Université de Montréal

Déconvolution des types cellulaires et quantification des infiltrats immunitaires dans les cancers pédiatriques

par

Mia Cherkaoui

Département de Biochimie et Médecine Moléculaire

Faculté de Médecine

Mémoire présenté en vue de l'obtention du grade

de Maître ès sciences (M. Sc.) en Bio-informatique

Août, 2021

Résumé

Les cancers pédiatriques sont responsables du plus haut taux de mortalité lié aux maladies à l'échelle mondiale. Parmi eux, la leucémie lymphoblastique aiguë reste le plus répandu et le plus meurtrier. Malgré un taux de survie avoisinant les 90 %, beaucoup d'enfants connaissent des rechutes et de lourds effets secondaires.

L'hétérogénéité en sous types de cette maladie complique les avancées. Les outils de nouvelle génération semblent être un moyen plus rapide pour améliorer cette caractérisation de sous-types à des fins thérapeutiques. Des outils de déconvolution sont alors utilisés pour estimer et vérifier la pureté tumorale des patients. Cette dernière, corrélée avec la croissance tumorale ou encore le pronostic, est un paramètre important à considérer.

Le projet s'est aussi penché sur l'implication des cellules immunitaires dans le microenvironnement tumoral de tumeurs solides. L'hypothèse étant que certaines cellules non cancéreuses influenceraient la progression tumorale. L'utilisation d'outils permettrait de distinguer les composants de cet environnement pour en déduire leur rôle spécifique.

Parmi les résultats obtenus, le choix d'outils adaptés aux données est un paramètre important. L'utilisation combinée de plusieurs outils favorise une meilleure interprétation des cancers. Ces derniers mettent en avant certains composants influençant le devenir de la tumeur ou encore le pronostic, et permettront ainsi de mieux classifier et prendre en charge le patient.

L'objectif final étant d'intégrer des approches plus personnalisées en fonction de la stratification effectuée, moins invasives, et qui mèneront à l'augmentation du taux de survie.

Mots-clés : cancers pédiatriques, leucémie lymphoblastique aiguë, outils bio-informatique, ARN-seq, déconvolution, infiltration immunitaire, tumeurs solides

Abstract

Pediatric cancers are responsible for the highest disease-related mortality rate worldwide. Among them, acute lymphoblastic leukemia remains the most common and the deadliest one. Despite a survival rate of 90 %, many children experience relapses and severe side effects.

The heterogeneity in subtypes of this leukemia slow down progress. Next generation tools appear to be a faster way to improve this subtypes' characterization for therapeutic ends. Deconvolution tools are used to estimate and verify tumor purity in patients. The latter, known to be correlated with tumor growth and prognosis, seems to be an important parameter to analyze.

This study also focused on the involvement of immune cells in the tumor microenvironment of solid tumors. The hypothesis being that some non-cancerous cells influence tumor progression. The use of tools can help to distinguish the components of this environment and deduce their specific role.

Among the results, the choice of tools suitable for our data is an important parameter to consider. The combined use of several tools promotes a better interpretation of cancers. By highlighting specific components influencing the tumors' future or the prognosis, they will allow a better classification and facilitate the patients' care.

The ultimate goal is to integrate more personalized therapies depending on the stratification performed, which will be less invasive, and which will hopefully lead to an increased survival rate.

Keywords: pediatric cancers, acute lymphoblastic leukemia, RNA-seq, bioinformatic tools, deconvolution, immune infiltrate, solid tumors

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

LLA : leucémie lymphoblastique aiguë

FPKM : Fragments per kilobase million : fragments par million de kilobase

Log2 : logarithme base 2

ARN-seq : séquençage ARN

scARN-seq : séquençage ARN « single cell »

Lymphocyte NK : lymphocytes « Natural Killer »

T reg : lymphocytes T régulateurs

CLP : cellule lymphoïde progénitrice

Terme GO : terme Gene Ontology

TARGET : Therapeutically Applicable Research To Generate Effective Treatments

DFCI : Dana-Farber Cancer Insitute

QcALL : Cohorte de la leucémie lymphoblastique aigue au Québec

TRICEPS : Personalized Targeted Therapy in Refractory/Relapsed Cancer in Childhood Study

SIGNATURE : omic-based translational research strategy for newly diagnosed cancers

TNM : Tumeur, Nodes (Ganglions Lymphatiques), Métastases

Remerciements

Je tiens à remercier tous ceux qui m'ont aidé à réaliser ce projet de près ou de loin.

Tout d'abord mon directeur de recherche Daniel Sinnett pour m'avoir permis d'intégrer son laboratoire le temps de ma maitrise, mais aussi les membres du laboratoire, spécialement l'équipe de bio-informatique pour toute l'aide apportée, Pascal St Onge et Maxime Caron.

Je remercie également toute ma famille, plus spécialement mes parents et mes grands-mères, mais aussi mes amis pour leur soutien sans faille tout au long de mes études malgré la distance.

- Chapitre 1 -

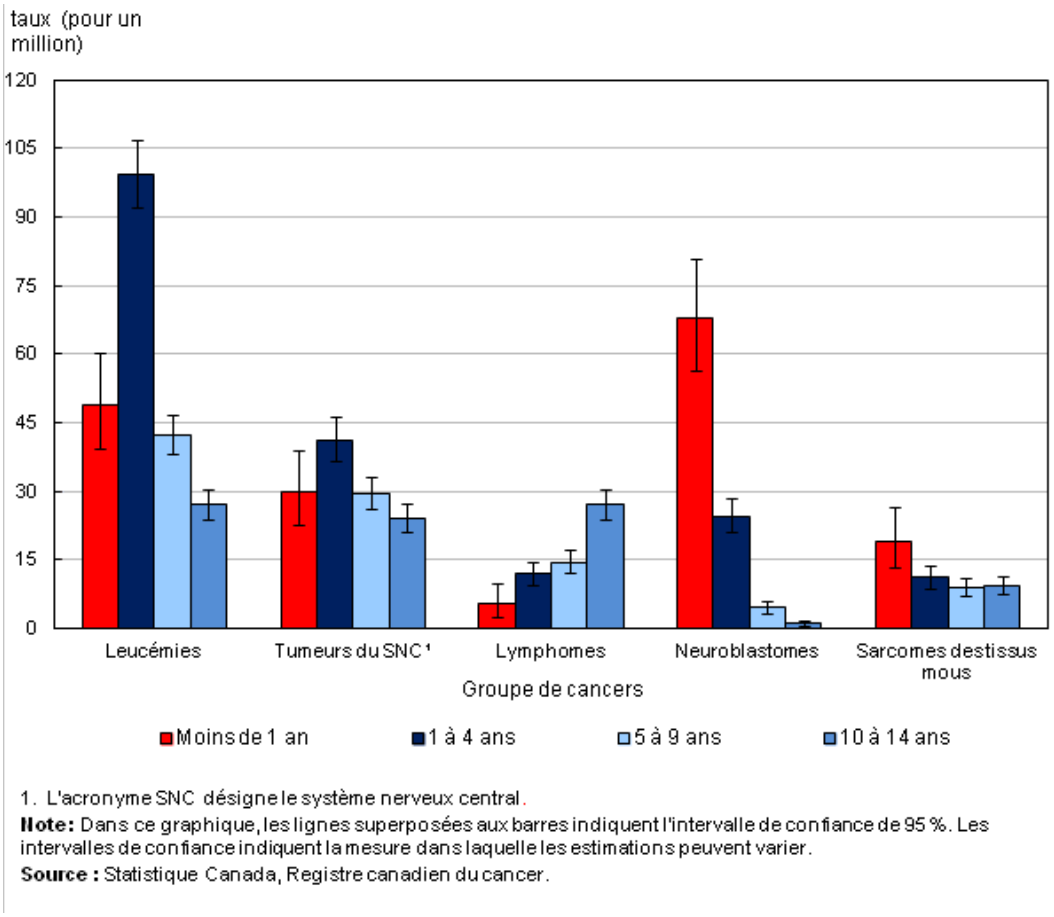
Introduction

1. Les cancers pédiatriques

Les cancers pédiatriques, même s'ils sont moins communs que chez les adultes, connaissent tout de même le plus haut taux de mortalité lié aux maladies chez les individus entre 6 mois d'âge et jusqu'au stade de l'âge adulte (1). Encore aujourd'hui, les facteurs déclenchant ces cancers chez les enfants sont pour beaucoup d'entre eux non élucidés ou même inconnus, notamment du fait que l'incidence de ces cancers, se situant majoritairement dans les cinq premières années de vie, réduit le potentiel effet causateur du mode de vie de l'individu ou de son environnement (2, 3).

Des études de Statistique Canada ont également rapporté que l'incidence de ces cancers a connu une hausse de 0,4% par année entre 1992 et 2010 au niveau national, avec des statistiques similaires rapportées au niveau international dans des pays comme les États-Unis, l'Australie ou d'autres pays d'Europe, amenant ainsi d'autant plus de chercheurs à trouver des solutions viables pour améliorer les chances de guérison de ces enfants (4).

De nos jours au Canada, plus de 80% des mineurs diagnostiqués d'un cancer survivent, mais la plupart d'entre eux connaitra des rechutes ou des effets secondaires importants pour le reste de leur vie (5). À l'échelle mondiale, c'est plus de 300 000 enfants et adolescents qui sont diagnostiqués chaque année, et une grande partie de ces enfants est dans des pays en voie de développement, où énormément d'entre eux meurent à cause du manque de diagnostic rapide et de traitements efficaces (5).



1. L'acronyme SNC désigne le système nerveux central.
Note: Dans ce graphique, les lignes superposées aux barres indiquent l'intervalle de confiance de 95 %. Les intervalles de confiance indiquent la mesure dans laquelle les estimations peuvent varier.
Source : Statistique Canada, Registre canadien du cancer.

Figure 1. – Incidence des cancers pédiatriques entre 2006 et 2010 au Canada. Incidence chez les enfants depuis la naissance jusqu'à leurs 14 ans, en fonction du groupe de cancer et de l'âge. Figure tirée de Ellison et Janz, 2015 (4).

Les tumeurs pédiatriques ont un mode de développement et de propagation plus rapide que ceux observés chez les adultes (2, 3). De ce fait, les tumeurs pédiatriques affectent les organes de manière différente, ce qui ajoute une part de complexité dans leur étude (2, 3). De plus, les incidences en fonction du groupe de cancer varient également (4). En pédiatrie, les cancers dits liquides, tels que les cancers du sang, sont plus fréquents que les cancers avec des tumeurs solides. Ces cancers liquides ont leurs propres spécificités et se manifestent différemment (6).

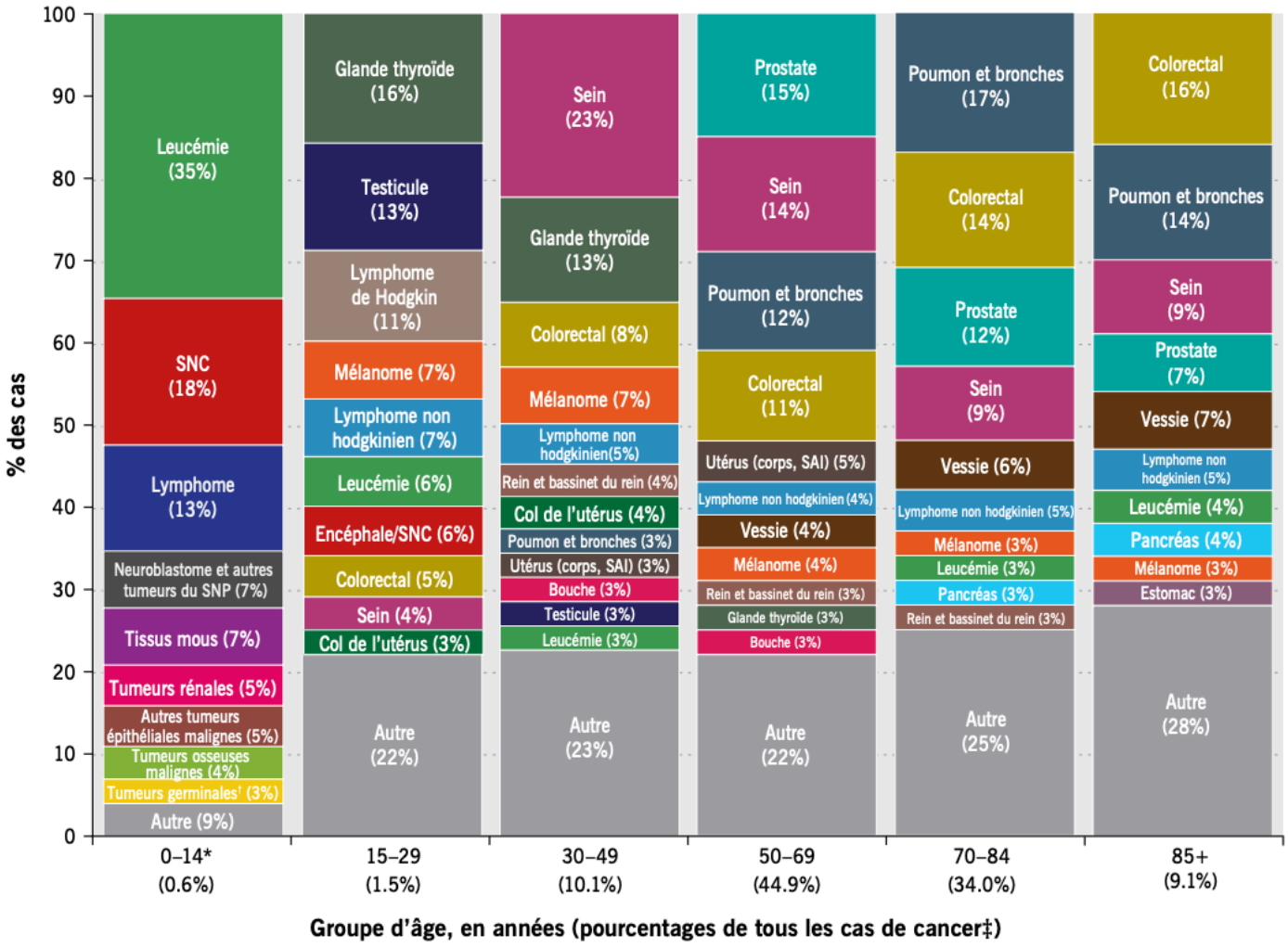


Figure 2. – Distribution des nouveaux cas de cancers selon le groupe d'âge au Canada entre 2011 et 2015. SNC : système nerveux central; SNP : système nerveux périphérique; NOS : sans autre indication. Le pourcentage relatif est calculé en fonction du nombre total de cas sur les 5 ans pour chaque groupe d'âge. Les cancers diagnostiqués chez les enfants (0 à 14 ans) ont été classés selon le Programme de Surveillance, épidémiologie et résultats finaux (SEER), mis à jour par la Classification internationale des cancers de l'enfant (ICCC). Les cancers diagnostiqués chez les personnes plus âgées ont été classés selon la Classification internationale des maladies pour l'oncologie (CIM-O). Figure tirée de *Statistiques canadiennes sur le cancer 2019* (1).

Dans les cas des leucémies, les cellules cancéreuses s'accumulent dans la moelle osseuse et le sang, ce qui leur donne une capacité à se répandre d'autant plus rapidement puisque ces cellules circulent directement dans le sang (6).

2. La leucémie lymphoblastique aiguë

Lors d'un processus d'hématopoïèse normal, les cellules souches hématopoïétiques se différencient d'abord en deux types cellulaires majeurs : la lignée myéloïde ou la lignée lymphoïde (7). La suite de la différenciation de la lignée myéloïde est responsable des globules rouges, des plaquettes, des monocytes et des granulocytes, tandis que la lignée lymphoïde se différencie en lymphocytes cytotoxiques naturels (NK) ou en lymphocytes B et T (7).

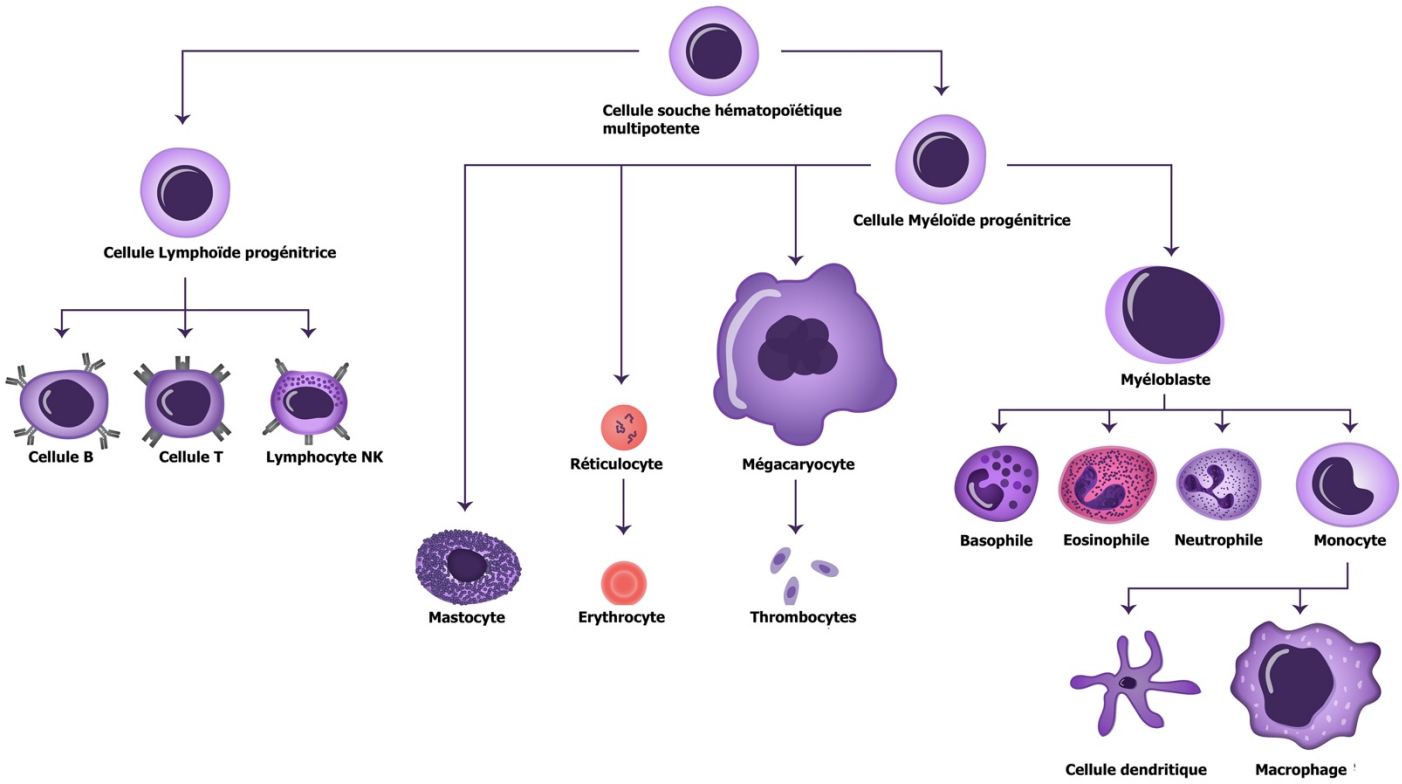


Figure 3. – Schéma de la différenciation cellulaire hématopoïétique. Figure adaptée de Janeway CA Jr et al., 2001 (8).

Lorsqu'une perturbation se produit dans une de ces lignées, ici particulièrement la lignée des lymphocytes B ou T, c'est une cellule B ou T immature qui est alors génétiquement modifiée, et qui va entrainer la perturbation du processus de maturation subséquent, ainsi que sa prolifération incontrôlée (9). Une accumulation de ces clones dans la moelle provoquera éventuellement la suppression de l'hématopoïèse normale et laissera place à une population majoritaire de blastes leucémiques (9).

Mon projet de recherche s'est alors d'abord intéressé à un type de cancer en particulier, la leucémie lymphoblastique aiguë (LLA), soit le cancer pédiatrique le plus répandu puisqu'il représente plus de 30% des cancers pédiatriques, qui, malgré un taux de survie dépassant les 90 %, présente un grand nombre de rechutes et de complications liées aux traitements, notamment le décès (10). De nouvelles avenues thérapeutiques sont requises, mais le manque d'une caractérisation complète dû à la multitude de sous types existants dans cette leucémie rend la tâche complexe (9, 10).

Il faut savoir qu'il existe plusieurs sous-types LLA, associés à divers pronostics (7, 9).

2.1 Sous types majeurs

Dans 85% des cas de LLA, les patients sont atteints de leucémie impactant la lignée lymphoïde des cellules B. La leucémie de type T est quant à elle moins fréquente mais plus agressive (10, 11).

Modifications moléculaires fréquentes

Les LLA possèdent une multitude de sous types caractérisée par des altérations moléculaires qui apparaissent lors de la leucémogenèse (i.e. événements somatiques) (12).

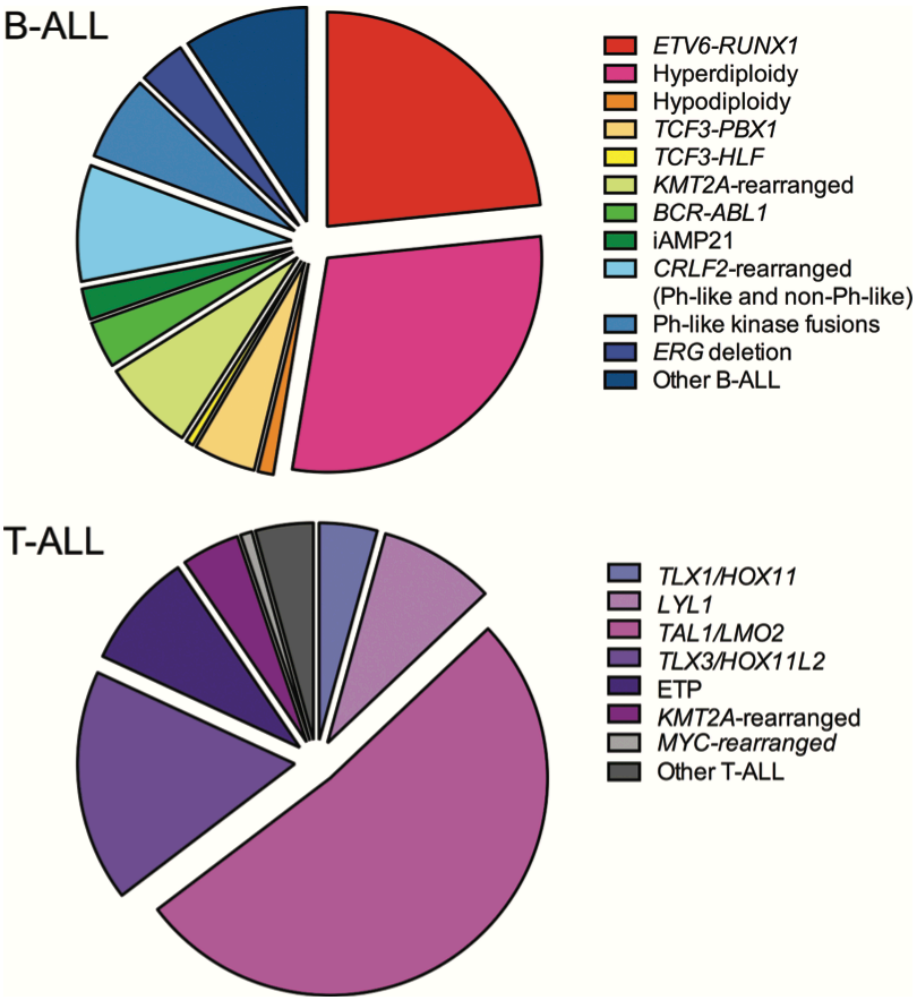


Figure 4. – Fréquence des sous-types cytogénétiques dans les LLA pédiatriques. Chaque altération cytogénétique représente un sous-type ayant une réponse différente dans les leucémies de type B et T. Les données des altérations génétiques submicroscopiques ne sont pas représentées. Figure tirée de Tasian et al., 2015 (11).

D’un point de vue moléculaire, les cellules leucémiques sont affectées par de multiples modifications génétiques et épigénétiques (10-12). La grande diversité de ces changements génétiques et la perturbation des voies biologiques associées, fait de ce type de cancer une maladie hétérogène.

Plusieurs centaines de ces altérations génétiques ont déjà été identifiées. Parmi celles-ci, on compte des changements du nombre de chromosomes (aneuploïdie), des réarrangements chromosomiques tels que des translocations (fusions de gènes), des mutations ponctuelles ou encore des insertions et délétions (10-12). Les translocations chromosomiques et les réarrangements intra-chromosomaux sont d’ailleurs souvent considérer comme évènements initiateurs des leucémogenèses (10-12). Toutefois, même si l’événement initiateur de ces mutations est connu dans le processus leucémique, les éléments enclenchant cette leucémie ne sont pas encore tous établis (10-12).

