix factorization (NMF) [12]. We take each obtained embeddings and train a battery of standard classifier using only 2% and present the accuracy of the classification on the remaining 98%. We have chosen k-Nearest Neighbours (1-NN) [19], a Decision Tree [84] with and without Adaboost [33], a Random Forest [11], and a Gaussian Naive Bayes [85] classifier. The numbers shown (Table 4.I) are the average result of 100 random splits. We observe that t-SNE (at perplexity 30, chosen via a hyper-parameter search) and our method generate embeddings that contain similar intra-group information (Table 4.I). Although both methods perform similarly in the semi-supervised classification task, given the difference in organization of tissues in encoding space (Figure 4.6), we find our embedding model to offer better division of tissue embedding space.

4.9.3 Examining the gene embedding space

During training, embedding coordinates move from the (0,0) initialization to their current coordinates. We examined the gene embedding space and found that low expressed genes aggregated around the starting coordinates (0,0), while highly expressed genes were pushed out in the periphery. Given this, we hypothesized that correlated

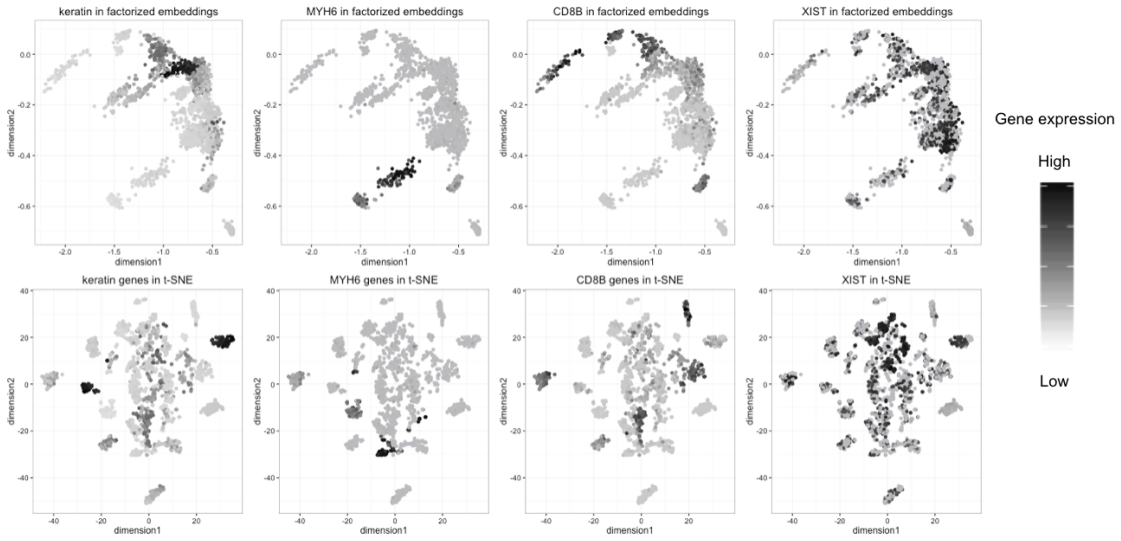


Figure 4.6 : Factorized embeddings but not t-SNE exploits tissue embedding space as a function of relative gene expression
Representation of tissue embeddings coloured by gene expression values for four chosen genes or gene groups : keratin, myosin heavy chain 6 (MYH6), CD8b and X-inactive specific transcript (XIST)

genes would aggregate together in gene embedding space.

To verify this hypothesis, we calculated all pair-wise correlations (Pearson correlation) for all genes and sorted the Pearson's r values. We then sampled pairs of genes within the top 0.1-50% of all correlations and reported their euclidean distance in embedding space. We found that genes with a higher correlation value were generally located closer to each other than less correlated genes (Figure 4.8).

Given the observed better division of tissue embedding space by our method, we hypothesized that the performance is due to the model simultaneously learning about gene functions in cells. Similar to what was observed in tissue embedding space, we expected the model to aggregate genes participating in similar functions in the gene embedding space. To verify this, we obtained tissue-specific gene sets, calculated by Pierson and colleagues [80] and reported by the Gene Network Annotation Tool (GNAT). For each tissue, the tissue-specific gene list was obtained and euclidean distances in embedding space were measured between these genes. Euclidean distances within a matching in size set of randomly selected genes was also obtained and values were compared with a

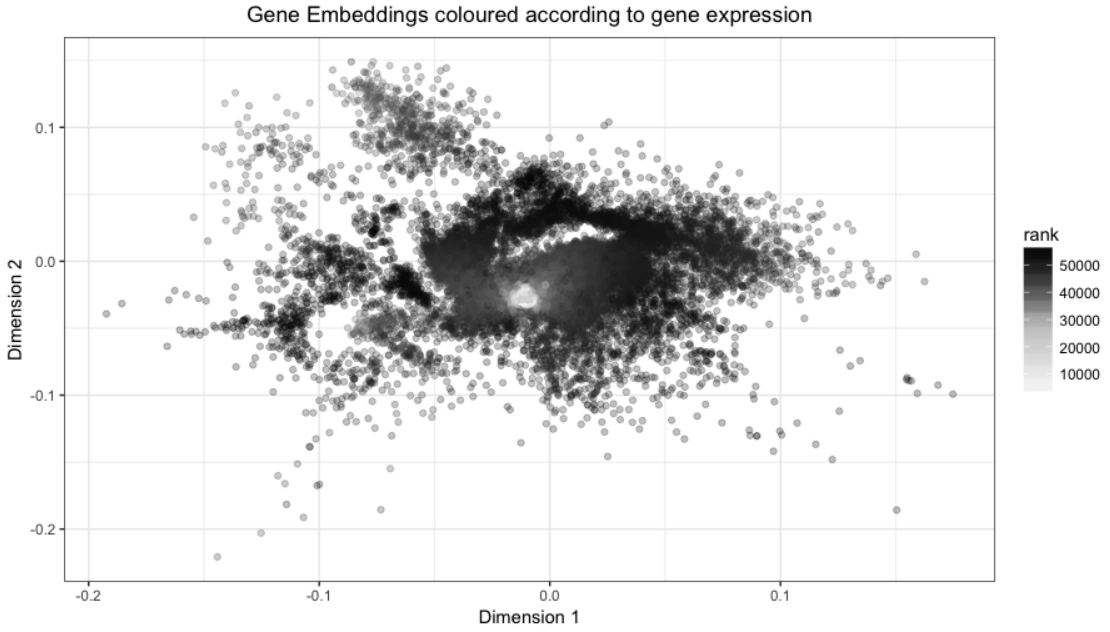


Figure 4.7 : 2-dimensional gene embeddings according to gene expression rank (sorted in ascending order)

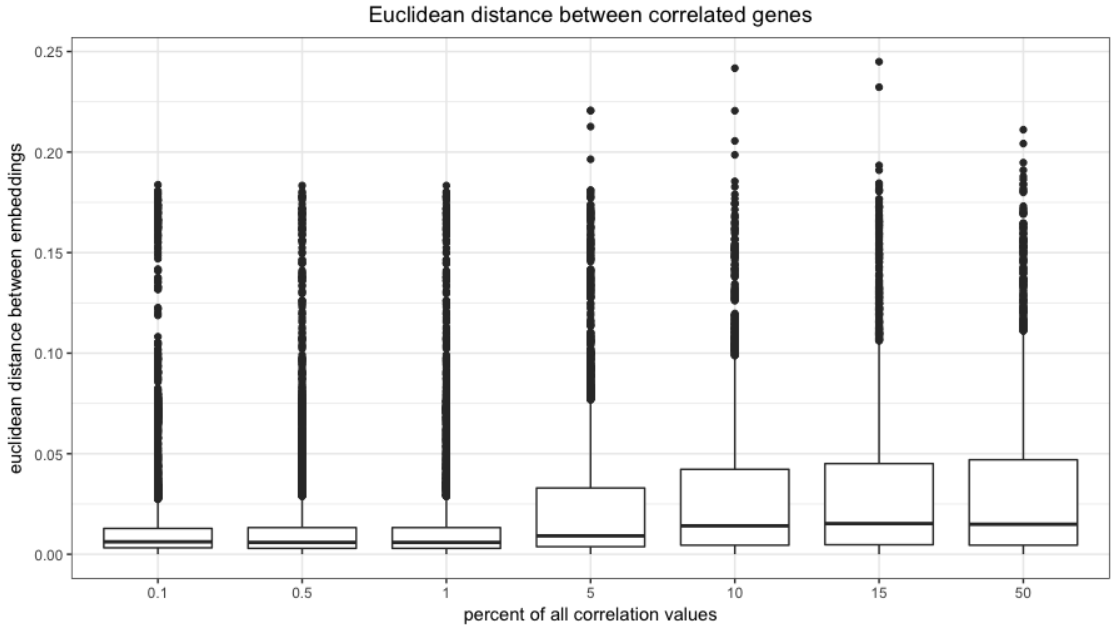


Figure 4.8 : Euclidean distance between genes, according to correlation threshold, across all samples

two-tailed Student's t test (Figure 4.9). We show four tissues : stomach, skin, lung and pancreas, chosen for their well defined clusters in Figure 4.5. For all four tissues, distances between genes in the tissue-specific gene sets were significantly smaller than in the randomly selected genes (p-values <0.0001). Taken together, these results show that gene embeddings are arranged in space according to the magnitude of gene expression, capture gene correlations in samples and recover reported tissue-specific gene sets [80].

