L'utilisation des outils bioinformatiques pour caractériser le paysage immunologique du cancer de la prostate

Thèse

Benjamin Vittrant

Doctorat en médecine moléculaire
Philosophiæ doctor (Ph. D.)

Québec, Canada

L'utilisation des outils bioinformatiques pour caractériser le paysage immunologique du cancer de la prostate

Thèse

Benjamin Vittrant

Sous la direction de

Yves Fradet

Arnaud Droit

Résumé

Dans le cadre de mon doctorat j'ai développé des approches appliquées d'analyse de données pour effectuer une analyse multi omique du cancer de la prostate (CaP). Mon projet s'est défini en deux parties distinctes correspondant aux deux articles intégrés dans le corps de mon document.

Une première partie du travail a consisté à récupérer des données omiques de différents types (RNA-Seq, Methylation, CNA, SNA, miRNA, données cliniques) associées au CaP et à les préparer avec un pipeline bioinformatique adapté. Ensuite j'ai eu pour objectif de chercher à mettre en avant de nouveaux points de contrôles de l'immunité associés à la récidive biochimique (BCR) dans le CaP au travers de ces données. Pour remplir cet objectif j'ai utilisé une approche particulière basée sur des algorithmes d'analyse en composante principale (PCA) et de régression des moindres carrés (PLS). Cela a permis de faire ressortir une famille spécifique de points de contrôle de l'immunité, la famille des LILR, qui peut potentiellement être une famille cible en immunothérapie.

Dans un second temps, j'ai utilisé ces mêmes données pour développer un protocole d'analyse d'apprentissage machine (ML). Le but de ce travail était de montrer qu'il était possible de prédire si des patients allaient récidiver ou pas à partir de données RNA-Seq. J'ai montré que même avec des petits jeux de données on pouvait atteindre des scores de prédiction très bon et que les algorithmes actuels de ML prenaient bien en compte la variabilité technique de la diversité des sources de données dans le CaP. Il est donc possible d'utiliser les biobanques actuelles possédées par les structures de recherches à travers le monde pour créer des jeux de données plus importants.

Abstract

As part of my PhD, I developed applied data analysis approaches to perform a multi-omic analysis of prostate cancer (CaP). My project was split into two distinct parts corresponding to the two articles integrated into the body of my document.

A first part of the work consisted in recovering omics data of different types (RNA-Seq, Methylation, CNA, SNA, miRNA, clinical data) associated with CaP and preparing them with an adapted bioinformatics pipeline. Then, my goal was to seek to highlight new immunity checkpoints associated with biochemical recurrence (BCR) in CaP through these data. To fulfill this objective, I used a special approach based on Principal Component Analysis (PCA) and Partial Least Squares Regression (PLS) algorithms. This has brought out a specific family of immunity checkpoints, the LILR family, which can potentially be a target family in immunotherapy.

Second, I used the same data to develop a machine learning (ML) analysis protocol. The aim of this work was to show that it was possible to predict whether or not patients would relapse from RNA-Seq data. I have shown that even with small datasets, one can achieve very good prediction scores and that current ML algorithms take into account the technical variability of the diverse data sources in the CaP. It is therefore possible to use current biobanks owned by research structures around the world to create larger datasets.

Table des matières

Liste des figures

Abréviations

Les abréviations sont classées par ordre alphabétique de l'acronyme utilisé:

ADT : *Androgen deprivation therapy*
AKT : Protéine kinase B
AR : *Androgen receptor*
ATM : *Ataxia telangiectasia serine/threonine kinase mutation*
AUC : *Area under the curve*
BET : *Bromodomain and extraterminal*
CaP : Cancer de la prostate
Cor : Corrélation
Cov : Covariance
CRPC : *Castration resistant prostate cancer*
CTLs : *Cytotoxic T lymphocytes*
DCs : *Dendritic cells*
DDR : *DNA damage Repair*
EZH2 : *Enhancer of zeste homolog 2*
FGF : *Fibroblast growth factor*
HLA : *Human leukocyte antigens*
HSCs : *Hematopoetic stem cells*
IE : *Immunoediting*
mCRPC : *Metastatic castration resistant prostate cancer*
MFS : *Metastasis-free survival*
MSI-H : *Micro-sattelite instability high*
mHSPC : *Metastatic hormone sensitive prostate cancer*
MLL : *Mixed lineage leukemia*
NLR : *Neutrophil-to-lymphocyte ratio*
nmCRPC : *Non metastatic castration resistant prostate cancer*

PARP : *Poly ADP ribose polymerase*
PRC2 : *Polycomb repressive complex 2*
PSA : *Prostate specific antigen*
PSMA : *Prostate-Specific Membrane Antigen*
PTEN : *Phosphatase and tensin homolog*
SD : Déviation standard
SI : Système immunitaire
TAA : *Tumor-associated antigens*
TAM : *Tumor-associated macrophages*
***Tc** : T-cell*
TCR : *T-cell receptor*
Tex : *T-cell exhausted*
Teff : *T-cell effector*
TLRs : *Toll-like receptors*
TMA : *Transcription mediated assay*
TIME : *Tumor immune micro-environnement*
TIC : *Tumor-immunity cycle*
UICC : *Union for International Cancer Control's*
Var : Variance
Wnt : *Wingless-related integration*

Remerciements

N’ayant jamais les bons mots pour ce genre de choses j’irai droit au but !

Tout d’abord j’aimerais remercier Arnaud et Yves de m’avoir fait confiance dans le cadre de ce projet de doctorat. Je n’ai jamais été un étudiant modèle et être arrivé aussi loin dans mes études n’a pu être possible que grâce à des professeurs prêts à prendre des risques. Je rajoute donc la grande Marie-Laure Martin-Magniette à cette liste et qui a aussi accepté de faire de l’encadrement scientifique sur mon projet.

Je remercie aussi toute l’équipe de l’AD lab et du Laboratoire d’Uro-Oncologie Expérimentale avec qui j’ai pu passer de très bons moments à diverses occasions. J’espère que chacune des personnes que j’ai pu croiser trouvera sa place professionnelle dans la recherche ou ailleurs mais aussi son bonheur d’une façon ou d’une autre.

J’aimerais remercier tout particulièrement Alain Bergeron, le ministre de l’ombre, qui s’est beaucoup investi dans mon projet que ce soit humainement ou scientifiquement. La régularité des cafés à la vanille a valu bien plus que toutes les relectures et corrections à mes yeux.

Dans un autre registre, je remercie trois femmes qui ont fait le déplacement pour venir me voir au Canada et me soutenir. Ma conjointe Sophie, mon énergique mère et ma fidèle amie Félicia.

Pour le reste, le temps fera toujours son office.

Avant-propos

Article 1

Concernant le premier article je suis le premier auteur et principal rédacteur. J'ai effectué tout le le travail de récupération, traitement et d'analyse de données. Alain Bergeron est deuxième auteur et a énormément participé à la rédaction de l'article. Arnaud Droit, Mickael Leclercq et Yves Fradet ont participé au travail et à la rédaction de l'article en plus. Oscar Eduardo Molina, Valérie Picard et hélène Hovington ont apporté leur expertise en biologie et effectué les expérimentations nécessaires. Marie-Laure Martin-magniette a apporté son expertise sur tous les aspects statistiques de l'article. Julie Livingstone et Paul boutros ont fournis les données CPC-gene et aidé aux analyses comparatives concernant ces données. Colins Collins a fourni une partie des données utilisées. Cet article a été soumis à la revue *OncoImmunology* et est présentement à l'étape de révision de la nouvelle version, *i.e*. celle incluse dans cette thèse.

Article 2

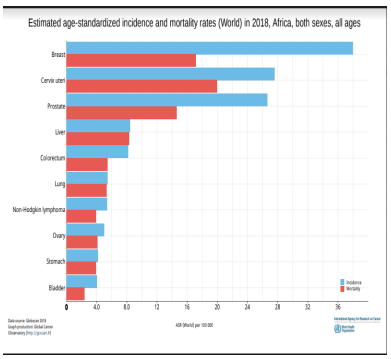
Concernant le second article je suis le premier auteur et principal rédacteur. J'ai effectué tout le le travail de récupération, traitement et d'analyse de données. Mickael Leclercq est deuxième auteur et a énormément participé à la rédaction de l'article. Arnaud Droit, Alain Bergeron et Yves Fradet ont participé au travail et à la rédaction de l'article en plus. Oscar Eduardo Molina a apporté son expertise en biologie et effectué les expérimentations nécessaires. Marie-Laure Martin-magniette a apporté son expertise sur tous les aspects statistiques de l'article. Colins Collins a fourni une partie des données utilisées. L'article a été soumis et accepté dans la revue *Frontiers in Genetic.*

Introduction

1.1 Le cancer de la prostate

1.1.a Généralités et statistiques

Le cancer est une maladie qui affecte les hommes depuis les temps anciens. Des preuves nous en sont apportées par des fossiles humains datant de la préhistoire[1]. Les civilisations ultérieures de l'Égypte ancienne et la Grèce antique nous en donnent aussi des preuves par l'archéologie[2]. Le mot cancer lui-même porte l'histoire de la maladie. Décrit dans l'Hippocratic Corpus, le mot cancer signifie crabe ou chancre en latin et est dérivé du grec καρκινος. L'étymologie exacte reste encore débattue aujourd'hui mais Hippocrate lui aurait donné ce nom par ce que le cancer “a des veines étendues de tous côtés, de même que le crabe a des pieds”.



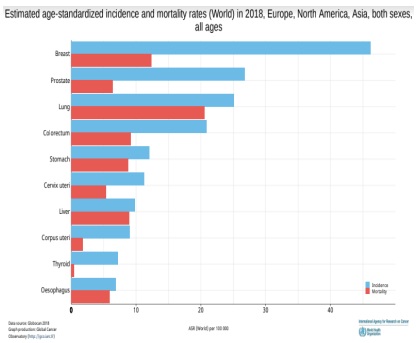


Figure 1. Incidence et mortalité : Incidence et mortalité par zone géographique et par cancer adapté du site de l’organisation mondial de la santé.

Concernant le terme prostate, en grec ancien le terme masculin '*prostates*' signifie 'président', celui qui protège, et était exclusivement utilisé comme un terme non médical. Ce n'est qu'à la Renaissance que les

anatomistes ont découvert cet organe et l'ont nommé corps glandulaire. En 1600, le physicien français De Laurens introduit le terme métaphorique 'prostatae'. Cependant, lui et ses contemporains ont mal interprété l'histoire de l'organe et le terme associé. Ils ont choisi le mauvais genre en le traduisant en latin et croyaient qu'il désignait un double organe. Ce n'est seulement dans les années 1800 que cette erreur anatomique fut corrigée tandis que l'erreur grammaticale a subsisté[3].

La prostate est une glande de la taille d'une grosse noix de Grenoble située sous la vessie. Elle est initialement petite et grossit à la puberté avec les autres caractéristiques masculines en devenir. La prostate joue un rôle dans la reproduction car elle produit un liquide qui fournit protection et nutriments aux spermatozoïdes ainsi que des enzymes facilitant leur pénétration à travers le col utérin. L'ensemble uro-génital masculin est montré en Figure 2.

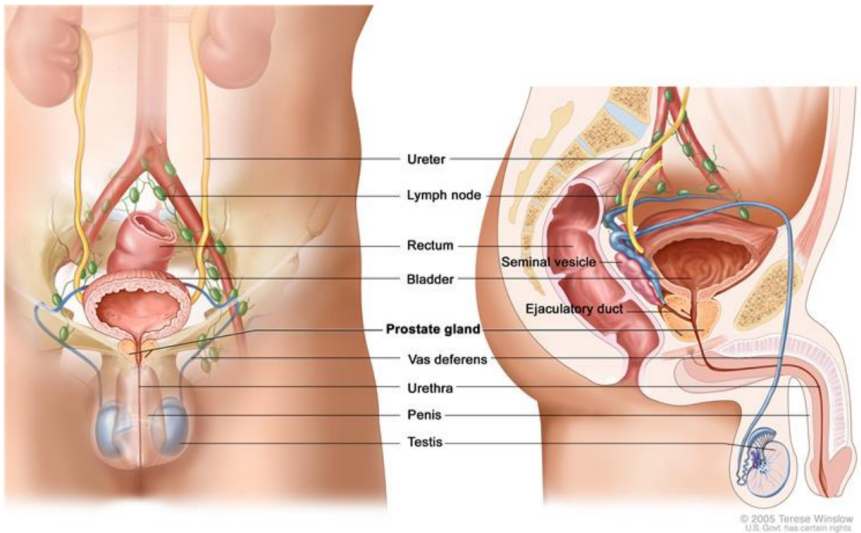


Figure 2. Système uro-genital mâle : Anatomie du système uro-génital masculin montrant entre autres, la prostate.

Le cancer de la prostate (CaP) est un des cancers les plus fréquents chez l'homme (Figure 1). En 2018, on a compté 1 276 106 nouveaux cas de

CaP dans le monde et 358 989 personnes en sont décédées. Rien qu'aux États-Unis d'Amérique, d'après le *National Institute of Health* (NIH) et la grande étude du SEER[4], on a dénombré cette même année environ 165 000 nouveaux cas de CaP et environ 30 000 décès liés à cette maladie (Tab. 1). Mais la prévalence et la mortalité sont très disparates selon les régions géographiques (Figure 3). On retrouve une plus forte prévalence du CaP dans les pays occidentaux dû notamment aux habitudes de vie et à l'alimentation ainsi qu'à une meilleure détection de la maladie, mais une mortalité plus forte dans les pays en développement à cause notamment d'une prévention et prise en charge de moins bonne qualité.

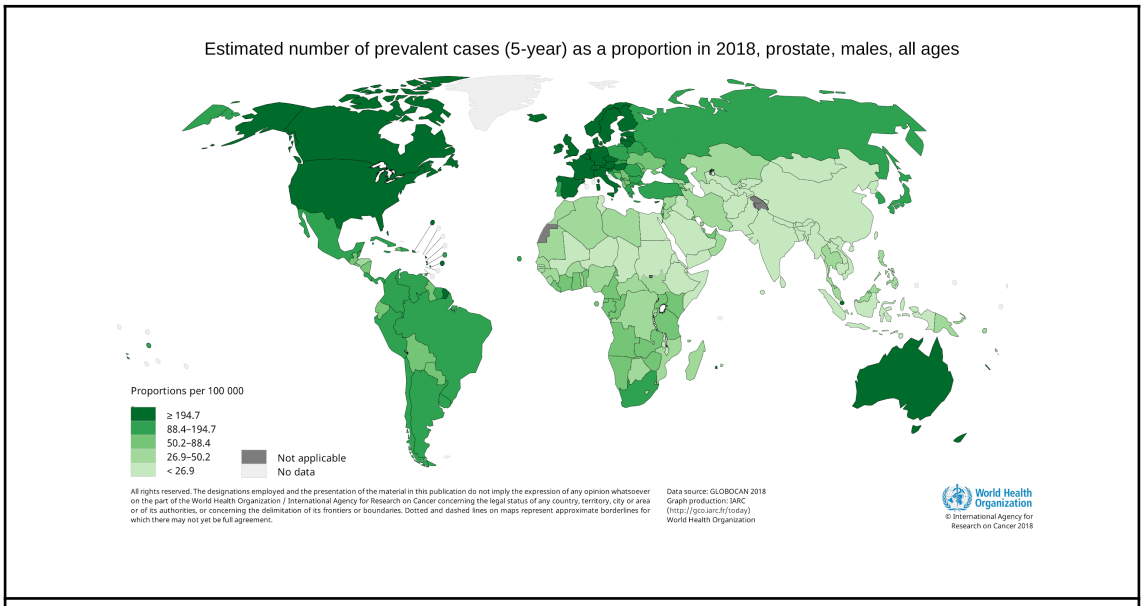


Figure 3. Incidence et mortalité : Incidence, mortalité et prévalence du CaP par zone géographique adapté du site de l'organisation mondial de la santé.

Les facteurs de risques connus[5–8] sont l'âge, les antécédents familiaux[9–11], la génétique, les races, les populations et le milieu et les habitudes de vie. Parmi celles-ci, l'alimentation, au travers la surconsommation de viande rouge, de graisses animales, le surplus de calcium et la déficience en vitamine D entre autres, joue un rôle très important. L'inflammation chronique via des infections répétées, ou des expositions régulières à des irritants pouvant créer des lésions serait un autre facteur étiologique important[12,13]. Finalement, les

taux anormaux d'hormones sexuelles mâles (*i.e*. les androgènes) constituent aussi un facteur de risque important puisque le CaP est un cancer dépendant des hormones mâles (hormono-dépendant) tel que démontré par Charles Brenton Huggins, médecin canado-américain qui a remporté le prix nobel de médecine en 1966.

1.1.b Diagnostic et évolution du CaP

Après 50 ans, les hommes doivent faire contrôler leur prostate régulièrement pour détecter précocement la présence d'un cancer. La détection peut aussi se faire au cours d'un examen préventif ou lorsque le patient consulte pour certains signes. Une suspicion peut-être émise en présence des situations suivantes:

- Faible ou flux interrompu d'urine
- Envie urgente d'uriner ou trop fréquente, particulièrement la nuit
- Difficultés à commencer à uriner
- Difficultés à vider complètement son urine
- Douleur pendant la miction
- Sang dans l'urine ou le sperme
- Douleurs continues dans le bas du dos
- Symptômes classiques de maladie: Fatigues, fièvres, anémies etc

Le test de base est le toucher rectal qui donne une première indication par un contact direct avec la prostate de la présence d'une induration. Ce test peut être accompagné d'un dosage de l'antigène prostatique spécifique ou PSA (pour *Prostate specific antigen*). La PSA est utilisée comme biomarqueur du CaP.

La PSA, aussi appelée kallikréine-3, est une sérine protéase de la famille des kallikréines codée par le gène *KLK3*[20]. Elle a été détectée puis purifiée à

partir du sérum pour la première fois en 1979[21] et 1980[22], respectivement. La PSA est une protéine de 3-33 kDa responsable de la solubilisation du sperme[23,24]. Elle est sécrétée dans les fluides séminaux par les cellules épithéliales luminales des conduits et acini de la prostate. Elle n'est pas un indicateur parfait du CaP mais donne déjà une indication de l'état de la prostate. Une PSA élevée peut-être simplement due à une hypertrophie bénigne de la prostate ou à une prostatite, par exemple.

Lorsque le taux de PSA sérique dépasse un certain seuil de référence (généralement 4.0 ng/ml) des examens plus poussés sont proposés. Le seuil de référence de la PSA peut varier selon les patients et leurs caractéristiques (*i.e.* âge, antécédents etc).

Lorsque le seuil est dépassé, les médecins vont effectuer une biopsie sous échographie transrectale afin de confirmer la présence de cancer et définir le niveau de risque via l'évaluation du grade et du stade du cancer. Depuis quelques années, les patients peuvent aussi passer un examen d'imagerie par résonance magnétique (IRM) avant la biopsie pour confirmer un problème prostatique. La biopsie prostatique reste cependant la seule méthode qui permet de poser un diagnostic précis du cancer, d'évaluer son niveau de risque et de définir les traitements à proposer au patient.

Dans certains cas, lorsqu'il y a suspicion de maladie très avancée, le médecin traitant prescrira des examens d'imagerie par tomodensitométrie et/ou scintigraphie osseuse pour détecter la présence de métastases mais il faut savoir que dans 90% des cas, le cancer est confiné à la prostate au moment du diagnostic[14]. Ainsi la détection de métastases au diagnostic est peu fréquente.

1.1.c Traitement du CaP localisé

Si le risque est faible et que le patient est éligible, il peut choisir la surveillance active. La surveillance active consiste à monitorer l'évolution possible du CaP

par la mesure du PSA aux 3-6 mois, des touchers rectaux, des examens d'IRM et/ou des biopsies prostatiques répétées jusqu'à la présence d'une évolution marquée du cancer qui nécessite une intervention plus radicale. Si le risque est moyen ou élevé, deux principales options sont offertes au patient: soit le traitement par radiothérapie ou par chirurgie[15] (Figure 4). Le choix dépend des moyens à disposition de la structure de soin et de la décision du patient suite à sa réflexion sur les avantages et inconvénients de chacune des approches pour le traitement de son cancer. Comme le travail de cette thèse est focusé uniquement sur les patients qui auront choisi le traitement chirurgical, les enjeux reliés au traitement par radiothérapie ne seront pas détaillés ici.

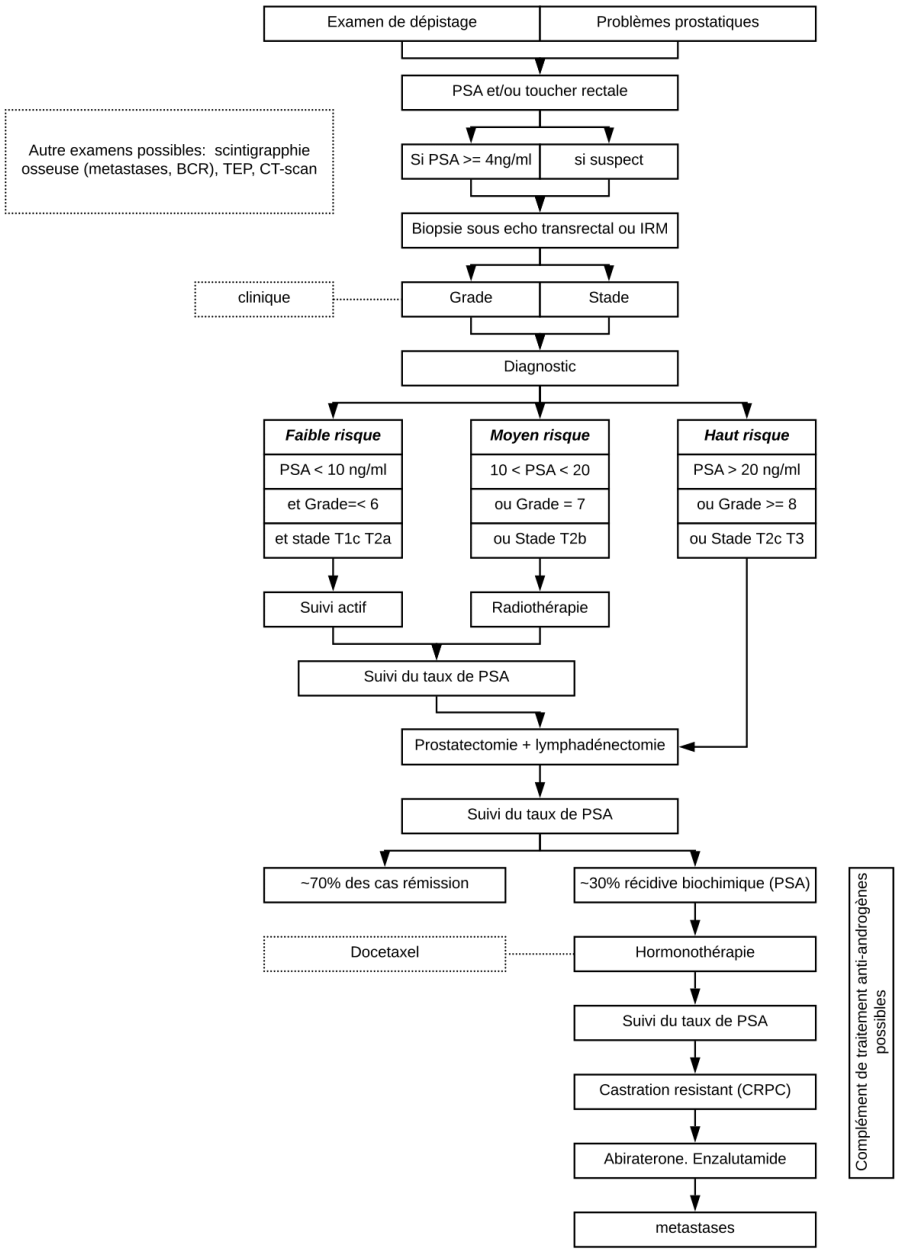


Figure 4. Évolution du cancer de la prostate : Diagramme de décision général et simplifié de l'orientation des patients dans le processus clinique du CaP.

Les patients ayant choisi le traitement chirurgical seront généralement traités par prostatectomie radicale (*i.e.* résection complète de la prostate) avec ou sans lymphadénectomie des chaînes ganglionnaires drainant la région prostatique. Suite à cette chirurgie le niveau sérique de PSA tombe généralement en deçà de 0.02 ng/ml. Environ 70% des patients bénéficieront

d’une rémission complète et durable suite à la prostatectomie. Les autres patients vont malheureusement subir une récidive biochimique (BCR, pour *biochemical recurrence*) dans un laps de temps variable pouvant aller de quelques semaines à plusieurs années. C'est à dire que leur taux de PSA va remonter au-delà de 0.02 ng/ml ce qui signifie que des cellules prostatiques cancéreuses sont encore présentes dans le lit prostatique (récidive locale) ou que des ganglions régionaux sont envahis par des cellules cancéreuses (récidive régionale) ou que des micrométastases sont présentes ailleurs dans le corps. Un faible pourcentage de patients récidivera très rapidement après la chirurgie. On parle alors dans ces cas d'échec à la prostatectomie. Ces patients vont subir une radiothérapie de rattrapage avant de recevoir les traitements classiques des patients en récidive.

Lorsqu’il y a récidive et puisque le CaP est hormono dépendant, le premier traitement à effectuer est de priver les cellules cancéreuses des androgènes (ADT pour *androgen deprivation therapy*) pour ralentir leur croissance en réduisant les niveaux de testostérone circulant. La réduction peut se faire avec une castration directe (orchidectomie), une castration chimique ou avec des traitements anti-androgènes et n’est pas recommandée dans le CaP localisé de bas grade[16–18].

L’effet de ce traitement est temporaire car le cancer va toujours évoluer vers une phase de la progression pour devenir hormono-résistant *i.e.* résistant à la castration (CRPC, pour *Castration Resistant Prostate Cancer*). Ce type de cancer finira toujours par se diffuser dans l'ensemble du corps et deviendra métastatique (mCRPC).

Les cancers hormono-résistants sont classiquement traités par chimiothérapie. Les traitements actuels pour les CRPC et mCRPC incluent le docetaxel et le cabazitaxel qui sont donnés en injections intra-veineuses et préviennent la dépolymérisation de la tubuline empêchant ainsi la division mitotique[19,20]. Le docetaxel a aussi été évalué dans le cas des cancers

métastatiques hormonosensibles mHSPC[21]. Le récepteur aux androgènes joue toujours un rôle sur la croissance des cellules cancéreuses même pour un cancer CRPC. Depuis quelques années une seconde génération d'anti-androgènes a été développée et permet de retarder l'évolution de la maladie. Parmi ceux-ci on compte l'abiratérone, un inhibiteur de la CYP17A1, ainsi que l'enzalutamide et l'apalutamide, des inhibiteurs directs du récepteur aux androgènes.

1.1.d Les cancers métastatiques

Comme tout cancer, le CaP, qu'il soit hormonosensible ou résistant à la castration, peut évoluer vers une forme métastatique. Le CaP se répand le plus souvent dans les os. Certaines formes peuvent aussi se retrouver dans le bas ventre ou vers les reins. Même si les traitements ont beaucoup évolué ces dernières années les molécules de premières lignes contre les CaP métastatiques restent les agents hormonaux comme la prednisone, et les anti-androgènes de seconde génération comme l'abiratérone, l'enzalutamide, ou l'apalutamide. Ces traitements sont actuellement testés seuls ou en synergie pour déterminer les doses et les meilleures combinaisons [22,23].

Si les métastases se retrouvent dans les os, le traitement de choix est le radium-223 (R223). Le R223 est un alpha-radionucléide qui se lie préférentiellement à l'hydroxyapatite qui est une composante majeure des os[24].

Quoi qu'il en soit, cette forme est malheureusement morbide dans presque 100% des cas. Les traitements évoluent donc souvent vers des formes palliatives et la minimisation des effets secondaires dûs aux métastases et particulièrement le traitement de la douleur du patient.

1.1.d Classification des tumeurs: *TNM* et *Gleason*

Dans le but de mieux différencier les cancers et établir un pronostic plus précis, les tumeurs sont classifiées selon leur stade et leur grade. L'Union

Internationale du Contrôle du Cancer (UICC) et différentes organisations travaillant sur le cancer ont développé un système de stadification particulier. Le système TNM (*Tumor, Node and Metastases*) a été initialement inventé par le docteur français Pierre Denoix entre 1943 et 1952. L'UICC a gardé et constamment amélioré ce système avec l'avancée des connaissances. La dernière version du système TNM est la 8ème et elle date de 2016[25]. Le système TNM est adapté à chaque type de cancer et ainsi il existe une définition spécifique du TNM pour le CaP.

Le T dans le système indique la taille de la tumeur primaire et le degré de dispersion dans les tissus locaux (invasion locale):

- TX: La tumeur principale ne peut être mesurée.
- T0: La tumeur principale ne peut être trouvée.
- T1,2,3,4: Indique la taille de la tumeur. Plus le nombre est grand, plus la tumeur est grosse et diffusée aux tissus adjacents.

Le N indique si le cancer s'est dispersé aux ganglions adjacents au site primaire, la taille des ganglions et combien sont envahis:

- NX: La tumeur ne peut être mesurée dans les ganglions.
- N0: Il n'y a pas de cancer dans les ganglions proches.
- N1,2,3,4: Précise si le cancer s'est diffusé dans les ganglions lymphatiques et à quel degré.

Le M indique si le cancer s'est dispersé dans différents organes (métastases):

- MX: Les métastases ne peuvent être mesurées.
- M0: Le cancer n'a pas fait de métastases dans d'autres organes.
- M1: Le cancer a métastasé.

Additionnellement des lettres sont placées après les T, N et M pour fournir des informations plus spécifiques. Chaque type de tumeurs solides (comme le cancer du sein, côlon, poumon ou prostate) à son propre système

de classification secondaire. En diagnostic clinique, après avoir déterminé la classification TNM, le médecin assigne un stade entre I et IV. Cela aide à déterminer le traitement adapté pour le patient:

- Stade 0 – Carcinoma *in situ*. Un stade peu avancé du cancer dans lequel les cellules tumorales n'ont pas envahi les tissus environnant.
- Stades I et II – Le cancer est limité à un organe ou à la location de son commencement. Il peut aussi avoir envahi les tissus adjacents.
- Stade III – Le cancer s'est dispersé plus loin dans les structures environnantes ou aux nœuds régionaux.
- Stade IV – Le cancer s'est dispersé dans d'autres organes ou tissus. Il est métastatique.

Tous les cancers peuvent être aussi classés de façon plus large dans 5 catégories. Elles peuvent être utilisées en pratique par le personnel soignant pour dialoguer avec le patient:

- *In situ*: Il y a des cellules anormales mais elles n'ont pas diffusé aux tissus adjacents.
- Localisé: Le cancer est limité à sa zone de développement primaire et ne montre pas de signe de diffusion.
- Régionale: Le cancer s'est diffusé aux noeuds lymphatiques et tissus proches.
- Distant: Le cancer a métastasé.
- Inconnue: Il n'y a pas assez d'informations.

Le score de Gleason est une mesure spécifique au CaP. Il est défini lors d'une analyse de coupe de tissu prostatique par un pathologiste. Le pathologiste va donner un score pour les deux organisations tissulaires dominantes (*i.e.* les plus représentées). Le score final est obtenu en additionnant les deux chiffres qui peuvent aller de 1 à 5 pour ainsi donner un score final allant en principe de 2 à 10 (Figure 5). En 2009 une modification du système de gradation a été introduite afin d'éliminer les scores finaux inférieurs à 6. La façon d'assigner le

score de Gleason a été modifiée à nouveau en 2016[26,27] en accord avec différents travaux de l'*International Society of Urological Pathology* (ISUP) de 2014[28].

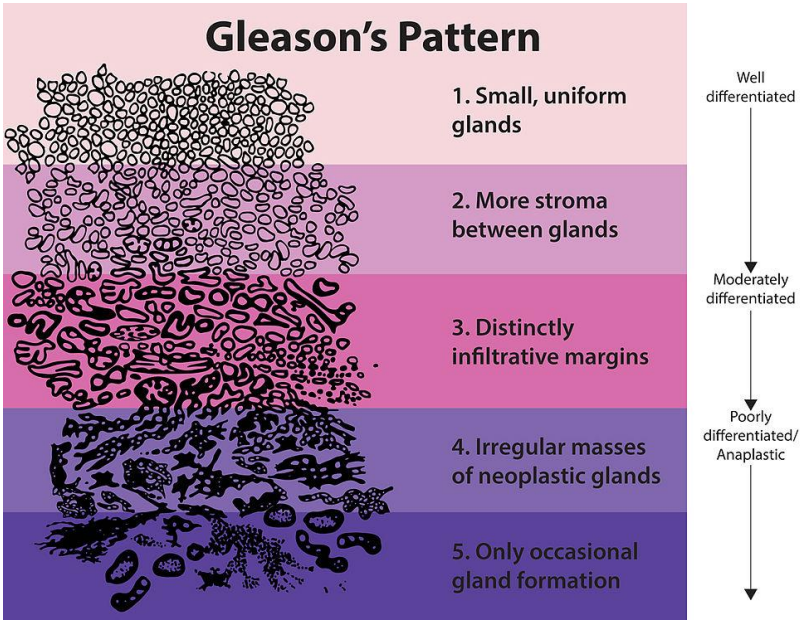


Figure 5. Score de Gleason : Le score de Gleason est un nombre correspondant à la somme de 2 autres. Ces chiffres sont des scores attribués à une biopsie de tumeur examinée par un histopathologiste et varient de 2 à 10. Le premier correspond à l'organisation dominante dans le tissu observé et le second l'organisation secondairement représentée. En score élevé est associé à un mauvais pronostic adapté de Epstein J.I et al[28].

Ce nouveau score de gleason peut être encore mieux relié à la survie à 5 ans et la récidive biochimique du CaP en définissant 5 sous-groupes. Pour les groupes 1-5 les moyennes de temps de survie sont de 96, 88, 63, 48 et 26 mois respectivement[29].

1.1.e Autres outils diagnostiques et pronostiques

Cette section traite des outils diagnostiques et pronostiques disponibles à l'heure actuelle. Cette liste est non exhaustive mais montre la diversité des approches et des finalités de ces outils pour les patients.

ARN PCA3: Cet ARN a été découvert en 1999 à partir d'une analyse d'expression différentielle entre des tissus tumoraux et des tissus sains adjacents [29]. Le gène PCA3 est localisé en position 9q21-22 et est constitué de 4 exons qui peuvent donner 3 transcrits différents. Très rapidement il a été mis en évidence une forte association entre l'expression de PCA3[30] et le CaP. Suite à ces découvertes des techniques de RT-PCR (pour *Reverse-Transcriptase-polymerase chain reaction*) dédiées ont été développées à travers des tests urinaires[31]. Avec une très bonne sensibilité et spécificité (AUC (pour *Area under the curve*) respective de 0.94 et 0.98)[31,32] des recherches ont été menées pour réduire la complexité du pipeline initial de détection[33]. *Gen-Probe Inc* a finalement pu commercialiser un test TMA approuvé par la FDA. Il est donc aujourd'hui utilisé en clinique en complément de certains autres outils[34].

Tables de Partin: Vers la fin des années 1990 le médecin Alan W. Partin, du centre médical Johns Hopkins aux États-Unis, a mis au point un outil pronostique combinant le grade de malignité selon l'échelle de Gleason, le stade et le taux de PSA pour aider les cliniciens et les patients à mieux prédire l'évolution de la maladie. Cet outil est encore très utilisé de nos jours en clinique pour établir le pronostic de la maladie (Figure 6).

Gleason score	Clinical stage T1c	Clinical stage T2a	Clinical stage T2b	Clinical stage T2c
Prediction of probability of organ-confined disease:				
Serum PSA = 0.0–2.5 ng/mL				
2–4	95 (89–99)	91 (79–98)	88 (73–97)	86 (71–97)
5–6	90 (88–93)	81 (77–85)	75 (69–81)	73 (63–81)
3 + 4 = 7	79 (74–85)	64 (56–71)	54 (46–63)	51 (38–63)
4 + 3 =7	71 (62–79)	53 (43–63)	43 (33–54)	39 (26–54)
8–10	66 (54–76)	47 (35–59)	37 (26–49)	34 (21–48)
Serum PSA = 2.6–4.0 ng/mL				
2–4	92 (82–98)	85 (69–96)	80 (61–95)	78 (58–94)
5–6	84 (81–86)	71 (66–75)	63 (57–69)	61 (50–70)
3 + 4 = 7	68 (62–74)	50 (43–57)	41 (33–48)	38 (27–50)
4 + 3 = 7	58 (48–67)	39 (30–48)	30 (22–39)	27 (18–40)
8–10	52 (41–63)	33 (24–44)	25 (17–34)	23 (14–34)
Serum PSA = 4.1–6.0 ng/mL				
2–4	90 (78–98)	81 (63–95)	75 (55–93)	73 (52–93)
5–6	80 (78–83)	66 (62–70)	57 (52–63)	55 (44–64)
3 + 4 = 7	63 (58–68)	44 (39–50)	35 (29–40)	31 (23–41)
4 + 3 = 7	52 (43–60)	33 (25–41)	25 (18–32)	21 (14–31)
8–10	46 (38–56)	28 (20–37)	21 (14–29)	18 (11–28)
Serum PSA 6.1–10 ng/mL				
2–4	87 (73–97)	76 (56–94)	69 (47–91)	67 (45–91)
5–6	75 (72–77)	58 (54–61)	49 (43–54)	46 (36–56)
3 + 4 = 7	54 (49–59)	35 (30–40)	26 (22–31)	24 (17–32)
4 + 3 = 7	43 (35–51)	25 (19–32)	19 (14–25)	16 (10–24)
8–10	37 (28–46)	21 (15–28)	15 (10–21)	13 (8–20)
Serum PSA > 10 ng/mL				
2–4	80 (61–95)	65 (43–89)	57 (35–86)	54 (32–85)
5–6	62 (58–64)	42 (38–46)	33 (28–38)	30 (21–38)
3 + 4 = 7	37 (32–42)	20 (17–24)	14 (11–17)	11 (7–17)
4 + 3 =7	27 (21–34)	14 (10–18)	9 (6–13)	7 (4–12)
8–10	22 (16–30)	11 (7–15)	7 (4–10)	6 (3–10)

Figure 6. Tables de Partin : Tables de Partin adapté des publications publiques de l'International Society of Urological Pathology (ISUP).

Les nomogrames de Kattan: Ces outils ont été créés par le biostatisticien américain Mike Kattan pour évaluer le risque de récidive après 5 ans. Après 5 ans le risque de rechute est très faible, il est donc important de pouvoir prédire cet état car cela peut aider à choisir vers quel protocole de soin orienter les patients. Cette échelle de calcul (Figure 7) se base sur les mêmes données que les tables de Partin.

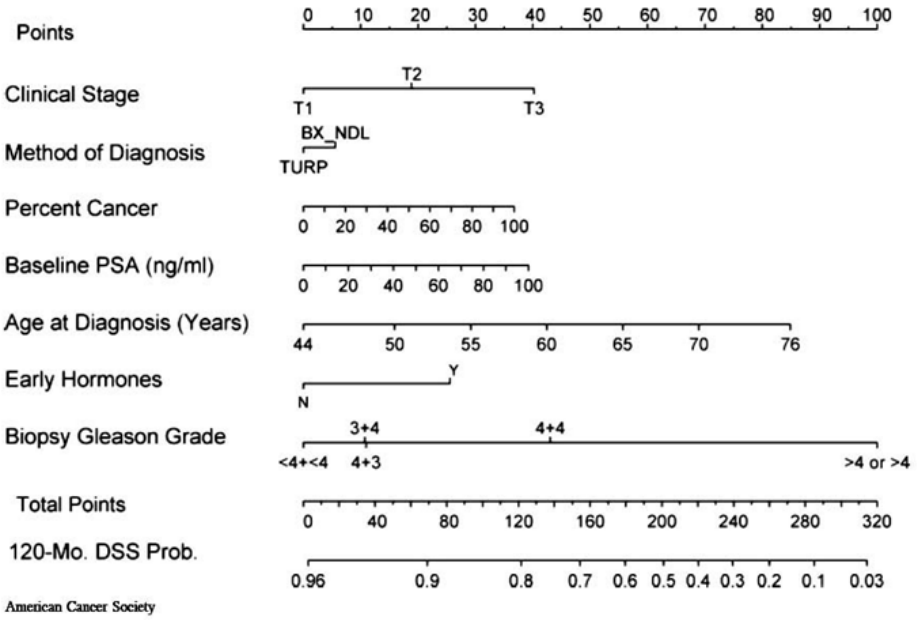


Figure 7. Nomogramme de Kattan : Nomogramme de Kattan adapté des publications publiques de l'International Society of Urological Pathology (ISUP).

Tables de survie d'Albertsen: La table de survie d'Albertsen a été mise au point par le médecin Peter Albersten en 1999. Son but était de prédire le risque de mortalité par cancer de la prostate ou mortalité par causes autres des patients ne recevant aucun traitement à visée curative (chirurgie ou radiothérapie) autre que de l'hormonothérapie lors de métastases s dans les 15 années suivant le diagnostic par rapport à d'autres maladies.

Score de Capra: D'autres nomogrammes ont été développés, comme celui de l'Université de San Francisco en Californie (UCSF), appelé score de « Capra » (pour *Cancer of the Prostate Risk Assessment*» ou évaluation du risque de cancer de la prostate). Il est utilisable pour les suites de nombreux traitements, y compris cette fois, dans les cas de protocoles de surveillance active ou de recours à la cryothérapie (Figure 8). Ce score est aujourd'hui considéré comme étant le meilleur outil pour prédire l'évolution de la maladie.

Variable	Level	CAPRA-S points
Pre-surgical PSA level (ng/ml)	0.00 to 6.00	0
	6.01 to 10.00	1
	10.01 to 20.00	2
	> 20.00	3
Pathologic Gleason score	≤ 3 + 3 = 6	0
	3 + 4 = 7	1
	4 + 3 = 7	2
	≥ 4 + 4 = 8	3
Surgical margin status	Negative	0
	Positive	2
Extracapsular extension	No	0
	Yes	1
Seminal vesicle invasion	No	0
	Yes	2
Lymph node invasion	No	0
	Yes	1

Figure 8. Score de Capra : Score de Capra adapté des publications publiques de l'International Society of Urological Pathology (ISUP).

1.1.f Caractéristiques moléculaires du CaP

Le CaP est un cancer hormono-dépendant fortement lié à la voie du récepteur aux androgènes (AR pour *Androgen receptor*). L'AR est un facteur de transcription ligand-dépendant appartenant à la grande famille des récepteurs nucléaires[35]. En absence de ligand (dihydrotestosterone, testostérone ou des

stéroïdes androgéniques) l'AR est localisé dans le cytoplasme sous forme complexe avec des protéines chaperones. Après interaction avec son ligand, l'AR et son ligand vont former un homodimère qui va se déplacer dans le noyau. Dans le noyau l'AR dimérisé va reconnaître des séquences régulatrices dans certaines régions régulatrices de l'ADN des gènes cibles des androgènes[36,37]. Suite à cela, il va recruter différents cofacteurs régulant certaines protéines ou favoriser des modifications épigénétiques pour former un complexe de régulation de la transcription. Son rôle connu est majoritairement activateur mais des recherches plus récentes viennent compléter nos connaissances sur son rôle de répresseur[38]. Bien entendu l'AR peut subir des modifications post-traductionnelles[39,40] et se décline sous différents variants avec des activités biologiques spécifiques[41].

Les cellules cancéreuses prostatiques sont sensibles à l'ADT. En effet, lors d'une récidive suivant un traitement par radiothérapie ou prostatectomie radicale, ou encore d'emblée si la maladie est déjà très avancée lors du diagnostic on utilise un traitement anti-androgénique par castration chimique ou plus rarement chirurgicale pour diminuer le taux de testostérone circulant. L'efficacité de ces traitements est temporaire et les patients vont presque toujours évoluer vers la forme CRPC i.e. la forme de CaP résistante à la castration. Les voies de l'AR restent toujours un axe thérapeutique important dans le CRPC, ne serait-ce que pour ralentir la progression de la maladie. Des traitements pour accentuer l'ADT via la compétition directe avec la DHT ou en réduisant la synthèse des androgènes avec des inhibiteurs type CYP17A1 ont montré des effets bénéfiques pour les patients[42]. L'abiraterone acetate est un inhibiteur irréversible de CYP17A1 qui convertit des précurseurs protéiques en stéroïdes dont des précurseurs d'androgènes. Ce composé bloque donc la production d'androgènes dans les testicules, les glandes surrénales et aussi dans la tumeur prostatique diminuant sa croissance. L'abiratérone a été autorisé en 2011 pour les CRPC peu avancés et en concomitance ou non avec du docetaxel[43]. L'enzalutamide est un autre anti-androgène de seconde génération approuvé par la FDA dans le traitement du mCRPC toujours en

concomitance avec le Docetaxel. Un autre composé relatif à l'enzalutamide comme l'apalutamide[44,45] a aussi été approuvé récemment par Santé Canada. Le Darolutamide fait aussi partie de cette nouvelle génération d'AR mais sa structure diffère de celle de l'enzalutamide[46,47]. Cette structure le rendrait plus efficace contre les mutants de l'AR et moins de pénétration de la barrière hémato-encéphalique ce qui réduirait ses effets secondaires[48,49]. Cet anti-androgène a été approuvé par la FDA (Federal Drug Agency) et Santé Canada.