De manière plus spécifique, dans environ 30% des cas, les leucémies impliquant les cellules B sont caractérisées par des évènements de gains chromosomiques comme l’hyperdiploïdie (>50 chromosomes) dans environ 30% des cas (10-12). Ce niveau de ploïdie a généralement un impact sur le pronostic et est associé à une réponse plutôt positive suite à une chimiothérapie dans à peu près 90 à 95% des cas (10, 12). Au contraire, une hypodiploïdie (< 44 chromosomes) est quant à elle associée à de mauvaises réponses, mais cette dernière survient beaucoup plus rarement, dans environ 1 à 2% des cas (10, 11). Concernant les translocations spécifiques, la plus commune dans les leucémies de la lignée B est la translocation t(12 ;21) qui se produit dans 25% des cas (11). Cette translocation créée un gène de fusion entre le facteur de transcription RUNX1 et le répresseur de transcription ETV6, étant tous les deux des régulateurs de l’hématopoïèse; avec ETV6 étant essentiel au niveau de la moelle osseuse pour l’établissement de cette hématopoïèse (11). Cette translocation est également associée à des pronostics plutôt favorables (13).

Du côté des leucémies de type T, on retrouve également des évènements récurrents. Le gène NOTCH1 par exemple, est activé par une mutation oncogénique dans plus de 60% des cas (14, 15). Ce dernier est un élément clé dans le développement du thymocyte ainsi que dans le destin des cellules T (14). L’inactivation de CDKN2A, un suppresseur de

tumeurs, et l'activation de NOTCH1, constituent deux altérations fréquentes qui prédisent une réponse plutôt favorable dans les leucémies de type T (10, 14, 15). D'autres réarrangements chromosomiques, comme ceux impliquant les récepteurs des cellules T peuvent également mener à l'expression de facteurs de transcription capables d'agir à titre d'oncogènes, mentionnons par exemple les translocations activant des facteurs comme TLX3 ou TAL1, provoquant un arrêt la différenciation (14).

Il existe également plusieurs translocations qui impliquent un réarrangement du gène MLL (10). Ce dernier code pour une méthyl-transférase impliquée dans la transcription, et plus particulièrement, dans la régulation épigénétique (10). Ces translocations sont présentes dans environ 80% des leucémies infantiles (< 1 an) dans lesquelles les partenaires de fusions sont différents dépendamment du sous type (10). En particulier, la translocation AF4-MLL est présente chez les leucémies infantiles dans 75% des cas, et est associée à un mauvais pronostic (10). Le tableau ci-dessous répertorie les modifications génétiques les plus fréquentes, ainsi que leurs pronostics associés.

Tableau I. – Modifications cytogénétiques fréquentes dans les LLA pédiatriques

Sous types	Fréquence pédiatrique	Gènes impliqués	Conséquences cliniques
Leucémie de type B :			
Haute hyperdiploïdie (> 50 chromosomes)	15 - 25%	NRAS, KRRAS, FLT3, PAX5, IKZF3, PAG1	Pronostic favorable
ETV6-RUNX1	15 - 25%	ETV, RUNX1, WHSC1, KRAS, NRAS	Pronostic favorable
KMT2A	3 - 4%	KMT2A, AFF1, MLLT1, MLLT3, MLLT10, EPS15…	Mauvais pronostic
Ph	2 - 5%	BCR, ABL1, RUNX1, PAX5, IKZF1, CDKN2	Mauvais pronostic, amélioré avec TK1
Ph-like	6 - 15%	CRLF2, ABL1/ABL2, JAK2, EPOR, JAK2…	Mauvais pronostic, modifiable avec TK1
Haute hypodiploïdie (40 - 43 chromosomes)	1 - 2%		Mauvais pronostic
DUX4	4 - 7%	ERG, IGH CD2, CD371^{+}	Favorable avec délétion ERG, intermédiaire sans délétion ERG
Leucémie de type T :			
Mutation NOTCH1 (9q34.3)	60 - 70%		Pronostic favorable dans l'ensemble
Translocations du TCR avec plusieurs oncogènes t(1;14), t(10;14), t(5;14)	~ 35%	LMO1, LMO2, TAL1, TLX1, TLX3	Pas d'impact défini
Del(1)(p32)	~ 10%	SIL-TAL1	Pas clairement établi
Délétion 9p	20 - 30%	CDKN2A, CDKN2B	Pas d'impact défini
Réarrangements 11q23	6%	MLL avec plusieurs gènes	Mauvais pronostic
JAK1 (1p32.3-p31.3)	2%		Mauvais pronostic

Tableau adapté de Li et al., 2021 et de Chiaretti et al., 2014 (13, 15).

2.2 Symptômes et traitements

D'un point de vue clinique, les patients qui présentent une leucémie ont plusieurs symptômes distinctifs et récurrents. Habituellement, 75% des cas de LLA sont diagnostiqués avant l'âge de 6 ans, ce qui semble correspondre au moment où les cellules immatures pré-B sont les plus nombreuses dans la moelle (16, 17). Pour les leucémies de type T, celles-ci apparaissent plutôt à l'adolescence lorsque le thymus a atteint sa taille maximale (16, 17). De manière générale, plus les patients sont âgés et plus ils sont

considérés à haut risque (17). Le critère de risque clinique (donc l'agressivité de la leucémie) dépend d'un certain nombre de facteurs (voir tableau II), notamment de l'âge, du sexe et du taux de globules blancs (17). Le traitement des enfants atteints de LLA sera ajusté en fonction de cette classification liée au risque de récidive (18, 19). Le traitement sera donc plus agressif chez les patients ayant une plus grand risque de rechute (18, 19). Ce traitement se divise en 3 grandes étapes s'étalant sur plus de deux ans, suivi d'une période de suivi de 3 ans pour un total de 5 ans (19).

Tableau II. – Facteurs de risques dans les LLA pédiatriques

Variables	Facteurs favorables	Facteurs non favorables
Facteurs cliniques et démographiques		
Age	De 1 à < 10 ans	< 1 an ou ≥ 10 ans
Sexe	Féminin	Masculin
Race ou groupe ethnique	Blanc, Asiatique	Noir, Amérindien, Hispanique
Nombre initial de globules blancs	Bas (< 50 000/mm^3)	Haut (≥ 50 000/mm^3)
Facteurs biologiques ou génétiques des cellules leucémiques		
Immunophénotype	Lignée des lymphocytes B	Lignée des lymphocytes T
Facteurs cytogénétiques	ETV6-RUNX1, hyperdiploïdie, trisomies favorables	BCR-ABL1, réarrangements MLL, hypodiploïdie
Facteurs génétiques	Délétions ERG	Délétions ou mutations de IKZF1, LLA Ph-like avec des altérations de gène kinase
Réponses au traitement		
Réponse après 1 semaine de traitement de glucocorticoïdes	Bonne réponse à la prednisone (<1000 blastes /mm^3)	Mauvaise réponse à la prednisone (≥ 1000 blastes /mm^3)
Blastes présents dans la moelle après 1-2 semaines de thérapies	Moelle M1 (< 5% de blastes) au jour 8 ou 15	Pas de moelle M1 (≥ 5% de blastes) au jour 8 ou 15
MRM pendant ou à la fin de l'induction	MRM faible (<0,01%) ou indétectable à des moments précis	MRM persistante ≥ 0,01% à des moments précis, (plus c'est haut, pire est le pronostic)
MRM à 3-4 mois	Bas (<0,01%), de préférence indétectable	MRM persistante ≥ 0,01%

MRM : maladie résiduelle minimale; Ph-like : chromosome de Philadelphie. Tableau adapté de Hunger et Mullighan, 2015 (17).

La première étape consiste à l'éradication des blastes leucémiques par induction de glucocorticoïdes afin de restaurer l'hématopoïèse (18, 19). Ensuite vient une période de 6 à 8 mois appelée la consolidation, durant laquelle de fortes doses de chimiothérapie sont administrées pour éradiquer les cellules leucémiques restantes (18, 19). Enfin, la plus longue et troisième étape consiste en un traitement à base d'immunosuppresseurs de plus faible intensité (18, 19).

Au cours des dernières décennies, l'optimisation des traitements, entre autres, grâce à une amélioration de la stratification des groupes à risque, a permis de faire passer le taux de survie de 10% à plus de 90% (20). Il reste tout de même 20% des patients qui feront une rechute et, de ceux-ci, la moitié décédera de leur maladie (20). Une meilleure stratification des patients pourrait mener à une amélioration du taux de survie. Par ailleurs, un nombre considérable de survivants est aussi confronté à des effets secondaires importants dus à la toxicité des traitements tels que des troubles métaboliques, des déficits neurocognitifs, des risques augmentés de morbidité ou encore une mort prématurée causé par des maladies cardiovasculaires (16, 17).

En résumé, même si les différents sous types de cette leucémie possèdent des similarités morphologiques et moléculaires, ils renferment également diverses distinctions, tant au niveau clinique, qu'au niveau moléculaire où des altérations génétiques s'y produisent.

Une stratification des patients basée sur les différents sous types semble alors être une stratégie attirante pour adapter plus précisément les traitements. Un autre enjeu important est le caractère polyclonal de la leucémie au diagnostic (16, 17). Il est alors difficile de prédire si un clone mineur, présent au diagnostic, résistera au traitement ou bien si cette résistance est due à l'apparition d'une nouvelle mutation (16, 17).

2.3 Organisation du système immunitaire

Un autre aspect important de la réponse aux traitements est la réponse immunitaire.

Lorsque l'organisme est exposé à l'apparition de cellules tumorales, des mécanismes anti-tumoraux sont enclenchés par le système immunitaire (21). Des cellules développées à partir de cellules souches hématopoïétiques dans la moelle sont alors différenciées pour remplir les différents rôles nécessaires au bon fonctionnement du système immunitaire (voir Figure 3) (21).

Tel que mentionné plus haut, un des types cellulaires majeur dans la LLA sont les lymphocytes, puisque ce sont généralement ces cellules impliquées dans la lignée lymphoïde qui sont anormales.

Lymphocytes B

Les lymphocytes B en font partie et sont le résultat de la différenciation de plusieurs types cellulaires. En effet, le processus de différenciation commence au niveau des cellules souches hématopoïétiques comme les autres cellules, puis passe au stade de cellules progénitrices multipotentes (MPP), pour ensuite donner lieu à un type de cellules capable de se différencier en cellules lymphoïdes ou myéloïdes (LMPP) (22). La lignée des lymphocytes est ensuite amorcée par un type plus spécifique de progéniteurs, soit les progéniteurs lymphocytaires (CLP) (22). Ces derniers se différencient une dernière fois en lymphocytes, dont les lymphocytes B.

Une fois les lymphocytes B activés, certains se divisent en cellules mémoires, et d'autres en plasmocytes : des cellules spécialisées dans la production d'anticorps (21). Elles sécrètent donc des anticorps qui vont se fixer à des antigènes spécifiques pour protéger l'organisme (22). Leur fixation permet ainsi d'agir à titre de marqueurs pour les autres cellules du système qui vont pouvoir être alertés, et ainsi remplir leur rôle (22). Certains de ces lymphocytes B ont aussi la capacité de se lier directement aux pathogènes, ou de se lier à d'autres phagocytes pour augmenter l'efficacité de la réponse (22).

Lymphocytes T

De leur côté, les lymphocytes T vont migrer vers le thymus pour leur maturation (21). Le thymus est essentiel dans le développement de ces lymphocytes, et ces derniers quitteront ce thymus pour migrer vers d'autres organes une fois qu'ils seront matures (21). Ces lymphocytes matures auront alors la capacité de détruire les agents pathogènes associés à un danger et d'alerter les autres types cellulaires (21). Ces lymphocytes T sont divisés en groupes distincts, ceux qui sont le plus retrouvés sont les lymphocytes T cytotoxiques (CD8), auxiliaires (CD4) et régulateurs (Treg) (21).

Les cellules T auxiliaires sont des intermédiaires servant à coordonner la réponse immunitaire. En effet, ces dernières servent généralement à établir une communication avec les autres cellules, notamment en attirant plus de cellules capables de détruire les pathogènes (21). D'autres vont stimuler les lymphocytes B pour qu'ils produisent plus d'anticorps (21).

Les lymphocytes T régulateurs quant à eux aident à maintenir l'homéostasie cellulaire (21). Ces derniers vont réprimer l'activité des cellules immunitaires qui ont fini leurs actions dans la réponse immunitaire, ou bien les cellules qui sont auto-immunes, c'est à dire celles qui attaquent les cellules saines de l'organisme (21).

Les lymphocytes T cytotoxiques sont les cellules anti-tumorales les plus importantes (21, 22). Ces lymphocytes vont migrer vers le site porteur d'antigènes spécifiques et vont se rattacher à sa cible pour la détruire (21).

Par ailleurs, lorsqu'assez d'immunogènes sont produits lors de l'initiation tumorale, les lymphocytes T n'ayant pas encore d'antigènes spécifiques (naïfs) ont la capacité de migrer vers le microenvironnement tumoral et construire une réponse immunitaire protectrice, éliminant ainsi des cellules cancéreuses (22). Ceci explique alors pourquoi les interactions avec les lymphocytes T sont de plus en plus étudiées dans divers types de cancer. Il a été montré qu'un haut taux de cellules T infiltrant les tumeurs était plutôt relié à des pronostics favorables dans plusieurs cancers (22).

Lymphocytes NK

Les cellules « Natural Killer », sont généralement efficaces contre les virus (21). Ces cellules sont dérivées de cellules de la moelle osseuse et sont présentes en faible nombre dans le sang et dans les tissus (21). Par ailleurs, elles sont nommées « tueuses naturelles » du fait qu'elles n'ont pas besoin d'une maturation et d'un développement aussi avancé que les lymphocytes T pour agir (21). La surface de ces cellules est composée de différents types de récepteurs, tant bien stimulateurs qu'inhibiteurs, ayant des rôles dans la surveillance immunitaire (22).

Phagocytes

Les phagocytes sont une catégorie moins présente dans la lignée lymphoïde et donc dans la LLA. Ces cellules sont constamment à la recherche de pathogènes, se multiplient lorsqu'une cible est détectée et envoient également des signaux aux autres types cellulaires pour qu'ils fassent de même (23).

Monocytes

Les monocytes circulent dans le sang et représentent environ 5 à 10% des globules blancs (21). Ils ont plusieurs rôles distincts associés à la protection, et sont par exemple impliqués dans la réparation des tissus, dans la présentation des antigènes aux lymphocytes, ou encore dans la phagocytose (23). Ils sont également retrouvés dans les parois des vaisseaux sanguins de certains organes, notamment la rate ou le foie, pour détruire les bactéries et microorganismes circulants (23). Lorsque ces monocytes s'infiltrent dans un tissu, ils changent de conformation et de taille pour ainsi se développer en macrophage (21).

Macrophages

Les macrophages quant à eux, sont un type de globules blancs dérivés des monocytes et importants dans le système immunitaire (21, 22). Ils sont responsables de l'élimination de débris cellulaires, mais également de la réparation des tissus ou encore du maintien de l'homéostasie (22, 23). Les macrophages sont alors impliqués dans la reconnaissance et la réponse aux infections et aux lésions. (22).

De plus, ils sont retrouvés tout au long de la progression des cancers, depuis la transformation des premières métastases jusqu'à leur résistance aux thérapies (22). Ces macrophages ingèrent alors les corps étrangers et délivrent des éléments toxiques à ces derniers pour les éliminer (22).

De manière générale, les macrophages se subdivisent en plusieurs sous types en fonction des signaux du microenvironnement (22). Le plus commun est le type M1, un type plutôt pro-inflammatoire, impliqué dans la présentation d'antigènes et la sécrétion de cytokines et chimiokines pour tuer les microorganismes pathogènes (22). Le type alternatif M2 quant à lui est plutôt engagé dans des processus tels que la réduction des inflammations et la réparation des lésions tissulaires (22).

Lors de l'installation d'une tumeur, plusieurs cellules et facteurs biochimiques vont être sécrétées et les monocytes circulant dans le sang, recrutés par le microenvironnement vont se différencier en macrophages associés aux tumeurs (TAM) (22). Ces derniers sont plutôt prédicteurs d'un mauvais pronostic et d'une réduction de la survie (22). Lorsque ces macrophages ne possèdent plus leur fonction phagocytaire, ils vont alors réguler la progression la tumeur (22).

Neutrophiles

Les neutrophiles sont également des cellules développées à partir des souches hématopoïétiques de la moelle (21). Ils représentent le plus grand nombre parmi les globules blancs dans le sang (21). Ces derniers sont des éléments clés lors d'inflammation puisqu'ils ont tendance à se déplacer et s'accumuler rapidement pour attaquer les microorganismes en les ingérant (21).

Cependant, dans différents types de cancers, les hauts taux de neutrophiles associés aux tumeurs ont été associé à des pronostics plutôt défavorables (22).

Autres types de cellules

D'autres types cellulaires moins présents jouent également des rôles similaires dans le système immunitaire.

Parmi ceux-là, on retrouve les cellules dendritiques qui jouent un rôle dans le déclenchement de la réponse immunitaire (22, 23). Ces dernières sont capables de reconnaître les antigènes et activer les lymphocytes T caractéristiques des pathogènes pour enclencher la réponse (22).

D'autres cellules comme les mastocytes interviennent également en ligne de défense face aux antigènes, notamment en recrutant d'autres types cellulaires, ou bien en se liant à un type d'anticorps libéré par les lymphocytes B (21). L'activation de ces mastocytes libère des granules qui renferment des médiateurs chimiques pour éliminer le pathogène (23).

Le tableau III présente les principaux marqueurs de surfaces utilisées pour distinguer ces différents types cellulaires.

Tableau III. – Principaux marqueurs des cellules immunitaires chez l'humain

Type cellulaire	Marqueurs de surface
Cellule souche hématopoïétique	$CD45^+$, $CD34^+$
Cellule lymphoïde progénitrice	CD45
Cellule myéloïde progénitrice	$CD11b^+$
Lymphocyte B	$CD19^+$
Lymphocyte T auxiliaire	$CD3^+$, $CD4^+$
Lymphocyte T cytotoxique	$CD3^+$, $CD8^+$
Lymphocyte T régulateur	$CD25^+$
Lymphocyte NK	$CD56^+$, CD3
Macrophage	$CD68^+$, $CD11b^+$
Monocyte	$CD14^+$
Neutrophile	$CD16^+$, $CD11b^+$
Cellule dendritique	$CD11c^+$, HLA-DR^+

+ : marqueur exprimé par le type cellulaire. Tableau adapté de *Immune cell markers poster* (24).

3. Le microenvironnement tumoral

Les composants du système immunitaire interagissent intimement avec les tumeurs tout au long du développement de la maladie (22, 25). Ces interactions complexes peuvent à la fois inhiber et favoriser la progression tumorale (22, 25).

De ce fait, en plus des cellules impliquées dans la tumeur même, il faut aussi considérer l'implication des molécules autour de la tumeur et de l'éventuelle infiltration d'autres cellules non cancéreuses (26).

La compréhension de ce microenvironnement tumoral aide à mieux comprendre la progression de la tumeur et prédire les réponses aux traitements (25, 27). En effet, la composition de cet environnement, ainsi que les interactions qui s'y produisent, ont permis de distinguer des groupes de patients pouvant avoir une meilleure réponse aux traitements, et ce, en fonction de la présence de biomarqueurs caractéristiques, ou d'autres facteurs comme l'infiltration immunitaire (25, 28). Ce phénomène est généralement étudié dans des tumeurs dites solides du fait que les tumeurs liquides, comme les leucémies, ne possèdent pas de masses et qu'une infiltration à proprement parlé n'est donc pas forcément observable et quantifiable (29).

Par ailleurs, les cellules tumorales développent des mécanismes pour résister à l'action du système immunitaire (25, 30). Ces cellules cancéreuses pourront alors échapper au contrôle immunitaire (25, 31). L'activation de ces points de contrôle spécifiques permet à la tumeur d'échapper à l'immunité (22).

Puisque les interactions entre les cellules tumorales et le système immunitaire sont gouvernées par un réseau d'interactions complexes entre les cellules, une connaissance approfondie de la composition des cellules immunitaires dans une tumeur peut s'avérer essentielle pour prédire la réponse d'un patient à une immunothérapie (25). Ainsi, la quantification de l'infiltration immunitaire a un potentiel informatif sur les mécanismes de l'évasion immunitaire et sur les potentiels marqueurs associés pour restimuler le système immunitaire (31, 32).

Chaque sous type de ces cellules immunitaires semble avoir un impact plus ou moins différent sur le développement de la tumeur (28). Par exemple, les cellules T CD8 sont

des cellules primordiales dans l'immunité anticancéreuse puisqu'elles sont capables de reconnaître et détruire des cellules tumorales porteuses d'antigènes spécifiques formés à partir de gènes mutés (nommés néoantigènes) (21, 26, 33). D'autres sous types cellulaires quant à eux, comme les cellules T régulatrices (Treg), vont plutôt exercer des fonctions immunosuppressives qui vont favoriser l'évasion immunitaire et supporter la tumorigenèse (25). En plus des cellules tumorales et des cellules immunitaires, le microenvironnement est aussi composé de cellules stromales, pouvant elles aussi favoriser la croissance tumorale (32).

La quantification de ces différents sous types cellulaires peut aider à l'identification des mécanismes et des effets thérapeutiques, autant en identifiant ceux responsables des réponses anticancéreuses qu'en identifiant ceux responsables du développement tumoral (28, 32).

Il est donc important d'analyser le rôle potentiel du microenvironnement, plus précisément le rôle des cellules immunitaires pouvant s'infiltrer dans les tumeurs. Dans mon projet, notre attention s'est portée sur 3 types de tumeurs solides : les Neuroblastomes, les Osteosarcomes et les tumeurs de Wilms (34).

3.1 Tumeurs solides

Neuroblastomes

Les neuroblastomes sont des cancers solides du système nerveux périphérique (35). Ce type de cancer représente 6% des cancers pédiatriques en Amérique du Nord, et se développe généralement avant l'âge 10 ans, avec la majorité des cas diagnostiqués chez des enfants de moins de 5 ans (36).

Ostéosarcomes

Les ostéosarcomes font partie des cancers des os les plus communs (37). Ils représentent 2% des cancers chez les enfants, et 3% des cancers chez les adolescents (36). Les ostéosarcomes sont diagnostiqués entre l'âge de 10 et 30 ans, avec les tumeurs primitives apparaissant majoritairement vers l'adolescence (36).

Tumeurs de Wilms

La tumeur de Wilms est le type de cancer du rein le plus courant (38). Elle représente environ 5% de tous les cancers en pédiatrie (36, 38). Ces tumeurs sont principalement retrouvées chez les enfants de 2 à 4 ans (38).

Le microenvironnement de ces 3 types de de cancer reste encore peu étudié (34, 39).

L'analyse des altérations génétiques dans les tumeurs solides et leur impact sur les voies biologiques peut servir à stratifier les patients en sous-groupes avec des pronostics spécifiques. Toutefois, comme expliqué auparavant, les cancers solides sont aussi définis par les différentes cellules immunitaires recrutées dans la tumeur lors de son établissement, ce qui laisse alors place à des approches plus novatrices.

Il est maintenant connu que l'hétérogénéité tumorale peut agir comme un facteur biologique ayant des implications dans plusieurs processus tels que le développement du cancer ou encore sa progression (26, 40). La plupart des tumeurs solides montre une forme d'hyper-mutabilité menant à une augmentation de la variabilité génomique au fur et à mesure de l'expansion des populations de cellules tumorales (26) .

Ainsi, pour prédire efficacement la réponse d'un patient à une thérapie, il faut d'abord réussir à comprendre comment la tumeur échappe l'immunosurveillance mise en place, mais aussi comprendre comment générer une réponse anti-tumorale efficace avec des effecteurs immunitaires et des antigènes fonctionnels dans des quantités adéquates (34, 40). Il existe diverses approches utilisant des outils de nouvelle génération pour déterminer le profil immunitaire d'une tumeur.

4. Les outils de nouvelle génération

Au cours des années, les analyses avec micropuces à ADN ont été remplacées par le séquençage de nouvelle génération (41). Le séquençage de nouvelle génération permet d'étudier l'entièreté du génome et son réseau de régulation de façon plus rapide et non biaisée (42). Ce dernier devient de plus en plus accessible, notamment de par la baisse

des coûts reliés à son utilisation (42). Nous pouvons ainsi effectuer une analyse simultanée de l'expression de l'ensemble des gènes (transcriptome), appelée ARN-seq (42-44).

Les données d'ARN-seq peuvent être employées pour plusieurs types d'analyses (45). 1) Identification des gènes différentiellement exprimés (45). Ce type d'analyse sert notamment à obtenir une meilleure compréhension des mécanismes impliqués dans les processus biologiques (45). 2) Identification d'évènements génétiques comme l'épissage alternatif, les mutations, les nouveaux transcrits, les ARN non codants ou encore les transcrits de fusion (45).

L'ARN-seq permet également de découvrir de nouveaux transcrits et isoformes, par opposition aux micropuces qui ne contiennent que des séquences d'intérêt à priori connues (41).

D'autres technologies d'ARN-seq, dites unicellulaires (scARN-seq), permettent la description des molécules ARN dans des cellules individuelles avec une plus haute résolution à l'échelle du génome (46, 47). Cependant, les données de cette méthode scARN-seq sont hautement variables et contiennent plus de bruit de fond que les données ARN-seq (46). De plus, en fonction de la complexité des échantillons à analyser, il faut fournir un plus grand nombre de cellules, et donc augmenter la profondeur de la couverture de reads, pour pallier le manque de détection de transcrits spécifiques (46, 47).

4.1 Applications dans les leucémies

Les cellules leucémiques ont été, et sont toujours, étudiées par le biais de plusieurs méthodes, allant de l'étude du caryotype, qui identifie les modifications chromosomiques, jusqu'au séquençage du génome entier, qui lui, identifie le moindre changement dans les séquences (18).

