Zinara LIDAMAHASOLO

Master Bio-informatique - Parcours M2 Méthodes Informatiques et Statistiques pour les OMICS

RAPPORT DE STAGE

Période de stage : 01 Mars 2022 au 31 Août 2022

Mutations somatiques et trajectoires cancéreuses dans la muqueuse orale saine

Maître de stage : Pierre MARTINEZ

Tuteur : Mikaël SALSON

Lieu de stage : Centre de Recherche en Cancérologie de Lyon - Equipe « Analyse intégrée de la dynamique du cancer »

LIEU DE STAGE

Mon stage s'est déroulé au Centre de Recherche en Cancérologie de Lyon (CRCL), au sein de l'équipe « Analyse intégrée de la dynamique du cancer » dirigée par le Dr. Pierre Saintigny. Le CRCL est une structure de recherche sur le cancer labellisée par l'Université de Lyon 1, l'Inserm (UMR 1052), le CNRS (5286), le Centre Léon Bérard et avec pour partenaire hospitalier les Hospices Civils de Lyon. Le centre est composé de 24 équipes de recherche, et compte près de 500 personnels dont 150 chercheurs et enseignants-chercheurs. Un des objectifs majeurs du CRCL est de soutenir le développement d'une recherche translationnelle forte, au service des personnes malades. Le centre compte également 5 startups œuvrant dans la recherche et le développement de nouvelles thérapies. L'équipe qui m'a accueilli est composée de chercheurs venus d'horizon divers : techniciens de laboratoires, pharmacien, médecins, biologistes, et bio-informaticiens. Mon maître de stage était le Dr. Pierre Martinez, bio-informaticien et chargé de recherche à l'INSERM. L'équipe travaille sur différents types de cancers, principalement les cancers de la bouche, du poumon et du sein. Afin de mieux comprendre le cancer, les travaux de l'équipe sont axés principalement sur la compréhension des mécanismes influençant les différentes étapes tout au long du développement tumoral, aussi bien sur l'initiation du cancer que sur sa manière de s'adapter au traitement.

ABRÉVIATIONS ET ACRONYMES

ADN : Acide Désoxyribonucléique

AF : « *Allele Frequency* » ou fréquence allélique du variant

AO : « *Allele Observation Count* » ou nombre de reads portant le variant

ARN : Acide Ribonucléique

BAM : « *Binary Alignment Map* »

BED : « *Browser Extensible Data* »

CECO : Carcinomes Épidermoïdes de la Cavité Orale

dN/dS : « Rates of Nonsynonymous and Synonymous Substitutions » ou Ratio entre mutations non-synonymes et synonymes

DP : « *Read Depth* » ou profondeur de lecture

EMT : « *Epithelial mesenchymal transition* » ou transition épithélio-mésenchymateuse

LOPM : Lésions Orales à Potentielles Malin

NGS : « *Next-generation sequencing* » ou séquençage de nouvelle génération

PCR : « *Polymerase Chain Reaction* » ou réaction de polymérisation en chaîne

QUAL : *« Phred-scaled quality score* » ou qualité associée à l'inférence du variant

UTR : « *UnTranslated Region* » ou région non traduite

VCF : « *Variant Call Format* »

WES : « *Whole-exome sequencing* » ou séquençage d'exome entier

TABLE DES MATIÈRES

I. INTRODUCTION

Le séquençage de nouvelle génération (NGS), également connu sous le nom de séquençage à haut débit, correspond à différentes technologies de séquençage modernes. Ces technologies permettent le séquençage de l'ADN et de l'ARN beaucoup plus rapidement et à moindre coût que le séquençage Sanger précédemment utilisé, et ont ainsi révolutionné l'étude de la génomique et de la biologie moléculaire.

Le séquençage de l'exome entier (WES) consiste à séquencer seulement les régions du génome codant pour les protéines. L'exome humain représente moins de 2% du génome, mais contient environ 85% des variants connus liés à des maladies **(van Dijk et al. 2014)**, faisant de cette méthode une alternative rentable au séquençage du génome entier. Le séquençage de l'exome entier s'organise en deux étapes successives : l'enrichissement des cibles, et le séquençage. L'enrichissement des cibles consiste généralement à capturer les régions d'intérêt dans un échantillon d'ADN puis à les amplifier par une réaction de polymérisation en chaîne (PCR) ; pour obtenir des « librairies » d'oligonucléotides de petite taille aisément séquençables. De nombreuses solutions, ou « kits » existent dans le commerce pour ce faire, avec différentes spécificités.

Les avancées en termes de séquençage ont entre autres offert la possibilité d'étudier les effets des mutations qui surviennent dans les cellules primaires chez un organisme eucaryote. Les organismes eucaryotes ont deux types de cellules primaires : les cellules germinales, à l'origine des gamètes et de la transmission du patrimoine génétique à la génération suivante ; et les cellules somatiques, qui se répliquent et se différencient au cours du développement et de la maintenance de l'organisme. Si une mutation est présente dans une cellule germinale, elle est appelée mutation germinale et sera présente dans toutes les cellules de l'individu. Les mutations qui surviennent dans les cellules somatiques sont quant à elles appelées mutations somatiques. Elles ne se trouvent pas dans toutes les cellules de l'organisme et leur fréquence dépend du nombre de cellules descendantes produites par la cellule où la mutation est apparue. Les mutations somatiques sont la cause la plus fréquente de cancer **(Knudson 1971)**.

Une cellule peut devenir cancéreuse quand une ou plusieurs mutations impactent la fonction de gènes clés, communément appelés « *drivers* ». Ces gènes confèrent aux cellules un avantage de croissance, survie et/ou invasion lorsqu'ils sont mutés **(Hanahan 2022)**, favorisant ainsi la tumorigenèse. L'identification de ces gènes est une étape cruciale vers la

personnalisation du traitement du cancer. Il existe deux types de *drivers* : les proto-oncogènes, et les suppresseurs de tumeurs. Les proto-oncogènes tels que BRAF, KRAS et MYC participent à la régulation de la croissance et de la division cellulaires, et peuvent provoquer le développement tumoral via des gains de fonction provoqués par des mutations. Les gènes suppresseurs de tumeurs tels que TP53, PTEN et CDKN2A sont antiprolifératifs et nécessitent généralement l'inactivation des deux allèles pour devenir tumorigènes **(Piraino et al. 2018)**. A l'heure actuelle, 3347 gènes drivers du cancer sont recensés, dont 591 canoniques et 2756 candidats **(Dressler et al. 2022)**.

Selon la localisation et l'organe affecté, les gènes drivers impliqués peuvent diverger entre les cancers, mais certains sont partagés par presque tous les cancers. Un gène driver peut-être identifié par la sur-représentation ou la sous-représentation de mutations dans une cohorte de cancers génotypés **(Cibulskis et al. 2013)**. D'autres approches, basées sur le ratio dN/dS entre mutations non-synonymes et synonymes dans un gène, permettent de quantifier l'avantage sélectif d'un gène lorsqu'il est muté **(Iñigo Martincorena et al. 2017; M. J. Williams et al. 2020)**. Enfin, l'étude des mutations présentes dans différents types de tissus sains a démontré non seulement la forte présence de mutation *drivers* mais a aussi permis d'identifier des gènes conférant un avantage prolifératif aux cellules les présentant, même en l'absence de développement (pré-)tumoral **(Thomas 2019; Inigo Martincorena 2019)**. Les mutations des drivers TP53 et NOTCH1 sont particulièrement proéminentes dans le tissu sain **(Abby et al. 2021)**, ce dernier apparaissant même comme ayant un effet protecteur contre la tumorigenèse malgré le fait qu'il confère un avantage sélectif aux cellules lorsque muté.

Les cancers tête-et-cou font partie de ceux où on a identifié les plus faibles nombres de gènes drivers **(Dressler et al. 2022)**. Pourtant, environ 550 000 nouveaux cas de carcinomes de la tête et du cou sont recensés chaque année dans le monde, responsable d'une mortalité avoisinant les 275 000 cas par an **(Ferlay et al. 2015)**, avec une forte mortalité avec un taux de survie au-delà de 5 ans inférieur à 50%. Le carcinome squameux est le type de tumeur le plus fréquent, représentant à lui seul plus de 90% des tumeurs, et le site de cancer le plus commun est la cavité orale. Contrairement à d'autres régions du corps humain, l'accessibilité de la cavité orale lui offre l'avantage d'être plus facile à étudier avec des moyens beaucoup moins invasifs via la muqueuse orale.

Une étude **(The Cancer Genome Atlas Network 2015)** a montré que le carcinome épidermoïde de la tête et du cou présente un ensemble particulièrement mixte d'anomalies génomiques avec très peu de gènes drivers identifiés. Jusqu'à présent, seulement 62 gènes sont identifiés comme drivers du cancer pour ce type de cancer **(« IntOGen - Cancer driver**

mutations in Head and neck squamous cell carcinoma » s. d.). Les carcinomes épidermoïdes de la cavité orale (CECO) peuvent se développer à partir de lésions orales à potentielles malin (LOPM). Ce processus évolutif depuis un lit de dysplasie vers un carcinome est un processus très bien décrit aujourd'hui, mais nous manquons toujours de moyens efficaces pour stratifier efficacement les lésions qui progresseront de manière maligne ou non. L'évolution somatique à partir d'un tissu sain en passant par le stade précancéreux jusqu'à une maladie totalement invasive reste de plus très mal caractérisée. Ce grand manque dans la compréhension de l'évolution somatique et de la trajectoire cancéreuse est à l'origine de mon projet de stage, qui vise à décrire la dynamique somatique évolutive de la muqueuse orale. Le but est d'identifier les mutations et les gènes sous sélection positive dans ce tissu, et d'estimer la pertinence des méthodes utilisées tant pour le prélèvement que pour l'analyse des données. Pour rappel, une mutation particulière est sélectionnée positivement si elle confère un avantage prolifératif à la cellule, ou négativement lorsqu'elle a un effet négatif sur la cellule, qui est ensuite éliminée. Les gènes sous sélection positive sont des gènes dont la fonction est récurremment modifiée par des mutations, car ils participent à des processus clé pour la prolifération et la survie de la cellule (dans un contexte donné).

II. MATERIELS ET METHODES

A. Echantillons

Nos jeux de données sont constitués de 27 échantillons provenant de 15 sujets et séquencés par séquençage d'exome entier. Les échantillons sont répartis en 2 batches. Le premier batch est composé de 6 paires d'échantillons qui proviennent de 6 individus : un échantillon prélevé par cytobrosse et un autre par biopsie. La cytobrosse est un dispositif permettant de réaliser des prélèvements cytologiques moins invasifs sur les muqueuses que la biopsie. Le kit de capture *SureSelect Human All Exon V6* de Agilent avec UTR, permettant de cibler les exons et les régions adjacentes non traduites à capturer lors d'un séquençage d'exome entier, a été utilisé sur ce premier batch. La couverture de séquençage moyenne des échantillons dans ce batch est de 656,6X. Le second batch est constitué de 15 échantillons. Ceux-ci ont été obtenus par cytobrosse sur 9 sujets dont 3 possédants chacun 3 prélèvements réalisés lors des suivis à des dates différentes. Le kit de capture *SureSelect Human All Exon V8* de Agilent sans UTR a été utilisé sur ce second batch. La couverture de séquençage

moyenne des échantillons dans ce batch est de 327,6X. Les 27 échantillons avaient déjà été séquencés, démultiplexés, alignés (génome hg38), dédupliqués et stockés au format BAM avant le début de mon stage. Toutes les étapes réalisées dans ce stage sont résumées dans la figure suivante :

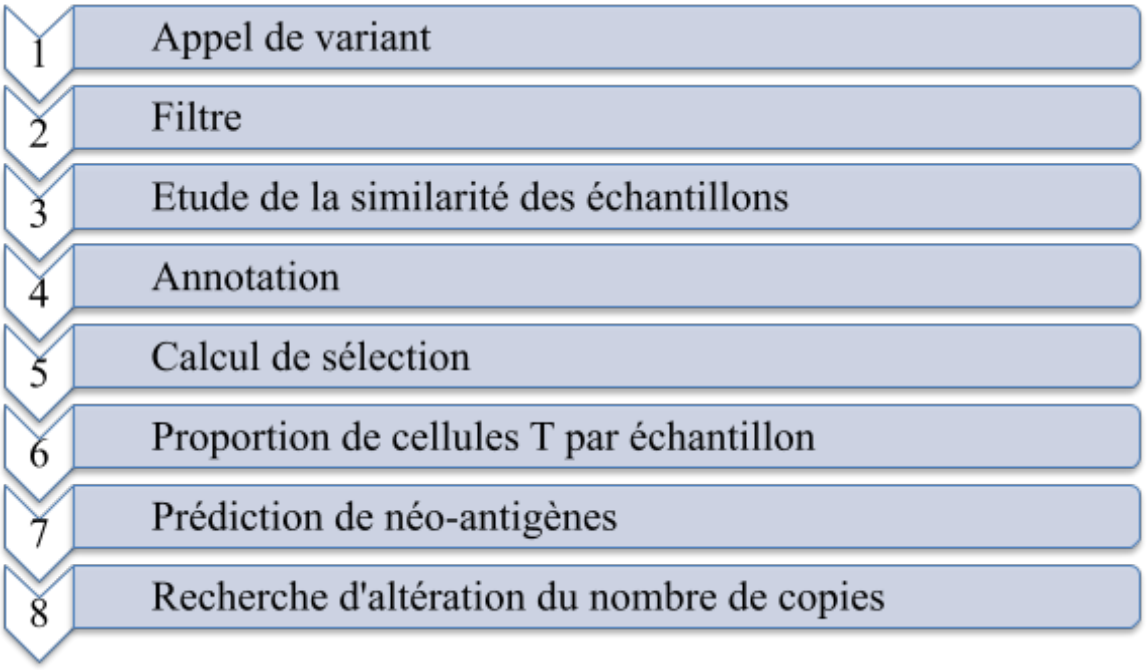

Figure 1 : Principales étapes du stage.

B. Appel de variants

Pour identifier les mutations somatiques et les différencier des mutations germinales, nous avons utilisé l'outil *Needlestack* **(Delhomme et al. 2020)** sur nos jeux de données. *Needlestack* est un outil ultra-sensible qui permet de faire un appel de variants somatiques sur les données de séquençage multi-échantillons de nouvelle génération comme dans notre cas. A la différence des autres outils d'appel de variant comme *mutect2* de *GATK*, *Needlestack* se base sur le principe que le fait d'analyser plusieurs échantillons en même temps donne une meilleure estimation de la distribution des erreurs de séquençage afin de mieux identifier les variants. *Needlestack* est un *pipeline* basé sur d'autres outils tels que *bedtools*, *samtools*, *R*, et *mpileup2readcount* (Figure 2). L'exécution du pipeline *Needlestack* requiert de plus le gestionnaire de flux *nextflow* **(Di Tommaso et al. 2017)** et des technologies conteneurs comme *Docker/Singularity* pour assurer l'évolution et la reproductibilité des tâches. Pour des raisons de droit utilisateur sur le cluster de calcul auquel j'ai eu accès, la solution *Singularity* a été utilisée **(Kurtzer, Sochat, et Bauer 2017)**.