Le développement du CaP implique un certain nombre de voies métaboliques qui nécessiteraient plusieurs pages de description pour chacune d'entre elles. Dans cette partie je présenterai brièvement celles qui pourraient être ciblées par des interventions thérapeutiques

PI3K/AKT/mTOR: Le rôle de la voie PI3K/AKT/mTOR dans l'initiation et la progression du cancer est maintenant bien connue[50,51]. La perte de *PTEN* est l'événement clé rendant hyperactif la signalisation PI3K et est associée avec un mauvais pronostic du CaP[52]. Un modèle murin a récemment montré toute l'évolution de ce processus[53].

Fibroblast growth factor (FGF): Cette voie joue un rôle important notamment dans les derniers stades de la maladie[54,55]. Des traitements de xénogreffes avec un inhibiteur du récepteur FGF, soit l'AZD4547, réduisent la croissance des tumeurs et l'effet est amplifié si l'on ajoute l'AZD5363, un inhibiteur de la voie AKT. Le Dovitinib est un inhibiteur de la voie FGF, une étude en phase 2 est en cours avec des résultats limités[56].

Wingless-related integration (WNT): Cette voie métabolique est impliquée dans le cancer à différents niveaux[57,58]. L'évaluation d'un peptide mimant l'effet de WNT-5A a montré des résultats intéressants dans des modèles murins[59]. La dérégulation de cette voie favorise aussi la croissance tumorale en jouant sur la voie mTOR[60].

Réparation de l'ADN (DDR): Comme dans beaucoup de cancers les mécanismes de réparation de l'ADN dans le CaP peuvent être altérés[61]. Des médicaments ont été développés pour inhiber la poly-ADP ribose polymerase (PARP) qui joue un rôle primordial dans la détection des erreurs dans l'ADN et de leur réparation[62]. L'Olaparib est un inhibiteur oral de la PARP qui a montré une activité antitumorale significative lors d'un récent essai clinique de phase 2 chez des patients atteints de mCRPC. Les patients portant des mutations dans les gènes de susceptibilité au cancer du sein (*BRCA1/2* pour *Breast cancer 1/2*) ou de l'ataxia telangiectasia serine/threonine kinase (*ATM*) ont des meilleures réponses à ces traitements.

Épigénétique (Régulation de la transcription): Le complexe *polycomb repressive complex 2* (PRC2) réprime l'expression de certains gènes suppresseurs de tumeur qui sont essentiels à l'initiation et la progression du CaP[63]. Une des protéines constitutives de PRC2 est *enhancer of zeste homolog 2* (EZH2) une histone methyltransférase qui catalyse la triméthylation de la lysine 27 des histone H3 causant la répression de la transcription. EZH2 a été rapportée comme étant différentiellement exprimée dans la plupart des CaP métastatiques[64]. D'autres complexes ou familles jouant un rôle dans les structures chromatiniennes comme le *mixed lineage leukemia* (MLL)[65], la démméthylase LSD1[66] ou encore la famille BET[67,68].

L'antigène membranaire spécifique de la prostate (PSMA): Le PSMA est exprimé sur les membranes des cellules prostatiques et est ciblé pour livrer des agents cytotoxiques de façon spécifique comme l'auristatin ou la maytansinoid. Les études montrent une grande efficacité de ces composés mais aussi une très grande variabilité selon les patients[69,70]. Des traitements combinant des thérapies alphas labellisées PSMA sont en cours d'évaluation[71–73] et sont prometteurs dans le traitement des mCRPC.

Fusions ETS: Une des principales caractéristiques du CaP au niveau moléculaire est son faible taux de mutations et de copies de gènes focalisées alors que son taux de nombre de copies étendues et de fusions de gènes est élevé[74]. Les altérations les plus communes dans le CaP sont les fusions du gène ETS *(E26 transformation-specific)* liées à la régulation des androgènes. Ces fusions viennent au début du développement du cancer et amènent une surexpression des gènes fusionnés ETS[75]. Un des exemples les plus connus est la fusion entre la protéase sérine 2 transmembranaire (TMPSS2) et le facteur de transcription ERG (de la famille des ETS) TEMPRESS-2ERG, que l'on trouve dans 40-80% des tumeurs[76]. Cette fusion semble jouer un rôle dans le développement primaire du cancer mais a aussi été rapportée comme ayant un pronostic favorable dans certains types de tumeurs (*wild type PTEN*)[77] .

1.1.g Analyse omiques du CaP: *TCGA*

Comme décrit dans les parties précédentes, il y a une grande hétérogénéité des réponses aux traitements selon les patients. Cela suggère que les tumeurs pourraient être classées en différents sous-groupes et que les cliniciens pourraient adapter les propositions de traitement en fonction de cette classification. Les récentes avancées dans les technologies de séquençage ont permis de proposer de larges études ayant pour but de caractériser les tumeurs au niveau génomique afin de les classifier en sous-groupes de tumeurs. Le projet TCGA PRAD est le plus conséquent en la matière et ne concerne pas seulement le CaP[78]. Dans le CaP 7 sous-types de tumeurs ont été définis. Quatre sont reliés au problème des fusion ETS qui ont pour conséquence la surexpression de 4 familles de gènes: *ERG, ETV1, ETV4* et *FLI1*. Les 3 autres sous-types sont caractérisés par des mutations dans les gènes *SPOP, FOXA1* ou *IDH1* (Figure 9). Il est à noter que parmi les cas composant la cohorte TCGA-PRAD, une vingtaine de cas proviennent du Québec via la participation de la Biobanque PROCURE à ce projet.

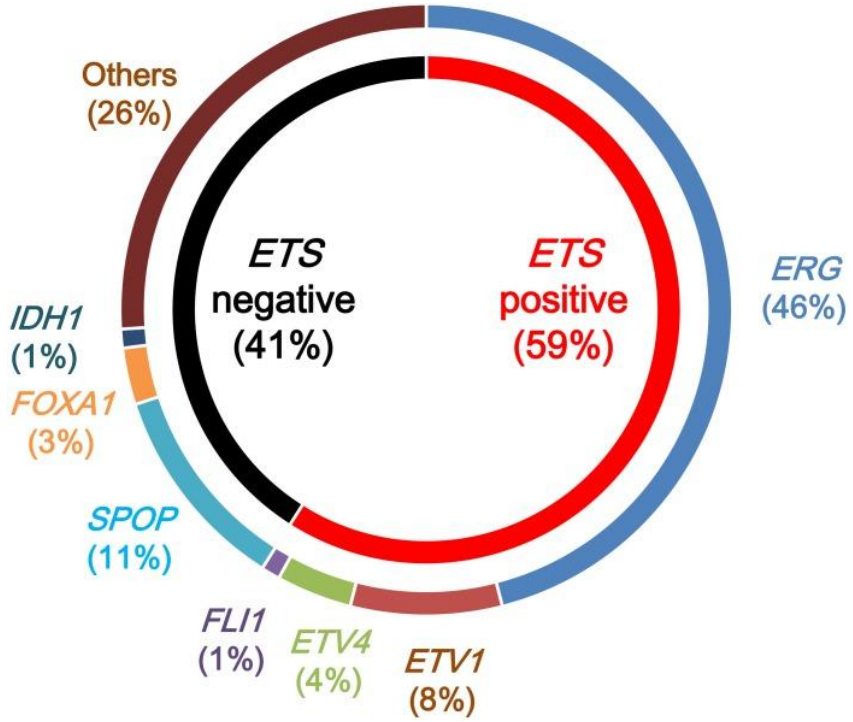


Figure 9. Sous-types moléculaires du CaP primaire TCGA adapté de Abeshouse A, et al[78] : Le CaP primaire peut-être classé dans deux catégories de réarrangement des facteurs de transcriptions ETS. Une catégorie avec réarrangement dite ETS positive et une sans réarrangement dite ETS négative. Les CaP ETS+ sont classifiés par des fusion ETS spécifiques: ERG, ETV1, ETV4, ou FLI1. Les ETS- sont définis par des mutations dans les gènes SPOP, FOXA1 et IDH1.

1.1.h Analyse omique du CaP: *CPC-GENE*

Une autre équipe a récemment publié une analyse multi-omique[79] pour prolonger les approches déjà lancées par TCGA mais en se concentrant sur le CaP de risque intermédiaire. Le groupe de recherche en uro-oncologie du CHU de Québec-Université Laval a participé à ce projet en fournissant 143 cas de la cohorte CPC-GENE. La cohorte CPC-GENE se distingue notamment par un suivi médian très long (~8 ans) ce qui permet d'étudier les événements de BCR

avec confiance chez ces patients. En résumé, l'étude fait ressortir certains liens entre la BCR est des altérations génomiques:

- CNA: une translocation inter-chromosomique au niveau du centromère du chromosome 7 (chr7:61–62 Mbp) et une amplification de MYC.
- SNV: une seule mutation dans le gène ATM
- Méthylation: une hyper-methylation du gène ACTL6B et une hypo-methylation du gène TCERGL1,

Prises ensemble, ces variables permettent de prédire la BCR avec une AUC de 0.83 dans la cohorte de validation.

1.2 Immunologie

1.2.a Rappels généraux

Le système immunitaire est un système biologique complexe de défense impliquant plusieurs types cellulaires qui a la capacité de discriminer entre le soi et le non-soi. Son rôle est de protéger l'organisme contre les agressions endogènes et exogènes. Il régule aussi les populations bactériennes utiles et les cellules corporelles (*e.g.* destruction des tissus fins inter palmaires avant la naissance) des individus. En réalité, le système immunitaire est constamment en train de réguler la prolifération cellulaire dans notre corps via l'action des leukocytes[80,81]. Par ces mécanismes la tumeur n'est pas seulement le fait des cellules cancéreuses, mais c'est plutôt le résultat d'un jeu complexe d'interactions entre les cellules cancéreuses et le système immunitaire[82]. Quand le cancer est diagnostiqué cela fait, en réalité, plusieurs semaines voire des mois que la maladie a commencé. Les cellules cancéreuses commencent à devenir pathologiques quand elles échappent à cette régulation. La tumeur qui en résulte est en fait le produit de cette interaction avec le système immunitaire, un phénomène qui est appelé *immunoediting (IE)*[83–85]. L'IE est l'ensemble des

mécanismes par lesquels le système immunitaire promeut et guide le développement tumoral. Ce développement est à l'heure actuelle défini en 3 phases : élimination, équilibre et échappement[83] (Figure 10).

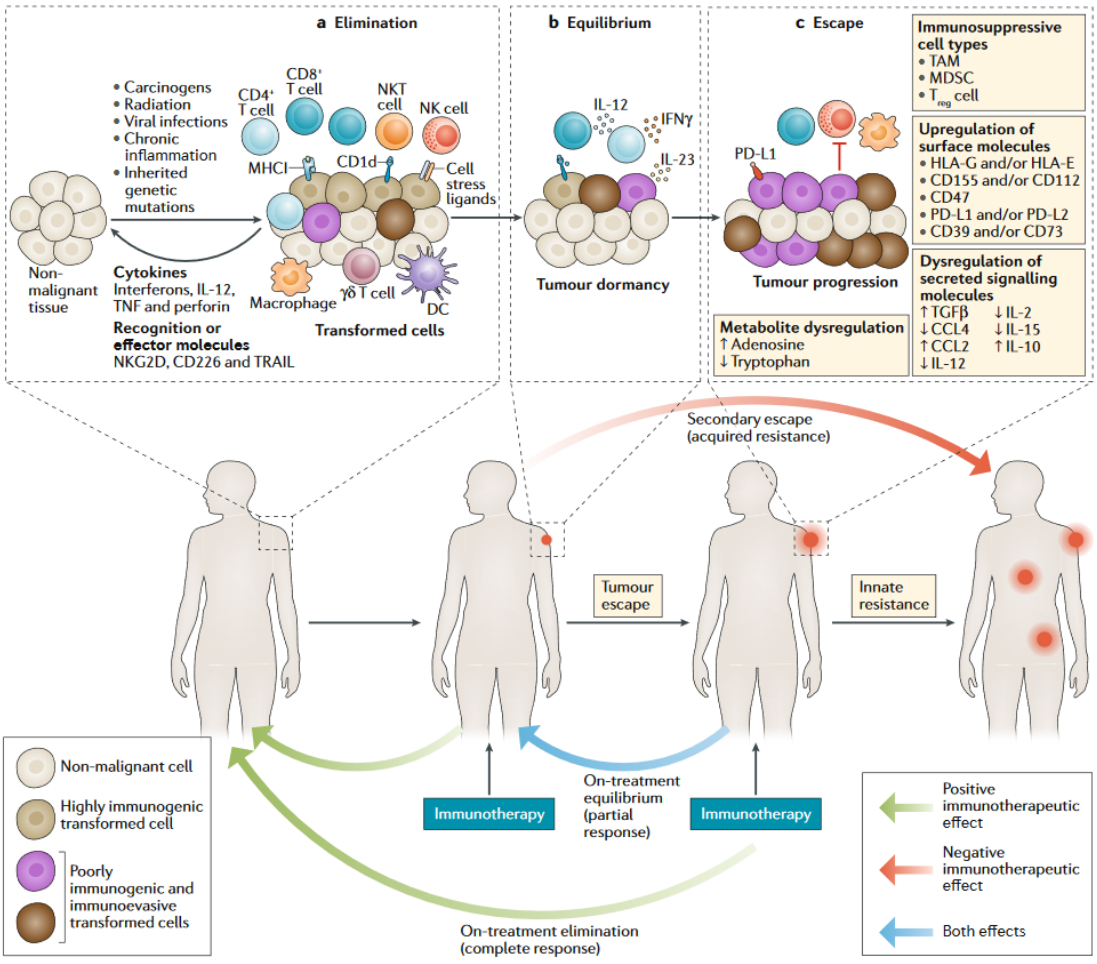


Figure 10. Différentes phases de la tumorigenèse adapté de O'Donnell et al [83] : La tumorigenèse est un processus patient spécifique dépendant du système immunitaire du patient et de son environnement. L'exposition à des mutagènes (produits chimiques, UV etc) peut provoquer la sélection de cellules cancéreuses plus agressives.

C'est pourquoi les symptômes et les réactions aux traitements médicaux diffèrent selon les personnes. L'équilibre entre le contrôle de la prolifération, la pathogénicité, l'apoptose et l'invasion métastatique est une des clés pour comprendre le cancer.

1.2.b Anticorps

Les anticorps sont des protéines appartement à la super famille des immunoglobulines (Ig). Ces protéines sécrétées par les lymphocytes B sont reconnues pour être une composante majeure du système de défense immunitaire. Le rôle des anticorps est de reconnaître et de se lier à leur antigène correspondant pour que ceux-ci soient reconnus par certaines cellules du système immunitaire. Les anticorps sont divisés en 5 isotypes - IgG, IgM, IgD, IgA et IgD . La différence entre leur structure leur confère des fonctions spécifiques au sein du système immunitaire. Les Ig sécrétées par les cellules plasmatiques circulent dans le sang et peuvent rentrer dans les tissus infectés. Cette composante du système immunitaire est appelée le système humoral. Ce terme vient du mot “humeur” qui est une ancienne appellation pour les fluides corporels comme la lymphe ou le sang où ils ont été initialement découverts. Les anticorps peuvent agir par neutralisation ou opsonisation (Figure 11). Bien que très efficace pour lutter contre les infections bactériennes et virales, leur rôle dans la réponse anti-tumorale est plutôt limité comparativement au rôle de la réponse dite “cellulaire” impliquant notamment les lymphocytes T.

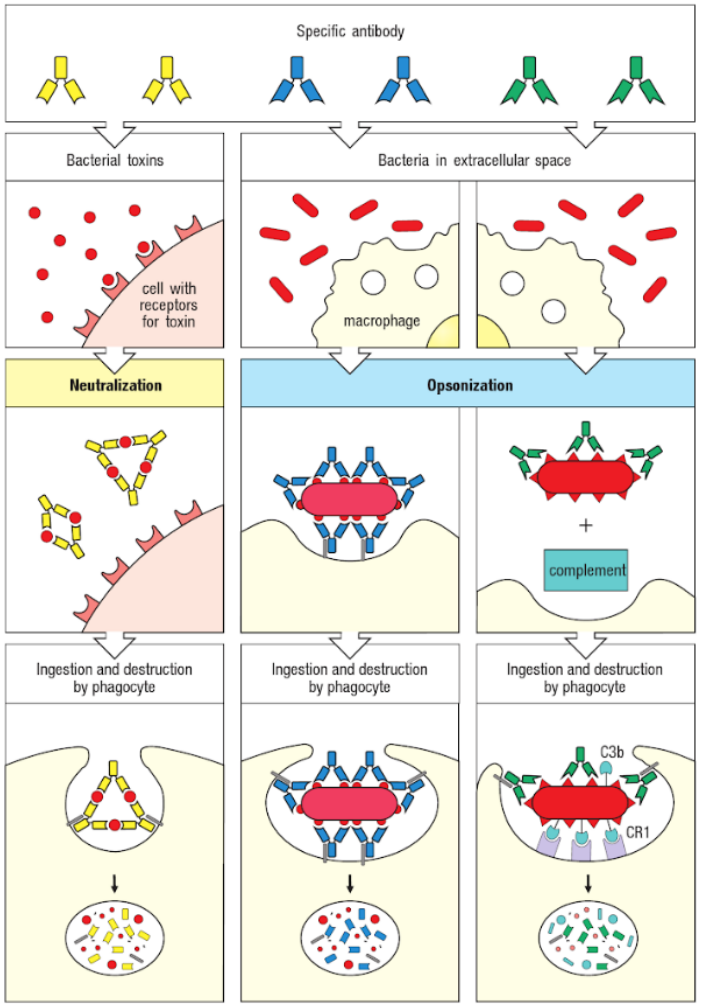


Figure 11. Action des Ig adapté du livre de Parham P. [86] **:** Mécanismes d'action des Ig selon les éléments moléculaires cibles. Quelque soit l'élément cible la réaction implique une phagocytose finale afin de décomposer le produit à éliminer et pouvoir récupérer des Ag pour la mémoire immunitaire.

1.2.c Complexe majeur d'histo-compatibilité

Les lymphocytes possèdent à leur surface des récepteurs capables de reconnaître des peptides antigéniques présentés par un ensemble de molécules appelé complexe majeur d'histocompatibilité (MHC pour *Major histocompatibility complex*). Chez l'humain les molécules composant le MHC sont appelées HLAs pour *Human Leukocyte Antigens*. Ces protéines membranaires ont la capacité de lier, via leur partie flexible, des peptides antigéniques (Ag) provenant de leur environnement. De cette façon, les cellules présentent en continu des antigènes qu'elles produisent ou récupèrent dans leur environnement (Figure 12). Ce système est très efficace car il est difficile de le tromper au niveau moléculaire. Une bactérie produit des protéines spécifiques différentes d'une cellule animale. Une cellule infectée par un virus va se comporter différemment au niveau moléculaire par rapport à une cellule normale.

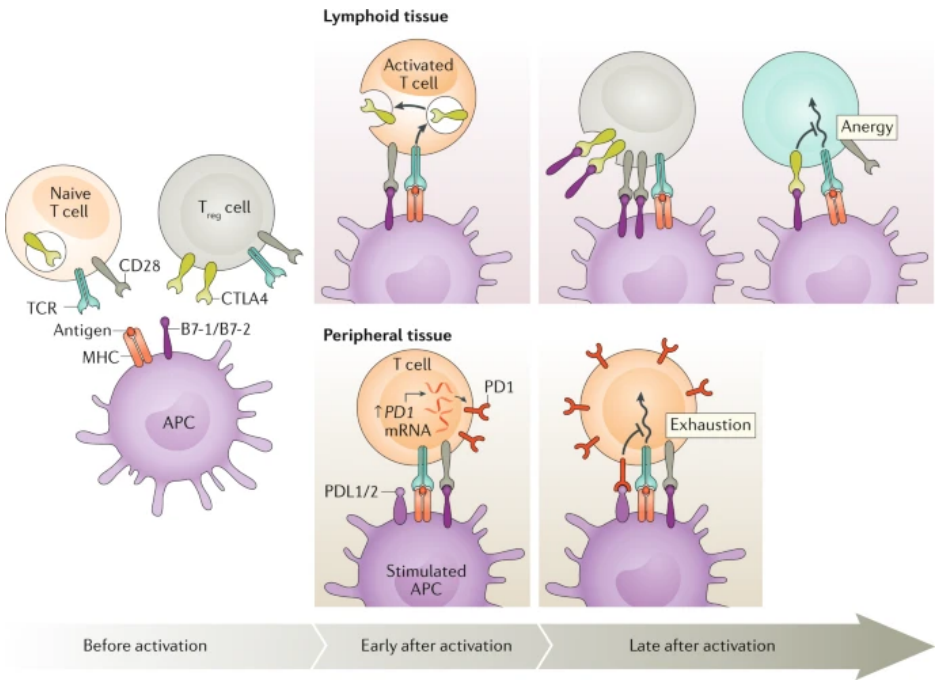


Figure 12. Activation des Tc adapté de Kirsty Minton [86] : Les Ag sont récupérés dans l'environnement et présentés aux Tc par les APC.

Les peptides antigéniques présentés par le MHC peuvent provenir de l'intérieur ou de l'environnement extérieur de la cellule. Seuls les globules rouges n'ont pas la capacité de présenter des peptides Ag internes. Ce contrôle de l'environnement cellulaire est fait en continue ce qui permet d'avoir à la surface des APC (*Antigen Presenting Cell*) une description très précise et dynamique de son état physiologique.

C'est ce système qui participe au rejet des greffes lors des transplantations par exemple. On comprend que plus les personnes sont proches au niveau généalogique, plus elles ont de chances d'être histo-compatibles. Ce système est donc une construction dynamique du soi. Le système immunitaire est en permanence en train de contrôler cette identité moléculaire afin de contrer tout éventuel dysfonctionnement. Une suractivité de ce système est aussi à l'origine de maladies auto-immunes.

Au niveau des lymphocytes T la molécule qui reconnaît le peptide Ag lié à la molécule HLA est appelé le récepteur de cellule T (TCR pour *T-cell receptor*) et sa structure assez complexe ne sera pas détaillée ici. Il y a donc autant besoin de diversité au niveau des TCR (pour être capable de reconnaître les Ag présentés) qu'il y a de diversité dans l'environnement. Les gènes codant pour les TCR subissent un processus complexe de recombinaisons génétiques leur permettant de créer cette diversité. Ce processus de recombinaison est aléatoire et permet de créer une très grande variété de TCR. Par contre, toutes ces cellules ne peuvent être maintenues dans le répertoire des lymphocytes. La sélection et la maturation des lymphocytes sont des étapes essentielles et délicates. Pour être retenus, les lymphocytes doivent être capables d'ignorer les Ag du soi mais de pouvoir répondre à un Ag du non soi. Le processus de sélection est coûteux car plus de 95% des lymphocytes seront détruits.

Finalement , seuls les lymphocytes présentant des TCR liant faiblement le soi et fortement le non soi seront conservés (Figure 13). Cette sélection permet donc d'éviter la création des réponses auto-immunes mais permet la protection en assurant des populations de lymphocytes qui pourront potentiellement se lier à n'importe quel Ag reconnu comme néfaste. Lors d'une infection, les cellules de l'immunité innée soient les macrophages, cellules dendritiques et neutrophiles participeront à la reconnaissance et à l'élimination des pathogènes. Via leur fonction présentatrices d'antigènes, ces APC présenteront les Ag de ces pathogènes aux lymphocytes pour induire la réponse adaptative. Sur les millions de lymphocytes disponibles, quelques-uns auront le bon TCR et seront activés suite à la reconnaissance de l'antigène. Il s'en suivra alors l'enclenchement de la réponse immunitaire adaptative classique avec la multiplication et spécialisation des lymphocytes en cellules CD8+ (cytotoxic *T cells*) et CD4+ (helper *T cell*).

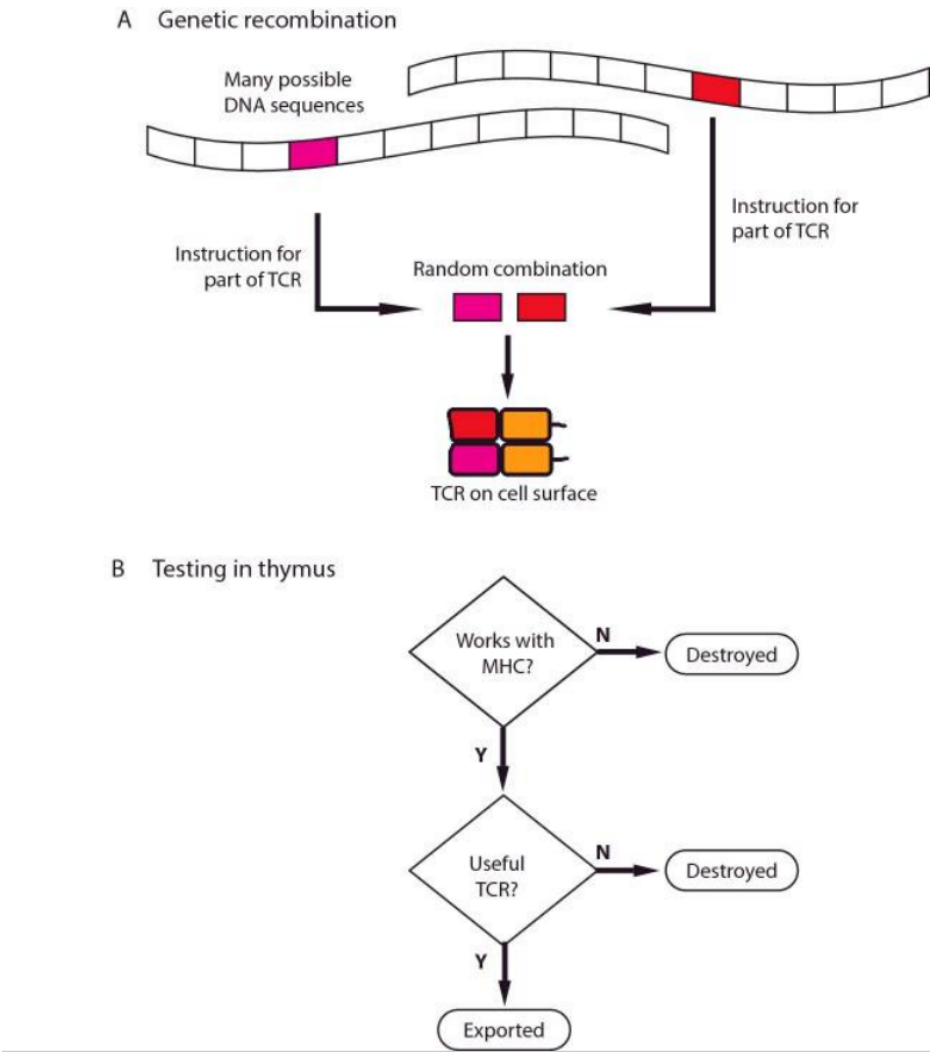


Figure 13. Production des TCR adapté du livre de Parham P. [86]: La diversité génétique est le moteur de la production des TCR. La résistance à de nouveaux pathogènes est donc le fruit du hasard puis de la fixation par sélection d'individus.

1.2.d Inflammasome et micro-environnement tumoral.

Les molécules de l'inflammation sont des complexes macromoléculaires qui déclenchent une réponse centrale et locale lors d'un changement dans la composition cytosolique. Les molécules de l'inflammation, que nous

appellerons inflammasome, sont classées selon leur structure en *nucleotide-binding oligomerization domain and leucine-rich repeat receptors* (NLRs), *absent in melanoma 2 (AIM2)-like receptors* (ALRs) et la pyrine récemment identifiée. La famille NLR est divisée en deux sous-familles, les NLRPs et NLRCs en fonction de si l'amino acide terminal contient un domaine pyrin ou un domaine de recrutement de caspase (CARD). NLRP3, NLRP3, NLR et les protéines inhibitrices de l'apoptose (NAIP-NLRC4) sont connues pour s'assembler et former les complexes de l'inflammasome[87,88]. D'autres NLR et non-NLR effecteurs comme les interférons-gamma de type 16 (IFI16), l'acide rétinoique influent aussi sur l'inflammasome de différentes manières[89–91].

Bien que le rôle primaire de l'inflammation soit de lutter contre des infections pathogéniques, une activation persistante de ces mécanismes conduit souvent à des pathologies chroniques. Il est maintenant bien connu que cette inflammation chronique favorise la tumorigenèse à différents niveaux[92] (Figure 14).

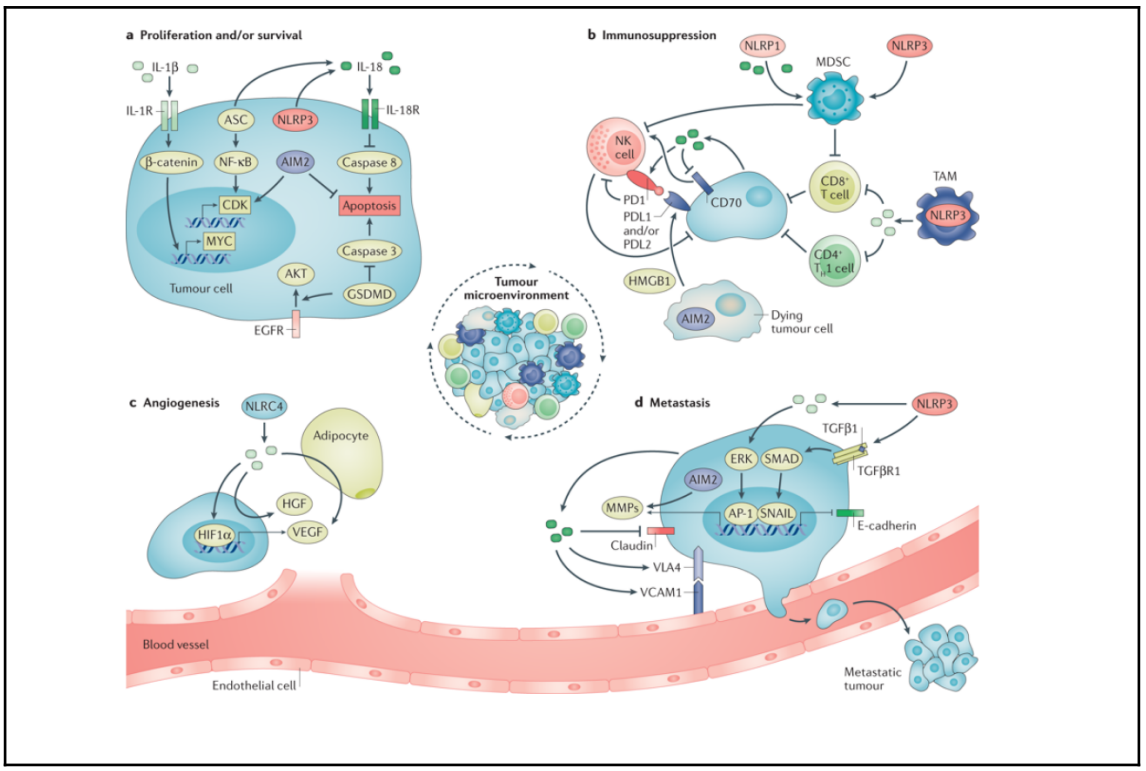


Figure 14. Inflammation et tumorigenèse adapté de Grivennikov et al [92] : *"L'inflammation peut favoriser la tumorigenèse via la production de certaines molécules comme les TGF, les cytokines ou la création d'un environnement hypoxique."*

Initialement, en réponse à la prolifération des cellules tumorales, le système immunitaire déclenche les mécanismes anti-tumoraux avec une vague de cellules inflammatoires dans le micro-environnement tumoral. Les cellules tumorales vont réagir en conséquence pour échapper à cette réaction inflammatoire qui est la première partie de la réaction immunitaire globale. La sécrétion d'IL-18, IL1-ß et d'autres molécules effectrices comme les cytokines (provenant des cellules immunitaires ou tumorales) est bien connue pour créer un environnement immunosuppresseur dans la tumeur. L'inflammation joue donc un rôle essentiel dans le développement du cancer et la construction du micro-environnement tumoral et par extension du micro-environnement immunitaire tumoral[93–96].

1.2.e Microenvironnement immunitaire tumoral

L'immunothérapie est intrinsèquement liée à la compréhension du micro-environnement immunitaire tumoral (TIME pour *Tumor immune microenvironment*) [97]. La compréhension du TIME a permis d'identifier des populations de patients pouvant répondre positivement ou non aux différents traitements en fonction de la présence de biomarqueurs spécifiques ou de caractéristiques histologiques comme l'infiltration de cellules immunitaires.

Cette infiltration peut-être quantifiée par des outils bioinformatiques comme CIBERSORT[98] et XCELL[99] (calcul de l'immunoscore)[100] en estimant l'abondance de certains types de cellules immunitaires. Ces estimations sont faites à partir de signatures de gènes. Elles permettent d'avoir une idée de la composition générale de l'infiltrat mais il est encore très difficile d'aller dans le

détail des sous types de cellules immunitaires. Ces outils aident au pronostic des patients en les stratifiant.

Il existe 3 principales classes de TIME (Figure 15).

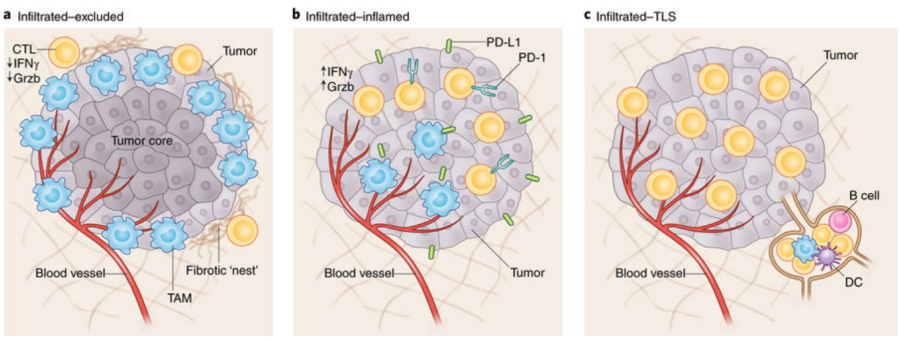


Figure 15. Principaux TIME adaptés de Binnewies M. et al [97] : (a) Les CTLs sont exclues (b) Tumeur infiltrée et inflammée (c) Tumeur inflammées avec des structures types TLS.

Un TIME non infiltré où les lymphocytes cytotoxiques CD8+ (CTLs pour *Cytotoxic T lymphocytes*) sont exclus du cœur de la tumeur. Dans ces tumeurs les CTLs sont retrouvés en périphérie de la tumeur en contact avec des macrophages Ly6Clo F4/80+ ou alors coincés dans des zones fibrosées.

Un TIME infiltré et inflammé où les cellules tumorales et myéloides expriment granzyme B (Grzb), IFN-γ et PD-1. Dans certaines de ces tumeurs on retrouvera beaucoup de mauvais appariements de l'ADN et de microsatellites instables (MSI-H pour *Microsatellite instability-high*) à cause de défectuausités des mécanismes de réparation de l'ADN . Ces problèmes d'appariement vont conduire à la création de nouveaux déterminants antigéniques appelés néoépitopes.

Un TIME type inflammé-TLS. Ces tumeurs contiennent des structures lymphoides tertiaires (TLSs) qui contiennent différents types de cellules immunitaires. Ces structures ressemblent à des ganglions lymphatiques.

Les tumeurs mettent en place un environnement qui leur est bénéfique en favorisant l'immunosuppression. Cela leur permet de promouvoir leur propre croissance et d'échapper au système immunitaire (Figure 16).

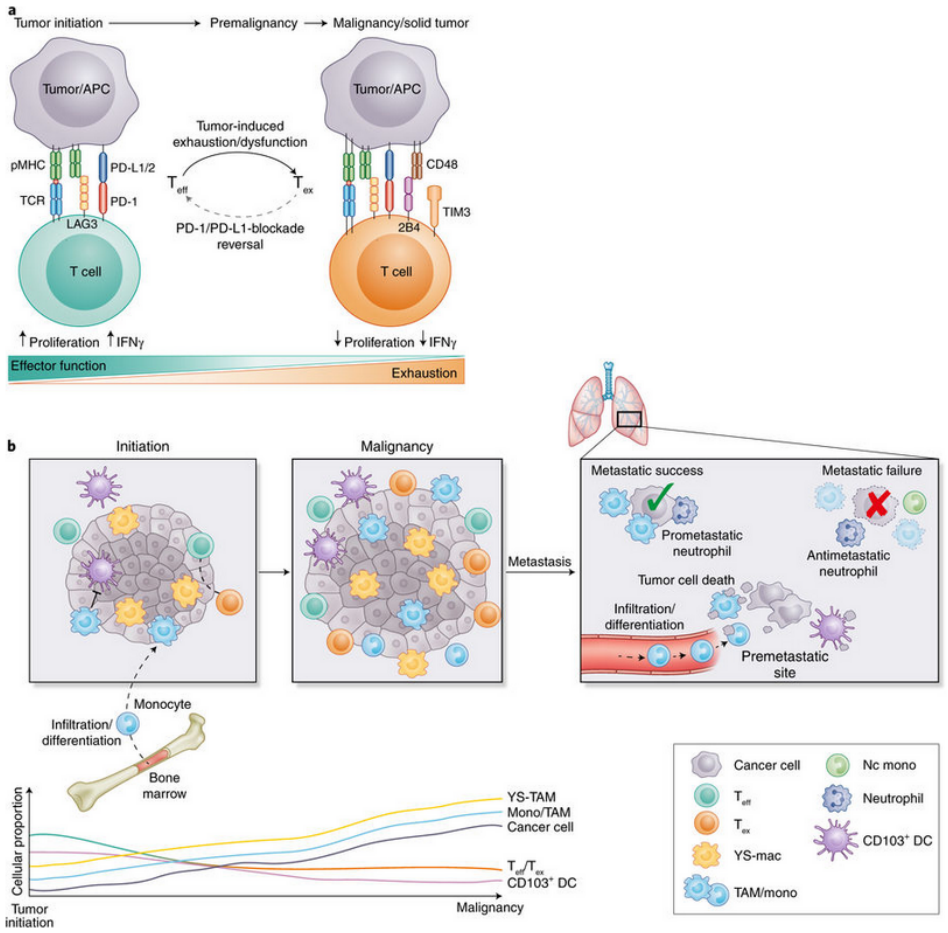


Figure 16. TIME et développement tumoral adaptés de Binnewies M. et al [97]: Mécanisme de l'épuisement au niveau moléculaire.

Les mécanismes de changement d'environnement regroupent des voies métaboliques allant par exemple de l'inflammation, au contrôle de l'hypoxie, la multiplication et la migration des cellules (Figure 17).

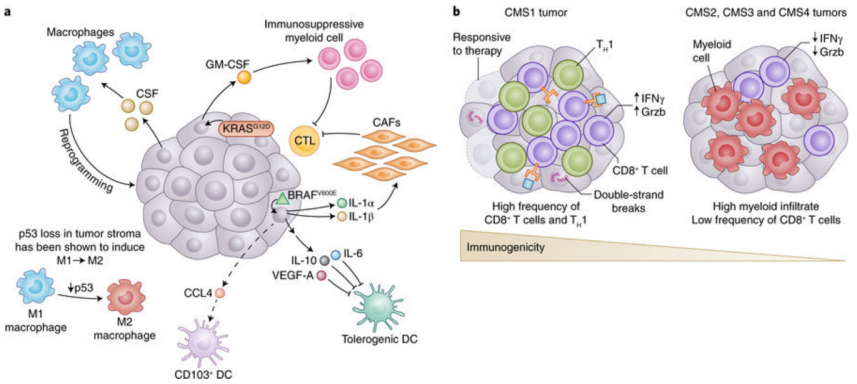


Figure 17. Mise en place du TIME par les tumeurs adaptés de Binnewies M. et al [97]: Quelques exemples de micro-environnement tumoraux et leurs voies métaboliques.

Une des caractéristiques du développement progressif du TIME est l'épuisement des lymphocytes T[101,102]. Cet état d'épuisement est caractérisé par une réponse plus faible des cellules T aux antigènes et une transformation des cellules T effectrices (Teff) en cellules épuisées (Tex). Ce changement se caractérise par une augmentation de l'expression de molécules dites points de contrôles immunologiques tels que PD-1, CTLA-4, LAG3, 2B4 et TIM3 et la diminution de la sécretion de cytokines ou d'interféron-γ. D'un point de vue thérapeutique, il y a des évidences qui montrent que certains types d'épuisement seraient irréversibles. Cela expliquerait en partie pourquoi certains patients ne répondent pas aux thérapies visant l'inhibition des points de contrôles immunologiques.

1.2.f Immunothérapie

Les origines de l'immunothérapie remonteraient à 1893 avec les observations de William B. Coley qui montra que l'injection de bactéries dans une tumeur pouvait induire sa régression[101]. Depuis cette époque, et pendant tout le XXième siècle, le développement de l'immunothérapie du cancer comme nouvelle modalité de traitement au même titre que la chimiothérapie et la radiothérapie a été un objectif qui a été poursuivi par de

nombreux chercheurs mais qui a été caractérisé par très peu de véritables succès cliniques. Par contre, les nombreuses approches qui ont été testées pendant cette période ont quand même permis des avancées scientifiques qui ont mené à une meilleure compréhension de l'immunologie tumorale et des facteurs qui contrôlent la réponse anti-tumorale. En 2016 un article de BMC medecine titrait *Cancer immunotherapy: the beginning of the end of cancer?*[103]. Bien que l'on soit toujours loin de cet horizon, les avancées en immunothérapie de cette dernière décennie ont été impressionnantes. Le cycle de l'immunité tumorale[99,100] (TIC) se définit en 3 phases.

Dans la première phase les DCs (cellules dendritiques) immatures vont détecter les signaux inflammatoires et des molécules pathogéniques. Cette reconnaissance se fait entre autres par des récepteurs de type *Toll-like (*TLRs pour *Toll-like receptors)* qui vont ensuite déclencher la maturation et l'augmentation de la phagocytose pour internaliser les antigènes. Les antigènes sont ensuite dégradés et réduits à des peptides pour pouvoir être présentés via les molécules du CMH.

Dans un deuxième temps, les DCs vont changer de forme et migrer vers les organes lymphoïdes pour permettre l'activation et amplification de cellules natives T CD8+ cytotoxiques spécifiques à l'Ag (CTLs). Dans une moindre mesure les DCs peuvent aussi favoriser la réponse via des Ac et l'activation de cellules NK.

Dans la dernière phase, les cellules T CD8+ vont migrer vers la tumeur à travers le réseau lymphatique pour rejoindre le TIME. Les tumeurs peuvent échapper aux CTLs en réduisant la présentation des Ag tumoraux ou en diminuant l'expression des molécules du CMH de classe I. Les tumeurs peuvent aussi présenter des molécules de surface comme PD-L1 qui cause l'engagement puis l'épuisement des CTLs via la surexpression du point de contrôle immunologique PD-1 à leur surface. Les interactions entre la tumeur et le tissu stromal hypoxique favorise la sécrétion de molécules

immunosuppressives comme par exemple le TGF-ß et l'indoleamine -pyrrole 2,3-dioxygenase (IDO).

Les agents immnothérapeutiques prennent donc place dans ce contexte à différents niveaux du TIC. Les virus oncolytiques comme le virus taliminogene laherparepvec (T-VEC), approuvés par la FDA pour le traitement du mélanome, favorisent la réponse anti-tumorale en améliorant la présentation des Ag par les DCs et en augmentant la production d'interféron de type 1 (IFN) ce qui réduit l'inflammation[104].

Une autre approche intéressante provient du travail de bio-ingégnierie consistant à fabriquer des lymphocytes T tueurs spécifiques d'un cancer. Ces cellules tueuses sont appelées CAR T-cell (pour *Chimaeric antigen receptor T-cell)*[105]. Ces cellules sont soient prises directement chez le patient puis modifiées en laboratoire avant d'être réinjectées au patient, on parle alors de traitement autologue, soient prises chez des porteurs sains puis injectées chez le patient, on parle alors de traitement allogénique. Globalement la première approche montre le moins de risque de rejet ou de toxicité mais elle est plus longue à mettre en place et les cellules du patient peuvent avoir déjà des spécificités dues au cancer ou aux traitements subis. La seconde approche amène plus de risque de toxicité mais elle permet de créer une banque de cellules congelées à partir d'un donneur sain qui sont donc disponibles à tout moment.