Un des défis majeurs de la génomique des cancers vient de l'hétérogénéité cellulaire observée dans une tumeur, qui est caractérisée par la coexistence de plusieurs populations cellulaires et de clones tumoraux (41).

L'utilisation de séquençage de nouvelle génération, notamment l'ARN-seq, permet d'identifier et découvrir des sous types de leucémies pour avancer un pronostic, mais permet également d'améliorer notre capacité à évaluer le risque qu'une thérapie échoue (18, 41).

En effet, lors de l'apparition et du développement de la maladie, l'expression des gènes aide à comprendre comment les processus de différenciation et les fonctions cellulaires ont été perturbés puisque ces profils de gènes sont hautement régulés et spécifiques dépendamment des tissus, des conditions physiologiques ou pathologiques ou encore des stades de développement (48). En réponse à ces modifications, la cellule tumorale va remanier la formation de plusieurs protéines et ainsi engendrer une différence dans l'expression des gènes par rapport aux cellules saines (48). Cette différence va alors être observable et quantifiable, notamment à l'aide d'outils (31, 48).

Ce type de méthode, réuni avec d'autres approches de profilage moléculaire, a permis au cours des dernières années d'identifier des marqueurs leucémiques uniques qui n'étaient pas détectables avec les méthodes standards, soulignant la nécessité des approches multimodales (9, 11, 19, 42).

4.1.1 Pureté tumorale

Du fait de la présence de cellules normales dans les tumeurs, les échantillons tumoraux ne peuvent pas être considérés comme parfaitement purs (49). La pureté tumorale fait donc référence à la proportion des cellules cancéreuses présente dans un tissu (29, 49).

Pendant de nombreuses années, cette pureté était estimée en clinique visuellement ou par l'analyse d'images (49). L'accès aux technologies de séquençage génomique permet aujourd'hui d'appliquer des méthodes computationnelles, basées sur l'utilisation des informations moléculaires, pour déduire la pureté tumorale et effectuer des classifications de sous-types plus précises (49).

Ces méthodes de profilage moléculaire ont été appliquées dans plusieurs types de cancers, notamment dans le cadre de programmes de médecine de précision (41).

4.1.2 Déconvolution

La méthode standard pour quantifier les types cellulaires est la cytométrie de flux (50). Cette méthode connaît cependant des limites significatives et peut rapidement devenir complexe. La cytométrie de flux peut par exemple devenir inutile, et ses résultats faussés, si certains anticorps spécifiques aux données ne sont pas incorporés dans l'analyse (51). Par ailleurs, cette méthode nécessite des échantillons récents qui n'ont pas eu le temps d'être détériorés par l'apoptose naturelle des cellules (51). Enfin, des limites importantes au niveau des coûts et de la charge de travail ont généré un intérêt capital pour l'analyse du transcriptome (52).

Ainsi, l'analyse des profils d'expression de marqueurs de surfaces de différents types cellulaires (voir tableau III ci-dessus) par les algorithmes de déconvolution permet de prédire les fractions cellulaires présentes dans un échantillon (31, 48). Les méthodes de déconvolution assument que l'expression des gènes spécifiques à chaque type cellulaire est en fait proportionnelle à sa proportion dans le mélange, en supposant que chaque sous type a des niveaux d'expressions similaires à travers les différents échantillons (50, 52). Ces méthodes avec des matrices de référence utilisent un maximum de gènes spécifiques aux différents sous types cellulaires pour qu'un minimum de gènes signatures soit représentatif d'un type en particulier, et puisse ainsi garantir une certaine distinction et spécificité de la méthode (50, 52).

Dans un échantillon hétérogène, si la matrice d'expression signature pour chaque type cellulaire est connue, le problème de déconvolution revient à la formulation d'un système d'équations linéaires indépendantes, décrivant l'expression de chaque gène du mélange, où chaque gène est une combinaison linéaire des différents sous types cellulaires de l'échantillon (50, 52). L'hypothèse de linéarité d'expression entre les données pures et hétérogènes a déjà été validée comme étant raisonnable dans de précédentes études, où les gènes d'un même type cellulaire sont généralement hautement corrélés (52).

Plusieurs outils de déconvolution ont été développés au fils des années, particulièrement pour avoir une meilleure connaissance de la relation entre la composition cellulaire et le degré d'une maladie (48, 53). Ces outils ont été développés avec des méthodes de calcul

différentes, avec ou sans référence, et avec des matrices signatures (lorsque présentes) composées des types cellulaires d'intérêt (31, 52).

Il existe différents types d'approches pour quantifier les types cellulaires avec les outils (voir tableau ci-dessous) (52). Il existe des approches basées sur les gènes marqueurs, sur des déconvolutions partielles (nécessitant une matrice de référence), et des approches de déconvolution complète (52).

Dans notre cas, une revue de la littérature de ces outils a été faite pour déterminer lequel serait le mieux adapté à nos données. Notre choix s'est arrêté sur 3 outils en particulier : Cibersort, DeconRNAseq et Quantiseq.

Tableau IV. – Caractéristiques d'outils de quantification cellulaire

Outils	Types	Méthodes	Types cellulaires	Références
TIminer	M	Enrichissement de gènes pré-classés	Différents ensembles avec 28, 31 et 64 types cellulaires	(54)
xCell	M	Enrichissement de gènes ajustés	64 types immunitaires et non immunitaires	(55)
MCP-counter	M	Moyenne géométrique de l'expression de gènes marqueurs	8 types immunitaires, fibroblastes et cellules endothéliales	(56)
DeconRNAseq	P	Régression des moindres carrés sous contraintes	-	(57)
PERT	P	Maximum de vraisemblance non négative	-	(58)
CIBERSORT	P	Régression avec vecteur de support Nu	22 types immunitaires	(59)
TIMER	P	Régression linéaire des moindres carrés	6 types immunitaires	(60)
EPIC	P	Régression des moindres carrés sous contraintes	6 types immunitaires, fibroblastes, cellules endothéliales et non caractérisés	(61)
Quantiseq	P	Régression des moindres carrés sous contraintes	10 types immunitaires, et cellules non caractérisées	(62)
deconf	C	Matrice de factorisation non négative	-	(63)
DSA	C	Programmation quadratique	-	(64)
ssKL	C	Matrice de factorisation non négative	-	(65)

M : méthode gènes marqueurs, P : déconvolution partielle, C : déconvolution complète. Tableau adapté de Finotello et Trajanoski, 2018 (52).

Les approches d'enrichissement de gènes tels que xCell ou MCP-counter nécessitent des marqueurs hautement spécifiques où les gènes sont exprimés exclusivement pour le type cellulaire en question, ce qui n'est pas applicable dans notre cas puisque le contenu des échantillons n'est pas nécessairement connu (43, 52). De plus, comme chaque gène est pris en compte indépendamment des autres signatures, il est plus difficile de faire la différence entre des types plus ou moins liés (52).

Additionnellement, ces deux outils donnent les résultats sous forme de scores, et ne permettent donc pas de comparer les proportions des types cellulaires dans les échantillons (31, 52). Par ailleurs, xCell connaît plusieurs limites reliées aux mélanges (52). Des études ont rapporté que xCell ne détectait aucun signal lorsque des échantillons non hétérogènes étaient présents dans les données (31).

Pour ce qui est des approches de déconvolution complète, ces dernières n'utilisent pas de référence pour estimer les proportions (52). Cependant, avec leur méthode de calcul, il n'est pas garanti de trouver les caractéristiques spécifiques aux types cellulaires nécessaires pour estimer les fractions cellulaires dans nos données (31).

Ainsi, notre choix s'est plutôt porté vers les approches de déconvolution partielle. Ces approches utilisent une matrice signature composée de l'expression des gènes de divers types cellulaires pour pouvoir les quantifier (52).

Parmi les outils disponibles, TIMER est un outil qui permet des déconvolutions de six types cellulaires (52, 66). Cependant les résultats de cet outil ne peuvent pas être comparés entre les différents types, et ne peuvent pas être interprétés directement comme des fractions cellulaires mais plutôt comme des scores (52). De plus, cette méthode, comme celle de xCell, analyse tous les échantillons de manière dépendante, et un même échantillon peut alors présenter des résultats différents lorsqu'il est soumis avec d'autres échantillons (31). EPIC utilise les données ARN-seq pour estimer les fractions relatives à tout le mélange (66). L'utilisation de cet outil est limitée car il ne considère malheureusement pas certains types cellulaires jouant un rôle important dans les cancers, tels que les cellules dendritiques, les cellules T régulatrices ainsi que les macrophages (66).

DeconRNAseq, un des outils choisis dans mon projet, permet de fournir sa propre matrice signature, et est donc plus facilement adaptable aux données disponibles (57). Quantiseq, un autre des outils choisis, est conçu pour les données ARN-seq, et estime les proportions cellulaires de 10 types différents en se référant au contenu total des échantillons, permettant ainsi une comparaison intra- et inter-échantillon (52, 62, 66). De plus, des hautes corrélations ont été observées entre les résultats de cet outil et ceux obtenus par les cytométries de flux (62). Cibersort, un des outils le plus cité dans la littérature, et le 3^{e} outil choisi dans mon projet, répertorie 22 types cellulaires différents et utilise une méthode qui semble être robuste vis à vis des potentielles populations cellulaires inconnues (52, 59, 67).

Les résultats sous forme de calculs de proportion sont plus faciles à interpréter que les scores. Les valeurs de proportion étant positives et entre 0 et 100 sont directement interprétables, tandis que les scores n'ont pas toujours les mêmes intervalles et sont parfois négatifs, sans unité (31). De plus, avec les proportions, il est généralement possible de faire des comparaisons aussi bien entre les sous types cellulaires et entre les échantillons, tandis que le score permet seulement les comparaisons entre échantillons (31).

4.2 Détermination du profil immunitaire

Certaines approches expérimentales standard existent pour quantifier ces cellules infiltratrices, mais ces méthodes ont souvent des limites au niveau du nombre de marqueurs, du débit ou du nombre d'échantillons (31). Maintes études ont prouvé que la localisation ainsi que l'abondance de ces cellules immunitaires sont des valeurs pronostiques intéressantes (40). Jusque récemment, les méthodes utilisées pour étudier les infiltrats immunitaires étaient principalement l'immunohistochimie, l'immunofluorescence et la cytométrie, mais ces dernières souffrent de certaines limitations considérables (52, 62). L'immunohistochimie et l'immunofluorescence par exemple, établissent des profils sur des lames de tissus tumoraux, limitant alors le nombre de biomarqueurs pouvant être accessibles simultanément, et ne sont donc pas forcément des analyses représentatives de l'ensemble de la masse tumorale et de son environnement (32). De plus, l'immunohistochimie a souvent du mal à capturer les

phénotypes fonctionnels, tandis que la cytométrie de flux elle, fait face à des manipulations de tissus laborieuses du fait de la possible dégradation de ces derniers, résultant parfois en la perte de certains types cellulaires ou en l'altération des profils d'expression (67).

La manière d'interpréter les sous types immunitaires peut également varier en fonction des institutions et des pathologistes, créant ainsi le manque d'une réelle méthode pour définir ces facteurs (27).

Avec l'essor des technologies de nouvelle génération, on dénote de plus en plus d'études en oncologie qui utilisent les données d'ARN-seq pour décrire le microenvironnement tumoral (39).

Des méthodes utilisant des gènes marqueurs provenant généralement d'une analyse préalable d'enrichissement des gènes (GSEA), et basée sur un score d'enrichissement de spécificité des gènes envers un certain type cellulaire, sont alors parfois utilisées (39, 52). Cependant, ces approches basées sur l'enrichissement des gènes calculent au final un score semi-quantitatif qui est élevé lorsque l'enrichissement d'un type cellulaire est notable dans un échantillon (52).

Les méthodes de déconvolution quant à elles, sont capables d'estimer quantitativement les fractions relatives des différents types cellulaires d'intérêt (39). L'analyse de la composition du microenvironnement semble être réalisable en étudiant le transcriptome. En effet, les algorithmes de déconvolution estiment les proportions cellulaires en utilisant des matrices signatures représentatives des profils d'expression spécifiques pour chaque type de cellule (39, 50). De plus, comme énoncé précédemment, ces proportions cellulaires sont parfois à privilégier par rapport aux scores, d'autant plus si les gènes marqueurs spécifiques aux sous-types cellulaires ne sont pas connus (48). En effet, le calcul du score quantitatif nécessite la connaissance de ces gènes marqueurs spécifiques à nos données, qui ne sont pas toujours connus (48).

En examinant des gènes exprimés par des types de cellules immunitaires spécifiques, il est alors possible pour une méthode computationnelle de mesurer l'abondance de ces cellules dans un échantillon donné, en faisant en sorte que l'outil reflète le plus possible

la composition de ce type de tumeur (39). En théorie, la déduction de cette composition immune, stromale et tumorale, est possible à partir des profils d'expression des gènes si un profil de référence est possiblement établi pour chaque type cellulaire associé à la tumeur analysée (29).

Comme mentionné précédemment, plusieurs études ont déjà montré des corrélations entre les données cliniques et les infiltrations immunitaires (26, 28, 68). Ces méthodes computationnelles sont alors un moyen de construire un score pour fournir des informations sur la composition du microenvironnement à partir des données ARN-seq de la tumeur, dans le but éventuel d'étudier leur impact clinique (68).

5. Problématiques et hypothèses

Malgré le succès thérapeutique des dernières décennies, plus de 20% des enfants et adolescents atteints d'un cancer succomberont à cette terrible maladie (10). Ainsi, le cancer demeure la principale cause de décès par la maladie chez les moins de 18 ans (10). On note également chez plus de 75% des survivants d'un cancer pédiatrique, l'apparition à l'âge adulte de nombreuses complications médicales liés aux traitements (10). Il est donc important de raffiner la manière dont les patients sont stratifiés afin de recevoir une régime thérapeutique optimal. Dans le cadre de mon projet de maitrise, j'ai appliqué divers outils bio-informatiques pour analyser des données de transcriptome afin de répondre à certains de ces défis dans la leucémie lymphoblastique aigue, et dans certaines tumeurs solides (neuroblastomes, ostéosarcomes et tumeurs de Wilms).

Leucémie lymphoblastique aigue (LLA)

La LLA est une maladie hétérogène pouvant être composée d'une multitude de sous clones et de types cellulaires(12).

Il est important d'améliorer notre compréhension de cette hétérogénéité, ainsi que sa dynamique suite aux traitements afin d'offrir des thérapies cliniques plus efficaces, et si possible moins toxiques.

Une des limitations est l'analyse de la pureté tumorale.

Hypothèse : L'utilisation de certains outils bio-informatiques permet de quantifier la pureté tumorale de manière plus rapide et précise, et d'identifier les différents types cellulaires nécessaires pour classifier les sous types de leucémies.

Dans mon projet, j'ai utilisé des données de transcriptome de patients LLA afin de valider cette hypothèse.

Tumeurs solides

Les traitements utilisés jusqu'à présent connaissent beaucoup de cas de résistance, d'évasion immunitaire, de rechutes ou d'effets secondaires sur le long terme. L'immunothérapie, impliquant le système immunitaire du patient, ou un nouveau système allogénique, offre une alternative aux patients qui ne répondent pas à la chimiothérapie (34).

Un défi est de caractériser le répertoire de cellules immunitaires afin d'évaluer l'impact du microenvironnement tumoral dans la réponse aux traitements.

Hypothèse : Le taux d'infiltration de certaines de ces cellules immunitaires favoriserait ou empêcherait l'avancée de ces cancers.

Dans ce volet de mon projet, je me suis concentré sur 3 tumeurs solides : neuroblastome, ostéosarcome et tumeur de Wilms.

6. Objectifs du projet

Mon projet se base sur 3 méthodes de déconvolution appliqués sur des jeux de données synthétiques, leucémiques et de tumeurs solides. Ce projet est divisé en deux grandes parties. La première partie relate la caractérisation d'un cancer hématopoïétique, soit la leucémie lymphoblastique aiguë, et ce, à l'aide d'outils de déconvolution. La deuxième partie quant à elle, traite de tumeurs solides, et se concentre sur l'interaction entre la tumeur et l'infiltration de cellules immunitaires du microenvironnement.

Les objectifs de mon projet sont les suivants :

1. Estimer l'efficacité des outils de déconvolution sur des échantillons synthétiques;

2. Analyser et vérifier la pureté tumorale de patients leucémiques en quantifiant les composants des tumeurs par déconvolution et les comparer aux données cliniques;

3. Évaluer l'effet d'une correction des proportions de types cellulaires lors d'une analyse différentielle;

4. Étudier l'impact de l'infiltration immunitaire des cellules T dans des tumeurs solides en fonction de différents critères cliniques.

<u>Résultats attendus :</u>

- Une identification claire et cohérente de la composition des tumeurs leucémiques, notamment des sous types cellulaires spécifiques, et ce, grâce aux outils de déconvolution;

- Un impact positif de l'infiltration immunitaire dans les tumeurs solides pour prédire le pronostic clinique.

- Chapitre 2 -

Matériel et méthodes

1. Échantillons synthétiques – ARN-seq

Des échantillons synthétiques d'ARN-seq (100 millions de reads) ont été créé à partir de lignées cellulaires de différents types cellulaires hématopoïétiques.

Les lignées cellulaires sont disponibles publiquement et proviennent de prélèvements de moelle chez des patients atteints de leucémie myéloïde aigue (69). Des cellules souches hématopoïétiques (dont plus de 80% étaient pré-leucémiques), des cellules souches leucémiques, et des blastes, ont été isolés de ces patients (69). Plusieurs autres types cellulaires hématopoïétiques normaux ont été isolés de donneurs sains (cellules B, cellules T CD4+ et CD8+, cellules NK, monocytes, érythrocytes …) (69).

En partant de 100 millions de reads par échantillon, afin de pouvoir manipuler plus facilement les sous types comme des pourcentages (100M = 100%), il est possible de reproduire divers niveaux d'hétérogénéité cellulaire. Par exemple, 80 millions de reads d'un type cellulaire mélangés avec 20 millions de reads d'un autre type correspondrait à une pureté cellulaire de 80 %.

Pour ce faire, j'ai utilisé l'outil seqtk, installé sur le serveur de calcul Graham dans Calcul Canada afin d'effectuer mes analyses. Avec cet outil, il est possible à partir d'une ligne de commande sur Linux, de sélectionner un nombre de reads spécifique à partir du fichier FASTQ pour chaque lignée cellulaire. Cette commande est effectuée plusieurs fois sur les fichiers de séquençage bruts (fastq) des types cellulaires primaires hématopoïétiques selon des proportions variant de 5% à 80%. Une fois ces fichiers formés, il suffit de faire des combinaisons des différentes lignées cellulaires à l'aide d'une simple commande de concaténation sur Linux pour obtenir des fichiers totaux de 100 millions de reads. Dans notre cas, 13 échantillons synthétiques composés de différents types cellulaires avec

différentes proportions ont été générés (voir tableau V ci-dessous). Pour que ces échantillons synthétiques soient représentatifs des données étudiées, les types cellulaires les plus présents étaient les lymphocytes B et T. Les autres types cellulaires ont été introduits par soucis de ressemblance aux échantillons tumoraux hétérogènes, mais également pour voir la variation des estimations des outils avec des échantillons plus ou moins purs.

Par soucis de clarté, les noms de chacun des échantillons ont été formés en précisant les pourcentages des principaux types cellulaires attendus. Chaque ensemble cellulaire de reads provient d'un même échantillon de reads, par exemple c'est toujours le même fichier de reads pour les 5M d'érythrocytes. De plus, par soucis de simplicité les reads sont toujours pris dans le même ordre par la commande, donc les 5M de reads d'érythrocytes sont également inclus dans les 10M d'érythrocytes.

Tableau V. – Combinaisons des types cellulaires dans les échantillons synthétiques.

Échantillons	Lymphocytes B	Lymphocytes T	Monocytes	Lymphocytes NK	GMP	Érythrocytes	CMP	CSH	MPP	LMPP	MEP	CLP
80B_20T	80 M	20 M CD4	-	-	-	-	-	-	-	-	-	-
80B_20Mo	80 M	-	20 M	-	-	-	-	-	-	-	-	-
80B_5T_15others	80 M	5 M	5 M	-	-	5 M	-	5 M	-	-	-	-
80B_0T_20others	80 M	-	5 M	5 M	5 M	5 M	-	-	-	-	-	-
65B_20T_15others	65 M	20 M	5 M	-	-	5 M	-	5 M	-	-	-	-
65B_0T_35others	65 M	-	10 M	5 M	5 M	5 M	5 M	-	-	-	5 M	-
20B_65T_15others	20 M	65 M	5 M	-	-	5 M	-	5 M	-	-	-	-
20B_0T_80others	20 M	-	20 M	15 M	5 M	10 M	5 M	5 M	5 M	5 M	5 M	5 M
5B_80T_15others	5 M	80 M	5 M	-	-	5 M	-	5 M	-	-	-	-
0B_80T_20others	-	80 M	5 M	5 M	5 M	5 M	-	-	-	-	-	-
0B_65T_35others	-	65 M	10 M	5 M	5 M	5 M	5 M	-	-	-	5 M	-
0B_20T_80others	-	20 M	20 M	15 M	5 M	10 M	5 M	5 M	5 M	5 M	5 M	5 M
0B_20CD8_80others	-	20 M CD8	20 M	15 M	5 M	10 M	5 M	5 M	5 M	5 M	5 M	5 M

- : absent, M : millions, GMP : progéniteur granulocytaire monocytaire, CMP : progéniteur myéloïde, CSH : cellule souche hématopoïétique, MPP : progéniteur multipotent, LMPP : progéniteur lymphoïde multipotent, MEP : progéniteur érythroblastique et mégacaryocytaire, CLP : progéniteur lymphoïde.

2. Cohortes et données de transcriptome

2.1 Leucémies

Les échantillons de moelle osseuse ont été obtenus pour 184 enfants nouvellement diagnostiqués avec la LLA et inscrits dans des projets de recherche comme DFCI, QcALL, TRICEPS ou encore SIGNATURE (voir tableau VI) (12, 70, 71). Ces projets ont pour but d'offrir une approche moderne de médecine personnalisée aux patients afin de mieux les stratifier, d'ajuster leurs thérapies et d'augmenter leur probabilité de guérison. Les échantillons de patients sont traités au Centre Hospitalier Sainte-Justine (Montréal, Canada) et ont subi une analyse cytogénétique (caryotype, FISH) selon les normes, ainsi que des analyses de séquençage ARN avec une couverture moyenne de 179 millions de reads (71).

Tableau VI. – Cohorte des patients leucémiques

Variables	Type B	Type T
Total des patients	148	36
Projet		
DFCI	60	24
QcALL	64	7
SIGNATURE	14	2
TRICEPS	10	3
Age au diagnostic		
1-10 ans	102	20
≥ 10 ans	46	16
Sexe		
Féminin	81	7
Masculin	66	29
Non disponible	1	0
Altérations génétiques		
ETV6-RUNX1	40	0
HHD	36	0
BCR-ABL1	4	0
Ph-like	16	0
MLL	5	0
DUX4	11	0
TCF3-PBX1	7	2
Pre-T	1	28
Autres	28	6

DFCI : menée au « Dana-Farber Cancer Institute » pour des enfants nouvellement diagnostiqués d'une LLA, regroupant 9 autres institutions aux États-Unis et au Canada, et cherchant de nouveaux traitements pour réduire les effets à long terme. **QcALL:** enfants de descendance européenne, diagnostiqués au Québec avec une LLA. Les échantillons sont pris au moment du diagnostic, pendant les deux ans de traitement et lors de la rechute ou avant/après transplantation. Étude s'appuyant sur les méthodes -omiques pour identifier de nouveaux facteurs génétiques pour améliorer les outils diagnostics et pronostics. **SIGNATURE:** depuis 2017, pour performer des analyses moléculaires d'enfants nouvellement diagnostiqués du cancer dans la province de Québec et en collaboration avec 3 autres centres pédiatriques. **TRICEPS**: depuis 2014, implique la caractérisation de génomes tumoraux pour identifier des mutations actionnables et fournir une thérapie ciblée personnalisée pour les enfants atteints de cancers qui rechutent.

2.2 Tumeurs solides

Les données d'expression des gènes des tumeurs solides utilisées proviennent de la cohorte TARGET (https://ocg.cancer.gov/programs/target/data-matrix), préalablement mis à notre disposition sur le serveur Graham par une équipe de notre laboratoire. Il y a parmi cela 164 Neuroblastomes, 171 Ostéosarcomes et 100 tumeurs de Wilms.

Les informations sur les critères clinique des patients ont également été prélevés sur le site du consortium TARGET en fonction des échantillons sélectionnés précédents, dans le répertoire des fichiers des métadonnées cliniques des tumeurs en question.

Seulement une partie des paramètres cliniques présents dans les fichiers ont été retenus, soit ceux qui pouvaient potentiellement avoir une association avec l'infiltration immunitaire. Ainsi, les variables cliniques sont les suivantes : le sexe, le pronostic vital (vivant ou décédé), les évènements (pas d'évènements ou mauvais présage), le temps sans évènements, l'âge au diagnostic et le temps de survie. Ce sont ces données qui seront corrélées par la suite avec les scores d'infiltration.

Une fois ces deux ensembles de données réunis, en joignant les identifiants SRR et les identifiants TARGET et en enlevant les échantillons dupliqués (différents SRR pour un même patient), il y a au final 146 échantillons de neuroblastomes, 86 cas d'ostéosarcomes et 50 échantillons de tumeurs de Wilms (voir tableau VII).