4.10 Conclusion

In this work we have shown that our method learns simultaneously i) sample embeddings by stratifying the space by tissue similarity and ii) gene embeddings, by aggregating genes participating in common functions. Finally, because of the way the data is input, our method is robust to partial data. This allows for incorporation of many datasets or single-cell RNA-Seq experiments.

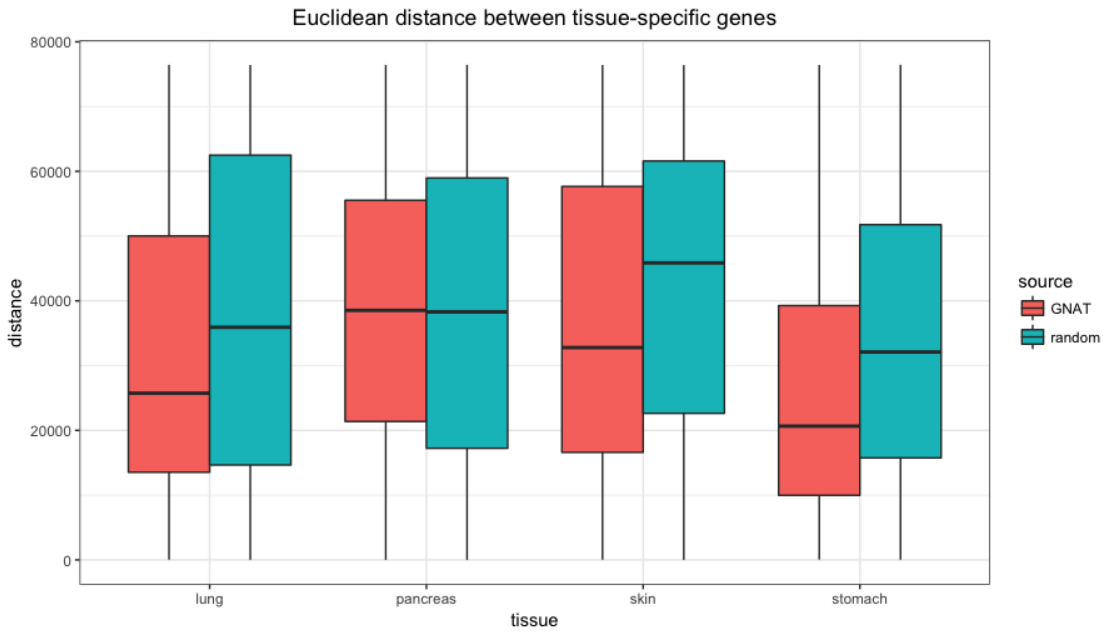


Figure 4.9 : Euclidean distance in embedding space for genes in the same tissue-specific gene list, proposed by the Genetic Network Annotation Tool (GNAT), compared to euclidean distance of randomly selected genes

CHAPITRE 5

DISCUSSION

5.1 La problématique des méthodes clairsemées (sparse)

Depuis que l'arrivée de la première signature génique de Golub et collègues en leucémie [37], plus de 150 000 signatures géniques ont été rapportées dans le domaine de la recherche sur le cancer et moins de 100 sont employées actuellement en clinique [83]. Une équipe a rapporté qu'avec le même jeu de données, ils ont pu extraire 1789 signatures géniques uniques et ils estiment qu'ils pourraient trouver plus de 500000 signatures prognostiques vérifiables dans ce même jeu de données[10].

J'ai observé un phénomène similaire dans les analyses de co-partitionnement (Chapitre 3). En effet, lors de la construction du réseau de corrélation, j'ai choisi de n'admettre que les top 1% des corrélations dans le jeu de données. Suite à l'enrichissement des GO terms dans chaque partition de gènes extraite, j'ai découvert un ensemble de fonctions décrivant ces gènes. J'ai basé le choix du seuil de 1% sur un seuil rapporté adéquat dans la littérature, dans des analyses similaires ([82]). Cependant, en accord avec la littérature, il était suffisant de changer le pourcentage des corrélations admises pour dramatiquement changer les résultats ([36]). En changeant le seuil des corrélations choisies pour le graphe, ceci changerait les partitions retrouvées par l'algorithme de co-partitionnement et donc modifierait les ensembles de gènes fournies à l'algorithme d'enrichissement de GO terms, peignant un paysage fonctionnel probablement différent.

De plus, la méthode employée compare un sous-ensemble de AML, les AML monocytaires aux autres AML contenus dans la cohorte. Il serait envisageable que la méthode soit biaisée vers les AML monocytaires, vu que l'échantillon est plus petit. En effet, un débalancement de classes est généralement néfaste pour les algorithmes d'apprentissage et rend les résultats moins fiables [24]. Finalement, le sous-groupe des AML monocytaires est uniforme en phénotype, comparé à l'autre groupe comparé, qui contenait les AML de classification M0-M4, dont les AML promyelocitaires, connues pour être dra-

matiquement différentes.

Malgré que ces limites de la méthode peuvent avoir peu d'impact sur les conclusions que mes co-auteurs et moi avons faits dans le manuscrit I (Chapitre 3), ces considérations pointent vers une faille possible de la méthode, qui requiert un réglage fin des seuils.

De cette constatation émerge la question sur la nature de la signature génique. Quelle est la taille réelle de la signature génique ?

5.2 La réelle taille d'une signature génique

Le chapitre 4 détaille notre méthode d'interpolation en espace de représentation, suite à un encodage des données de transcriptomique dans un espace latent de plus petite dimensionnalité que les données d'origine.

J'ai montré qu'il est possible de transformer un exemple d'un phénotype à un autre, sans créer d'exemples adversariaux [39]. Malgré que je n'ai rapporté que les 10 gènes avec la plus haute magnitude de transformation, j'ai observé que tous les gènes nécessitent une transformation. Ceci a également été observé dans le jeu de données original de Golub (Figure 5.1). Malgré que Golub et al. n'aient séléctionné que 50 gènes différentiellement exprimés pour distinguer les trois sous-types de leucémies [37], notre modèle (VAE) rapporte que la plupart des gènes sont différentiellement exprimés. Ces résultats suggèrent que les signatures géniques sont des phénomènes impliquant une grande partie des gènes, sinon la totalité.

Les limites de la méthode de *warp* sont, cependant, de nature technique. En effet, l'autoencodeur variationel (VAE) utilise un encodage tiré d'une gaussienne [26]. Bien que ceci apporte une certaine aisance dans les analyses subséquentes de cet espace, ceci apporte également un désavantage majeur à la méthode. Il est, en effet, communément rapporté que les VAE produisent une reconstruction floue des données d'entrée, justement à cause de l'encodage tiré d'une gaussienne [51] (plus de détails dans la section 2.6.2.4).