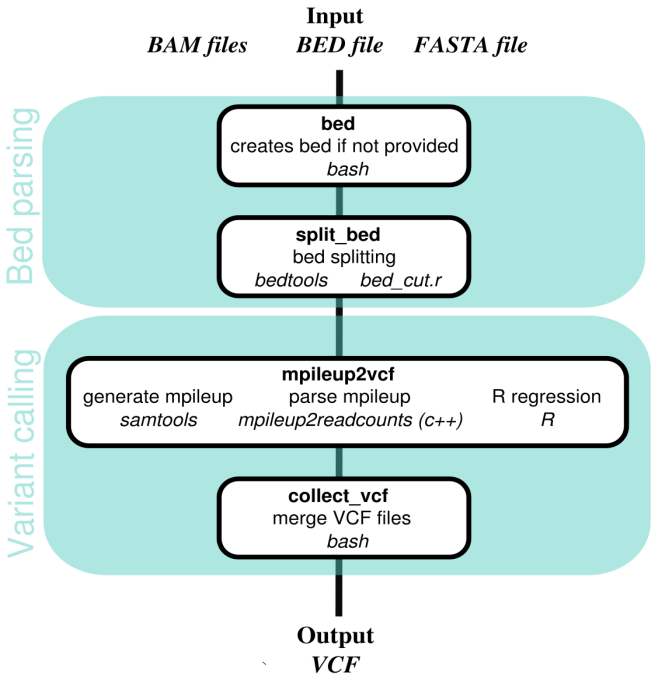

Figure 2 : Le pipeline des étapes implémentées dans Needlestack.
[https://github.com/IARCbioinfo/needlestack/blob/master/needlestack.png]

Needlestack utilise tous les échantillons pour estimer le taux d'erreur de séquençage par locus génomique, ce qui peut mener à des calculs intensifs. Afin d'optimiser le temps de calcul, les fichiers BAM individuels ont été séparés par chromosome, et l'appel de mutations pour chaque chromosome a été réalisé en parallèle. Un fichier BED a été utilisé pour restreindre l'appel de variants à l'intersection des deux kits de capture utilisés pour le séquençage.

Pour filtrer nos résultats et ne garder que les mutations pouvant être déterminées comme somatiques avec confiance, nous avons utilisé le package R *vcfR* **(Knaus et Grünwald 2017),** destiné à faciliter la manipulation des fichiers d'appel de variants, selon les paramètres suivants :

- la fréquence allélique de variants (AF) : nous avons gardé les variants avec une fréquence allélique inférieure à 0,3 afin d'éliminer les variants germinaux ;
- la profondeur de lecture (DP) et le nombre de reads portant le variant (AO) : nous n'avons gardé que les variants avec une profondeur de lecture supérieure à 50 et portés par au moins 3 reads, afin d'enlever les potentiels faux positifs ;
- la qualité associée à l'inférence du variant (QUAL) : nous avons fixé un seuil batch-spécifique minimum de qualité des variants, selon la distribution de leur qualité.

C. Similarité des échantillons

Pour mesurer la similarité entre 2 échantillons, nous avons utilisé l'indice de Jaccard, défini comme le rapport de la taille de l'intersection et de la taille de l'union pour 2 ensembles : $J(A,B) = (A \cap B) / (A \cup B)$. Nous avons ensuite réalisé 3 comparaisons différentes : les échantillons d'un même patient obtenus par biopsie et ceux obtenus par cytobrosse pour les 6 patients du premier batch ; les échantillons obtenus par cytobrosse de même patient à différentes dates pour les 3 seuls patients du deuxième batch possédant chacun 3 échantillons espacés dans le temps ; et les échantillons obtenus par cytobrosse de patients différents entre eux.

D. Annotation et calcul de sélection

Nos variants filtrés ont été annotés avec l'outil *VEP* **(McLaren et al. 2016),** afin d'identifier les gènes impliqués et l'impact fonctionnel de chaque mutation. Nous avons de plus utilisé les données d'expression en cellules uniques d'une étude récente sur la muqueuse orale **(D. W. Williams et al. 2021)**, pour définir les gènes exprimés et non-exprimés dans les cellules épithéliales de la muqueuse orale. Les gènes présentant au moins un transcript dans au moins 5% des cellules sont considérés comme exprimés, et les gènes présentant des transcripts dans moins de 1% des cellules sont considérés comme non-exprimés. En croisant ces informations avec les variants annotés, nous avons pu identifier 1) les 500 gènes les plus mutés dans nos échantillons qui sont exprimés dans l'épithélium de la muqueuse orale saine et 2) les 1000 gènes les plus mutés dans nos échantillons qui ne sont pas exprimés dans ce tissu. L'outil *dNdScv* **(Iñigo Martincorena et al. 2017)** a été utilisé pour calculer les avantages sélectifs de ces 500 et 1000 gènes, ainsi que des 62 gènes *drivers* du cancer épidermoïde de la tête et du cou. Cet outil permet de quantifier la sélection dans le cancer et l'évolution somatique en calculant les ratios dN/dS pour les mutations faux-sens, les mutations non-sens, et celles impactant les sites d'épissage.

E. Proportion de cellules T

Nous avons utilisé le package R *TCellExTRECT* pour le calcul de la fraction de cellules T infiltrant chaque échantillon. Ce processus est basé sur la couverture (en nombre de reads) dans la région TRA (*T Cell Receptor Alpha Locus*), localisée sur le chromosome 14.

F. Identification des néo-antigènes

Afin de déterminer les mutations donnant lieu à de nouveaux antigènes reconnaissables par le système immunitaire acquis du patient, il est nécessaire de premièrement déterminer le profil HLA (*Human Leukocyte Antigen*) de chaque échantillon avec l'outil *Polysolver* **(Shukla et al. 2015)**. Cet outil infère les allèles pour les 3 gènes majeurs (HLA-A, -B, -C) qui codent pour le complexe majeur d'histocompatibilité de classe 1 à partir de données de séquençage d'exome entier. *Polysolver* commence par récupérer les reads appartenant à la région HLA en se référant à une base de données de tous les allèles HLA connus (Figure 3). Ces reads sont ensuite alignés sur une bibliothèque génomique en conservant les alignements les mieux notés. Puis, les 2 allèles pour chaque gène HLA sont inférés à l'aide d'une approche bayésienne qui prend en compte la qualité des reads retenus, les tailles d'insert observées, ainsi que les probabilités dépendantes de l'ethnicité pour chaque allèle.

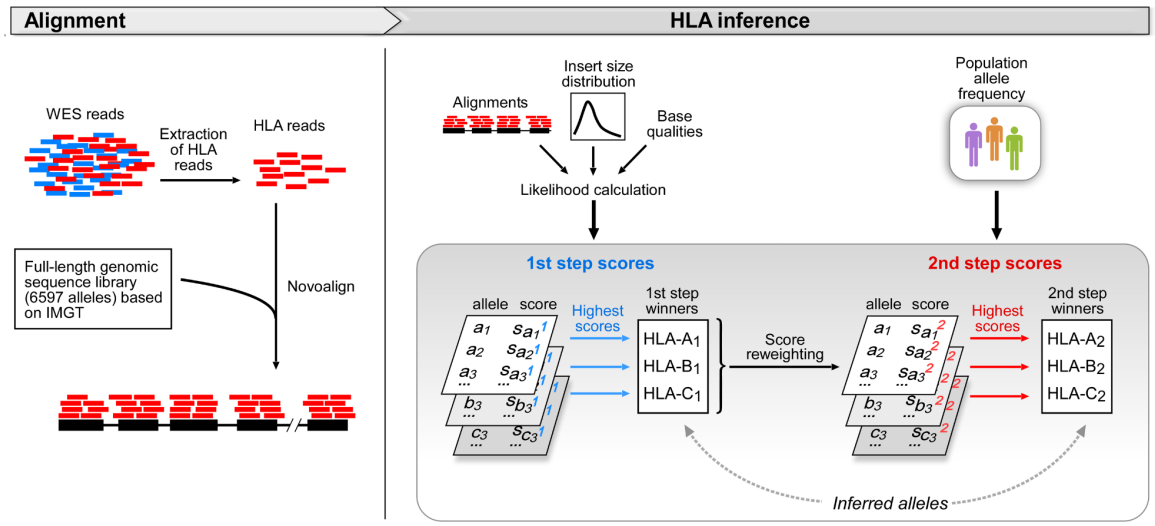

Figure 3 : Le pipeline des étapes implémentées dans Polysolver.
[https://software.broadinstitute.org/cancer/cga/sites/default/files/data/tools/polysolver/POLYSOLVER.png]

Une fois les profils HLA des échantillons déterminés, il est possible de prédire les néo-antigènes. Pour cela, nous avons utilisé l'outil *NeoPredPipe* **(Schenck et al. 2019)** qui permet la prédiction de néo-antigènes à partir d'un fichier VCF et dépend des outils *ANNOVAR* **(Wang, Li, et Hakonarson 2010)**, *netMHCpan* **(Hoof et al. 2009)**, et *PeptideMatch* **(C. Chen et al. 2013)** pour fonctionner (Figure 4).

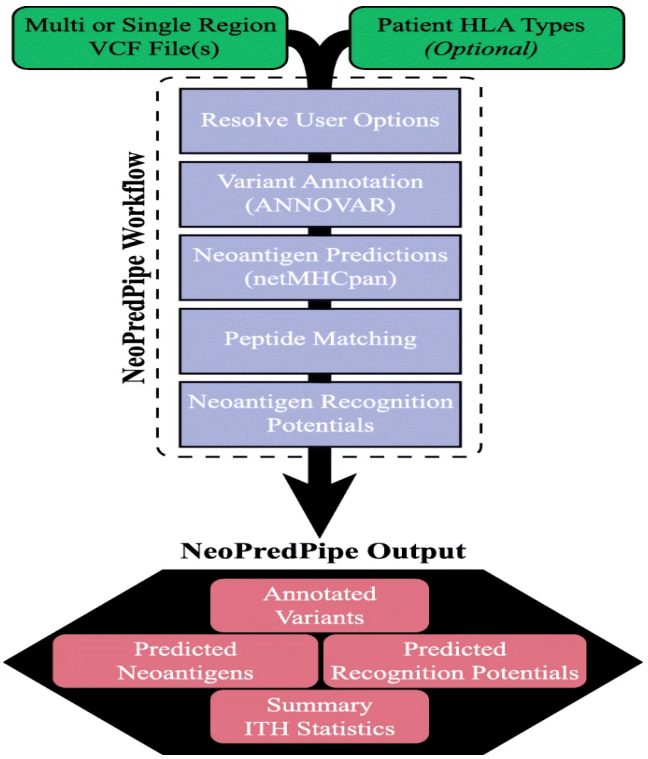

Figure 4 : Le pipeline des étapes implémentées dans NeoPredPipe **(Schenck et al. 2019)**.

Grâce à l'outil *ANNOVAR*, *NeoPredPipe* va annoter les variants pour identifier ceux étant non-synonymes. La séquence d'acide aminé mutée est prédite à partir des variants non-synonymes annotés, et les séquences peptidiques entourant l'acide aminé nouvellement introduit sont extraites pour la prédiction de l'épitope. L'outil *netMHCpan* utilise les haplotypes HLA précédemment déterminés pour prédire les néo-antigènes primaires. Nous nous sommes concentrés sur les valeurs par défaut pour rechercher des néo-antigènes dans les 8-, 9-, ou 10-mers contenant toute nouvelle mutation. Ce processus va nous donner des néo-antigènes prédits que *PeptideMatch* va référencer avec des peptides normaux pour déterminer si les épitopes candidats sont nouveaux ou pas par rapport à un protéome de référence que nous allons fournir. Les néo-antigènes vont aussi être étiquetés comme « ligant fort » si leur liaison avec le complexe majeur d'histocompatibilité est forte, et « ligant faible » dans le cas contraire.

G. Altérations du nombre de copies

Nous avons recherché l'éventuelle présence d'altérations de nombre de copies dans les échantillons à l'aide de l'outil *PureCN* **(Riester et al. 2016)** (Figure 5).

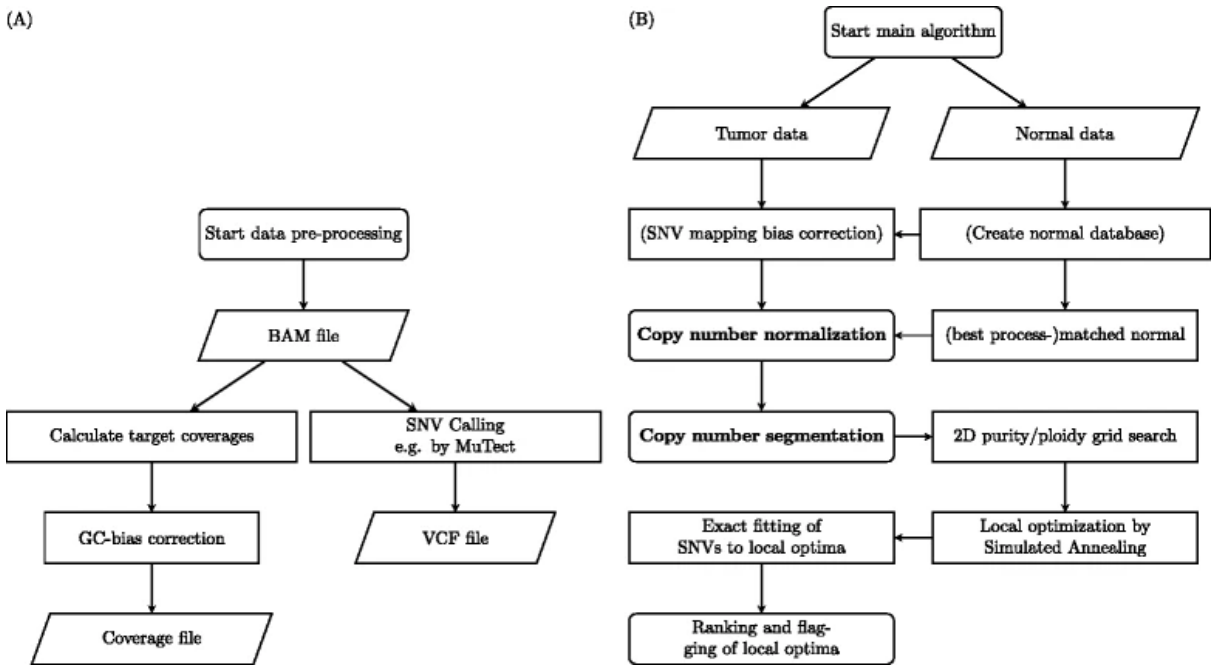

Figure 5 : Organigramme (A) du pipeline et algorithme (B) de prétraitement de données PureCN ***(26)****.*

Ces altérations sont typiques des génomes tumoraux et nous voulions estimer si leur présence est détectable dans la muqueuse saine environnante. Contrairement à d'autres solutions qui nécessitent un échantillon normal apparié, *PureCN* est un package R permettant de déterminer ces altérations du nombre de copies dans des données de séquençage d'exome sans normal apparié, ce qui correspond au design de l'étude.

Cela se base sur l'utilisation des couvertures normalisées des autres échantillons du même batch afin de déterminer les attentes correspondant à un ensemble de contrôles normaux, mais non appariés. Cela permet de déterminer la pureté et la ploïdie de l'échantillon, de détecter les segments du génome comportant un signal anormal et de déterminer leur caryotype. *PureCN* a été conçu à la base pour intégrer les pipelines standards basés sur GATK. Il fonctionne sans reformatage complémentaire ni de risques de complications avec les fichiers VCF issus de *mutect* **(Cibulskis et al. 2013)** de *GATK*, que nous avons utilisés pour ces analyses.