En 2017 un traitement pour la leucémie aiguë lymphoblastique (ALL pour *Acute lymphoblastic leukemia*) chez l'enfant a été approuvé par la FDA sous le nom de Yescarta (axicabtagene ciloleucel).

1.3 Points de contrôle immunologique

1.3.a Les points de contrôle immunologique majeurs : CTLA4 et PD1/PD-L1

Au courant du XXème siècle les avancées en immunobiologie amenèrent les découvertes dans les années 1990 du rôle inhibiteur de CTLA-4, par James Allison , puis de la voie PD-1/PD-L1 par Tasuku Honjo. Ces chercheurs reçurent conjointement le prix Nobel de Physiologie ou Médecine en 2018 pour ces importantes découvertes.

Vers 2015 l'utilisation d'inhibiteurs de CTLA-4 puis de PD-1[103] a permis de mettre en pratique la levée de l'inhibition du système immunitaire afin de restaurer une réponse immunitaire efficace contre les cellules tumorales. De ces deux points de contrôle immunologique découlent actuellement une série de molécules sur le marché (Nivolumab, Pembrolizumab, Cémiplimab, Atézolizumab, Durvalumab, Avélumab, Ipilimumab, Trémélimumab). Mais ces premières molécules ont aussi montré des effets secondaires délétères très forts résultant de la levée de l'inhibition immunitaire qui n'est pas spécifique à la tumeur dans le corps humain. Depuis, plusieurs autres points de contrôle immunologique sont étudiés dans différents types de cancer.

CTLA4

CTLA-4[106–108] modifie la réponse des Tc via des voies métaboliques intrinsèques et extrinsèques incluant l'inhibition de la translation, le recrutement de phosphatases, l'activation de ligases ou l'inhibition de récepteurs à cytokines.

CTLA-4 rentre en compétition avec CD28 pour la liaison à CD80/86, le relargage de l'indoleamine-dioxygenase (IDO) et la modulation des

Lymphocytes T régulateurs (Treg) (Figure 18). Mais les effets varient grandement chez les patients et il nous manque toujours un schéma clair des mécanismes associés à CTLA4.

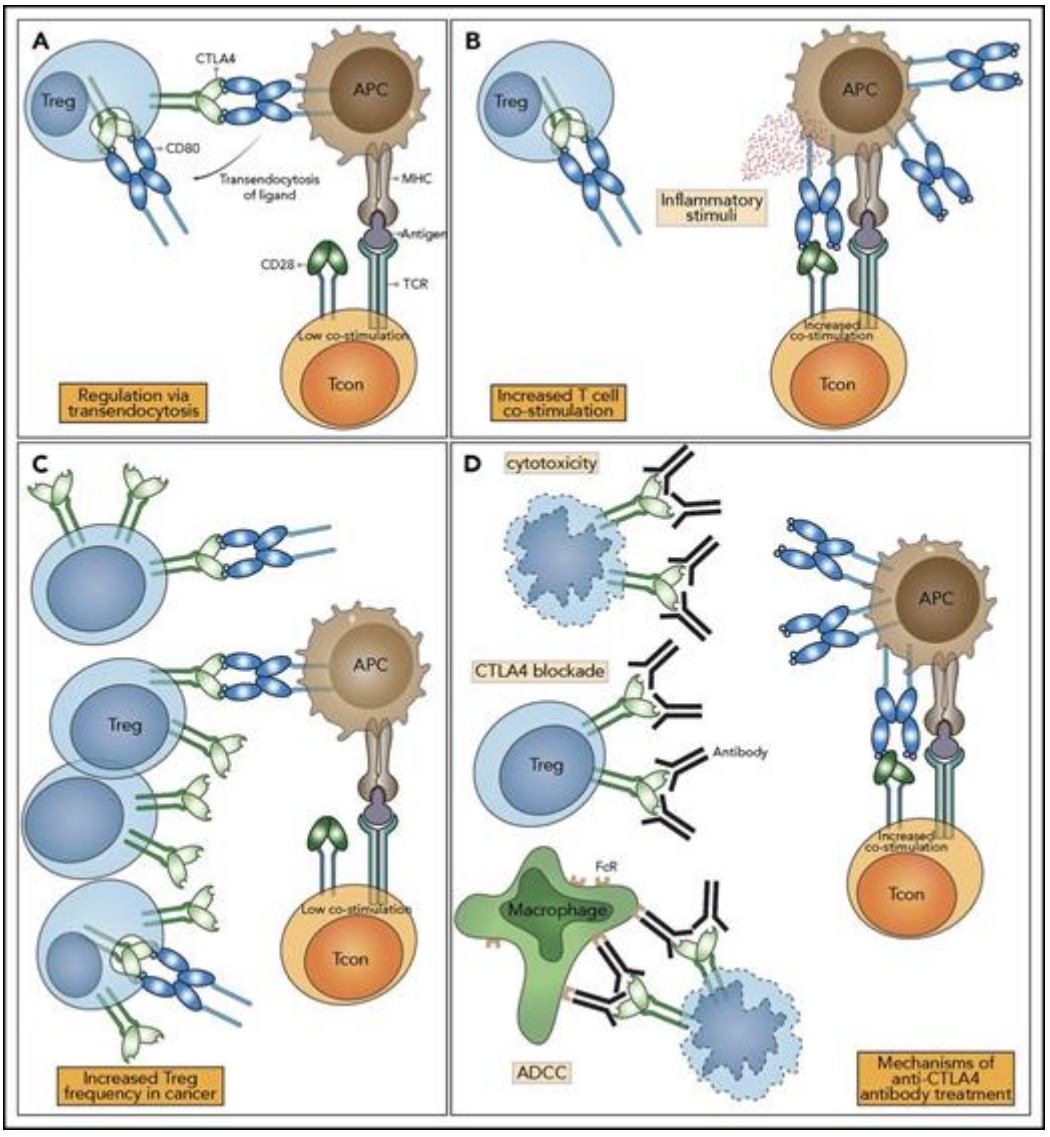


Figure 18. Modèles de CTLA-4 adapté de Rowshanravan, B et al [109]: Schéma général montrant les mécanismes connus de CTLA-4 et les traitements anti-CTLA-4.

PD1/PD-L1

PD-1[106,110] est un membre de la superfamille CD28 qui régule négativement l'activation des Tc. Le blocage de PD1 ou de son ligand PD-L1a montré un effet dans 20 à 30% des cas de différents cancer. Le signal est généré par un immuno-récepteur tyrosine cytoplasmique qui lie la protéine SHP-2 et limite la signalisation avec les Bc et Tc. On sait que PD-1-SHP-2 inhibe les TCR et la signalisation (Figure 19) CD28 mais on ne sait pas comment PD-1 agit sur les sous-types de Tc (Figure PD-1/PD-L1).

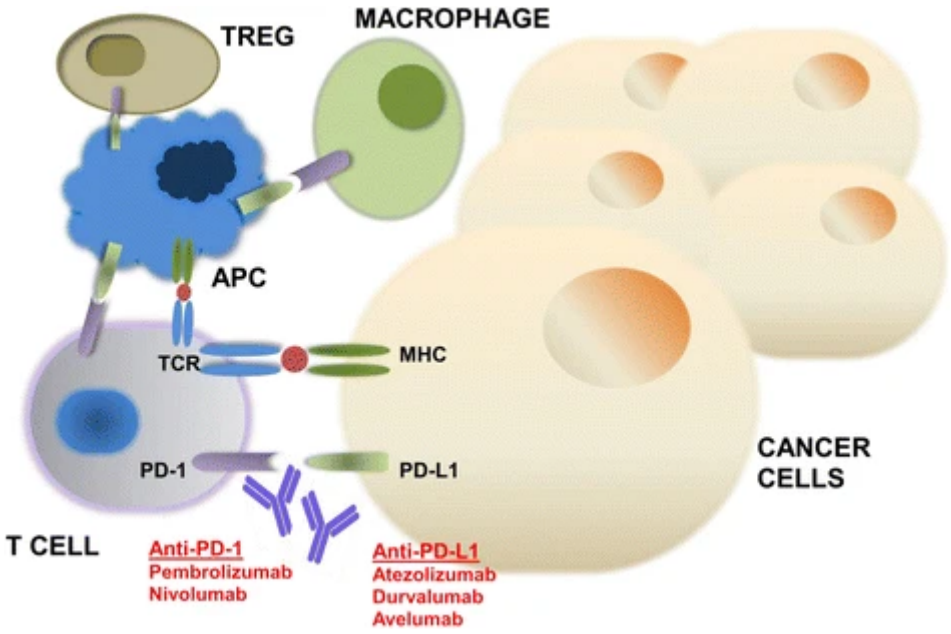


Figure 19. Modèle de PD-1/PD-L1 adapté de Gong J. et al [111] : Schéma général des mécanismes d'action connus de PD-1/PD-L1 .

1.3.a Les points de contrôle immunologique mineurs

Après les majeurs d'autres points de contrôle immunologique ont commencé à être étudiés :

- **LAG-3 (CD223) :** LAG-3 est un point de contrôle immunologique exprimé sur les CD4+ et CD8+ activé, Treg, *Natural Killer cells* (NK) et les cellules

B. Cette molécule est située près du récepteur CD4 et lie le MHC2 avec une forte affinité. LAG-3 est considéré comme un point de contrôle immunologique à fort potentiel[112].

- **OX40/OX40L**[113,114] **:** OX40 est une molécule appartenant à la classe des TNFR. Il est exprimé sur les Tc, NK, neutrophiles et sur-exprimé sur les Treg humains. Les études sur l'expression et fonctions d'OX40 dans les TILs ont montré que sa stimulation induisait l'activité effectrice des CD4+ et CD8+ et réduisait l'expression d'OX40 sur les Treg FOXP3+ infiltrant. Chez la souris l'administration d'agoniste d'OX40 favorise la prolifération des Tc, la production de cytokines et améliore la mémoire des Tc. Il est donc aussi un candidat à l'étude.
- Il existe aussi des traitements dont le but est d'activer des sentiers biologiques stimulant l'action des Tc. Notamment via la famille des *Tumor necrosis factor* (TNF) qui sont des molécules de la famille des glucocorticoides (TNFRSF18, TNFRSF9 ou encore le TNFRSF4)[104]. On parle alors de points de contrôle immunologique activateurs.
- Il existe beaucoup d'autres points de contrôle immunologique et sans être exhaustif on peut citer TIM-3, VISTA, B7-H3, ICOS/ICOSL etc.

1.3.c Les points de contrôle immunologique nouveaux : La famille des LILR

La liste des points de contrôle immunologique s'est allongée au cours des dernières années. Certaines molécules connues pour jouer un rôle dans la régulation de la réponse immune sont maintenant considérées comme des points de contrôle immunologique. C'est le cas entre autres des LILR. La description de cette famille est choisie par rapport aux résultats de l'article 1 inséré dans la thèse. Les connaissances sur cette famille évoluent vite depuis l'intérêt croissant pour l'immunothérapie de façon générale. La description de la

famille ne se veut donc pas exhaustive mais présente brièvement l'état des connaissances actuellest.

La famille des LILR, acronyme de *Human Leukocyte Immunoglobulin-like receptors*, est une famille de gènes codant pour 6 récepteurs activateurs (LILRA1-6) dont un soluble LILRA3 et 6 récepteurs immunosuppresseurs (LILRB1-5)[115,116]. Le nombre de copies de ces gènes est stable sauf pour LILRA3 et 6. Il existe aussi deux pseudo gènes (LILRP1, LILRP2) que nous ne détaillerons pas.

Comme tous les gènes liés aux mécanismes immunitaires sujet à la reconnaissance antigénique ces familles divergent rapidement au niveau phylogénétique. Il est donc difficile de faire des parallèles avec des modèles animaux. Néanmoins chez la souris le gène PirB est connu comme étant l'orthologue des gènes LILRB2/LILRB3.

Chez l'humain cette famille est nommée de différentes façons :

- CD85 antigen-like family
- ILT: Ig-like transcript 2
- LILR ou LIR: leukocyte immunoglobulin-like receptor

Une des approches possibles (et c'est celle considérée ici) consiste à rechercher tous les noms de gènes associés à un locus génomique spécifique. Dans notre cas la région d'intérêt est la 19q13.4 (Figure 20) et c'est à partir de ces coordonnées que nous avons pu récupérer les différents noms cités plus haut.

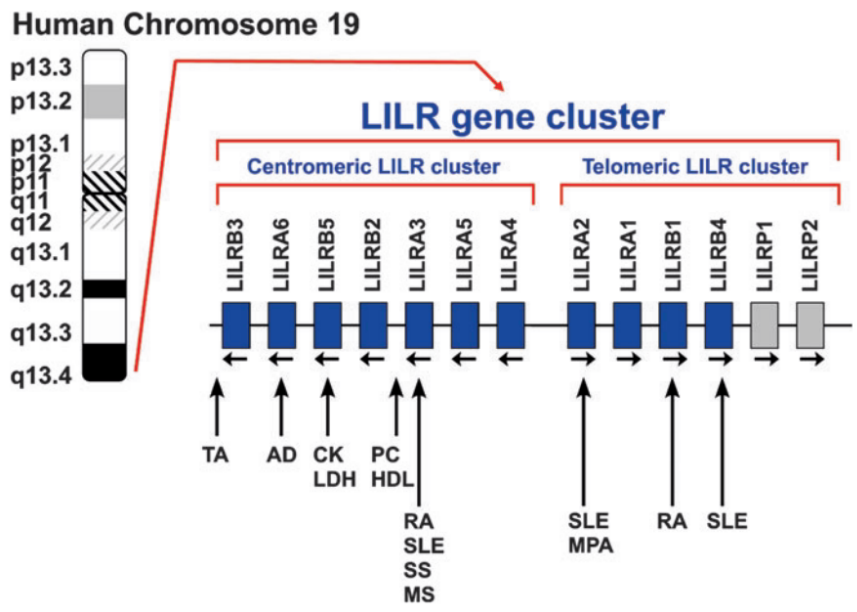


Figure 20. Organisation chromosomique de la famille LILR adapté de Hirayasu K. et al [117] : L'ensemble de la famille est situé sur le chromosome 19 et sous-tend une évolution homologue.

La famille des LILR est donc une famille pairée (activatrice et inhibitrice) dont la partie extracellulaire est fortement homologue mais dont la partie intracellulaire diffère grandement[118]. Les LILRB (inhibiteurs) possèdent 2 ou 4 domaines Ig et une longue partie cytoplasmique basée sur des immunorécepteurs inhibiteurs à tyrosine. Les LILRA (activateur) à l'inverse possèdent une courte portion cytoplasmique et des motifs activateurs tyrosine à chaîne FcRγ (Figure 21).

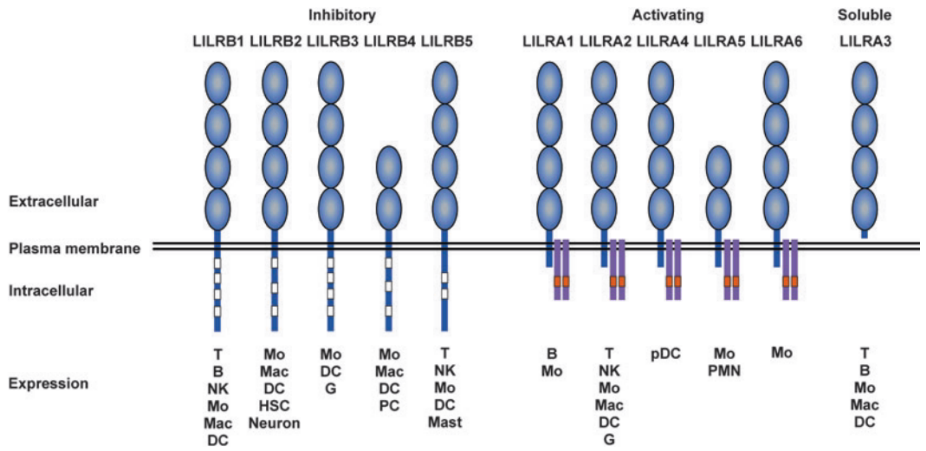


Figure 21. Structure globale des LILR adapté de Hirayasu K. et al [117] : Les différents gènes LIL impliquent des interactions avec différentes cellules immunitaires. Ces interactions ne sont pas décrites en détail expérimentalement et la structure des LIL n'est pas résolue.

La forme soluble est obtenue par épissage alternatif[119]. Les LILR sont classés en 2 groupes: le groupe 1 (LILRB1-2 et LILRA1-3) et le groupe 2 (LILRB3-5 et LILRA4-6). Cette classification est basée sur le résidu d'interaction LILRB1 qui interagit avec la molécule HLA[120].

LILRB

LILRB1[121]**:** LILRB1 reconnaît une large gamme des molécules HLA de classe I (HLA1) mais ne reconnaît pas les formes libres des ß2-microglobulines (B2M). La reconnaissance dépend d'un domaine A3 non polymorphique d'une chaîne lourde d'une HLA1 et B2M. Ce domaine est en compétition avec CD8 ce qui laisse à penser que LILRB1 module les CD8+ Tc en bloquant la liaison au CD8. LILRB1 est exprimé sur les cellules B, myéloïdes et une partie des cellules T et NK. Les HLA1 sont donc des ligands importants de LILRB1 qui permettent d'éviter des réactions auto-immunes grâce aux cellules du soi

exprimant les HLA1 et en permettant d'éliminer le non soi ou des cellules dérégulées qui n'expriment pas les HLA1 (Figure 22).

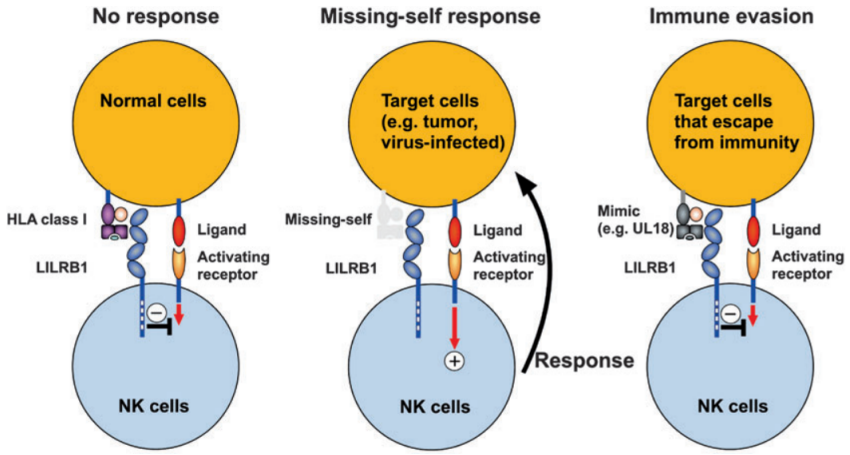


Figure 22. Modèle d'inhibition simplifié de LILRB1 adapté de Davidson, C. L. [121] **:** Les cellules immunitaires comme les NK, transmettent un signal inhibiteur et n'attaquent pas les cellules normales en reconnaissant les molécules HLA1 sur les cellules normales via LILRB1. Mais les NK tuent leur cible, comme des cellules infectées par des virus ou tumorales qui n'ont pas d'HLA1. Pour éviter le système immunitaire les virus ou bactéries expriment des répliques de LILRB1 pour inhiber directement les LILRA.

LILRB2: LILRB2 reconnaît quant à lui les formes libres des HLA1 et B2M mais aussi les chaînes lourdes, incluant HLA-B27 et HLA-G, et ce de façon plus forte que les HLA1 classiques. Le degré de liaison est modulé par l'Ag présenté par la molécule HLA1 à LILRB2. LILRB1 reconnaît aussi une protéine similaire aux HLA, CD1d, qui se trouve sur la surface des cellules et dans les compartiments lysosomaux et endosomaux pour bloquer le relargage des Ag lipidiques sur CD1d. LILRB2 reconnaît aussi des ligands non-HLA comme angiopoetin-like (ANGPTL), la B-amyloid, Nogo66, MAG et OMgp. LILRB2 est exprimé à la surface des cellules myéloïdes, des HSCc et des

neurones mais pas sur les cellules T, B ou NK. L'homologue murin semble être le gène PIR-B, la seule forme inhibitrice qui coïncide avec les formes et fonctions humaines. Les souris déficientes en PIR-B sont donc souvent utilisées pour analyser les fonctions de LILRB2. Ses fonctions inhibitrices dépendent des ligands et types cellulaires visés. LILRB2 est impliqué dans l'inhibition fonctionnelle des myélomonocytes en reconnaissant des HLA1. Pendant la différentiation cellulaire ANGPTLs inhibe la différenciation et permet l'expansion des HSCs via LILRB2. Dans la fonction neuronale, les inhibiteurs de myéline, incluant Nogo66, MAG et OMgp sont associés avec LILRB2 et régulent la régénération des axones. La b-amyloide enclenche la maladie d'alzheimer via LILRB2.

LILRB3: LILRB3 reconnaît les ligands associés à la cytokératine-8 sur des cellules glandulaires épithéliales nécrosées. La fonction générale de LILRB3 reste peu connue. LILRB3 est exprimé au niveau des monocytes, DCs et granulocytes. Des Ac allogéniques ont aussi été détectés sur des HSC transplantées ce qui pourrait affecter LILRB3. LILRB3 étant très polymorphique la production d'Ag est suspectée de mener à des allo-Ac.

LILRB4: LILRB4 est exprimé sur les monocytes, macrophages, DCs et les cellules plasmatiques. Une régulation positive de LILRB4 avec LILRB2 est associée avec la tolérisation des cellules présentant les Ag. Mais ses fonctions et ligands sont peu connues.

LILRB5: LILRB5 est exprimé sur les monocytes, DCs, mastocytes granulés, cellules T et reconnaît des chaînes lourdes HLA1. Mais ses fonctions ne sont pas connues.

LILRA

LILRA1: LILRA1 est exprimé sur les monocytes et les cellules B. LILRA1 se lie préfenretiellement aux chaînes lourdes d'HLA-B27 et HLA-C mais moins fortement que LILRB1 et LILRB2. On ne sait pas pourquoi à l'heure actuelle.

LILRA2: LILRA2 est exprimé sur les monocytes, macrophages, DCs, granulocytes et certains sous-groupes de Tc et NK. Son ligand endogène est inconnu mais sa fonction est de reconnaitre les situations immunitaires ou les anticorps sont détruits par les pathogènes.

LILRA3[122,123]**:** LILRA3 est le seule forme soluble de la famille. Elle est sécrétée par les monocytes, les cellules B et une fraction des Tc et jamais retrouvée à la surface des cellules. LILRA3 se lie aux molécules HLA1. LILRA3 a une très forte affinité pour le récepteur Nogo66 et promeut la formation des synapses en ayant une inhibition compétitive de Nogo66 par LILRB2.

LILRA4: LILRA4 est exprimé uniquement sur les DCs plasmacytoides (pDCs). LILRA4 reconnaît BST2 sur différentes cellules cancéreuses comme ligand et supprime l'activation des pDCs. Comme pour la plupart des LILR, le mécanisme est peu connu.

LILRA5: LILRA5 est retrouvé sur les monocytes et neutrophiles mais ses ligands et fonction sont inconnus. La structure crystaline diffère grandement de LILRB1 et 2 ce qui indiquerait qu'il ne lie pas les molécules HLA1.

LILRA6: L'expression de LILRB6 a été déterminé au niveau transcriptomique dans les monocytes. Ses fonctions ne sont pas connues.

Relation avec le CaP

Finalement il y a très peu de recherches qui sont mises en place spécifiquement pour comprendre les interactions entre le CaP et cette famille

de gènes spécifiques. Quelques articles pointent vers la propension des cellules à activer LILRB1 sur certaines cellules immunitaires pour échapper à leur contrôle[124,125] mais pour les autres on ne trouve pas d'articles bien ficelés et cohérents. Cette famille et son rôle dans le cancer reste donc largement inexplorée et inexploitée.

1.3.d Autres approches

D'autres approches vont essayer de stopper les complexes favorisant l'immunosuppression. Par exemple certains métabolites (kynurenine et adénosine) et des cytokines (TGF-ß) qui régulent les mouvements et la maturation des cellules myéloides suppressives (MDSCs pour *Myeloid-derived suppressor cells*) et les Treg sont sujets à différentes études.

On peut aussi citer les recherches effectuées avec les double anticorps qui redirigent les cellules NK ou les Tc vers la tumeur cible. Les constructions *ex vivo* de lymphocytes infiltrant la tumeur (TILs). Les constructions de Tc avec des Ac chimériques (CATs) ou des Tc recombinants (TCRs) qui se lient à la tumeur pour la détruire en sautant complètement les étapes de réponse immunitaire endogénique[126].

À l'heure actuelle la recherche se fait sur l'effet combinatoire d'un panel de molécules pouvant être utilisées ensemble plutôt qu'indépendamment[127]. Dans cette optique, il y a un besoin de connaître plus finement les fonctions des différentes molécules, leurs caractéristiques thérapeutiques (toxicité, tolérabilité, efficacité, index thérapeutique, dosage) et l'implication pour les patients ainsi que la viabilité financière de tels traitements[128].

1.4 L'immunothérapie du CaP

1.4.a Infiltration immunitaire

Le CaP n'est généralement pas un cancer à évolution rapide ce qui en fait théoriquement un candidat idéal pour des traitements d'immunothérapie car contrairement à la chimiothérapie et à la radiothérapie, l'immunothérapie nécessite plus de temps pour son action et ainsi elle est "présumée" être plus efficace pour les cancers à faible charge tumorale et évolution lente. Je dis présumée parce que comme expliqué plus haut les cancers qui ont le mieux répondu à l'immunothérapie ciblant les points de contrôle immunologique sont les cancers les plus agressifs tels que mélanome, poumon, rein, vessie etc. En ce sens, certaines études se sont appliquées à caractériser l'infiltration des cellules immunes dans les tumeurs prostatiques et à décrire l'impact de la présence de ces cellules sur le comportement des tumeurs.

On sait que les lymphocytes T infiltrent la tumeur en plus grande quantité que les lymphocytes B[129]. En particulier les Treg se retrouvent en plus grandes quantités dans les stades avancés de la maladie. Ces lymphocytes expriment une grande quantité de PD-1 indiquant un épuisement de ces cellules[130,131].

D'autres études ont montré que les lymphocytes B participaient à la progression du CaP en activant IKK-alpha, STAT3 et BMI1 dans les CRPC[130,132]. Il a aussi été montré que les cellules CD20+ se retrouvaient en plus grande quantité dans les tissus malins que bénins[133].

Le rôle du ratio neutrophiles/lymphocytes (NLR pour *neutrophil/lymphocyte ratio*) est encore mal compris. Certaines études ont montré que c'était un marqueur de mauvais pronostic post-docetaxel[134], un prédicteur de l'évolution de la PSA mais pas après une chimiothérapie[135] ou encore un indicateur bénin dans certains cas[136].

Des études ont aussi été faites sur les macrophages infiltrant les tumeurs (TAM pour *tumor-associated macrophages*). Ces cellules ont aussi été associées au pronostic mais dans des sous-types spécifiques de tumeurs et dans des groupes de patients spécifiques[137,138].

1.4.b Traitements

Le Sipuleucel-T est le premier traitement d'immunothérapie à être approuvé pour le traitement du CaP. Cette thérapie cellulaire personnalisée a été approuvée par la FDA en 2010 pour le traitement du mCRPC[139], Le traitement consiste à premièrement récupérer par leukaphérèse les cellules mononucléaires sanguines du patient. Ensuite les cellules sont cultivées en présence d'une protéine recombinante composée de l'antigène acide phosphatase prostatique (PAP) fusionné à la cytokine GM-CSF (pour *Granulocyte-macrophage-colony stimulating factor*). Suite à une période d'incubation, les cellules dendritiques se retrouvent chargées avec les peptides antigéniques de la PAP et la maturation des cellules est activée par le GM-CSF permettant ainsi une présentation optimale des Ag. Les cellules sont ensuite réinjectées au patient. Les effets à long terme sont positifs et les effets secondaires sont médicalement gérables[139]. Dans une étude de phase III randomisée chez des patients avec mCRPC minimalement symptomatiques, le Sipuleucel-T a augmenté la survie de 21.7 à 25.8 mois (HR 0.78 95%CI 0.61-0.98) sans par contre augmenter significativement la survie sans progression. Le problème est le ratio coût bénéfice de la thérapie qui limite son utilisation.

> "*Le GVAX est un vaccin cellulaire constitué de cellules tumorales prostatiques PC-3 et LNCaP irradiées et modifiées pour exprimer de manière constitutive le GM-CSF. PC-3 et LNCaP sont des lignées cellulaires établies et représentatives de cancers de la prostate humain. En théorie cette approche permettrait d'induire des réponses*

immunologiques à plusieurs antigènes associés aux tumeurs. Toutefois, 2 études cliniques de phase III VITAL-1 et VITAL-2 qui comparaient le GVAX au docétaxel plus prednisone chez des patients mCRPC ont été arrêtées pour des problèmes de faisabilité et des risques accrus de décès chez les patients recevant le GVAX. Le développement de cette thérapie a été ultimement arrêté"[140].

En dehors de ces thérapies il n'existe pas ou peu d'immunothérapies actives pour le CaP. Cela s'explique par le fait que le CaP présente peu de néo-antigènes en pratique comparativement à d'autres cancers[141,142]. Le CaP semble concentrer les mutations principalement dans les gènes de réparation de l'ADN plutôt que d'avoir un taux de mutation globalement élevé, et ce à tous les stades de la maladie[61,143,144].

Mais la plus grosse difficulté pour le développement de l'immunothérapie dans le cancer de la prostate est l'environnement immunosuppressif de la tumeur elle-même[145]. Les molécules hormonales dérivées des stéroïdes situées localement influencent aussi le TIME[146]. Cette combinaison de facteurs rend la compréhension des traitements particulièrement difficile et hétérogène.

Malgré cela les différents traitements testés pour d'autres cancers sont aussi évalués dans le CaP. Les premières études sur l'anti-CTLA4 (Ipilimumab) ont montré des résultats cliniques positifs sur certains patients mais n'ont pas été retrouvés dans des études avec des cohortes plus grandes[147]. Le pembrolizumab (anti-PD-1) est actuellement testé dans différents essais cliniques et certains patients mCRPC, notamment ceux présentant une déficience dans les mécanismes de réparation des mésappariements sembleraient bien répondre (Ref PMID: 29641940f). Pour les autres molécules anti-PD1 on peut citer en cours d'essai l'atezolizumab, le durvalumab et l'avelumab pour le traitement du mCRPC. L'essai clinique le plus avancé (phase III) compare l'atezolizumab utilisé comme agent principal en

combinaison avec l'enzalutamide. Un récent bilan de l'ensemble des traitements en cours d'essai a été publié[148].

1.5 Bioinformatique

1.5.a Analyses des omiques

L'évolution des technologies a apporté avec elle la production de quantité de données. Ce changement a fait entrer la bioinformatique dans l'aire des données massives (*Big data*)[149] et donc les problématiques qui y sont liées. La production massive de données nécessite une structure et une organisation intelligente (formatage, niveau de compression, organisation) de l'information sinon sous peine d'être rapidement inutilisées[150]. De ce besoin sont nés de multiples bases de données et systèmes d'organisation[151] (Matériel supplémentaire 1).

Les évolutions technologiques en matière de production de données omiques ont permis d'approcher la recherche en biologie de manière beaucoup plus large. Nos capacités à produire des données de plus en plus précises et interprétables concernant les expressions de gènes (RNA-seq), les micro-RNA (miRNA-seq), les mutations (SNP/SNV), les copies de génome (CNV/CNA) ou les données de méthylation apportent de nouvelles ambitions en terme scientifiques et techniques. La possibilité d'avoir une vue des relations entre ces données permet de mieux comprendre des mécanismes comme la transcription, la traduction et la fonction des protéines d'un point de vue global.

A ce jour la plupart des analyses omiques se concentrent soit sur un seul type de données soit mettent en corrélation des biomarqueurs sélectionnés indépendamment les uns des autres. Par exemple, on sélectionne un gène A

relié significativement à la maladie puis une mutation B. La sélection et l'association sont des étapes séparées et spécifiques de chaque omique. Cette simplification biologique a amené les chercheurs à se poser la question de comment pourraient être utilisées ces données en les analysant soit d'un bloc soit en plusieurs blocs tout en ayant une description relationnelle biologiquement pertinente.

De récents travaux ont tracé les grandes lignes de ce que pourrait être un protocole idéal dans le cadre de l'intégration de données multi-omiques[152–155]. D'autres ont étudié les aspects mathématiques[156] et certains ont déjà mis en place des applications pratiques en adaptant des méthodes déjà connues[157–160].

Selon Joerg Martin Buescher *et al.*[152] on peut isoler quelques axes importants d'analyse qui seraient communs à toutes approches multi-omiques:

- Comprendre le comportement statistique des différentes sorties de chaque type de données omiques de façon indépendante et en détail.
- Comprendre les relations entre les différents types de données omiques dans un contexte biologique particulier.
- Prendre en compte le temps d'analyse pour rester dans une échelle de temps raisonnable au vu des projets planifiés.

Dans le cas de ce projet de thèse nous serons toujours dans l'optique d'une approche supervisée puisque nous cherchons à comprendre pourquoi des patients avec CaP récidivent alors que d'autres ne récidivent pas en identifiant des caractéristiques propres aux patients qui récidivent. L'approche générale pour étudier des données dont la finalité est la prédiction d'un état ou l'association de variables à un état suit 3 étapes[161]:

- Association spécifique avec la ou les variables à prédire
- Vérification de la reproductibilité et de la précision du résultat

- Vérification de l'indépendance des variables explicatives vis à vis des autres signaux

Dans le cadre des analyses multi-omiques appliquées à la santé la finalité est très souvent la recherche de biomarqueurs et de nouvelles cibles thérapeutiques. Un biomarqueur peut se définir comme n'importe quelle entité biologique mesurable comme un gène via son niveau d'expression; de mutation ou d'amplification, ou encore des miRNA, cRNA, une protéine etc. Un biomarqueur peut aussi être une combinaison de différentes entités qui définissent une signature diagnostique, pronostique ou autre. Depuis plusieurs années, un certain nombre de biomarqueurs[162–164] diagnostiques, pronostiques ou théranostiques sont proposés dans la littérature scientifique. Ce nouvel essor est bercé par les espoirs que porte en elle la médecine personnalisée. Ce nouvel espace de recherche promet, en oncologie et ailleurs, d'améliorer autant la prévention que les soins des individus concernés. Cette science implique donc la question du développement des biomarqueurs de plus en plus spécifiques. Dans le CaP la PSA est utilisée depuis longtemps en tant que biomarqueur. La révolution en biologie moléculaire qu'est l'avènement du séquençage de haut débit et la caractérisation de plus en plus fine des tissus tumoraux permettent d'aller de plus en plus loin vers la découverte des sentiers biologiques aberrants au niveau moléculaire. Cet engouement apporte la découverte de biomarqueurs qui ne passent que rarement l'examen pratique et la validation expérimentale. Ce décalage s'explique en partie par le coup inhérent à la recherche en biologie moléculaire mais aussi par un manque de confiance dans les résultats qui semblent loin de la réalité expérimentale malgré un besoin croissant. Ce décalage peut s'expliquer par certains facteurs:

- La réplicabilité et la spécificité des expériences faisant émerger de nouveaux biomarqueurs.
- Les coûts de développement inhérents à la biologie expérimentale.

- Le développement des champs bioinformatique et statistiques de plus en plus indépendant des équipes cliniques (manque de collégialité).

Il convient donc avant tout de souligner certains points pour s'assurer de la qualité des découvertes et donc de minimiser les risques d'échec clinique en aval. La mise en place d'une procédure adaptée pour augmenter les chances *d'utilité* de nos futurs biomarqueurs est donc essentielle.

Des points proposés par Lisa M McShane *et al.*[163] nous retranscrivons ici ceux qui concernent notre travail et sont possibles à contrôler:

- Rechercher dans toutes les données disponibles si des problèmes ou des indications méthodologiques ont été publiés vis à vis des données de travail.
- Évaluer les données fournies et répartir des données brutes si les critères d'évaluation ne sont pas satisfaits.
- Avoir connaissance des biais expérimentaux et d'analyse pour pouvoir les intégrer dans des processus de correction ou de modélisation. Cela inclut par exemple les effets de lot, les manipulations d'échantillons, les spécificités des plateformes, manipulateurs ou des produits utilisés. Cette vérification est essentielle pour éviter les artéfacts menant à des liens entre les données omiques ou les résultats cliniques.
- Choisir et évaluer des méthodes statistiques appropriées.
- S'assurer de la compatibilité des outils utilisés et de leur constance en termes de résultats pour la durée effective du projet.
- Documenter les sources de variations qui pourraient affecter la reproductibilité des résultats finaux ainsi que la variabilité globale au cours du travail de recherche.
- Travail de revue des résultats pour évaluer leur pertinence dans le domaine de recherche.

Si cet ensemble de points est bien défini alors la faisabilité du passage à la clinique a plus de chance de réussir.

1.5.b Projet internationaux

L'évolution des technologies et des méthodes d'analyse ont permis aux scientifiques de penser des grands projets de recherche. Ces projets comme le TCGA (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga) ou encore Allan brain atlas (https://portal.brain-map.org/) ont mis en avant les capacités des chercheurs à travailler en collaboration, proposé des résultats accessibles à tous (sous certaines conditions), favorisé la recherche pluridisciplinaire et permis des milliers de publications.

Les gros projets posent des difficultés que tout chercheur va rencontrer s'il touche de près ou de loin aux données générées par ces projets. Si l'on se concentre sur le projet TCGA il y a plusieurs points à noter.

Gestion des données : Les données multi omiques du projet TCGA vont de centaines de To à des centaines de Go en fonction du sous projet sur lequel on travaille. Il est donc nécessaire de mettre en place des systèmes de stockage adaptés ou de découper son travail en parcelles pour pouvoir gérer ces données. Le stockage informatique redondant pour avoir une meilleure sécurité coûte très cher et cet aspect technique est peu abordé dans la mise en place des projets de thèse.

La modification des données : Sur des projets aussi gros, il y a toujours des mises à jour des données. Il est très important de trouver un moyen de récupérer ce flux d'information qui peut aller d'un simple ajout de patient, d'information mais aussi de retrait de données pour cause de mauvaise qualité. En général les publications primaires (groupe de chercheurs ayant eu la primauté sur les données du projet) s'occupent de faire un premier gros

nettoyage, comme pour le projet PRAD, et si le site du projet est bien fait il y a toujours une page de notes pour les données (https://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/). Dans le cas de TCGA cette page a été ajoutée à la refonte du site et il était initialement très dur de s'y retrouver initialement.

Type des données : Selon le projet les formats de données peuvent être très uniformes ou très hétérogènes selon le nombre de technologies utilisées. Pour chacun de ces formats il existe des outils dédié pour l'analyse et le traitement. Dans l'idéal, il faut partir des données brutes si elles sont disponibles. Cela permet d'uniformiser les pipelines tout en utilisant les outils les plus à jour à chaque fois. La suite de logiciels samtools[165] propose un ensemble d'outils permettant l'analyse de fichiers provenant de différentes technologies de séquençage haut débit par exemple. Des instituts se spécialisent aussi dans la création d'outils dédiés. Le Broad institute (https://www.broadinstitute.org/) développe régulièrement de très bons outils d'analyse utilisés par un grand nombre de scientifiques et souvent utilisés par défaut dans des projets internationaux.

Sécurité des données : Dans le cadre du projet TCGA une partie des données sont en accès libre complet et une autre part en accès restreint. Le fait de restreindre l'accès aux données a deux objectifs: permettre de communiquer de façon spécifique certaines recommandations et notamment les termes de citation du projet dans les travaux publiés mais aussi, le plus important, lié moralement et juridiquement les utilisateurs aux données. Ces données restent des données médicales de patients. Ces patients font confiance aux scientifiques pour gérer leurs données de façon sécuritaire et responsable. Cela implique au niveau IT un minimum d'organisation et de sécurité qui sont rappelés lorsque l'accès aux données restreintes est donné. En général, le niveau de sécurité minimum implique de travailler dans des espaces réseaux fermés et de ne communiquer les données que par des canaux dédiés.

1.5.c Analyse de données en grande dimension

L'analyse de données en grande dimension repose sur plusieurs aspects qui vont de la préparation des données, à la sélection de variables et le choix des algorithmes d'analyse.

La préparation des données

Cette étape peut sembler trivial mais elle représente 80% du travail. Définir le type de problème (classification, régression, etc.), nettoyer les données, transformer les données sont des étapes essentielles et souvent spécifiques des méthodologies qui vont être utilisées. Typiquement la vérification des hypothèses de normalité si l'on veut utiliser des modèles de régression peut devoir obliger à effectuer des transformations de Box-Cox[166]. Certains algorithmes sont sensibles à des transformations usuelles comme le fait de centrer-réduire les données ou la normalisation Z. Il faut donc en avoir conscience pour anticiper les variations de résultats.

Sélection de variables

La sélection des variables en amont de l'analyse a plusieurs avantages. Elle permet d'éviter le surapprentissage en enlevant les variables redondantes et diminue le bruit d'apprentissage (Matériel supplémentaire 2,3,4[178]). Cette réduction du bruit entraîne généralement *de facto* un gain de performance mais aussi un gain de temps d'analyse.

On découpe en général les méthodes de sélection de variables en 3 groupes:

- **Filtre simple:** Une mesure d'association est choisie et les variables filtrées dessus. Typiquement la corrélation ou le score de Chi peuvent être utilisés.
- **Méthodes emballées (Wrapper-based):** Ces méthodes considèrent la sélection de variables comme un problème en soi et y appliquent des algorithmes particuliers. Typiquement la méthode d'élimination des ariables redondantes (*Recursive Features Elimination*).
- **Intégrées (Embedded):** Ces méthodes sont intégrées dans les algorithmes ou les approches elles-mêmes. La méthode Lasso pour la régression ou les forêts aléatoires sont de ce type.

Comprendre ses données : Spécificités des données médicales

Les données médicales possèdent bien souvent, du fait de la difficulté à recruter des patients, une structure dite horizontale. C'est à dire que le nombre d'échantillons (*e.g*. patients) est très inférieur au nombre de variables explicatives (*e.g*. gènes). Cela s'explique par le coût inhérent de la recherche médicale et la difficulté d'intégrer de larges cohortes de patients dans des études. Cette forme particulière des données pose des problèmes méthodologiques et analytiques.

Dans le cas des approches classiques par des méthodes de régression si l'on a I observations (*e.g*. patients) décrites par K variable dépendante (*e.g.* BCR) et l'on note Y=I*K d'un côté et qu'on a J un ensemble de variables (*e.g.* gènes) décrivant ces observations I et qu'on note X = I*J alors le but d'une analyse multivariée est de décrire la structure reliant Y et X. Quand Y est un vecteur et X une matrice alors la structure peut être définie par des régressions multiples.

Dans le cas des données omiques, on a un nombre d'observations J très grand devant I. Dans ce cas, il arrive souvent que des variables se retrouvent

liées par une relation linéaire alors qu'il n'y a pas de lien réel entre elles. Mathématiquement cela se traduit par une relation de type $Z = aX + bY$ et définit la multicolinéarité. Si une variable indépendante est une combinaison linéaire d'autres variables alors, en régression multiple, il peut devenir difficile voire impossible d'inverser la matrice des moments X^TX. La multicollinearité tend en fait à rendre la matrice X singulière. Son déterminant ne peut donc être calculé et la matrice inverse X^{-1} n'existe pas et donc la solution classique $(X^TX)^{-1}X^T$ en régression n'est pas utilisable. Les approches par régression deviennent donc très complexes à utiliser.

D'autres limites pour ce type d'approche ont été bien définies par Nicolas Verzelen[167]. Il a démontré que, quelle que soit la méthodologie choisie, il y a une limite à la pertinence de l'estimation des paramètres quand $n << k*Log(p/k)$ où k est le nombre de variables potentielles incluses dans le modèle de régression.

Algorithmes d'analyse

Il existe un grand nombre d'algorithmes et je ne vais pas essayer d'en faire un inventaire exhaustif. De mon point de vue le choix se fait en fonction des données elles-mêmes. Dans le cas des données médicales, le problème de multicollinéarité est central.

Une façon de contrer le problème de multicolinéarité est d'utiliser des algorithmes d'apprentissage machine. Par ce terme on entend des algorithmes plus récents qui s'opposent aux régressions classiques. En pratique ces algorithmes (*e.g*. forêts aléatoires [168], réseaux de neurones[169], séparateurs à vaste marge (*SVM*)[170] sont nombreux et sont connus depuis plusieurs décennies. Ce qui les a remis en vue de tous est le développement de l'informatique et les capacités calculatoires qui en découlent. L'avènement d'internet et l'accumulation des données qui nous a fait entrer dans l'ère des données massives (*Big data*) a permis de les rendre utiles comparativement

aux méthodes de régression classiques. Ces approches ont en outre l'avantage de pouvoir gérer un grand nombre de variables.