De plus, bien que la méthode offre une piste de solution pour l'interprétabilité des réseaux de neurones et permet de mieux comprendre les données, elle ne fait que fournir

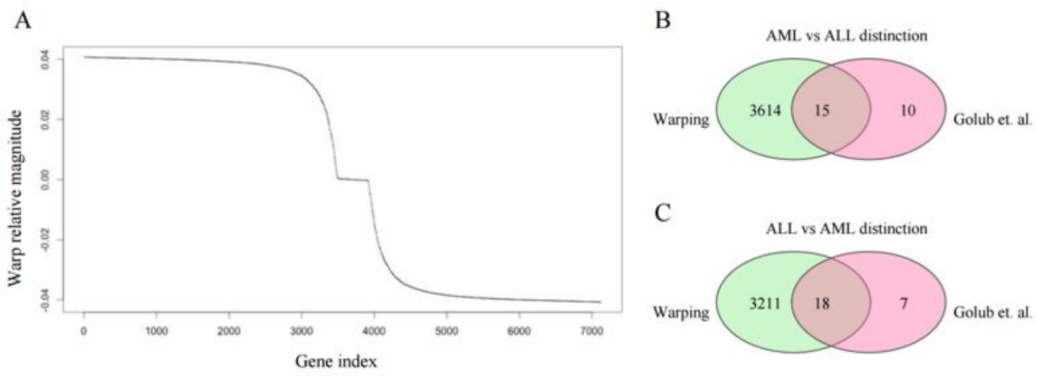


Figure 5.1 : Analyse par warp des données de Golub et collègues
A) Magnitude triée de la transformation pour chaque gène du jeu de données Golub pour changer de classe de AML vers ALL.B) Ensembles des gènes rapportés ayant une magnitude de transformation positive par les deux méthodes ainsi que leur chevauchement.

une liste de gènes différentiellement exprimées. L'information sur les mécanismes géniques ou procédés cellulaires présents dans les cellules observées ne sont accessibles que par analyses subséquentes, par exemple par enrichissement pour des GO terms. Or, la plupart des expériences de séquençages sont faites avec une attente de découvrir un mécanisme génique nouveau dans le phénotype à l'étude. Je me suis donc concentrée à trouver une méthode permettrant d'extraire simultanément des informations sur les échantillons et les gènes.

5.3 Apprentissage simultané d'un espace de gènes et d'échantillons

Mes co-auteurs et moi avons détaillé dans le chapitre 5 notre méthode permettant d'extraire à la fois une représentation des échantillons ainsi que celle des gènes. J'ai appliqué cette méthode sur le jeu de données GTEx et j'ai examiné les espaces d'encodage des échantillons et des gènes et je rapporte que i) dans l'espace des échantillons, les tissus s'aggrègent dans l'espace par type mais aussi par similitude d'expression génique, et ii) dans l'espace des gènes, les gènes s'aggrègent en fonction de leur co-expression et les réseaux de gènes tissu-spécifiques sont conservés.

5.3.1 Encodage des échantillons dans l'espace latent

J'ai observé une performance similaire à t-SNE dans l'espace des échantillons, évalué par une tâche d'apprentissage semi-supervisé (Tableau 4.I). Je m'attendais à ce que t-SNE performe mieux que notre méthode, car c'est une technique qui n'encode qu'un espace d'echantillons et pour obtenir l'encodage des gènes, il faudrait appliquer t-SNE à la matrice de données transposée. En effet, je m'attendais à avoir une performance moindre des *embeddings* factorisés parce que cette méthode gère simultanément deux espaces et est donc moins spécialisée. Vu que l'entrainement des classifieurs ne se fait que sur 2% des exemples (stratifié par classe), ceci demeure une tâche difficile, car pour certaines classes, il n'y avait qu'un point d'entrainement (minimum 1, maximum 6).

Donc si l'arrangement des points d'une classe se fait en deux groupes, se trouvant à des coordonnées opposées de l'espace d'encodage, ceci mènerait à la mauvais classification de la part des points qui ne se trouve pas dans le voisinnage du point d'entrainement choisi. Il est important de noter, également qu'aucune optimisation des hyper-paramètres n'a été faite pour cette tâche de classification, ce qui pourrait expliquer au moins en partie, la performance très inférieure du perceptron multi-couches sur l'espace des *embeddings*, cette méthode étant particuliérement dépendante des hyper-paramètres.

De plus, j'ai appliqué cette méthode au jeu de données de AML-ALL de Golub et collègues [37] et je rapporte que dans l'espace des échantillons, notre méthode sépare entriérement les 38 échantillons de leucémies en sous-types (Figure 5.2). Or, Golub et collègues n'ont travaillé que sur 36 des 38 échantillons, disant que 2 des 38 sont difficiles à classifier [37]. Je suppose que ceci est plutôt dû au fait qu'ils se limitent à 50 gènes pour la classification et suggére que la vraie distinction entre les deux phénotypes est en réalité beaucoup plus complexe (et probablement non-linéaire).

J'ai également observé que le modèle de *embeddings* exploite l'espace très différement de t-SNE. En effet, j'observe que les échantillons s'aggrégent selon un gradient d'expression génique (Figure 4.6). Il était attendu que les gènes sélectionnés soient exprimés dans plusieurs tissus à la fois (plus de détails : section 4.9.2). Je suppose qu'une mécanique similaire au *warp* (Section 4.1) serait envisageable ici, en extrapolant pour

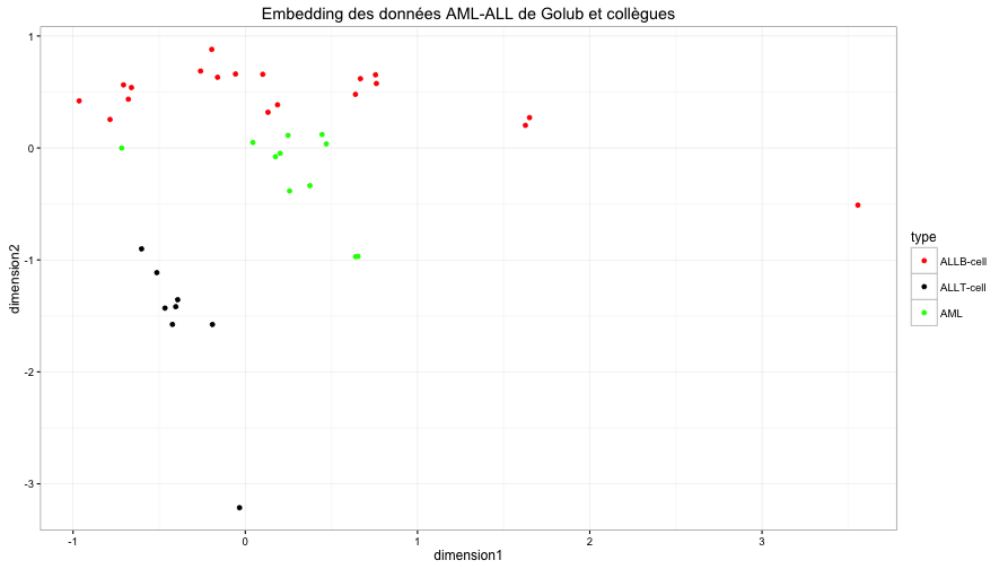


Figure 5.2 : Espace des échantillons obtenu par embedding factorisés des données AML-ALL de Golub et collègues.

l'espace entier le niveau d'expression génique nécessaire.

5.3.2 Encodage des gènes dans l'espace latent

Afin d'analyser l'espace d'encodage des gènes, j'ai été confrontée à une tâche difficile, vu qu'à la différence des échantillons, qui sont divisés en classes de tissus, il n'existe pas de classes définitives de gènes.

En examinant l'espace d'encodage des gènes, j'ai tout d'abord conclu que le modèle aggrège les gènes qui sont correlés (Figure 4.8). Ensuite j'ai voulu éxaminer où se placent dans l'espace des gènes exclusivement exprimés dans des tissus spécifiques. Pierson et collègues [80] ont construit un outil d'annotation de gènes permettant d'extraire des groupes de gènes spécifiques à certains tissus. J'ai obtenu ces groupes de gènes et je rapporte que les coordonnées de ces gènes sont plus proches les unes des autres que des coordonnées de gènes choisis aléatoirement (Figure 4.9). Ceci évoque les résultats de Mikolov et collègues, où des mots ayant un sens similaire étaient proches dans l'espace d'encodage [67]. Ma conclusion est donc que l'espace des gènes est représentatif des interactions géniques fonctionelles rapportées dans la littérature.

5.3.3 Biais relatifs aux méthodes de prétraitement des données

Dans ce travail j'ai utilisé des jeux de données pré-quantifiés. Ceci signifie que mes méthodes proposées (Chapitre 4) font abstraction de la notion d'épissage alternatif (alternative splicing) et de variant muté. En effet, comme la séquence brute n'est pas pris en compte, la quantification est basée uniquement sur l'annotation et le génome de référence. L'étape de mapping devient problématique, où les observations sur lesquelles je bâtis le modèle sont largement biaisées par l'outil de mapping choisi.

5.4 Conclusion

J'ai observé que i) la variabilité d'une signature génique est très dépendante du seuil d'inclusion de données choisi et ii) qu'il y a peu de reproducibilité dans les études transcriptomiques [8, 10, 83]. J'en conclus que les signatures géniques dans les maladies complexes, telles le cancer, seraient plutôt omnigéniques que polygéniques ou monogéniques et que les méthodes d'apprentissage machine peuvent offrir une certaine interprétabilité des données biologiques.