III. RÉSULTATS

A. Variants somatiques

L'appel de variants a donné au total 2 457 587 variants dont la majorité (95%) appartient au batch 2 (Figure 6).

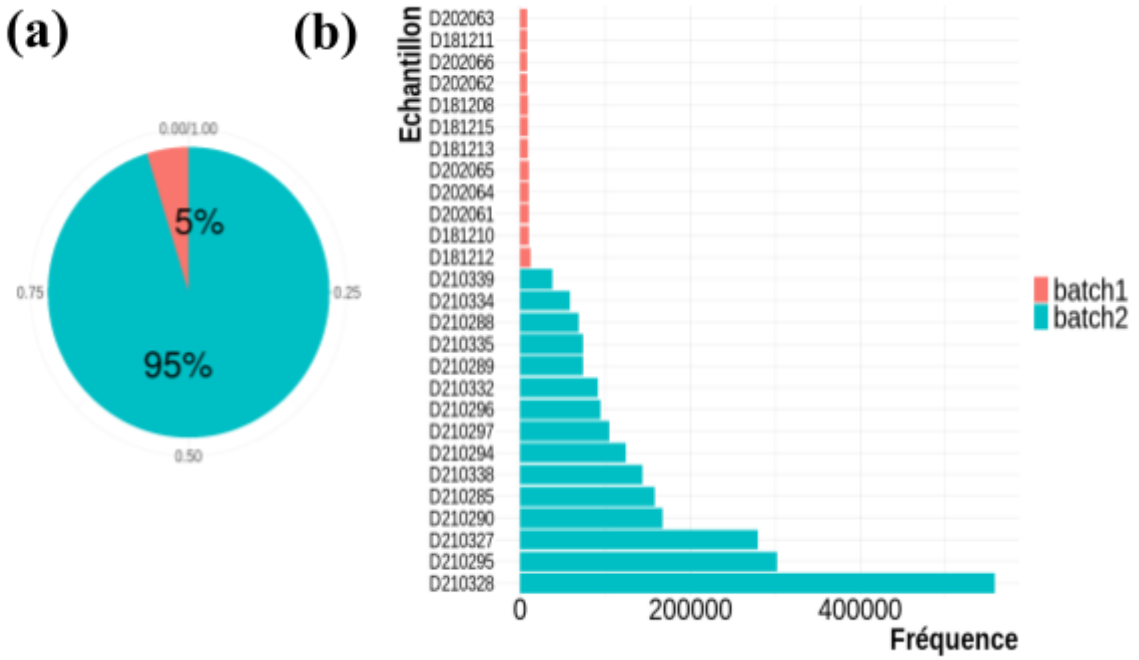

*Figure 6 : (**a**) Répartition des variants selon les batches. (**b**) Répartition des variants selon les échantillons.*

L'analyse de la qualité des variants a montré que la majorité des variants de batch 2 était de mauvaise qualité avec une médiane de 81,24, alors que celle de batch 1 était de 1000 (Figure 7). Les 2 batches avaient une distribution de profondeur de lecture assez proche, avec des valeurs de 3 quartiles de [321, 619, 1104] pour batch 1 et [635, 824, 1045] pour batch 2, mais batch 2 présentait des valeurs extrêmes plus hautes, allant jusqu'à 43 800 reads, contre maximum 15 271 pour batch 1. Les variants de batch 2 présentaient beaucoup plus de fréquences alléliques très faibles que ceux de batch 1.

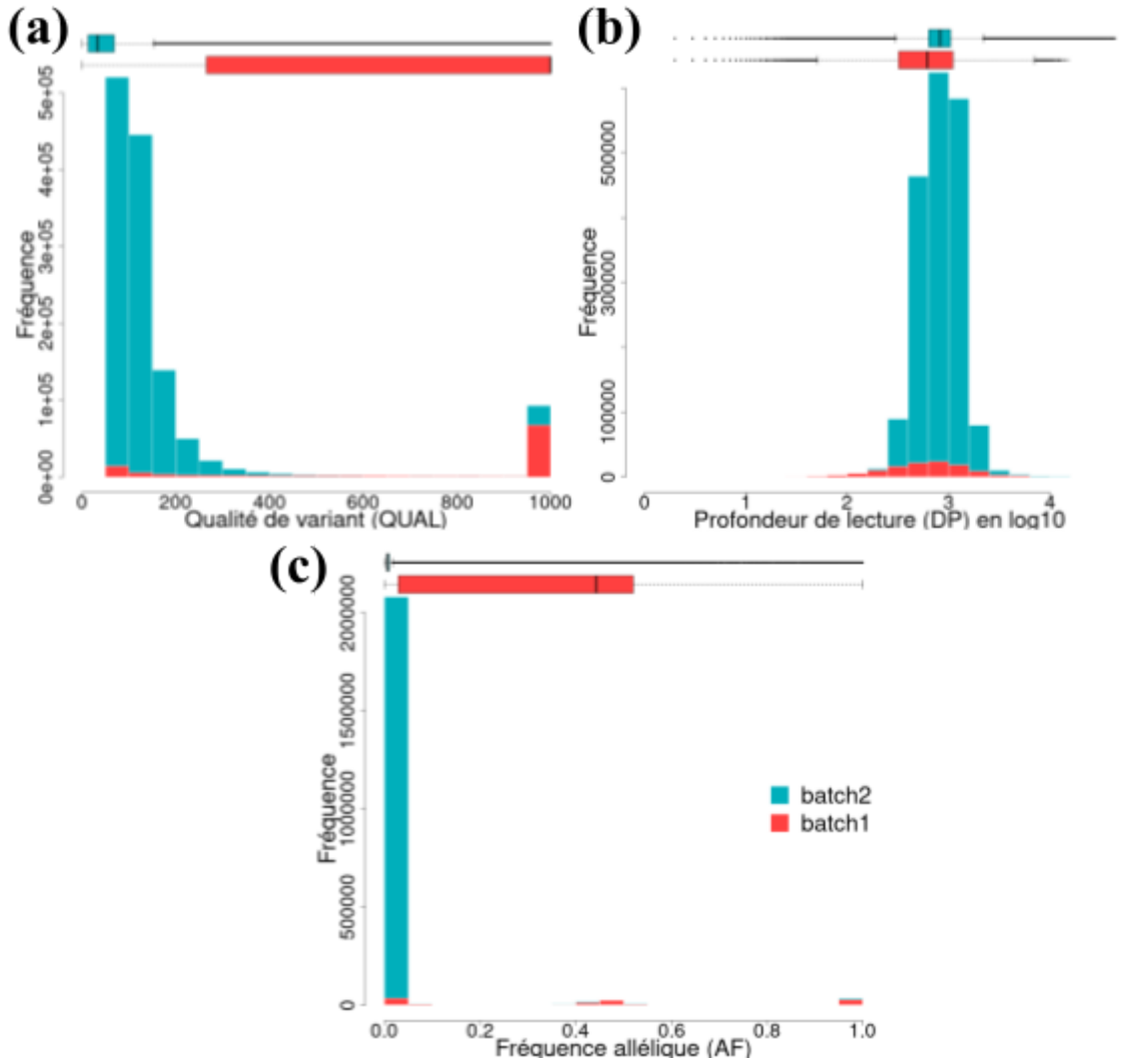

*Figure 7 : **(a)** Répartition de la qualité des variants selon les batches. **(b)** Répartition de la profondeur de lecture selon les batches. **(c)** Répartition de la fréquence allélique des variants selon les batches.*

Afin d'éviter la sur-représentation de mutations de moindre qualité issues du batch 2, nous avons décidé de fixer pour le filtre un seuil minimal du paramètre de qualité de 100 pour batch 1 et de 400 pour batch 2. Le filtre a éliminé 97,3% des variants appelés dont la majorité appartenait à batch 2 (Figure 8). Après filtre, il ne restait plus que 28,4% des variants initiaux de batch 1 et 1,4% de batch 2. Au total, 51% des variants restants appartenaient à batch 1 et 49% à batch 2. La différence nette observée de la distribution des variants entre les échantillons des batches a aussi disparu.

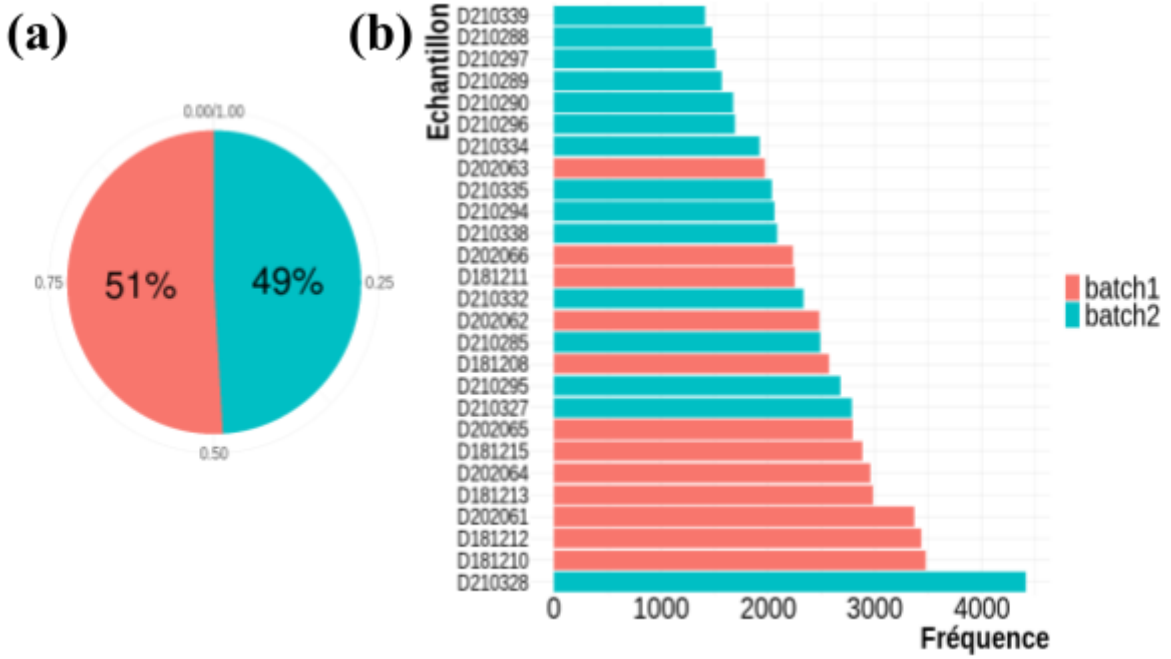

*Figure 8 : **(a)** Répartition des variants selon les batches. **(b)** Répartition des variants selon les échantillons.*

Le filtre a eu pour impact une nette amélioration de la qualité des variants du batch 2, faisant passer sa médiane de 81,24 avant filtre à 1000 après filtre (Figure 9). La distribution de la profondeur de lecture des batches s'est légèrement rapprochée puisque la distance entre les médianes est passé de 205 avant filtre à 22 après filtre. Nous avons aussi observé un net rapprochement de la répartition de la fréquence allélique des batches car la distance entre les médianes est passée de 0,44 avant filtre à 0,01 après filtres.

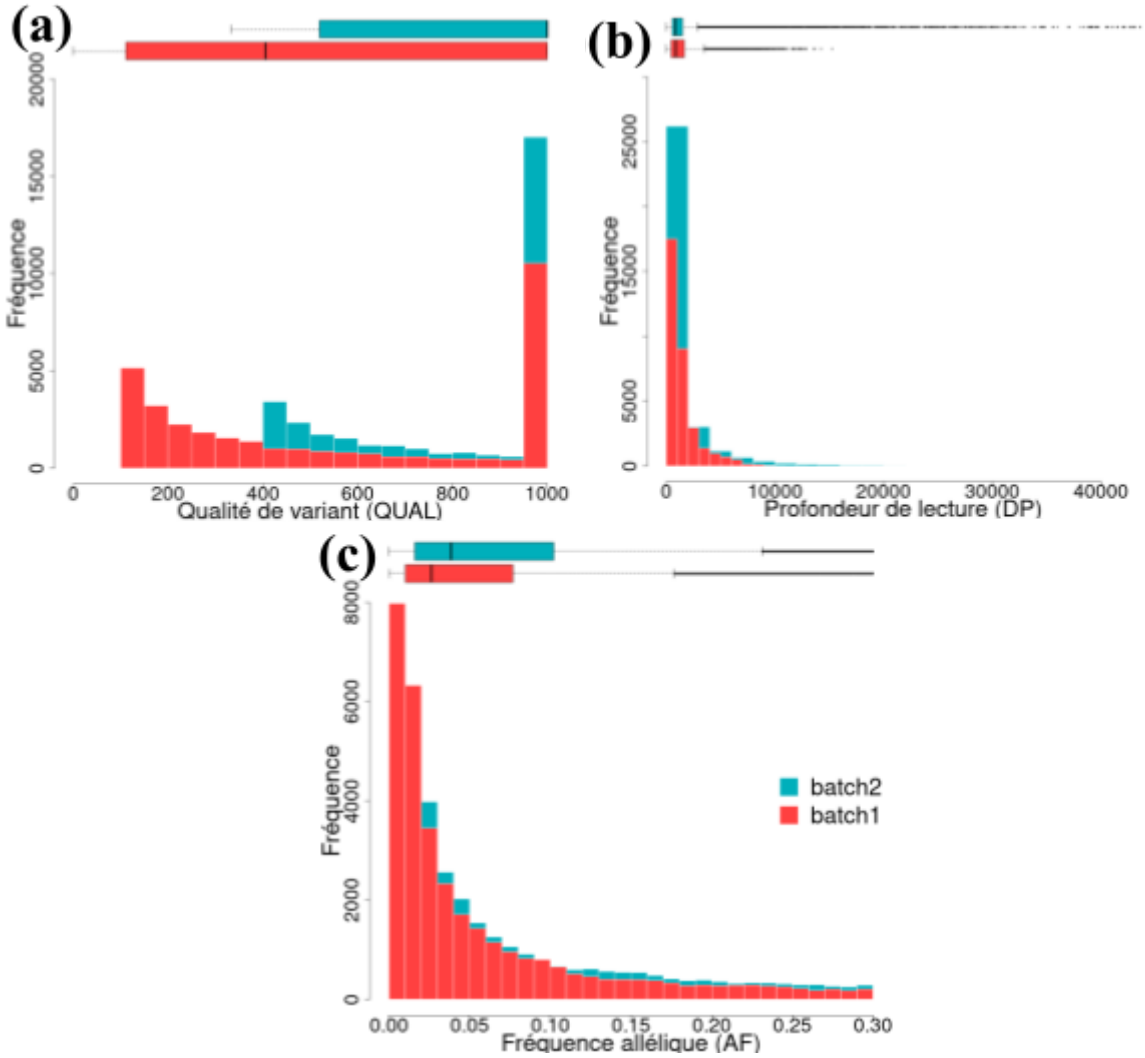

*Figure 9 : (**a**) Répartition de la qualité des variants selon les batches après filtre. (**b**) Répartition de la profondeur de lecture selon les batches après filtre. (**c**) Répartition de la fréquence allélique des variants selon les batches après filtre.*

B. Similarité des échantillons

Comme attendu, la similarité entre échantillons issus des mêmes patients était significativement plus élevée qu'entre échantillons de patients différents (Figure 10). Des tests Wilcoxon ont donnée des valeurs p=0,86 pour les comparaisons biopsie-brosse et brosse-brosse de même patient, $p=4{,}1\times10^{-5}$ pour les comparaisons biopsie-brosse de même patient et brosse-brosse de patient différent, et $p=7{,}1\times10^{-7}$ pour les comparaisons brosse-brosse de même patient et brosse-brosse de patient différent.