Tableau VII. – Cohorte des trois types de tumeurs solides TARGET

Variables	Neuroblastomes	Ostéosarcomes	Tumeurs de Wilms
Total des patients	146	86	50
Race/Ethnicité			
Blanc	105	59	32
Noir ou Afro-américain	27	7	11
Inconnu	14	10	7
Age moyen au diagnostic	3,5	15	4,6
Sexe			
Féminin	61	37	25
Masculin	85	49	25
Évènements			
Aucun	0	32	16
Rechute	42	37	32
Progression	19	0	2
Mort	14	2	0
Autres	15	1	0
Inconnu	56	14	0
Statut vital			
Vivant	75	55	28
Décédé	71	29	22
Inconnu	0	2	0

3. Outils de déconvolution

J'ai utilisé 3 outils de déconvolution : Cibersort, DeconRNAseq et Quantiseq, qui peuvent estimer les résultats sous forme de fractions cellulaires (57, 66, 67).

Un modèle de déconvolution d'expression des gènes avec matrice de référence peut être schématisé par la figure ci-dessous. Il peut être exprimé comme un problème matriciel approximatif comme suit : $M \approx S \times F$. Ce problème peut ensuite être résolu comme une équation linéaire indépendante ayant une solution (le nombre de solutions peut être trouvé avec le théorème de Capelli-Fontené-Frobenius-Kronecker-Rouché) (72). Ainsi, le

mélange tumoral *M*, composé d'un certain nombre *i* de gènes exprimés dans différents types cellulaires, peut être décomposé sous la forme d'une équation linéaire (52). Cette équation correspond au produit entre les profils d'expression de gènes dans des lignées cellulaires d'une matrice signature, généralement fournie par l'outil, et le facteur *F* recherché, correspondant à la proportion relative des types cellulaires observés dans la tumeur hétérogène (52). Si la condition d'avoir plus d'équations que d'inconnus est respectée; c'est à dire dans notre cas d'avoir plus de gènes que de type cellulaires dans le système; alors, en résolvant cette équation pour *F*, les algorithmes de déconvolution sont capables d'estimer ces fractions (52).

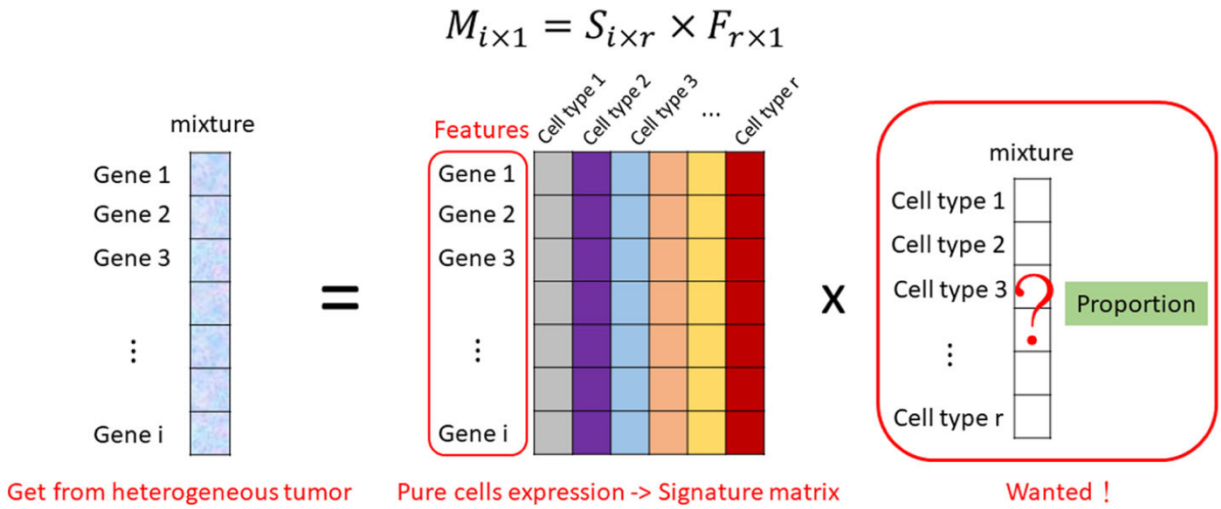


Figure 5. – Représentation d'un modèle de déconvolution d'expression des gènes. M désigne la matrice du mélange tumoral avec l'expression pour chaque gène, S correspond à la matrice signature de référence avec l'expression des gènes de référence dans les lignées pures, et F représente ce qui est recherché, soit la fraction des différents types cellulaires dans le mélange tumoral. Figure tirée de Chen et al., 2018 (50).

3.1 Cibersort

Cibersort est un outil de déconvolution souvent cité dans la littérature (59, 67). Le pipeline de traitement des données ARN-seq décrit précédemment, fourni à la fin de son exécution un fichier dans lequel on retrouve l'expression en FPKM de tous les gènes présents pour

chaque échantillon tumoral. Une matrice d'expression composée des identifiants HUGO des gènes sur chaque ligne et les différents échantillons sur chaque colonne est ainsi créée et donnée en entrée à Cibersort (59). Dans mon projet, j'ai exécuté l'algorithme de Cibersort à l'aide d'un code R, avec le fichier source disponible sur le portail web (http://Cibersort.stanford.edu/). Il est également possible d'exporter sa matrice sur le site et de lancer le programme sur la plateforme (59). À savoir que l'exécution sur R donne une matrice des résultats, tandis que le lancement sur le site fourni automatiquement une visualisation sous forme de graphiques.

Cet outil utilise une méthode v-SVR (support vector regression) pour trouver le vecteur *F*, soit la fraction de chaque type cellulaire (59, 67). Cette méthode définit un espace vectoriel capturant le plus de points possibles en fonction de contraintes données, et en pénalisant les points à l'extérieur d'un rayon d'erreur, faisant d'elle une méthode plutôt robuste aux populations cellulaires inconnues (67). De plus, cette méthode de calcul introduit une norme de façon à minimiser la variance entre des sous-types pouvant être corrélés entre eux (67).

L'algorithme a été opéré en utilisant la matrice d'expression de nos mélanges, la matrice signature LM22 étant celle par défaut pour Cibersort, et en sélectionnant 100 permutations, soit le minimum recommandé pour obtenir une meilleure rigueur statistique (59). Cette matrice par défaut LM22 contient l'expression de 547 gènes dans 22 sous types cellulaires hématopoïétiques purs d'origine humaine, conçue et validée sur des données de micropuces, mais applicable aux données d'ARN-seq (59). Il est aussi possible de créer sa propre matrice signature et la fournir à CIBSERSORT (59).

Les résultats de Cibersort sont représentés sous forme de matrice dans laquelle l'addition de tous les types cellulaires présents dans chaque échantillon donne 1, pouvant alors être interprétés comme la proportion de chaque type cellulaire. Quelques données statistiques telles que la valeur-p, la valeur RMSE et le coefficient de corrélation de Pearson sont également données pour chaque échantillon (59).

3.2 DeconRNAseq

DeconRNAseq est un outil de déconvolution applicable aux données d'ARN-seq, un peu moins cité, également disponible à partir d'une librairie R nommée « DeconRNAseq »

installable à partir de Bioconductor (57). Il faut fournir à ce dernier la même matrice d'expression normalisée en FPKM des mélanges tumoraux que celle employée avec Cibersort, en utilisant toutefois la nomenclature ENSEMBL pour les identifiants des gènes (57). Cette nomenclature est utilisée pour pouvoir correspondre à la nomenclature des identifiants de gènes de la matrice signature obtenue publiquement, composée de 13 lignées cellulaires hématopoïétiques pures (69). Cet outil peut être couplé avec n'importe quelle matrice signature ayant un lien avec les données tumorales à analyser (57).

Lors de l'exécution, la matrice signature et la matrice de nos échantillons tumoraux sont fournies sous la forme de tableaux de données R. Elles sont ensuite soumises à un algorithme utilisant une optimisation quadratique résolvant un problème de contrainte des moindres carrés non négatif (57). Ce dernier permet généralement d'obtenir une solution globalement optimale en préservant la non-négativité des proportions cellulaires estimées (57). Les résultats de l'exécution sont enregistrés dans une variable de R sous forme de matrice, et sont sauvegardés dans un fichier résultats par la suite. Encore une fois, la somme des différents types cellulaires est aussi égale à 1 pour chaque échantillon, pour pouvoir être considéré comme des pourcentages cellulaires.

3.3 Quantiseq

Quantiseq quant à lui, est un outil conçu spécifiquement pour les données d'ARN-seq (62) qui permet d'effectuer directement toute l'analyse à partir des « reads » d'ARN-seq au format FASTQ (62, 66). Ainsi, ce dernier exécute le pipeline pour les étapes de quantification et de normalisation, et jusqu'à l'estimation des fractions cellulaires par déconvolution (62, 66). À noter qu'il est également possible de fractionner ce pipeline et de le débuter directement au niveau de la déconvolution. Pour mon projet, j'ai choisis de commencer dès le traitement des reads (62). C'est donc des fichiers de reads qui ont été donnés en entrée. L'exécution du pipeline complet permettrait d'augmenter ainsi la robustesse de l'outil, en diminuant les impacts des changements entre les différentes étapes, pour obtenir moins d'incohérences dans les estimations (62, 66).

Dans notre cas, par soucis d'efficacité et de rapidité, un script bash généralisé est écrit pour effectuer ce pipeline sur tous nos échantillons en même temps. Les fichiers d'entrée sont des tableaux de trois colonnes, contenant chacune respectivement, le nom de

l'échantillon, le chemin vers le fichier du read « forward » et le chemin vers le fichier du read « reverse » (62).

L'outil utilise une matrice signature de 153 gènes provenant de données d'ARN-seq et 170 gènes de données de micropuces, pour pallier le potentiel manque de gènes signatures, et ce, pour 10 sous types cellulaires (62).

L'algorithme est alors exécuté selon la méthode décrite sur le site web (https://icbi.i-med.ac.at/software/quantiseq/doc/). Quelques modifications au niveau des paramètres par défaut ont été effectuées dans notre cas. Notamment, l'utilisation de la méthode de déconvolution choisie a été changé pour « bisquare », en référence à la fonction de régression linéaire robuste de Tukey, car la méthode « lsei » par défaut donne un grand nombre de cellules estimées comme « others », ne spécifiant pas leur type spécifique (62, 66). Le nombre de « threads » a été fixée à 8 dans notre cas, ce dernier est à adapter en fonction de la taille des fichiers, et l'option « tumor =TRUE » a été spécifiée pour correspondre à l'analyse d'échantillon de provenance tumorale (62, 66). Les résultats sont encore une fois fournis sous forme de matrice pouvant être considérer comme des proportions cellulaires.

4. Analyse d'expression différentielle de gènes

Dans le volet leucémies, les deux sous types majoritaires, soit HHD et ETV6-RUNX1 ont été sélectionnés pour effectuer cette analyse. Ainsi, 5 échantillons dans chaque sous-groupe sont choisis, en veillant à choisir des échantillons avec des pourcentages en cellules B et T estimés tout de même assez variables, et ce, à partir des résultats obtenus avec DeconRNAseq.

Tableau VIII. – Fichier de métadonnées des échantillons

échantillons	sous-types	% lymphocytes B	% lymphocytes T
152103_SIGN0050_T	ETV6_RUNX1	90,8	5,3
152425_DFCI16-074_T	ETV6_RUNX1	95,4	1,5
155830_SIGN0126_T	ETV6_RUNX1	81,1	15,3
156880_SIGN0173_T	HHD	81,3	5,2
158780_DFCI16-238_T	ETV6_RUNX1	60,3	12,2
158813_SISJ0223_T	HHD	90,9	5,7
777_T	HHD	64,9	10,2
901_T	ETV6_RUNX1	34,7	0
981_T	HHD	24,8	48
DFCI16-015_T	HHD	95,3	1,9

Un fichier de métadonnées, représenté par le tableau ci-dessus, est ainsi créé, dans lequel les sous-types et les pourcentages sont spécifiés pour chaque échantillon. Des fichiers de comptes normalisés des différents gènes pour chaque échantillon provenant du pipeline exécuté auparavant sont également récupérés, et placés ensemble dans une matrice avec les identifiants de gènes sur les lignes et les échantillons en colonnes. Un code R préalablement écrit par notre laboratoire, et disponible sur le serveur, permet d'enlever les gènes trop faiblement exprimés (ayant un nombre de compte trop bas < 3) ou d'enlever certains identifiants non voulus, notamment à l'aide d'un fichier contenant ces identifiants spécifiques.

Les expressions différentielles sont ensuite quantifiées à partir de la librairie DESeq2, dans laquelle les 10 échantillons (5 de chaque type) sont corrigés ou non avec le pourcentage estimé pour les types cellulaires sélectionnés. Ceci se fait en ajoutant ces données en tant que co-variables au début du paramètre de design dans la commande lors de la formation de l'objet DESeq (73). La correction du pourcentage cellulaire permettrait de réduire l'expression de gènes liés spécifiquement aux différents types cellulaires et plutôt faire ressortir les différences entre les sous-types. Le changement de condition entre les analyses provient donc du fait que dans le premier design, la seule co-variable est celle des sous-types (HHD ou ETV6), tandis que dans la deuxième, les pourcentages en lymphocytes B et T sont ajoutés.

Une fois les résultats obtenus dans une matrice, des librairies de R ont été utilisées pour créer des graphiques pour visualisation. Ainsi, des graphiques de corrélation avant et après correction, pour le logFoldChange, indice du changement d'expression, puis pour la valeur-p, indice de changement de significativité, ont été générés. Des graphiques volcans avant et après correction sont aussi produits. Des seuils de significativité sont ensuite choisis pour identifier les gènes les plus différentiellement exprimés dans les deux conditions. Nous avons utilisé un seuil qui permet d'identifier des gènes significativement différentiellement exprimés lorsque -log10(p-value) est au-dessus de 5, c'est à dire la p-value en dessous de 0,00001 (soit 10^{-5}). Toutes les p-values sont transformées sous base de -log10(p), donc plus la valeur de -log10(p) augmente et plus la valeur p diminue. Pour le log2FoldChange, qui est en log2, le seuil est aux valeurs inférieures à -5 pour les gènes sous-exprimés, et supérieures à 5 pour les gènes surexprimés. Les symboles du top 20 de ces gènes le plus différentiellement exprimés sont ensuite affichés sur les graphiques, en affichant un maximum de gènes possibles sans qu'il y ait de superpositions.

5. Prédiction de l'impact fonctionnel

Une fois les résultats de l'analyse différentielle obtenue, les 100 premiers gènes les plus différentiellement exprimés avant et après correction sont sélectionnés. Des fichiers textes contenant la liste des identifiants Ensembl de ces gènes sont ensuite créés. Ces listes sont fournies au logiciel web Metascape afin de les associer aux termes GO et prédire leur fonction biologique (74). Il faut s'assurer de garder seulement le système de classification de la base de données Gene Ontology dans les paramètres. Les résultats sont générés sous forme de graphiques avec les termes GO les plus représentés et significatifs, leur description, ainsi que la valeur p sous forme de -log10(p). Il est possible de télécharger les résultats de tous les termes GO sous forme de tableau Excel avec les termes, leurs descriptions, leurs valeurs p et les gènes auxquels ils sont associés. Dans notre cas une corrélation de la significativité de ces termes GO a été effectuée avant et après correction pour avoir une meilleure visibilité sur les changements.

6. Scores d'infiltration immunitaire

Les scores d'infiltration immunitaire obtenus seront par la suite corrélés avec les données cliniques des tumeurs solides.

6.1 Score absolu de Cibersort

Cibersort est aussi capable de calculer un score mesurant l'abondance de chaque type cellulaire (59). Ce score immunitaire est estimé directement par Cibersort en fournissant les mêmes paramètres que précédemment, c'est à dire la matrice d'expression des tumeurs solides, la même matrice signature par défaut LM22, mais en rajoutant également cette fois-ci, l'option « absolute=TRUE » pour avoir les résultats sous forme de scores et non sous forme de pourcentages de types cellulaires (59, 67). Ce score provient d'un calcul impliquant la médiane de l'expression de tous les gènes dans la matrice de référence, divisée par la médiane de l'expression de tous les gènes dans le mélange (59, 67). Cette méthode semble capturer le contenu immunitaire de chaque type cellulaire présent, et donne le résultat de ces scores sous forme de matrice avec les échantillons sur les lignes et chaque type cellulaire en colonne (67). Des valeurs statistiques et une colonne supplémentaire représentant le score immun global de tous les types cellulaires sont également présentes. Dans notre cas, le score retenu est celui représentant l'ensemble des lymphocytes T.

6.2 Score d'expression des cellules T

Une autre méthode de score a été proposée dans un article de Danaher et al., pour estimer le taux d'infiltration des cellules immunitaires dans les tumeurs solides (75). Cette fois, le score est basé sur l'expression de gènes marqueurs spécifiques à chaque type cellulaire. Le score avec les gènes marqueurs pour l'infiltration des lymphocytes T est celui qui donne les résultats les plus cohérents et vraisemblables (75). Ce score est basé sur la proposition que la moyenne des logarithmes en base 2 de l'expression normalisée des gènes marqueurs est reliée à la proportion de ces types cellulaires spécifiques dans la tumeur (75).

Les six gènes marqueurs reportés pour les lymphocytes T sont ceux disponibles dans leur tableau récapitulatif, soit : CD6, CD3D, CD3E, SH2D1A, TRAT1, CD3G (75).

Les gènes marqueurs sont souvent associés à un type cellulaire lors d'analyse de données réalisée par un laboratoire et une méthode en particulier. De ce fait, plus le nombre de gènes marqueurs est important et plus il y a de chances de retrouver ces gènes dans nos données analysées, et ainsi garantir une meilleure spécificité en termes de sous type (76). Intuitivement, un nombre plus élevé de marqueurs indique des aspects uniques dans le programme transcriptionnel d'une population, ce qui devrait augmenter sa chance d'être identifié à travers les ensembles de données (76).

Le tableau ci-dessous répertorie les marqueurs identifiés dans l'étude pour les autres types cellulaires.

Tableau IX. – Gènes marqueurs candidats pour identifier les types cellulaires

Sous types	Nombre de gènes candidats	Statistique de similarité moyenne par paire dans TCGA	Gènes marqueurs sélectionnés
Lymphocytes B	34	0,59	BLK, CD19, FCRL2, MS4A1, KIAA0125, TNFRSF17, TCL1A, SPIB, PNOC
CD45	1	[a]NA	PTRPC
Lymphocytes cytotoxiques	18	0,69	PRF1, GZMA, GZMB, NKG7, GZMH, KLRK1, KLRB1, KLRD1, CTSW, GNLY
Cellules dendritiques	7	0,46	CCL13, CD209, HSD1B1
CD8 épuisées	5	0,44	LAG3, CD244, EOMES, PTGER4
Macrophages	33	0,71	CD68, CD84, CD163, MS4A4A
Mastocytes	31	0,74	TPSB2, TPSAB1, CPA3, MS4A2, HDC
Neutrophiles	32	0,48	FPR1, SIGLEC5, CSF3R, FCAR, FCGR3B, CEACAM3, S100A12
Cellules NK CD56dim	14	0,40	KIR2DL3, KIR3DL1, KIR3DL2, IL21R
Lymphocytes NK	36	0,47	XCL1, XCL2, NCR1
Lymphocytes T	13	0,81	CD6, CD3D, CD3E, SH2D1A, TRAT1, CD3G
Lymphocytes Th1 (auxiliaire)	27	[a]NA	TBX21
Lymphocytes T régulateurs	18	[a]NA	FOXP3
Lymphocytes T CD8	35	0,51	CD8A, CD8B
Cellules CD4	20	NA	

[a]Un seul gène marqueur, impossible d'obtenir un score. TCGA : Atlas du Génome du Cancer. Les types cellulaires qui n'ont pas de gènes marqueurs acceptables sont exclus. Un score de similarité statistique de 1 indique une linéarité parfaite avec ces marqueurs. Tableau adapté de Danaher et al., 2017 (75).

Ainsi, à partir de la matrice d'expression de nos gènes, la même que celle fournie pour le calcul du score avec Cibersort, le calcul du score d'expression est effectué avec la méthode prescrite, mais en y apportant tout de même de légères modifications (75). En effet, pour des soucis de visualisation, et que pour tous les scores obtenus soient supérieurs à 0, la formule devient alors : moyenne des log2(FPKM gènes marqueurs +1).

7. Analyses des données et graphiques

Toutes les analyses de données de ce projet sont effectuées à l'aide de logiciels et outils existants et de codes écrits et exécutés en langage R. Les figures des résultats proviennent du traitement de données sur R également, notamment par l'usage de librairies telles que ggplot et cowplot.

Des composantes statistiques ont aussi été calculées sur R, par exemple les droites linéaires et les coefficients de corrélation selon la méthode de Pearson. Des valeurs p, avec la méthode de Kruskal-Wallis lorsque plus de deux paramètres sont comparés, ou bien avec la méthode de Wilcoxon lorsque la comparaison s'effectue entre deux variables seulement; ont toutes été générées avec des fonctions dans R, notamment dans la librairie statistique « dplyr ».

- Chapitre 3 -

Utilisation de données synthétiques pour évaluer l'efficacité des outils de déconvolution

a. Déconvolution

J'ai généré 13 échantillons synthétiques représentatifs de diverses compositions et proportions en types cellulaires. Ces données synthétiques ont ensuite été utilisées pour tester l'efficacité de trois outils de déconvolution (DeconRNAseq, Cibersort, Quantiseq) qui seront utilisés pour analyser nos données leucémiques ultérieurement.

13 échantillons synthétiques ont été générés à partir de types cellulaires primaires hématopoïétiques selon des proportions variant de 5% a 80% (Figure 6) (69). Des matrices d'expression de gènes de ces 13 échantillons synthétiques sont alors fournies à DeconRNAseq et Cibersort, et les fichiers de séquençage bruts à Quantiseq.

L'outil DeconRNAseq (Figure 6A) estime une proportion considérable de cellules lymphoïdes progénitrices (CLP), largement supérieure à ce qui est attendu, et ce, à chaque fois qu'un taux significatif de lymphocytes B est également présent dans le mélange. Les estimations de cellules B dépassent rarement les 25%, alors que plus de 65% sont attendus dans les six premiers échantillons. On remarque également un estimé plus bas de cellules T que ce qui est attendu. Cet écart est plus grand lorsque des cellules B sont présentes dans le mélange (20B_65T_15others par exemple). Par ailleurs, certains échantillons présentent des estimations pour des types cellulaires qui ne sont pas censés être dans la composition de base. À titre d'exemple la présence d'érythrocytes dans l'échantillon 80B_20T qui ne contient que des cellules B et T, ou la présence de cellules NK, qui sont estimées à plus de 40% dans l'échantillon 0B_20T_80others, alors qu'un maximum de 15% est attendu pour ce type cellulaire.

Pour Cibersort (Figure 6B), certains types cellulaires sont encore une fois surestimés dans quelques échantillons. Cette fois, il s'agit plutôt des lymphocytes T ou des monocytes, ces derniers étant d'autant plus surestimés lorsque les lymphocytes B sont en quantité moindre dans les échantillons. Par exemple dans l'échantillon 0B_20T_80others, l'outil estime la présence de 40% de cellules T et 40% de monocytes, alors que dans les deux cas il n'est pas attendu plus de 20% de ces types cellulaires. De même pour l'échantillon 20B_0T_80others qui affiche environ 20% de lymphocytes T et plus de 40% de monocytes, qui est donc encore une fois surestimé puisqu'il ne doit normalement pas être composé de lymphocytes T, et pas plus de 20% de monocytes.

Quantiseq quant à lui est l'outil parmi les trois qui surestime le plus les lymphocytes T, particulièrement lorsque ce type cellulaire est attendu en grande quantité dans l'échantillon et que les autres types sont moindres (Figure 6C). Par exemple, dans les échantillons 5B_80T_20others et 0B_65T_35others qui contiennent respectivement 80% et 65% de cellules T, l'outil estime la proportion de cellules T entre 95 et 98%. Cependant, Quantiseq est l'outil qui introduit le moins de types cellulaires qui ne sont pas présents dans les échantillons de base.

En résumé, les trois outils se comparent au niveau des estimations, aucun ne semble se démarquer avec ces échantillons à ce stade-ci. Ils seront donc tous utilisés pour les analyses subséquentes de déconvolution sur des données de patients leucémiques.

b. Score d'expression des cellules T

J'ai ensuite voulu déterminer la corrélation entre un score d'expression des cellules T et les estimations en lymphocytes T faites par chacun des outils. L'intérêt de prendre les pourcentages estimés par l'outil et non les pourcentages attendus est de savoir si les estimations faites dépendent de la méthode utilisée par l'outil, ou plutôt d'un processus antérieur déjà présent dans la matrice d'expression.

Les trois outils ont une corrélation positive et significative avec des valeurs p en dessous de 0,05 (Figure 6D-F). La corrélation la plus forte et la plus significative est obtenue avec Cibersort, qui montre un coefficient de Pearson de 0,94 entre les cellules estimées par l'outil et le score de la matrice d'expression (Figure 6E), suivi par Quantiseq (Figure 6F) et enfin DeconRNAseq (Figure 6D).