De plus, par leur grande capacité, les méthodes d'apprentissage profond seraient les candidates idéales pour mettre en place un système pouvant intégrer les données biologiques acquises dans le passé aux données présentes et futures permettant ainsi de tirer profit de tout ce qui a été fait en génomique à ce jour. J'anticipe donc la possibilité d'intégrer plusieurs types d'expériences dans le modèle de *embeddings*, par l'ajout d'un espace de *embeddings* supplémentaire, pour obtenir une vue d'ensemble sur ce qui est connu en biologie. Par exemple, le modèle d'*embeddings* offre une piste de solution pour les problèmes courants en séquençage de cellule unique (single-cell, scRNASeq). En effet, ce domaine fleurissant se heurte à la problématique de manquer de couverture et de profondeur de séquençage et de nombreux efforts sont déployés pour tenter d'extrapoler la valeur d'expression de gènes qui ne sont pas quantifiés. Or, l'architecture de du modèle de *embeddings* ne requiert pas une valeur d'expression pour tous les gènes et est robuste aux valeurs manquantes et donc pourrait être une solution viable à une embûche d'une technique ayant tant à offrir. En effet, les expériences scRNA-Seq

offrent une polyvalence incomparable, dans l'étude de sous-types cellulaires, que ce soit le microenvironnement tumoral ou même la niche hématopoïétique.

De plus, un système d'intelligence artificielle qui intègre la biologie dans tous ses détails pourrait aider dans la caractérisation de phénotypes et maladies complexes, le choix du meilleur traitement ou même, dans le contexte de greffe d'organes, faire le choix entre un donneur compatible ou non. Il serait donc impératif de poursuivre les efforts vers un système d'intelligence artificielle tenant compte de toutes les données biologiques disponibles.

Finalement, je compte explorer dans un projet futur des options permettant d'éviter les biais reliés aux traitements des données post-séquençage (mapping, normalisation, quantification) [83, 86]. Ceci sera exploré en modifiant mon approche pour utiliser directement les séquences produites par le séquenceur, sans prétraitement ni alignement, probablement utilisant une approche par k-mers (n-grams) [63]. Utiliser la séquence brute via une approche par k-mers exposerait l'algorithme à une multitude d'informations supplémentaires. En effet, il est envisageable que la différence entre un phénotype malade et sain serait encodée dans la séquence même, et non seulement dans le niveau d'expression. Aussi, il serait possible que la signature génique est également exprimée dans l'ARN non-aligné ou non-présent dans l'annotation choisie. Finalement, une approche par k-mers introduirait la notion de similarité de séquence entre les gènes, permettant de catégoriser par domaines ou même familles de séquences.

BIBLIOGRAPHIE

[1] 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. *Nature genetics*, 25(1):25–9, may 2000.

[2] Stuart P Atkinson, Joseph Collin, Neganova Irina, George Anyfantis, Bo Kim Kyung, Majlinda Lako et Lyle Armstrong. A putative role for the immunoproteasome in the maintenance of pluripotency in human embryonic stem cells. *Stem cells (Dayton, Ohio)*, 30(7):1373–84, jul 2012.

[3] Michael Basler, Christopher J Kirk et Marcus Groettrup. The immunoproteasome in antigen processing and other immunological functions. *Current opinion in immunology*, 25(1):74–80, feb 2013.

[4] Michael Basler, Sarah Mundt, Tony Muchamuel, Carlo Moll, Jing Jiang, Marcus Groettrup et Christopher J Kirk. Inhibition of the immunoproteasome ameliorates experimental autoimmune encephalomyelitis. *EMBO Molecular Medicine*, 6(2): 226–238, jan 2014.

[5] Stephen B Baylin. DNA methylation and gene silencing in cancer. *Nature clinical practice. Oncology*, 2 Suppl 1:S4–11, dec 2005.

[6] Richard Bellman. *Dynamic programming*. Dover Publications, 2003. ISBN 0486428095.

[7] Caterina Bendotti, Marianna Marino, Cristina Cheroni, Elena Fontana, Valeria Crippa, Angelo Poletti et Silvia De Biasi. Dysfunction of constitutive and inducible ubiquitin-proteasome system in amyotrophic lateral sclerosis : implication for protein aggregation and immune response. *Progress in neurobiology*, 97(2): 101–26, may 2012.

[8] Christophe Benoist, Ronald N. Germain et Diane Mathis. A Plaidoyer for 'Systems Immunology'. *Immunological Reviews*, 210(1):229–234, apr 2006.

[9] Philip S. Bernard, Joel S. Parker, Michael Mullins, Maggie C U Cheung, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenhush, Inge J. Stijleman, Juan Palazzo, J. S. Matron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis et Charles M. Perou. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, mar 2009.

[10] Paul C Boutros, Suzanne K Lau, Melania Pintilie, Ni Liu, Frances A Shepherd, Sandy D Der, Ming-Sound Tsao, Linda Z Penn et Igor Jurisica. Prognostic gene signatures for non-small-cell lung cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 106(8):2824–8, feb 2009.

[11] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[12] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub et Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(12):4164–9, mar 2004.

[13] Cancer Genome Atlas Research Network, Timothy J Ley, Christopher Miller, Li Ding, Benjamin J Raphael, Andrew J Mungall, A Gordon Robertson, Katherine Hoadley, Timothy J Triche, Peter W Laird, Jack D Baty, Lucinda L Fulton, Robert Fulton, Sharon E Heath, Joelle Kalicki-Veizer, Cyriac Kandoth, Jeffery M Klco, Daniel C Koboldt, Krishna-Latha Kanchi, Shashikant Kulkarni, Tamara L Lamprecht, David E Larson, Ling Lin, Charles Lu, Michael D McLellan, Joshua F McMichael, Jacqueline Payton, Heather Schmidt, David H Spencer, Michael H Tomasson, John W Wallis, Lukas D Wartman, Mark A Watson, John Welch, Michael C Wendl, Adrian Ally, Miruna Balasundaram, Inanc Birol, Yaron Butterfield, Readman Chiu, Andy Chu, Eric Chuah, Hye-Jung Chun, Richard Corbett, Noreen Dhalla, Ranabir Guin, An He, Carrie Hirst, Martin Hirst,

Robert A Holt, Steven Jones, Aly Karsan, Darlene Lee, Haiyan I Li, Marco A Marra, Michael Mayo, Richard A Moore, Karen Mungall, Jeremy Parker, Erin Pleasance, Patrick Plettner, Jacquie Schein, Dominik Stoll, Lucas Swanson, Angela Tam, Nina Thiessen, Richard Varhol, Natasja Wye, Yongjun Zhao, Stacey Gabriel, Gad Getz, Carrie Sougnez, Lihua Zou, Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Frederick Applebaum, Stephen B Baylin, Rehan Akbani, Bradley M Broom, Ken Chen, Thomas C Motter, Khanh Nguyen, John N Weinstein, Nianziang Zhang, Martin L Ferguson, Christopher Adams, Aaron Black, Jay Bowen, Julie Gastier-Foster, Thomas Grossman, Tara Lichtenberg, Lisa Wise, Tanja Davidsen, John A Demchok, Kenna R Mills Shaw, Margi Sheth, Heidi J Sofia, Liming Yang, James R Downing et Greg Eley. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine*, 368(22):2059–74, may 2013.

[14] Chiao-Nan Joyce Chen, Ted G Graber, Wendy M Bratten, Deborah A Ferrington et LaDora V Thompson. Immunoproteasome in animal models of Duchenne muscular dystrophy. *Journal of muscle research and cell motility*, 35(2):191–201, apr 2014.

[15] Frederic Chibon. Cancer gene expression signatures-The rise and fall ? *European Journal of Cancer*, 49(8):2000–2009, may 2013.

[16] Francois Chollet. Keras, 2015.

[17] Judith K Christman. 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation : mechanistic studies and their implications for cancer therapy. *Oncogene*, 21(35):5483–5495, aug 2002.

[18] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang et Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome biology*, 17:13, jan 2016.

[19] T. Cover et P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, jan 1967.

[20] Nico P. Dantuma et Laura C. Bott. The ubiquitin-proteasome system in neurodegenerative diseases : precipitating factor, yet part of the solution. *Frontiers in Molecular Neuroscience*, 7, jul 2014.