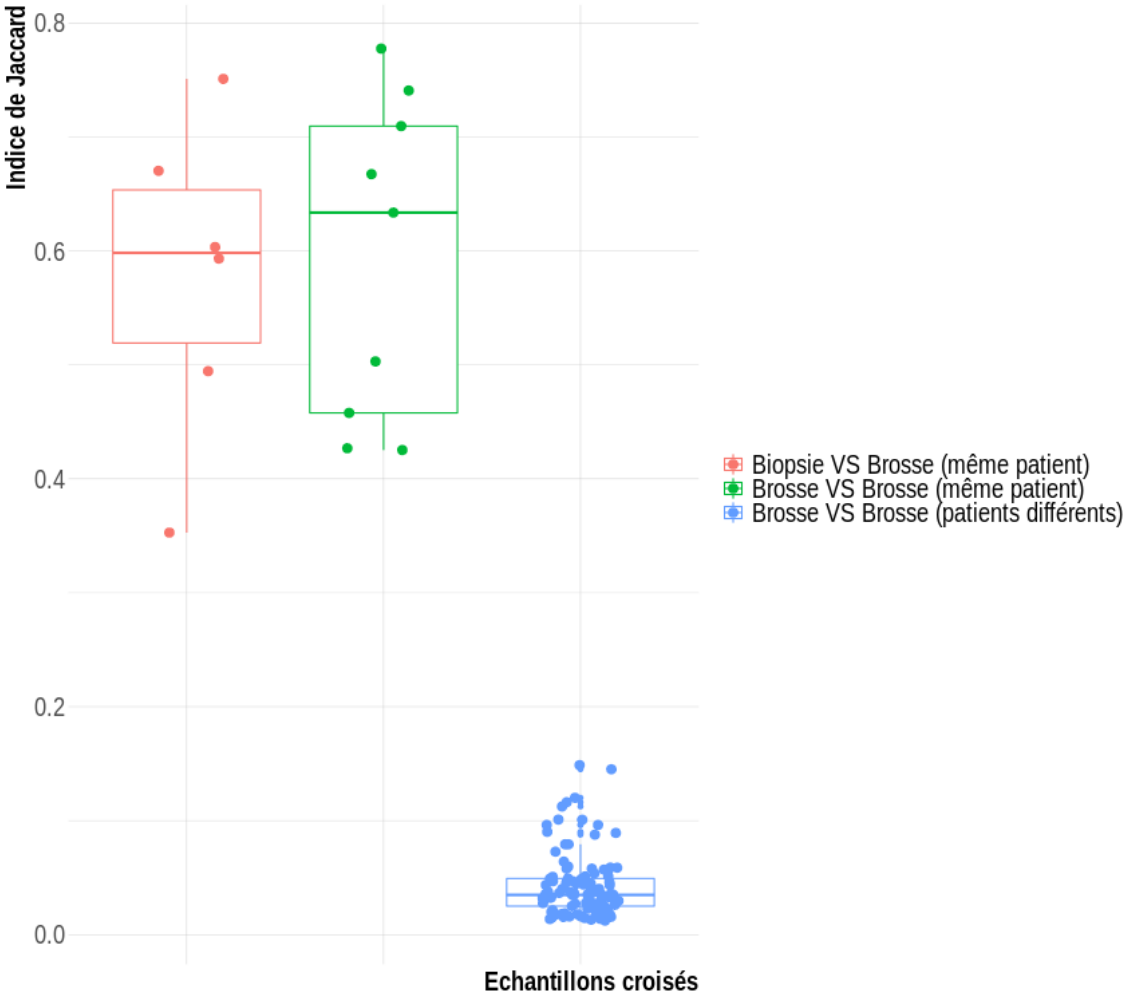

Figure 10 : Indice de Jaccard entre les échantillons.

C. Annotation et calcul de sélection

L'annotation des mutations avec l'outil *VEP* nous a informé que seulement 31.5% des variants codants étaient des synonymes, c'est-à-dire des variants qui n'induisent pas le changement des acides aminés associés (Figure 11). 62,5% des variants codants étaient des faux-sens (*missense* en anglais), qui sont des variants ponctuels (une seule base ADN affectée) non-synonymes induisant un changement de l'acide aminé associé. Le reste des mutations correspondait à d'autres sous-types de la classe non-synonyme.

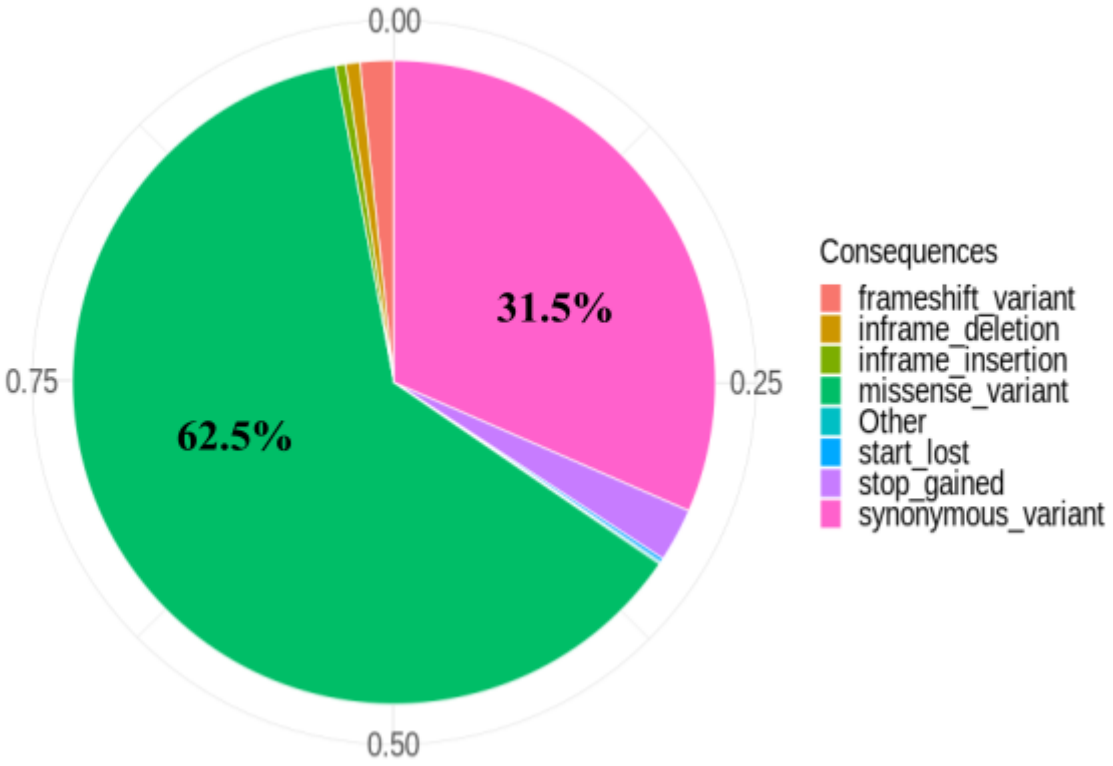

Figure 11 : Conséquences des variants.

Sur les 62 gènes drivers du cancer épidermoïde de la tête et du cou, 9 étaient significativement sous sélection dans nos échantillons (Tableau 1). Les gènes CREBBP, KDM5C, FAT3, et MYH9 étaient fortement sous élection négative pour les variants faux-sens et non-sens. Le gène CASP8 était le seul sous sélection positive à la fois pour les variants faux-sens et non-sens.

Tableau 1 : Gènes sous sélection significative (Plus c'est rouge, plus c'est sous sélection positive, et plus c'est bleu, plus c'est sous sélection négative)

Gène	nombre_de_mutations	dNdS_faux_sens	dNdS_non_sens	p_globale_ajustée
CREBBP	10	0.0000000	0.0000000	0.0006718274
CASP8	9	1.2133029	40.8414933	0.0006718274
FAT1	5	0.1296048	2.4311507	0.0007435047
NOTCH1	19	0.8875176	8.4462262	0.0032685884
KDM5C	8	0.0000000	0.0000000	0.0064674674
FAT3	11	0.2627164	0.0000000	0.0087696291
HLA-B	25	11.1075040	0.0000000	0.0217561387
HLA-A	45	1.2874269	0.0000000	0.0427825393
MYH9	15	0.1241248	0.4862293	0.0439383395

Les gènes FAT1 et NOTCH1 étaient sous sélection positive pour les variants non-sens, mais sous sélection négative pour les faux-sens. Les gènes HLA-A et HLA-B étaient retrouvés sous sélection positive pour les variants faux-sens, mais sous sélection négative pour les non-sens. Le gène TP53 n'était pas significatif (Tableau 2) mais a présenté une forte tendance vers une sélection positive à la fois pour les variants faux-sens et non-sens.

Tableau 2 : Gènes sous sélection non-significative (Plus c'est rouge, plus c'est sous sélection positive, et plus c'est bleu, plus c'est sous sélection négative)

Gène	nombre_de_mutation	dNdS_faux_sens	dNdS_non_sens	p_valeur_ajustée
KMT2C	12	0.6372499	0.000000	0.1466849
TP53	6	2.7857984	18.93655	0.2102199
BRCA2	5	0.4000035	0.000000	0.3675089
PIK3CA	6	1.8106351	0.000000	0.8660165
DDX3X	5	2.0057472	0.000000	0.8660165
NOTCH2	14	0.7701560	0.000000	0.9401093

Concernant les 500 gènes les plus mutés parmi ceux exprimés dans les cellules épithéliales de la muqueuse orale saine et n'étant pas des drivers, 40 avaient un avantage sélectif significatif (Annexe 1) dont la majorité sous sélection négative. Les gènes HLA-DRB1 et FRG1 étaient sous sélection positive à la fois pour les variants faux-sens et les non-sens. Les gènes RBMXL1, SASH1, SETD2, et UBC étaient sous sélection positive uniquement pour les variants non-sens. Le gène HLA-DQB1 était le seul sous sélection positive uniquement pour les variants faux-sens. Les gènes TUSC1, HLA-DPB1, ANKRD11, ITGB4, et PRKRA n'étaient pas significatifs (Annexe 2) mais présentaient également une tendance vers une sélection positive.

Parmi les 1000 gènes les plus mutés parmi ceux non-exprimés dans les cellules épithéliales de la muqueuse orale saine et n'étant pas des drivers, 182 avaient un avantage sélectif significatif (Annexe 3). 37 gènes avaient un avantage sélectif positif pour les variants faux-sens et non-sens. Les gènes ANKRD20A1, POTED, NPIPA8, PPIAL4C, GAGE12E, GAGE12D, KRTAP4-3, SAA2, OR4F21, OR2L8, OR2T3, CT45A2, CT45A3, CT45A8, CT45A9, et OR51A4 étaient très fortement et uniquement sous sélection positive pour les variants faux-sens. Les gènes CDRT15, OR2T2, et FAM47A étaient très fortement et uniquement sous sélection positive pour les variants non-sens. Les gènes TCP10L2, OR2T35, GAGE10, LCN1, DEFB104A, DEFB104B, et PRR23D1 présentaient une forte tendance vers une sélection positive mais n'étaient pas significatifs (Annexe 4). La pertinence de ces

résultats sera étudiée dans la suite des travaux du laboratoire, en particulier après l'analyse d'échantillons supplémentaires.

D. Proportion de cellules T

Les proportions de cellules T étaient significativement plus élevées dans les échantillons de batch 1 (médiane=0,08) que dans ceux de batch 2 (médiane=0) (Figure 12). Dans batch 1, les échantillons obtenus par biopsie présentaient une proportion de cellules T significativement plus élevée (médiane=0,12) que ceux obtenus par cytobrosse (médiane=0,06). Cela confirme les résultats préliminaires du labo obtenus par cytologie, qui avaient rapporté une forte pureté des échantillons de muqueuse orale prélevés par cytobrosse (98% de cellules épithéliales).

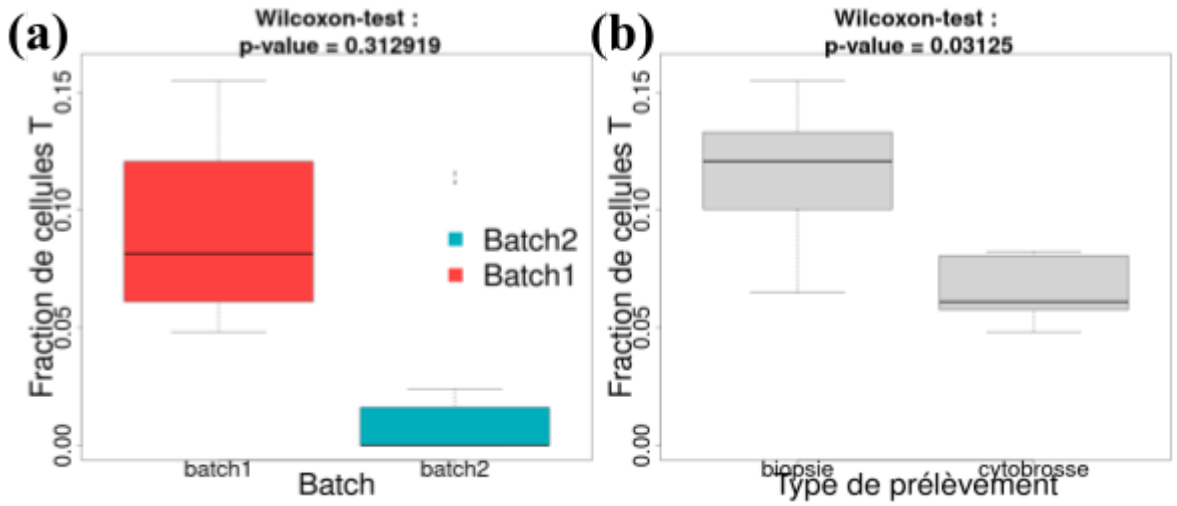

*Figure 12 : **(a)** Fractions de cellules T selon le batch. **(b)** Fractions de cellules T selon le type de prélèvement dans batch 1.*

Nous avons de plus trouvé une liaison significative et inverse entre les proportions de cellules T et la fréquence des variants des gènes HLA-A et HLA-C pour chaque échantillon. Pour le gène HLA-B, cette liaison n'était pas significative (Figure 13). Cela suggère que ces mutations ne proviennent pas des cellules T infiltrantes, mais bien des cellules épithéliales.