Mais ces approches apportent peu d'information sur les processus biologiques sous-jacents lorsqu'elles sont utilisées. La complexité algorithmique et mathématique derrière rend leur interprétation difficile au-delà de leur excellente performance en termes de prédiction. À ce niveau, ces approches sont d'ailleurs les plus performantes et chaque année différents concours de d'apprentissage machine permettent (*e.g.* Kaggle) de situer l'évolution des différentes méthodes.

Quoiqu'il en soit, les approches par apprentissage machine nécessitent une méthodologie particulière quel que soit le contexte dans lequel elles sont utilisées[171,172].

Langages

Le choix du langage de code est loin d'être anodin. Au delà du fait de posséder les compétences dans un langage il y a un choix technique important à faire. Une bonne façon d'avoir une vue globale de l'utilisation des langages utilisés est de se référer au résumé annuel de StackOverflow (https://insights.stackoverflow.com/survey/). Il n'existe bien évidemment pas de langage parfait mais on distingue surtout des tendances par domaine d'expertise.

Typiquement les personnes ayant des formations en statistiques ou purement universitaires auront tendance à privilégier R comme langage, celui-ci ayant été inventé en milieu universitaire pour des statistiques.

Les personnes ayant des formations pluridisciplinaires se dirigeront principalement vers python et Java car ce sont des langages versatiles et aussi les plus utilisés à l'heure actuelle.

Les personnes ayant des formations en informatique pures auront tendance à privilégier le C/C++ sur des projets de développement.

Le choix dépend donc aussi de la population cible. Si un projet n'a pas vocation d'être un outil ou une méthodologie généralisable alors le choix du langage est spécifique à celui qui étudie les données. En revanche si le code doit être publié sous forme d'un outil ou d'un script réutilisable alors là des questions se posent. Les langages les plus recommandés dans ce cas la sont Java ou python. Java étant le seul langage utilisable sur n'importe quelle machine et python étant le langage le plus utilisé en industrie à l'heure actuelle. Les packages R sont à l'heure actuelle principalement limité au milieu académique. Ils sont peu utilisés en dehors du milieu universitaire même si de plus en plus les industries de recherche se tournent vers ce langage. L'entreprise Rstudio poussant et finançant ce développement en proposant des outils et interfaces web de plus en plus ergonomiques et pratiques.

Plus récemment le langage Julia (*https://julialang.org/*) créé au MIT commence à proposer différentes librairies utilisables d'apprentissage machine avec une typologie orientée pour une parallélisation simple. C'est donc un nouvel outil à surveiller avec attention

Librairies

Différentes librairies d'analyse sont disponibles sous différents langages. Les trois principales sont MLR[173] pour R, scikit-Learn[174] pour Python et Weka[175] pour Java. Le choix d'une librairie dépend plus d'une commodité par rapport au langage que de différences réelles entre les méthodes proposées à l'intérieur.

Des librairies comme mixOmics[176] ont été développées pour des problèmes de données omiques même si dans l'absolu les algorithmes utilisés sont exportables à tout type de problème.

1.5.d Régression des moindres carrés partiels

Pour contrer le problème de multicolinéarité et du trop grand nombre de variables, différentes approches ont été proposées. La plus simple est de sélectionner des variables en amont de l'analyse avec un à priori conceptuel. Dans le cadre de régression complexe on peut utiliser des approches de type ridge/lasso par exemple.

Une autre possibilité est d'utiliser les propriétés réductrices de l'analyse en composante principale (ACP). En décomposant X en $X = R\Delta V^t$ avec R et V les matrices gauche et droite des vecteurs singuliers et Δ une matrice diagonale contenant les valeurs singulières. On peut ainsi utiliser R comme composantes principales pour décrire Y. Les propriétés orthogonales de la décomposition en valeurs singulières permettent d'éliminer une majeure partie des problèmes de multicolinéarité. Comme ces décompositions expliquent X il faut par la suite trouver quelles composantes de X décrivent le mieux Y.

Les approches de régression des moindres carrés partiels (PLS pour *Partial least square*)[177] ont été développées dans ce cadre. Elles conviennent donc parfaitement à l'analyse des multi-omics ou le nombre de variables (*e.g.* gènes, protéines etc) est très supérieur au nombre d'échantillons (*e.g*. patients, cellules etc). La PLS va chercher un groupe de composantes, appelés vecteurs latents, en faisant une décomposition simultanée de X et Y tout en ayant la contrainte de maximiser la covariance entre X et Y.

Les approches développées dans le package mixOmics sont des méthodes de régression de ce type. Ces approches vont chercher à maximiser la covariance entre les quantités Xa et u Yb sachant que les matrices X et Y ont déjà été réduites grâce à l'ACP. Cela implique de fixer $a^Ta=1$ et $b^Tb = 1$ ce qui rend Xa et Yb maximaux. Quand le premier vecteur latent est trouvé, il est enlevé de X et Y. Ce processus est répété jusqu'à réduire X à une matrice nulle.

Ce processus permet donc de définir des vecteurs latents décrivant au mieux la structure de covariance entre X et Y pour des composantes définies.

L'approche *Multi Integrative method* (MINT)[179] est dérivée de ces méthodes vectorielles. Dans les analyses multi-omiques il est toujours difficile de reproduire les études et donc d'obtenir des résultats fiables. Une des solutions est de combiner des analyses indépendantes ou de comparer leurs résultats. Le problème est que la variabilité technique empêche souvent ce type d'approche en rendant difficile la comparaison des résultats. L'approche MINT a justement été développée pour pouvoir faire cette intégration de données tout en prenant en compte ces effets techniques. C'est une analyse de PLS étendue à plusieurs groupes qui sélectionne des variables d'intérêt liées à une réponse. Le détail et la logique mathématiques sont les suivant :

PLS-DA	$\max_{\|\|a_h\|\|_2=\|\|b_h\|\|_2=1} cov(X_h a_h, Y_h b_h)$
mgPLS-DA	$\max_{\|\|a_h\|\|_2=\|\|b_h\|\|_2=1} \sum_{m=1}^{M} n_m cov\left(X_h^{(m)} a_h, Y_h^{(m)} b_h\right)$
MINT	$\max_{\|\|a_h\|\|_2=\|\|b_s\|\|_2=1} \sum_{m=1}^{M} n_m cov(X_h^{(m)} a_h, Y_h^{(m)} b_h) + \lambda_h \|\|a_h\|\|_1$

Ces approches sont particulièrement utiles dans un contexte d'analyse biologique puisqu'elles permettent de travailler avec des jeux de données provenant de sources différentes tout en gardant des résultats interprétables. Ce qui est de leur avantage par rapport aux approches d'apprentissage machine.

1.5.e Gain d'information et forêt aléatoires (apprentissage machine)

Dans le cadre de mon travail, une approche de réduction de variables basée sur l'entropie, le gain d'information, et un algorithme, les forêts aléatoires (*random forest - RF*) m'ont particulièrement intéressés.

Gain d'information

Le gain d'information est une méthode permettant de réduire le nombre de variables en choisissant celles qui apportent le plus d'information. Cette information est en fait dérivée du calcul de l'entropie statistique. Le calcul de l'entropie suit la formule:

$$E = -\sum_i p_i log_2 p_i$$

Où E est l'entropie du système et p_i la probabilité de tirer l'attribut i d'une classe de variable. Dans un exemple simple, si l'on a un fichier clinique avec une colonne contenant l'information sur la récidive dans le cancer (qui prend une valeur binaire, le patient récidive ou non) p_i est la probabilité de tirer l'une ou l'autre des catégories.

En métagénomique l'entropie de Shannon est simplement la somme des entropies de chaque attribut pour une classe donnée c'est donc une mesure de diversité.

Concrètement si l'on prend une variable qui ne prend qu'une seule valeur alors son entropie est $E = - 1*log_2 1 = 0$, on voit rapidement que cette variable n'apporte pas d'information, elle est non informative. Dans le cadre des

packages utilisés la sélection de variables par gain d'information est défini comme:

$$GI(attribut) = E(Classe) + E(Attribut) - E(Classe, Attribut)$$

Le gain d'information d'un attribut nous donne donc la quantité d'information par rapport à la classe cible. Elle mesure la différence d'information entre les cas où l'on connaît la valeur de l'attribut et le cas où on ne la connaît pas. En soit n'importe quelle mesure d'information peut-être utilisée mais c'est l'entropie de Shannon qui est la plus répandue en pratique.

Dans un cas extrême ou une classe ne prendrait qu'une valeur, le gain d'information serait toujours 0 puisqu'aucun attribut nous apporterait de l'information.

Après avoir attribué un score GI à chaque attribut on peut sélectionner ceux qui nous intéressent le plus et réduire ainsi la dimensionnalité de notre jeu de données.

Forêts aléatoires

Les forêts aléatoires (RF pour *Random forest*) ont initialement été proposées en 1995 (Tin Kan Ho, AT&T Bell Laboratories)[180] puis améliorées par divers groupes de recherche.

Pour comprendre le RF il faut déjà comprendre le principe d'un arbre de décision. Pour cela je vais définir quelques termes en premier lieu:

- Noeud racine (*root node*) : C'est le nœud initial contenant toute la population qui va devoir être subdivisée en nœud de population de plus en plus homogène.
- Feuille ou nœud terminal : C'est un nœud final que l'on ne divise plus.

Dans les algorithmes d'arbre de décision, la population est découpée et répartie dans chaque nœud en partant de la racine jusqu'aux feuilles. Pour chaque division une variable critère est utilisée. Par exemple, si l'on veut

prédire si un individu est Québécois au sein de la population Canadienne on pourra utiliser la variable “aime la poutine - oui/non” qui va porter beaucoup d’information. Cette première séparation en deux noeud va permettre de rendre plus homogène les noeuds suivants et ainsi de suite jusqu'à ce qu’on retrouve dans chaque nœud un population unique. Bien sûr, le nombre de nœuds et de branches sont des paramètres à définir pour éviter le surapprentissage par exemple.

Pour résumer, un arbre de décision calcule l’entropie de chaque variable par rapport à une réponse et construit un nœud à partir de cette valeur. Cet ensemble va donner un arbre composé de nœuds qui vont mener à une prise de décision via l’entropie de chaque branche.

Les RF vont utiliser ce principe en agrégeant en fait différents arbres de décision, c’est le principe du *bagging* (mise en sac). En plus du principe de *bagging*, les RF ajoutent de l’aléatoire au niveau de la population initiale. Pour chaque arbre on sélectionne un échantillon *bootstrap* d’individus et à chaque étape, la construction d’un nœud de l’arbre se fait sur un sous-ensemble de variables tirées aléatoirement.

On se retrouve donc avec plusieurs arbres (modèles), une composante aléatoire et des prédictions différentes pour chaque individu. Comment obtenir l’estimation finale? Dans le cas d’une réponse qualitative on regarde la variable la plus fréquente et pour une réponse quantitative on fait simplement la moyenne des valeurs prédites.

De plus, cet algorithme est très populaire de par ses propriétés le rendant insensible à diverses transformations des variables, notamment le redimensionnement (*scaling),* ce qui permet de mélanger des données de différentes valeurs mais de même nature[181]. Cette propriété est aussi très pratique pour discuter avec les experts des domaines d’intérêt et amener à des questions sur les variables elles-mêmes.

Les modèles produits sont inspectables dans une certaine mesure. Cela dépend du nombre de nœuds et de forêts de l'ensemble du système.

1.5.f Analyse multi-omiques appliquées au CaP

Avant de pousser les analyses mutli-omiques j'ai cherché quels pourraient être les jeux de données potentiellement utilisables. Au début du projet j'ai pu répertorier :

- 2016 - mCRPC - CGH/RNA-Seq - 176 samples - Fred Hutchinson [182]
- 2016 - nPC - Meth - 114 samples - multi institute [183]
- 2015 - PCa - SNV/CNA/RNA-Seq/Meth/miRNA - 333 samples - TCGA [78]
- 2015 - mCRPC - SNV/CNA/RNA-Seq - 150 samples - SU2C/PCF dream team [144]
- 2014 - PCa - CNA - 104 samples - MSKCC [184]
- 2013 - PCa - SNV/CNA - 57 samples - Broad/Cornell [185]
- 2012 - mCRPC - SNV - 61 samples - MCTP [186]
- 2012 - PCa - SNV - 112 samples - Broad.Cornell [187]
- 2010 - PCa - CNA/RNA-Seq - 240 samples - MSKCC [188]

Si l'on considère qu'une analyse multi-omique comporte RNA-Seq/CNA/SNV/Meth/miRNA il n'y a à l'heure actuelle que le projet TCGA qui propose un tel ensemble de données en association avec des informations cliniques pour un nombre relativement grand de patients (~500). La plupart des autres jeux de données ont tiré avantage des technologies de séquençage en analysant par exemple uniquement le transcriptome (*Whole exome sequencing*) puis en analysant les mutations/expression sur ces données. L'autre facteur déterminant est le temps de suivi des patients qui pour nous était un critère de sélection essentiel vu l'importance de la survenue de la BCR dans le temps pour déterminer la progression de la maladie. Ce facteur réduit

considérablement le nombre de cas analysables (ou éligibles) lorsqu'on considère les données cliniques disponibles dans ces jeux de données.

Au moment où j'écris ces lignes, j'ai pu trouver d'autres jeux de données :

- 2020 - PCa - CNA/CDK12 focus - 1465 samples - MSK [189]
- 2019 - mCRPC - SNV/RNA-Seq - 444 samples - SU2C/PCF dream team [190]
- 2018 - PCa - SNV/CNA - 1013 samples - MSKCC/DFCI [191]
- 2017 - PCa - SNV/CNA/RNA-seq/Meth - 477 samples - CPC-Gene [79]
- 2017 - PCa - SNV/RNA-Seq - 65 samples - SMMU [192]
- 2017 - PCa - SNV/CNA - 504 samples - MSKCC [193]
- Et un programme large comme PIONEER[194]

Il n'existe donc pas d'équivalent de TCGA à l'heure actuelle. Par contre si l'on se concentre sur les SNV/CNA on trouve beaucoup de données.

1.6 Problématique, hypothèses et objectifs du projet

1.6.a Problématique

D'une part, le CaP répond peu aux traitements immunothérapeutiques actuels [195–197] il y a donc nécessité d'améliorer la compréhension du micro-environnement tumoral pour pouvoir, soit mieux utiliser les traitements déjà développés, soit proposer de nouvelles pistes de solution pour augmenter les chances de survie des patients à risque élevé de progression.

D'autre part, il y a un besoin de développement d'outils pouvant prédire efficacement l'évolution de la maladie d'un patient. L'orientation des patients pour améliorer leur prise en charge et les orienter rapidement dans les traitements adéquats est une des clefs pour augmenter les chances de survie.

1.6.b Hypothèses

Hypothèse 1 ou hypothèse exploratoire : Les études actuelles en génomique du CaP retrouvées dans la littérature se limitent à l'analyse de série de données de même type. Mon hypothèse est que l'analyse de plusieurs séries de données soit de même type ou de type différents par des approches multi-omiques novatrices apportera des résultats plus probants ou significatifs sur le plan biologique et permettra d'ouvrir des pistes d'exploration pour améliorer la compréhension du micro-environnement tumoral. De plus, en concentrant nos analyses sur un ensemble de variables liées à l'immunité notre travail sera orienté vers la recherche de sentiers métaboliques spécifiques pouvant être utilisés dans le contexte de l'immunothérapie.

Hypothèse 2: La majorité des praticiens hospitaliers actuels se limitent à l'utilisation du grade, stade et du taux de PSA majoritairement pour établir le pronostic de la maladie car actuellement ces paramètres sont encore ceux qui prédisent le mieux l'évolution de la maladie. Mon hypothèse est qu'en utilisant des données omiques seules de type RNAseq obtenues via des approches par séquençage (et non simple quantification comme le microarray) il est possible d'obtenir, par des approches d'apprentissage machine, des signatures géniques ayant des scores de prédiction supérieurs à ces données clinico-pathologiques.

1.6.c Objectifs

Pour répondre aux hypothèses définies, mes objectifs sont:

- **Objectif 1:** D'une part, mettre en place une analyse multi-omique pour essayer de proposer des cibles pour le développement de nouvelles approches immunothérapeutiques du CaP. Cette recherche est présentée au chapitre 1.
- **Objectif 2:** Et d'autre part mettre en place un ensemble de méthodes d'apprentissage machine pour prédire la BCR à partir de données RNA-Seq. Cette recherche fait l'objet du chapitre 2.

Les objectifs sous-jacents sont :

- Récupérer des jeux de données adéquats en cherchant ceux qui peuvent répondre à la problématique et aux hypothèses. Dans notre cas des jeux de données omiques pour des patients avec et sans récidive en essayant d'équilibrer au maximum les classes.
- Mettre en place un protocole d'analyse complet et automatisé pour préparer les données récupérées.
- Il faut choisir la méthode statistique la plus appropriée par rapport aux hypothèses et à la nature des données.

Chapitre 1 : Article 1

2.1 Résumé

Jusqu'à présent l'immunothérapie du cancer de la prostate (CaP) a montré peu d'efficacité. Les chances de réussite de tels traitements pourraient être améliorées en utilisant des approches plus adaptées à l'immunobiologie du CaP. L'objectif de cette étude est de mettre en place une analyse multi-omiques pour identifier des biomarqueurs de progression du CaP et mieux comprendre la biologie de la maladie tout en proposant de nouvelles cibles immunothérapeutiques. Les jeux de données RNA-Seq, miRNA, de méthylation, variation du nombre de copies (CNV) et les mutations (SNV) provenant du projet PRAD de TCGA ont été récupéré, nettoyé et ré-analysé après avoir filtré sur un ensemble de variables liées à l'immunité. Deux autres jeux de données RNA-Seq ont ensuite été utilisés pour confirmer les résultats trouvés dans l'analyse TCGA. La régression des moindres carrés avec sélection de variables (Sparse Partial Least Squares-Discriminant, sPLS-DA) a été utilisée pour identifier les variables associées avec la récidive biochimique (BCR) dans chaque type de données omiques. Le niveau d'expression des variables sélectionnées a ensuite été contrôlé dans les différents jeux de données RNA-Seq en utilisant la méthode Multivariate INTegrative (MINT). Des courbes de survie ont ensuite été calculées pour les gènes sélectionnés afin de vérifier l'association avec la BCR. Les variables sélectionnées prédisent la BCR avec un taux d'erreur équilibré (BER) de 0.20 à 0.51 dans les différentes données omiques individuelles et de 0.05 avec les données omiques regroupées. La famille des Immunoglobulin Ig-like Receptor (LILR) est associée avec la BCR et le temps à la récidive que ce soit dans les analyses multi-omiques ou RNA-Seq. Les patients avec une forte expression des LILR ont plus de chance d'avoir une BCR et plus rapidement. Des agents immunothérapeutiques ciblant cette famille de marqueurs pourraient donc être développés pour ralentir la progression du CaP.

2.2 Titre et auteurs

Immune-Focused Multi-Omics Analysis of Prostate Cancer: Leukocyte Ig-Like Receptors are Associated with Disease Progression

Benjamin Vittrant[1,2], Alain Bergeron[2], Oscar Eduardo Molina[2], Mickael Leclercq[1], Marie-Laure Martin-Magniette[3,4,5], Colin Collins[6], Yves Fradet[2], Arnaud Droit[1]*

[1] Laboratoire de bioinformatique, Centre de recherche du CHU de Québec-Université Laval, Québec, Canada

[2] Laboratoire d'Uro-Oncologie Expérimentale, Axe Oncologie, Centre de recherche du CHU de Québec-Université Laval, Québec, Canada

[3] UMR MIA-Paris, AgroParisTech, INRAE, Université Paris-Saclay, Paris, France

[4] Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91405, Orsay, France.

[5] Université de Paris, CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), 91405, Orsay, France

[6] Vancouver Prostate Cancer centre, Vancouver, British Columbia, Canada

*Correspondence should be addressed to: arnaud.droit@crchuq.ulaval.ca

Running title: LILR association with prostate cancer progression

Keywords: Prostate Cancer, Biochemical Recurrence, Multi-Omics Analysis, Leukocyte Ig-Like Receptors, Immunotherapy

Funding: This research was realized with internal funds from the Laboratoire d'Uro-Oncologie Expérimentale. The production of RNA-seq data at VPCC was realized with funds from the Terry Fox Research Institute New Frontier Program Project Grant #1062.

Disclosure statement: No potential conflicts of interest were disclosed.

2.3 Abstract

Prostate cancer (PCa) immunotherapy has shown limited activity so far. The success rate of PCa immunotherapy could be improved by approaches more adapted to the immunobiology of the disease. The objective of this study was to perform a multi-omics analysis to identify immune genes associated with PCa progression to better understand PCa immunobiology and propose new immunotherapeutic targets. The mRNA, miRNA, methylation, copy number alteration, and single nucleotide variation datasets from The Cancer Genome Atlas PRAD cohort were analyzed after filtering for genes associated with immunity. Sparse partial least squares-discriminant analyses were performed to identify features associated with biochemical recurrence (BCR) in each type of omics data. Selected features predicted BCR with a balanced error rate (BER) of 0.20 to 0.51 in single-omics and of 0.05 in multi-omics analyses. Among features associated with BCR were genes from the Immunoglobulin Ig-like Receptor (LILR) family which are immune checkpoints with immunotherapeutic potential. Using the Multivariate INTegrative (MINT) analysis, the association of five LILR genes with BCR was found in a combination of three RNA-seq datasets and further confirmed with Kaplan-Meier curves in the combined RNA-seq datasets as well as in an independent RNA-seq dataset. Finally, immunohistochemistry showed that a high number of LILRB1 positive cells within the tumors predicted long term adverse outcomes. Thus, tumors characterized by abnormal expression of LILR genes have more chances to recur. The immunotherapeutic potential of these immune regulators to stimulate the immune response against PCa should be evaluated in pre-clinical models.

2.4 Introduction

Prostate cancer (PCa) immunotherapy has been mostly attempted with therapeutic anti-cancer vaccines using either dendritic cell-based, whole cell-based or vector-based vaccines as well as with some other approaches but always with limited efficacy [1–3]. In recent years, the development of immune checkpoint inhibitors has revolutionized the field of cancer immunotherapy. Immune checkpoints are a series of receptors/ligands that either inhibit or activate the function of immune cells [4]. CTLA-4 and PD-1 or its ligand PD-L1 are the most well-known immune checkpoints but many others have been identified [5,6]. The inhibition of CTLA-4 and PD-1/PD-L1 has shown impressive anti-tumor activity in cancers such as melanoma, lung, kidney and bladder cancers and efforts are made to identify strategies to improve their efficacy, notably through the identification of biomarkers for the selection of patients more likely to respond or through combination with other drugs or therapies [7].

First attempts of PCa immunotherapy using immune checkpoint inhibitors have been however rather disappointing. Two Phase III trials testing Ipilimumab (anti-CTLA-4) showed no effect on the overall survival compared to placebo although it had some positive impact on progression free survival [8,9]. Phase I and II trials testing PD-1 or PD-L1 inhibitors have also been conducted. These studies showed that a higher anti-tumor activity could be observed in subsets of patients with tumors showing DNA repair abnormalities or when the inhibitors were used in combination with other drugs such as enzalutamide or olaparib[10]. Although these results are encouraging, PCa immunotherapy is still suboptimal and more effective approaches must be identified. A better understanding of the antitumor immune defects associated with PCa progression will help to develop immunotherapies more adapted to this disease.

In this study, we performed a multi-omics analysis using The Cancer Genome Atlas (TCGA) PCa datasets to identify immune-related features associated with biochemical recurrence (BCR; i.e a rise in PSA level after radical prostatectomy) with the presumption that the identification of features

common to different types of omics data would support their relevance and importance in the immunobiology of PCa. These analyses led us to identify some leukocyte immunoglobulin-like receptors (LILR) as candidate biomarkers of BCR.

LILR are a family of immune receptors that either activate (LILRA members) or suppress (LILRB members) the immune cell functions. These Type 1 transmembrane glycoproteins are composed of 2 or 4 extracellular Ig-like domains that bind ligands and either a short cytoplasmic tail with an immunoreceptor tyrosine-based activation motif (ITAM), for a LILRA members of the family, or a long cytoplasmic tail with an immunoreceptor tyrosine-based inhibitory motif (ITIM), for the LILRB members of the family [11,12]. These receptors are widely expressed in hematopoietic-lineage cells but their exact function is still not well understood. LILRA and LILRB have been shown to bind to various ligands including membrane bound proteins such as MHC class I molecules (most strongly with HLA-G molecules) and soluble proteins such as angiopoietin-like proteins (ANGPTLs), myelin inhibitors and S100A8/9 proteins[12]. LILR up- or downregulation was shown to impact the response to bacterial and viral infections as well as to influence the outcomes in diseases such as autoimmunity, inflammatory diseases and cancer[13].

We report here the details of our multi-omics analyses and discuss the potential of LILR and especially LILRB1 as targets for PCa immunotherapy.

2.5 Results

Eligibility and preparation of data

The TCGA PRAD project comprises a total of 498 men treated by radical prostatectomy. From these cases a large number are inadequate in terms of quality of RNA sequencing as indicated by TCGA PRAD team[14]. Moreover, as our objective was to identify features associated with BCR which may happen several years after radical prostatectomy, we selected cases with a minimum of 5 years of clinical follow-up and combined them with those cases where BCR

happened in less than 5 years. We therefore discarded several cases that did not encounter our eligibility criteria (see details in Material and Methods). This imposed an important selection as only 45 cases were conserved for this analysis (Supplementary Figure S1). The resulting cohort is enriched in high-risk PCa which may help to identify features associated with BCR (Supplementary Table S1).

mRNA, miRNA, methylation, CNV, and SNV datasets from TCGA PRAD project were recovered and curated. Non-informative data in each dataset were discarded. RNA-seq data were completely reanalyzed. The numbers of CNV and SNV data were considerably reduced by the selection of features. Following this first step, the number of RNA, miRNA, Methylation, CNV and SNV features were 29820, 1211, 20112, 13925 and 928 respectively. We next applied our filter to select features associated with a set of 812 immune genes. After filtering for immune genes the resulting number of RNA, miRNA, Methylation, CNV and SNV features were 768, 1211, 768, 138 and 6 respectively (Supplementary Figure S2).

Features associated with BCR

To identify immune gene-related features associated with BCR, prediction modeling was performed using sparse partial least squares-discriminant analysis (sPLS-DA). These analyses were first performed on each type of omics separately. Supplementary Table 2 provides the list of features identified by the sPLS-DA in each omics dataset. Overall, 51 mRNA, 44 miRNA, 36 methylation loci, 32 CNV and 6 SNV were identified. The selected mRNA, miRNA and methylation loci predicted well occurence of a BCR with BER of 0.20, 0.23 and 0.26, respectively while the selected CNV and SNV, with BER of 0.46 and 0.51, respectively, did not predict as well BCR (Figure 1). We next merged all those features into one single set of data and performed a general sPLS-DA analysis. This resulted in an almost perfect prediction of BCR with a BER of 0.05 (Figure 2).

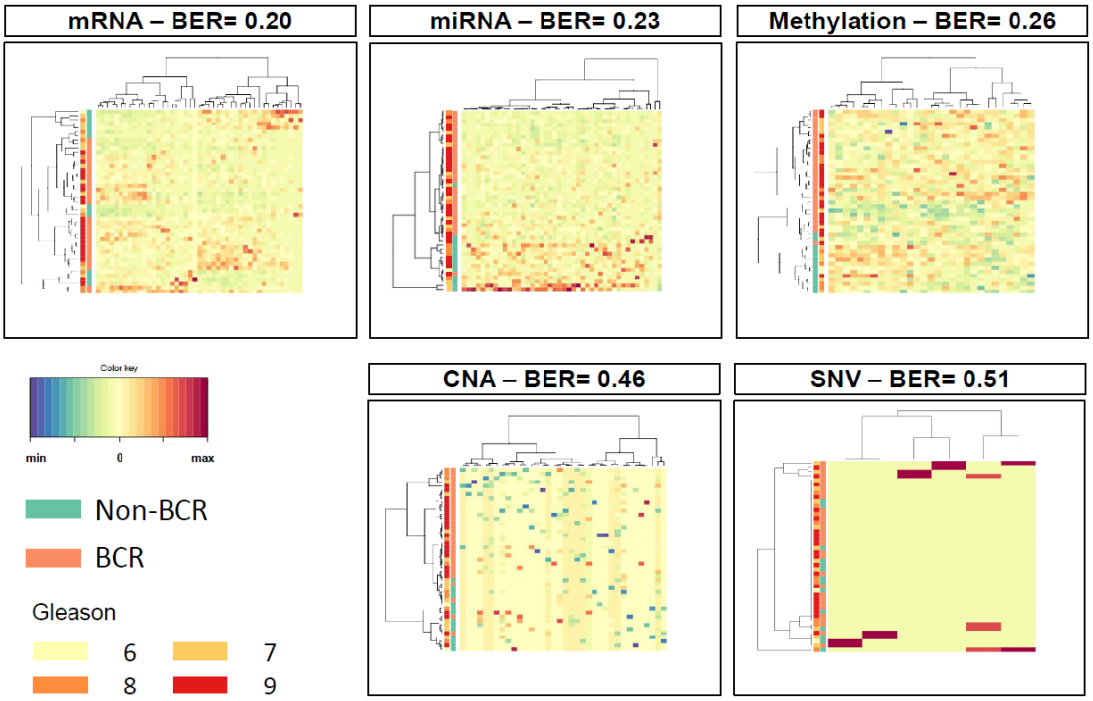


Figure 1: Results of sPLS-DA in individual omics datasets from TCGA PRAD. To identify immune-related features associated with BCR, prediction modeling was performed using sPLS-DA. Overall, 51 mRNA, 44, miRNA, 36 methylation loci, 32 CNA and 6 SNV were identified. The selected mRNA, miRNA and methylation loci predicted well BCR with BER of 0.20, 0.23 and 0.26, respectively. With BER of 0.46 and 0.51, the selected CNA and SNV, respectively, did not predict well BCR because of insufficient power.

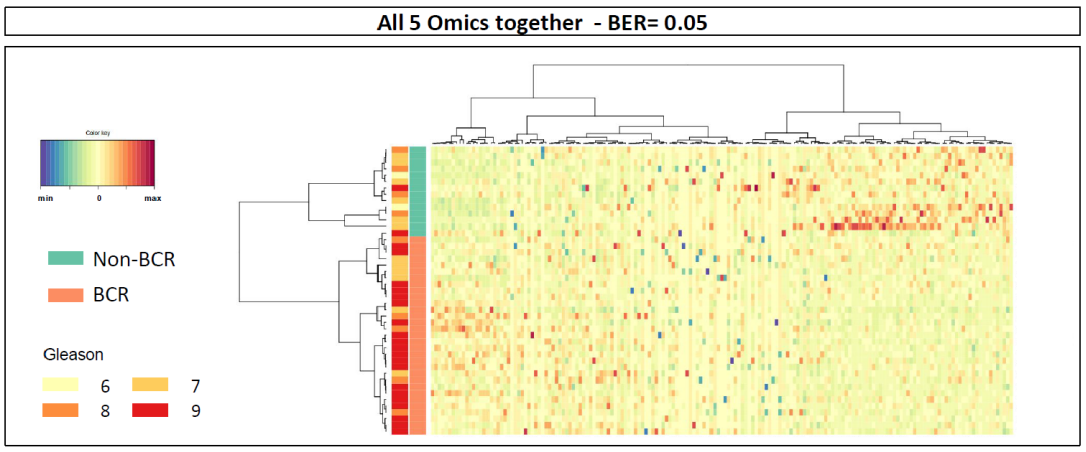


Figure 2: Results of sPLS-DA of the combined omics dataset from TCGA PRAD. Following the sPLS-DA of individual omics datasets, here we merged those features into one single set of data and performed a general sPLS-DA analysis. This resulted in an almost

perfect prediction of BCR with a BER of 0.05

Analysis of mRNA selection showed that many features were associated with leukocyte activation, cell activation, regulation of catalytic activity, immune system process, intracellular signal transduction etc. (Supplementary Figure S3). Among these were some features associated with antigen processing and presentation that were also retrieved in the other types of omics data. These comprised predominantly HLA molecules, killer-cell immunoglobulin-like receptor (KIR) and leukocyte immunoglobulin-like receptor (LILR) genes (Supplementary Table S2). Being receptors of MHC molecules, LILRs are a family of immune regulators with either activator or suppressor activities that are showing some potential as targets for cancer immunotherapy[15,16]. We thus further analyzed the association between 30 LILR gene-associated features and BCR (Supplementary Table S3). As shown in Figure 3, the 30 LILR gene-associated features alone could predict BCR in sPLS-DA analysis with a BER of 0.34 suggesting a role of this family of receptors in the progression of PCa. Similar analyses were performed with features associated with the KIR genes or the HLA genes, alone or in combination with those of LILR genes. However these analyses did not provide better BER indicating that genes of the LILR family were more strongly associated with BCR (results not shown).

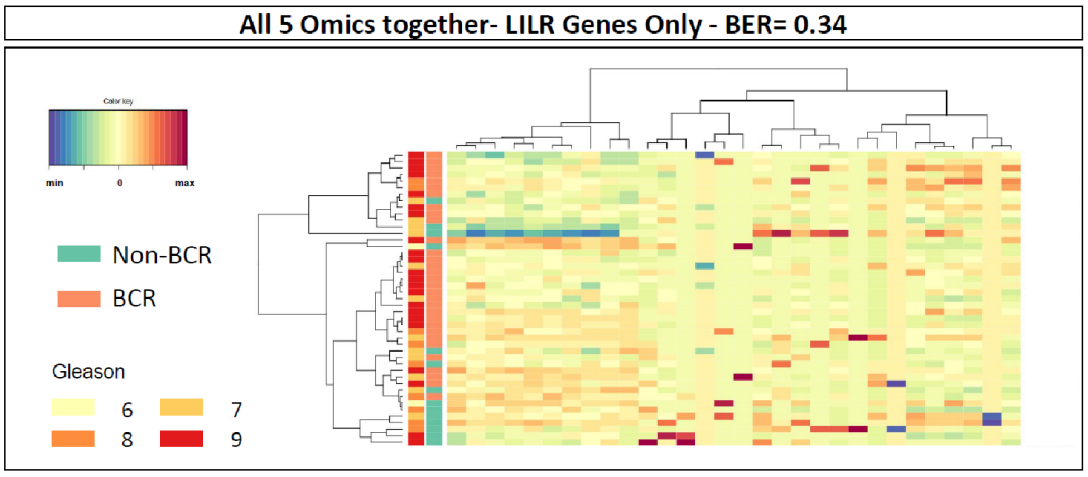


Figure 3: Results of sPLS-DA of LILR gene-related features in the combined omics dataset of TCGA PRAD. The analysis of the 30 LILR gene-related features in the combined dataset resulted in the prediction of BCR with a BER of 0.34.

To further validate the association of LILR genes with BCR and because of the paucity of PCa multiomics datasets beside TCGA, we sought to validate this association in a combination of RNA-seq datasets of PCa that would represent more than 150 patient tumors to ensure statistical power. Therefore we selected the GSE54460 RNA-Seq dataset from Long et al.[17] and a RNA-seq dataset from C. Collins[18] with the objective to combine them with the RNA-seq dataset from the 52 selected cases of the TCGA PRAD project. We reanalyzed these new RNA-seq data in the same way as we did for the TCGA PRAD RNA-seq data to ensure proper assembly of the datasets (Supplementary Figure S4). Thereafter we used the Multivariate INTegrative (MINT) approach to assess whether the LILR genes were associated with BCR in the combined RNA-seq dataset of 171 tumors (Supplementary Table S4). Figure 4 shows that five genes from the family of LILR genes (LILRB1, LILRB2, LILRB3, LILRB5 and LILRA3) altogether were associated with BCR with a BER of 0.34 confirming an association between LILR genes and BCR.

All 3 RNA-seq datasets- LILR Genes Only - BER= 0.34

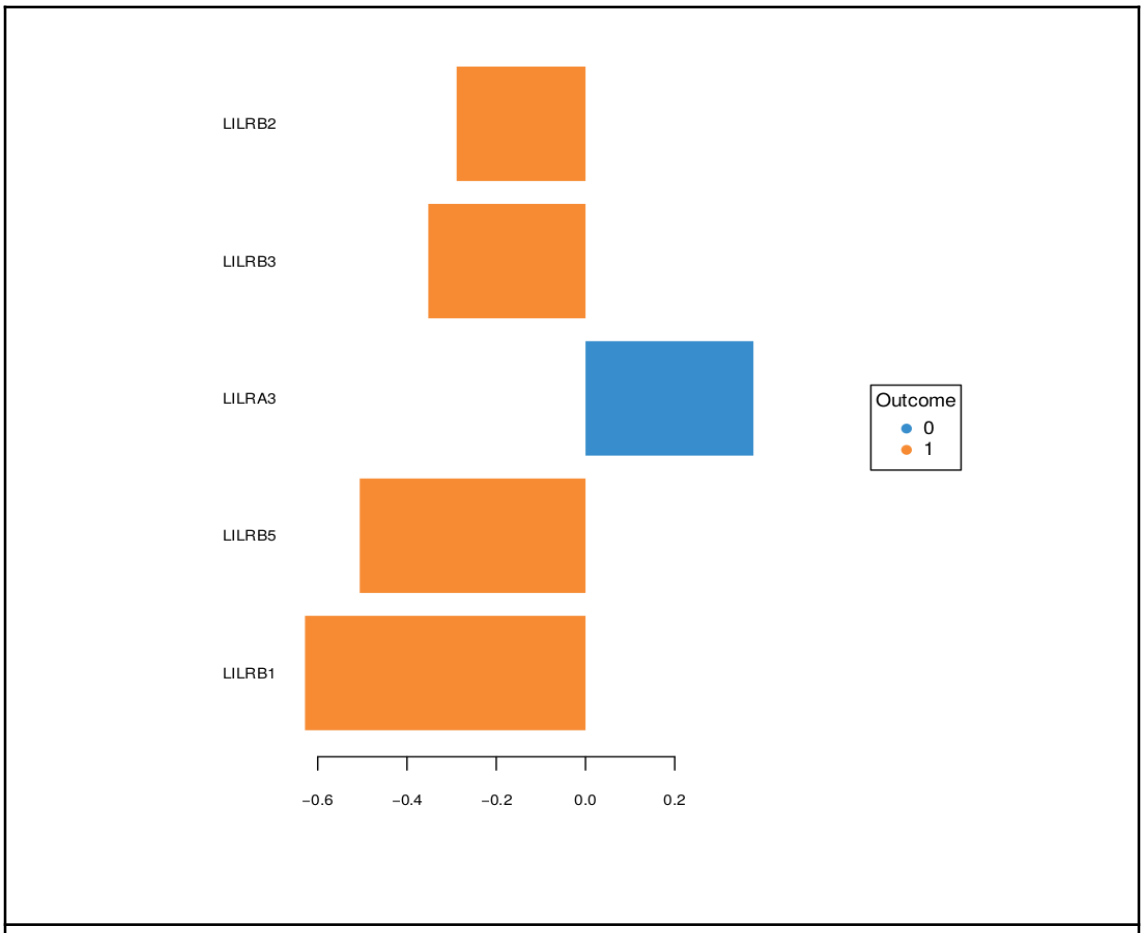


Figure 4: Results of the MINT analysis. Using the combined TCGA-GSE54460-VPCC RNA-seq dataset, five LILR genes were found to be associated together with BCR. The five LILR genes were associated with BCR with a BER of 0.34

In order to further characterize the association of these LILR genes with occurence of BCR, we next performed Kaplan-Meier analyses. Figure 5 shows that, as revealed by the MINT analysis, LILRB1 is a gene that is strongly associated with BCR as a high expression of LILRB1 is associated with a shorter BCR-free survival (Figure 5A; log-rank $p < 0.0001$). High expression of LILRB2 is also associated with shorter BCR-free survival but this association is less significant (Figure 5B; log-rank $p=0.04$). High expression of LILRB5 is nearly significantly associated with a shorter BCR-free survival (Figure 5D; Log-rank $p=0.06$) while that of LILRB3 alone is not associated with BCR-free survival (Figure 5C; log-rank $p=0.27$). When the expression of these four genes is summed, the association with BCR-free survival is not better (Figure 5E; log-rank $p=0.008$) than that of LILRB1 alone indicating that LILRB1 is the main

driver of the association. Moreover, removing LILRB1 from the combination greatly affects the significance of the association (not shown), supporting the importance of LILRB1. At the opposite, high expression of LILRA3 was significantly associated with a better BCR-free survival (Figure 5E; log-rank p=0.003).

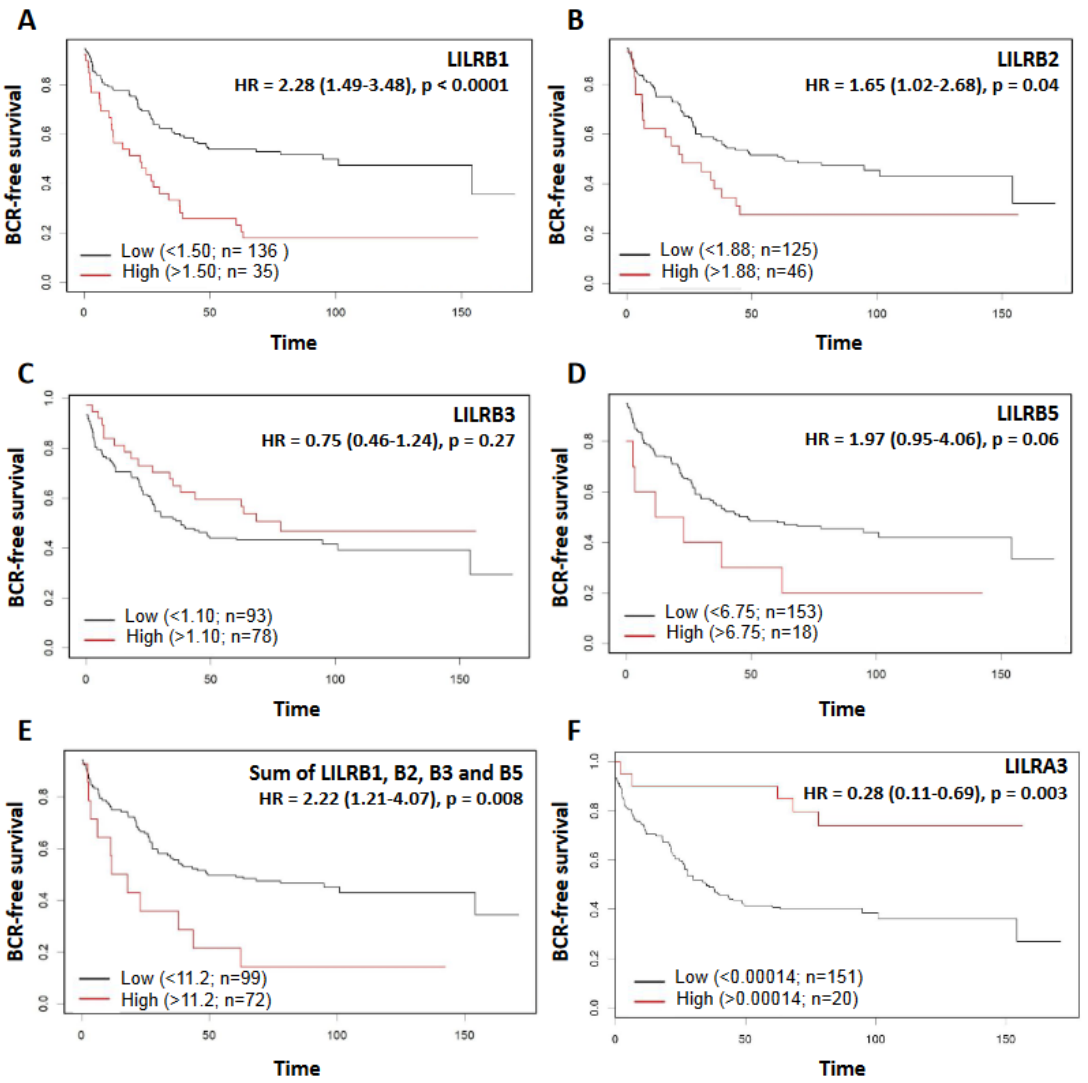


Figure 5: Kaplan-Meier analysis of dichotomized expression of LILRB1 (A), LILRB2 (B), LILRB3 (C), LILRB5 (D) and LILRA3 (F) genes in the combined TCGA-GSE54460-VPCC RNA-seq dataset. High level expression of LILRB genes have a tendency to be associated with shorter BCR-free survival but only LILRB1 and LILRB2 have a significant p value. The sum of the expression of these genes (E) was also associated with a significantly shorter BCR-free survival. At the opposite, higher expression of LILRA3 (F) gene is associated with a higher BCR-free survival.