[21] Gerjan de Bruin, Eva M. Huber, Bo-Tao Xin, Eva J. van Rooden, Karol Al-Ayed, Kyung-Bo Kim, Alexei F. Kisselev, Christoph Driessen, Mario van der Stelt, Gijsbert A. van der Marel, Michael Groll et Herman S. Overkleeft. Structure-Based Design of β1i or β5i Specific Inhibitors of Human Immunoproteasomes. *Journal of Medicinal Chemistry*, 57(14):6197–6209, jul 2014.

[22] D.A. De Verteuil, A. Rouette, M.-P. Hardy, S. Lavallée, A. Trofimov, E. Gaucher et C. Perreault. Immunoproteasomes shape the transcriptome and regulate the function of dendritic cells. *Journal of Immunology*, 193(3), 2014.

[23] Danielle de Verteuil, Tara L Muratore-Schroeder, Diana P Granados, Marie-Hélène Fortier, Marie-Pierre Hardy, Alexandre Bramoullé, Etienne Caron, Krystel Vincent, Sylvie Mader, Sébastien Lemieux, Pierre Thibault et Claude Perreault. Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules. *Molecular & cellular proteomics : MCP*, 9(9):2034–47, sep 2010.

[24] Susan C. Hu Der-Chiang Li, Chiao-Wen Liu. A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medecine*, 40:509–518, may 2010.

[25] Christine Desmedt, Benjamin Haibe-Kains, Pratyaksha Wirapati, Marc Buyse, Denis Larsimont, Gianluca Bontempi, Mauro Delorenzi, Martine Piccart et Christos Sotiriou. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(16):5158–65, aug 2008.

[26] Carl Doersch. Tutorial on Variational Autoencoders. jun 2016.

[27] Hartmut Döhner, Daniel J. Weisdorf et Clara D. Bloomfield. Acute Myeloid Leukemia. *New England Journal of Medicine*, 373(12):1136–1152, sep 2015.

[28] Guilherme Augusto Dos Santos, Lev Kats et Pier Paolo Pandolfi. Synergy against PML-RARa : targeting transcription, proteolysis, differentiation, and self-renewal in acute promyelocytic leukemia. *The Journal of experimental medicine*, 210(13): 2793–802, dec 2013.

[29] Cheryl M Ethen, Stacy A Hussong, Cavan Reilly, Xiao Feng, Timothy W Olsen et Deborah A Ferrington. Transformation of the proteasome with age-related macular degeneration. *FEBS letters*, 581(5):885–90, mar 2007.

[30] Cédric Févotte et Jérôme Idier. Algorithms for Nonnegative Matrix Factorization with the β-Divergence. *Neural Computation*, 23(9):2421–2456, sep 2011.

[31] Karin Flick et Peter Kaiser. Protein degradation and the stress response. *Seminars in Cell & Developmental Biology*, 23(5):515–522, jul 2012.

[32] Manfred Fliegauf, Christian Fröhlich, Judit Horvath, Heike Olbrich, Friedhelm Hildebrandt et Heymut Omran. Identification of the human CYS1 gene and candidate gene analysis in Boichis disease. *Pediatric nephrology (Berlin, Germany)*, 18(6):498–505, jun 2003.

[33] Yoav Freund et Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*. *Journal of Computer and System Sciences*, 55, 1997.

[34] Wolf Herman Fridman, Jérôme Galon, Franck Pagès, Eric Tartour, Catheriné Sautès-Fridman et Guido Kroemer. Prognostic and predictive impact of intra- and peritumoral immune infiltrates. *Cancer research*, 71(17):5601–5, sep 2011.

[35] Mathew J. Garnett, Elena J. Edelman, Sonja J. Heidorn, Chris D. Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I. Richard Thompson, Xi Luo,

Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy J. Milano, Graham R. Bignell, Ah T. Tam, Helen Davies, Jesse A. Stevenson, Syd Barthorpe, Stephen R. Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xeni Mitropoulos, Tatiana Mironenko, Helen Thi, Laura Richardson, Wenjun Zhou, Frances Jewitt, Tinghu Zhang, Patrick O'Brien, Jessica L. Boisvert, Stacey Price, Wooyoung Hur, Wanjuan Yang, Xianming Deng, Adam Butler, Hwan Geun Choi, Jae Won Chang, Jose Baselga, Ivan Stamenkovic, Jeffrey A. Engelman, Sreenath V. Sharma, Olivier Delattre, Julio Saez-Rodriguez, Nathanael S. Gray, Jeffrey Settleman, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Sridhar Ramaswamy, Ultan McDermott et Cyril H. Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483 (7391):570–575, mar 2012.

[36] Kimberly Glass et Michelle Girvan. Annotation Enrichment Analysis : An Alternative Method for Evaluating the Functional Properties of Gene Sets. *Scientific Reports*, 4(1):4191, may 2015.

[37] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield et E S Lander. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, 286(5439):531–7, oct 1999.

[38] Ian Goodfellow, Yoshua Bengio et Aaron Courville. *Deep learning*. 2016. ISBN 9780262035613.

[39] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio. Generative Adversarial Networks. 2014.

[40] Diana P Granados, Pierre-Luc Tanguay, Marie-Pierre Hardy, Etienne Caron, Danielle de Verteuil, Sylvain Meloche et Claude Perreault. ER stress affects processing of MHC class I-associated peptides. *BMC immunology*, 10(1):10, feb 2009.

[41] Benoît Guillaume, Jacques Chapiro, Vincent Stroobant, Didier Colau, Benoît Van Holle, Grégory Parvizi, Marie-Pierre Bousquet-Dubouch, Ivan Théate, Nicolas Parmentier et Benoît J Van den Eynde. Two abundant proteasome subtypes that uniquely process some antigens presented by HLA class I molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 107(43): 18599–604, oct 2010.

[42] Benoît Guillaume, Vincent Stroobant, Marie-Pierre Bousquet-Dubouch, Didier Colau, Jacques Chapiro, Nicolas Parmentier, Alexandre Dalet et Benoît J Van den Eynde. Analysis of the processing of seven human tumor antigens by intermediate proteasomes. *Journal of immunology (Baltimore, Md. : 1950)*, 189(7):3538–47, oct 2012.

[43] Sylvia Heink, Daniela Ludwig, Peter-M Kloetzel et Elke Krüger. IFN-gamma-induced immune adaptation of the proteasome system is an accelerated and transient response. *Proceedings of the National Academy of Sciences of the United States of America*, 102(26):9241–6, jun 2005.

[44] B A Hills. An alternative view of the role(s) of surfactant and the alveolar model. *Journal of applied physiology (Bethesda, Md. : 1985)*, 87(5):1567–83, nov 1999.

[45] Mohit Jain, Roland Nilsson, Sonia Sharma, Nikhil Madhusudhan, Toshimori Kitami, Amanda L Souza, Ran Kafri, Marc W Kirschner, Clary B Clish et Vamsi K Mootha. Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science (New York, N.Y.)*, 336(6084):1040–4, may 2012.

[46] Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, sep 1967.

[47] Andrew J Kale et Bradley S Moore. Molecular mechanisms of acquired proteasome inhibitor resistance. *Journal of medicinal chemistry*, 55(23):10317–27, dec 2012.

[48] Sander Kelderman, Ton N M Schumacher et John B A G Haanen. Acquired and intrinsic resistance in cancer immunotherapy. *Molecular oncology*, 8(6):1132–9, sep 2014.

[49] Ilona E Keller, Oliver Vosyka, Shinji Takenaka, Alexander Kloß, Burkhardt Dahlmann, Lianne I Willems, Martijn Verdoes, Hermen S Overkleeft, Elisabeth Marcos, Serge Adnot, Stefanie M Hauck, Clemens Ruppert, Andreas Günther, Susanne Herold, Shinji Ohno, Heiko Adler, Oliver Eickelberg et Silke Meiners. Regulation of immunoproteasome function in the lung. *Scientific reports*, 5(1): 10230, may 2015.

[50] S Khan, M van den Broek, K Schwarz, R de Giuli, P A Diener et M Groettrup. Immunoproteasomes largely replace constitutive proteasomes during an antiviral and antibacterial immune response in the liver. *Journal of immunology (Baltimore, Md. : 1950)*, 167(12):6859–68, dec 2001.

[51] Diederik P Kingma et Max Welling. Auto-Encoding Variational Bayes. dec 2013.