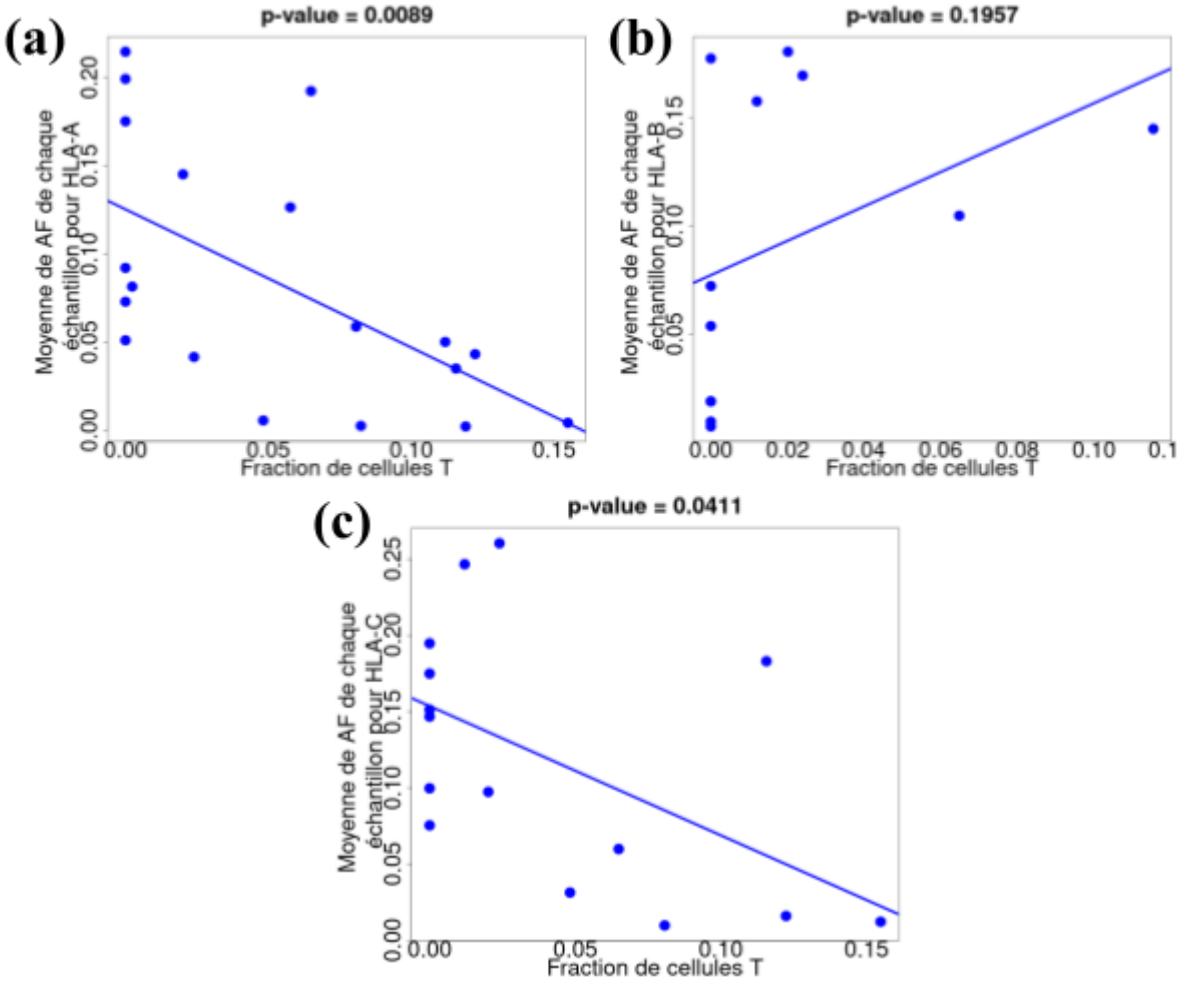

*Figure 13 : **(a)** Liaison entre proportion de cellules T et fréquence allélique des variants du gène HLA-A. **(b)** Liaison entre proportion de cellules T et fréquence allélique des variants du gène HLA-B. **(c)** Liaison entre proportion de cellules T et fréquence allélique des variants du gène HLA-C.*

E. Identification des néo-antigènes

En moyenne, 30,93% des variants somatiques dans chaque échantillon conduisaient à des néo-antigènes (Figure 14). Parmi ces néo-antigènes, en moyenne 89,80% étaient nouveaux dans chaque échantillon. Et parmi ces nouveaux néo-antigènes, en moyenne 73,34% étaient des « ligands faibles » et 26,66% des « ligands forts » dans chaque échantillon.

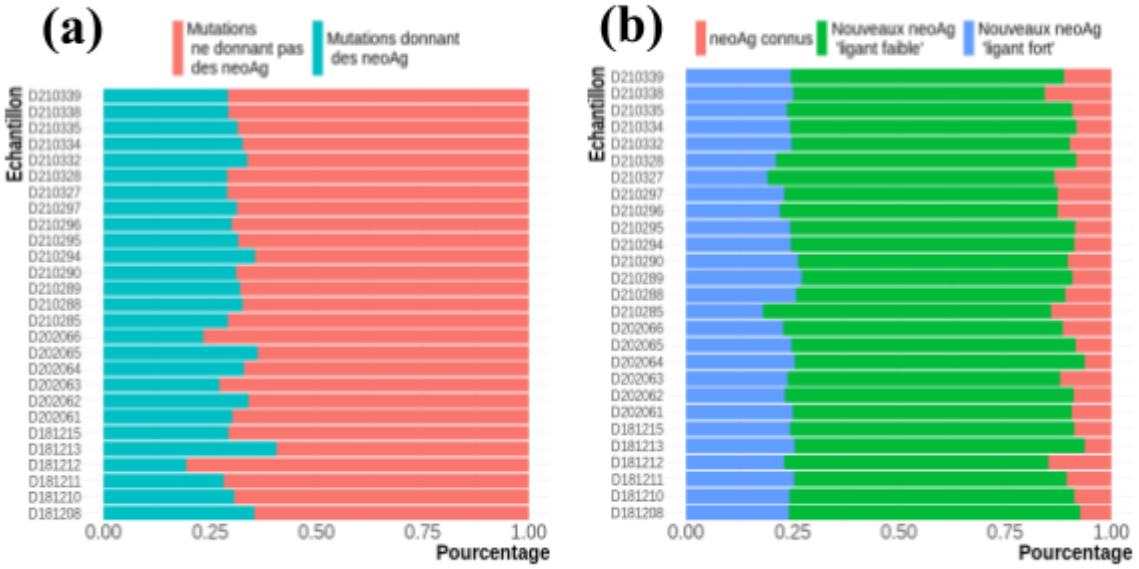

*Figure 14 : **(a)*** Variants *donnant des néo-antigènes dans chaque échantillon. **(b)** Nature des nouveaux néo-antigènes dans chaque échantillon.*

En moyenne, 3,10% des variants dans chaque échantillon donnaient des néo-antigènes et se situaient sur les gènes exprimés dans les cellules épithéliales de la muqueuse orale saine (Figure 15). Cela représente 9,59% des néo-antigènes totaux étaient issus de ces gènes exprimés. Le pourcentage moyen dans chaque échantillon des nouveaux néo-antigènes « ligands faibles » issus des gènes exprimés était de 6,52%, qui étaient plus fréquents que les « ligands forts » (2,31%).

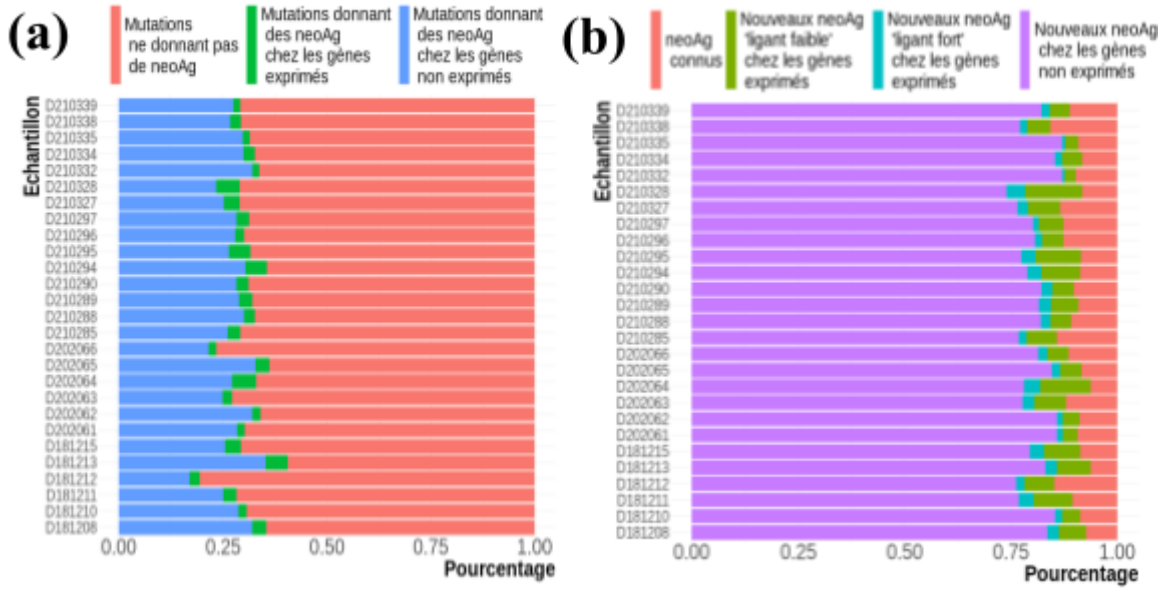

*Figure 15 : **(a)** Variants donnant des néo-antigènes dans chaque échantillon. **(b)** Nature des nouveaux néo-antigènes dans chaque échantillon.*

F. Altérations du nombre de copies

Comme attendu pour des échantillons de tissu sain, la pureté tumorale (proportion des cellules étant tumorales) était estimée comme faible, avec une moyenne de 23%. La médiane de la ploïdie était de 2 (Figure 16). Aucun échantillon n'a présenté d'altération significative de nombre de copies. Cela suggère qu'il n'y a pas d'effet champ de cancérisation, où des altérations du nombre de copies typiques du cancer seraient aussi présentes dans le tissu sain environnant (dans notre cas, la muqueuse orale contro-latérale).

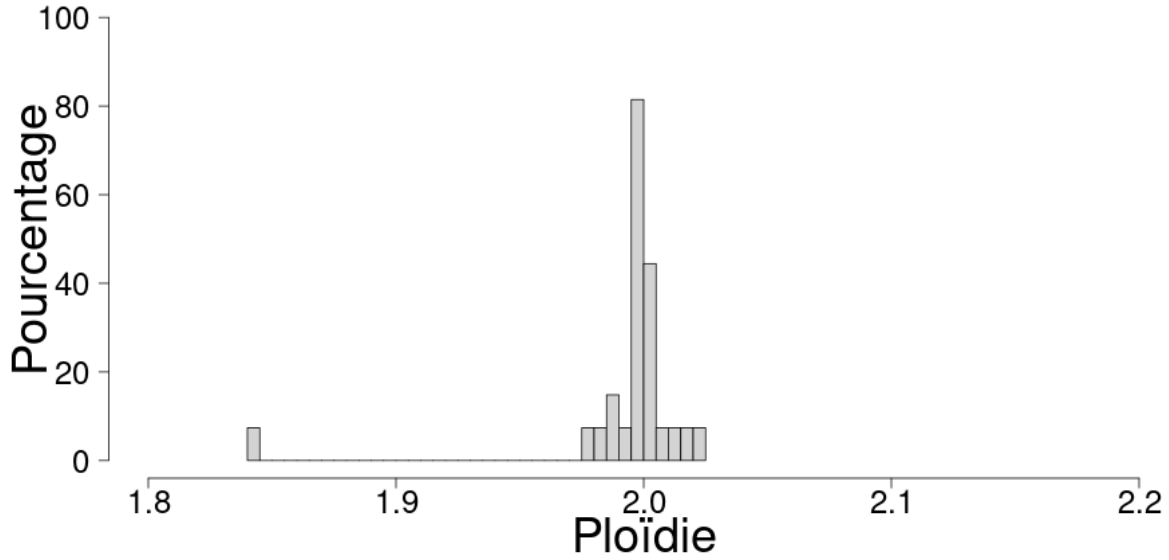

Figure 16 : Ploïdie tumorale des échantillons.

IV. DISCUSSION

A. Appel de variants et filtre

A l'issue de l'appel de variants, les échantillons du batch 2 présentaient plus de variants que ceux du batch 1 probablement à cause de la différence de profondeur de séquençage entre les 2 batches. En effet, les échantillons de batch 1 ont été séquencé avec une profondeur moyenne de 656,6X tandis que ceux de batch 2 ont été séquencé à 327,6X. Avec une profondeur de lecture plus importante, les informations provenant d'un maximum de reads conduisaient à une meilleure qualité des variants en nombre plus faible, mais probablement plus de vrais-positifs dans les échantillons du batch 1. Pour les échantillons du batch 2, la faible profondeur de séquençage impliquait une prédiction moins précise des

variants à cause de l'impact plus élevé des erreurs de séquençage, faisant augmenter le nombre de faux-positifs et de variants prédits. En majorité, ce sont ces variants faux-positifs et erreurs de séquençage ainsi que les artéfacts d'alignement qui avaient des mauvais scores de qualité, des profondeurs de lectures et fréquences alléliques très faibles et qui n'ont pas passé le filtre mis en place. Ici, nous avons utilisé un filtre dur (*Hard filter*) pour chercher les vrais variants somatiques rares. C'est une méthode beaucoup moins précise car les paramètres du filtre sont fixés de manière arbitraire. Il existe d'autres méthodes plus précises que nous n'avons pas utilisées car inadaptées à notre cas comme la méthode « *Ensemble calling* » **(Chapman et al. 2021)** qui est un appel à consensus établi sur un vote majoritaire sur les variants candidats par plusieurs outils d'appel de variants, la méthode « *Blacklisting* » qui permet d'exclure la liste des positions génomiques et des mutations en se basant sur un panel de normaux, et la méthode par « *Machine learning* » qui consiste à entraîner des algorithmes pour reconnaître les caractéristiques qui distinguent les vrais variants. A cause du petit nombre d'échantillons disponibles, nous n'avons également pas pu tester et calibrer le filtre sur des données séparées de tests pour obtenir les seuils qui distinguent le mieux les variants vrais-positifs des variants faux-positifs.