Ces résultats suggèrent que certaines surestimations ne sont pas complètement dépendantes de la méthode employée par l'outil et sont parfois déjà présentes avant son exécution.

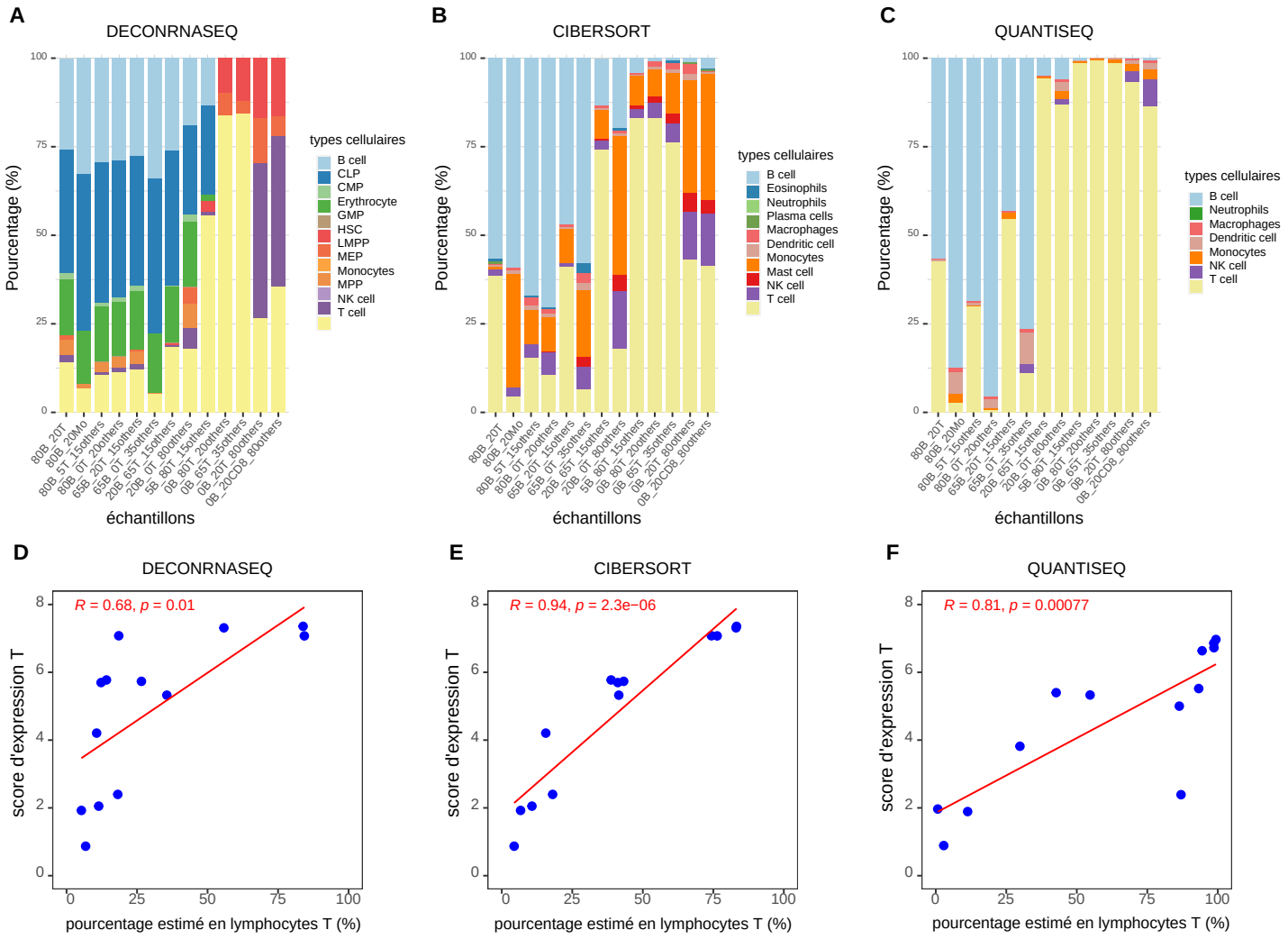


Figure 6. – Déconvolution de données synthétiques et corrélation des cellules T. Résultats de déconvolution pour les estimations des différents types cellulaires avec DeconRNAseq (A), Cibersort (B) et Quantiseq (C). Les noms des échantillons synthétiques en abscisse représentent la composition et la proportion types cellulaires attendus (voir Matériel et Méthodes). Corrélation entre le score d'expression T et les pourcentages de lymphocytes T estimés par DeconRNAseq (D), Cibersort (E) et Quantiseq (F). L'axe des abscisses représente le pourcentage estimé en cellules T pour chaque outil. Le score d'expression T représente la moyenne des log2 (de l'expression de gènes marqueurs de cellules T + 1). La droite rouge suit une fonction de régression linéaire simple, la valeur R représente le coefficient de corrélation selon la méthode de Pearson, et la significativité est représentée par la valeur p.

- Chapitre 4 -

Estimation de la pureté tumorale par déconvolution dans des échantillons leucémiques

1. Déconvolution

J'ai utilisé les outils DeconRNAseq, Cibersort et Quantiseq pour estimer la pureté tumorale dans 184 patients leucémiques de type B (n=148) et T (n=36) (57, 59, 62). Les résultats des estimations sont représentés à la Figure 7.

DeconRNAseq et Quantiseq affichent des pourcentages élevés de lymphocytes B, avec plus de la moitié des échantillons présentant des pourcentages à plus de 75% pour ce type cellulaire (Figure 7A,C). Quantiseq estime cependant des plus hauts taux de lymphocytes T, tandis que DeconRNAseq présente quant à lui un nombre considérable d'échantillons de type B plutôt hétérogènes avec plusieurs types cellulaires présents dans un même pourcentage.

Cibersort de son côté estime rarement plus de 60% de lymphocytes B dans les échantillons, mais ces proportions peuvent atteindre 80% pour les lymphocytes T (Figure 7B). Cibersort arrive mieux à caractériser les type T, mais surtout à différencier et séparer distinctement les types B et types T, comparativement aux deux autres outils.

En résumé, chaque outil de déconvolution a ses points forts et ses faiblesses vis à vis de l'estimation des différents types cellulaires.

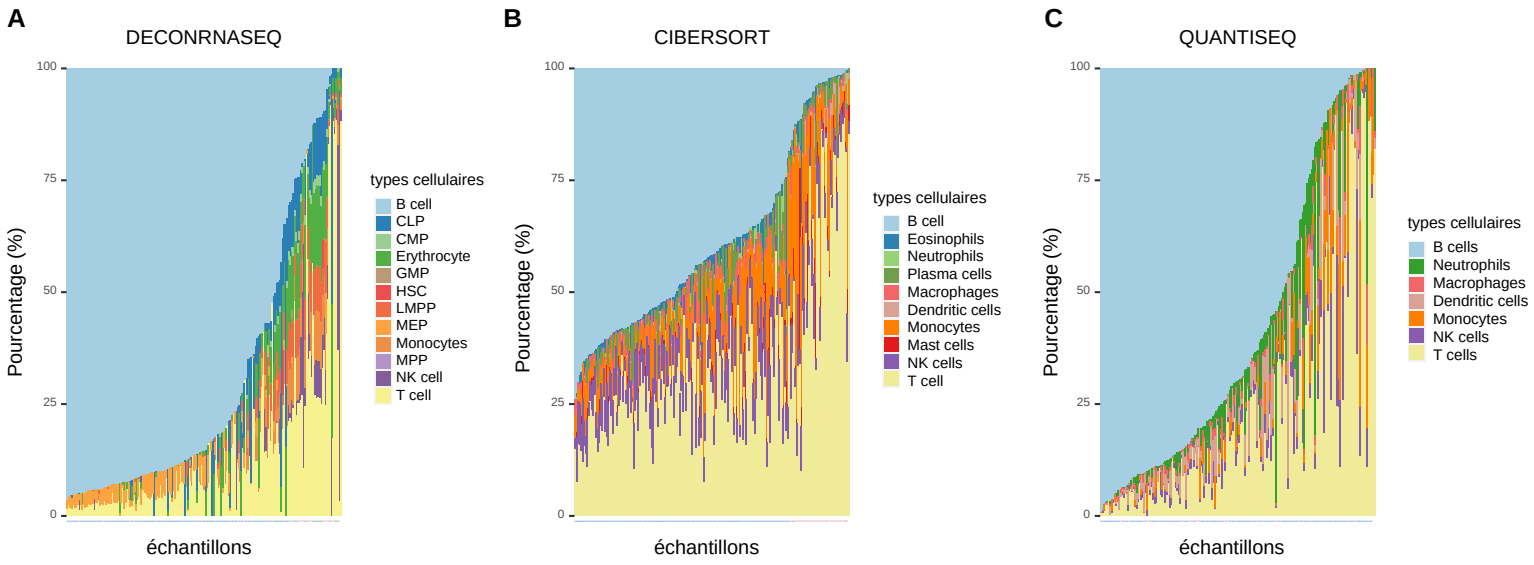


Figure 7. – Estimation de la composition tumorale des LLA par déconvolution. Résultats de déconvolution pour chaque outil : DeconRNAseq (A), Cibersort (B) et Quantiseq (C). Les échantillons en abscisses sont ici représentés et colorés en fonction de leur type, soit leucémie de type B en bleu et leucémie de type T en rouge. La légende de chaque graphique indique les divers types cellulaires hématopoïétiques présents en fonction de la matrice signature de chaque outil.

2. Corrélation entre la déconvolution et la pureté tumorale clinique

J'ai ensuite déterminé la corrélation entre les estimations de types cellulaires obtenues par les trois outils de déconvolution et les pourcentages de blastes leucémiques recueillis au niveau clinique. Les résultats pour les leucémies B (graphiques A,B,C) et T (graphique D,E,F) sont présentés à la figure 8.

DeconRNAseq présente une bonne corrélation avec les données cliniques pour les lymphocytes B (en haut à droite de la Figure 8A). Pour Cibersort, même si une légère corrélation significative est présente avec les pourcentages cliniques, aucun estimé ne dépasse les 75% de lymphocytes B (Figure 8B). Quantiseq quant à lui est plutôt similaire à DeconRNAseq, il affiche même une meilleure corrélation pour les lymphocytes B avec une valeur-p significative (Figure 8C).

Pour ce qui est de l'estimation des cellules T, il est possible d'observer que DeconRNAseq les estiment moins bien, avec la majorité des échantillons estimés à moins de 25% de lymphocytes T, même lorsque les pourcentages cliniques sont proches des 100% (Figure 8D). Cibersort au contraire, produit des estimations un peu plus élevées dépassant pour certaines les 75% de lymphocytes T, tout en étant en cohérence avec les données cliniques (Figure 8E). Les résultats de Quantiseq sont plutôt similaires entre les deux types (B et T), avec un bon nombre d'échantillons estimé à plus de 75% de lymphocytes également (Figure 8F). Les valeurs-p ne permettent pas ici de supporter la corrélation entre les pourcentages de lymphocytes T exprimés et les pourcentages cliniques.

En résumé, la performance des outils est variable selon le type de leucémie, mais peut également varier en fonction des différences de purification entre les échantillons cliniques et nos échantillons.

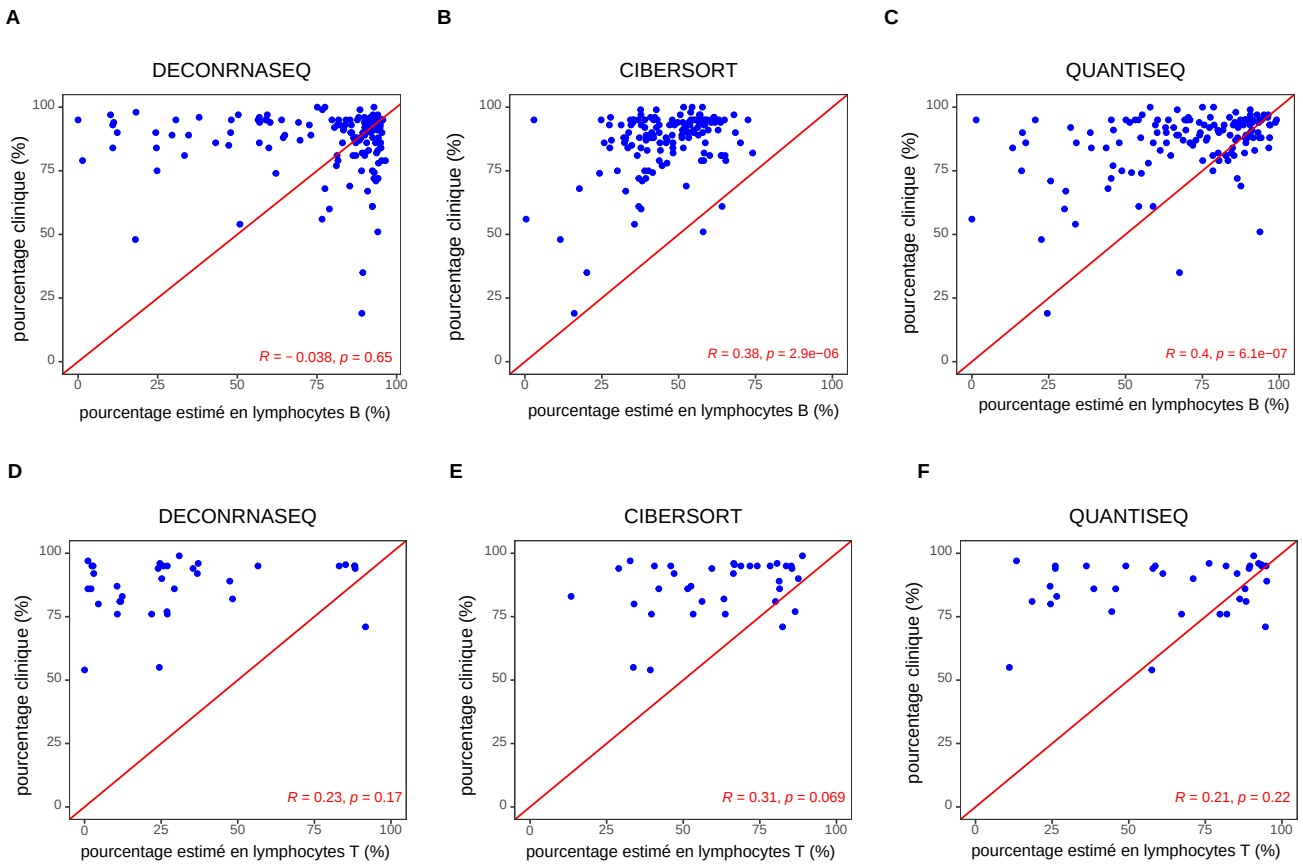


Figure 8. – Corrélations entre les estimations des outils et les données cliniques pour les 2 types majeurs de LLA. Corrélations entre les pourcentages de blastes cliniques pour les 148 patients de type B et les estimations en lymphocytes B obtenues avec DeconRNAseq (A), Cibersort (B) et Quantiseq (C). Corrélations avec les pourcentages de blastes cliniques de 36 patients de type T et les estimations en lymphocytes T obtenues avec les mêmes outils : DeconRNAseq (D), Cibersort (E) et Quantiseq (F). Le coefficient de corrélation R est calculé avec la méthode de Pearson, avec la p-value associée, et la droite représente le cas idéal qui suit la fonction x=y.

3. Score d’expression des cellules T

Nous utilisons le score d’expression des cellules T afin de vérifier si les différences observées au niveau des outils de déconvolution proviennent des outils eux-mêmes. Ainsi, de manière similaire qu’avec les données synthétiques, le score d’expression T est calculé une nouvelle fois à partir des matrices d’expression, puis corrélé avec l’estimation des lymphocytes T des outils, et ce, pour tous les échantillons (Figure 9).

Parmi les trois outils, DeconRNAseq est le moins performant pour l’estimation des lymphocytes T car plusieurs échantillons sont estimés proches de 0 et peu dépassent les 25%, alors que le score d’expression oscille entre 0 et 7 (Figure 9A). De plus, l’outil affiche un coefficient de corrélation de 0,4, qui est inférieur aux deux autres outils. Cibersort et Quantiseq ont des coefficients de corrélation proches de 0,8, ayant en plus de cela, des valeurs p largement inférieure à 0,05 (Figure 9B,C). Cibersort et Quantiseq exhibent une corrélation positive.

Ces résultats suggèrent que la méthodologie derrière les outils de déconvolution a une influence considérable sur certaines estimations des proportions cellulaires.

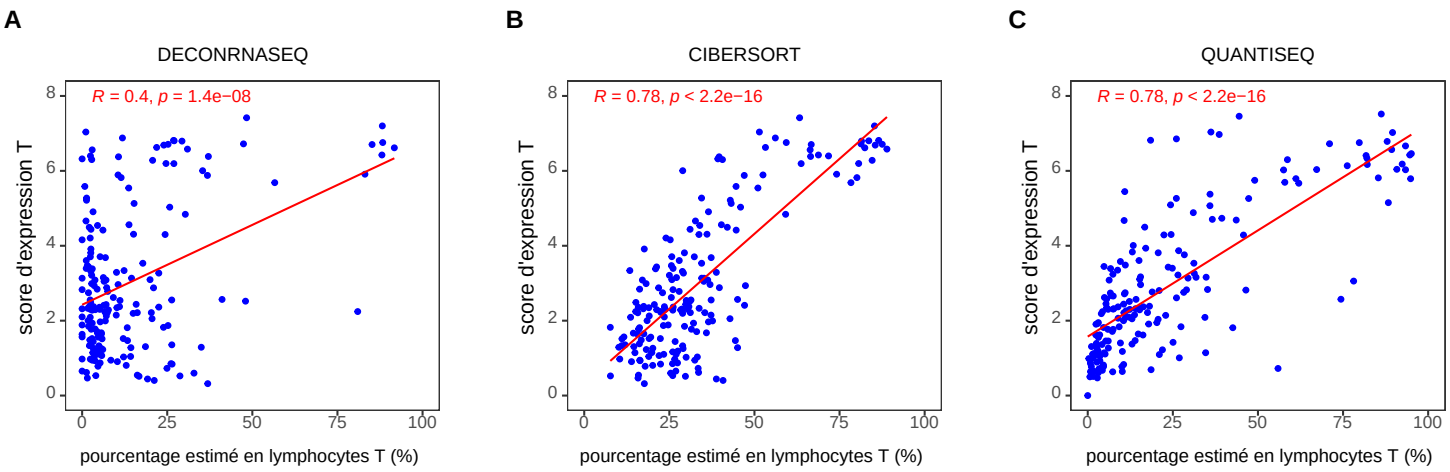


Figure 9. – Corrélations entre le score d'expression T et les estimations des lymphocytes T par les outils de déconvolution. Les pourcentages estimés de cellules T obtenues avec les outils DeconRNAseq (A), Cibersort (B) et Quantiseq (C) sont en abscisses, corrélés avec le score d'expression T en log2 en ordonnée. Le score d'expression T représente la moyenne des log2 (de l'expression de gènes marqueurs de cellules T + 1). La droite rouge suit une fonction de régression linéaire simple, et la valeur R représente le coefficient de corrélation selon la méthode de Pearson, avec la p-value associée.

4. Évaluation de l'effet des proportions cellulaires sur l'analyse d'expression différentielle de gènes

J'ai voulu vérifier l'implication des pourcentages des types cellulaires estimés sur l'émergence ou la disparition des gènes différentiellement exprimés. Pour ce faire, j'ai sélectionné dix échantillons de deux sous types de leucémie, soit HHD (n=5) et ETV6-RUNX1 (n=5), tout en sélectionnant des proportions variables en lymphocytes B et T parmi les résultats estimés par DeconRNAseq pour chaque groupe. Deux analyses différentielles ont été réalisées, la première impliquant seulement la différence d'expression de gènes entre les sous-types, et la deuxième en incluant les proportions cellulaires estimées comme co-variables dans la formule du design de l'analyse différentielle afin de corriger pour les résultats de la déconvolution. Cette correction devrait mieux faire ressortir les différences entre les sous-types, et moins les variances entre les types cellulaires.

Nous avons effectué les corrélations entre les logFoldChange avant et après correction, et les corrélations pour les valeurs p, avant et après correction (Figure 10A,B). On observe que les valeurs avant et après correction sont fortement corrélées positivement avec des coefficients supérieurs à 0,9, et de façon très significative, de l'ordre de 10^{-16}. En effet, en considérant l'ensemble des résultats, la correction des proportions cellulaires a peu d'effets. La majorité des valeurs ne fluctue pas trop avant et après correction. Il y a tout de même certains points pour lesquels ces données oscillent et qui vont nécessiter une attention plus particulière.

Des graphiques volcans présentent les résultats des deux analyses différentielles, respectivement avant puis après correction (Figure 10C,D). Les points verts sur les graphiques représentent les gènes les plus différentiellement exprimés, il est alors intéressant de comparer les changements au niveau de ces gènes entre les deux conditions. Par exemple, des gènes comme PTPRG (récepteur protéine kinase) gagnent de la significativité après correction. Ce dernier passe d'une valeur d'environ 25 à 29 au niveau de -log10(p-value). De même, le gène SPRED1 (suppresseur de tumeur) gagne en significativité également, tandis que certains gènes comme DSC2 perdent de la significativité.

Ces observations signifient que la correction pour les pourcentages cellulaires influence l'expression de certains gènes spécifiques, mais globalement n'affecte pas drastiquement la majorité des résultats.

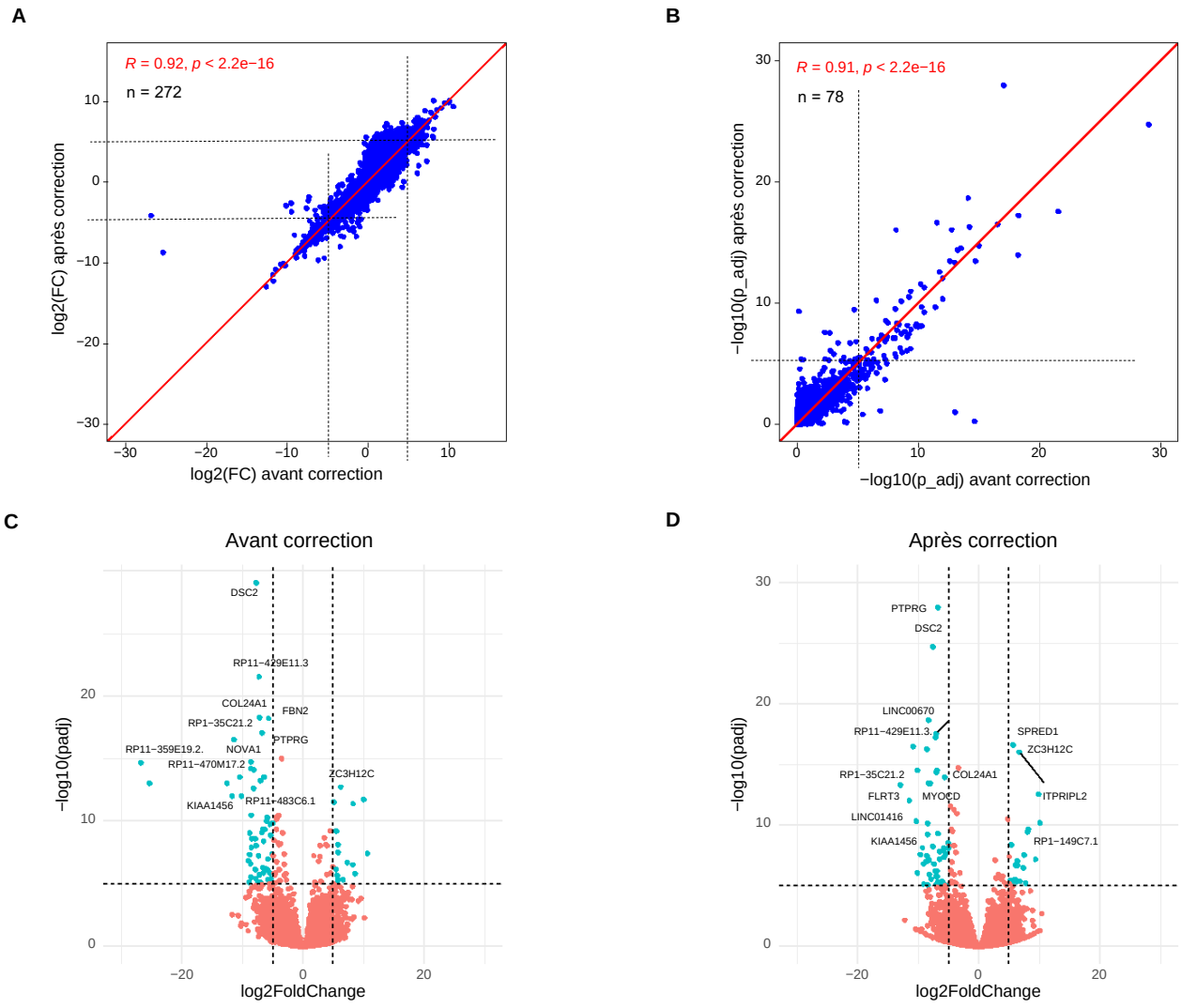


Figure 10. – Analyse d'expression différentielle des gènes avant et après correction pour la proportion des types cellulaires estimés par déconvolution. Corrélations entre (A) log2FC et (B) p-value, avant et après correction. Le coefficient de corrélation R est calculé avec la méthode de Pearson et la droite représente le cas idéal qui suit la fonction x=y. Graphiques volcan des gènes différentiellement exprimés (en vert) (C) avant et (D) après correction. Les gènes sont considérés différentiellement exprimés si log2FC > 5 ou log2FC < -5 et que -log10(p-adj) > 5. Les valeurs p sont donc transformées en -log10(p) et les FoldChange en log2. Le top 20 des gènes les plus significatifs est affiché sur le graphique (en évitant les chevauchements lorsque possible). Les pointillés sur les graphiques correspondent aux seuils choisis, et n au nombre de gènes respectant nos conditions.