[52] Christiaan Klijn, Steffen Durinck, Eric W Stawiski, Peter M Haverty, Zhaoshi Jiang, Hanbin Liu, Jeremiah Degenhardt, Oleg Mayba, Florian Gnad, Jinfeng Liu, Gregoire Pau, Jens Reeder, Yi Cao, Kiran Mukhyala, Suresh K Selvaraj, Mamie Yu, Gregory J Zynda, Matthew J Brauer, Thomas D Wu, Robert C Gentleman, Gerard Manning, Robert L Yauch, Richard Bourgon, David Stokoe, Zora Modrusan, Richard M Neve, Frederic J de Sauvage, Jeffrey Settleman, Somasekar Seshagiri et Zemin Zhang. A comprehensive transcriptional portrait of human cancer cell lines. *Nature biotechnology*, 33(3):306–12, mar 2015.

[53] Eleni D Lagadinou, Alexander Sach, Kevin Callahan, Randall M Rossi, Sarah J Neering, Mohammad Minhajuddin, John M Ashton, Shanshan Pei, Valerie Grose, Kristen M O'Dwyer, Jane L Liesveld, Paul S Brookes, Michael W Becker et Craig T Jordan. BCL-2 inhibition targets oxidative phosphorylation and selectively eradicates quiescent human leukemia stem cells. *Cell stem cell*, 12(3): 329–41, mar 2013.

[54] Ben Langmead, Cole Trapnell, Mihai Pop et Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[55] Vincent-Philippe Lavallée, Irène Baccelli, Jana Krosl, Brian Wilhelm, Frédéric Barabé, Patrick Gendron, Geneviève Boucher, Sébastien Lemieux, Anne Marinier, Sylvain Meloche, Josée Hébert et Guy Sauvageau. The transcriptomic landscape and directed chemical interrogation of MLL-rearranged acute myeloid leukemias. *Nature genetics*, 47(9):1030–7, sep 2015.

[56] Vincent-Philippe Lavallée, Patrick Gendron, Sébastien Lemieux, Giovanni D'Angelo, Josée Hébert et Guy Sauvageau. EVI1-rearranged acute myeloid leukemias are characterized by distinct molecular alterations. *Blood*, 125(1):140–3, jan 2015.

[57] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard et V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. 1995.

[58] Heng Li et Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, jul 2009.

[59] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino,

Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok et Helen F Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

[60] Ji Luo, Nicole L. Solimini et Stephen J. Elledge. Principles of Cancer Therapy : Oncogene and Non-oncogene Addiction, mar 2009.

[61] J Macqueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(233):281–297, 1967.

[62] S.C. Madeira et A.L. Oliveira. Biclustering algorithms for biological data analysis : a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, jan 2004.

[63] Guillaume Marçais et Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, 27 (6):764–70, mar 2011.

[64] Daniel Markovich et Heini Murer. The SLC13 gene family of sodium sulphate/carboxylate cotransporters, feb 2004.

[65] Silke Meiners et Oliver Eickelberg. What shall we do with the damaged proteins in lung disease. Ask the proteasome !, nov 2012.

[66] Marie-Christine Meunier, Jean-Sébastien Delisle, Julie Bergeron, Vincent Rineau, Chantal Baron et Claude Perreault. T cells targeted against a single minor histocompatibility antigen can cure solid tumors. *Nature medicine*, 11(11):1222–1229, nov 2005.

[67] Tomas Mikolov, Kai Chen, Greg Corrado et Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv :1301.3781v3 [cs.CL]*, jan 2013.

[68] Zachary Miller, Wooin Lee et Kyung Kim. The Immunoproteasome as a Therapeutic Target for Hematological Malignancies. *Current Cancer Drug Targets*, 14 (6):537–548, 2014.

[69] Michele Mishto, Fabio Luciani, Hermann Georg Holzhütter, Elena Bellavista, Aurelia Santoro, Kathrin Textoris-Taube, Claudio Franceschi, Peter M Kloetzel et Alexey Zaikin. Modeling the in Vitro 20S Proteasome Activity : The Effect of PA28-ab and of the Sequence and Length of Polypeptides on the Degradation Kinetics. *Journal of Molecular Biology*, 377(5):1607–1617, apr 2008.

[70] Bernhard Mlecnik, Gabriela Bindea, Helen K Angell, Pauline Maby, Mihaela Angelova, David Tougeron, Sarah E Church, Lucie Lafontaine, Maria Fischer, Tessa Fredriksen, Maristella Sasso, Amélie M Bilocq, Amos Kirilovsky, Anna C Obenauf, Mohamad Hamieh, Anne Berger, Patrick Bruneval, Jean Jacques Tuech, Jean Christophe Sabourin, Florence Le Pessot, Jacques Mauillon, Arash Rafii, Pierre Laurent-Puig, Michael R Speicher, Zlatko Trajanoski, Pierre Michel, Richard Sesboüe, Thierry Frebourg, Franck Pagès, Viia Valge-Archer, Jean Baptiste Latouche et Jérôme Galon. Integrative Analyses of Colorectal Cancer Show Im-

munoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability. *Immunity*, 44(3):698–711, mar 2016.

[71] Bernhard Mlecnik, Gabriela Bindea, Amos Kirilovsky, Helen K Angell, Anna C Obenauf, Marie Tosolini, Sarah E Church, Pauline Maby, Angela Vasaturo, Mihaela Angelova, Tessa Fredriksen, Stéphanie Mauger, Maximilian Waldner, Anne Berger, Michael R Speicher, F. Pages, Viia Valge-Archer et Jérôme Galon. The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis. *Science Translational Medicine*, 8(327):327ra26–327ra26, feb 2016.

[72] Tony Muchamuel, Michael Basler, Monette A Aujay, Erika Suzuki, Khalid W Kalim, Christoph Lauer, Catherine Sylvain, Eileen R Ring, Jamie Shields, Jing Jiang, Peter Shwonek, Francesco Parlati, Susan D Demo, Mark K Bennett, Christopher J Kirk et Marcus Groettrup. A selective inhibitor of the immunoproteasome subunit LMP7 blocks cytokine production and attenuates progression of experimental arthritis. *Nature Medicine*, 15(7):781–787, jul 2009.

[73] L Muñoz, J F Nomdedéu, N Villamor, R Guardia, D Colomer, J M Ribera, J P Torres, J J Berlanga, C Fernández, A Llorente, M P Queipo de Llano, J M Sánchez, S Brunet et J Sierra. Acute myeloid leukemia with MLL rearrangements : clinicobiological features, prognostic impact and value of flow cytometry in the detection of residual leukemic cells. *Leukemia*, 17(1):76–82, jan 2003.

[74] Veronique Nogueira et Nissim Hay. Molecular pathways : Reactive oxygen species homeostasis in cancer cells and implications for cancer therapy. *Clinical Cancer Research*, 19(16):4309–4314, aug 2013.

[75] Rebecca Nugent et Marina Meila. An overview of clustering applied to molecular biology., 2010.

[76] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa

Bono et Minoru Kanehisa. KEGG : Kyoto encyclopedia of genes and genomes, jan 1999.

[77] Alicia Oshlack, Mark D Robinson et Matthew D Young. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):220, 2010.

[78] Alberto Pascual-Montano, Pedro Carmona-Saez, Monica Chagoyen, Francisco Tirado, Jose M Carazo et Roberto D Pascual-Marqui. bioNMF : a versatile tool for non-negative matrix factorization in biology. *BMC bioinformatics*, 7:366, jul 2006.

[79] Jeffrey Pennington, Richard Socher et Christopher D Manning. GloVe : Global Vectors for Word Representation. URL `https://nlp.stanford.edu/pubs/glove.pdf`.

[80] Emma Pierson, Daphne Koller, Alexis Battle, Sara Mostafavi et Sara Mostafavi. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLOS Computational Biology*, 11(5):e1004220, may 2015.

[81] Harun Pirim, Burak EkÅ§ioÄ§lu, Andy Perkins et Cetin Yüceer. Clustering of High Throughput Gene Expression Data. *Computers & operations research*, 39 (12):3046–3061, dec 2012.

[82] Rosario Michael Piro, Ugo Ala, Ivan Molineris, Elena Grassi, Chiara Bracco, Gian Paolo Perego, Paolo Provero et Ferdinando Di Cunto. An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. *European Journal of Human Genetics*, 19(11):1173–1180, nov 2011.

[83] George Poste. Bring on the biomarkers. *Nature*, 469(7329):156–157, jan 2011.

[84] J R Quinlan. Induction of Decision Trees. *Machine Learning*, 1:81–106, 1986.

[85] I. Rish et I. Rish. An empirical study of the naive bayes classifier. 2001.

[86] Christelle Robert et Mick Watson. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biology*, 16(1):177, jan 2015.