B. Les gènes sous sélection positive

Sur les 62 gènes drivers du carcinome épidermoïde de la tête et du cou, CASP8, HLA-A, HLA-B, FAT1, et NOTCH1 ont été retrouvés sous sélection positive dans notre étude. Ce faible nombre de gènes détectés est probablement dû à la petite taille de notre cohorte. Le gène CASP8 est un gène pro-tumoral ou tumeur-suppresseur selon le contexte cellulaire, impliqué dans la protéolyse et la voie de signalisation apoptotique extrinsèque via les récepteurs de mort (récepteurs cellulaires qui induisent l'apoptose lorsqu'ils sont activés par leurs ligands). Ce gène est muté dans 10% des cancers épidermoïdes de la tête et du cou **(Cerami et al. 2012; Hoadley et al. 2018)**. Les mutations de ce gène comprennent à la fois des mutations faux-sens et non-sens comme dans notre cas, sont largement distribuées dans toute la région codante et sont associées à une survie réduite des patients **(Cerami et al. 2012; Uzunparmak et al. 2020)**. Cependant, une étude récente a démontré que les mutants du gène CASP8 associés au cancer épidermoïde de la tête et du cou conservent quand même des propriétés pouvant influencer l'apoptose médiée par TRAIL (médiateur de l'apoptose) et l'induction de cytokines, ainsi que la composition du microenvironnement tumoral **(Cui et al. 2021)**. Les gènes HLA-A et HLA-B codent pour les molécules du complexe

d'histocompatibilité de classe I impliquées dans la présentation des peptides intracellulaires aux lymphocytes T, permettant au système immunitaire de reconnaître et de détruire les cellules infectées ou cancéreuses. Les mutations somatiques au niveau des gènes HLA sont un mécanisme d'évasion immunitaire car ces mutations sont associées à une charge plus élevée de néo-antigènes qui devraient être présentés via les complexes majeurs d'histocompatibilité **(Castro et al. 2019)**. La majorité des mutations touchant les gènes HLA sont des non-synonymes avec une prédominance des faux-sens **(Castro et al. 2019)**, un fait qui a été confirmé par notre étude. Cela suggère de plus une pression sélective pour que les complexes HLA ne soient plus en mesure de présenter des néo-antigènes à la surface de la cellule, lui permettant ainsi d'échapper à la surveillance immunitaire de l'organisme. Le gène FAT1 est impliqué dans le contrôle de la prolifération cellulaire. Son inactivation favorise l'EMT, les caractéristiques invasives et les métastases dans le carcinome épidermoïde de la peau, le cancer du poumon, et les tumeurs de la tête et du cou **(Pastushenko et al. 2021)**. L'analyse fonctionnelle réalisée lors d'une étude **(Lin et al. 2018)** sur le cancer de la tête et du cou a aussi suggéré que les mutations non-sens dominent dans FAT1 et entraînent la perte de la suppression de la progression tumorale. Cela a été corroboré par notre étude car nous avons trouvé une sélection positive uniquement pour les mutations faux-sens. Le gène NOTCH1 est l'un des plus célèbres gènes du cancer dans les études de tissu sain. Il code pour des protéines du récepteur transmembranaire à passage unique. Dans notre étude, le gène est retrouvé sous sélection positive uniquement pour les non-sens. Ce modèle est retrouvé dans des études dans lesquelles le biais inféré pour les mutations non-sens était le plus dominant pour le gène NOTCH1 **(M. J. Williams et al. 2020)**, et que les ratios dD/dS pour les mutations tronquées (non-sens et sur les sites d'épissage) étaient les plus élevés **(Iñigo Martincorena et al. 2018)**. Son influence a récemment été démontrée comme permettant une plus forte prolifération dans l'œsophage sain, tout en ayant au final un effet protecteur contre le développement tumoral **(Abby et al. 2021)**. Le gène tumeur-suppresseur TP53 fait partie des gènes drivers du cancer les plus célèbres, et est surnommé le « gardien du génome ». Il est retrouvé dans nombreuses études avec une sélection positive pour les mutations tronquées **(M. J. Williams et al. 2020; Iñigo Martincorena et al. 2018)**, mais dans notre étude il présente cette tendance sans être significative probablement à cause du faible nombre d'échantillons. Tout ceci suggère que les données et la méthodologie sont probablement correctes et appropriées.

Sur les autres gènes qui ne sont pas drivers du cancer épidermoïde de la tête et du cou, nous avons trouvé une centaine de gènes sous sélection positive. Parmi ces gènes, une

cinquantaine sont fortement exprimés dans l'épithélium de la muqueuse orale, ce qui les rend potentiellement importants dans l'évolution somatique et la tumorigenèse. Le gène HLA-DRB1 que nous avons retrouvé sous sélection positive pour les faux-sens et les non-sens est un gène HLA-II, qui code pour les complexes d'histocompatibilité de classe II. Son polymorphisme est le plus abondant du gène HLA-II et est le facteur décisif du polymorphisme de l'antigène immunitaire. Sa relation avec le cancer du sein a été démontrée **(Liu et al. 2021; Zouré et al. 2021)**. Le gène FRG1 sous sélection positive également pour les mutations faux-sens et non-sens joue un rôle dans la tumorigenèse et l'angiogenèse. Son expression est associée à la survie des patients dans le cancer du col de l'utérus, gastrique, du poumon, et du foie **(Khan, Palo, et Dixit 2021)**. Sa restriction favorise aussi la progression du cancer de la prostate et affecte la migration et l'invasion des cellules cancéreuses de la prostate **(Tiwari et al. 2019)**. Le gène RBMXL1 sous sélection positive pour les variants non-sens est un rétrogène des protéines de liaisons à l'ARN. Il est surexprimé chez les personnes atteintes de leucémies myéloïdes aiguës, sa restriction retarde le développement de la leucémie **(Prieto et al. 2021)**. Le gène SASH1 sous sélection positive aussi pour les variants non-sens est un tumeur-suppresseur impliqué dans l'apoptose et la prolifération cellulaire. C'est un indicateur pronostique et une cible thérapeutique potentielle dans le cancer du poumon non à petite cellule **(Burgess et al. 2020)**. Il supprime l'invasion des cellules cancéreuses du sein triple négatives **(Jiang et al. 2020)**, un type de cancer extrêmement agressif et sans traitement efficace. Le gène SETD2 qui est sous sélection positive pour les variants non-sens est impliqué dans la régulation épigénétique de la transcription, la réparation de l'ADN, et les fonctions liées aux protéines non histones. C'est un tumeur-suppresseur potentiel dans de nombreux cancers solides **(R. Chen et al. 2020)**. Le gène UBC également sous sélection positive pour les variants non-sens code pour des protéines précurseurs de la poly-ubiquitination (polymérisation sur une protéine cible, d'une chaîne de molécule d'ubiquitine, permettant ainsi l'adressage de la protéine modifiée au protéasome pour dégradation). La restriction inhibe la prolifération et la radiorésistance des cellules du cancer du poumon non à petite cellule **(Tang et al. 2015)**. Le gène HLA-DQB1 sous sélection positive pour les variants faux-sens est un gène HLA-II. Son expression est un nouveau facteur pronostique favorable pour la rechute dans l'adénocarcinome pulmonaire à un stade précoce **(Zhang et al. 2019)**.

C. Proportion de cellules T et néo-antigènes

La proportion élevée de cellules T dans les échantillons obtenus par biopsie signifie que cette méthode de prélèvement n'était pas la plus adaptée à notre étude. En effet, afin de déterminer les gènes sous sélection positive dans les cellules épithéliales orales, il est important d'avoir une grande pureté des échantillons, avec les moins de cellules tierces possible. La méthode de prélèvement par cytobrosse va donc être la seule méthode utilisée pour les prochaines études sur l'évolution somatique dans le laboratoire. Ici, nous avons réalisé une simple description et avons démontré que l'accumulation des mutations somatiques dans les tissus sains conduisent à des néo-antigènes avec une proportion non négligeable d'antigène fort (26,66% des nouveaux néo-antigènes, 23,95% des néo-antigènes totaux) probablement déjà immunogène.

V. CONCLUSION

Cette étude d'exploration nous a conduit à des pistes et informations intéressantes dans la compréhension de l'évolution somatique dans la muqueuse orale. Dans ce stage, nous avons implémenté tous les pipelines qui feront office de socle pour les prochaines analyses plus poussées sur l'évolution somatique de la muqueuse orale au sein de l'équipe. Un financement a de plus récemment (juillet 2022) été obtenu par le laboratoire Saintigny pour analyser des échantillons de muqueuse orale saine chez des sujets sains fumeurs et non-fumeurs, ainsi que des patients atteints de carcinome épidermoïde de la cavité orale. Les développements et mises en place d'outils effectués durant mon stage faciliteront grandement ces analyses.

Nous avons trouvé une différence de répartition des variants entre les échantillons des deux batches, mais cela a été corrigé par un filtre. Nous avons établi le filtre avec les paramètres qualité du variant, fréquence allélique du variant, profondeur de lecture, nombre de reads portant le variant, dont les seuils ont été fixés de manière arbitraire sans contrôle qualité et recalibration ultérieure. Nous avons cependant pu constater l'impact du filtre avec une amélioration nette de la qualité des variants, principalement du batch 2, et un rapprochement de la profondeur de lecture et la fréquence allélique entre les batches. L'étude de la similarité entre les échantillons n'a pas détecté d'artefacts de séquençage susceptibles d'influencer nos analyses. L'annotation des variants nous a montré une proportion majoritaire de mutations non-synonymes. Malgré les diversités liées au batch et la petite taille de la cohorte, nous avons tout de même pu retrouver les gènes candidats les plus attendus

(NOTCH1, TP53), confirmant la pertinence de nos approches. De plus, des gènes non associés jusqu'à présent au cancer ont été identifiés comme pertinents dans l'évolution somatique, même si leur véracité devra être étudiée plus profondément dans les études à venir au sein de l'équipe. Dans notre étude, deux méthodes de prélèvements ont été utilisées, et nous avons trouvé que la méthode la moins invasive était aussi la plus adaptée pour une catégorie d'étude comme la nôtre. Nous avons aussi démontré la présence de néo-antigènes avec une antigénicité élevée au sein des tissus sains. Comme attendu pour ces tissus sains, nous n'avons pas trouvé d'altération du nombre de copies. L'absence de données patients n'a pas permis de faire une modélisation et une caractérisation mathématique avec des facteurs explicatifs et des facteurs tiers du processus d'évolution somatique.

Dans l'optique d'exploiter ces résultats, un projet de thèse a été soumis pour financement auprès de la *Ligue Nationale Contre le Cancer,* pour lequel je suis le candidat identifié. Ce projet consistera à étudier et caractériser l'évolution somatique et pré-maligne de la muqueuse orale. Il aura pour principal objectif de mieux comprendre les processus évolutifs somatiques à l'œuvre en amont de l'initiation tumorale, afin d'élaborer de meilleures stratégies de prévention.

APPORT DU STAGE

Dans ce stage, j'ai implémenté les pipelines d'analyse dans leur totalité. Cela m'a permis d'améliorer mes compétences en informatique, surtout en termes de programmation en langage *R* et *Bash*. J'ai également pu maîtriser le système d'exploitation UNIX et les clusters de calcul, manipuler les gestionnaires de flux comme *nextflow* et les technologies conteneurs comme *Singulatity*, et manipuler des outils de bio-informatiques de dernière génération. Le stage m'a aussi permis d'apprendre à faire de la documentation sur les outils. Cela m'a permis de gagner en autonomie et en persévérance, de renforcer le côté relationnel et le travail en équipe, et surtout de découvrir le monde de la cancérologie qui est très vaste, multidisciplinaire, avec un fort besoin en bio-informatique. Pour l'équipe, mes scripts seront utilisés comme base de référence pour des études similaires, et mes résultats ont soulevé certaines pistes à suivre et à approfondir. Après ce stage, j'envisage d'évoluer et faire carrière dans l'étude du cancer, un problème de santé humaine qui convient à mon profil de pharmacien, à travers une thèse ou bien un contrat professionnel pour un poste d'ingénieur.

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Abby, Emilie, Stefan C Dentro, Michael W J Hall, Joanna C Fowler, Swee Hoe Ong, Roshan Sood, Christian W Siebel, Moritz Gerstung, Benjamin A Hall, et Philip H Jones. 2021. « *Notch1* Mutation Drives Clonal Expansion in Normal Esophageal Epithelium but Impairs Tumor Growth ». Preprint. Cancer Biology. https://doi.org/10.1101/2021.06.18.448956.
2. Burgess, Joshua T., Emma Bolderson, Mark N. Adams, Pascal H. G. Duijf, Shu-Dong Zhang, Steven G. Gray, Gavin Wright, Derek J. Richard, et Kenneth J. O'Byrne. 2020. « SASH1 Is a Prognostic Indicator and Potential Therapeutic Target in Non-Small Cell Lung Cancer ». *Scientific Reports* 10 (1): 18605. https://doi.org/10.1038/s41598-020-75625-1.
3. Castro, Andrea, Kivilcim Ozturk, Rachel Marty Pyke, Su Xian, Maurizio Zanetti, et Hannah Carter. 2019. « Elevated Neoantigen Levels in Tumors with Somatic Mutations in the HLA-A, HLA-B, HLA-C and B2M Genes ». *BMC Medical Genomics* 12 (S6): 107. https://doi.org/10.1186/s12920-019-0544-1.
4. Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, et al. 2012. « The CBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data ». *Cancer Discovery* 2 (5): 401‑4. https://doi.org/10.1158/2159-8290.CD-12-0095.
5. Chapman, Brad, Rory Kirchner, Lorena Pantano, Sergey Naumenko, Matthias De Smet, Luca Beltrame, Tetiana Khotiainsteva, et al. 2021. « bcbio/bcbio-nextgen »: Zenodo. https://doi.org/10.5281/ZENODO.5781867.
6. Chen, Chuming, Zhiwen Li, Hongzhan Huang, Baris E. Suzek, Cathy H. Wu, et UniProt Consortium. 2013. « A Fast Peptide Match Service for UniProt Knowledgebase ». *Bioinformatics* 29 (21): 2808‑9. https://doi.org/10.1093/bioinformatics/btt484.
7. Chen, Rui, Wei-qing Zhao, Cheng Fang, Xin Yang, et Mei Ji. 2020. « Histone Methyltransferase SETD2: A Potential Tumor Suppressor in Solid Cancers ». *Journal of Cancer* 11 (11): 3349‑56. https://doi.org/10.7150/jca.38391.
8. Cibulskis, Kristian, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, et Gad Getz. 2013. « Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples ». *Nature Biotechnology* 31 (3): 213‑19. https://doi.org/10.1038/nbt.2514.
9. Cui, Zhibin, Hadas Dabas, Brandon C. Leonard, Jamie V. Shiah, Jennifer R. Grandis, et Daniel E. Johnson. 2021. « Caspase-8 Mutations Associated with Head and Neck Cancer Differentially Retain Functional Properties Related to TRAIL-Induced Apoptosis and Cytokine Induction ». *Cell Death & Disease* 12 (8): 775. https://doi.org/10.1038/s41419-021-04066-z.
10. Delhomme, Tiffany M, Patrice H Avogbe, Aurélie A G Gabriel, Nicolas Alcala, Noemie Leblay, Catherine Voegele, Maxime Vallée, et al. 2020. « Needlestack: An Ultra-Sensitive Variant Caller for Multi-Sample next Generation Sequencing Data ». *NAR Genomics and Bioinformatics* 2 (2): lqaa021. https://doi.org/10.1093/nargab/lqaa021.
11. Di Tommaso, Paolo, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, et Cedric Notredame. 2017. « Nextflow Enables Reproducible Computational Workflows ». *Nature Biotechnology* 35 (4): 316‑19. https://doi.org/10.1038/nbt.3820.