5. Prédiction de l'impact fonctionnel des gènes différentiellement exprimés

Pour connaître les fonctions des gènes qui sont altérés lors de la correction pour les proportions de types cellulaires, une analyse fonctionnelle a été réalisée à l'aide des 100 gènes les plus différentiellement exprimés avant puis après correction. Ainsi, les listes de ces gènes ont été fournies à l'outil web Metascape pour une analyse à travers la base de données d'ontologie de gènes (Gene Ontology, GO) (74). La Figure 11 ci-dessous répertorie les fonctions les plus significatives avant correction (A), puis la même chose après correction (B), et la corrélation entre les termes GO avant et après (C).

On note que le terme GO :0007169, impliqué dans la voie de signalisation des protéines tyrosine kinase transmembranaire, apparaît parmi les termes GO les plus significatifs après correction (Figure 11B). On retrouve également certains termes GO qui gagnent en significativité entre les graphiques A et B, par exemple le terme GO :0098742 (impliqué dans l'adhésion entre les cellules) (Figure 11A,B).

Le terme GO :0008543 (impliqué dans la voie des récepteurs de facteurs de croissance des fibroblastes), n'est pas représenté sur les deux premiers graphiques (Figure 11A,B), mais il gagne en significativité après correction (Figure 11C). Dans d'autres cas, il y a une apparition de termes après correction, comme par exemple le terme GO :0071900, impliqué dans la régulation de l'activité de la sérine/thréonine protéine kinase, et qui n'est pas visible sur le graphique avant correction.

D'autres termes GO ont plus tendance à perdre de la significativité après correction, notamment le terme GO :0050890 (impliqué dans la cognition), le terme GO :0030512 ou encore le terme GO :0090101 (relié à la régulation négative de la voie de signalisation du récepteur transmembranaire de la sérine/thréonine protéine kinase).

Ces observations laissent suggérer que la correction a un certain impact sur quelques gènes et leurs fonctions, et qu'il est nécessaire d'analyser le rôle de ces derniers plus en détail.

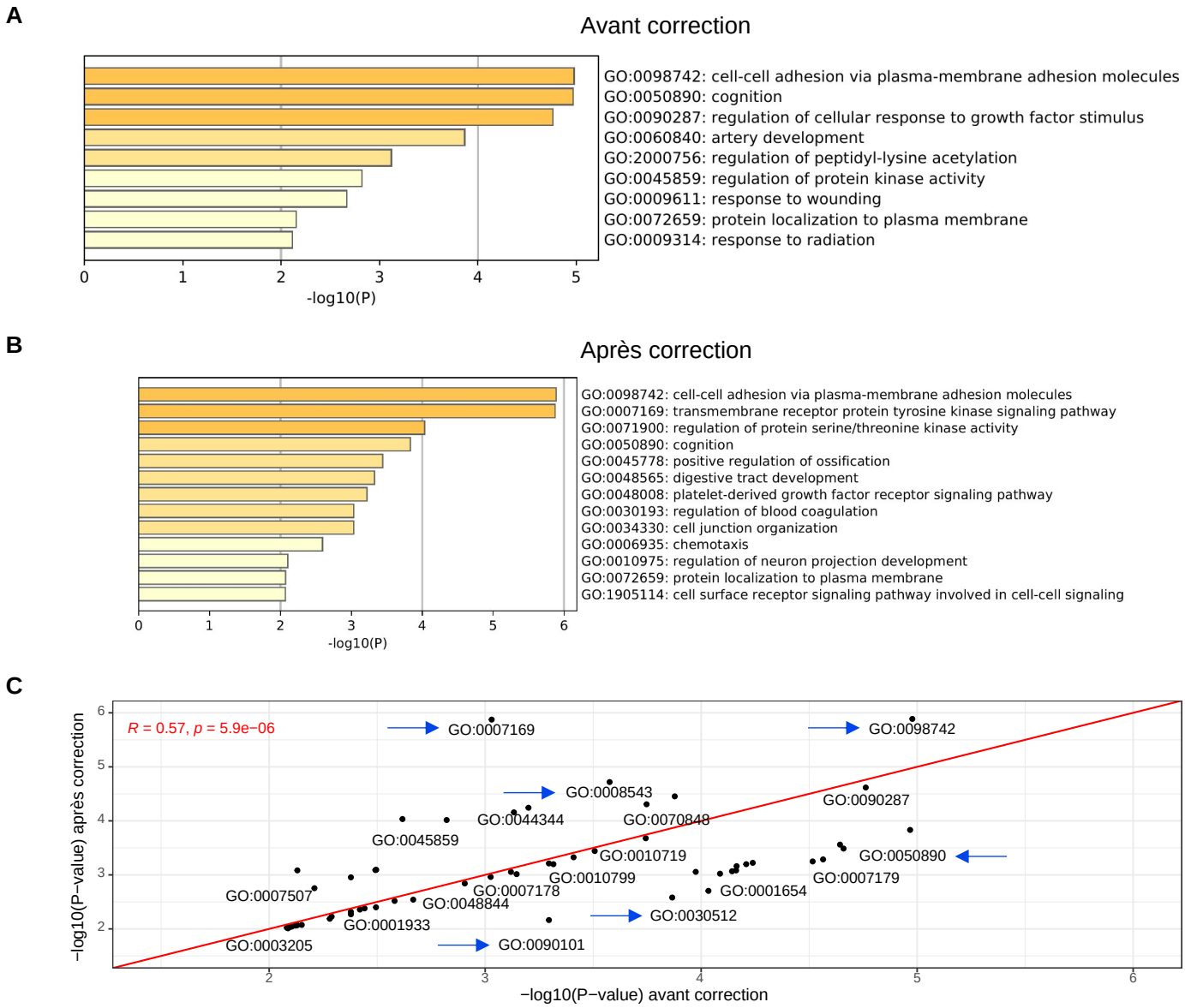


Figure 11. – Analyse fonctionnelle des gènes les plus différentiellement exprimés avant et après correction pour les proportions cellulaires. Les termes GO les plus significatifs obtenus avec Metascape pour les gènes les plus différentiellement exprimés (A) avant et (B) après correction. (C) Corrélation entre la significativité (p-value) des termes GO principaux avant et après correction. Les valeurs p sont transformées en -log10(p). Le coefficient de corrélation R est calculé avec la méthode de Pearson, et la droite représente le cas idéal qui suit la fonction x=y. Les flèches bleues sont les termes GO choisis pour être décrits plus en détail du fait qu'ils ont gagné ou perdu de la significativité.

- Chapitre 5 -

Détermination du profil immunitaire dans les tumeurs solides

1. Quantification de l'infiltration immunitaire

Les interactions avec le microenvironnement peuvent avoir des incidences sur la progression tumorale (28, 32). J'ai donc essayé d'évaluer le répertoire de cellules immunitaires dans trois tumeurs pédiatriques solides : Neuroblastomes, Ostéosarcomes et tumeurs de Wilms. Pour ce faire, j'ai déterminé deux scores pour les cellules T en utilisant des données publiques pour 282 échantillons provenant du consortium TARGET (https://ocg.cancer.gov/programs/target) : 146 Neuroblastomes, 86 Ostéosarcomes et 50 tumeurs de Wilms.

En utilisant le mode « absolu » de Cibersort, j'ai généré des scores quantifiant chaque type cellulaire présent dans les échantillons (59). Un second score a été généré en utilisant la moyenne d'expression de gènes marqueurs spécifiques aux cellules T (comme dans le Chapitre 3) (75).

À la figure 12, on observe la distribution, la comparaison entre les trois types de tumeurs, ainsi que la corrélation entre ces deux scores. On observe d'abord l'étendue de la distribution des scores (Figure 12A,B). Les scores de Cibersort oscillent entre 0 et 2 (Figure 12A), tandis que le score d'expression T va jusqu'à 4 (Figure 12B), ce qui n'est pas inattendu car la méthode de calcul de ces deux scores diffère. Plus de 80% des scores de Cibersort ne dépasse pas 1 (Figure 12A). Pour les scores basés sur la moyenne d'expression des cellules T, ceux-ci sont en dessous de 2 pour les ostéosarcomes et tumeurs de Wilms. Les deux scores sont plus élevés pour les neuroblastomes (Figure 12A,B).

Les taux d'infiltration calculés avec Cibersort entre les types de tumeurs semblent visuellement assez proches (Figure 12C). Cependant, les valeurs-p sont tout de même significatives, notamment entre les ostéosarcomes et les deux autres types de tumeurs, soulignant ainsi une

différence d'infiltration. Les différences d'infiltration entre les trois types de tumeurs sont davantage visibles avec le score d'expression T, notamment avec des valeurs-p beaucoup plus significatives, de l'ordre de 10^{-10} entre les trois types (Figure 12D).

Malgré les différences, la corrélation positive entre les deux scores est bonne avec une valeur de 0,8 (Figure 12E). On observe également une bonne corrélation lorsque l'on considère séparément les types tumoraux : neuroblastomes avec un coefficient de 0,91 ; ostéosarcomes à 0,78 ; et tumeurs de Wilms à 0,75 (Figure 12F).

En résumé, nos résultats montrent une bonne corrélation entre les deux scores étudiés et que le taux d'infiltration peut varier en fonction du type de tumeurs.

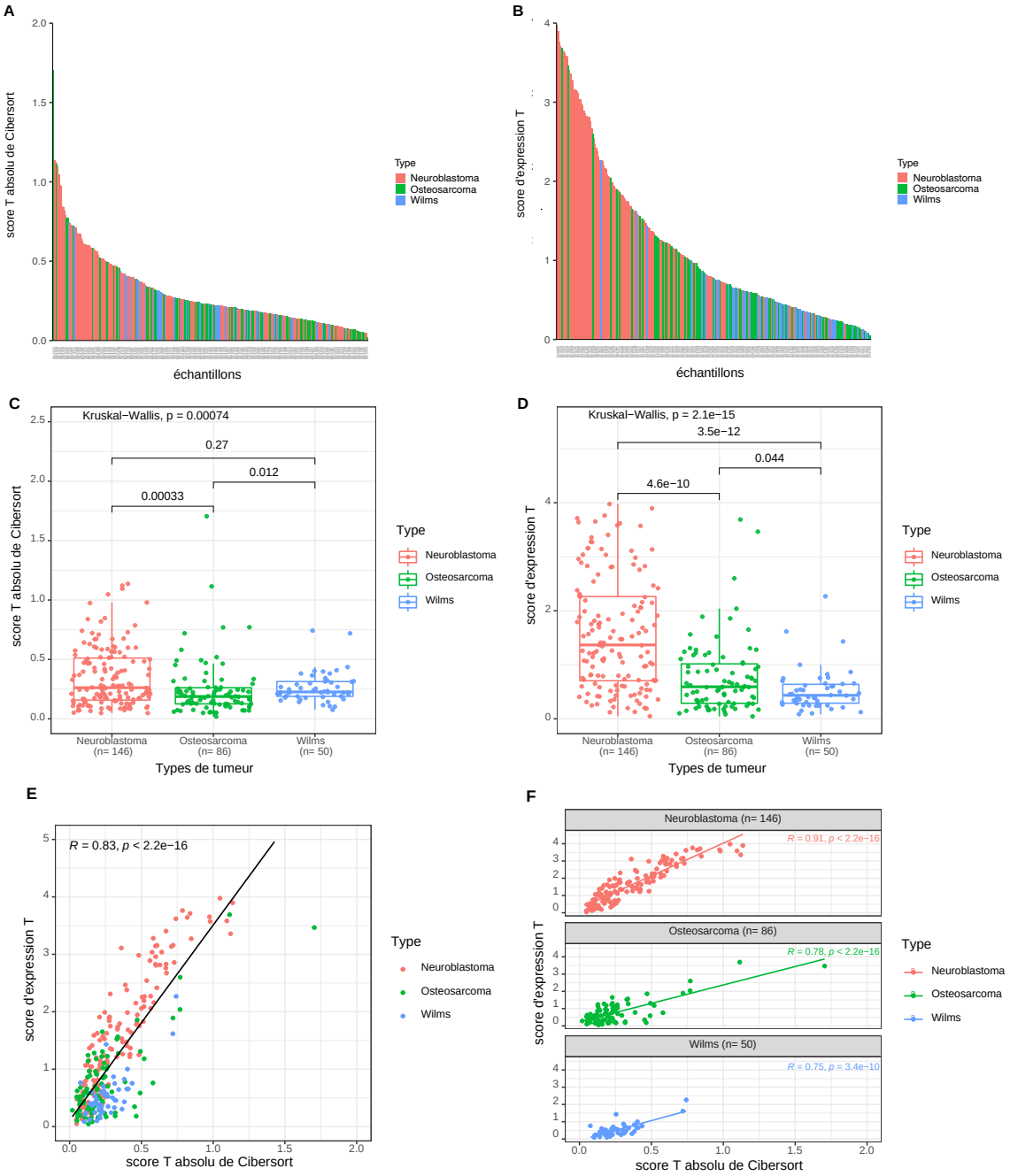


Figure 12. – Distributions et corrélations des scores d'infiltration des lymphocytes T. Distribution du score (A) des lymphocytes T calculé par Cibersort en mode « absolu » et (B) du score en log2 d'expression des cellules T avec les gènes marqueurs. Représentation des scores en fonction des trois types de tumeurs pour (C) le score T calculé par Cibersort et (D) le score d'expression T. La valeur p calculée par la méthode Kruskal-Wallis compare les trois types de tumeurs, tandis que la méthode de Wilcoxon calcule les valeurs p entre deux paramètres. Corrélation entre les deux scores d'infiltration T (E) pour toutes les tumeurs et (F) de manière séparée en fonction du type de tumeurs. Les droites suivent une fonction de régression linéaire simple et les valeurs R représentent le coefficient de corrélation selon la méthode de Pearson, avec la p-value associée.

1.1 Corrélation entre les scores d'infiltration et les données cliniques

J'ai mesuré les fluctuations des scores dépendamment de différents paramètres et conditions cliniques. Les scores obtenus avec Cibersort ont été comparés dans les trois types de tumeurs à travers des paramètres conditionnels tels que le sexe du patient, le pronostic, les évènements, mais aussi des paramètres temporels tels que le temps sans évènements, l'âge au diagnostic ou encore le temps de survie (Figure 13).

Je n'ai observé aucune différence significative entre les paramètres cliniques et le score d'infiltration de Cibersort (Figure 13A-F), (toutes les valeurs-p sont supérieures à 0,05 et les coefficients de corrélation proches de 0).

Ces résultats suggèrent alors que le score d'infiltration des lymphocytes T calculé avec Cibersort n'a pas d'impact sur les critères cliniques étudiés

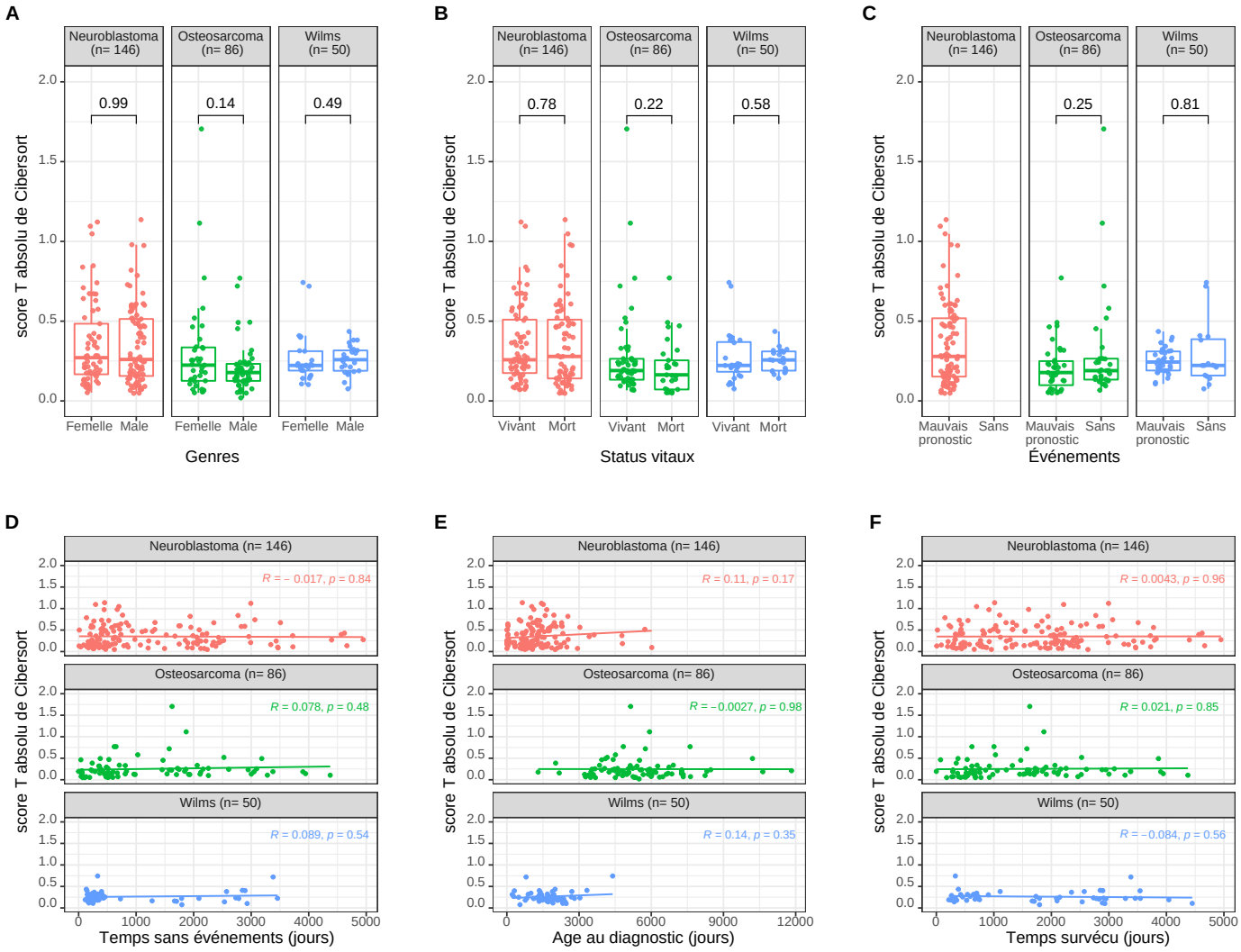


Figure 13. – Corrélations du score d'infiltration des cellules T de Cibersort avec différents paramètres cliniques chez les patients TARGET (n=282). Variations des scores par types de tumeurs pour différentes variables cliniques en abscisses : (A) le sexe, (B) le statut vital, (C) les évènements, (D) le temps sans évènements, (E) l'âge au diagnostic, et (F) le temps survécu. Pas de score disponible pour les neuroblastomes lorsqu'il n'y a pas d'évènements dans les données. Les valeurs p entre deux conditions sur les trois premiers graphiques sont calculées avec la méthode de Wilcoxon. Les droites sur les trois derniers graphiques suivent une fonction de régression linéaire simple et les valeurs R représentent le coefficient de corrélation selon la méthode de Pearson, avec la p-value associée. Le nombre n entre parenthèses représente le nombre d'échantillon pour chaque type de tumeur.

Les mêmes analyses ont été effectuées pour le score d'expression des cellules T produit à partir de gènes marqueurs (Figure 14) (75). Ce score a été corrélé pour les trois types de tumeurs avec les mêmes paramètres que le score de Cibersort.

Pour les ostéosarcomes, on observe une corrélation avec la survie (valeur p = 0,0054), avec une médiane du score d'infiltration en lymphocytes T plus élevée pour les personnes encore vivantes (Figure 14B). On observe également une corrélation avec la présence d'un événement chez les ostéosarcomes, où les patients qui ne présentent pas d'évènements ont un score T plus haut. (Figure 14C).

Aucune autre corrélation n'a été observée (Figure 14D-F).

Ces résultats suggèrent que le score d'expression T a une valeur prédictive dans certaines tumeurs, notamment pour la survie et la présence d'événements. Ceci souligne donc un potentiel rôle de l'infiltration des lymphocytes T sur le pronostic clinique.

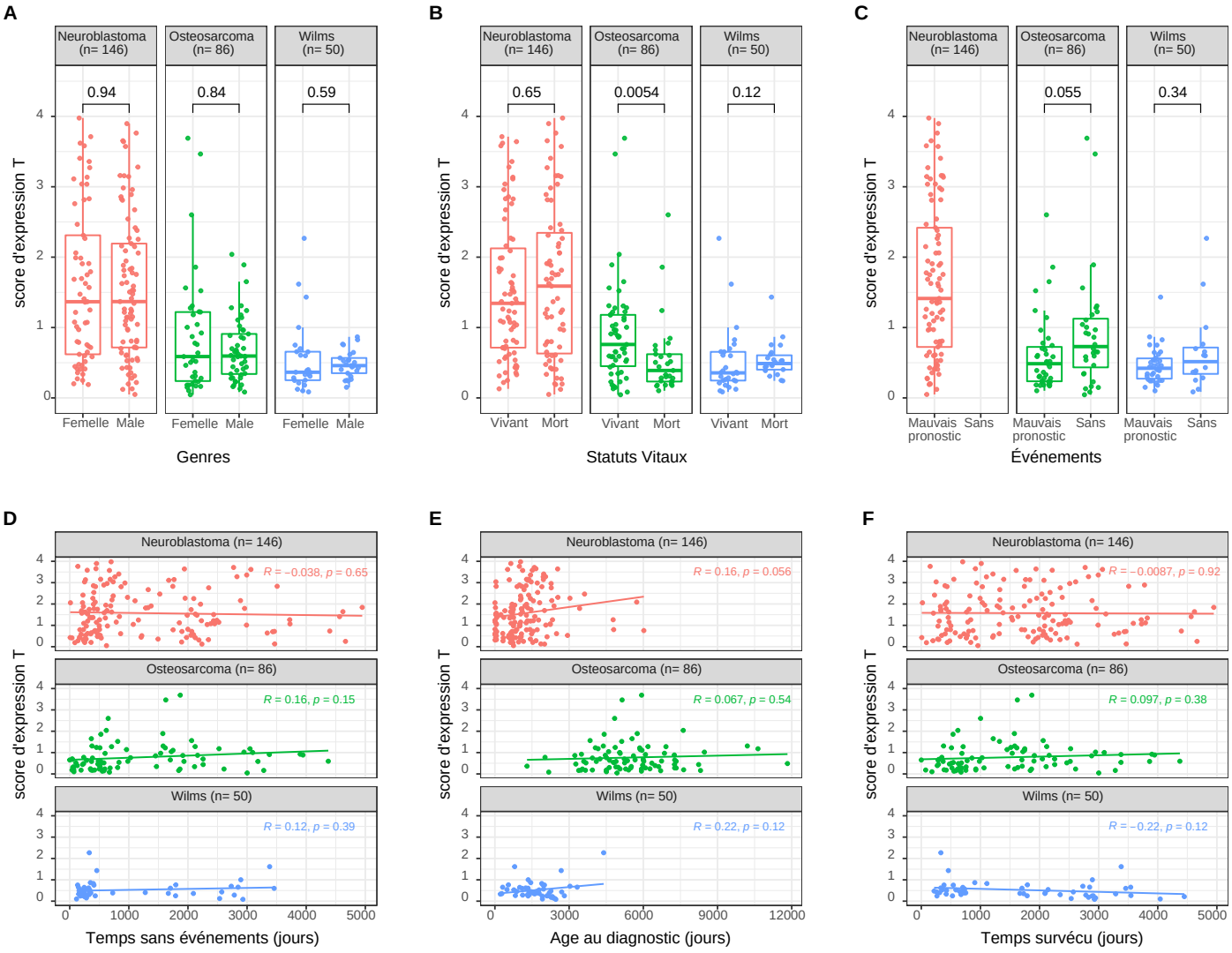


Figure 14. – Corrélation du score d'infiltration des cellules T des gènes marqueurs avec différents paramètres cliniques chez les patients TARGET (n=282). Variations des scores d'expression par types de tumeurs pour différentes variables cliniques en abscisses : (A) le sexe, (B) le statut vital, (C) les évènements, (D) le temps sans évènements, (E) l'âge au diagnostic, et (F) le temps survécu. Pas de score disponible pour les neuroblastomes lorsqu'il n'y a pas d'évènements dans nos données. Les valeurs p entre deux conditions sur les trois premiers graphiques sont calculées avec la méthode de Wilcoxon. Les droites sur les trois derniers graphiques suivent une fonction de régression linéaire simple et les valeurs R représentent le coefficient de corrélation selon la méthode de Pearson, avec la p-value associée. Le nombre n entre parenthèses représente le nombre d'échantillon pour chaque type de tumeur.

- Chapitre 6 -

Discussion

Dans le cadre de mon projet, j'ai utilisé des outils bio-informatiques pour obtenir une meilleure caractérisation des cancers pédiatriques, et ce, à d'éventuelles fins thérapeutiques. Le fil conducteur de mon projet est basé sur l'utilisation de ces outils pour la quantification des types cellulaires, l'analyse de la pureté tumorale, la quantification de l'infiltrat immunitaire, et leurs divers rôles dans l'évolution des tumeurs.