As the combined TCGA-GSE54460-VPC cohort is composed of high risk tumors, we next sought to determine whether the association of these genes with BCR was maintained in a cohort of intermediate-risk PCa samples. We therefore performed new Kaplan-Meier analyses using a sub-cohort of the CPC-GENE project[19] comprising 144 men operated at our institution (Supplementary Table S5). Figure 6B-D shows that high expression of LILRB2, LILRB3 and LILRB5 genes was significantly associated with a shorter BCR-free survival. However LILRB1 gene expression was not significantly associated with BCR in this cohort although a trend can be observed. No data of LILRA3 gene expression was found in the CPC-GENE dataset, therefore the association of this gene with BCR could not be assessed. When taken all together, the sum of the expression of LILRB1, LILRB2, LILRB3 and LILRB5 genes was still associated with BCR (Figure 6E; log-rank p=0.009).

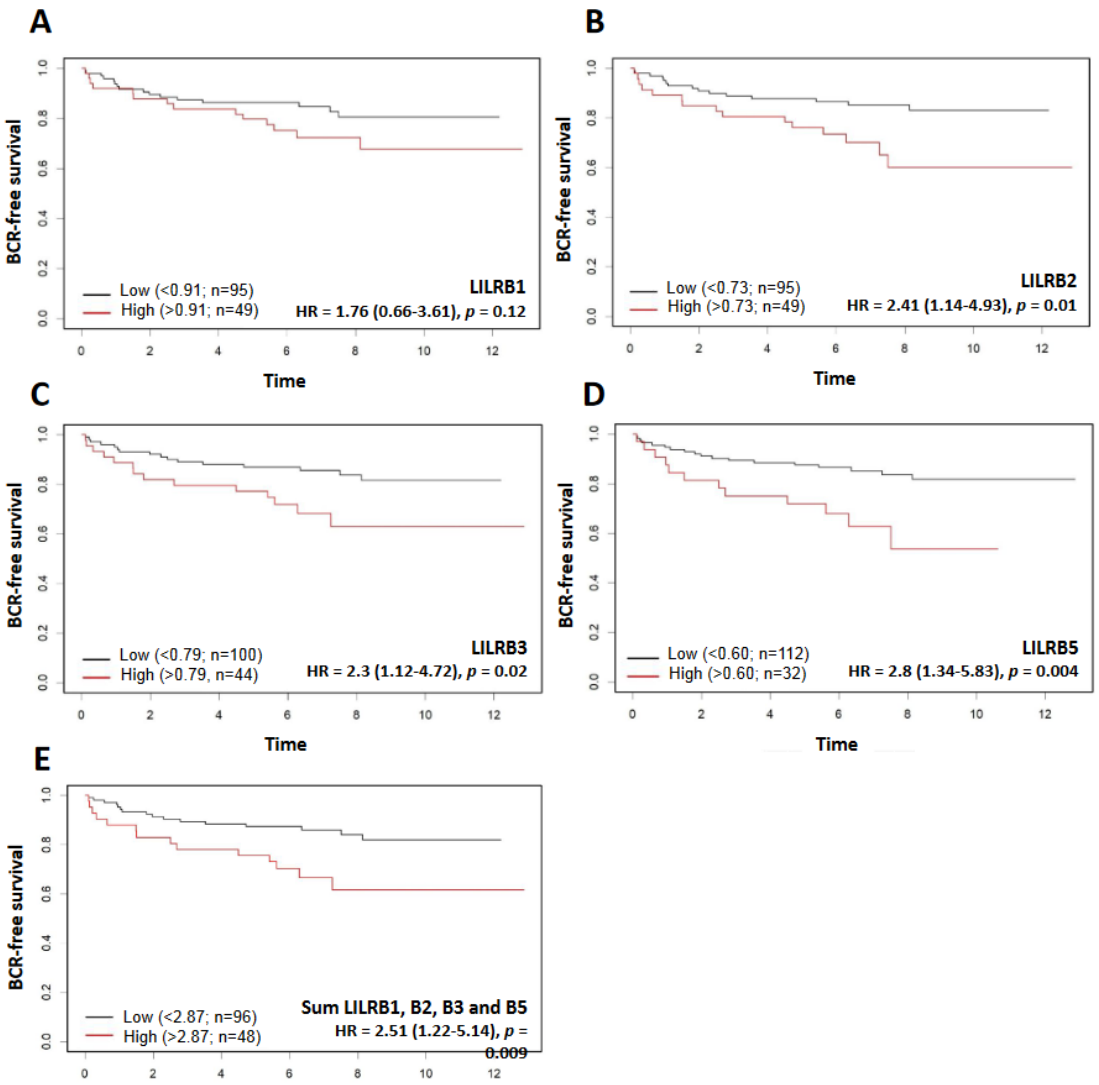


Figure 6: Kaplan-Meier analysis of dichotomized expression of LILRB1 (A), LILRB2 (B), LILRB3 (C), and LILRB5 (D) in the CPC-GENE RNA-seq dataset. High level expression of LILRB2, LILRB3, LILRB5 genes, but not that of LILRB1, was significantly associated with shorter BCR-free survival. The sum of the expression of these genes (E) was also associated with a significantly shorter BCR-free survival (HR=2.51, p=0,009). The association of LILRA3 gene expression with BCR could not be analyzed as there was no LILRA3 mRNA expression data in the CPC-GENE dataset.

The absence of statistically significant association of LILRB1 expression with BCR in the intermediate risk cohort suggests that LILRB1 gene expression could be associated with grade. Spearman correlation showed indeed that LILRB1 mRNA was correlated with grade in the TCGA-GSE54460-VPC cohort (r_s=0.53, p=0.01; Supplementary Table S6), while expression of LILRB2, LILRB3, LILRB5 and LILRA3 was not associated with grade in the combined

cohort. In order to further assess the association of LILRB1 with BCR, we analyzed the expression of LILRB1 protein in a series of 20 high-risk prostate tumors by immunohistochemistry. LILRB1 was found on immune cells scattered between tumor glands (Supplementary Figure S5 and Table S7). No tumor cells or stromal cells expressed the protein. In Kaplan-Meier analyses, a high level of $LILRB1^{+}$ cells infiltrating the tumor was found to be associated with poor clinical outcomes such as the need for definitive androgen deprivation therapy (Figure 7B; log-rank 0=0.009) and having lethal PCa defined as PCa that has already led to death or metastatic castration-resistant PCa that will eventually lead to death by PCa. All those results suggest that higher expression of LILRB1, LILRB2, LILRB3 and LILRB5 genes is associated with progression whereas a higher expression of LILRA3 would protect against progression of PCa.

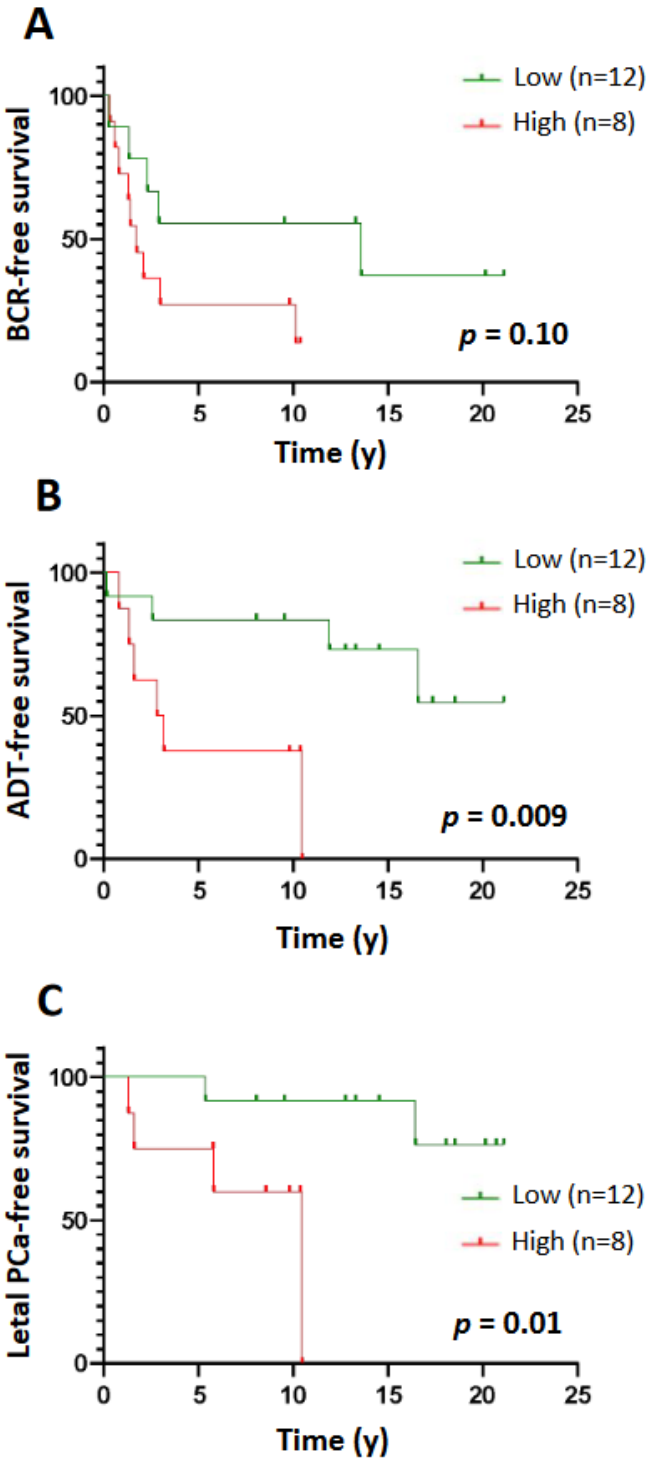


Figure 7: Kaplan-Meier curves of BCR-free (A), ADT-free (B) and letal PCa-free (C) survival according to high (level 3) vs low (levels 1-2) levels of LILRB1$^+$ cells in the tumor area of high-risk PCa samples. The number of LILRB1$^+$ cells in the tumor area was analyzed using immunohistochemistry on a series of 20 formalin-fixed and paraffin-embedded T2-T3 stage prostate tumor samples with long clinical follow-up. Expression of LILRB1 was classified as level 1, 2 and 3 corresponding to low, intermediate and high number of positive cells. Statistical significance was determined by the log-rank test.

2.6 Discussion

PCa immunotherapy using the first generation immune checkpoint inhibitors (i.e. against CTLA-4 and PD-1/PD-L1) has shown poor success in initial clinical trials. Some explanations for this might be the limited immunogenicity of PCa cells or immunosuppression mechanisms other than those involving these major immune checkpoints[20]. To explore the immunobiology of PCa, we performed a bioinformatic analysis using different types of omics data to identify immune-related features associated with the first event of PCa progression, i.e. the BCR consisting in an elevation of seric PSA after radical prostatectomy. Our hypothesis is that the identification of biological features selected in different types of omics would support their biological relevance and might provide candidate molecular targets for immunotherapeutic intervention.

To perform these multi-omics analyses involving large datasets we used for variable selection sPLS-DA, a multivariate exploratory approach, that provides more insight into cell biology, biological pathways or complex traits than other commonly used approaches such as machine learning approaches[21]. We first used this approach on the TCGA PRAD data within each type of omics to select features related to BCR and then we merged the selected features and applied the same approach on all the selected features. BCR was used as the clinical outcome of PCa progression instead of late outcomes associated with aggressive cancer such as detection of metastases or PCa-specific death. A major limitation of our approach is the size of the cohort available after the selection based on quality criteria. As reported, the TCGA PRAD cohort has a very short median follow-up which limits the correlation analyses between genomic features and clinical outcomes, especially the late ones. In order to increase the accuracy of the association between features and BCR we selected 5 years as a minimal follow-up knowing that it would considerably reduce the size of the final cohort.

The sPLS-DA led to the identification of an association between a series of genes involved in antigen presentation and regulation of immune cell activity and PCa progression. HLA antigens, KIR and LILR receptors form a complex system of molecules involved in the recognition of self/non-self antigens which can have an impact on various immunological responses and impact for example the outcome of viral infections, and diseases such as autoimmunity and cancer[11,22]. Among these, LILRs were those that were the most strongly associated with BCR. The association was further demonstrated in the combined RNA-seq datasets of 171 tumors using the MINT method which revealed an association between the combination of LILRB1, LILRB2, LILRB3, LILRB5 and LILRA3 genes and BCR. Kaplan-Meier analyses further confirmed this association and identified LILRB1 as the gene with the strongest association with BCR. However, validation in a cohort of intermediate risk PCa with very few high grade tumors showed no significant association of LILRB1 with BCR while the association was observed for LILRB2, LILRB3 and LILRB5. The absence of significant association of LILRB1 with BCR in this cohort is concordant with the association of LILRB1 with Gleason grade observed in the combined TCGA-GSE54460-VPCC cohort (supplementary Table S5). This positive relationship with grade might be explained by the fact that LIIRB1 is known to be predominantly expressed in macrophages and higher levels of M2 macrophages have been shown to be positively associated with Gleason grade and worst outcome[23,24]. Our immunohistochemistry analysis of 20 high-risk tumors also supported this as an association with adverse long term outcomes was observed. Multi-parametric analyses would be needed to confirm whether the immune cells expressing LILRB1 are indeed macrophages.

In cancer, LILRBs and especially LILRB1 immunosuppressive activity have been shown to contribute to cancer evasion. For example, Raji cells proliferation was proportionally inhibited by increasing amounts of HLA-G aggregated on nanoparticles and this inhibition was reversed when LILRB1 expression was inhibited using small interfering RNA or antagonistic mAb demonstrating that HLA-G inhibition is depending on LILRB1 expression[12]. The

immunosuppressive function of the HLA-G/LILRB1 signalling pathway has led to the identification of LILRB1 and HLA-G as new immune checkpoints that are potential targets for immunotherapy[25]. More recently the LILRB1/MHC-I signalling pathway was identified as a second “Don’t Eat Me” signal in tumor-associated macrophages[15]. Studies of the primary “Don’t Eat Me” signal, the CD47/SIRP-α signalling pathway, showed that inhibition of LILRB1 or MHC-I molecules potentiate the phagocytosis of tumor cells by macrophages in a manner that is independent of the inhibition of the CD47/SIRP-a axis[15]. Further analysis of the LILRB1/MHC-1 pathway may lead to the development of therapies to help restore macrophage function in the tumor microenvironment. Such therapies could complement the CD47/SIRP-α-based therapies that are already showing great potential in pre-clinical and early clinical studies[26,27] .

In conclusion, we performed a multi-omics analysis using PCa datasets that led us to identify a series of immune features that all together were strongly associated with BCR. Further analysis of these features allowed us to identify some candidate molecular targets that could be prioritized for immunotherapeutic intervention in PCa. Our data point toward a role for LILRB molecules and especially LILRB1 and suggest that these receptors could play a role in the resistance of PCa to anti-tumor immune response. Immunotherapeutic interventions aiming at the inhibition of the LILRB1/MHC-I pathway alone or in combination with therapies targeting complementary pathways (e.g. CD47/SIRP-α; PD-1/PD-L1 etc.) may provide a more adapted immunotherapeutic treatment to PCa immunobiology and would hopefully lead to better clinical response. Such an approach should be tested in pre-clinical models to further assess the immunotherapeutic potential of LILRB1 inhibition to stimulate immune response against PCa.

2.7 Material and methods

Patients and datasets

This study was approved by the Research Ethics Committee of the CHU de Québec-Université Laval (Project 2018-3670). Next generation sequencing (NGS) RNA-seq, miRNA, methylation, copy number variation (CNV), and single nucleotide variants (SNV) data from the TCGA PRAD project (498 samples) along with their associated clinical data were downloaded from the Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov/). The GSE54460 RNA-Seq and clinical data (106 samples) published by Long et al.[17] were downloaded from the Gene Expression Omnibus (GEO) website (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54460). The Vancouver Prostate Cancer Centre (VPCC) RNA-Seq and clinical data (85 samples) were provided by C. Collins (University of British Columbia, Vancouver, BC, Canada)[18]. A sub-cohort from the Canadian Prostate Cancer Genome Network (CPC-GENE) (n=144) was used as a validation cohort[19]. This cohort corresponds to men that were operated in our institution. Data were downloaded from https://ega-archive.org/. All patients had localized disease and were treated by radical prostatectomy. For each patient, available clinical data comprised at least the pathological characteristics of the tumor (grade and stage), PSA level at diagnosis, occurrence of BCR, time between radical prostatectomy and BCR, occurrence of death and date of death as well as the date of last follow-up.

Eligibility criteria

Eligibility was set by criteria of minimal quality. The TCGA PRAD project comprised 498 participants. However according to the TCGA Research Network, 131 participants must be omitted because of excessive RNA degradation[14]. The TCGA cohort is also characterized by a short follow-up. Patients with less than 60 months of follow-up were discarded. We also ignored tumors with less than 40% of tumor cell content. Patients treated with neoadjuvant or concomitant hormonal therapy were not conserved for the study. The same selection criteria were applied to GSE54460, VPCC and CPC-GENE

cohorts except for the selection based on the percentage of tumor cell content as this information was not provided.

OMIC data processing

RNA-Seq data

The RNA-Seq data were completely re-analyzed to avoid variability in data processing. The use of a common pipeline of analysis ensures accuracy of integrative analyses of transcriptomic datasets. The quality of the raw fastq files was assessed using FastQC[28] (v0.11.5) and trimmed with Trimmomatic[29] (v0.32). A threshold quality per base of 30 (based on Phred 33) and a minimal length of 40 bases were necessary otherwise the read was not conserved for analysis. The sequences were then mapped on GrCH38.p7 using Kallisto (v0.43.0)[30]. Kallisto was used to estimate isoform counts, adjusted for the amount of bias in the experiment to ensure a coherent non-naive mapping. Gene counts were computed with tximport[31]. The Ensembl Gene ID were converted with Biomart tools[32,33] from transcript ID to gene ID. The RNA-seq counts were then normalized to negative control genes (housekeeping genes) using the RUVg method[34,35]. In order to perform this normalization, we selected from the literature a series of six housekeeping genes that could be candidates for control reference genes in PCa experiments[36–39]. These genes were: RRN18S, ACTB, PPIA, GAPDH, PGK1, and GUSB. The expression of these genes was tested by RT-qPCR in a series of 50 prostate tumors and they were shown to be stably expressed between tumor samples (Data not shown). However, we excluded from the final list the ribosomal gene RRN18S because ribosomal RNAs were removed from our RNA-seq datasets. We also excluded PGK1 as it was shown that hypoxia in PCa alters the expression of this gene[36]. Therefore, we finally used GUSB, PPIA, GAPDH and ACTB as negative control genes for the normalization of the counts. The same process was applied to GSE554460 and VPCC RNA-seq datasets. The dataset corresponding to 144 tumors from the CPC-GENE cohort was used for validation. The mRNA

expression data were not processed as were the data from TCGA, GSE554460 and VPCC. Fastq files were downloaded and directly used for statistical analyses.

miRNA data

The level 3 NGS miRNA data from the TCGA PRAD project were provided as normalized counts in reads-per-million-miRNA-mapped. MiRNAs with a normalized count of zero or with no value were removed from the miRNA dataset. mirWalk 2.0[40], which rely on different databases (mirTarBase, mirDB and TargetS), was used to assign genes to miRNA according to predicted target genes.

Methylation data

Genome-wide methylation data from the TCGA PRAD project were generated using the Illumina Infinium HumanMethylation27 and HumanMethylation450 BeadChip platform (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Methylation_LO_Pipeline/). The level 3 methylation data were preprocessed and provided as Beta (ß) values. ß values (0 for unmethylated allele to 1 for fully unmethylated allele) are the estimate of methylation level using the ratio of intensities between methylated and unmethylated alleles. Genes with no ß values were removed from the dataset.

CNV data

The level 3 CNV data from the TCGA PRAD project were preprocessed using Birdsuite[23] from the Broad Institute and the R package DNAcopy[41]. The data were cleaned, normalized, segmented and log transformed. From these data we created an index with all unique regions found. These regions were annotated with ChipSeeker[42] which associates the closest genomic object to the region's coordinates.

SNV data

The level 3 SNV data from the TCGA were retrieved and cleaned with the VCFtools suite[43]. An index with the mutation found in all patients at the base level was created and all unique mutations were kept as final variables. For each mutation we kept the information of location (Chromosome and coordinates) and mutation type (e.g. from pattern x to pattern y).

Set of immune genes

The multivariate analyses were focused on immune genes. The set of selected immune genes for these analyses is composed of 812 Ensembl genes. This set of genes was derived from a meta-analysis which targeted the anti-genome of tumor cells in interaction with the immune system[44].

Statistical analyses

To identify features associated with BCR in every set of omics, sparse partial least square-discriminant analysis (sPLS-DA) models were calculated using the mixOmics package[21]. In each sPLS-DA analysis, the BCR was defined as the Y response. The Mfold validation strategy with a fold of 5 and 200 repetitions was used to ensure the stability of the model. To assess the predictive potential of mixed omics data, the selected omics features were merged and again sPLS-DA were calculated as above. The association between selected features and BCR was further analyzed by multivariate regression analysis in three RNA-seq datasets (i.e. TCGA PRAD, GSE54460 and VPCC) using the Multivariate INTegrative (MINT)[45] method from the mixOmics package. Again, the Mfold validation strategy with a fold of 5 and 200 repetitions was used as in the sPLS-DA analyses. Within all R analyses, the seed parameter was defined as 2543 for reproducibility. Code used for the PLS-DA model can be found here http://mixomics.org/methods/pls-da/ and the

one for the MINT model can be found here http://mixomics.org/mixmint/stemcells-example/.

The balanced error rate (BER= 1-0.5*(sensitivity + specificity)) measured at the centroid distance was used in sPLS-DA and MINT analyses to assess the quality of the association between the BCR and the omic features. A BER of 0 means a perfect classification while a BER over 0.5 means no association with the response variable. We considered a BER<0.4 as a good score value.

To perform Kaplan-Meier survival analysis, the expression data for each mRNA of interest were optimally dichotomized using the Cutoff finder tool[46]. The method fitting Cox proportional hazard models to dichotomize variable was used to define a threshold expression value. Then the survival and survminer packages were used to perform the survival analysis within R.

Immunohistochemistry

Analysis of LILRB1 expression was performed on prostate tumors obtained from our local biobank URO-1. This analysis was approved by the Research Ethics Committee of the CHU de Québec-Université Laval (Project 2012-1059). Briefly, five µm-thick sections of formalin-fixed and paraffin-embedded tumors were deparaffinized and submitted to heat-induced antigen retrieval (97°C, 20 min) in Tris/EDTA, pH 9 (Dako Code K8004: EnVision™ FLEX, High pH buffer) using a PT Link, Pre-Treatment Module for Tissue Specimens (Dako, Burlington, ON, Canada). Endogenous peroxidases were blocked by incubation in 3% peroxide solution for 10 min. Bound antibodies were revealed using the IDetect super stain HRP polymer kit (ID labs, London, Ontario, Canada) as follows. Slides were initially incubated for 10 minutes at room temperature with Super block solution to avoid nonspecific background staining. Then, slides were with anti-LILRB1 rabbit monoclonal antibody (mAb)(clone EPR21007, dilution 1:500, Abcam, Toronto, ON) during 1 hour at room temperature. After washes, slides were incubated for 30 min with HRP-Polymer Conjugate according to manufacturer's recommendations. Final revelation was performed by a 5 min of incubation with DAB. Finally slides were

rinsed, counterstained with hematoxylin, dehydrated and mounted with coverslip using MM 24 low viscosity mounting medium (Leica Microsystems, Durham, USA). Slides were digitalized using a Nanozoomer (Hamamatsu Photonics, Bidgewater NJ, USA) and visualized using the NDP.view2 software (Hamamatsu Photonics). Scoring of the relative number of positive cells in the tumor area was performed by a trained technician and was reported on a scale from 1 to 3.

ACKNOWLEDGEMENT

The authors would like to thank Fan Mo and Antonio Hurtado-Coll for the preparation of the data from VPCC.

Funding

This research was realized with internal funds from the Laboratoire d'Uro-Oncologie Expérimentale. The production of RNA-seq data at VPCC was realized with funds from the Terry Fox Research Institute New Frontier Program Project Grant #1062.

Disclosure statement

No potential conflicts of interest were disclosed.

Author contribution

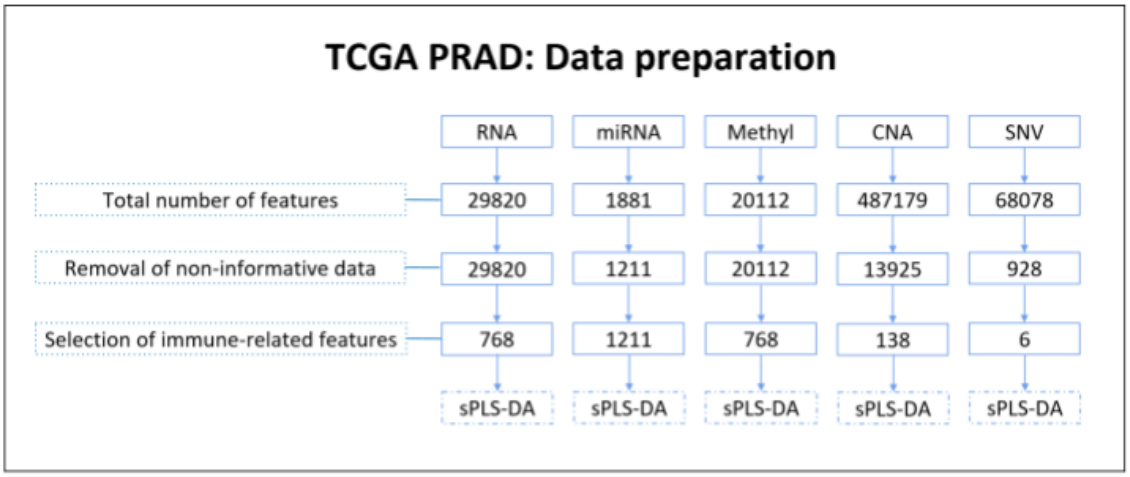
BV, YF, AB, ML and AD conceived the study. YF, AB and AD supervised the study. BV performed all bioinformatics and statistical analyses. JL and PB performed some bioinformatic analyses. OEM performed the analysis of housekeeping gene expression. OEM, XPL, HH and VP were involved in the immunohistochemistry analyses. AB, ML, MLMM, PB, CC, YF and AD interpreted data and provided advice. BV, AB and ML wrote the manuscript. All authors critically reviewed the manuscript.

.

SUPPLEMENTARY MATERIAL

Supplementary Table S1. Baseline characteristics of the selected TCGA sub-cohort.

Clinico-Pathological Characteristics	TCGA n (%)
N	45 (100)
Grade	
6	2 (4)
7	12 (27)
8	9 (20)
9	22 (49)
10	0 (0)
Stage	
T1c	0 (0)
T2	0 (0)
T2a	1 (2)
T2b	2 (5)
T2c	8 (18)
T3	0 (0)
T3a	15 (33)
T3b	19 (42)
T4	0 (0)
Pre-Op PSA	
≤10	26 (58)
10-20	15 (33)
≥20	4 (9)
BCR	
Yes	31 (69)
No	14 (31)

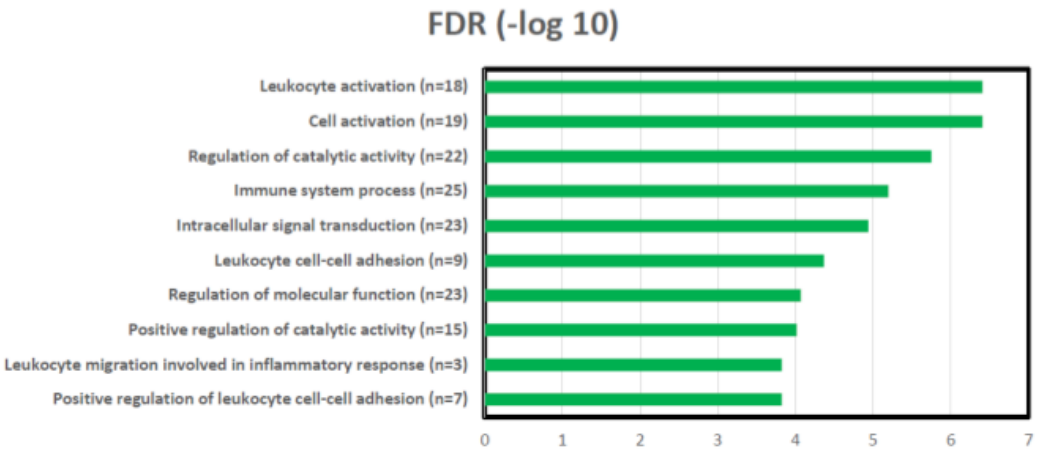


Supplementary Figure S2: Data preparation. For each of the 45 selected participants, mRNA, miRNA, CNA, methylation, and SNV datasets were recovered and curated. Non-informative data in each datasets were discarded. RNA-seq data were completely reanalyzed from raw data. We next applied a filter to select for 812 genes associated with immunity. After this preparation, the data were ready for the sparse partial least square-discriminant analyses (sPLS-DA).

Supplementary Table S2. Results of sPLS-DA analysis in each dataset.

mRNA	miRNA	Methylation	CNA	SNV
ADA2	hsa-mir-485	IARS	**LILRA4**	**chr19_54574711_A_G_intron_variant_LILRA2**
ADAM8	hsa-mir-4728	PTRH2	NXN	**chr19_54274966_C_T_intron_variant_LILRB2**
ADORA3	hsa-mir-450a-1	KRT80	CDH2	**chr19_54251083_A_G_transcript_variant_LILRB5**
BMI1	hsa-mir-664a	SERTAD2	EFNA5	**chr19_54786359_C_T_gene_variant_KIR2DL1**
BRIP1	hsa-mir-26b	ECT2	PLA2G4A	**chr19_54775213_T_A_missense_variant_KIR2DL1**
BTN2A2	hsa-mir-141	PRC1	**HLA-DQA1**	**chr19_54832736_C_T_gene_variant_KIR3DL1**
BTN3A2	hsa-mir-6736	**HLA-DMB**	MAD2L1	
C1orf162	hsa-mir-581	SIGLEC5	KRT1	
CBX1	hsa-mir-505	TBX21	FCRL5	
CCDC65	hsa-mir-6511b-2	IL17RC	RBMS3	
CCL14	hsa-mir-582	GRP	PYHIN1	
CCNA2	**hsa-mir-5702**	EIF2S1	KLRD1	
CCNB1	hsa-mir-125a	CTR9	MYC	
CCT6B	hsa-mir-1976	KLRK1	IFI16	
CEACAM3	hsa-mir-320a	CCL4	CD53	
CEP55	hsa-mir-377	TACSTD2	EPHA4	
CKAP2L	hsa-mir-152	CD48	DAB2	
CTSG	hsa-mir-219a-1	MMD	RUNX3	
DARS	hsa-mir-362	MTHFD2	ETS1	
DYNLL1	hsa-mir-5588	ADAMTS3	STAP1	
EGR3	hsa-mir-6715b	TNFSF8	**HLA-DRB1**	
EXO1	hsa-mir-4676	SDCBP	**HLA-DRB5**	
FBXO5	hsa-mir-187	NCOA4	SLC36A1	
FLT3	hsa-mir-204	DAPP1	IPCEF1	
GADD45B	hsa-mir-145	FGFBP2	**LILRA5**	
GIMAP7	hsa-mir-3202-2	GZMH	TRAF3IP3	
GPR84	hsa-mir-574	JAK2	SKAP1	
HELLS	hsa-mir-504	SETD7	C3AR1	
ICAM1	hsa-mir-605	TLR5	**LILRA3**	
IGF1	**hsa-mir-6087**	FURIN	**HLA-DMB**	
IKZF1	hsa-mir-4521	SERPINB2	JAK2	
IL3RA	hsa-mir-1301	CREB3	AHNAK	
ITGB4	hsa-mir-487a	TSLP		
KIF20A	hsa-mir-186	PDCD1		
KNTC1	hsa-mir-29b-2	KLF5		
KRIT1	hsa-mir-339	**KIR3DL2**		
LILRA2	hsa-mir-423			
LILRA5	hsa-mir-506			
MRC1	**hsa-mir-5694**			
NEIL3	hsa-mir-3653			
NFATC1	**hsa-mir-1248**			
NLRP3	hsa-mir-197			
PCNA	hsa-mir-450a-2			
PHRF1	hsa-mir-342			
S100A8				
SELE				
SLA				
SPOCK2				
TMED2				
TNF				
TRIB2				

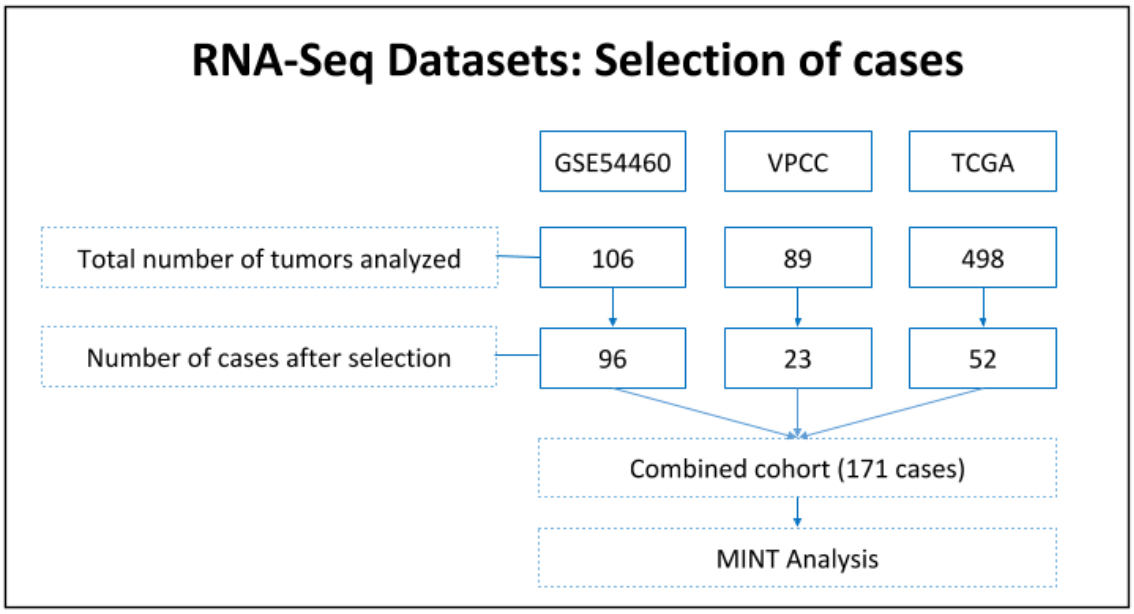
* In bold and underlined are the features associated with HLA, LILR and KIR gene families, all implicated in antigen presentation .



Supplementary Figure S3. Gene ontology analysis of features identified with the sPLS-DA of mRNA. The bar chart shows the top 10 GO terms ranked by fold enrichment (FDR; -log 10). The number of genes out of 51 involved in each pathway is indicated between parentheses.

Supplementary Table S3. List of 30 LILR-gene associated features used in sPLS-DA.

mRNA	miRNA	Methylation	CNA	SNV
rna_LILRA1	mir-1248_LILRA4	met_LILRA1	cna_LILRA3	chr19_54574711_A_G_intron_variant_LILRA2
rna_LILRA2	mir-5694_LILRA2	met_LILRA2	cna_LILRA4	chr19_54274966_C_T_intron_variant_LILRB2
rna_LILRA3	mir-5702_LILRB3	met_LILRA3	cna_LILRA5	chr19_54251083_A_G_transcript_variant_LILRB5
rna_LILRA4	mir-6087_LILRA4	met_LILRA4		
rna_LILRA5		met_LILRA5		
rna_LILRB1		met_LILRB1		
rna_LILRB2		met_LILRB2		
rna_LILRB3		met_LILRB3		
rna_LILRB4		met_LILRB4		
rna_LILRB5		met_LILRB5		



Supplementary Figure S4: Selection of cases from RNA-seq datsets. To validate the predictive value of LILR genes to predict BCR, the genes were analyzed in a series of three RNA-seq datasets. The same eligibility criteria and data preparation pipeline used to prepare RNA-seq data of TCGA was applied to the GSE54460 and the VPCC datasets. This resulted in the elimination of several cases of the VPCC and some cases of the GSE54460 datasets. The resulting data from 171 participants were combined and analyzed using the MINT approach.

Supplementary Table S4. Baseline characteristics of the individual cohorts and the combined cohort used for the MINT analysis.

Clinico-Pathological Characteristics	TCGA n (%)	GSE54460 n (%)	VPCC n (%)	Combined n (%)
N	52 (100)	96 (100)	23 (100)	171 (100)
Grade				
5	0 (0)	1 (1)	3 (13)	4 (2)
6	2 (4)	9 (9)	12 (52)	23 (13)
7	14 (27)	72 (75)	4 (17)	90 (53)
8	9 (17)	9 (9)	1 (4)	19 (11)
9	27 (52)	5 (5)	2 (8)	34 (20)
10	0 (0)	0 (0)	1 (4)	1 (1)
Stage				
T1c	0 (0)	14 (15)	0 (0)	14 (8)
T2	0 (0)	7 (7)	0 (0)	7 (4)
T2a	1 (2)	21 (22)	3 (13.)	25 (15)
T2b	2 (4)	10 (10)	0 (0)	12 (7)
T2c	9 (17)	26 (27)	17 (74)	52 (30)
T3	0 (0)	2 (2)	0 (0)	2 (1)
T3a	16 (31)	5 (5)	2 (9)	23 (13)
T3b	24 (46)	9 (9)	1 (4)	34 (20)
T4	0 (0)	1 (1)	0 (0)	1 (1)
NA	0 (0)	1 (1)	0 (0)	1 (1)
Pre-Op PSA				
<=10	31 (59)	64 (67)	21 (92)	116 (68)
10-20	16 (31)	17 (18)	1 (4)	35 (20)
>=20	5 (10)	12 (12)	1 (4)	18 (10)
NA	0 (0)	3 (3)	0 (0)	3 (2)
BCR				
Yes	38 (73)	42 (44)	18 (78)	98 (57)
No	14 (27)	54 (56)	5 (22)	73 (43)

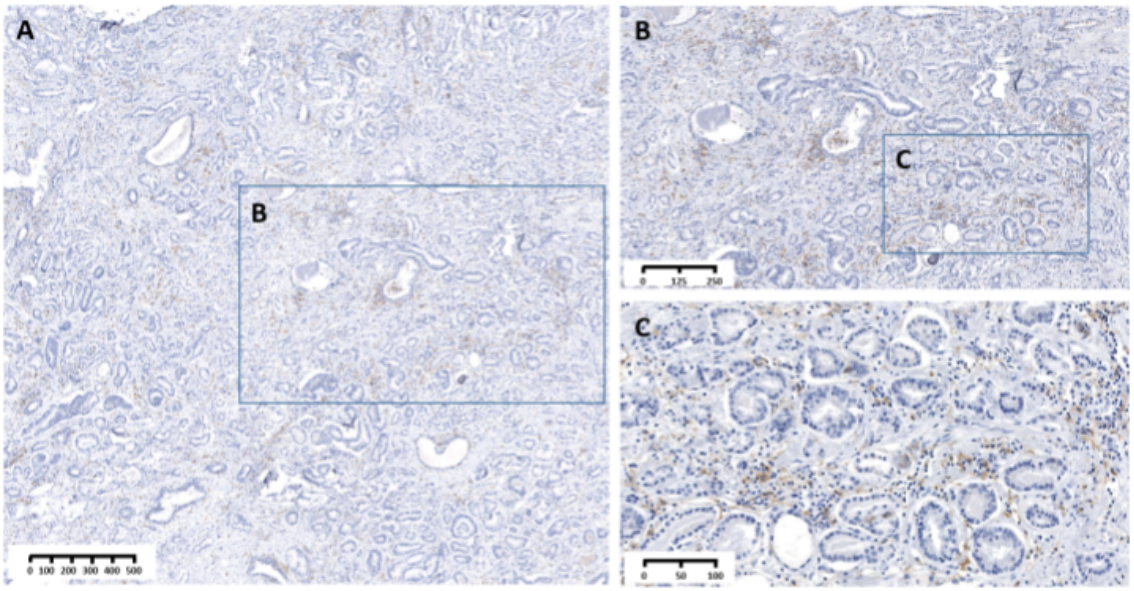
Supplementary Table S5. Baseline characteristics of the selected CPC-gene cohort.

Clinico-Pathological Characteristics	CPC-Gene n (%)
N	144
Grade	
5	
6	6 (4)
7	123 (85)
8	6 (4)
9	3 (2)
10	0 (0)
Stage	
T1a	1 (<1)
T1b	1 (<1)
T1C	81 (56)
T2	0 (0)
T2a	37 (26)
T2b	21 (15)
T2c	3 (2)
T3	0 (0)
T3a	0 (0)
T3b	0 (0)
T4	0 (0)
Pre-Op PSA	
<=10	110 (76)
10-20	33 (23)
>=20	1 (<1)
BCR	
Yes	30 (21)
No	114 (79)

Supplementary Table S6. Spearman correlation between LILR gene expression and tumor Grade in each datatset as well as in the TCGA/GSE54460/VPCC combined datatset.

Gene	ENSG #	Dataset	Spearman*	*p* value
LILRB1	ENSG00000104972	TCGA	**0.77**	<0.001
		GSE54460	0.45	0.09
		VPCC	0.45	0.09
		Combined	**0.53**	**0.01**
LILRB2	ENSG00000131042	TCGA	0.21	0.40
		GSE54460	0.10	0.72
		VPCC	0.11	0.71
		Combined	0.13	0.39
LILRB3	ENSG00000204577	TCGA	0.13	0.45
		GSE54460	-0.06	0.75
		VPCC	**0.54**	**0.03**
		Combined	0.13	0.24
LILRB5	ENSG00000105609	TCGA	-0.02	0.80
		GSE54460	0.04	0.07
		VPCC	-0.06	0.02
		Combined	-0.07	0.39
LILRA3	ENSG00000275841	TCGA	-0.80	0.88
		GSE54460	NA	0.74
		VPCC	-0.31	0.26
		Combined	-0.22	0.35

*Correlation coefficients higher than 0.5 or lower than -0.5 and with significant *p* value are in bold characters and underlined.



Supplementary Figure S5: Immunohistochemical analysis of LILRB1 expression in prostate cancer. Example of a high level of LILRB1^{+} cells within the core of the tumor. Signal is found in immune cells interspersed between the tumor glands. Three different magnifications are shown. Area shown in panel B is delineated by a bleu rectangle in panel A while area shown in panel C is delineated by a bleu rectangle in panel B. Scale bars are presented.

Supplemmentary Table S7. Clinico-pathological characteristics of the 20 tumors analyzed by immunohistochemistry for LILRB1 expression and the relative quantity of LILRB1 positive cells (levels 1 to 3) observed are provided.