12. Dijk, Erwin L. van, Hélène Auger, Yan Jaszczyszyn, et Claude Thermes. 2014. « Ten Years of Next-Generation Sequencing Technology ». *Trends in Genetics* 30 (9): 418‑26. https://doi.org/10.1016/j.tig.2014.07.001.
13. Dressler, Lisa, Michele Bortolomeazzi, Mohamed Reda Keddar, Hrvoje Misetic, Giulia Sartini, Amelia Acha-Sagredo, Lucia Montorsi, et al. 2022. « Comparative Assessment of Genes Driving Cancer and Somatic Evolution in Non-Cancer Tissues: An Update of the Network of Cancer Genes (NCG) Resource ». *Genome Biology* 23 (1): 35. https://doi.org/10.1186/s13059-022-02607-z.
14. Ferlay, Jacques, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, et Freddie Bray. 2015. « Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012: Globocan 2012 ». *International Journal of Cancer* 136 (5): E359‑86. https://doi.org/10.1002/ijc.29210.
15. Hanahan, Douglas. 2022. « Hallmarks of Cancer: New Dimensions ». *Cancer Discovery* 12 (1): 31‑46. https://doi.org/10.1158/2159-8290.CD-21-1059.
16. Hoadley, Katherine A., Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, et al. 2018. « Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer ». *Cell* 173 (2): 291-304.e6. https://doi.org/10.1016/j.cell.2018.03.022.
17. Hoof, Ilka, Bjoern Peters, John Sidney, Lasse Eggers Pedersen, Alessandro Sette, Ole Lund, Søren Buus, et Morten Nielsen. 2009. « NetMHCpan, a Method for MHC Class I Binding Prediction beyond Humans ». *Immunogenetics* 61 (1): 1‑13. https://doi.org/10.1007/s00251-008-0341-z.
18. « IntOGen - Cancer driver mutations in Head and neck squamous cell carcinoma ». s. d. Consulté le 31 juillet 2022. https://www.intogen.org/search?cancer=HNSC.
19. Jiang, Ke, Peng Liu, Huizhe Xu, Dapeng Liang, Kun Fang, Sha Du, Wei Cheng, et al. 2020. « SASH1 Suppresses Triple-Negative Breast Cancer Cell Invasion through YAP-ARHGAP42-Actin Axis ». *Oncogene* 39 (27): 5015‑30. https://doi.org/10.1038/s41388-020-1356-7.
20. Khan, Rehan, Ananya Palo, et Manjusha Dixit. 2021. « Role of FRG1 in Predicting the Overall Survivability in Cancers Using Multivariate Based Optimal Model ». *Scientific Reports* 11 (1): 22505. https://doi.org/10.1038/s41598-021-01665-w.
21. Knaus, Brian J., et Niklaus J. Grünwald. 2017. « VCFR : A Package to Manipulate and Visualize Variant Call Format Data in R ». *Molecular Ecology Resources* 17 (1): 44‑53. https://doi.org/10.1111/1755-0998.12549.
22. Knudson, Alfred G. 1971. « Mutation and Cancer: Statistical Study of Retinoblastoma ». *Proceedings of the National Academy of Sciences* 68 (4): 820‑23. https://doi.org/10.1073/pnas.68.4.820.
23. Kurtzer, Gregory M., Vanessa Sochat, et Michael W. Bauer. 2017. « Singularity: Scientific Containers for Mobility of Compute ». Édité par Attila Gursoy. *PLOS ONE* 12 (5): e0177459. https://doi.org/10.1371/journal.pone.0177459.
24. Lin, Shu-Chun, Li-Han Lin, Ssu-Yu Yu, Shou-Yen Kao, Kuo-Wei Chang, Hui-Wen Cheng, et Chung-Ji Liu. 2018. « FAT1 Somatic Mutations in Head and Neck Carcinoma Are Associated with Tumor Progression and Survival ». *Carcinogenesis*, août. https://doi.org/10.1093/carcin/bgy107.
25. Liu, Linlin, Xu Sun, Chenxi Yuan, et Huaimin Liu. 2021. « Relationship between HLA-DRB1 Gene Polymorphism and Breast Cancer: A Protocol for Systematic Review and Meta-Analysis ». *Medicine* 100 (12): e25078. https://doi.org/10.1097/MD.0000000000025078.

26. Martincorena, Inigo. 2019. « Somatic Mutation and Clonal Expansions in Human Tissues ». *Genome Medicine* 11 (1): 35. https://doi.org/10.1186/s13073-019-0648-4.
27. Martincorena, Iñigo, Joanna C. Fowler, Agnieszka Wabik, Andrew R. J. Lawson, Federico Abascal, Michael W. J. Hall, Alex Cagan, et al. 2018. « Somatic Mutant Clones Colonize the Human Esophagus with Age ». *Science* 362 (6417): 911‑17. https://doi.org/10.1126/science.aau3879.
28. Martincorena, Iñigo, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, et Peter J. Campbell. 2017. « Universal Patterns of Selection in Cancer and Somatic Tissues ». *Cell* 171 (5): 1029-1041.e21. https://doi.org/10.1016/j.cell.2017.09.042.
29. McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, et Fiona Cunningham. 2016. « The Ensembl Variant Effect Predictor ». *Genome Biology* 17 (1): 122. https://doi.org/10.1186/s13059-016-0974-4.
30. Pastushenko, Ievgenia, Federico Mauri, Yura Song, Florian de Cock, Bob Meeusen, Benjamin Swedlund, Francis Impens, et al. 2021. « Fat1 Deletion Promotes Hybrid EMT State, Tumour Stemness and Metastasis ». *Nature* 589 (7842): 448‑55. https://doi.org/10.1038/s41586-020-03046-1.
31. Piraino, Scott W., Valentina Thomas, Peter O'Donovan, et Simon J. Furney. 2018. « Driver Versus Passenger Mutations in Tumors ». In *Reference Module in Biomedical Sciences*, B9780128012383650000. Elsevier. https://doi.org/10.1016/B978-0-12-801238-3.65045-6.
32. Prieto, Camila, Diu T. T. Nguyen, Zhaoqi Liu, Justin Wheat, Alexendar Perez, Saroj Gourkanti, Timothy Chou, et al. 2021. « Transcriptional Control of CBX5 by the RNA-Binding Proteins RBMX and RBMXL1 Maintains Chromatin State in Myeloid Leukemia ». *Nature Cancer* 2 (7): 741‑57. https://doi.org/10.1038/s43018-021-00220-w.
33. Riester, Markus, Angad P. Singh, A. Rose Brannon, Kun Yu, Catarina D. Campbell, Derek Y. Chiang, et Michael P. Morrissey. 2016. « PureCN: Copy Number Calling and SNV Classification Using Targeted Short Read Sequencing ». *Source Code for Biology and Medicine* 11 (1): 13. https://doi.org/10.1186/s13029-016-0060-z.
34. Schenck, Ryan O., Eszter Lakatos, Chandler Gatenbee, Trevor A. Graham, et Alexander R.A. Anderson. 2019. « NeoPredPipe: High-Throughput Neoantigen Prediction and Recognition Potential Pipeline ». *BMC Bioinformatics* 20 (1): 264. https://doi.org/10.1186/s12859-019-2876-4.
35. Shukla, Sachet A, Michael S Rooney, Mohini Rajasagi, Grace Tiao, Philip M Dixon, Michael S Lawrence, Jonathan Stevens, et al. 2015. « Comprehensive Analysis of Cancer-Associated Somatic Mutations in Class I HLA Genes ». *Nature Biotechnology* 33 (11): 1152‑58. https://doi.org/10.1038/nbt.3344.
36. Tang, Yiting, Yangyang Geng, Judong Luo, Wenhao Shen, Wei Zhu, Cuicui Meng, Ming Li, Xifa Zhou, Shuyu Zhang, et Jianping Cao. 2015. « Downregulation of Ubiquitin Inhibits the Proliferation and Radioresistance of Non-Small Cell Lung Cancer Cells in Vitro and in Vivo ». *Scientific Reports* 5 (1): 9476. https://doi.org/10.1038/srep09476.
37. The Cancer Genome Atlas Network. 2015. « Comprehensive Genomic Characterization of Head and Neck Squamous Cell Carcinomas ». *Nature* 517 (7536): 576‑82. https://doi.org/10.1038/nature14129.
38. Thomas, Hugh. 2019. « Mutation and Clonal Selection in the Ageing Oesophagus ». *Nature Reviews Gastroenterology & Hepatology* 16 (3): 139‑139. https://doi.org/10.1038/s41575-019-0117-y.

39. Tiwari, Ankit, Bratati Mukherjee, Md. Khurshidul Hassan, Niharika Pattanaik, Archita Mohanty Jaiswal, et Manjusha Dixit. 2019. « Reduced FRG1 Expression Promotes Prostate Cancer Progression and Affects Prostate Cancer Cell Migration and Invasion ». *BMC Cancer* 19 (1): 346. https://doi.org/10.1186/s12885-019-5509-4.
40. Uzunparmak, Burak, Meng Gao, Antje Lindemann, Kelly Erikson, Li Wang, Eric Lin, Steven J. Frank, et al. 2020. « Caspase-8 Loss Radiosensitizes Head and Neck Squamous Cell Carcinoma to SMAC Mimetic–Induced Necroptosis ». *JCI Insight* 5 (23): e139837. https://doi.org/10.1172/jci.insight.139837.
41. Wang, K., M. Li, et H. Hakonarson. 2010. « ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data ». *Nucleic Acids Research* 38 (16): e164‑e164. https://doi.org/10.1093/nar/gkq603.
42. Williams, Drake Winslow, Teresa Greenwell-Wild, Laurie Brenchley, Nicolas Dutzan, Andrew Overmiller, Andrew Phillip Sawaya, Simone Webb, et al. 2021. « Human Oral Mucosa Cell Atlas Reveals a Stromal-Neutrophil Axis Regulating Tissue Immunity ». *Cell* 184 (15): 4090-4104.e15. https://doi.org/10.1016/j.cell.2021.05.013.
43. Williams, Marc J, Luis Zapata, Benjamin Werner, Chris P Barnes, Andrea Sottoriva, et Trevor A Graham. 2020. « Measuring the Distribution of Fitness Effects in Somatic Evolution by Combining Clonal Dynamics with DN/DS Ratios ». *ELife* 9 (mars): e48714. https://doi.org/10.7554/eLife.48714.
44. Zhang, Liang, Mengxia Li, Bo Deng, Nan Dai, Yan Feng, Jinlu Shan, Yuxin Yang, et al. 2019. « HLA-DQB1 Expression on Tumor Cells Is a Novel Favorable Prognostic Factor for Relapse in Early-Stage Lung Adenocarcinoma ». *Cancer Management and Research* Volume 11 (avril): 2605‑16. https://doi.org/10.2147/CMAR.S197855.
45. Zouré, Abdou Azaque, Lanyo Jospin Amegnona, Nayi Zongo, Isabelle Touwendpoulimdé Kiendrebeogo, Pegdwendé Abel Sorgho, Fabienne Ingrid Zongo, Albert Théophane Yonli, et al. 2021. « Carriage of HLA-DRB1*11 and 1*12 Alleles and Risk Factors in Patients with Breast Cancer in Burkina Faso ». *Open Life Sciences* 16 (1): 1101‑10. https://doi.org/10.1515/biol-2021-0113.

ANNEXES

Annexe 1 : Sélection significative chez les 500 gènes les plus mutés parmi ceux exprimés dans les cellules épithéliales de la muqueuse orale saine (en rouge : sous sélection positive, en bleu : sous sélection négative).

Gène	nombre de mutations	dNdS faux sens	dNdS non sens	pglobale ajustée
TUBA1A	39	0	0	0
TUBA1B	35	0	0	0
TUBA4A	46	0,066430074	0	1,10E-12
NDUFS7	26	0	0	1,76E-09
AHNAK2	164	0,391028417	0	1,35E-06
RBMXL1	6	0	58,61270048	1,78E-05
TUBB2A	55	0,213348781	0	4,19E-05
TUBA1C	13	0	0	0,00012905
CSNK1E	13	0	0	0,000281209
SASH1	5	0	12,68362973	0,000503154
EPHA2	15	0,032419288	0	0,000707572
BOD1	26	0,108161858	0	0,000803181
BPTF	7	0	0	0,000803181
RNF213	5	0,11141418	0	0,000950075
CREBBP	10	0	0	0,001144543
CASP8	9	1,2133029	40,84149328	0,001223967
FAT1	5	0,129604792	2,431150672	0,00155479
SETD2	5	0	6,636784454	0,001584979
KLC1	10	0	0	0,001677825
UBC	14	0,03853545	1,388672551	0,003924971
CNTNAP3B	32	0,300477476	0	0,003924971
MEF2D	10	0	0	0,004379485
MEF2C	13	0,040776295	0	0,005009977
HERC2	44	0,250420933	0,194023655	0,005916451
NOTCH1	19	0,887517575	8,446226249	0,005916451
HLA-DQB1	25	24,27614911	0	0,006856237
RYR1	7	0,212867607	0	0,006856237
SYNE1	15	0,435839997	0	0,006978579
HLA-DRB1	87	1,29647117	4,557500034	0,009464965
KDM5C	8	0	0	0,011209525
SDHA	14	0,147915532	0	0,013898429
FRG1	10	9,058940048	21,2951012	0,013898429
PLEC	7	0,218885538	0	0,013898429
HIPK2	8	0	0	0,014717785
GJA1	8	0	0	0,014717785
TNRC6A	7	0	0	0,015009367
MED27	8	0	0	0,016973201
NUP54	11	0,121594645	0	0,018006688
NFIX	14	0,085452855	0	0,021646127
AHNAK	7	0,323910349	0	0,024930865
AHCTF1	5	0	0	0,028506768
PKD1	14	0,205302655	0	0,029499418
HSPA1B	60	0,386569307	0	0,036020297
FAM98B	7	0	0	0,036020297
HLA-B	25	11,10750402	0	0,037371432

HUWE1	7	0,345225236	0	0,038042551
MAP2K7	7	0	0	0,048842977
TNRC18	8	0,266898918	0	0,048866758

Annexe 2 : Sélection non significative chez les 500 gènes les plus mutés parmi ceux exprimés dans les cellules épithéliales de la muqueuse orale saine (en rouge : sous sélection positive, en bleu : sous sélection négative).

Gène	nombre de mutations	dNdS faux sens	dNdS non sens	pglobale ajustée
TUSC1	6	5,232801475	0	0,050993338
UBB	7	0	0	0,066260213
COL18A1	6	0,156353915	0	0,066260213
CKMT1A	6	0	0	0,069497297
NUMA1	6	0,287093041	0	0,076322619
CBFA2T2	6	0	0	0,076322619
HLA-DPB1	13	4,590378256	0	0,076492644
POLR2J3	7	0	0	0,076712579
HLA-A	45	1,287426917	0	0,076712579
RAB40C	10	0,063540845	0	0,080640038
MYH9	15	0,124124823	0,486229323	0,081280477
ANKRD11	25	2,625480273	0	0,083397203
CNTNAP3	18	0,329994133	0	0,083397203
USF3	5	0,289685431	0	0,084824037
MAU2	6	0	0	0,091647598
ITGB4	8	0,430033043	7,648030017	0,09234652
PRKRA	13	4,119228392	0	0,09234652
RBM25	5	0	0	0,097688418
TRAPPC2L	6	0	0	0,104939175
GNA11	6	0	0	0,106522797
CARD16	6	0	0	0,106522797
NPEPPS	17	4,753128654	9,795718472	0,108885508
CCZ1	12	7,515555881	0	0,109705436

Annexe 3 : Sélection significative chez les 1000 gènes les plus mutés parmi ceux exprimés dans les cellules épithéliales de la muqueuse orale saine (en rouge : sous sélection positive, en bleu : sous sélection négative).