[87] A. Rouette, A. Trofimov, D. Haberl, G. Boucher, V P Lavallee, G. D'Angelo, J Hebert, G. Sauvageau, S. Lemieux et C. Perreault. Expression of immunoproteasome genes is regulated by cell-intrinsic and -extrinsic factors in human cancers. *Scientific reports*, 6(September):34019, 2016.

[88] Ruslan Salakhutdinov et Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, jul 2009.

[89] Florian A Salomons, V. Menendez-Benito, C. Bottcher, Brett A McCray, J Paul Taylor et Nico P Dantuma. Selective Accumulation of Aggregation-Prone Proteasome Substrates in Response to Proteotoxic Stress. *Molecular and Cellular Biology*, 29(7):1774–1785, apr 2009.

[90] Jonathon Shlens. A tutorial on principal component analysis. *Internet Article*, pages 1–13, apr 2005.

[91] Karen Simonyan, Andrea Vedaldi et Andrew Zisserman. Deep Inside Convolutional Networks : Visualising Image Classification Models and Saliency Maps. 2013.

[92] Robert Sokal et Charles Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin.*, 38:1409–1438, 1958.

[93] T Sørlie, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matese, P O Brown, D Botstein, P E Lønning et A L Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19): 10869–74, sep 2001.

[94] Shrivani Sriskanthadevan, Danny V Jeyaraju, Timothy E Chung, Swayam Prabha, Wei Xu, Marko Skrtic, Bozhena Jhas, Rose Hurren, Marcela Gronda, Xiaoming Wang, Yulia Jitkova, Mahadeo A Sukhai, Feng Hsu Lin, Neil Maclean, Rob Laister, Carolyn A Goard, Peter J Mullen, Stephanie Xie, Linda Z Penn, Ian M Rogers,

John E Dick, Mark D Minden et Aaron D Schimmer. AML cells have low spare reserve capacity in their respiratory chain that renders them susceptible to oxidative metabolic stress. *Blood*, 125(13):2120–2130, mar 2015.

[95] Charles St-Pierre, Assya Trofimov, Sylvie Brochu, Sébastien Lemieux et Claude Perreault. Differential Features of AIRE-Induced and AIRE-Independent Promiscuous Gene Expression in Thymic Epithelial Cells. *The Journal of Immunology*, 195(2):498–506, 2015.

[96] Fran Supek, Matko Bošnjak, Nives Škunca et Tomislav Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7):e21800, jul 2011.

[97] Jaanus Suurväli, Luc Jouneau, Dominique Thépot, Simona Grusea, Pierre Pontarotti, Louis Du Pasquier, Sirje Rüütel Boudinot et Pierre Boudinot. The Proto-MHC of Placozoans, a Region Specialized in Cellular Stress and Ubiquitination/Proteasome Pathways. *The Journal of Immunology*, 193(6):2891–2901, sep 2014.

[98] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow et Rob Fergus. Intriguing properties of neural networks. 2013.

[99] Pablo Tamayo, Daniel Scanfeld, Benjamin L Ebert, Michael A Gillette, Charles W M Roberts et Jill P Mesirov. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proceedings of the National Academy of Sciences of the United States of America*, 104(14):5959–64, apr 2007.

[100] Theano Development Team. Theano : A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, page 19, may 2016.

[101] R E M Toes, A K Nussbaum, S Degermann, M Schirle, N P N Emmerich, M Kraft, C Laplace, A Zwinderman, T P Dick, J Müller, B Schönfisch, C Schmid, H J Feh-

ling, S Stevanovic, H G Rammensee, H Schild, J Muller, B Schonfisch, C Schmid, H J Fehling, S Stevanovic, H G Rammensee et H Schild. Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J Exp Med*, 194(1):1–12., jul 2001.

[102] L J P Van Der Maaten et G E Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[103] L J P Van Der Maaten, E O Postma et H J Van Den Herik. Dimensionality Reduction : A Comparative Review. *Journal of Machine Learning Research*, 10:1–41, 2009.

[104] Roel G.W. Verhaak, Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, Li Ding, Todd Golub, Jill P. Mesirov, Gabriele Alexe, Michael Lawrence, Michael O'Kelly, Pablo Tamayo, Barbara A. Weir, Stacey Gabriel, Wendy Winckler, Supriya Gupta, Lakshmi Jakkula, Heidi S. Feiler, J. Graeme Hodgson, C. David James, Jann N. Sarkaria, Cameron Brennan, Ari Kahn, Paul T. Spellman, Richard K. Wilson, Terence P. Speed, Joe W. Gray, Matthew Meyerson, Gad Getz, Charles M. Perou, D. Neil Hayes et Cancer Genome Atlas Research Network. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110, jan 2010.

[105] Pascal Vincent, Hugo Larochelle, Yoshua Bengio et Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. Dans *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1096–1103, 2008. ISBN 9781605582054.

[106] Shiyu Wang et Randal J. Kaufman. The impact of the unfolded protein response on human disease, jun 2012.

[107] X Wang, Z Yan, M Fulciniti, Y Li, M Gkotzamanidou, S B Amin, P K Shah, Y Zhang, N C Munshi et C Li. Transcription factor-pathway coexpression analysis

reveals cooperation between SP1 and ESR1 on dysregulating cell cycle arrest in non-hyperdiploid multiple myeloma. *Leukemia*, 28(4):894–903, apr 2014.

[108] Nathaniel M Weathington et Rama K Mallampalli. Emerging therapies targeting the ubiquitin proteasome system in cancer, jan 2014.

[109] John R. Webb, Katy Milne, Peter Watson, Ronald J. DeLeeuw et Brad H. Nelson. Tumor-infiltrating lymphocytes expressing the tissue resident memory marker cd103 are associated with increased survival in high-grade serous ovarian cancer. *Clinical Cancer Research*, 20(2):434–444, jan 2014.

[110] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber et Bernhard Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, may 2014.

[111] Gerald Willimsky et Thomas Blankenstein. The adaptive immune response to sporadic cancer, dec 2007.

[112] Pratyaksha Wirapati, Christos Sotiriou, Susanne Kunkel, Pierre Farmer, Sylvain Pradervand, Benjamin Haibe-Kains, Christine Desmedt, Michail Ignatiadis, Thierry Sengstag, Frédéric Schütz, Darlene R Goldstein, Martine Piccart et Mauro Delorenzi. Meta-analysis of gene expression profiles in breast cancer : toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, 10(4):R65, aug 2008.

[113] F. Y. Wu. The Potts model. *Reviews of Modern Physics*, 54(1):235–268, jan 1982.

[114] Tianlei Xu, Ruixin Zhu, Qi Liu et Zhiwei Cao. Quantitatively integrating molecular structure and bioactivity profile evidence into drug-target relationship analysis. *BMC Bioinformatics*, 13(1):75, may 2012.

[115] Koon-Kiu Yan, Daifeng Wang, Joel Rozowsky, Henry Zheng, Chao Cheng et Mark Gerstein. OrthoClust : an orthology-based network framework for clustering data across multiple species. *Genome Biology*, 15(8):R100, 2014.

[116] Liubin Yang, Rachel Rau et Margaret A. Goodell. DNMT3A in haematological malignancies. *Nature Reviews Cancer*, 15(3):152–165, feb 2015.

Annexe I

Matériel supplémentaire pour le Chapitre 3

Expression of immunoproteasome genes is regulated by cell-intrinsic and -extrinsic factors in human cancers

Alexandre Rouette[1,2,6], Assya Trofimov[1,2,3,6], David Haberl[1], Geneviève Boucher[1], Vincent-Philippe Lavallée[1,2,4,5], Giovanni D'Angelo[4,5], Josée Hébert[1,2,4,5], Guy Sauvageau[1,2,4,5], Sébastien Lemieux[1,3] and Claude Perreault[1,2,4*]

[1]Institute for Research in Immunology and Cancer; [2]Department of Medicine and [3]Department of Computer Science and Operations Research, Université de Montréal; [4]Division of Hematology-Oncology and [5]Quebec Leukemia Cell Bank, Maisonneuve-Rosemont Hospital, Montreal, Quebec, Canada; [6]These authors contributed equally to this work