Sample #	Gleason #1	Gleason #2	Gleason Sum	T Stage	N Stage	Positive Margin	RP* failure	BCR*	Definitive ADT*	CRPC*	MET*	Death	Level of LILRB1**
A	3	5	8	T3b	N2	Yes	No	Yes	Yes	Yes	Yes	Yes	1
B	4	3	7	T3a	N0	Yes	No	No	No	No	No	No	3
C	3	3	6	T2c	N0	Yes	No	No	No	No	No	No	1
D	3	4	7	T2c	N0	Yes	No	Yes	Yes	No	No	No	1
E	.	.	7	T2c	N0	Yes	No	Yes	No	No	No	No	2
F	5	4	9	T3b	N2	Yes	Yes	No	Yes	Yes	Yes	Yes	3
G	3	4	7	T3b	N1	Yes	No	Yes	Yes	Yes	Yes	Yes	3
H	4	3	7	T3b	N0	No	No	Yes	Yes	Yes	Yes	No	1
I	4	4	8	T3b	N2	Yes	No	Yes	Yes	No	No	Yes	3
J	4	3	7	T3b	N1	No	Yes	Yes	Yes	No	No	Yes	3
K	4	3	7	T2c	N0	Yes	No	Yes	No	No	No	Yes	2
L	4	4	8	T2a	N0	Yes	No	No	No	No	No	Yes	1
M	3	3	6	T2c	N0	Yes	No	Yes	No	No	No	Yes	1
N	4	5	9	T3b	N0	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3
O	3	4	7	T3a	N0	Yes	No	Yes	No	No	No	Yes	2
P	4	5	9	T3b	N1	Yes	No	No	Yes	No	No	No	1
Q	4	4	8	T3b	N0	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3
R	3	4	7	T3a	N2	Yes	No	No	No	No	No	No	1
S	3	4	7	T4	N0	Yes	Yes	No	No	No	No	No	1
T	4	3	7	T3a	N2	No	No	No	No	No	No	No	3

*RP: Radical prostatectomy; BCR: Biochemical recurrence; ADT: Androgen deprivation therapy; CRPC: Castration-resistant prostate cancer; MET: Metastases

** 1: Low; 2: Intermediate; 3: High

2.8 Bibliographie

1. Comiskey MC, Dallos MC, Drake CG. Immunotherapy in Prostate

Cancer: Teaching an Old Dog New Tricks. *Curr Oncol Rep*. 2018;20(9):75. doi:10.1007/s11912-018-0712-z

2. Bilusic M, Madan RA, Gulley JL. Immunotherapy of Prostate Cancer: Facts and Hopes. *Clin Cancer Res Off J Am Assoc Cancer Res*. 2017;23(22):6764-6770. doi:10.1158/1078-0432.CCR-17-0019

3. Scheid E, Major P, Bergeron A, et al. Tn-MUC1 DC Vaccination of Rhesus Macaques and a Phase I/II Trial in Patients with Nonmetastatic Castrate-Resistant Prostate Cancer. *Cancer Immunol Res*. 2016;4(10):881-892. doi:10.1158/2326-6066.CIR-15-0189

4. Wei SC, Duffy CR, Allison JP. Fundamental Mechanisms of Immune Checkpoint Blockade Therapy. *Cancer Discov*. 2018;8(9):1069-1086. doi:10.1158/2159-8290.CD-18-0367

5. Dempke WCM, Fenchel K, Uciechowski P, Dale SP. Second- and third-generation drugs for immuno-oncology treatment-The more the better? *Eur J Cancer Oxf Engl 1990*. 2017;74:55-72. doi:10.1016/j.ejca.2017.01.001

6. Marin-Acevedo JA, Dholaria B, Soyano AE, Knutson KL, Chumsri S, Lou Y. Next generation of immune checkpoint therapy in cancer: new developments and challenges. *J Hematol OncolJ Hematol Oncol*. 2018;11(1):39. doi:10.1186/s13045-018-0582-8

7. Longo V, Brunetti O, Azzariti A, et al. Strategies to Improve Cancer Immune Checkpoint Inhibitors Efficacy, Other Than Abscopal Effect: A Systematic Review. *Cancers*. 2019;11(4). doi:10.3390/cancers11040539

8. Kwon ED, Drake CG, Scher HI, et al. Ipilimumab versus placebo after radiotherapy in patients with metastatic castration-resistant prostate cancer that had progressed after docetaxel chemotherapy (CA184-043): a multicentre, randomised, double-blind, phase 3 trial. *Lancet Oncol*. 2014;15(7):700-712. doi:10.1016/S1470-2045(14)70189-5

9. Beer TM, Kwon ED, Drake CG, et al. Randomized, Double-Blind, Phase III Trial of Ipilimumab Versus Placebo in Asymptomatic or Minimally Symptomatic Patients With Metastatic Chemotherapy-Naive Castration-Resistant Prostate Cancer. *J Clin Oncol Off J Am Soc Clin Oncol*. 2017;35(1):40-47. doi:10.1200/JCO.2016.69.1584

10. Jafari S, Molavi O, Kahroba H, et al. Clinical application of immune checkpoints in targeted immunotherapy of prostate cancer. *Cell Mol Life Sci CMLS*. Published online January 31, 2020. doi:10.1007/s00018-020-03459-1

11. Hudson LE, Allen RL. Leukocyte Ig-Like Receptors - A Model for MHC Class I Disease Associations. *Front Immunol*. 2016;7:281. doi:10.3389/fimmu.2016.00281

12. Naji A, Menier C, Maki G, Carosella ED, Rouas-Freiss N. Neoplastic B-cell growth is impaired by HLA-G/ILT2 interaction. *Leukemia*. 2012;26(8):1889-1892. doi:10.1038/leu.2012.62

13. Brown D, Trowsdale J, Allen R. The LILR family: modulators of innate and adaptive immune pathways in health and disease. *Tissue Antigens*. 2004;64(3):215-225. doi:10.1111/j.0001-2815.2004.00290.x

14. Abeshouse A, Ahn J, Akbani R, et al. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*. 2015;163(4):1011-1025. doi:10.1016/j.cell.2015.10.025

15. Barkal AA, Weiskopf K, Kao KS, et al. Engagement of MHC class I by the inhibitory receptor LILRB1 suppresses macrophages and is a target of cancer immunotherapy. *Nat Immunol*. 2018;19(1):76-84. doi:10.1038/s41590-017-0004-z

16. Zhao J, Zhong S, Niu X, Jiang J, Zhang R, Li Q. The MHC class I-LILRB1 signalling axis as a promising target in cancer therapy. *Scand J Immunol*. 2019;90(5):e12804. doi:10.1111/sji.12804

17. Long Q, Xu J, Osunkoya AO, et al. Global Transcriptome Analysis of Formalin-Fixed Prostate Cancer Specimens Identifies Biomarkers of Disease Recurrence. *Cancer Res*. 2014;74(12):3228-3237. doi:10.1158/0008-5472.CAN-13-2699

18. Wyatt AW, Mo F, Wang K, et al. Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer. *Genome Biol*. 2014;15(8):426. doi:10.1186/s13059-014-0426-y

19. Fraser M, Sabelnykova VY, Yamaguchi TN, et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature*. 2017;541(7637):359-364. doi:10.1038/nature20788

20. Bryant G, Wang L, Mulholland DJ. Overcoming Oncogenic Mediated Tumor Immunity in Prostate Cancer. *Int J Mol Sci*. 2017;18(7). doi:10.3390/ijms18071542

21. Rohart F, Gautier B, Singh A, Cao K-AL. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput Biol*. 2017;13(11):e1005752. doi:10.1371/journal.pcbi.1005752

22. Ivarsson MA, Michaëlsson J, Fauriat C. Activating killer cell Ig-like

receptors in health and disease. *Front Immunol*. 2014;5:184. doi:10.3389/fimmu.2014.00184

23. Shimura S, Yang G, Ebara S, Wheeler TM, Frolov A, Thompson TC. Reduced infiltration of tumor-associated macrophages in human prostate cancer: association with cancer progression. *Cancer Res*. 2000;60(20):5857-5861.

24. Erlandsson A, Carlsson J, Lundholm M, et al. M2 macrophages and regulatory T cells in lethal prostate cancer. *The Prostate*. 2019;79(4):363-369. doi:10.1002/pros.23742

25. Carosella ED, Ploussard G, LeMaoult J, Desgrandchamps F. A Systematic Review of Immunotherapy in Urologic Cancer: Evolving Roles for Targeting of CTLA-4, PD-1/PD-L1, and HLA-G. *Eur Urol*. 2015;68(2):267-279. doi:10.1016/j.eururo.2015.02.032

26. Hayat SMG, Bianconi V, Pirro M, Jaafari MR, Hatamipour M, Sahebkar A. CD47: role in the immune system and application to cancer therapy. *Cell Oncol Dordr*. Published online August 14, 2019. doi:10.1007/s13402-019-00469-5

27. Sikic BI, Lakhani N, Patnaik A, et al. First-in-Human, First-in-Class Phase I Trial of the Anti-CD47 Antibody Hu5F9-G4 in Patients With Advanced Cancers. *J Clin Oncol Off J Am Soc Clin Oncol*. 2019;37(12):946-953. doi:10.1200/JCO.18.02018

28. Simon A. *FastQC : A Quality Control Tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

30. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519

31. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*. 2015;4. Accessed May 11, 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4712774/

32. Smedley D, Haider S, Durinck S, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res*. 2015;43(W1):W589-W598. doi:10.1093/nar/gkv350

33. Kinsella RJ, Kähäri A, Haider S, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*. 2011;2011. doi:10.1093/database/bar030

34. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*. 2014;32(9):896-902. doi:10.1038/nbt.2931

35. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*. 2012;13(3):539-552. doi:10.1093/biostatistics/kxr034

36. Vajda A, Marignol L, Barrett C, et al. Gene expression analysis in prostate cancer: The importance of the endogenous control. *The Prostate*. 2013;73(4):382-390. doi:10.1002/pros.22578

37. Chua SL, See Too WC, Khoo BY, Few LL. UBC and YWHAZ as suitable reference genes for accurate normalisation of gene expression using MCF7, HCT116 and HepG2 cell lines. *Cytotechnology*. 2011;63(6):645-654. doi:10.1007/s10616-011-9383-4

38. de Kok JB, Roelofs RW, Giesendorf BA, et al. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Lab Invest*. 2005;85(1):154-159. doi:10.1038/labinvest.3700208

39. Ohl F, Jung M, Xu C, et al. Gene expression studies in prostate cancer tissue: which reference gene should be selected for normalization? *J Mol Med*. 2005;83(12):1014-1024. doi:10.1007/s00109-005-0703-z

40. Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat Methods*. 2015;12(8):697-697. doi:10.1038/nmeth.3485

41. Seshan VE, Olshen A. *DNAcopy: DNA Copy Number Data Analysis*. Bioconductor version: Release (3.10); 2020. doi:10.18129/B9.bioc.DNAcopy

42. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*. 2015;31(14):2382-2383. doi:10.1093/bioinformatics/btv145

43. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330

44. Angelova M, Charoentong P, Hackl H, et al. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol*. 2015;16(1). doi:10.1186/s13059-015-0620-6

45. Rohart F, Eslami A, Matigian N, Bougeard S, Lê Cao K-A. MINT: a

multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics*. 2017;18(1). doi:10.1186/s12859-017-1553-8

46. Budczies J, Klauschen F, Sinn BV, et al. Cutoff Finder: A Comprehensive and Straightforward Web Application Enabling Rapid Biomarker Cutoff Optimization. van Diest P, ed. *PLoS ONE*. 2012;7(12):e51862. doi:10.1371/journal.pone.0051862

Chapitre 2 : Article 2

3.1 Résumé

Déterminer quel traitement donner aux hommes avec un cancer de la prostate (CaP) est un défi pour les médecins. Actuellement la classification clinique du cancer est basée sur des critères histo-pathologiques comme les score de Gleason (grade), le stade (TNM) et la mesure du niveau de l'antigène spécifique de la prostate (PSA). Mais les données transcriptomiques ont le potentiel de permettre le développement de méthodes pouvant prédire avec précision l'évolution de la maladie. Cependant il existe peu de jeux de données RNA-seq de bonne qualité ou répondant à des critères cliniques d'intérêt, notamment un long suivi des patients, ce qui rend difficile la mise en place d'algorithmes dédiés. Dans cette étude nous proposons une approche d'apprentissage machine à partir de jeux de données RNA-Seq de sources différentes. Nous montrons qu'il est possible d'obtenir des résultats prédictifs de la BCR performants tout en utilisant des petits jeux de données d'un seul bloc. Cela montre donc qu'il est possible de retravailler et regrouper tous les travaux d'équipes et de sources différentes pour mettre en place des algorithmes prédictifs à moindre coût.

3.2 Titres et auteurs

Identification of a transcriptomic prognostic signature by machine learning using a combination of small cohorts of prostate cancer

Benjamin Vittrant [1,2], Mickael Leclercq [1,2], Marie Laure Martin-Magniette [3, 4], Colin Collins [5,6], Alain Bergeron [1,7], Yves Fradet [1,7] and Arnaud Droit [1,2,*]

[1] *Centre de Recherche du CHU de Québec – Université Laval, Québec, Québec, Canada*

[2] *Département de Médecine Moléculaire, Université Laval, Québec, Canada*

[3] *Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay, Paris Diderot, Sorbonne Paris-Cité, Bâtiment 630, 91405 Orsay, France.*

[4] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*

[5] *Vancouver Prostate Cancer Centre, Vancouver, British Columbia, Canada*

[6] *University of British Columbia, Dept. of Urologic Sciences, Vancouver, British Columbia, Canada*

[7] *Département de Chirurgie, Oncology axis, Université Laval, Québec, Canada*

* To whom correspondence should be addressed: arnaud.Droit@crchudequebec.ulaval.ca

Correspondence may also be addressed to mickael.Leclercq@gmail.com

Present Address: Arnaud Droit, Ph.D., Computational Biology Laboratory, Centre de Recherche du CHU de Québec – Université Laval, 2705 boul. Laurier, R-4773, Québec (QC), G1V 4G2, CANADA.

3.3 Abstract

Background: Determining which treatment to provide to men with prostate cancer (PCa) is a major challenge for clinicians. Currently, the clinical risk-stratification for PCa is based on clinico-pathological variables such as Gleason grade, stage and prostate-specific antigen (PSA) levels. But transcriptomic data have the potential to enable the development of more precise approaches to predict evolution of the disease. However, high quality RNA sequencing (RNA-seq) datasets along with clinical data with long follow-up allowing discovery of biochemical recurrence (BCR) biomarkers are small and rare. In this study, we propose a machine learning approach that is robust to batch effect and enables the discovery of highly predictive signatures despite using small datasets.

Methods: Gene expression data were extracted from three RNA-Seq datasets cumulating a total of 171 PCa patients. Data were reanalyzed using a unique pipeline to ensure uniformity. Using a machine learning approach, a total

of 14 classifiers were tested with various parameters to identify the best model and gene signature to predict BCR.

Results: Using a random forest model, we have identified a signature composed of only three genes (JUN, HES4, PPDPF) predicting BCR with better accuracy (74.2%, BER=27%) than the clinico-pathological variables (69.2%, BER=32%) currently in use to predict PCa evolution. This score is in the range of the studies that predicted BCR in single-cohort with a higher number of patients.

Conclusions: We showed that it is possible to merge and analyze different small and heterogeneous datasets altogether to obtain a better signature than if they were analyzed individually, thus reducing the need for very large cohorts. This study demonstrates the feasibility to regroup different small datasets in one larger to identify a predictive genomic signature that would benefit PCa patients.

3.4 Introduction

Prostate Cancer (PCa) is the most common non-cutaneous cancer in American men. Around 160 000 men were diagnosed with PCa in 2017 (1) and around 27 000 died of it. The burden of this disease on public health is important and expected to grow as a recent study revealed that the incidence of advanced PCa increased in the last few years (2). PCa is a complex and heterogeneous disease (3,4) since the risk of relapse and death after treatment differs among cancers with the same clinico-pathological features, namely the grade (Gleason score), stage (TNM; Tumor, Node, Metastasis) (5,6) and the level of prostatic specific antigen (PSA) (7).

Current treatments for localized PCa mainly include surgical removal or external beam radiation therapy of the prostate. If the initial treatments did not succeed to cure the patient then a recurrence will occur, revealed by an increase in seric PSA level, an event called biochemical recurrence (BCR). After surgery, about 70% of the patients will be cured and about 30% will

relapse to a BCR. Since prostate tumor cells depend on androgens to grow, recurrences are treated with androgen deprivation therapy consisting in chemical or surgical castration either alone or in association with administration of anti-androgens. However, the cancer will inevitably recur and will then be called castration-resistant prostate cancer (CRPC). To treat CRPC, docetaxel (8) was introduced in 2004, but more recently, second generation of androgen-deprivation therapies resulted in better survival (8,9). Ultimately all these tumors will relapse and patients will be offered palliative therapy. Consequently, in order to offer better treatments to these patients, there is a pressing need to identify earlier those tumors that will recur after surgery and evolve to become lethal.

One problem generally inherent to cancer care is to orient people to the adequate treatment corresponding to the stage of the disease and the individual characteristics of the patient (10). In PCa, the stage, grade and PSA level are currently the best standards to drive patients in the different treatment options. Currently, after radical prostatectomy the PSA level is actively monitored to assess the BCR, but there is no biomarker that is used clinically to predict a future BCR.

To reduce costs and continue to improve prognostic, omics data are promising. With the decreasing price of RNA sequencing, the accessibility of affordable technologies (e.g. MinION from Oxford Nanopore Technologies (11)), the available PCa cohorts and the efficient computational approaches, transcriptomics is becoming a valuable resource to identify biomarkers (12). The rapid development of omics technology has led to the availability of many omics databases (13–15), including The Cancer Genome Atlas Program (TCGA) (16) and those of the International Cancer Genome Consortium (ICGC) (17), thus opening an opportunity to apply and test machine learning algorithms (18). These algorithms have been utilized as an aim to model the progression and treatment of cancerous conditions, and resulted in effective and accurate decision-making (19). However, many of the datasets results from patients

cohorts that were either rather small and/or had insufficient follow-up of clinical history which limit their use for clinical outcome prediction.

Hence, there is a challenge to set up predictive models that could anticipate the event of BCR, thus predicting the evolution of cancer, immediately after surgery. Consequently, we propose here a method to discover a transcriptomic signature that could be used to predict BCR events using a combination of datasets to increase the discovery potential. To this purpose, we applied specific preprocessing and cleaning steps on three RNA-seq datasets and established a machine learning protocol.

3.5 Materials and methods

Research pipeline

After recovering the raw data from the different studies, we processed them in a pipeline composed of three main steps: Samples quality control and selection, sequencing data processing, machine learning analysis (Figure 1). All developed scripts are available in the github repository (See data availability).

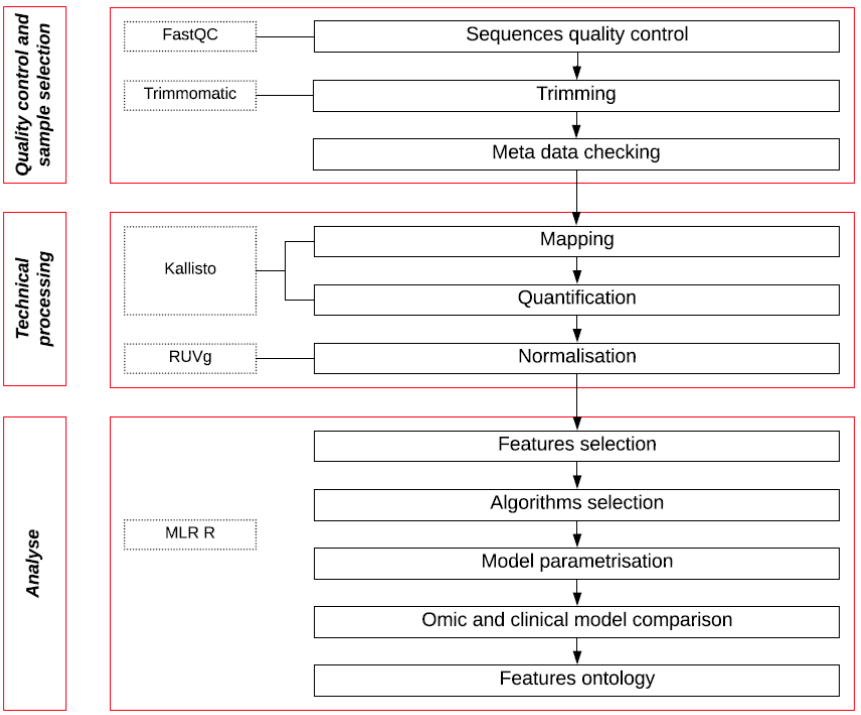


Figure 1: Pipeline workflow. Quality control of raw data sequencing files is measured, then trimmed to remove their adaptors. Patient metadata are then filtered to keep only BCR patients with long follow-up. Retained sequences are then mapped, quantified and normalized. Finally, a machine learning approach is used to analyze the data to obtain a gene expression predictive signature and a model.

Datasets

We retrieved three different RNA-Seq datasets of radical prostatectomy specimens with the associated clinical features. The first dataset is from The Genome Cancer Atlas (TCGA) cohort in the Prostate Adenocarcinoma (PRAD) project. The second dataset (GSE54460) is from a cohort constituted by Long et al. (20) and the third dataset was provided by Prof C. Collins from the Vancouver Prostate Cancer Center (VPCC) (21).

Quality of the BCR event data is dependent on patient clinical follow-up. A patient followed only a few weeks or months after surgery without showing BCR would be considered as a non-BCR case. These cases are a bias since the patient could have experienced a BCR event after the period of follow-up. Consequently, we discarded from our analysis the patients with no BCR whose follow-up was inferior to 60 months.

TCGA-PRAD dataset: Data from 498 samples were initially recovered from the PRAD project on the TCGA data portal (https://portal.gdc.cancer.gov/). According to the TCGA Research Network (22) 131 samples must be discarded because of the presence of RNA degradation, as we did. We also ignored samples with less than 40% of tumor cells (column percent_tumor_cells in clinical file) and follow-up inferior to 60 months. We ended up with 52 samples after these filters.

GSE54460 dataset: The data were downloaded from NCBI website (GEO accession GSE54460) where sequencing and clinical data from 106 patients were recovered. After selecting cases with a minimum of 60 months of follow-up, we ended-up with 96 patients of whom 54 had a BCR.

VPCC dataset: We obtained the raw fastq files and clinical data from 85 patients, available at European Nucleotide Archive of the EMBL-EBI under accession PRJEB6530. Patients treated with hormonal therapy before radical prostatectomy were removed because this treatment strongly alters RNA expression. After selecting patients for minimal follow-up we ended up with 23 patients of whom 5 experienced a BCR.

The baseline characteristics of the resulting individual and combined cohorts after selection of eligible cases are summarized in Table 1.

Table 1: Baseline characteristics of the cohorts

		TCGA	GSE54460	VPCC	Total
Patients					
		52	96	23	171
Grade					
Low grade					
5		0	1	3	4
6		2	9	12	23
7		14	72	4	90
	Total	16	82	19	117
High grade					
8		9	9	1	19
9		27	5	2	34

10		0	0	1	1
NA		0	0	0	0
	Total	36	14	4	54
	Total	52	96	23	171
Stage					
T1C		0	14	0	14
T2		0	7	0	7
T2A		1	21	3	25
T2B		2	10	0	12
T2C		9	26	17	52
T3		0	2	0	2
T3A		16	5	2	23
T3B		24	9	1	34
T4		0	1	0	1
NA		0	1	0	1
	Total	52	96	23	171
BCR					
NO		14	54	5	73
YES		38	42	18	98
	Total	52	96	23	171
PSA at dx/preop					
<= 10		31	64	21	116
10 - 20		16	17	1	34
>=20		5	12	1	18
NA		0	3	0	3
	Total	52	96	23	171

Quality control, alignment and gene expression

The quality of the raw fastq files from the TCGA cohort was measured using FastQC (23) (v0.11.5) and Trimmomatic (24) (v0.32). A threshold quality per base of 30 (based on Phred 33) and a minimal length of 40 bases were applied. The transcriptomes were then mapped on GrCH38.p7 using Kallisto (25) (v0.43.0). The software Kallisto was used to estimate isoform counts, adjusted for the amount of bias in the experiment to ensure a coherent no-naive mapping.Default paired end parameters indicated in kallisto's manual were used. The index needed to run Kallisto is provided on the official github repository (https://github.com/pachterlab/kallisto-transcriptome-indices/releases), but can be manually created. Consequently, we computed gene counts with tximport (26). The Ensembl gene identifiers were converted with Biomart tools (27,28)

from transcript ID to gene ID. For both GSE54460 and VPCC datasets, we processed the raw fastq files using the same method as for the TCGA dataset. However, in GSE54460 the ribosomal sequences were still present within the reads, so we separated these sequences from the mapped reads and removed them. After mapping procedure, 29820 Ensembl genes were found in TCGA-PRAD dataset, 28704 in GSE54460 dataset and 32334 in VPCC dataset. The difference of number of Ensembl genes detected is explained by the sequencing depth of the datasets. A total of 25504 Ensembl genes were common to all sets and were retained for the analysis.

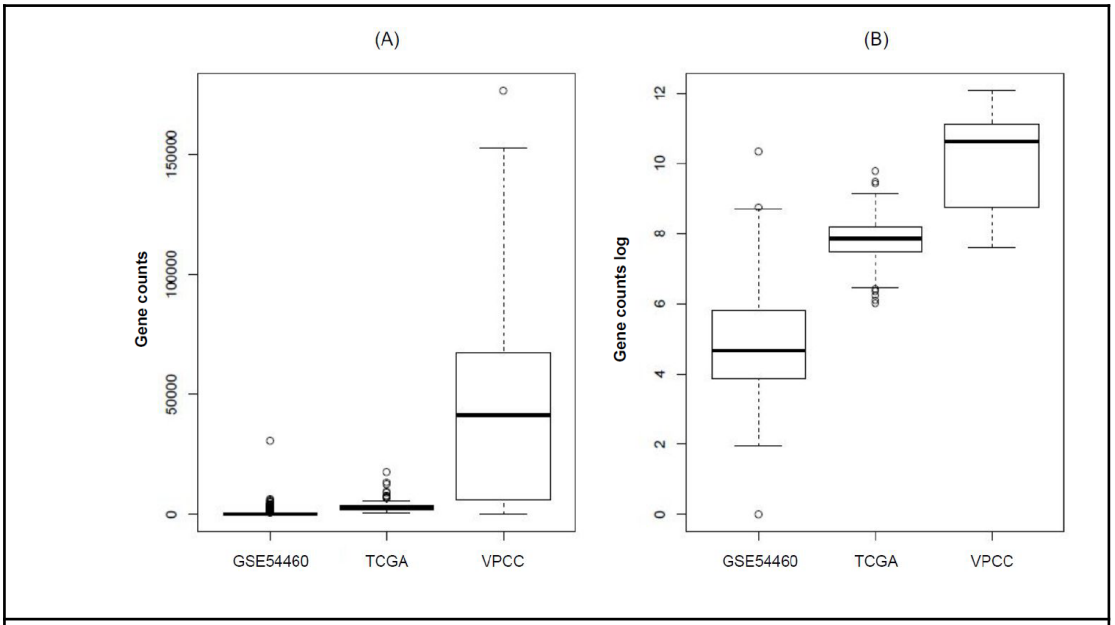


Figure 2: Summary of gene expression value in each dataset (A) or log of the expression value (B).

Normalization

The gene expression data were normalized with the RUV method (29,30) in each dataset separately following the default protocol indicated in the

RUVseq package vignette. RUVg uses negative control genes (housekeeping genes), assumed not to be differentially expressed. In order to normalize properly we selected in the literature a set of specific housekeeping genes (HKG) candidates for PCa (31–34): ACTB, PPIA, GAPDH, PGK1, GUSB, RRN18S, and RPL13A. The expression of these genes was tested by RT-qPCR in a series of 50 prostate tumors and the genes were shown to be stably expressed between tumor samples. We excluded from the final list the ribosomal genes RRN18S and RPL13A because ribosomal RNAs were removed from our RNA-seq datasets. PGK1 was also excluded according to recent results (31). Finally, four genes were chosen: GUSB, PPIA, GAPDH and ACTB.

Machine learning

There are multiple approaches to treat biological data in a machine learning workflow (35,36). Many machine learning libraries exist, in various programming languages, such as MLR in R (37), ScikitLearn (38) in python and WEKA (39) in Java. We chose the MLR (v2.8) package in R to set up our work. Our general workflow is described in Figure 3.

Validation strategy

We performed a resampling to assess the performance of the learning algorithm, avoid over-optimistic results and get a more robust measure of the performance of our model. The entire dataset was split into a random stratified (i.e. class balance preserved) training and testing sets, 1000 times, hence the classification algorithm is trained and tested on different sets. The measure of performance is an aggregated value (e.g. average) of the individual performance on the test set. Because we have no repeated measures and independent variables (i.e. the patients) we chose the subsampling method which is also the best in general in different benchmarks but is less effective computationally (40). The resampling strategy was run 200 times with a split of 2/3 for training and 1/3 for test sets. In the resampling methods the split is

usually 4/5 or 9/10. In our case we wanted to avoid over-optimistic results then we chose a smaller train set closer to a classical cross validation (CV) approach.

Performance metric

To evaluate the performance we used the balanced error rate (BER), the Matthews correlation coefficient (MCC) and the Mean misclassification error (MMCE). The BER is calculated as the average proportion of wrongly classified samples in each class and weights up small sample size classes (Table 2). The area under the curve (AUC) was also reported.

Table 2: Performance measures. The detailed formula of our metrics.

Performance metric	Formula
Sensitivity	TP/(TP + FN)
Specificity	TN/(TN + FP)
Accuracy	(TP + TN)/(TP + TN + FP + FN)*100
MCC	$\sqrt{(TP + TN)/(TP + TN + FP + FN)} * 100$
BER	1 - 0.5(Sensitivity + Specificity)
MMCE	mean(response != truth)

Feature selection

Feature selection was performed to reduce dimensionality to improve prediction performances by removing uninformative features, which has been proven successful in other studies (41). There are different approaches to identify relevant features (42–44). We chose information gain ranking, an entropy based method, that can handle both numerical (e.g. gene expression) and categorical data (e.g. clinical data). In MLR this method relies on the package FSelector which is an entropy based selection method (45,46).

Classifier hyper-parametrization

Algorithms typically require to change the settings of parameters to optimize their performance. The optimization method was the Irace method (47) which is automated and implemented in an R package. We also work with a grid search algorithm for some specific parameters, which span the space in a number of chosen steps. These methods are also available within the MLR package to be used directly with the created tasks. The hyperparameters search depends on the algorithm iterated, defined in the MLR related man page.

3.6 Results

Model and features selection

Following our machine learning pipeline (Figure 3), we first reduced the dimension of the dataset and removed non-informative features to obtain 400 top ranked features to train and benchmark 13 models (Figure 4). We observed that the random forest (RF) algorithm (48) performed best on our data. The classical RF was chosen as the main model for our further analysis.

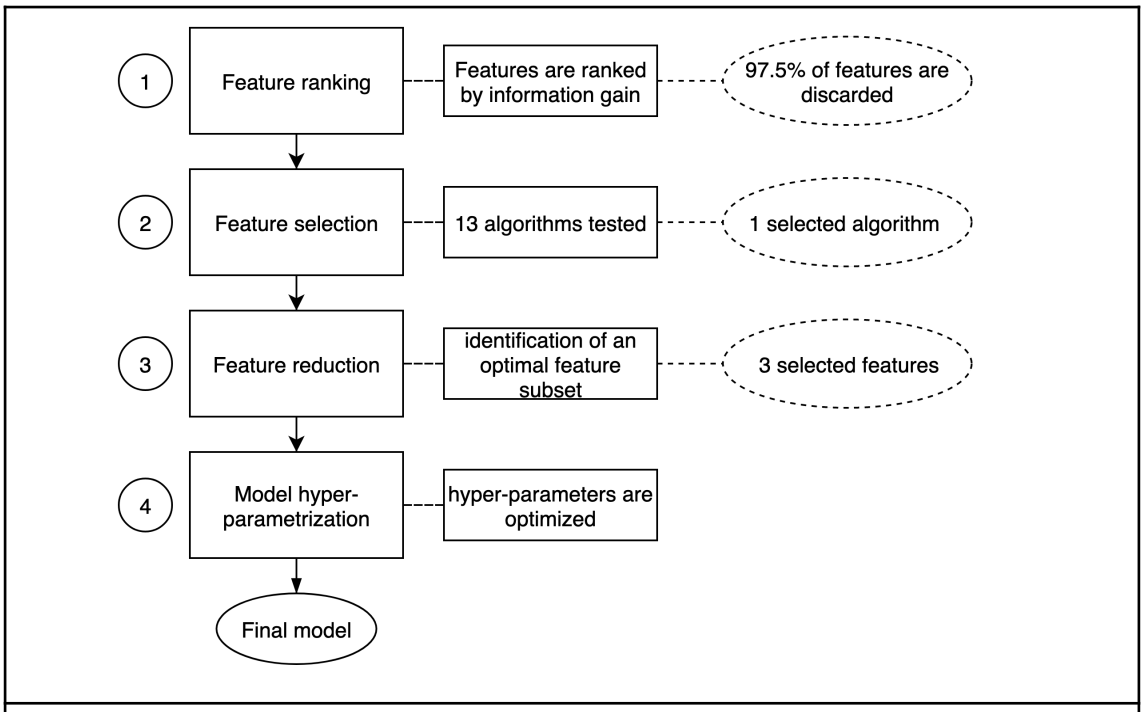


Figure 3: Machine learning feature selection and model evaluation workflow

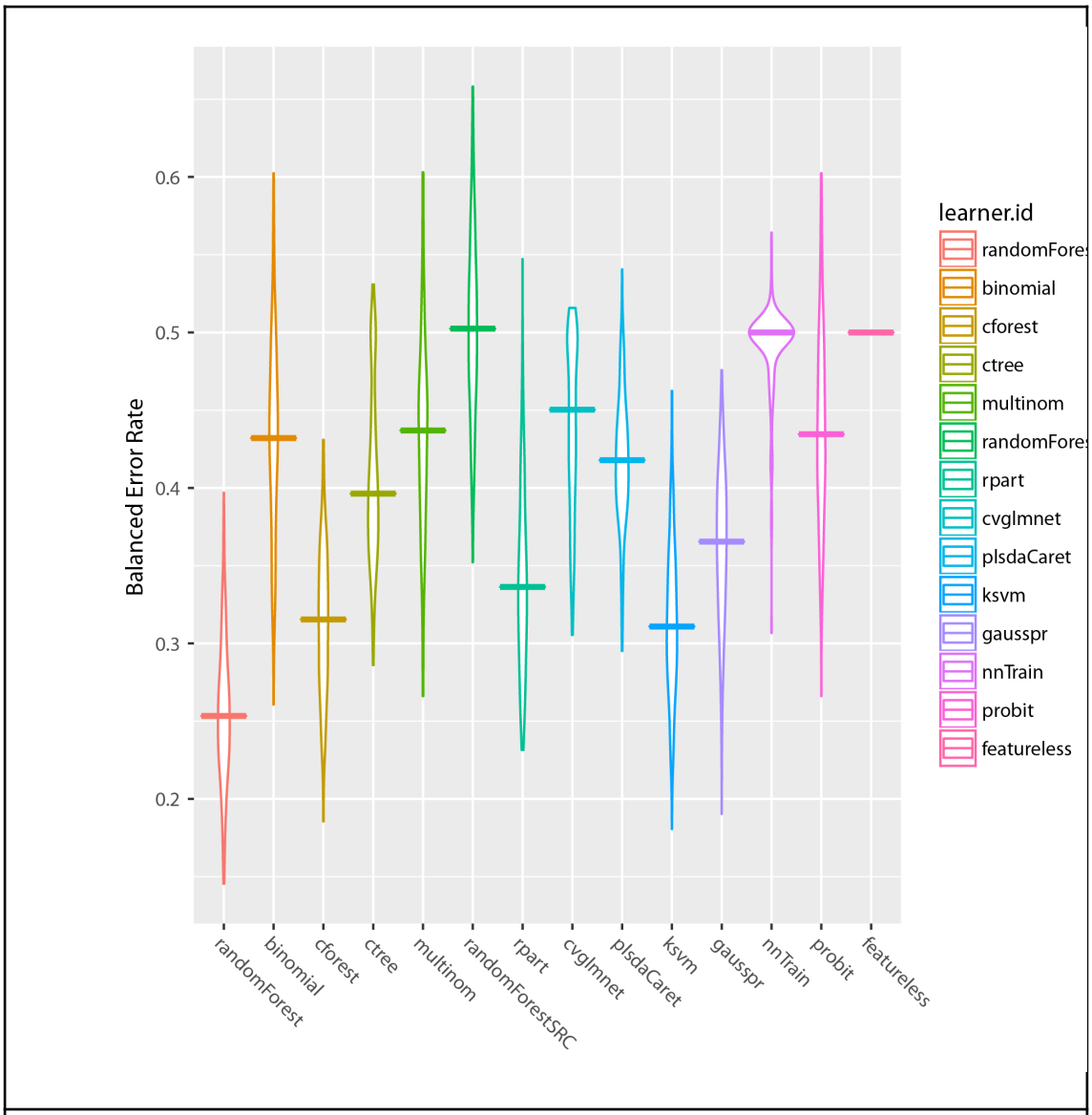


Figure 4: Machine learning algorithms comparison. The BER results of our 13 benchmarked algorithms are presented. The last model is a featureless control case.

Since our goal was to identify a very short genomic signature we looked up the BER rate and other metrics while varying the number of selected features, from 1 to 400, used in the model. We observed that the BER and MMCE dropped rapidly with a few features selected (<3) then oscillated around 0.27 (Figure 5).

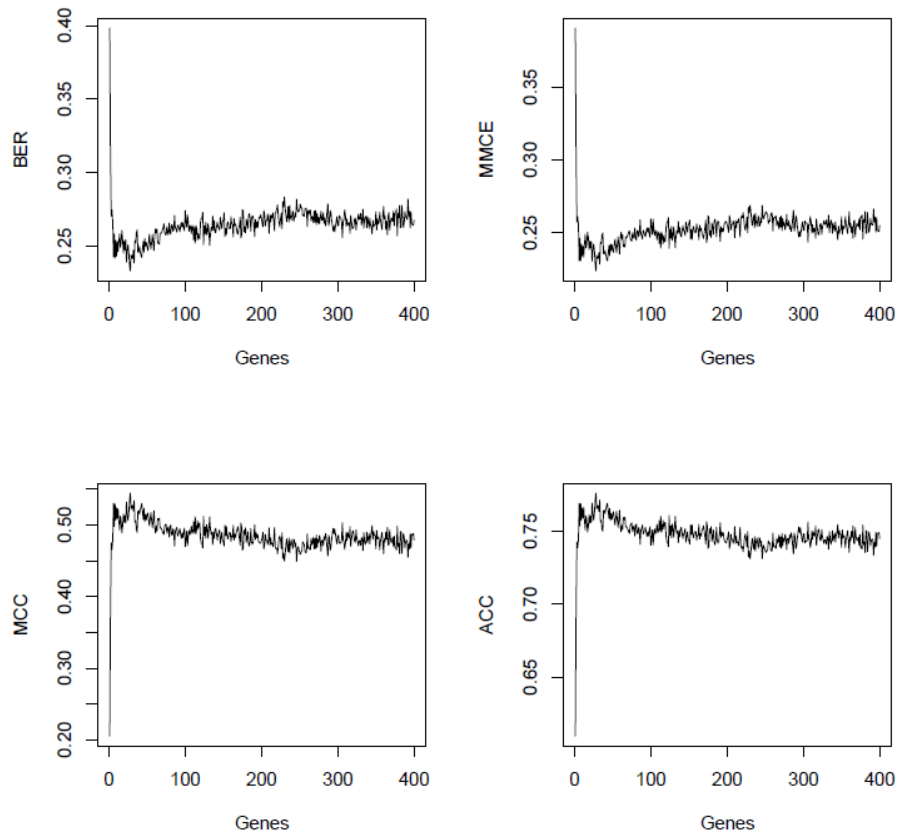


Figure 5: For the 400 genes tested the best genes/performance ratio is obtained with less than 20 genes in our model.

The MCC and the ACC went up rapidly and stabilized in the same way. After these observations, we focused the analysis on the first eight genes. The results are shown in Table 3. We observed a shift in BER value after adding the third most predictive gene to the signature. Afterwards, BER begins to stabilize around 0.25-0.28 despite adding more informative genes. Consequently, we decided to keep the first three genes for the rest of the analysis. These genes are ENSG00000125534 (PPDPF), ENSG00000177606 (JUN) and ENSG00000188290 (HES4).

Table 3: Feature selection benchmark. Benchmark on specific number of features has been performed and results of the performance metrics are presented.

Nb of features	BER	MMCE	MCC	ACC	Gene name	ENSG
1	0.40	0.39	0.20	0.60	PPDPF	ENSG00000125534
2	0.32	0.30	0.38	0.69	HES4	ENSG00000188290
3	0.28	0.28	0.48	0.74	JUN	ENSG00000177606
4	0.28	0.26	0.47	0.73	GNB2	ENSG00000172354
5	0.28	0.26	0.48	0.74	PYROXD2	ENSG00000119943
6	0.25	0.23	0.53	0.77	MAP3K2	ENSG00000169967
7	0.27	0.25	0.50	0.75	RPL28	ENSG00000108107
8	0.25	0.23	0.53	0.77	DHCR24	ENSG00000116133

Hyper-parameters optimization and final model

Four hyper-parameters of the RF classifier were optimized: ntree, mtry, maxnode and nodesize. Ntree refers to the number of decision trees in the model, mtry the number of variables selected from a decision split for the next split, maxnodes the maximal number of nodes in the forest and nodesize the minimal number of samples allowed in a node. Because we selected only three features, the parametrization step was not expected to drastically change the performance of our optimization task. First we used a grid search method to define the best setting for each parameter taken individually, letting the others at default. The grid search provided us 500 (ntree),1 (mtry), 24 (maxnodes) and 5 (nodesize) (Figure 6).

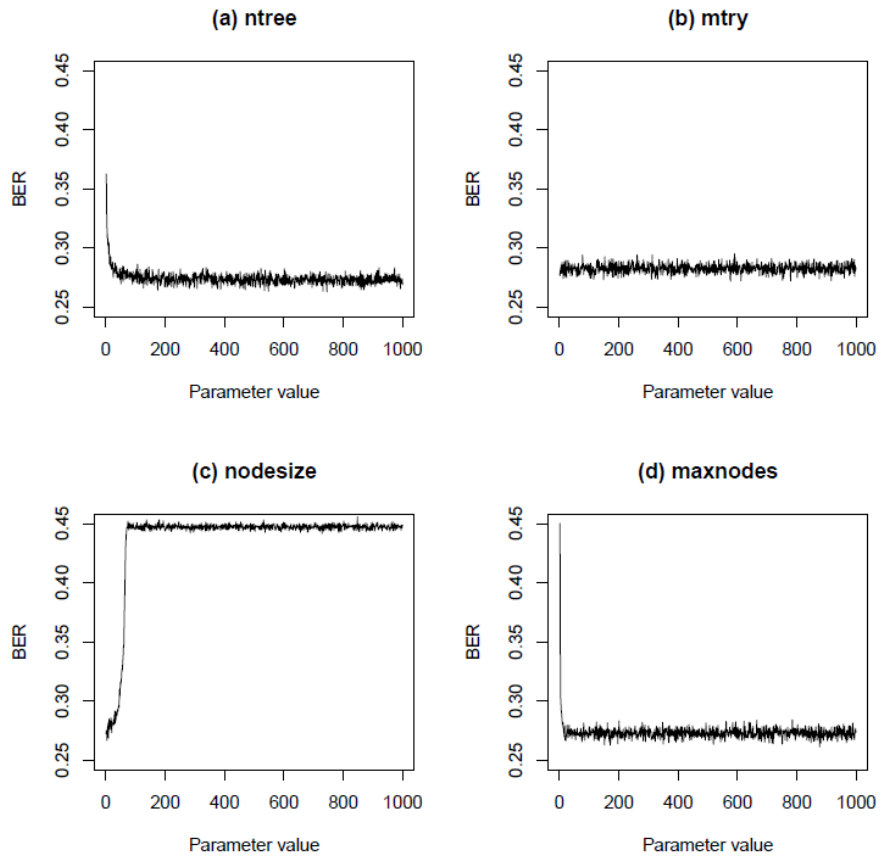


Figure 6: BER evolution according to modulation of RF parameters. Four different RF hyper-parameters were tested while keeping the others at default value in a grid search approach. The results were then used in an Irace search to find optimal parameters.

From these hyper-parameters an Irace search was performed around the space of those values. The best value was obtained with ntree, mtry, maxnodes and nodesize at 187, 1, 881 and 1 resp. for a BER of 0.27. We observed relative stability despite the modification of the hyperparameters.

To ensure the stability of our three-gene model, a subsampling test was done 100000 times for the last part of our work. From this subsampling, the results obtained are ber=0.274, mmce=0.26, mcc=0.468, fpr=0.368, tpr=0.82, acc=0.739. Then we calculated the associated AUC (0.761) and plotted the ROC curve Figure 7.

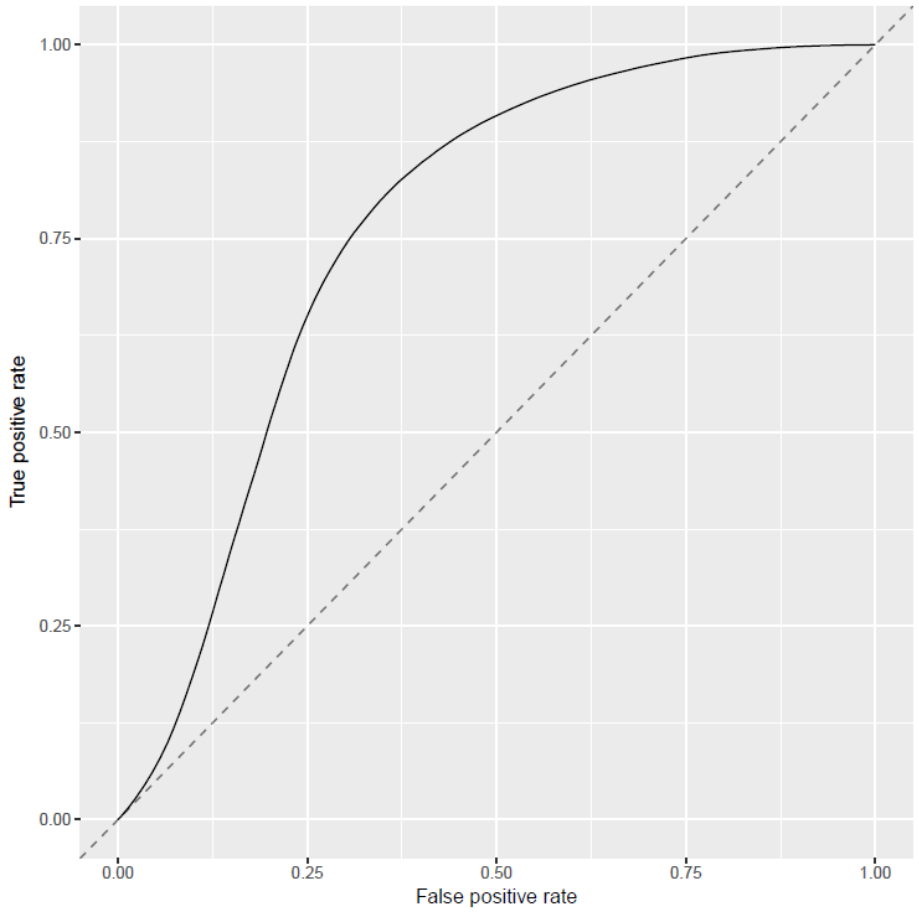


Figure 7: ROC curve for the three-gene model

The proposed three genes signature (see gene distribution for each cohort in Figure 8) model can be retrained using the training data provided in the github repository (see data availability section), and new data must be processed following the indications in Materials and Methods before being submitted to the model.