Gène	nombre de mutations	dNdS faux sens	dNdS non sens	pglobale ajustée
TTN	44	0,304480678	0	0
NPIPB4	386	0,404410712	0,284220212	0
NBPF20	1147	2,294764132	1,455751956	0
NBPF10	368	2,011506845	0	0
TUBA3E	29	0	0	1,54E-11
ANKRD20A1	65	25,41085274	0	7,15E-10
USP17L17	370	1,866942725	3,501132353	6,72E-09
USP17L22	347	1,382802493	3,530200274	1,32E-08
TUBB2B	54	0,135483261	0	3,64E-08
MUC17	529	1,656735512	0	5,30E-08
USP17L19	314	1,811429116	3,600784784	5,94E-08
MGAM	21	0,020265799	0	6,60E-08
TBC1D3L	144	2,611449617	0	1,29E-07

USP17L24	326	1,65024983	3,371811934	2,26E-07
USP17L20	402	1,398642185	3,15455176	2,26E-07
USP17L27	331	1,821502081	2,924487933	2,59E-07
USP17L15	366	1,323540852	3,376911892	2,84E-07
USP17L10	94	5,512072193	11,56173058	4,71E-07
USP17L29	269	2,158174166	2,575865116	5,59E-07
TBC1D3H	130	2,575006978	0	6,65E-07
USP17L18	375	1,42742901	3,193306717	9,50E-07
USP17L13	243	1,552787703	3,749293161	1,00E-06
USP17L25	325	1,806693121	2,308806263	1,50E-06
USP17L26	324	1,518690194	3,03289975	1,86E-06
USP17L5	326	1,580866042	2,811459068	3,04E-06
PRAMEF25	150	0,43068858	0,254816148	4,03E-06
CDRT15	9	0	44,72666643	4,39E-06
IGHV3-30	19	0	3,568255874	4,39E-06
USP17L30	234	2,094479233	1,419747636	4,57E-06
ATP2B2	20	0,024364926	0	4,57E-06
USP17L11	415	1,159697292	2,472588915	4,73E-06
OR2T2	14	0,871983612	51,57432161	5,10E-06
TBC1D3C	146	2,279062705	0	9,19E-06
PRDM9	25	0,075826033	0	9,74E-06
TAS2R43	10	4,39316441	156,0547823	1,23E-05
USP17L28	302	1,621581121	2,770567515	1,29E-05
KIR2DL4	221	1,076625323	0	1,40E-05
MUC4	857	1,093209051	0,14371738	1,43E-05
NOMO2	22	2,800006024	24,34217227	1,77E-05
LILRB4	21	0,051804203	0	3,59E-05
TBC1D3D	129	1,757075945	0	5,16E-05
OR11H12	44	8,064154673	0	5,16E-05
LPA	12	0	0	5,87E-05
TUBB4A	24	0,071725386	0	6,05E-05
FAM205A	34	0,262334983	3,390583837	6,05E-05
CR1	20	0,067043075	0	6,76E-05
MUC16	21	0,347113626	0	7,34E-05
GAGE2A	8	4,951381632	140,8754279	9,08E-05
TBC1D3E	126	2,230899633	0	9,51E-05
TBC1D3I	155	1,981020445	0,319261796	0,000124441
PRAMEF26	158	0,517108222	0,265881799	0,000128014
PRAMEF9	150	0,475273073	0,255554192	0,000133498
PRAMEF4	156	0,49003205	0,244246385	0,000173313
USP17L21	289	1,341506207	2,199286682	0,000189866
POTED	32	17,58530277	0	0,000249052
FOXD4L6	18	7,297286743	97,9458659	0,000396681
PRAMEF6	173	0,569790171	0,253755706	0,000396681
GOLGA8J	96	2,344227581	0	0,000396681
USP17L12	290	1,242882178	2,080478127	0,000432079
TBC1D3	126	1,902859744	0	0,000490284
GAGE13	20	37,98690867	115,5696004	0,000536451
GAGE12G	14	28,17291862	45,43989669	0,000548675
NBPF15	91	2,494109636	0	0,000706884
TSPY3	17	0,30063257	5,289724043	0,001037617
KRTAP9-2	7	0	0	0,001046745
ZNF705A	10	0	0	0,001046745
PCDH11X	41	0,250641942	0,34728121	0,001275448
PRAMEF12	53	0,424577705	0	0,001275448
SOX1	10	3,416570209	0	0,00133087

TUBB3	22	0,107778429	0	0,001404577
GOLGA8F	96	1,362337288	0	0,001406126
NPIPB3	101	0,515011768	0	0,001468032
PRAMEF15	174	0,575500394	0,231932	0,001468032
TSPY10	19	0,320091093	4,49484511	0,001468032
TAF7L	11	0	0,440838847	0,001535568
TCAF2	21	6,90693275	0	0,001541762
OR2T33	24	0,125111215	0	0,001562099
FLG2	28	0,174101858	0	0,001630102
FOXD4L3	20	5,872331093	72,96295414	0,00165355
GAGE12F	10	17,92822094	45,43989669	0,001801533
CCDC168	6	0,19298518	0	0,001806475
LILRB3	25	0,152568468	0	0,002044529
MUC20	70	0,935486892	5,417479097	0,002135893
TUBA3D	10	0	0	0,002455847
TAS2R46	11	7,73256544	92,17196308	0,002602831
KIR2DL1	148	0,689478915	0	0,002660115
POTEC	9	0	0	0,002675589
TSPY9P	16	0,884120263	10,37938433	0,002793198
GOLGA8G	95	1,01516371	0	0,002967411
ARFGEF3	5	0	0	0,002967411
TSPY4	17	0,66701779	7,903272953	0,003115688
KIR3DL3	69	0,92084656	0	0,003115688
MUC5B	163	0,639478324	0	0,003132043
GOLGA8O	101	0,544468469	1,204431284	0,00328848
NPIPA8	28	21,52344945	0	0,003474518
GAGE1	8	10,89584458	43,93991243	0,003495675
GOLGA8T	75	1,349353706	0	0,004026017
TPSB2	16	0,079994572	0	0,005284554
SPDYE6	63	2,079884494	0	0,005360308
GAGE12E	7	15,72703616	0	0,005360308
TBC1D3G	150	1,844048693	0,793038976	0,005552397
PRAMEF27	109	0,427806255	0,332015077	0,005703983
LRRC37A	21	3,537617894	0	0,005760977
KRTAP4-3	10	13,52175956	0	0,006605216
GOLGA6L6	22	0,298707297	0	0,006605216
MUC5AC	46	2,254542703	0	0,006954184
GPR173	9	0	0	0,007069921
MYH4	13	0,071939094	0	0,007525071
LRRC37A3	27	4,438422522	0	0,007707719
FAM86B2	75	0,469124073	1,052171832	0,007707719
TSPY1	14	1,309756896	12,31655071	0,009006394
GAGE12D	6	13,10586347	0	0,009006394
TSPY2	15	0,222052091	5,240736924	0,009012711
POTEB3	30	0,236827466	0	0,009385923
SAA2	6	12,18133739	0	0,009787613
OR4F21	33	13,05307853	0	0,009787613
NPIPA5	57	0,588993343	0	0,009787613
ZFHX4	6	0,084703959	0	0,01008163
FAT3	11	0,262716397	0	0,010625957
CLEC18B	8	0	0	0,011702824
PRAMEF11	115	0,725822999	0	0,011770992
ANKRD30A	16	1,079122431	11,99298129	0,01180126
GOLGA8S	112	1,08104756	0	0,012283093
PRB2	57	0,600113183	0	0,012377898

KIR3DL2	102	0,599492395	0	0,012377898
ZNF813	7	0	0	0,013342549
OR2L8	17	16,8050514	0	0,01461974
VN1R4	24	0,306928743	0	0,01461974
KRT86	8	0	0	0,01461974
RGPD5	36	0,439620293	0	0,015065434
SPDYE2B	67	1,499275229	0	0,015065434
USH2A	6	0,240742107	0	0,015773021
C10orf90	7	1,119836836	23,03560817	0,016328966
BMP8A	8	0	0	0,018115276
OR2T3	16	14,88054548	0	0,018115276
NPIPA2	16	0,242804249	0	0,018262143
FAM47A	5	0,34014214	28,38958553	0,019372565
MAGEA12	8	0	0	0,020087023
OR2T5	17	14,62497275	26,03702324	0,020107788
RGPD8	43	0,549580277	0	0,022907177
TSPY8	12	0,192442921	0	0,023031004
MUC12	538	1,220544114	0,490101713	0,024867761
PRAMEF8	80	0,477008622	0	0,025599791
HRC	64	0,612189538	0	0,026130121
SAA2-SAA4	6	7,12504391	0	0,026291085
RGPD6	40	0,538718148	0	0,027199769
PRAMEF19	7	0	0	0,027263193
GAGE12J	12	26,06457322	45,34721085	0,027392813
FAM153A	24	8,709546778	0	0,029107476
PRAMEF5	218	0,792084843	0,242500959	0,029107476
PPIAL4C	13	22,10439113	0	0,030330609
PLXNA4	12	0,088150905	0	0,030495248
PRB3	12	8,296074296	0	0,030517362
RFPL1	7	0	0	0,030737886
CYP2D6	11	3,679644325	16,67554548	0,031676562
CDH23	5	0,314046904	0	0,031676562
SDK1	7	0,242679093	6,083246844	0,031676562
ADGRG4	6	0,210289987	0	0,03181199
SMN1	7	0	0	0,032397305
KIR3DL1	96	0,629145367	0	0,032639715
ANKRD30B	6	0	0	0,033064085
SPATA31A6	89	2,157190471	0	0,033135488
SMN2	7	0	0	0,0332658
KRT81	7	0	0	0,033374413
SPDYE2	63	0,983172995	0	0,033423133
UGT2B11	16	9,923362387	0	0,03424551
CT45A3	12	19,0644499	0	0,03424551
CT45A9	12	19,1274255	0	0,03424551
CT45A2	12	19,1274255	0	0,03424551
CT45A8	12	19,1274255	0	0,03424551
KCNC3	18	0,154920006	0	0,038066107
OR4F5	15	0,279189746	0	0,038806214
SPATA31A1	94	1,971125705	0	0,042523205
CFHR5	6	0	0	0,042523205
GOLGA8H	83	0,997765688	0	0,043535077
GOLGA6A	38	1,480597708	0	0,044326257
NUTM2B	34	1,071504834	3,815056696	0,044975788
NBPF26	164	1,316835044	0,621761908	0,045869589
HOXB1	9	6,617357153	0	0,046905418

SPDYE16	13	10,43675661	0	0,046990091
XIRP2	6	0,173693207	0	0,04837377
OR51A4	13	13,05325594	0	0,04837377
SPATA31A7	96	0,831942785	0	0,049720623

Annexe 4 : Sélection non significative chez les 1000 gènes les plus mutés parmi ceux exprimés dans les cellules épithéliales de la muqueuse orale saine (en rouge : sous sélection positive, en bleu : sous sélection négative).

Gène	nombre de mutations	dNdS faux sens	dNdS non sens	pglobale ajustée
LRRC43	9	3,45735728	0	0,050691835
LRRC37A2	72	1,350107237	0	0,052923844
EEF1A2	19	0,187501937	0	0,052923844
POLR2J2	7	0	0	0,05430765
PSG8	24	9,170772376	0	0,05430765
NUTM2E	24	0,895983362	3,763873124	0,05430765
OR4F16	37	4,938151919	0	0,054534722
GOLGA8R	88	0,632226226	1,232338245	0,057085773
GOLGA6B	39	1,143103298	0	0,057085773
LRP2	5	0,319134693	0	0,05749672
TCP10L2	6	2,519129223	28,34637184	0,060296078
KIAA0513	9	1,719199175	4,131710357	0,062314134
BOD1L2	7	0	0	0,063649167
FCGR2A	6	0	0	0,06382158
OR2T35	6	3,584307917	60,39346124	0,064514667
KCND3	6	0	0	0,0651697
GOLGA8M	74	1,013140739	0	0,067357263
KRTAP3-3	7	0	0	0,06782479
MYH1	8	0,060055179	0	0,067867007
OR11H2	28	1,209230793	4,995764635	0,068353211
NUTM2A	32	0,737029902	0	0,069742738
ZNF423	11	0,099407126	0	0,069786554
TUBA3C	6	0	0	0,071684309
SLC35G6	11	9,728802311	0	0,072473436
NLGN3	11	0,102151125	0	0,072473436
GOLGA8Q	70	1,171962241	2,464405099	0,072569699
FOXD4L4	13	1,095441117	0	0,073364346
PYGM	19	0,668360309	0	0,075069066
PGA5	6	0	0	0,075069066
NEB	16	0,432491901	0,689059811	0,079648774
PCDHB15	6	0	0	0,080982556
OR10H5	6	0	0	0,081928237
OR2T12	31	0,309140602	0	0,081928237
NOMO3	14	0,71567725	7,435182533	0,08283638
ADRA1B	5	0,454772792	0	0,083131544
DNAH3	6	0,29489186	0	0,083879133
GAGE10	7	13,20946001	46,85834509	0,085362997
GOLGA6L1	35	0,917966058	0	0,08572051
GOLGA6C	31	0,94501334	0	0,090336382
SSX2	11	0,116009748	0	0,091348277
OR10H2	11	9,57949019	0	0,09207983
NPIPA7	28	5,088170146	0	0,096231131
LCN1	8	12,43708724	0	0,096554182
TRIM64B	13	3,694890632	0	0,096809794

PCDHB8	35	0,343868497	0	0,097380609
IGHE	12	6,159056141	0	0,097380609
DEFB104A	6	24,91879687	0	0,097380609
DEFB104B	6	24,91879687	0	0,097380609
SPDYE5	34	0,904483666	0	0,097394445
PRR23D1	10	10,88413078	0	0,098589768
AKR7A3	14	0,470679151	0	0,099028377
GOLGA8N	89	0,718569178	1,333301055	0,099028377
KRTAP10-7	11	8,442831974	0	0,099394826
LILRB1	15	0,187565307	0	0,099701314
GOLGA6L22	29	1,398702511	0	0,099701314
PRAMEF10	63	2,232293137	0	0,109956744
UBTFL1	29	1,353368732	0	0,109956744

RÉSUMÉ

L'évolution somatique est caractérisée par l'accumulation de mutations somatiques lors des réplications cellulaires. Une cellule mutante devient cancéreuse quand les mutations impactent la fonction de gènes clé lui permettant de proliférer de manière incontrôlée dans son environnement. L'identification de ces gènes est cruciale pour la personnalisation du traitement. La muqueuse orale, accessible par des moyens non-invasifs, est un bon candidat pour l'étude de l'évolution somatique en amont de la tumorigenèse, et qui en dicte le contexte et la dynamique initiale. De plus, c'est un processus encore mal décrit dans cette région, d'où la mise en place de mon projet de stage. Nous avons retrouvé sous sélection positive les principaux gènes drivers (NOTCH1, TP53, FAT1) du principal cancer affectant cette zone, le cancer épidermoïde de la tête et du cou. Nous avons également identifié d'autres gènes non-drivers sous sélection positive, qui méritent d'être étudiés plus en détail. Nous avons trouvé que la méthode de prélèvement d'échantillon par cytobrosse est la plus adaptée pour cette catégorie d'étude. La présence de néo-antigènes à forte antigénicité a été confirmée au sein des cellules saines. Aucune altération de nombre de copies n'a été trouvée, comme attendu pour des cellules saines.

ABSTRACT

Somatic evolution is characterized by the accumulation of somatic mutations during cell replication. A mutant cell becomes cancerous when mutations impact the function of key genes allowing the cell to acquire a growth and invasion advantage, promoting tumorigenesis. The identification of these genes is crucial for the personalization of treatment. The oral mucosa, accessible by non-invasive means, is a good candidate for the study of somatic evolution. Moreover, it is a process that is still poorly described in this region, hence the implementation of my internship project. We found positive selection operating on the main driver genes (NOTCH1, TP53, FAT1genes) of the main cancer affecting this area, the head and neck squamous cell cancer. We have also identified other non-driver genes under positive selection, which deserve further investigation. We have found that the cytobrush sample collection method is the most suitable for this type of study. The presence of neo-antigens with strong antigenicity has been confirmed in healthy cells. No copy number alterations were found, as expected for healthy cells.