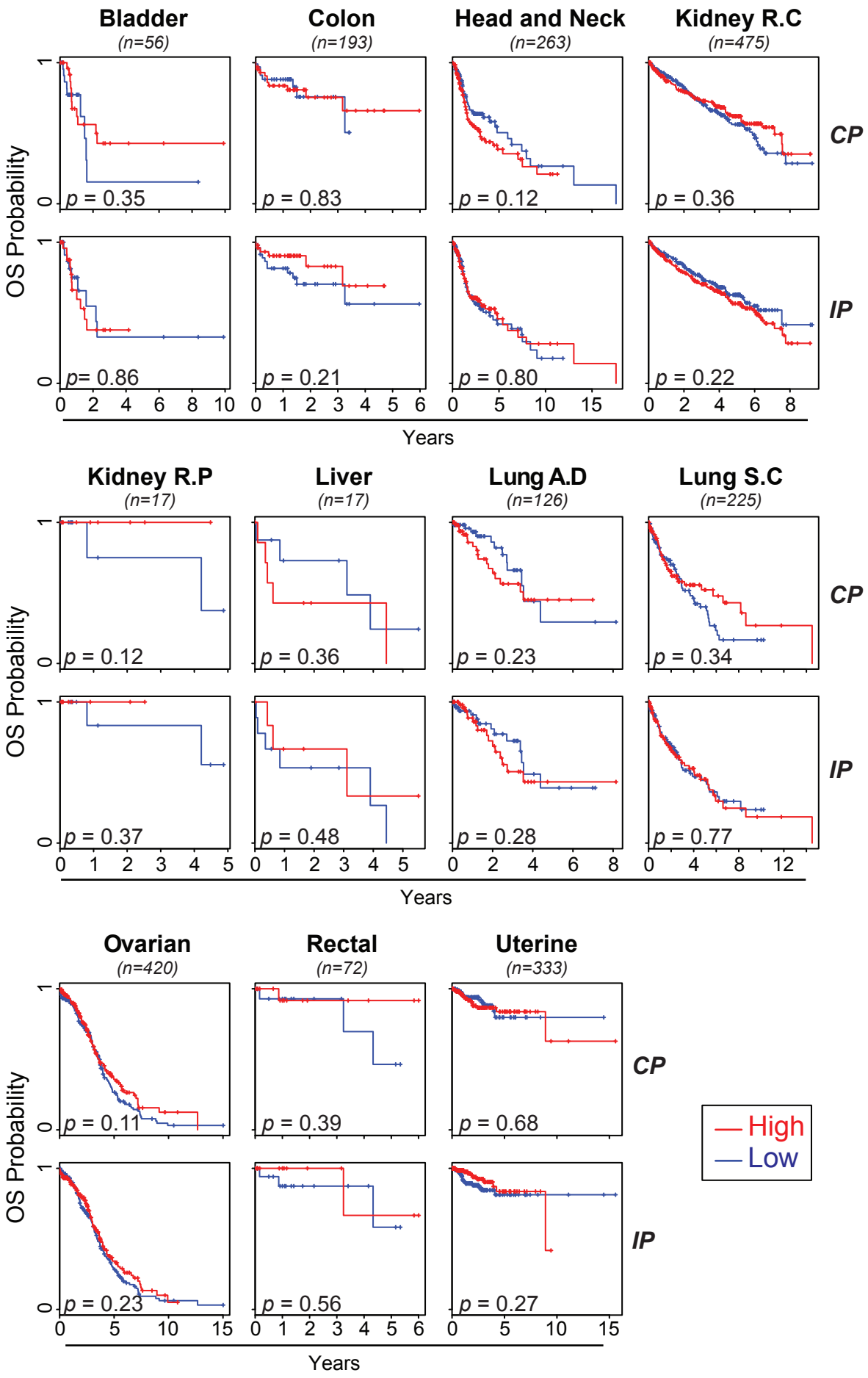
This Appendix includes :

Breast cancer		
Gene Name	Hazard Ratio [95% Confidence Interval]	*p*-value
PSMB5	0.56 [0.15 – 2.12]	0.393
PSMB6	0.47 [0.11 – 2.06]	0.316
PSMB7	1.49 [0.29 – 7.65]	0.634
PSMB8	0.34 [0.18 – 0.63]	0.001
PSMB9	0.60 [0.34 – 1.04]	0.067
PSMB10	0.31 [0.14 – 0.66]	0.003
AML		
Gene Name	Hazard Ratio [95% Confidence Interval]	*p*-value
PSMB5	2.95 [0.82 – 10.7]	0.099
PSMB6	2.82 [0.66 – 12.1]	0.162
PSMB7	2.14 [0.19 – 24.4]	0.541
PSMB8	6.30 [1.56 – 25.6]	0.001
PSMB9	2.25 [0.88 – 5.73]	0.088
PSMB10	2.85 [0.91 – 8.93]	0.072

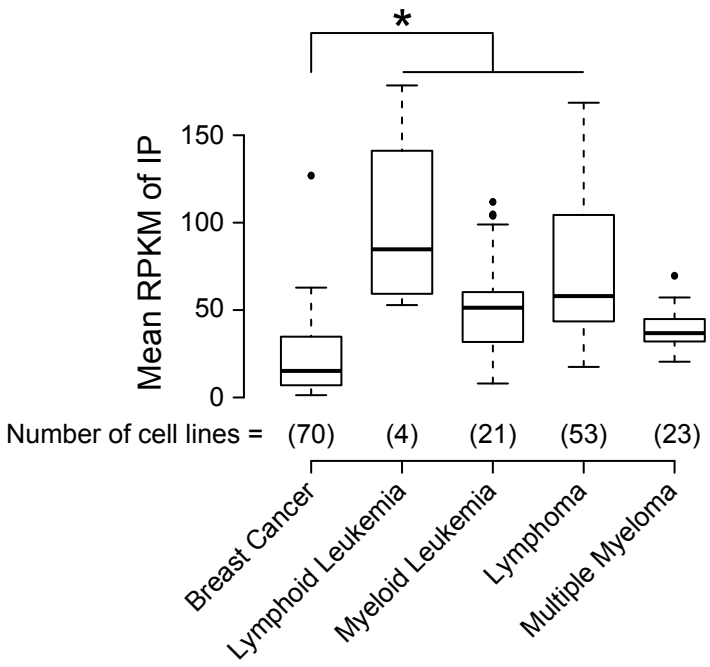
Supplementary Table S1. Risk of death associated to expression of proteasome encoding genes. Cox proportional hazards models were used to estimate hazard ratios, 95% confidence intervals and *p*-values associated with expression of proteasome-encoding genes as linear covariates.

Functional Category	M5 *PSMB8*	M5 *PSMB9*	M5 *PSMB10*	non-M5 *PSMB8*	non-M5 *PSMB9*	non-M5 *PSMB10*
Apoptotic process	3.3	1.5	0.0	3.0	2.1	0.0
Cell cycle	11.5	9.1	0.0	5.1	5.2	7.0
Death	1.6	1.5	0.0	0.0	0.0	0.0
DNA damage response	16.4	0.0	0.0	0.0	4.2	1.8
Immune process	6.6	10.6	0.0	74.8	59.4	26.3
Macromolecule catabolism	19.7	0.0	0.0	13.1	5.2	15.8
Membrane invagination	11.5	0.0	2.8	0.0	0.0	0.0
Mitochondria	23.0	0.0	0.0	0.0	0.0	3.5
Multi-organism process	1.6	1.5	0.0	1.0	1.0	5.3
Viral process	4.9	1.5	0.0	0.0	0.0	0.0
Metabolism and gene expression	0.0	56.1	75.0	1.0	11.5	28.1
Stress response	0.0	18.2	0.0	0.0	1.0	3.5
Cell signaling	0.0	0.0	22.2	2.0	6.3	3.5
Extracellular transport	0.0	0.0	0.0	0.0	4.2	0.0
Cellular component organisation	0.0	0.0	0.0	0.0	0.0	5.3

Supplementary Table S2. Functional categories associated to genes specifically correlating with IP in M5 or non-M5 AMLs. Enriched gene-ontology terms for genes specifically correlated to IP genes in M5 and non-M5 AMLs were assigned to functional categories using Revigo (see Figure 6 and Methods). Each functional category is depicted as a percentage of total categories for its column and plotted in Figure 6.



Supplementary Figure S1. Correlation of CP or IP expression with survival outcome. Patients were divided in two equal groups based on CP or IP z-score. Kaplan-Meier plots of overall survival (OS) for CP^{high}/CP^{low} patients, or IP^{high}/IP^{low} patients are shown for the indicated cancer type. *p*-values were calculated using the log-rank test. R.C: Renal Cell; R.P: Renal Papillary; A.D: Adenocarcinoma; S.C: Squamous Cell.



Supplementary Figure S2. IP expression in human cancer cell lines.
Boxplots of the mean RPKM of *PSMB8*, *PSMB9* and P*SMB10*. Numbers in parenthesis indicate the number of cell lines per disease type. Differences between means were determined by one-way analysis of variance (ANOVA) followed by Tukey's post-hoc test. * indicates $p < 0.01$.