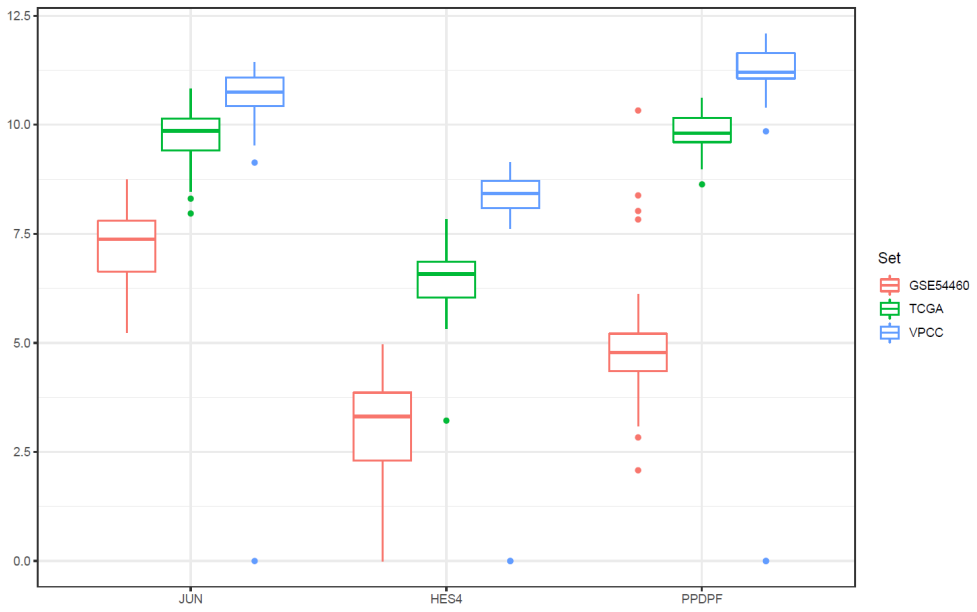


Figure 8: Log2 transformed distribution of normalized read counts for the three genes signature in each cohort.

Comparison of omics and clinic models

We compared the potential of omic data versus clinical data to assess the accuracy of our omics model. A RF model for the clinical data (Grade, stage and PSA) and a merged model combining clinic and omics data were set up following the same protocol used for the omics data. For the clinical model the best BER obtained was 0.311 and for the mixed model the best BER obtained was 0.276 (Table 4).

Table 4: Comparison of model performance using clinic or omics data or both. The omics model is based on three genes and the clinic model is a model based on the grade, stage and PSA. The omics + clinic model integrates all the selected features together.

Metric	Omics	Clinic	Omics + Clinic
BER	0.27	0.32	0.28

MMCE	0.257	0.306	0.265
MCC	0.474	0.373	0.457
ACC	0.742	0.692	0.734
Parameters			
ntree	187	1402	667
mtry	1	3	1
maxnodes	881	30	25
nodesize	1	4	6

Single cohort performance

To further assess the performance of the three-gene model obtained with the combined dataset, we also performed the analysis with the individual cohorts. We used the RF algorithm iterated on the 50 best features from Information Gain on the three datasets evaluated by leave one out group validation (i.e. 2 datasets for training, one for testing), and the combined dataset evaluated by resampling (see Validation strategy in Methods). The results are displayed in Figure 9 and show that the combined dataset offers better performances and specially more stable performances

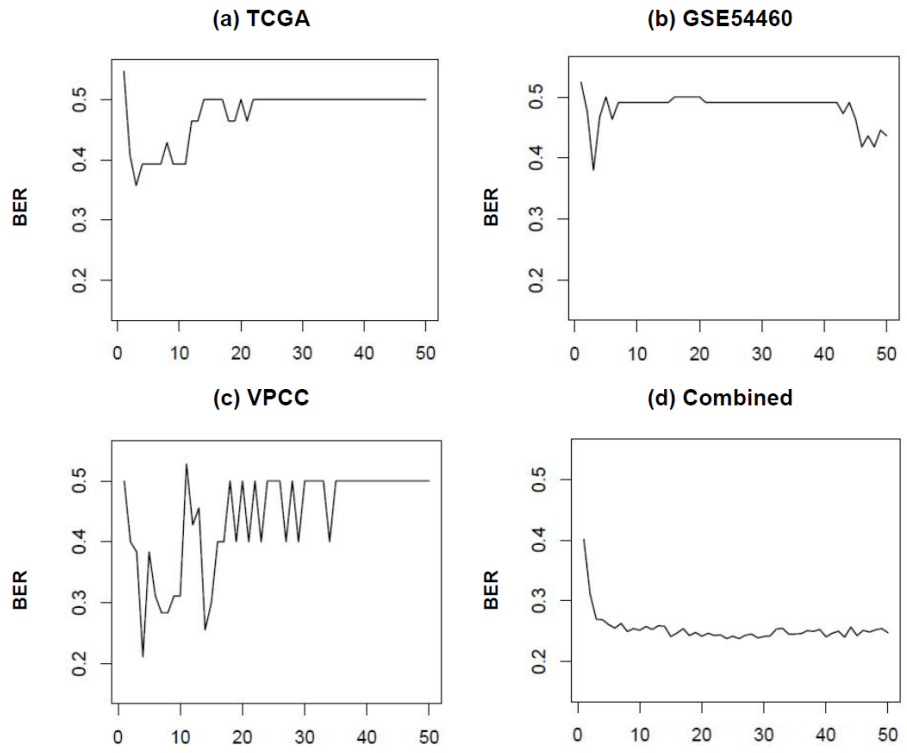


Figure 9: Performance obtained using leave one out group validation. (a) Model trained on GSE54460 and VPCC then tested on TCGA. (b) Model trained on TCGA and VPCC then tested on GSE54460 . (c) Model trained on GSE54460 and TCGA then tested on VPCC. (d) Combined dataset evaluated by subsampling method described in Validation strategy in Methods.

3.7 Discussion

Machine Learning is one of the fastest growing fields in bioinformatics (49) and its application to healthcare is a challenge. In the past decade, various mathematical methods using combination of omics biomarkers (50,51), including non-coding RNAs, PCA3, TMPRSS2:ERG (52) were developed to improve PCa diagnosis (53,54), define the grade (55), define the risk (56) and predict survival time (57). Machine learning approaches to predict BCR or other characteristics demonstrated good performances in various situations. Lalonde & al. (58,59) built a 100 loci-DNA (CNV) signature for low to high risk cohorts with 563 patients and a 60-month follow-up for BCR. The obtained AUC was 0.74, which is similar to our performance but with another technology (CNV

assay) and for much fewer biomarkers. Moreover, a model containing so many features can be suspected of overfitting. Regnier-Courdert et al. (60) built a model on Partin table from a large cohort of 1700 patients to improve cancer grading and staging, and obtained an AUC of 0.68. Mangiola et al. (61) focused on gene expression but chose to predict dichotomous cohorts with low versus high risk patients. With a cohort of 80 patients and an average follow-up of 27-29 months they achieved an AUC of 0.72. Finally, Abou-Ouf et al. (62) used a large cohort of 545 patients to define a ten-gene signature from microarray exon chips to predict BCR, but couldn't exceed an AUC of 0.65. Thus, there was a large room for improvement in terms of predictive performance, and a lack of focus on small gene signature, much easier to reproduce, to predict BCR with recent technology (RNA-Seq).

In this study, we took advantage of the power of machine learning to identify a biomarker signature composed of three genes. We showed that such short signature from omics data performs better to predict BCR than clinico-pathological features or a combination of these data (i.e. clinico-pathological + omics data). We have explored many machine learning algorithms, since each has its advantages and drawbacks in terms of computational time, hyper-parameters and range of application (class, type and dimension) and also because their performance depends on the type of data and their composition (63). Using this approach, we ended with a Random Forest model with a 27% BER with a three genes signature.

The identified signature of the proposed model contains three genes: JUN, HES4 and PPDPF. Gene JUN is well known for being a transcription factor acting as an oncogene (64–67). Proteins of the JUN family combined with the Fos protein to form the heterodimeric AP-1 transcription factor. This complex can enter into the nucleus and bind specific DNA sequences to module targeted genes. AP-1 activity is induced by stimuli such as growth factors and cytokines that bind to specific cell surface receptors (68). Recently a miRNA targeting JUN has been identified as tumor suppressor (69).

HES4 (Hes Family BHLH Transcription Factor 4) is a gene related to the PI3K-Akt signaling pathway. This gene is a transcription factor binding DNA. It is

related to the NOTCH3 receptor and is a biomarker of PCa aggressiveness (70) and is also related to colorectal cancer in the same pathway (71). It was demonstrated as a high grade biomarker of osteosarcoma (72).

Finally, PPDPF is known to be expressed during pancreas development (Pancreatic Progenitor Cell Differentiation And Proliferation Factor (73)) and differentially expressed in several types of cancer (74,75). But it was not previously associated with prostate cancer.

We have attempted to understand the biological links between these three genes and the eventual relation with the BCR. This is not straightforward considering that Random Forest models tend to reflect a nonlinear approximation of statistical relationships, hence providing little insight of how elements of the signature are related. Thus, we have performed a protein-protein interaction networks functional enrichment analysis using String-DB (76) on the three identified genes, but no evident relations could be found, even after addition of intermediate protein nodes. We have also performed a gene list enrichment analysis and candidate gene prioritization based on functional annotations using ToppGene Suite (77) using the three identified genes. The only biologically relevant (i.e. cancer hormono-dependant as the Pca) and significant (q-value 2.1E-2 after FDR Benjamini–Yekutieli procedure correction) hit is that the three genes exist in the Human Breast Nam08 30genes UpregulatedGeneList signature (78), provided by GeneSigDB (79), but no evident and/or significant biological functions by ontology seem to link these three genes together. We have eventually expanded the list of three genes to 320 genes by retrieving correlated genes (>90% Pearson correlation) and observed that many genes were involved in mitochondrial functions, including mitochondrial translation, mitochondrial gene expression, mitochondrial translational termination and mitochondrial translational elongation, all having a q-value <5.9E-5 after FDR Benjamini–Yekutieli procedure correction. This observation is supported by other studies who have found a clear relation between mitochondrial genomic alterations and BCR (80–82).

This is not the first time that predictive three-genes signatures have been identified in various diseases (83–94), hence showing that extensive research is ongoing to identify multigenic signatures containing a reasonable number of potential targets. The identified genes could be eventually verified in other cohorts or by experimental validations. One key point should be to add gradually smaller datasets to control the signature stability with various experiments and technologies. Integrate too large cohorts in this approach will imbalance model parameters in favor of that cohort, then all the advantages of using several small dataset will be lost. This approach has the advantage of offering a small research team the opportunity to integrate their own work in a larger view. After integrating more dataset, a set up in a specific technology such as TaqMan probe to evaluate gene expression could be proposed as diagnosis and maybe to develop drugs (95,96). Conclusions

By using an appropriate data transformation strategy and machine learning pipeline, we have identified a three-gene signature. With the decreasing price of RNA sequencing and its growing accuracy there are opportunities for less invasive and faster exams if the right biological variables are chosen. Other investigations on other omics data using the same machine learning approach could be undertaken, such as using miRNAs (90,91). We also showed that it is possible to concatenate several cohorts to get stable and performing models from heterogeneous RNA-Seq PCa datasets, hence showing a robustness against batch effect. This study demonstrates the potential of taking advantage of many independent datasets produced on the same disease. Machine learning algorithms can handle the batch effect if there is the right preprocessing pipeline applied on the data.

Funding

This research was realized with internal funds from the Laboratoire d'Uro-Oncologie Expérimentale (Ulaval, Dr Fradet). The production of RNA-seq data at VPCC was realized with funds from the Terry Fox Research Institute New Frontier Program Project Grant #1062 (TFRI NF PPG, UBC - Dr Collins).

3.8 Bibliographie

1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. CA Cancer J Clin (2017) 67:7–30.

2. Weiner AB, Matulewicz RS, Eggener SE, Schaeffer EM. Increasing incidence of metastatic prostate cancer in the United States (2004–2013). Prostate Cancer and Prostatic Diseases (2016) 19:395–397. doi:10.1038/pcan.2016.30

3. Buyyounouski MK, Pickles T, Kestin LL, Allison R, Williams SG. Validating the interval to biochemical failure for the identification of potentially lethal prostate cancer. J Clin Oncol (2012) 30:1857–1863.

4. D'Amico AV, Moul J, Carroll PR, Sun L, Lubeck D, Chen M-H. Cancer-Specific Mortality After Surgery or Radiation for Patients With Clinically Localized Prostate Cancer Managed During the Prostate-Specific Antigen Era. Journal of Clinical Oncology (2003) 21:2163–2172. doi:10.1200/jco.2003.01.075

5. Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK, Gershenwald JE, Compton CC, Hess KR, Sullivan DC, et al. AJCC Cancer Staging Manual. Springer (2018).

6. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. Annals of Surgical Oncology (2010) 17:1471–1474. doi:10.1245/s10434-010-0985-4

7. Papsidero LD, Wang MC, Valenzuela LA, Murphy GP, Chu TM. A prostate antigen in sera of prostatic cancer patients. Cancer Res (1980) 40:2428–2432.

8. Tannock IF, de Wit R, Berry WR, Horti J, Pluzanska A, Chi KN, Oudard S, Théodore C, James ND, Turesson I, et al. Docetaxel plus Prednisone or Mitoxantrone plus Prednisone for Advanced Prostate Cancer. New England Journal of Medicine (2004) 351:1502–1512. doi:10.1056/nejmoa040720

9. Nevedomskaya E, Baumgart SJ, Haendler B. Recent Advances in Prostate Cancer Treatment and Drug Discovery. Int J Mol Sci (2018) 19: doi:10.3390/ijms19051359

10. Terada N, Akamatsu S, Kobayashi T, Inoue T, Ogawa O, Antonarakis ES. Prognostic and predictive biomarkers in prostate cancer: latest evidence and clinical implications. Therapeutic Advances in Medical Oncology (2017) 9:565–573. doi:10.1177/1758834017719215

11. Menegon M, Cantaloni C, Rodriguez-Prieto A, Centomo C, Abdelfattah A, Rossato M, Bernardi M, Xumerle L, Loader S, Delledonne M. On site DNA barcoding by nanopore sequencing. PLoS One (2017) 12:e0184741.

12. Nikitina AS, Sharova EI, Danilenko SA, Butusova TB, Vasiliev AO, Govorov AV, Prilepskaya EA, Pushkar DY, Kostryukova ES. Novel RNA biomarkers of prostate cancer revealed by RNA-seq analysis of formalin-fixed samples obtained from Russian patients. Oncotarget (2017) 8:32990–33001.

13. Almeida H, Meurs M-J, Kosseim L, Butler G, Tsang A. Machine learning for biomedical literature triage. PLoS One (2014) 9:e115892.

14. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big Data: Astronomical or Genomical? PLoS Biol (2015) 13:e1002195.

15. Marx V. The big challenges of big data. Nature (2013) 498:255–260. doi:10.1038/498255a

16. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (2015) 19:A68–77.

17. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, et al. International network of cancer genome projects. Nature (2010) 464:993–998.

18. Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. Briefings in Bioinformatics (2016)bbw113. doi:10.1093/bib/bbw113

19. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction.

Computational and Structural Biotechnology Journal (2015) 13:8–17. doi:10.1016/j.csbj.2014.11.005

20. Long Q, Xu J, Osunkoya AO, Sannigrahi S, Johnson BA, Zhou W, Gillespie T, Park JY, Nam RK, Sugar L, et al. Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. Cancer Res (2014) 74:3228–3237.

21. Wyatt AW, Mo F, Wang K, McConeghy B, Brahmbhatt S, Jong L, Mitchell DM, Johnston RL, Haegert A, Li E, et al. Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer. Genome Biol (2014) 15:426.

22. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell (2015) 163:1011–1025.

23. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC: a quality control tool for high throughput sequence data. Babraham Institute (2010) Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (2014) 30:2114–2120.

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

26. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res (2015) 4:1521.

27. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (2011) 2011:bar030.

28. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res (2015) 43:W589–98.

29. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics (2012) 13:539–552.

30. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology (2014) 32:896–902. doi:10.1038/nbt.2931

31. Vajda A, Marignol L, Barrett C, Madden SF, Lynch TH, Hollywood D, Perry AS. Gene expression analysis in prostate cancer: the importance of the endogenous control. Prostate (2013) 73:382–390.

32. Chua SL, See Too WC, Khoo BY, Few LL. UBC and YWHAZ as suitable reference genes for accurate normalisation of gene expression using MCF7, HCT116 and HepG2 cell lines. Cytotechnology (2011) 63:645–654.

33. de Kok JB, Roelofs RW, Giesendorf BA, Pennings JL, Waas ET, Feuth T, Swinkels DW, Span PN. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. Lab Invest (2005) 85:154–159.

34. Ohl F, Jung M, Xu C, Stephan C, Rabien A, Burkhardt M, Nitsche A, Kristiansen G, Loening SA, Radonić A, et al. Gene expression studies in prostate cancer tissue: which reference gene should be selected for normalization? J Mol Med (2005) 83:1014–1024.

35. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS One (2018) 13:e0194889.

36. Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K. Efficient Machine Learning for Big Data: A Review. Big Data Research (2015) 2:87–93. doi:10.1016/j.bdr.2015.04.001

37. Lesmeister C. Mastering Machine Learning with R. Packt Publishing Ltd (2015).

38. Garreta R, Moncecchi G. Learning scikit-learn: Machine Learning in Python. Packt Publishing Ltd (2013).

39. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. ACM SIGKDD Explorations Newsletter (2009) 11:10. doi:10.1145/1656274.1656278

40. Bischl B, Mersmann O, Trautmann H, Weihs C. Resampling methods for meta-model validation with recommendations for evolutionary computation. Evol Comput (2012) 20:249–275.

41. Novakovic J, Strbac P, Bulatovic D. Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research (2011) 21:119–135. doi:10.2298/yjor1101119n

42. Hira ZM, Gillies DF. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Adv Bioinformatics (2015) 2015:198363.

43. Singh RK, Sivabalakrishnan M. Feature Selection of Gene Expression Data for Cancer Classification: A Review. Procedia Computer Science (2015) 50:52–57. doi:10.1016/j.procs.2015.04.060

44. Raza MS, Qamar U. Introduction to Feature Selection. Understanding and Using Rough Set Based Feature Selection: Concepts, Techniques and Applications (2019)1–25. doi:10.1007/978-981-32-9166-9_1

45. Coifman RR, Wickerhauser MV. Entropy-based algorithms for best basis selection. IEEE Transactions on Information Theory (1992) 38:713–718. doi:10.1109/18.119732

46. Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory (1991) 37:145–151. doi:10.1109/18.61115

47. López-Ibáñez M, Dubois-Lacoste J, Cáceres LP, Birattari M, Stützle T. The irace package: Iterated racing for automatic algorithm configuration. Operations Research Perspectives (2016) 3:43–58. doi:10.1016/j.orp.2016.09.002

48. Ho TK. International Conference On Document Analysis and Recognition. Proceedings of 3rd International Conference on Document Analysis and Recognition (1995) doi:10.1109/icdar.1995.601943

49. Inza I, Calvo B, Armañanzas R, Bengoetxea E, Larrañaga P, Lozano JA. Machine Learning: An Indispensable Tool in Bioinformatics. Methods in Molecular Biology (2010)25–48. doi:10.1007/978-1-60327-194-3_2

50. Gaudreau P-O, Stagg J, Soulières D, Saad F. The Present and Future of Biomarkers in Prostate Cancer: Proteomics, Genomics, and Immunology Advancements. Biomark Cancer (2016) 8:15–33.

51. Halabi S, Small EJ, Kantoff PW, Kattan MW, Kaplan EB, Dawson NA, Levine EG, Blumenstein BA, Vogelzang NJ. Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer. J Clin Oncol (2003) 21:1232–1237.

52. Nilsson J, Skog J, Nordstrand A, Baranov V, Mincheva-Nilsson L, Breakefield XO, Widmark A. Prostate cancer-derived urine exosomes: a novel approach to biomarkers for prostate cancer. Br J Cancer (2009) 100:1603–1607.

53. Guo J, Yang J, Zhang X, Feng X, Zhang H, Chen L, Johnson H, Persson JL, Xiao K. A Panel of Biomarkers for Diagnosis of Prostate Cancer Using Urine Samples. Anticancer Res (2018) 38:1471–1477.

54. Wang X, An P, Zeng J, Liu X, Wang B, Fang X, Wang F, Ren G, Min J. Serum ferritin in combination with prostate-specific antigen improves predictive accuracy for prostate cancer. Oncotarget (2017) 8:17862–17872.

55. Arvaniti E, Fricker KS, Moret M, Rupp N, Hermanns T, Fankhauser C, Wey N, Wild PJ, Rüschoff JH, Claassen M. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci Rep (2018) 8:12054.

56. Paulo P, Maia S, Pinto C, Pinto P, Monteiro A, Peixoto A, Teixeira MR. Targeted next generation sequencing identifies functionally deleterious germline mutations in novel genes in early-onset/familial prostate cancer. PLoS Genet (2018) 14:e1007355.

57. Zupan B, Demsar J, Kattan MW, Beck JR, Bratko I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artif Intell Med (2000) 20:59–75.

58. Lalonde E, Alkallas R, Chua MLK, Fraser M, Haider S, Meng A, Zheng J, Yao CQ, Picard V, Orain M, et al. Translating a Prognostic DNA Genomic Classifier into the Clinic: Retrospective Validation in 563 Localized Prostate Tumors. Eur Urol (2017) 72:22–31.

59. Lalonde E, Ishkanian AS, Sykes J, Fraser M, Ross-Adams H, Erho N, Dunning MJ, Halim S, Lamb AD, Moon NC, et al. Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. The Lancet Oncology (2014) 15:1521–1532. doi:10.1016/s1470-2045(14)71021-6

60. Regnier-Coudert O, McCall J, Lothian R, Lam T, McClinton S, N'dow J. Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers. Artif Intell Med (2012) 55:25–35.

61. Mangiola S, Stuchbery R, Macintyre G, Clarkson MJ, Peters JS, Costello AJ, Hovens CM, Corcoran NM. Periprostatic fat tissue transcriptome reveals a signature diagnostic for high-risk prostate cancer. Endocrine-Related Cancer (2018) 25:569–581. doi:10.1530/erc-18-0058

62. Abou-Ouf H, Alshalalfa M, Takhar M, Erho N, Donnelly B, Davicioni E, Karnes RJ, Bismar TA. Validation of a 10-gene molecular signature for predicting biochemical recurrence and clinical metastasis in localized prostate cancer. J Cancer Res Clin Oncol (2018) 144:883–891.

63. Heung B, Ho HC, Zhang J, Knudby A, Bulmer CE, Schmidt MG. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma (2016) 265:62–77. doi:10.1016/j.geoderma.2015.11.014

64. Mariani O, Brennetot C, Coindre J-M, Gruel N, Ganem C, Delattre O, Stern M-H, Aurias A. JUN Oncogene Amplification and Overexpression Block Adipocytic Differentiation in Highly Aggressive Sarcomas. Cancer Cell (2007) 11:361–374. doi:10.1016/j.ccr.2007.02.007

65. Vogt PK, Bos TJ. jun:Oncogene and Transcription Factor. Advances in Cancer Research (1990)1–35. doi:10.1016/s0065-230x(08)60466-2

66. Wasylyk C, Schneikert J, Wasylyk B. Oncogene v-jun modulates DNA replication. Oncogene (1990) 5:1055–1058.

67. Maki Y, Bos TJ, Davis C, Starbuck M, Vogt PK. Avian sarcoma virus 17 carries the jun oncogene. Proc Natl Acad Sci U S A (1987) 84:2848–2852.

68. Yang JT, Bader BL, Kreidberg JA, Ullman-Culleré M, Trevithick JE, Hynes RO. Overlapping and independent functions of fibronectin receptor integrins in early mesodermal development. Dev Biol (1999) 215:264–277.

69. Liu J, Yan J, Zhou C, Ma Q, Jin Q, Yang Z. miR-1285-3p acts as a potential tumor suppressor miRNA via downregulating JUN expression in hepatocellular carcinoma. Tumour Biol (2015) 36:219–225.

70. Carvalho FLF, Simons B, Berman DM. Abstract B56: Notch signaling in prostate cancer progression. Cancer Research (2012) 72:B56–B56. doi:10.1158/1538-7445.prca2012-b56

71. Sikandar SS, Pate KT, Anderson S, Dizon D, Edwards RA, Waterman ML, Lipkin SM. NOTCH signaling is required for formation and self-renewal of tumor-initiating cells and for repression of secretory cell differentiation in colon cancer. Cancer Res (2010) 70:1469–1478.

72. McManus M, Kleinerman E, Yang Y, Livingston JA, Mortus J, Rivera R, Zweidler-McKay P, Schadler K. Hes4: A potential prognostic biomarker for newly diagnosed patients with high-grade osteosarcoma. Pediatr Blood Cancer (2017) 64: doi:10.1002/pbc.26318

73. Breunig M, Hohwieler M, Seufferlein T, Liebau S, Kleger A. PPDPF impacts pancreatic differentiation of human pluripotent stem cell derived pancreatic organoids. 72 Jahrestagung der Deutschen Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten mit Sektion Endoskopie – 11 Herbsttagung der Deutschen Gesellschaft für Allgemein- und Viszeralchirurgie gemeinsam mit den Arbeitsgemeinschaften der DGAV (2017) doi:10.1055/s-0037-1604922

74. Xue T-C, Zhang B-H, Ye S-L, Ren Z-G. Differentially expressed gene profiles of intrahepatic cholangiocarcinoma, hepatocellular carcinoma, and combined hepatocellular-cholangiocarcinoma by integrated microarray analysis. Tumour Biol (2015) 36:5891–5899.

75. Voena C, Di Giacomo F, Panizza E, D'Amico L, Boccalatte FE, Pellegrino E, Todaro M, Recupero D, Tabbò F, Ambrogio C, et al. The EGFR family members sustain the neoplastic phenotype of ALK+ lung adenocarcinoma via EGR1. Oncogenesis (2013) 2:e43.

76. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res (2019) 47:D607–D613.

77. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res (2009) 37:W305–11.

78. Nam DH, Jeon HM, Kim S, Kim MH, Lee YJ, Lee MS, Kim H, Joo KM, Lee DS, Price JE, et al. Activation of notch signaling in a xenograft model of brain metastasis. Clin Cancer Res (2008) 14: doi:10.1158/1078-0432.CCR-07-4039

79. Culhane AC, Schröder MS, Sultana R, Picard SC, Martinelli EN, Kelly C, Haibe-Kains B, Kapushesky M, St Pierre A-A, Flahive W, et al. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. Nucleic Acids Res (2012) 40:D1060–6.

80. Ellinger J, Müller SC, Wernert N, von Ruecker A, Bastian PJ. Mitochondrial DNA in serum of patients with prostate cancer: a predictor of biochemical recurrence after prostatectomy. BJU Int (2008) 102:628–632.

81. Xu J, Chang W-S, Tsai C-W, Bau D-T, Davis JW, Thompson TC, Logothetis CJ, Gu J. Mitochondrial DNA copy number in peripheral blood leukocytes is associated with biochemical recurrence in prostate cancer patients in African Americans. Carcinogenesis (2020) 41:267–273.

82. Kalsbeek AMF, Chan EFK, Grogan J, Petersen DC, Jaratlerdsiri W, Gupta R, Lyons RJ, Haynes A-M, Horvath LG, Kench JG, et al. Mutational load of the mitochondrial genome predicts pathological features and biochemical recurrence in prostate cancer. Aging (2016) 8:2702–2712.

83. Xiao K, Liu Q, Peng G, Su J, Qin C-Y, Wang X-Y. Identification and validation of a three-gene signature as a candidate prognostic biomarker for lower grade glioma. PeerJ (2020) 8:e8312.

84. Bao B, Zheng C, Yang B, Jin Y, Hou K, Li Z, Zheng X, Yu S, Zhang X, Fan Y, et al. Identification of Subtype-Specific Three-Gene Signature for Prognostic Prediction in Diffuse Type Gastric Cancer. Front Oncol (2019) 9:1243.

85. Ding T-T, Ma H, Feng J-H. A three-gene novel predictor for improving the prognosis of cervical cancer. Oncol Lett (2019) 18:4907–4915.

86. Saidak Z, Pascual C, Bouaoud J, Galmiche L, Clatot F, Dakpé S, Page C, Galmiche A. A three-gene expression signature associated with positive surgical margins in tongue squamous cell carcinomas: Predicting surgical resectability from tumour biology? Oral Oncol (2019) 94:115–120.

87. Yang Y, Lu Q, Shao X, Mo B, Nie X, Liu W, Chen X, Tang Y, Deng Y, Yan J. Development Of A Three-Gene Prognostic Signature For Hepatitis B Virus Associated Hepatocellular Carcinoma Based On Integrated Transcriptomic Analysis. J Cancer (2018) 9:1989–2002.

88. Chen H, Liu X, Jin Z, Gou C, Liang M, Cui L, Zhao X. A three miRNAs signature for predicting the transformation of oral leukoplakia to oral squamous cell carcinoma. Am J Cancer Res (2018) 8:1403–1413.

89. Li B, Feng W, Luo O, Xu T, Cao Y, Wu H, Yu D, Ding Y. Development and Validation of a Three-gene Prognostic Signature for Patients with Hepatocellular Carcinoma. Sci Rep (2017) 7:5517.

90. De Palma G, Sallustio F, Curci C, Galleggiante V, Rutigliano M, Serino G, Ditonno P, Battaglia M, Schena FP. The Three-Gene Signature in Urinary Extracellular Vesicles from Patients with Clear Cell Renal Cell Carcinoma. J Cancer (2016) 7:1960–1967.

91. Ibrahim MK, Salama H, Abd El Rahman M, Dawood RM, Bader El Din NG, Salem HF, Abdelrahim MEA, Omran D, Omran MH, El-Wakeel KH, et al. Three Gene Signature for Predicting the Development of Hepatocellular Carcinoma in Chronically Infected Hepatitis C Virus Patients. J Interferon Cytokine Res (2016) 36:698–705.

92. Wang W, Zhang L, Wang Z, Yang F, Wang H, Liang T, Wu F, Lan Q, Wang J, Zhao J. A three-gene signature for prognosis in patients with MGMT promoter-methylated glioblastoma. Oncotarget (2016) 7:69991–69999.

93. Thakkar A, Raj H, Ravishankar, Muthuvelan B, Balakrishnan A, Padigaru M. High Expression of Three-Gene Signature Improves Prediction of Relapse-Free Survival in Estrogen Receptor-Positive and Node-Positive Breast Tumors. Biomark Insights (2015) 10:103–112.

94. Sun L-L, Wu J-Y, Wu Z-Y, Shen J-H, Xu X-E, Chen B, Wang S-H, Li E-M, Xu L-Y. A three-gene signature and clinical outcome in esophageal squamous cell carcinoma. Int J Cancer (2015) 136:E569–77.

95. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer (2019) 19:133–150.

96. Laetsch TW, DuBois SG, Mascarenhas L, Turpin B, Federman N, Albert CM, Nagasubramanian R, Davis JL, Rudzinski E, Feraco AM, et al. Larotrectinib for paediatric solid tumours harbouring NTRK gene fusions: phase 1 results from a multicentre, open-label, phase 1/2 study. Lancet Oncol (2018) 19:705–714.

97. Kristensen H, Thomsen AR, Haldrup C, Dyrskjøt L, Høyer S, Borre M, Mouritzen P, Ørntoft TF, Sørensen KD. Novel diagnostic and prognostic classifiers for prostate cancer identified by genome-wide microRNA profiling. Oncotarget (2016) 7:30760–30771.

98. Matin F, Australian Prostate Cancer BioResource, Jeet V, Moya L, Selth LA, Chambers S, Clements JA, Batra J. A Plasma Biomarker Panel of Four MicroRNAs for the Diagnosis of Prostate Cancer. Scientific Reports (2018) 8: doi:10.1038/s41598-018-24424-w

Chapitre 3 : Discussion

Le CaP reste un problème de santé public majeur malgré les progrès effectués dans le diagnostic et le prognostic de la maladie au cours des dernières décennies. Une des raisons principales pour expliquer cela est la nature intrinsèque et la complexité du CaP. Jusqu'à maintenant, les meilleurs outils pour caractériser et prédire l'évolution des CaP ont été les caractéristiques clinico-pathologiques de base telles que le grade, le stade et le niveau de PSA. Malgré tout, ces critères ont une capacité limitée pour distinguer les différents CaP puisque des tumeurs ayant des caractéristiques clinico-pathologiques équivalentes présentent des comportements cliniques différents. L'avènement récent des techniques d'analyses génomiques a permis d'espérer une meilleure caractérisation des tumeurs de prostate. Par contre, ces analyses ont montré que la variabilité était encore plus grande qu'imaginée. Ainsi il apparaît clair qu'une meilleure compréhension et éventuellement catégorisation des tumeurs de prostate nécessitera des analyses multi-dimensionnelles faisant intervenir plusieurs caractéristiques ou jeux de données. Une autre avancée majeure des dernières années en oncologie est le développement des immunothérapies basées sur les points de contrôles immunologiques. La découverte de ces molécules capables d'inhiber la réponse immune (*e.g.* CTLA-4, PD-1 *etc*.) mais surtout la démonstration clinique que des inhibiteurs de ces molécules pouvaient réactiver la réponse immune anti-tumorale a été un point tournant dans la recherche en oncologie des dernières années. De nombreux patients atteints de cancer du poumon, de la peau, de la vessie et d'autres types de cancer ont pu profiter de ces découvertes. Cependant, les récents développement de l'immunothérapie ne bénéficient pas encore aux patients atteints du CaP. A ce jour, les approches

immunothérapeutiques basées sur les points de contrôle immunologiques les plus connus n'ont pas donné des résultats très encourageants De plus les interactions entre le système immunitaire et le cancer de la prostate ne sont pas très bien caractérisées. Une meilleure compréhension de ces interactions (réponse immune) pourrait permettre de mettre en évidence de nouveaux biomarqueurs prédictifs ou de nouvelles cibles pour l'immunothérapie du CaP.

Dans ce contexte ma thèse avait deux objectifs:

- d'une part d'utiliser différents jeux de données omiques pour identifier parmis des gènes de l'immunité de nouveaux marqueurs immunologiques pouvant être pertinents dans le traitement du CaP
- et d'autre part mettre en place des modèles prédictifs d'apprentissage machine pour prédire la BCR à partir de gènes de l'immunité.

Le premier objectif correspond au chapitre 1 de la thèse. Au travers d'une analyse multi-omique, j'ai pu faire ressortir une famille d'intérêt potentiellement utilisable dans des traitements d'immunothérapie. L'analyse multi-omique apporte des problèmes de validation des résultats dans le sens ou l'on a beaucoup de variables (expression de gènes, miRNA, CNA, méthylation) et peu de patients. Pour palier à ce problème nous avons d'abord sélectionné une famille à fort potentiel (LILR) puis nous avons vérifié son intérêt en effectuant une validation croisée sur 3 jeux de données RNA-Seq. Cela nous a permis de confirmer que la famille avait un intérêt et que ce n'était pas un artefact dû aux jeux de données multi-omiques.

Le problème majeur que j'ai pu voir à ce niveau est que les gènes de l'immunité sont en général peu exprimés puisque ceux-ci se retrouvent principalement dans les cellules immunes qui infiltrent les tumeurs. Leur détection est donc proportionnelle au niveau d'infiltration des tumeurs par les cellules immunes mais elle est aussi très dépendante de la qualité des données et de certains paramètres:

- Choix de la profondeur de séquençage : Une couverture de X100 semble être une valeur raisonnable car c'est à ce niveau que nous avons obtenu des comptes en milliers pour certains gènes de l'immunité qui nous intéressaient.
- Qualité du séquençage : La préparation des données de RNA-Seq au travers des étapes d'amplification ou de ribo déplétion impacte la proportion de comptes de gènes disponibles et donc influence les gènes à faible compte.
- La qualité de l'échantillon biologique : Le résultat du séquençage de l'ARN dépendant directement de la qualité du tissu. Un tissu qui n'a pas été congelé dans les minutes suivant sa résection peut être sujet à un niveau plus ou moins grand de dégradation de l'ARN ce qui peut avoir un impact important sur la détection des ARN exprimés à un plus faible niveau. De plus, un tissu qui contient une proportion plus ou moins grande de tumeurs pourrait avoir une proportion plus ou moins grande de cellules immunes infiltrant la tumeur. Ces facteurs pourraient tous jouer un rôle dans sur le niveau d'expression des gènes associés à l'immunité.

Dans la plupart des jeux de données que nous avons consultés, un bon nombre des gènes de l'immunité étaient enlevés de la matrice de comptes finaux tout simplement parce qu'ils ne passaient pas les filtres bioinformatiques avant même la partie analytique.

Dans le cadre du travail sur cet article nous nous étions initialement concentrés sur 3 familles de gènes : LILR, KIR[198–200] et HLA[201,202], tous des gènes associés au processus de présentation antigénique. Ces trois familles de gènes étaient celles qui étaient le plus représentées dans les résultats de l'analyse multi-omique. Après vérification, les gènes KIR et HLA ne présentaient pas des niveaux de BER suffisants pour que nous les présentions ou nous concentrions dessus. On parlera ici de tendance puisque ces familles sont liées. Les récepteurs LILR et KIR se lient aux molécules de présentation HLA. Dans ce contexte nous avions émis l'hypothèse que certaines mutations ou expression aberrantes dans les gènes HLA puissent favoriser ou non les

interactions avec les LILR ou KIR. Mais pour cela il aurait fallu pouvoir relier finement les types HLA avec les profils d'expression des gènes.

De façon générale il serait plus que nécessaire d'utiliser la protéomique pour caractériser l'expression des protéines LILR dans les tumeurs de prostate. Actuellement la technique de protéomique non ciblée ne permet pas d'étudier les protéines peu exprimées et il serait d'autant plus difficile de les détecter puisque ce sont des protéines transmembranaires. Mais il faut bien noter que comparativement au RNA-Seq, avec l'analyse de protéine on se rapproche plus de l'effet réel biologique si la protéine est observée . C'est un des problèmes inhérents à l'analyse RNA-Seq car on n'a pas d'information sur la régulation des protéines après l'étape de traduction. On n'a d'ailleurs pas réellement d'information sur la régulation de la traduction en tant que telle. Mais la protéomique est bien plus complexe à mettre en place comparativement au RNA-seq. Il faut donc effectuer des progrès techniques dans l'approche protéique avant de pouvoir l'utiliser en routine comparativement au RNA-Seq.

Concernant le deuxième objectif de ma thèse qui correspond au second article, il intègre les problèmes majeurs que j'ai rencontrés et qui sont liés à la disponibilités des données et leur annotation. De plus en plus les jeux de données sont disponibles et facilement récupérables. Mais il faut noter certaines limites :

- La qualité des échantillons biologiques: Tel que mentionné plus haut il peut y avoir une grande variabilité dans la qualité de l'échantillon biologique tant au niveau de l'intégrité biologique du tissus (dégradation) que de la qualité de l'échantillonnage (contenu de la zone prélevée). Malheureusement, il n'y a pas à ma connaissance de protocole normalisé de biobanking dans le cancer de la prostate qui permettrait de réduire cette variabilité d'un établissement à un autre. La variabilité de la qualité des tissus mis en banque n'est pas une particularité du cancer de la

prostate. C'est un problème commun à plusieurs types de biobanques de tissus biologiques[203].

- Choix du type de tissus : Devant la difficulté à trouver des échantillons congelés de qualité pour effectuer les analyses de séquençage, certains groupes se sont tournés vers l'utilisation de tissus fixés et inclus en paraffine comme choix de matériel de départ. Cependant une cryo-conservation dans de l'azote liquide et une fixation en formalin et un enrobage dans la paraffine n'ont pas le même effet physique sur le tissu et notamment sur les molécules d'ARN, induisant ainsi une variabilité au niveau de la performance pour la détection des ARN moins abondants. Cependant des analyses suggèrent que l'utilisation des échantillons FFPE peut être une alternative valable à condition d'ajuster le niveau de séquençage[204].

- La bioinformatique : Le pipeline de préparation au niveau bioinformatique influence énormément les valeurs finales des entités biologiques. Même si conceptuellement les outils font en principe la même chose (*e.g*. les outils d'alignement alignent tous), les choix des paramètres des algorithmes peuvent considérablement influencer les résultats.

- Qualité des données cliniques : Tous les modèles que nous mettons en place sont reliés à des réponses cliniques (e.g. Récidive, accident, événement etc). Les données cliniques sont donc essentielles au risque d'avoir des résultats qui n'auront aucun sens sans que personne ne s'en rende compte. Cette qualité initiale dépend des questionnaires donnés aux patients, ceux remplis par les équipes de soin et le temps pris pour analyser, comprendre et contrôler ces informations.

La limitation ne vient donc pas de la partie apprentissage machine ou les algorithmes disponibles sont puissants et la méthodologie bien connue mais des données elles-mêmes. Comprendre les données, le problème et le contexte sont les éléments majeurs pour qu'un algorithme prédictif soit pertinent.

D'ailleurs dans le cadre de notre article nos résultats en termes de performance sont globalement équivalents aux autres travaux fait sur le sujet. La réelle différence vient du fait que nous avons montré qu'en mélangeant des jeux de données provenant d'équipes différentes on pouvait tout de même avoir une prédiction stable de la BCR dans le CaP. Cela implique qu'il est possible de grouper tous les jeux de données disponibles si les métadonnées cliniques sont à jour au niveau de la BCR (il faut un suivi clinique de plusieurs années pour identifier avec assez de précision tous les événements de récidive biochimique). On peut donc tirer profit de tous les travaux déjà effectués en la matière et surtout à l'avenir il sera possible de faire ce type de travail sur des bases de données nationales mise à jour en continu. On aura donc un potentiel de diagnostic très puissant. On pourrait même pousser la logique en supposant que les modèles pourraient être entraînés pour un individu particulier à chaque fois qu'il est nécessaire.

Après tout ça se pose la question du RNA-Seq. Est-ce que le RNA-Seq est le type de données le plus pertinent pour faire de la recherche de cibles potentielles immunitaires ou de l'apprentissage machine?

Le RNA-Seq n'est probablement pas meilleur ou pire que n'importe quel autre type de données biologiques. En soit, il contient une information qui n'est pas indépendante des autres données (mutation, protéomique, ...) et donc si l'on voulait être rigoureux, il faudrait faire des analyses multi-omiques pour chaque projet. Mais à l'heure actuelle ce n'est pas possible financièrement. Par exemple dans le CaP le seul projet multi-omique réel, avec 5 types de omiques différents, est le projet TCGA. Et c'est un projet international étalé sur 10 ans qui a coûté plusieurs dizaines de millions de dollars.

Dans ce contexte, le RNA-Seq est la meilleure alternative sur laquelle se concentrer pour plusieurs raisons. L'information qu'il contient reste pertinente même si elle ne représente pas l'ensemble des processus biologiques qui amènent à une maladie.

Les données de mutations sont puissantes mais nécessitent une grande profondeur de séquençage pour être étudiées et sont donc chères. De plus, 15 ans de recherche sur les *Genome Wide Association Studies* (GWAS) n'ont donné que très peu de résultats significatifs au niveau des applications cliniques. En cancer on utilise principalement les mutations pour définir des niveaux de risque, i.e. un potentiel d'aller vers un état mais pas pour le diagnostic réel.

Pour les données de méthylation la plupart des gens considèrent que l'on pourra voir leurs effets au niveau de l'expression des ARN. Il est alors plus utile d'analyser directement l'ARN.

Les CNA sont très importants dans le cancer mais encore une fois cela coûte cher car il faut faire de WGS pour avoir une vue d'ensemble.

Les miRNA passent par la même technologie que le RNA-Seq et ainsi ils commencent à être étudiés ensemble. Le désavantage est que ce sont des molécules très instables. Il faut donc encore faire de la recherche au niveau des protocoles d'analyse.