1. Données synthétiques

1.1 Déconvolution

La première étape réalisée consistait en l'exécution des outils de déconvolution sur des données synthétiques pour tester leur efficacité. L'observation globale faite à partir de ces résultats réside dans le fait que les proportions estimées ne sont pas toujours exactement celles qui sont attendues. En effet, plusieurs proportions cellulaires diffèrent, en étant sous-estimées ou bien surestimées dans certains de nos échantillons. En se focalisant sur les résultats obtenus avec DeconRNAseq, on remarque un haut taux de CLP (progéniteur lymphoïde), surestimé notamment lorsque les lymphocytes B sont présents dans l'échantillon. Cette observation est potentiellement due à la ressemblance de certaines signatures entre les lymphocytes et leurs cellules progénitrices (59). Ainsi, certaines signatures des lymphocytes B seraient davantage similaires aux signatures références des cellules lymphoïdes progénitrices (CLP) dans la matrice de DeconRNAseq. Cette ressemblance fausse alors la reconnaissance des lymphocytes dans nos échantillons synthétiques qui proviennent de lignées cellulaires correspondant à des leucémies à un stade différent de maturation. Newman avait déjà fait ce genre d'observations en spécifiant que les résultats de déconvolution pouvaient être négativement affectés lorsque plusieurs types cellulaires étaient hautement corrélés (59,

77). Il en avait déduit que de plus hautes proportions étaient estimées pour des types cellulaires dont les profils se ressemblaient le plus (77). Dans notre cas, il faudrait peut-être additionner les proportions des lymphocytes B et des CLP pour avoir des estimations un peu plus vraisemblables. Par ailleurs, il apparaît aussi que lorsque les proportions des types cellulaires majoritaires (B et T) sont plus faibles, l'outil va introduire des estimations plus hautes pour d'autres types cellulaires non attendus dans le mélange.

Avec Cibersort, certains échantillons présentent aussi des surestimations, mais cette fois au niveau des lymphocytes T, et dans une moindre mesure, au niveau des monocytes. Les lymphocytes B sont légèrement sous-estimés. Des études ont mentionné que même si Cibersort possède un biais d'estimation plus faible que certaines autres approches, il a tendance à surestimer ou sous-estimer systématiquement certains types cellulaires (59). Puisque nos données synthétiques ne proviennent pas de tissus tumoraux, les signatures des types cellulaires sont aussi probablement différentes de celles utilisées par l'outil.

Des surestimations beaucoup plus extrêmes au niveau des lymphocytes T sont faites avec Quantiseq. Certaines surestimations sont également visibles au niveau des lymphocytes B avec cet outil, notamment lorsque les cellules T ne sont pas présentes, ou sont en faible quantité dans l'échantillon. Cependant, cet outil estime tout de même la plupart des échantillons avec un type cellulaire majoritaire, et introduit très peu d'autres types cellulaires non désirés. Ceci est possiblement dû au fait que cet outil possède une matrice signature moins diversifiée en types cellulaires (52, 62). Les types cellulaires présents sont donc plus généraux et diminuent ainsi le risque d'un biais lié à une forte ressemblance entre deux types différents. Cependant, ceci favorise potentiellement la surexpression des autres types cellulaires majoritaires.

En résumé, nos résultats ont montré que la déconvolution est variable et dépendante de l'outil utilisé, même à partir d'un jeu de données synthétiques. Ces résultats découlent notamment de la méthode de calcul, mais surtout de la matrice signature utilisée par l'outil, qui peut être composée de plusieurs types cellulaires, à des stades différents de maturation comparativement à nos données, et qui peuvent alors fausser les estimations (53). Comme exemple, les leucémies T sont principalement caractérisées par des précurseurs lymphoïdes T étant à un stade immature (pré-T), ainsi l'utilisation de types

cellulaires primaires T CD4+ ou CD8+ peut être sous-optimal pour déconvoluer les patients LLA T (78).

L'idéal serait; si tous les types cellulaires et leurs stades de maturation présents dans nos données sont connus; de créer une matrice de référence spécifique avec des signatures présentes dans nos données. Cette matrice peut être créée mais elle doit être adaptée aux données qui sont analysées, il faut donc déterminer les gènes marqueurs spécifiques avant de réaliser l'analyse, ce qui rend la tâche plus complexe.

1.2 Score d'expression des cellules T

Afin de vérifier si les surestimations observées proviennent de la méthode utilisée par ces outils, ou bien si d'autres facteurs antérieurs sont impliqués, un score pour les lymphocytes T a été généré à partir de l'expression de marqueurs de surface T (75). J'ai ainsi démontré que le score T est positivement corrélé avec les estimations en lymphocytes T des outils de déconvolution. Ce score est fortement corrélé avec les estimations de Cibersort. Ces résultats indiquent qu'une partie de la surestimation observée est déjà présente dans la matrice d'expression donnée à l'outil. Cependant, il faut noter que Cibersort et DeconRNAseq partagent la même matrice d'expression, et DeconRNAseq présente tout de même moins de surestimations. Ceci sous-entend que d'autres éléments spécifiques à l'outil, tels que la matrice signature ou la méthode de calcul, viennent influencer ces estimations (53). Même si Quantiseq effectue tout le pipeline de son côté, la matrice d'expression créée par l'outil présente aussi certaines surexpressions avant la déconvolution. Ainsi, ces surestimations peuvent avoir plusieurs origines, elles peuvent provenir d'une surexpression déjà présente dans la matrice d'expression fournie à l'outil, et dépendamment des signatures et méthodes utilisées, ces surexpressions peuvent persister ou disparaître une fois l'outil exécuté.

En résumé, les paramètres généralement pris en compte par l'outil ne sont pas forcément optimaux, d'autant plus que la méthode avec laquelle nos échantillons sont créés ne l'est peut-être pas également. Il aurait peut-être fallu faire des échantillonnages indépendants et créer à chaque fois plusieurs échantillons synthétiques avec les mêmes proportions pour quantifier la variabilité. Ces paramètres seraient donc à revoir afin d'optimiser

l'utilisation de ces outils. Comme les trois outils se valent, j'ai décidé de tous les utiliser pour l'étude des données leucémiques.

2. Leucémies lymphoblastiques aigues (LLA)

2.1 Déconvolution

J'ai démontré que DeconRNAseq estime une proportion majoritaire de lymphocytes B, ce qui est attendu puisque la cohorte est composée majoritairement de type B. Cependant, dans certains cas, des échantillons de type T sont présents avec une proportion de lymphocytes B estimée à plus de 80%. On peut expliquer ces résultats de plusieurs façons. Selon la matrice de référence utilisée par l'outil, il est possible que les types cellulaires dans le mélange tumoral se rapprochent plus des signatures de précurseurs lymphoïdes que des cellules T. Par ailleurs, il est possible de voir que des échantillons de type B n'ont pas une proportion de lymphocytes B majoritaire, mais plutôt une combinaison de différents types cellulaires à proportions égales. Cette hétérogénéité devrait être analysée plus en détail puisque cela suppose une pureté tumorale faible, et donc une potentielle complication dans les propositions de thérapies et analyses subséquentes (49, 78).

J'ai également démontré que Quantiseq est assez semblable à DeconRNAseq au niveau des résultats. Les estimations montrent une majorité de lymphocytes B. Cependant, Quantiseq affiche également des échantillons de type T avec de hautes proportions estimées en lymphocytes B, mais l'inverse est également observé.

Nos résultats suggèrent un autre biais lié à la déconvolution qui serait associé au contenu en ARNm (52). En effet, il est possible que les niveaux d'expression de certains gènes puissent fausser l'estimation des types cellulaires ayant un plus haut taux d'ARN (52). Par contre, comme les outils que j'ai testés réalisent des normalisations, ce biais ne devrait normalement pas être important dans nos résultats (52).

Cibersort quant à lui affiche peu d'échantillons avec plus de 70% de lymphocytes B. Il présente également plusieurs échantillons assez hétérogènes, et estime des échantillons avec de grandes quantités de cellules T. Par contre, Cibersort est le meilleur des trois

outils pour séparer distinctement les types B des types T. Ceci est probablement la conséquence de la matrice signature composée de 22 types cellulaires à différents stades de maturation de cellules hématopoïétiques (59). Comme mentionné précédemment, les signatures présentes dans nos données ne sont pas forcément les mêmes que celles utilisées lors de la construction de la matrice de référence de l'outil (67). De plus, si certains de ces sous types ne sont pas présents dans les données ARN-seq, il est possible que les résultats de déconvolution soient affectés négativement (67).

En résumé, les outils de déconvolution présentent certaines limites. Ces limites sont notamment au niveau de la disponibilité des types cellulaires de référence et de la corrélation de leur transcriptome, ou au niveau de la pureté tumorale des tissus à déconvoluer (79).

Beaucoup d'outils utilisent des matrices de référence développées et validées avec des micropuces à ADN (44). Il manque encore aujourd'hui des validations faites avec les données ARN-seq (44). Comme on a pu le voir, ces matrices de référence jouent un rôle considérable sur les résultats. Des études ont montré que le fait d'enlever un type cellulaire, ou de ne pas avoir des types cellulaires assez représentatifs du mélange tumoral dans celle-ci, empire en général les proportions des autres types cellulaires estimées par l'outil, ainsi que la performance de la méthode (31).

Une autre limitation connue vis à vis des outils qui prédisent les fractions cellulaires vient du fait que lorsque certains types cellulaires sont présents mais à très faible taux, les prédictions connaissent plus d'erreurs d'estimation (61). De plus, les estimations des types cellulaires peuvent être altérées lorsque leurs profils d'expression sont trop ressemblants (61). Dans ce cas, les approches d'enrichissement de gènes sont peut-être préférables, ou bien l'utilisation de méthodes combinant les deux approches.

2.2 Corrélation clinique

Afin d'étudier la cohérence de ces outils au niveau des estimations faites, ces dernières ont été comparées avec les estimations de blastes leucémiques calculées en clinique.

Pour les lymphocytes B, malgré le fait qu'aucun outil ne fournit une corrélation parfaite, j'ai déterminé la capacité d'estimer le plus d'échantillons quasiment purs/homogènes pour le bon type cellulaire, tout en étant en accord avec les estimations cliniques. Ce facteur de pureté est connu pour faciliter ensuite les critères de décision des thérapies et les pronostics (29, 49). En effet, une hétérogénéité tumorale pourrait souligner la présence de sous-clones pouvant avoir un impact non désiré sur le traitement (48). Aussi, en connaissant le type cellulaire majoritaire présent dans la tumeur, il est possible d'adapter les stratégies thérapeutiques en fonction de celui-ci, certains étant plus bénéfiques que d'autres (49).

J'ai ainsi démontré que DeconRNAseq est l'outil le plus performant pour l'estimation des lymphocytes B, même si ce dernier ne montre pas de corrélation apparente. Quantiseq se trouve juste après, avec un bon nombre d'échantillons entre 60 et 90% de lymphocytes B estimés. Cibersort quant à lui, comme remarqué précédemment, n'affiche jamais plus de 75% de lymphocytes B estimés, malgré le fait qu'il présente un meilleur coefficient de corrélation que DeconRNAseq par exemple. Ceci peut être encore une fois être dû à une limite retrouvée dans plusieurs méthodes de déconvolution basées sur la signature de gènes, soit la fidélité des profils de référence (59). En effet, des dérégulations induites par la maladie et des modifications au niveau des interactions entre les cellules peuvent parfois provoquer des différences considérables au niveaux des profils d'expression, et ainsi faire varier les estimations des outils en conséquence (59).

La tendance est différente dans le cas des leucémies de type T où DeconRNAseq performe moins bien. Toutes les valeurs sont estimées en dessous de 50% de lymphocytes T, alors que les données cliniques sont majoritairement au-dessus de 75% de lymphocytes T. Quantiseq reste lui à peu près constant dans ses estimations. Cibersort est le meilleur outil pour estimer les lymphocytes T. Ceci confirme nos résultats obtenus avec les données synthétiques : DeconRNAseq est un outil plus cohérent pour les estimations de lymphocytes B et Cibersort performe mieux avec les lymphocytes T. Étant

donné le petit nombre de leucémie de type T, il faudrait valider cette étude avec un plus grand échantillonnage.

Les outils de déconvolution ayant chacun leur propre matrice signature, les différences observées pourraient dépendre de l'expression de certains marqueurs spécifiques (31). Un autre aspect à prendre en compte est le fait que les analyses réalisées en clinique ne sont pas faites sur les mêmes prélèvements que ceux fournis pour la recherche. En effet, les échantillons de notre analyse proviennent de ponctions de moelles subséquentes, donc de moins bonne qualité en termes de cellularité. Ceci pourrait expliquer, en partie, la variabilité observée car l'échantillon reçu pour le séquençage ne possède pas forcément le même contenu en blastes que celui analysé en laboratoire clinique. Dans le cadre d'un transfert ARN-seq vers un laboratoire clinique, il serait important de quantifier cliniquement le contenu en blastes du prélèvement, et d'utiliser ce dernier pour l'analyse moléculaire.

En résumé, pour augmenter l'efficacité des analyses de données ARN-seq dans les tumeurs, des paramètres spécifiques doivent être pris en compte par les méthodes de déconvolution. Ces paramètres doivent pouvoir varier en fonction du type de cancer, du tissu concerné, et éventuellement en fonction des signatures de gènes spécifiques pour lesquelles l'expression change dans un certain type de tumeur.

2.3 Score d'expression des cellules T

Pour déterminer si les observations précédentes sont liées à la méthode de l'outil utilisé ou à la matrice d'expression fournie, j'ai calculé les scores d'expression des cellules T à partir de l'expression des marqueurs de surface (75).

La majorité des estimations des cellules T par DeconRNAseq est proche de 0 alors que le score d'expression lui varie. Ceci indique qu'il y a effectivement un facteur dans la méthode de cet outil qui entre en jeu dans les estimations des lymphocytes T. Ceci pourrait provenir des signatures de la matrice de référence utilisée, de l'équation, ou de la différence de maturation des cellules entre les échantillons et la signature. Tous ces facteurs peuvent mener à des estimations biaisées (52, 59).

Les scores d'expression et les estimations en cellules T avec Cibersort présentent une meilleure corrélation, alors que la matrice d'expression utilisée par l'outil est identique à celle fournie à DeconRNAseq. Ceci renforce alors l'hypothèse d'un facteur dépendant de l'outil en question, plutôt qu'un facteur relatif à la matrice des mélanges tumoraux.

Le coefficient de corrélation obtenu avec Quantiseq est le même que celui obtenu avec Cibersort, malgré le fait que Quantiseq utilise une matrice d'expression construite indépendamment des autres outils. Ces deux dernières corrélations permettent de souligner la cohérence entre les estimations et le choix des marqueurs pour les lymphocytes T.

2.4 Analyse d'expression différentielle

Une dimension additionnelle a voulu être explorée avec les données leucémiques. En introduisant les pourcentages des types cellulaires estimés par les outils, et donc en réduisant les variations d'expressions liés à ces derniers, il est possible de mettre en avant l'altération de gènes et fonctions plus spécifiques aux sous types de leucémie. Ce type d'analyse permettrait alors d'exposer certaines cibles et certains biomarqueurs potentiels.

Pour ce faire, une analyse différentielle des gènes a été réalisée avec et sans correction pour les types cellulaires sur un groupe de 10 échantillons de deux sous-types majoritaires différents : HHD (n=5) et ETV6-RUNX1 (n=5).

Ces analyses ont permis d'identifier les gènes les plus différentiellement exprimés pour les deux analyses. J'ai ensuite sélectionné certains gènes les plus significatifs, notamment PTPRG ou SPRED1, qui ont gagné en significativité.

PTPRG est un récepteur de protéine tyrosine kinase phosphatase de type G (80). Les protéines phosphatases sont des molécules connues pour avoir un rôle régulateur dans plusieurs processus cellulaires, certains d'entre eux étant même associés au développement des cancers, comme la transformation oncogénique et la croissance cellulaire (80). Un simple déséquilibre au niveau des activités de ces phosphatases peut rapidement mener à la transformation des cellules normales en cellules malignes (80). PTPRG a été identifié comme un gène candidat suppresseur de tumeur (80). Il inhibe la

signalisation de la kinase « Akt » et provoque ainsi la suppression des métastases et de la tumorigenèse (80).

SPRED1 fonctionne comme un suppresseur de tumeur (81). Son inactivation augmente la prolifération cellulaire et engendre une résistance aux traitements inhibant l'activité des tyrosine kinase (82). SPRED1 interagit avec plusieurs facteurs, et agit alors dans plusieurs processus tels que la prolifération ou la différenciation des cellules hématopoïétiques (82). Ces facteurs (comme NF1 ou VEGFR) vont promouvoir la leucémogenèse en supprimant certaines voies de signalisation (Ras-MAPK principalement) (82). Des mutations dans cette voie sont associées à des rechutes précoces et une chimiorésistance dans les LLA (82). La surexpression de SPRED1 a tendance à réduire la prolifération et induire l'apoptose, tandis que son inactivation peut à contrario engendrer cette prolifération et réduire l'apoptose de certaines cellules (82).

D'autre part, il est également possible de retrouver certains gènes qui perdent de la significativité, comme FBN2 dans les gènes sous-exprimés, ou LINC00887 dans ceux surexprimés.

FBN2 est impliqué dans la création de la fibrilline 2, prenant place au niveau de la matrice extracellulaire et se liant à d'autres molécules pour assurer le maintien des ligaments, vaisseaux, nerfs et autres composants durant le développement (83). Ces microfibriles se lient à des facteurs de croissance comme TGF-B pour les garder inactifs, puisqu'une fois relâchés, ces derniers affectent la croissance et la réparation des tissus dans le corps (83).

LINC00887 quant à lui, est un long ARN non codant, souvent surexprimé dans les carcinomes (84). Des hauts taux de cet ARN sont souvent associés à un mauvais pronostic et une augmentation de la prolifération dans ces carcinomes (85).

Ces gènes perdent potentiellement de la significativité du fait qu'ils sont plus associés à des mécanismes favorisant la progression tumorale, et ne sont donc pas des biomarqueurs à considérer.

Par ailleurs, d'autres gènes qui perdent de la significativité sont peut-être davantage impliqués dans des processus qui ne sont pas directement reliés à l'émergence de

cancer. Certains autres gènes perdent de la significativité car ils sont certainement exprimés dans un type cellulaire spécifique, et ne sont pas différentiellement exprimés dans les sous-types. Ces observations laissent penser que ces gènes seront donc moins caractérisés comme gènes cibles.

Il faut garder à l'esprit que mes résultats ont été obtenus sur un petit nombre d'échantillons, pour seulement deux sous-types, et avec des estimations faites par DeconRNAseq. Ils ne représentent donc pas forcément l'ensemble des données leucémiques.

Il serait intéressant de valider ces résultats sur un nombre plus grand d'échantillon, avec d'autres sous-types, et en introduisant les proportions estimées d'un autre outil. Il peut aussi être intéressant d'introduire les pourcentages de blastes cliniques dans le design, puis de comparer les résultats des gènes obtenus.

2.5 Impact fonctionnel des gènes différentiellement exprimés

Pour avoir plus d'informations sur les fonctions et les voies biologiques altérées, j'ai réalisé une analyse fonctionnelle en me concentrant sur les fonctions qui ont gagné ou perdu en significativité.

Le terme GO :0007169, significatif après correction, est impliqué dans la voie de signalisation des protéines tyrosine kinases transmembranaires, connue pour être elle-même impliquée dans la différenciation cellulaire, la croissance ou encore l'apoptose (86). C'est une voie très présente dans les cancers et souvent ciblée dans les thérapies (86). Le terme GO :0071900, significatif seulement après correction, est en lien avec les processus régulant l'activité de la protéine sérine/thréonine kinase, dont le récepteur joue un rôle vital dans les cancers, puisque certaines d'entre elles sont cruciales dans la régulation et le développement cellulaire (87).

Le terme GO :0098742, significatif après correction, est associé à l'adhésion entre les cellules via la membrane plasmique, soit des molécules qui ont déjà été reportées comme ayant un lien avec les cancers, tant à leur progression qu'à leur inhibition (88). Elles sont également capables d'activer des voies de signalisation de facteurs de croissance, soit

une des caractéristiques proéminentes dans les cancers (88). D'autres exemples, comme le terme GO :008543, ont connu une augmentation plus légère en termes de significativité, mais ce dernier est aussi relié à des processus du développement tumoral puisqu'il est associé au récepteur du facteur de croissance des fibroblastes (FGFR), jouant aussi un rôle dans la croissance cellulaire et la différenciation (89).

Certains termes GO qui ont perdu de la significativité sont également intéressants à analyser pour comprendre les fonctions dans lesquelles ils sont impliqués. Le terme GO :0050890 est en lien avec la cognition, et n'est donc pas directement relié aux cancers. D'autres comme le terme GO :0030512, ou encore le terme GO :0090101, sont quant à eux engagés dans la régulation négative de voies de signalisation de certains facteurs de transcription ou de protéines kinases.

En résumé, les gains de significativité sont plutôt associés à des gènes impliqués dans les processus retrouvés dans le développement des cancers tels que la progression, la prolifération ou encore la croissance cellulaire, tandis que ceux qui perdent de la significativité sont souvent liés à des processus qui empêchent, réduisent ou inhibent les voies de signalisation, ou bien qui ne sont pas reliés aux cancers en général, mais plutôt à d'autres fonctions dans l'organisme.

3. Détermination du profil immunitaire

Le dernier aspect envisagé dans mon projet concerne l'interaction entre les cellules du microenvironnement et le développement de tumeurs, et plus particulièrement le rôle de l'infiltration de certaines cellules immunitaires dans celui-ci.

Les cellules infiltrant les tumeurs sont la conséquence de réponses générées directement par le système immunitaire (22, 40). Des hypothèses ont été émises sur le fait que certaines de ces cellules pouvaient potentiellement aider à améliorer la condition des patients, et qu'il serait également possible d'intervenir cliniquement sur ces cellules (26, 68).

Plusieurs études ont suggéré que des cellules du microenvironnement peuvent prédisposer les patients à une diversité de réponses aux traitements, et qu'elles représentent également une opportunité pour avoir une vue plus globale sur des potentiels biomarqueurs, ainsi qu'une meilleure stratification des patients (32). Parmi ces cellules, les cellules immunitaires infiltrant les tumeurs sont le type majeur non tumoral pouvant influencer le pronostic (26, 34). La quantification des cellules immunitaires peut donc mettre en avant les mécanismes sous-jacents contrebalançant les cancers (31). Dépendamment du type de tumeur, mais surtout du type cellulaire hautement infiltré, l'aboutissement clinique peut grandement différer (34). Cependant, comme les leucémies lymphoïdes sont des tumeurs du sang, l'aspect d'infiltration en tant que tel, c'est à dire dans une masse, n'est pas visible comme dans les autres types de cancers solides. C'est pourquoi pour cette partie du projet, trois cancers solides fréquents en pédiatrie ont été choisis, soit les neuroblastomes, les ostéosarcomes et les tumeurs de Wilms.

Des analyses ont montré que le système de classification TNM mis en place se basant sur l'anatomie des cancers, distinguait moins bien la stratification des pronostics comparativement au modèle du score immunitaire (90). Les approches disponibles pour calculer ces infiltrations sous la forme d'un score sont globalement divisées en deux catégories : celles basées sur des gènes signatures spécifiques, et celles basées sur les méthodes de déconvolution (39). Ces méthodes informatiques sont avantageuses au niveau du coût, mais aussi au niveau du temps et de la charge de travail par rapport aux techniques expérimentales (52).

Dans plusieurs cancers tels que dans le cancer du sein, dans les mélanomes ou les métastases des cancers colorectaux, les lymphocytes T CD8+ sont associés à de bonnes réponses aux chimiothérapies et une prolongation de la survie (28, 90). Il a été rapporté que les cellules T CD4+ et CD8+, composantes majeures du microenvironnement, servent généralement à limiter la croissance cellulaire et supprimer l'infiltration tumorale (28). Dans certains types de cancers, ces impacts n'ont pas encore été démontrés ou trop peu étudiés (90). Il semblerait que le microenvironnement tumoral influencerait le système immunitaire de façon protectrice en développant une immunité anti-tumorale, ou bien de façon dommageable en entrainant la progression de la tumeur (28, 90).

Dans mon projet, j'ai mesuré deux scores relatifs aux lymphocytes T : un basé sur l'outil de déconvolution Cibersort, l'autre basé sur l'expression de gènes marqueurs.

Au niveau de la distribution des scores, j'ai remarqué que les scores de Cibersort sont moins élevés que ceux estimés par la méthode d'expression des gènes signatures. Cette différence serait due au fait que le score d'expression à partir d'un certain panel de gènes hautement spécifiques, serait plus important que celui obtenu à partir de la déconvolution qui considère plutôt l'ensemble des gènes (31). On peut également s'attarder sur le mode de fonctionnement de Cibersort qui dispose d'une fonction censée enlever la variance des gènes des cellules stromales et immunitaires, souvent surexprimées dans certains cancers, tout en essayant de considérer les gènes plus spécifiques à certains types cellulaire (39). Cependant, avec cette méthode, d'autres gènes potentiellement spécifiques dans notre cas peuvent être enlever par défaut (i.e. de manière non intentionnelle) (39).