In fine, c'est le rapport information/coût qui prime et à ce niveau le RNA-Seq reste en tête. De plus avec la possibilité d'intégrer différents types d'ARN (miRNA, cRNA,etc.) cette technologie a encore un potentiel inexploité. D'ailleurs si l'on regarde les récents développements industriels notamment au niveau des technologies de séquençages portables comme le MinION de la compagnie Oxford Nanopore Technologies on voit que les industriels se concentrent sur le RNA-Seq.

Chapitre 4 : Perspective

Dans le cadre du développement des traitements d'immunothérapie, il y a un besoin de tester de nouvelles molécules. L'association de la famille des LIL avec la BCR dans le CaP montre que cette famille de récepteurs pourrait être ciblée seule ou en combinaison avec d'autres points de contrôle immunologique. La famille des LILR se retrouve principalement sur les cellules d'origine myéloïdes, ils sont donc liés à des types de tumeurs spécifiques. Ce faisant, la connaissance histologique des tumeurs (Grade, stade) reste nécessaire pour pouvoir orienter les patients vers des traitements personnalisés.

La famille des LILR est peu connue de façon générale et comme pour beaucoup de mécanismes immunitaires le fonctionnement et le rôle des ces gènes n'est pas parfaitement connu. Il existe un homologue de LILRB1 chez la souris appelé PIR-B qui peut permettre d'en apprendre plus sur le rôle de LILRB1 dans les cancers humains. Comme expliqué en introduction, on voit de plus en plus de travaux sur cette famille et elle commence à être intégrée comme molécule à tester. Dans ce cadre, il serait intéressant de pouvoir lier les séquences HLA des patients avec la famille des LILR et KIR. Cela nécessiterait un séquençage avec une grande profondeur pour pouvoir définir précisément les motifs HLA associés à des caractéristiques cliniques tout en analysant l'expression des LILR et KIR. Cela pourrait permettre de préciser les mécanismes liés aux cellules immunes infiltrantes et le polymorphisme génétique des molécules HLA.

Une question que je n'ai jamais vu abordée dans la littérature est la question de la temporalité de l'expression des LILR. Le RNA-Seq correspond à un instant fixe de l'état d'une tumeur d'un patient. Mais la maladie passe par différentes phases de développement et on ne sait pas comment la famille des LILR s'insère dans ce processus. Il serait donc potentiellement pertinent de suivre des patients tout au long de la maladie pour mesurer le taux d'expression

des LILR sur les cellules immunes en fonction de la gravité du cancer dans le temps. Ce problème rejoint une approche plus globale de l'immunothérapie et de son efficacité.

Dans le cadre de l'application des approches d'apprentissage machine, la limitation actuelle provient de notre capacité à produire des données de qualité. Côté théorique nous avons des algorithmes prédictifs très puissants et une méthodologie associée globalement consensuelle qui a montré son efficacité sur des grands jeux de données. La recherche actuelle est faite non pas sur les modèles en eux-mêmes mais plutôt leur interprétabilité. Dans ce cas là on parle surtout des approches dites de *deep learning* qui correspondent à des réseaux de neurones plus ou moins complexes. Prédire est facile, comprendre est plus complexe dans le cadre des problèmes en grande dimension.

Les points importants à noter pour nous sont donc tout autres. Il y a 3 leviers d'activation réellement important pour nous au delà des modèles:

- La masse des données, sa volumétrie
- La qualité des données, leur annotation
- Leur organisation, disponibilité
- Transparence et reproductibilité

On peut donc se poser la question de comment récupérer plus de données et des données de qualité. Au niveau omique on pourrait voir arriver de plus en plus de données de séquençage comme résultat d'un examen de routine ce qui permettrait d'augmenter le volume de données disponibles. Mais pour ça il y a un temps d'appropriation de ces technologies par le corps médical qui sont pour l'instant plus utilisées par la recherche.

Annoter les données avec des informations est très important puisqu'en analyse de données on peut toujours avoir des résultats quelque soit la qualité des données. Il y a donc un travail et un engagement nécessaire de la

part du personnel de soin. Ce sont eux qui sont en première ligne pour récupérer l'information nécessaire et la trier en premier lieu.

En association avec le volume de données il faut une politique d'ingénierie des données. Avoir beaucoup de données ne signifie pas qu'elles sont facilement accessibles et disponibles. On n'interroge pas des To de données comme on ouvre un fichier texte. Le métier d'ingénieur de données est donc essentiel dans des projets impliquant les données omiques.

En dernier lieu la transparence des données est nécessaire et "*l'open data"* est une solution toute indiquée. Il faut donc en tant que chercheur favorisé les journaux privilégiant ce modèle. La question de la reproductibilité peut-être résolue en demandant des environnements de développement complets sous forme de machine virtuelle ou de conteneurs. Cela nécessite un travail supplémentaire mais conduirait grandement à améliorer la qualité de la recherche. Au minimum un *Github* contenant le code devrait être proposé.

Conclusion

Notre analyse multi-omique a permis de faire ressortir un groupe de variables biologiques pertinentes dont la famille des gènes LILR et plus particulièrement le gène LILRB1. Ce gène fait partie des molécules potentiellement utilisables dans le contexte du développement de l'immunothérapie du CaP et est donc très prometteur.

Nous avons aussi montré qu'il était possible de regrouper des petits (<100 patients) jeux de données RNA-Seq de sources différentes et de les analyser en bloc avec la méthodologie appropriée pour obtenir une signature prédictive de la BCR. Il n'est donc pas forcément nécessaire de dépenser beaucoup pour constituer des grosses cohortes alors qu'il est possible de constituer des bases de données à partir de multiples petits jeux provenant de sources différentes.

f you read this thesis send me an email ti bv;research@protonmail we'll try to organise a coffee in Paris ! I'll be please to meet you.

Matériel supplémentaire

Assemblage de métagénome	MetaQuast[207], MetaVelvet[208], MetaVelvet-SL[209], IDBA-UD[210], Meta Ray[211], MetAMOS[212]
Groupes et intervalle phylogénétiques	Phymm[213], SCIMM[214], MetaWatt[215],Concoct[216]
Prédiction de gène au niveau métagénomique	MetaGeneAnnotator[217],Orphelia[218], Glimmer-MG[219], FragGeneScan[220], Prokka[221], GeneMark[222], MetaGeneMark[223], GenScan[224]
Bases de données protéiques	InterPro[225], InterProScan[226]
Sentiers biologiques et associations	Reactome[227], UniProt[228], KEGG[229], WikiPathways[230]}, MetaCyc[231], STITCH[232]
Gène spécifique	Xander[233]
Partage de données et accès en ligne	Meta4[234], MG-Rast[235], EBI Metagenomics[236], IMG/M[237], EDGE[238]

Matériel supplémentaire 1. Outils bioinformatiques : Liste des outils d'analyses de bioinformatique au moment de mon travail de préparation des données.

Méthode	Description	Package R
PCA	Analyse en composante principale	stats, ade4, vegan, factomineR, psych
CA, COA	Analyse de correspondance	CA, FactomineR,ade4
NSC	Analyse de correspondance non symétrique	ade4, mixOmic
MDS	Positionnement multidimensionnel	ade4, stats, ape
NMF	Factorisation non négative matricielle	nmf
sPCA, nsPCA, pPCA	PCA avec sélection de variables	PMA, mixOmics, nsprcomp
NIPALS PCA	PCA pour données manquantes	ade4, pcaMethods, mixOmics
bPCA	PCA bayesienne	pcaMethods
MCA	Analyse de correspondance multiple	ade4, MASS
ICA	Analyse de composantes indépendantes	FastICA
sIPCA	combine sPCA et ICA	mixOmics

Matériel supplémentaire 2. ***Réduction de dimension - 1 :*** Méthodes de réduction de dimension pour un jeu de données seul.

Méthode	Description	Sélection de variables	Package R
CCA	Analyse canonique de corrélation (n>p)	Non	cca
CCA	Analyse canonique de correspondance	Non	cca, vegan
RDA	PCA redondante	Non	vegan
Procrutes	similitude de distance	Non	ade4, vegan
rCCA	corrélation canonique régularisée	Non	cca
sCCA	sparse CCA	Oui	pma
pCCA	CCA pénalisée	Oui	spCCA
WAPLS	Régression des moindres carrés balancée	Non	rioja,paltran
PLS	multi-blocks PLS	Oui	spls, mixOmics, caret
sPLS	sparse PLS	Oui	mixOmics
sPLS-DA	sPLS Discriminant Analysis	Oui	mixOmics
cPCA	analyse de co-inertie	Non	mogsa
CIA	Analyse de correspondance symetrique	Non	ade4, made4

Matériel supplémentaire 3***. Réduction de dimension - 2 :*** Méthodes de réduction de dimension pour des jeux de données en pair.

Méthode	Description	Sélection de variables	Package R
MCIA	Analyse de co-inertie multiple	Non	omicade4, ade4
gCCA	CCA généralisée	Non	dmt
rGCCA	gCCA généralisée régularisée	Non	dmt, rgcca
sGCCA	rGCCA sparse	Oui	rgcca, mixOmics
STATIS	Structuratio des Tableaux à Trois indices de la statistique	Non	ade4
candecomp	Higher order generalizations of SVD et PCA	Non	ThreeWay, PTak
PTA	analyse triadique partielle	Non	ade4
statico	statis + CIA	Non	ade4

Matériel supplémentaire 3. Réduction de dimension - 3 : Méthodes de réduction de dimension pour des jeux de données multiples.

Bibliographie

1. Strouhal, E. & Nemeckova, A. Paleopathological find of a sacral neurilemmoma from ancient Egypt. *Am. J. Phys. Anthropol.* **125**, 320–328 (2004).
2. David, A. R. & Zimmerman, M. R. Cancer: an old disease, a new disease or something in between ? *Nat. Rev. Cancer* **10**, 728–733 (2010).
3. Marx, F. J. & Karenberg, A. History of the Term Prostate. *The Prostate* **69**, 208–213 (2009).
4. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA. Cancer J. Clin.* **68**, 7–30 (2018).
5. Saad, F., McCormack, M. & McCormack, M. *Le cancer de la prostate*. (Annika Parance, 2012).
6. Giovannucci, E. *et al.* Calcium and fructose intake in relation to risk of prostate cancer. *Cancer Res.* **58**, 442–447 (1998).
7. Bostwick, D. G. *et al.* Human prostate cancer risk factors. *Cancer* **101**, 2371–2490 (2004).
8. Scherr, D., Swindle, P. W. & Scardino, P. T. National Comprehensive Cancer Network guidelines for the management of prostate cancer. *Urology* **61**, 14–24 (2003).
9. Pritchard, C. C. *et al.* Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. *N. Engl. J. Med.* **375**, 443–453 (2016).
10. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer

heritability. *Nat. Genet.* **48**, 30–35 (2016).

11. Hjelmborg, J. B. *et al.* The Heritability of Prostate Cancer in the Nordic Twin Study of Cancer. *Cancer Epidemiol. Prev. Biomark.* cebp.0568.2013 (2014) doi:10.1158/1055-9965.EPI-13-0568.

12. Fujita, K., Hayashi, T., Matsushita, M., Uemura, M. & Nonomura, N. Obesity, Inflammation, and Prostate Cancer. *J. Clin. Med.* **8**, 201 (2019).

13. Porter, C. M., Shrestha, E., Peiffer, L. B. & Sfanos, K. S. The microbiome in prostate inflammation and prostate cancer. *Prostate Cancer Prostatic Dis.* **21**, 345–354 (2018).

14. Yap, T. A. *et al.* Drug discovery in advanced prostate cancer: translating biology into therapy. *Nat. Rev. Drug Discov.* **15**, 699–718 (2016).

15. Keyes, M. *et al.* Treatment options for localized prostate cancer. *Can. Fam. Physician* **59**, 1269–1274 (2013).

16. Hoffman, K. E. *et al.* Physician variation in management of low-risk prostate cancer: a population-based cohort study. *JAMA Intern. Med.* **174**, 1450–1459 (2014).

17. Trinh, Q.-D. & Schrag, D. Measuring the effectiveness of androgen-deprivation therapy for prostate cancer in the medicare population: adequate data are neither the same as nor the enemy of perfect data. *JAMA Intern. Med.* **174**, 1468–1469 (2014).

18. Lu-Yao, G. L. *et al.* Fifteen-Year Survival Outcomes Following Primary Androgen-Deprivation Therapy for Localized Prostate Cancer. *JAMA Intern. Med.* **174**, 1460–1467 (2014).

19. Francini, E. & Sweeney, C. J. Docetaxel Activity in the Era of

Life-prolonging Hormonal Therapies for Metastatic Castration-resistant Prostate Cancer. *Eur. Urol.* **70**, 410–412 (2016).

20. Martin, S. K. & Kyprianou, N. Exploitation of the Androgen Receptor to Overcome Taxane Resistance in Advanced Prostate Cancer. *Adv. Cancer Res.* **127**, 123–158 (2015).

21. van Soest, R. J. & de Wit, R. Irrefutable evidence for the use of docetaxel in newly diagnosed metastatic prostate cancer: results from the STAMPEDE and CHAARTED trials. *BMC Med.* **13**, 304 (2015).

22. Sartor, O. & de Bono, J. S. Metastatic Prostate Cancer. *N. Engl. J. Med.* **378**, 645–657 (2018).

23. Dellis, A. *et al.* Management of advanced prostate cancer: A systematic review of existing guidelines and recommendations. *Cancer Treat. Rev.* **73**, 54–61 (2019).

24. Henriksen, G., Breistøl, K., Bruland, Ø. S., Fodstad, Ø. & Larsen, R. H. Significant Antitumor Effect from Bone-seeking, α-Particle-emitting 223Ra Demonstrated in an Experimental Skeletal Metastases Model. *Cancer Res.* **62**, 3120–3125 (2002).

25. *AJCC cancer staging manual*. (Springer, 2010).

26. Humphrey, P. A., Moch, H., Cubilla, A. L., Ulbright, T. M. & Reuter, V. E. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs—Part B: Prostate and Bladder Tumours. *Eur. Urol.* **70**, 106–119 (2016).

27. Inamura, K. Prostatic cancers: understanding their molecular pathology and the 2016 WHO classification. *Oncotarget* **9**, 14723–14737 (2018).

28. Epstein, J. I. *et al.* The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *Am. J. Surg. Pathol.* **40**, 244–252 (2016).

29. Bussemakers, M. J. *et al.* DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* **59**, 5975–5979 (1999).

30. de Kok, J. B. *et al.* DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. *Cancer Res.* **62**, 2695–2698 (2002).

31. Hessels, D. *et al.* DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. *Eur. Urol.* **44**, 8–15; discussion 15-16 (2003).

32. Schalken, J. A., Hessels, D. & Verhaegh, G. New targets for therapy in prostate cancer: differential display code 3 (DD3(PCA3)), a highly prostate cancer-specific gene. *Urology* **62**, 34–43 (2003).

33. Groskopf, J. *et al.* APTIMA PCA3 molecular urine test: development of a method to aid in the diagnosis of prostate cancer. *Clin. Chem.* **52**, 1089–1095 (2006).

34. Hessels, D. & Schalken, J. A. The use of *PCA3* in the diagnosis of prostate cancer. *Nat. Rev. Urol.* **6**, 255–261 (2009).

35. Dai, C., Heemers, H. & Sharifi, N. Androgen Signaling in Prostate Cancer. *Cold Spring Harb. Perspect. Med.* **7**, (2017).

36. Nadal, M. *et al.* Structure of the homodimeric androgen receptor ligand-binding domain. *Nat. Commun.* **8**, 14388 (2017).

37. van Royen, M. E., van Cappellen, W. A., de Vos, C., Houtsmuller, A. B. & Trapman, J. Stepwise androgen receptor dimerization. *J. Cell Sci.* **125**,

1970–1979 (2012).

38. Grosse, A., Bartsch, S. & Baniahmad, A. Androgen receptor-mediated gene repression. *Mol. Cell. Endocrinol.* **352**, 46–56 (2012).

39. van der Steen, T., Tindall, D. J. & Huang, H. Posttranslational modification of the androgen receptor in prostate cancer. *Int. J. Mol. Sci.* **14**, 14833–14859 (2013).

40. Faus, H. & Haendler, B. Post-translational modifications of steroid receptors. *Biomed. Pharmacother. Biomedecine Pharmacother.* **60**, 520–528 (2006).

41. Wadosky, K. M. & Koochekpour, S. Molecular mechanisms underlying resistance to androgen deprivation therapy in prostate cancer. *Oncotarget* **7**, 64447–64470 (2016).

42. De Maeseneer, D. J., Van Praet, C., Lumen, N. & Rottey, S. Battling resistance mechanisms in antihormonal prostate cancer treatment: Novel agents and combinations. *Urol. Oncol.* **33**, 310–321 (2015).

43. James, N. D., Spears, M. R. & Sydes, M. R. Abiraterone in Metastatic Prostate Cancer. *N. Engl. J. Med.* **377**, 1696–1697 (2017).

44. Smith, M. R. *et al.* Apalutamide Treatment and Metastasis-free Survival in Prostate Cancer. *N. Engl. J. Med.* **378**, 1408–1418 (2018).

45. Rathkopf, D. E. *et al.* Safety and Antitumor Activity of Apalutamide (ARN-509) in Metastatic Castration-Resistant Prostate Cancer with and without Prior Abiraterone Acetate and Prednisone. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **23**, 3544–3551 (2017).

46. Shore, N. D. Darolutamide (ODM-201) for the treatment of prostate

cancer. *Expert Opin. Pharmacother.* **18**, 945–952 (2017).

47. Fizazi, K., Albiges, L., Loriot, Y. & Massard, C. ODM-201: a new-generation androgen receptor inhibitor in castration-resistant prostate cancer. *Expert Rev. Anticancer Ther.* **15**, 1007–1017 (2015).

48. Borgmann, H. *et al.* Moving Towards Precision Urologic Oncology: Targeting Enzalutamide-resistant Prostate Cancer and Mutated Forms of the Androgen Receptor Using the Novel Inhibitor Darolutamide (ODM-201). *Eur. Urol.* **73**, 4–8 (2018).

49. Moilanen, A.-M. *et al.* Discovery of ODM-201, a new-generation androgen receptor inhibitor targeting resistance mechanisms to androgen signaling-directed prostate cancer therapies. *Sci. Rep.* **5**, 12007 (2015).

50. Crumbaker, M., Khoja, L. & Joshua, A. M. AR Signaling and the PI3K Pathway in Prostate Cancer. *Cancers* **9**, (2017).

51. Wise, H. M., Hermida, M. A. & Leslie, N. R. Prostate cancer, PI3K, PTEN and prognosis. *Clin. Sci. Lond. Engl. 1979* **131**, 197–210 (2017).

52. Jamaspishvili, T. *et al.* Clinical implications of PTEN loss in prostate cancer. *Nat. Rev. Urol.* **15**, 222–234 (2018).

53. Wang, S. *et al.* Prostate-specific deletion of the murine Pten tumor suppressor gene leads to metastatic prostate cancer. *Cancer Cell* **4**, 209–221 (2003).

54. Gallick, G. E., Corn, P. G., Zurita, A. J. & Lin, S.-H. Small-molecule protein tyrosine kinase inhibitors for the treatment of metastatic prostate cancer. *Future Med. Chem.* **4**, 107–119 (2012).

55. Corn, P. G., Wang, F., McKeehan, W. L. & Navone, N. Targeting

fibroblast growth factor pathways in prostate cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **19**, 5856–5866 (2013).

56. Choi, Y. J. *et al.* Phase II Study of Dovitinib in Patients with Castration-Resistant Prostate Cancer (KCSG-GU11-05). *Cancer Res. Treat. Off. J. Korean Cancer Assoc.* **50**, 1252–1259 (2018).

57. Schneider, J. A. & Logan, S. K. Revisiting the role of Wnt/β-catenin signaling in prostate cancer. *Mol. Cell. Endocrinol.* **462**, 3–8 (2018).

58. Syed Khaja, A. S. *et al.* Elevated level of Wnt5a protein in localized prostate cancer tissue is associated with better outcome. *PloS One* **6**, e26539 (2011).

59. Canesin, G. *et al.* Treatment with the WNT5A-mimicking peptide Foxy-5 effectively reduces the metastatic spread of WNT5A-low prostate cancer cells in an orthotopic mouse model. *PloS One* **12**, e0184418 (2017).

60. Jefferies, M. T. *et al.* PTEN loss and activation of K-RAS and β-catenin cooperate to accelerate prostate tumourigenesis. *J. Pathol.* **243**, 442–456 (2017).

61. Schweizer, M. T. & Antonarakis, E. S. Prognostic and therapeutic implications of DNA repair gene mutations in advanced prostate cancer. *Clin. Adv. Hematol. Oncol. HO* **15**, 785–795 (2017).

62. Ramakrishnan Geethakumari, P., Schiewer, M. J., Knudsen, K. E. & Kelly, W. K. PARP Inhibitors in Prostate Cancer. *Curr. Treat. Options Oncol.* **18**, 37 (2017).

63. Liu, X., Wu, Q. & Li, L. Functional and therapeutic significance of EZH2 in urological cancers. *Oncotarget* **8**, 38044–38055 (2017).

64. Xu, K. *et al.* EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* **338**, 1465–1469 (2012).

65. Malik, R. *et al.* Targeting the MLL complex in castration-resistant prostate cancer. *Nat. Med.* **21**, 344–352 (2015).

66. Metzger, E. *et al.* LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature* **437**, 436–439 (2005).

67. Zuber, V. *et al.* Bromodomain protein 4 discriminates tissue-specific super-enhancers containing disease-specific susceptibility loci in prostate and breast cancer. *BMC Genomics* **18**, 270 (2017).

68. Jung, M., Gelato, K. A., Fernández-Montalván, A., Siegel, S. & Haendler, B. Targeting BET bromodomains for cancer treatment. *Epigenomics* **7**, 487–501 (2015).

69. DiPippo, V. A. *et al.* Addition of PSMA ADC to enzalutamide therapy significantly improves survival in in vivo model of castration resistant prostate cancer. *The Prostate* **76**, 325–334 (2016).

70. Olson, W. Antibody-drug conjugates targeting prostate-specific membrane antigen. *Front. Biosci.* **19**, 12 (2014).

71. Ferdinandus, J., Violet, J., Sandhu, S. & Hofman, M. S. Prostate-specific membrane antigen theranostics: therapy with lutetium-177. *Curr. Opin. Urol.* **28**, 197–204 (2018).

72. Kratochwil, C. *et al.* Targeted α-Therapy of Metastatic Castration-Resistant Prostate Cancer with 225Ac-PSMA-617: Swimmer-Plot Analysis Suggests Efficacy Regarding Duration of Tumor Control. *J. Nucl.*

Med. Off. Publ. Soc. Nucl. Med. **59**, 795–802 (2018).

73. Hammer, S. *et al.* Abstract 5200: Preclinical pharmacology of the PSMA-targeted thorium-227 conjugate PSMA-TTC: a novel targeted alpha therapeutic for the treatment of prostate cancer. *Cancer Res.* **77**, 5200–5200 (2017).

74. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).

75. Gasi Tandefelt, D., Boormans, J., Hermans, K. & Trapman, J. ETS fusion genes in prostate cancer. *Endocr. Relat. Cancer* **21**, R143-152 (2014).

76. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).

77. Kim, S. H. *et al.* Overexpression of ERG and Wild-Type PTEN Are Associated with Favorable Clinical Prognosis and Low Biochemical Recurrence in Prostate Cancer. *PloS One* **10**, e0122498 (2015).

78. Abeshouse, A. *et al.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).

79. Fraser, M. *et al.* Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364 (2017).

80. Nicholson, L. B. The immune system. *Essays Biochem.* **60**, 275–301 (2016).

81. Parham, P. *The Immune System, Fourth Edition*. (Garland Science, 2014).

82. Finn, O. J. Immuno-oncology: understanding the function and dysfunction of the immune system in cancer. *Ann. Oncol.* **23**, viii6–viii9

(2012).

83. O'Donnell, J. S., Teng, M. W. L. & Smyth, M. J. Cancer immunoediting and resistance to T cell-based immunotherapy. *Nat. Rev. Clin. Oncol.* **16**, 151–167 (2019).

84. Efremova, M. *et al.* Targeting immune checkpoints potentiates immunoediting and changes the dynamics of tumor evolution. *Nat. Commun.* **9**, 32 (2018).

85. Dunn, G. P., Old, L. J. & Schreiber, R. D. The Immunobiology of Cancer Immunosurveillance and Immunoediting. *Immunity* **21**, 137–148 (2004).

86. Waldman, A. D., Fritz, J. M. & Lenardo, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat. Rev. Immunol.* 1–18 (2020) doi:10.1038/s41577-020-0306-5.

87. Kesavardhana, S. & Kanneganti, T.-D. Mechanisms governing inflammasome activation, assembly and pyroptosis induction. *Int. Immunol.* **29**, 201–210 (2017).

88. Schroder, K. & Tschopp, J. The inflammasomes. *Cell* **140**, 821–832 (2010).

89. Man, S. M. Inflammasomes in the gastrointestinal tract: infection, cancer and gut microbiota homeostasis. *Nat. Rev. Gastroenterol. Hepatol.* **15**, 721–737 (2018).

90. Gültekin, Y., Eren, E. & Özören, N. Overexpressed NLRC3 acts as an anti-inflammatory cytosolic protein. *J. Innate Immun.* **7**, 25–36 (2015).

91. Davis, B. K. *et al.* Cutting edge: NLRC5-dependent activation of the inflammasome. *J. Immunol. Baltim. Md 1950* **186**, 1333–1337 (2011).

92. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, inflammation, and cancer. *Cell* **140**, 883–899 (2010).

93. Karki, R. & Kanneganti, T.-D. Diverging inflammasome signals in tumorigenesis and potential targeting. *Nat. Rev. Cancer* **19**, 197–214 (2019).

94. Terme, M. *et al.* IL-18 induces PD-1-dependent immunosuppression in cancer. *Cancer Res.* **71**, 5393–5399 (2011).

95. Kang, J. S. *et al.* Interleukin-18 increases metastasis and immune escape of stomach cancer via the downregulation of CD70 and maintenance of CD44. *Carcinogenesis* **30**, 1987–1996 (2009).

96. Tu, S. *et al.* Overexpression of interleukin-1beta induces gastric inflammation and cancer and mobilizes myeloid-derived suppressor cells in mice. *Cancer Cell* **14**, 408–419 (2008).

97. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* 1 (2018) doi:10.1038/s41591-018-0014-x.

98. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

99. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, (2017).

100. Bindea, G. *et al.* Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* **39**, 782–795 (2013).

101. Wherry, E. J. & Kurachi, M. Molecular and cellular insights into T cell exhaustion. *Nat. Rev. Immunol.* **15**, 486–499 (2015).

102. Pauken, K. E. & Wherry, E. J. Overcoming T cell exhaustion in infection and cancer. *Trends Immunol.* **36**, 265–276 (2015).

103. Farkona, S., Diamandis, E. P. & Blasutig, I. M. Cancer immunotherapy: the beginning of the end of cancer? *BMC Med.* **14**, 73 (2016).

104. Andtbacka, R. H. I. *et al.* Talimogene Laherparepvec Improves Durable Response Rate in Patients With Advanced Melanoma. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **33**, 2780–2788 (2015).

105. Jackson, H. J., Rafiq, S. & Brentjens, R. J. Driving CAR T-cells forward. *Nat. Rev. Clin. Oncol.* **13**, 370–383 (2016).

106. Brunner-Weinzierl, M. C. & Rudd, C. E. CTLA-4 and PD-1 Control of T-Cell Motility and Migration: Implications for Tumor Immunotherapy. *Front. Immunol.* **9**, (2018).

107. Kwek, S. S., Cha, E. & Fong, L. Unmasking the immune recognition of prostate cancer with CTLA4 blockade. *Nat. Rev. Cancer* **12**, 289–297 (2012).

108. Hoos, A. Development of immuno-oncology drugs [mdash] from CTLA4 to PD1 to the next generations. *Nat. Rev. Drug Discov.* **15**, 235–247 (2016).

109. Rowshanravan, B., Halliday, N. & Sansom, D. M. CTLA-4: a moving target in immunotherapy. *Blood* **131**, 58–67 (2018).

110. Isaacsson Velho, P. & Antonarakis, E. S. PD-1/PD-L1 pathway inhibitors in advanced prostate cancer. *Expert Rev. Clin. Pharmacol.* **11**, 475–486 (2018).

111. Gong, J., Chehrazi-Raffle, A., Reddi, S. & Salgia, R. Development of PD-1 and PD-L1 inhibitors as a form of cancer immunotherapy: a comprehensive review of registration trials and future considerations. *J.*

Immunother. Cancer **6**, 8 (2018).

112. Ruffo, E., Wu, R. C., Bruno, T. C., Workman, C. J. & Vignali, D. A. A. Lymphocyte-activation gene 3 (LAG3): The next immune checkpoint receptor. *Semin. Immunol.* **42**, 101305 (2019).

113. Aspeslagh, S. *et al.* Rationale for anti-OX40 cancer immunotherapy. *Eur. J. Cancer Oxf. Engl. 1990* **52**, 50–66 (2016).

114. Curti, B. D. *et al.* OX40 is a potent immune-stimulating target in late-stage cancer patients. *Cancer Res.* **73**, 7189–7198 (2013).

115. Hirayasu, K. & Arase, H. Functional and genetic diversity of leukocyte immunoglobulin-like receptor and implication for disease associations. *J. Hum. Genet.* **60**, 703–708 (2015).

116. Brown, D., Trowsdale, J. & Allen, R. The LILR family: modulators of innate and adaptive immune pathways in health and disease. *Tissue Antigens* **64**, 215–225 (2004).

117. Hirayasu, K. & Arase, H. Leukocyte Immunoglobulin-Like Receptor (LILR). in *Encyclopedia of Signaling Molecules* (ed. Choi, S.) 1–8 (Springer New York, 2016). doi:10.1007/978-1-4614-6438-9_101689-1.

118. Arase, H. & Lanier, L. L. Specific recognition of virus-infected cells by paired NK receptors. *Rev. Med. Virol.* **14**, 83–93 (2004).

119. Jones, D. C. *et al.* Alternative mRNA splicing creates transcripts encoding soluble proteins from most LILR genes. *Eur. J. Immunol.* **39**, 3195–3206 (2009).

120. Willcox, B. E., Thomas, L. M. & Bjorkman, P. J. Crystal structure of HLA-A2 bound to LIR-1, a host and viral major histocompatibility complex

receptor. *Nat. Immunol.* **4**, 913–919 (2003).

121. Davidson, C. L., Li, N. L. & Burshtyn, D. N. LILRB1 polymorphism and surface phenotypes of natural killer cells. *Hum. Immunol.* **71**, 942–949 (2010).

122. An, H. *et al.* Soluble LILRA3, a Potential Natural Antiinflammatory Protein, Is Increased in Patients with Rheumatoid Arthritis and Is Tightly Regulated by Interleukin 10, Tumor Necrosis Factor-α, and Interferon-γ. *J. Rheumatol.* **37**, 1596–1606 (2010).

123. Norman, P. J., Carey, B. S., Stephens, H. A. F. & Vaughan, R. W. DNA sequence variation and molecular genotyping of natural killer leukocyte immunoglobulin-like receptor, LILRA3. *Immunogenetics* **55**, 165–171 (2003).

124. Pasero, C. *et al.* Inherent and Tumor-Driven Immune Tolerance in the Prostate Microenvironment Impairs Natural Killer Cell Antitumor Activity. *Cancer Res.* **76**, 2153–2165 (2016).

125. Barkal, A. A. *et al.* Engagement of MHC class I by the inhibitory receptor LILRB1 suppresses macrophages and is a target of cancer immunotherapy. *Nat. Immunol.* **19**, 76–84 (2018).

126. Fesnak, A. D., June, C. H. & Levine, B. L. Engineered T cells: the promise and challenges of cancer immunotherapy. *Nat. Rev. Cancer* **16**, 566–581 (2016).

127. Marshall, H. T. & Djamgoz, M. B. A. Immuno-Oncology: Emerging Targets and Combination Therapies. *Front. Oncol.* **8**, (2018).

128. Pollard, M. E., Moskowitz, A. J., Diefenbach, M. A. & Hall, S. J. Cost-effectiveness analysis of treatments for metastatic castration resistant

prostate cancer. *Asian J. Urol.* **4**, 37–43 (2017).

129. Ebelt, K. *et al.* Prostate cancer lesions are surrounded by FOXP3+, PD-1+ and B7-H1+ lymphocyte clusters. *Eur. J. Cancer Oxf. Engl. 1990* **45**, 1664–1672 (2009).

130. Ammirante, M., Luo, J.-L., Grivennikov, S., Nedospasov, S. & Karin, M. B-cell-derived lymphotoxin promotes castration-resistant prostate cancer. *Nature* **464**, 302–305 (2010).

131. Sfanos, K. S. *et al.* Human prostate-infiltrating CD8+ T lymphocytes are oligoclonal and PD-1+. *The Prostate* **69**, 1694–1703 (2009).

132. Ammirante, M. *et al.* An IKKα-E2F1-BMI1 cascade activated by infiltrating B cells controls prostate regeneration and tumor recurrence. *Genes Dev.* **27**, 1435–1440 (2013).

133. Woo, J. R. *et al.* Tumor infiltrating B-cells are increased in prostate cancer tissue. *J. Transl. Med.* **12**, 30 (2014).

134. Sonpavde, G. *et al.* Prognostic impact of the neutrophil-to-lymphocyte ratio in men with metastatic castration-resistant prostate cancer. *Clin. Genitourin. Cancer* **12**, 317–324 (2014).

135. Sümbül, A. T. *et al.* Neutrophil-to-lymphocyte ratio predicts PSA response, but not outcomes in patients with castration-resistant prostate cancer treated with docetaxel. *Int. Urol. Nephrol.* **46**, 1531–1535 (2014).

136. Fujita, K. *et al.* Low serum neutrophil count predicts a positive prostate biopsy. *Prostate Cancer Prostatic Dis.* **15**, 386–390 (2012).

137. Lanciotti, M. *et al.* The role of M1 and M2 macrophages in prostate cancer in relation to extracapsular tumor extension and biochemical

recurrence after radical prostatectomy. *BioMed Res. Int.* **2014**, 486798 (2014).

138. Nonomura, N. *et al.* Infiltration of tumour-associated macrophages in prostate biopsy specimens is predictive of disease progression after hormonal therapy for prostate cancer. *BJU Int.* **107**, 1918–1922 (2011).

139. Mulders, P. F., De Santis, M., Powles, T. & Fizazi, K. Targeted treatment of metastatic castration-resistant prostate cancer with sipuleucel-T immunotherapy. *Cancer Immunol. Immunother. CII* **64**, 655–663 (2015).

140. Janiczek, M. *et al.* Immunotherapy as a Promising Treatment for Prostate Cancer: A Systematic Review. *Journal of Immunology Research* https://www.hindawi.com/journals/jir/2017/4861570/ (2017) doi:10.1155/2017/4861570.

141. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* **18**, 1009–1021 (2017).

142. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).

143. Meeks, H. D. *et al.* BRCA2 Polymorphic Stop Codon K3326X and the Risk of Breast, Prostate, and Ovarian Cancers. *J. Natl. Cancer Inst.* **108**, (2016).

144. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **161**, 1215–1228 (2015).

145. Pasero, C. *et al.* Inherent and Tumor-Driven Immune Tolerance in the Prostate Microenvironment Impairs Natural Killer Cell Antitumor Activity.

Cancer Res. **76**, 2153–2165 (2016).

146. Boibessot, C. & Toren, P. Sex steroids in the tumor microenvironment and prostate cancer progression. *Endocr. Relat. Cancer* **25**, R179–R196 (2018).

147. Alaia, C. *et al.* Ipilimumab for the treatment of metastatic prostate cancer. *Expert Opin. Biol. Ther.* **18**, 205–213 (2018).

148. Schepisi, G. *et al.* Immunotherapy for Prostate Cancer: Where We Are Headed. *Int. J. Mol. Sci.* **18**, (2017).

149. Hynes, S. O., Pang, B., James, J. A., Maxwell, P. & Salto-Tellez, M. Tissue-based next generation sequencing: application in a universal healthcare system. *Br. J. Cancer* **116**, 553–560 (2017).

150. Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nat. Rev. Genet.* **14**, 333–346 (2013).

151. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I. & Watson, M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Front. Genet.* **8**, (2017).

152. Buescher, J. M. & Driggers, E. M. Integration of omics: more than the sum of its parts. *Cancer Metab.* **4**, (2016).

153. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **16**, 85–97 (2015).

154. Chadeau-Hyam, M. *et al.* Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers. *Environ. Mol. Mutagen.* **54**, 542–557 (2013).

155. Hamid, J. S. *et al.* Data Integration in Genetics and Genomics: Methods and Challenges. *Hum. Genomics Proteomics* **1**, (2009).

156. Bersanelli, M. *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* **17**, (2016).

157. Pineda, S. *et al.* Integration Analysis of Three Omics Data Using Penalized Regression Methods: An Application to Bladder Cancer. *PLOS Genet.* **11**, e1005689 (2015).

158. Serizawa, R. R. *et al.* Integrated genetic and epigenetic analysis of bladder cancer reveals an additive diagnostic value of FGFR3 mutations and hypermethylation events. *Int. J. Cancer* **129**, 78–87.

159. Li, Q. *et al.* Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci. *Cell* **152**, 633–641 (2013).

160. Greenawalt, D. M. *et al.* Integrating Genetic Association, Genetics of Gene Expression, and Single Nucleotide Polymorphism Set Analysis to Identify Susceptibility Loci for Type 2 Diabetes Mellitus. *Am. J. Epidemiol.* **176**, 423–430 (2012).

161. Chibon, F. Cancer gene expression signatures – The rise and fall? *Eur. J. Cancer* **49**, 2000–2009 (2013).

162. Goossens, N., Nakagawa, S., Sun, X. & Hoshida, Y. Cancer biomarker discovery and validation. *Transl. Cancer Res.* **4**, 256 (2015).

163. McShane, L. M. *et al.* Criteria for the use of omics-based predictors in clinical trials. *Nature* **502**, 317–320 (2013).

164. McShane, L. M. *et al.* REporting recommendations for tumour MARKer prognostic studies (REMARK). *Br. J. Cancer* **93**, 387–391 (2005).

165. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

166. Osborne, J. Improving your data transformations: Applying the Box-Cox transformation. **15**, 10 (2010).

167. Verzelen, N. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6**, 38–90 (2012).

168. Al-Allak, A., Bertelli, G. & Lewis, P. Random forests: The new generation of machine learning algorithms to predict survival in breast cancer. *Int. J. Surg.* **11**, 607 (2013).

169. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).

170. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

171. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE* **13**, e0194889 (2018).

172. Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K. & Taha, K. Efficient Machine Learning for Big Data: A Review. *Big Data Res.* **2**, 87–93 (2015).

173. Bischl, B. *et al.* mlr: Machine Learning in R. 5.

174. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

175. Hall, M. *et al.* The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.* **11**, 10–18 (2009).

176. Rohart, F., Gautier, B., Singh, A. & Cao, K.-A. L. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13**, e1005752 (2017).

177. Abdi, H. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 97–106 (2010).

178. Meng, C. *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17**, 628–641 (2016).

179. Rohart, F., Eslami, A., Matigian, N., Bougeard, S. & Lê Cao, K.-A. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* **18**, (2017).

180. Tin Kam, H. *Proceedings of the Third International Conference on*. vol. 1 (1995).

181. Hasti, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. (2017).

182. Kumar, A. *et al.* Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat. Med.* **22**, 369–378 (2016).

183. Beltran, H. *et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat. Med.* **22**, 298–305 (2016).

184. Hieronymus, H. *et al.* Copy number alteration burden predicts prostate cancer relapse. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 11139–11144 (2014).

185. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell*

153, 666–677 (2013).

186. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).

187. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).

188. Taylor, B. S. *et al.* Integrative Genomic Profiling of Human Prostate Cancer. *Cancer Cell* **18**, 11–22 (2010).

189. Nguyen, B. *et al.* Pan-cancer Analysis of CDK12 Alterations Identifies a Subset of Prostate Cancers with Distinct Genomic and Clinical Characteristics. *Eur. Urol.* (2020) doi:10.1016/j.eururo.2020.03.024.

190. Abida, W. *et al.* Genomic correlates of clinical outcome in advanced prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 11428–11436 (2019).

191. Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).

192. Ren, S. *et al.* Whole-genome and Transcriptome Sequencing of Prostate Cancer Identify New Genetic Alterations Driving Disease Progression. *Eur. Urol.* **73**, 322–339 (2018).

193. Abida, W. *et al.* Prospective Genomic Profiling of Prostate Cancer Across Disease States Reveals Germline and Somatic Alterations That May Affect Clinical Decision Making. *JCO Precis. Oncol.* **2017**, (2017).

194. Omar, M. I. *et al.* Introducing PIONEER: a project to harness big data in prostate cancer research. *Nat. Rev. Urol.* 1–11 (2020) doi:10.1038/s41585-020-0324-x.

195. Lu, X. *et al.* Effective combinatorial immunotherapy for castration-resistant prostate cancer. *Nature* **543**, 728–732 (2017).

196. Decker, W. K. *et al.* Cancer Immunotherapy: Historical Perspective of a Clinical Revolution and Emerging Preclinical Animal Models. *Front. Immunol.* **8**, (2017).

197. Gotwals, P. *et al.* Prospects for combining targeted and conventional cancer therapy with immunotherapy. *Nat. Rev. Cancer* **17**, 286–301 (2017).

198. Pende, D. *et al.* Killer Ig-Like Receptors (KIRs): Their Role in NK Cell Modulation and Developments Leading to Their Clinical Exploitation. *Front. Immunol.* **10**, 1179 (2019).

199. Purdy, A. K. & Campbell, K. S. Natural killer cells and cancer: Regulation by the killer cell Ig-like receptors (KIR). *Cancer Biol. Ther.* **8**, 2209–2218 (2009).

200. Uhrberg, M. The KIR gene family: life in the fast lane of evolution. *Eur. J. Immunol.* **35**, 10–15 (2005).

201. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.* **54**, 15–39 (2009).

202. Choo, S. Y. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *Yonsei Med. J.* **48**, 11–23 (2007).

203. Shabihkhani, M. *et al.* The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings. *Clin. Biochem.* **47**, 258–266 (2014).

204. Bossel Ben-Moshe, N. *et al.* mRNA-seq whole transcriptome profiling of

fresh frozen versus archived fixed tissues. *BMC Genomics* **19**, 419 (2018).

205. Moch, H., Cubilla, A. L., Humphrey, P. A., Reuter, V. E. & Ulbright, T. M. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part A: Renal, Penile, and Testicular Tumours. *Eur. Urol.* **70**, 93–105 (2016).

206. *Pathology and genetics of tumours of the urinary system and male genital organs: ... editorial and consensus conference in Lyon, France, December 14 - 18, 2002*. (IARC Press, 2006).

207. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).

208. Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* **40**, e155–e155 (2012).

209. Afiahayati, Sato, K. & Sakakibara, Y. MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Res.* **22**, 69–77 (2015).

210. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).

211. Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13**, R122 (2012).

212. Treangen, T. J. *et al.* MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* **14**, R2 (2013).

213. Brady, A. & Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* **6**, 673–676 (2009).

214. Kelley, David R and Salzberg, Steven L. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics* **11**, 544 (2010).

215. Strous, M., Kraft, B., Bisdorf, R. & Tegetmeyer, H. E. The Binning of Metagenomic Contigs for Microbial Physiology of Mixed Cultures. *Front. Microbiol.* **3**, (2012).

216. Alneberg, J. *et al.* CONCOCT: clustering contigs on coverage and composition. *ArXiv Prepr. ArXiv13124038* (2013).

217. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Res.* **15**, 387–396 (2008).

218. Hoff, K. J., Lingner, T., Meinicke, P. & Tech, M. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* **37**, W101–W105 (2009).

219. Kelley, D. R., Liu, B., Delcher, A. L., Pop, M. & Salzberg, S. L. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* **40**, e9–e9 (2012).

220. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38**, e191–e191 (2010).

221. Seemann, T. Prokka: rapid prokaryotic genome annotation.

Bioinformatics **30**, 2068–2069 (2014).

222. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).

223. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132–e132 (2010).

224. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA11Edited by F. E. Cohen. *J. Mol. Biol.* **268**, 78–94 (1997).

225. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).

226. Zdobnov, E. M. & Apweiler, R. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).

227. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016).

228. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).

229. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).

230. Kutmon, M. *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.* **44**, D488–D494 (2016).

231. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46**, D633–D639 (2018).

232. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).

233. Wang, Q. *et al.* Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* **3**, (2015).

234. Richardson, E. J., Escalettes, F., Fotheringham, I., Wallace, R. J. & Watson, M. Meta4: a web application for sharing and annotating metagenomic gene predictions using web services. *Front. Genet.* **4**, (2013).

235. Wilke, A. *et al.* The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.* **44**, D590–D594 (2016).

236. Mitchell, A. *et al.* EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.* **44**, D595–D603 (2016).

237. Chen, I.-M. A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).

238. Li, P.-E. *et al.* Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res.* **45**, 67–80 (2017).