Par ailleurs, j'ai observé que les neuroblastomes présentent des valeurs de scores plus élevées que les deux autres types de tumeur. Ce genre d'observations a déjà été considéré dans d'autres études de déconvolution, dans lesquelles il a été révélé que ce type de tumeur possède le plus haut taux d'infiltration en lymphocytes T parmi les cancers pédiatriques (34).

Lors de la comparaison des scores entre les trois types de tumeurs étudiés, il faut prendre en considération que les sarcomes et les ostéosarcomes montrent généralement une infiltration immunitaire plus grande avec l'implication des cellules T, alors que les tumeurs de Wilms ou le sarcome synovial montrent un plus faible taux d'infiltration en lymphocytes T (34).

Dans mon projet, le score calculé avec Cibersort affiche des médianes qui semblent assez proches, avec les neuroblastomes possédant le plus haut score, suivi étrangement par les tumeurs de Wilms puis finalement par les ostéosarcomes. Une analyse plus poussée m'a permis de mettre en évidence une plus grande différence entre les ostéosarcomes et les deux autres types de tumeurs. En regardant les résultats obtenus avec le second

score (score d'expression T), qui montre une plus haute spécificité des marqueurs T, on note que les tumeurs de Wilms présentent un score plus bas, ce qui est davantage en accord avec ce qui a été rapporté dans la littérature (34). Les différences les plus significatives sont maintenant entre les neuroblastomes et les deux autres types de tumeur. À partir de ces observations, il est possible de déduire que des différences de score, et donc d'infiltration, existent entre les différents types de tumeur. Ceci est cohérent avec plusieurs études qui ont soulignées le fait que ces infiltrations varient dépendamment du type cellulaire, mais également du type de cancer, du niveau de maturation, de leur statut fonctionnel, de leur localisation dans les tissus, et même de la façon dont ils interagissent (40).

Malgré ces différences, on note une bonne corrélation entre ces deux scores. J'ai ensuite comparé ces scores avec des paramètres cliniques pour déterminer si cette infiltration a un impact potentiel sur le pronostic.

La comparaison entre le score de Cibersort et les critères cliniques ne montre aucune différence significative, et ce, peu importe le type de tumeur, alors que plusieurs études suggèrent autrement (22, 33, 35, 37, 38). Par ailleurs, les neuroblastomes ne présentent pas de données pour les patients sans évènements, il s'agirait de savoir si cela est lié à un facteur spécifique dans ce type de cancer, ou s'il n'y a juste pas de patients pour ce cas dans nos données. Ceci peut suggérer plusieurs choses, soit le score calculé avec Cibersort n'est pas assez sensible pour détecter des différences d'infiltration, soit il n'y a tout simplement pas d'impact sur ces données avec l'infiltration des cellules T.

En théorie, les méthodes de déconvolution réussissent à quantifier les différents types cellulaires de manière relativement proche de la réalité, sans pour autant nécessairement avoir besoin de gènes spécifiques à un seul type cellulaire ou bien de connaître tous les types d'un ensemble de gènes (59). Cependant, les outils dépendent des données avec lesquelles ils ont été créés et validés. Ils ont plutôt du mal à s'adapter à des ensembles de gènes trop différents de leur référence. De ce fait, le microenvironnement étant lui-même dynamique, ce dernier peut venir perturber la biologie usuelle et compliquer la

tâche des outils, ou encore augmenter la probabilité d'erreur d'estimation de manière inaperçue (59).

Les méthodes basées sur un nombre plus restreint, mais plus spécifique, de gènes marqueurs permettrait peut-être de diminuer cette dépendance à la plateforme utilisée (42, 52).

Pour tester cette hypothèse, j'ai effectué les mêmes analyses mais cette fois en utilisant les scores basés sur les gènes signatures. Cette fois, j'ai observé une corrélation entre le score et le statut vital chez les ostéosarcomes. En effet, au niveau du statut vital, une différence d'infiltration est observée entre les personnes vivantes et décédées. Les personnes encore en vie présentent des taux d'infiltration en lymphocytes T supérieurs aux personnes défuntes. On note également que les patients n'ayant jamais eu d'évènements (p. ex. rechute) ont une médiane de score d'infiltration plus haute que les patients ayant connus des rechutes ou des progressions tumorales. Une tendance similaire avait été observée avec le score de Cibersort mais n'était pas significative. Des études effectuées auparavant avaient également rapporté qu'un haut score immunitaire favorisait l'amélioration des ostéosarcomes, où il était possible d'observer que les patients démontraient un taux de survie sur 5 ans plus long lorsqu'ils présentaient un score immun plus haut (91). Les deux sous types les plus communs analysés dans l'étude étaient les lymphocytes T et les macrophages (91). Ces observations laissent sous-entendre qu'une plus haute infiltration de cellules T favoriserait un meilleur pronostic pour les patients avec un ostéosarcome.

En résumé, l'infiltration des cellules T varie donc non seulement à travers les individus, mais également à travers les différents types de cancers. De plus, en corrélant ces scores d'infiltration avec les variables cliniques, une tendance a pu être observée chez les ostéosarcomes avec le score utilisant les gènes marqueurs pour les cellules T. Un plus haut taux de lymphocytes T fournit potentiellement de meilleures chances de survie, et dans une moindre mesure, permet d'être moins sujet à des évènements tels que les rechutes.

On s'attendait à des observations similaires pour les deux autres types de tumeurs. En effet, l'impact des lymphocytes T a été associé à une réponse bénéfique dans les neuroblastomes, il y a déjà plus de 40 ans (92). Cependant, il n'est pas encore tout à fait clair si les différences d'infiltrations observées étaient dues à des caractéristiques de la tumeur, à des effets du traitement ou bien à d'autres facteurs (92). Plusieurs études ont proposé que les lymphocytes T ont une valeur pronostic indépendante des indicateurs utilisés pour la stratification usuelle des patients (92). Les cellules T CD8+ ou T CD3+ sont généralement corrélées avec un bon pronostic dans les neuroblastomes, où les hautes infiltrations présentent de meilleures évolutions, comparativement aux infiltrations de monocytes (92). Des études dans les tumeurs de Wilms ont noté qu'un haut taux d'infiltration des cellules T CD8 est corrélé avec un nombre plus faible de récidive tumorale (93). Un faible score a quant à lui été associé à un mauvais pronostic et une durée plus courte sans évènements (93). Encore une fois ici, le niveau d'infiltration des cellules T était généralement associé à un meilleur résultat clinique.

La discordance entre ces études et la mienne pourrait s'expliquer, entre autres, par le petit échantillonnage de mon étude. L'absence de différence entre les patients à haute et basse infiltration peut aussi supposer que les lymphocytes infiltrant les tumeurs peuvent être présents, sans pour autant être assez spécifiques ou fonctionnels pour contrer les actions de la tumeur (94).

De plus, il n'y a pas réellement de contrôle dans ce genre d'études qui pourrait donner des références sur la norme, ce qui suppose que les résultats observés dans ce type d'analyses prédisent généralement les pronostics pour des patients ayant normalement déjà reçu un diagnostic. Une analyse sur un plus long terme permettrait de voir l'évolution de ces infiltrations en fonction des divers critères. Il serait également intéressant de quantifier d'autres types cellulaires et de les corréler aux paramètres cliniques pour voir si d'autres types cellulaires présents à plus haut taux ont aussi un impact clinique. Ces derniers peuvent faire varier les scores des autres infiltrations sans que l'on s'en rende forcément compte en analysant un seul type cellulaire. Les macrophages sont souvent enrichis dans les tumeurs solides et décrits comme étant associés à un mauvais pronostic dans plusieurs cancers (95). En effet, ils sont souvent générateurs d'une action pro-

tumorale et sont impliqués dans la suppression des fonctions lymphocytaires T, ou dans la production de facteurs de croissance contribuant à la progression tumorale (96). Certaines cellules du microenvironnement non cancéreuses comme les cellules stromales ou mesenchymales peuvent aussi promouvoir la progression de la tumeur (28, 68, 96). Il y a donc potentiellement d'autres facteurs pouvant influencer nos résultats.

- Chapitre 7 -

Conclusion et perspectives

Leucémies lymphoblastiques aigues

Les leucémies lymphoblastiques aigues présentent encore beaucoup de décès, de rechutes et d'effets secondaires liés à des thérapies généralisées (71). Des stratégies de stratification des patients sont nécessaires pour améliorer l'efficacité des thérapies (71). Une meilleure caractérisation et compréhension des sous-types de leucémies permettrait de faire avancer les pronostics et les classifications (71). Cependant, certains facteurs importants tels que la pureté tumorale notamment, augmentent la complexité des analyses (49, 78).

Dans mon projet, j'ai utilisé des outils de déconvolution pour analyser et quantifier cette pureté tumorale. J'ai ainsi démontré que les résultats sont variables selon l'outil utilisé. La méthodologie derrière les outils est un important facteur à considérer. Il faut choisir la méthode la plus adaptée en fonction de nos données. J'ai démontré que l'efficacité de l'outil varie dépendamment des sous-types cellulaires, et que des approches multimodales sont toujours intéressantes à favoriser. Ce projet a aussi démontré l'importance de la matrice signature (96). L'idéal serait de développer une matrice signature à partir de l'état des cellules dans nos données étant donné qu'une simple différence de marqueurs ou de maturation des types cellulaires peut mener à des estimations différentes (96).

Dans le futur, il serait intéressant de tester d'autres approches qui prennent en compte ces limites. L'utilisation de données « single-cell » permettrait d'identifier avec précision les types cellulaires normaux et cancéreux présents dans les échantillons leucémiques et utiliser leurs signatures pour la déconvolution (85). D'autres approches par apprentissage machine qui reconstruisent le transcriptome des types cellulaires de chaque échantillon, et qui ne nécessitent pas de matrice de référence peuvent aussi être intéressantes (97).

Enfin, il serait intéressant de pouvoir développer des outils de déconvolution capables de soustraire toutes les cellules non cancéreuses, pour quantifier seulement la partie tumorale.

Tumeurs solides

Dans mon projet, j'ai aussi démontré que le score immunitaire basé sur l'expression des gènes signatures est associé avec certains paramètres cliniques chez les ostéosarcomes. Les personnes vivantes et celles n'ayant pas eu d'évènements tels que des rechutes ou une progression tumorale, présentaient des infiltrations plus hautes en lymphocytes T. Cette observation est intéressante car l'estimation de l'infiltration cellulaire est importante dans la prise en charge d'un patient (30, 95).

En conclusion, la classification plus précise des groupes de patients à l'aide d'outils moléculaires, permettra de mieux gérer la trajectoire des enfants atteints de cancers, dans le but de diminuer la mortalité et la toxicité des traitements.

Références bibliographiques

1. Comité consultatif des statistiques canadiennes sur le cancer. Statistiques canadiennes sur le cancer 2019. Toronto (ON): Société canadienne du cancer; 2019.

2. Ellison LF, De P, Mery LS, Grundy PE. Canadian cancer statistics at a glance: cancer in children. Canadian Medical Association Journal. 2009;180(4):422-4.

3. Cancer in Children and Adolescents: National Cancer Institute; 2020 [Disponible: https://www.cancer.gov/types/childhood-cancers/child-adolescent-cancers-fact-sheet#what-are-the-possible-causes-of-cancer-in-children].

4. Ellison L, Janz T. Incidence du cancer et mortalité par cancer chez les enfants au Canada. Coup d'oeil sur la santé: Statistique Canada; 2015.

5. WHO Global initiative for childhood cancer : an overview. World Health Organization; 2020.

6. Hodgson CS. Le cancer du sang au Canada: Faits et statistiques. Société de leucémie & lymphome du Canada; 2016.

7. Colby-Graham MF, Chordas C. The childhood leukemias. Journal of Pediatric Nursing. 2003;18(2):87-95.

8. Janeway CA Jr, Travers P, Walport M, Shlomchik MJ. The components of the immune system. 2001. In: Immunobiology: The Immune System in Health and Disease [Internet]. New York: Garland Science. 5th edition.

9. Graux C. Biology of acute lymphoblastic leukemia (ALL): clinical and therapeutic relevance. Transfusion and Apheresis Science. 2011;44(2):183-9.

10. Bhojwani D, Yang JJ, Pui CH. Biology of childhood acute lymphoblastic leukemia. Pediatr Clin North Am. 2015;62(1):47-60.

11. Tasian SK, Loh ML, Hunger SP. Childhood acute lymphoblastic leukemia: Integrating genomics into therapy. Cancer. 2015;121(20):3577-90.

12. Lajoie M, Drouin S, Caron M, St-Onge P, Ouimet M, Gioia R, et al. Specific expression of novel long non-coding RNAs in high-hyperdiploid childhood acute lymphoblastic leukemia. PLoS One. 2017;12(3).

13. Li J, Dai Y, Wu L, Zhang M, Ouyang W, Huang J, et al. Emerging molecular subtypes and therapeutic targets in B-cell precursor acute lymphoblastic leukemia. Front Med. 2021;15(3):347-71.

14. Belver L, Ferrando A. The genetics and mechanisms of T cell acute lymphoblastic leukaemia. Nature. 2016;16(8):494-507.

15. Chiaretti S, Zini G, Bassan R. Diagnosis and subclassification of acute lymphoblastic leukemia. Mediterranean Journal of Hematology and Infectious Diseases. 2014;6(1):e2014073.

16. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. The Lancet. 2013;381(9881):1943-55.

17. Hunger SP, Mullighan CG. Acute Lymphoblastic Leukemia in Children. New England Journal of Medicine. 2015;373(16):1541-52.

18. Ching-Hon Pui, William E. Evans. Treatment of acute lymphoblastic leukemia. New England Journal of Medicine. 2006;354:166-78.

19. Pui C-H. Recent Research Advances in Childhood Acute Lymphoblastic Leukemia. Journal of the Formosan Medical Association. 2010;109(11):777-87.

20. Ries Lag, Smith MA, Gurney JG, Linet M, Tamra T, Young JL, et al. Cancer Incidence and Survival among Children and Adolescents: United States

SEER Program 1975-1995. National Cancer Institute, SEER Program. 1999.

21. Sullivan K. The immune system and primary immunodeficiency diseases. 2013. In: IDF Patient & Family Handbook for primary immunodeficiency diseases. 5th edition.

22. Hugo Gonzalez, Catharina Hagerling, Zena Werb. Roles of the immune system in cancer - from tumor initiation to metastatic progression. Genes & Development. 2018;32:1267-84.

23. Woodward W. Phagocytes: TeachMe Physiology; [modifié Mars 2021. Disponible: https://teachmephysiology.com/immune-system/cells-immune-system/phagocytes/.

24. Immune cell markers poster: abcam; [Disponible: https://www.abcam.com/primary-antibodies/immune-cell-markers-poster.

25. Vinay DS, Ryan EP, Pawelec G, Talib WH, Stagg J, Elkord E, et al. Immune evasion in cancer: Mechanistic basis and therapeutic strategies. Seminars in Cancer Biology. 2015;35 Suppl:S185-S98.

26. Fridman WH, Pages F, Sautes-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nature. 2012;12(4):298-306.

27. Hao Y, Yan M, Heath BR, Lei YL, Xie Y. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. PLoS Computational Biology. 2019;15(5).

28. Giraldo NA, Sanchez-Salas R, Peske JD, Vano Y, Becht E, Petitprez F, et al. The clinical role of the TME in solid cancer. British Journal of Cancer. 2019;120(1):45-53.

29. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communication. 2013;4:2612.

30. Koirala P, Roth ME, Gill J, Piperdi S, Chinai JM, Geller DS, et al. Immune infiltration and PD-L1 expression in the tumor microenvironment are prognostic in osteosarcoma. Scientific Reports. 2016;6.

31. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics. 2019;35(14):436-45.

32. Petitprez F, Sun CM, Lacroix L, Sautes-Fridman C, de Reynies A, Fridman WH. Quantitative Analyses of the Tumor Microenvironment Composition and Orientation in the Era of Precision Medicine. Frontiers in Oncology. 2018;8:390.

33. Yadav DK, Jain V, Dinda AK, Agarwala S. Tumor-Infiltrating Lymphocytes in Wilms Tumor. Indian Journal of Medical and Paediatric Oncology. 2021;41(01):34-8.

34. Grobner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature. 2018;555(7696):321-7.

35. Batchu S. Immunological landscape of Neuroblastoma and its clinical significance. Cancer Treatment and Research Communication. 2021;26:2468-942.

36. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA : a Cancer Journal for Clinicians. 2021;71(1):7-33.

37. Wu CC, Beird HC, Andrew Livingston J, Advani S, Mitra A, Cao S, et al. Immuno-genomic landscape of osteosarcoma. Nature Communication. 2020;11(1):1008.

38. Holl EK, Routh JC, Johnston AW, Frazier V, Rice HE, Tracy ET, et al. Immune expression in children with Wilms tumor: a pilot study. Journal of Pediatric Urology. 2019;15(5):441 e1- e8.

39. Jimenez-Sanchez A, Cast O, Miller ML. Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. Cancer Research. 2019;79(24):6238-46.

40. Jochems C, Schlom J. Tumor-infiltrating immune cells and prognosis: the potential link between conventional cancer therapy and immunity. Exp Biol Med. 2011;236(5):567-79.

41. Hong M, Tao S, Zhang L, Diao LT, Huang X, Huang S, et al. RNA sequencing: new technologies and applications in cancer research. Journal of Hematology & Oncology. 2020;13(1):166.

42. Coccaro N, Anelli L, Zagaria A, Specchia G, Albano F. Next-Generation Sequencing in Acute Lymphoblastic Leukemia. International Journal of Molecular Sciences. 2019;20(12).

43. Lin Y, Li H, Xiao X, Zhang L, Wang K, Yang W, et al. DAISM-DNN: Highly accurate cell type proportion estimation with in silico data augmentation and deep neural networks. 2020.

44. Vaske OM, Bjork I, Salama SR, Beale H, Tayi Shah A, Sanders L, et al. Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer. JAMA Network Open. 2019;2(10).

45. Wang Y, Mashock M, Tong Z, Mu X, Chen H, Zhou X, et al. Changing Technologies of RNA Sequencing and Their Applications in Clinical Oncology. Frontiers in Oncology. 2020;10:447.

46. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Medicine. 2017;9(75).

47. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Frontiers in Genetics. 2019;10:317.

48. Li H, Sharma A, Ming W, Sun X, Liu H. A deconvolution method and its application in analyzing the cellular fractions in acute myeloid leukemia samples. BMC Genomics. 2020;21(1):652.

49. Zhang W, Feng H, Wu H, Zheng X. Accounting for tumor purity improves cancer subtype classification from DNA methylation data. Bioinformatics. 2017;33(17):2651-7.

50. Chen SH, Kuo WY, Su SY, Chung WC, Ho JM, Lu HH, et al. A gene profiling deconvolution approach to estimating immune cell composition from complex tissues. BMC Bioinformatics. 2018;19(Suppl 4):154.

51. Jen WY, Chee YL, Cheong MA. Flow Cytometry: LearrnHaem; 2021

52. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunology, Immunotherapy. 2018;67(7):1031-40.

53. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nature Communication. 2020;11(1).

54. Tappeiner E, Finotello F, Charoentong P, Mayer C, Rieder D, Zlatko T. TIminer: NGS data mining pipeline for cancer immunology and immunotherapy Bioinformatics. 2017;33(19):3140-1.

55. Aran D, Hu Z, Butte A J. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biology. 2017;18.

56. Becht E, Giraldo N. A, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biology. 2016;17(218).

57. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013;29(8):1083-5.

58. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra P W. PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. PLoS Computational Biology. 2012;8(12).

59. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods. 2015;12(5):453-7.

60. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biology. 2016;17(174).

61. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife. 2017;6.

62. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. quanTIseq : quantifying immune contexture of human tumors. 2018.

63. Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, et al. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. BMC Bioinformatics. 2010;11(27).

64. Zhong Y, Wan Y-W, Pang K, Chow L ML, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics. 2013;14(89).

65. Brunet J-P, Tamayo P, Golub T R, Mesirov J P. Metagenes and molecular pattern discovery using matrix factorization. PNAS. 2004;101(12):4164-9.

66. Plattner C, Finotello F, Rieder D. Deconvoluting tumor-infiltrating immune cells from RNA-seq data using quanTIseq. Methods Enzymology. 2020;636:261-85.

67. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods Molecular Biology. 2018:243-59.

68. Chen Y, Zhao B, Wang X. Tumor infiltrating immune cells (TIICs) as a biomarker for prognosis benefits in patients with osteosarcoma. BMC Cancer. 2020;20(1):1022.

69. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016;48(10):1193-203.

70. Spinella JF, Richer C, Cassart P, Ouimet M, Healy J, Sinnett D. Mutational dynamics of early and late relapsed childhood ALL: rapid clonal expansion and long-term dormancy. Blood Advances. 2018;2(3):177-88.

71. Khater F, Vairy S, Langlois S, Dumoucel S, Sontag T, St-Onge P, et al. Molecular Profiling of Hard-to-Treat Childhood and Adolescent Cancers. JAMA Network Open. 2019;2(4).

72. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics. 2018;34(11):1969-79.

73. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550.

74. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communication. 2019;10(1):1523.

75. Danaher P, Warren S, Dennis L, D'Amico L, White A, Disis ML, et al. Gene expression markers of Tumor Infiltrating Leukocytes. J Immunother Cancer. 2017;5:18.

76. Fischerr S, Gillis J. How many markers are needed to determine a cell's type? iScience. 2021;24(11).

77. Francisco Avila Cobos, Jo Vandesompele, Pieter Mestdagh, Katleen De Preter. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics. 2018;34(11):1969–79.

78. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology. 2019;37(7):773-82.

79. Bhinder B, Elemento O. Computational methods in tumor immunology. Methods Enzymology. 2020;636:209-59.

80. Cheung A. K., Ip J. C., Chu A. C., Cheng Y., Leong M. M., Ko J. M., et al. PTPRG suppresses tumor growth and invasion via inhibition of Akt signaling in nasopharyngeal carcinoma. Oncotarget. 2015;6(15):13434-47.

81. SPRED1 Is a Tumor Suppressor in Mucosal Melanoma. Cancer Discovery. 2018;8(12):1507.

82. Pasmant E, Gilbert-Dussardier B, Petit A, de Laval B, Luscan A, Gruber A, et al. SPRED1, a RAS MAPK pathway inhibitor that causes Legius syndrome, is a tumour suppressor downregulated in paediatric acute myeloblastic leukaemia. Oncogene. 2015;34(5):631-8.

83. Hong Q, Li R, Zhang Y, K. G. Fibrillin 2 gene knockdown inhibits invasion and migration of lung cancer cells. Cell Molecular Biology. 2020;66(7):190-6.

84. Xie J, Zhong Y, Chen R, Li G, Luo Y, Yang J, et al. Serum long non-coding RNA LINC00887 as a potential biomarker for diagnosis of renal cell carcinoma. FEBS Open Biology. 2020;10(9):1802-9.

85. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communication. 2019;10(1):380.

86. Paul MK, AK. M. Tyrosine kinase - Role and significance in Cancer. International Journal of Medical Science. 2004;1(2):101-15.

87. Khushwant S. Bhullar, Naiara Orrego Lagarón, Eileen M. McGowan, Indu Parmar, Amitabh Jha, Basil P. Hubbard, et al. Kinase-targeted cancer therapies: progress, challenges and future directions. Molecular Cancer. 2018;17(48).

88. Janiszewska M, Primi MC, T. I. Cell adhesion in cancer: Beyond the migration of single cells. Journal of Biological Chemistry. 2020;295(8):2495-505.

89. Zhou WY, Zheng H, Du XL, JL. Y. Characterization of FGFR signaling pathway as therapeutic targets for sarcoma patients. Cancer Biology & Medicine. 2016;13(2):260-8.

90. Jérôme Galon, Franck Pagès, Francesco M Marincola, Helen K Angell, Magdalena Thurin, Alessandro Lugli, et al. Cancer classification using the Immunoscore: a worldwide task force. Journal of Translational Medicine. 2012;10:205.

91. Yoshida K, Okamoto M, Aoki K, Takahashi J, N. S. A Review of T-Cell Related Therapy for Osteosarcoma. International Journal of Molecular Science. 2020;21(14).

92. Mina M, Boldrini R, Citti A, Romania P, D'Alicandro V, De Ioris M, et al. Tumor-infiltrating T lymphocytes improve clinical outcome of therapy-resistant neuroblastoma. Oncoimmunology. 2015;4(9).

93. Mardanpour K, Rahbar M, Mardanpour S, Mardanpour N, Rezaei M. CD8+ T-cell lymphocytes infiltration predict clinical outcomes in Wilms' tumor. Tumour Biology. 2020;42(12):1010428320975976.

94. Szanto CL, Cornel AM, Vijver SV, S. N. Monitoring Immune Responses in Neuroblastoma Patients during Therapy. Cancers 2020;12(2):519.

95. Zhang Y, Zhang Z. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. Cellular and Molecular Immunology. 2020;17(8):807-21.

96. Rhee JK, Jung YC, Kim KR, Yoo J, Kim J, Lee YJ, et al. Impact of Tumor Purity on Immune Gene Expression and Clustering Analyses across Multiple Cancer Types. Cancer Immunology Research. 2018;6(1):87-97.

97. Kevin Menden, Mohamed Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, et al. Deep learning–based cell composition analysis from tissue expression profiles. Sciences Advances. 2020;6(30).