Université de Montréal

Classification moléculaire des Tumeurs de Wilms par analyse RNA-Seq

Par

Cédric Roux

Département de Biochimie, Université de Montréal, Faculté de Médecine

Mémoire présenté en vue de l'obtention du grade de Maitrise en bio-informatique

Octobre 2020

Université de Montréal

Unité académique : Département de Biochimie, Université de Montréal, Faculté de Médecine

---

Ce mémoire (ou cette thèse) intitulé(e)

Classification moléculaire des Tumeurs de Wilms par analyse RNA-Seq

Présenté par

Cédric Roux

A été évalué(e) par un jury composé des personnes suivantes

François Major
Président-rapporteur

Daniel Sinnett
Directeur de recherche

Sébastien Lemieux
Codirecteur

Franz Bernd Lang
Membre du jury

Résumé

La tumeur de Wilms (TW) est un cancer du rein retrouvé principalement chez les enfants âgés de 2 à 4 ans. Elle représente 90% des cancers pédiatriques du rein. Le taux de survie des TW est supérieur à 90%. Ce dernier est rendu possible grâce à une stratification des patients en fonction du risque de récidive. Les TW sont classées dans différents groupes de risque selon le stade, l'histologie, la taille de la tumeur et la perte d'hétérozygotie aux loci 1p et 16q.

Deux régions sont importantes dans la génétique des TW, WT1 et WT2. WT1 est un gène qui code pour un facteur de transcription important dans différentes phases du développement rénal. Des anomalies de WT1 sont retrouvées dans certains syndromes humains tels que WAGR et Denys-Drash qui sont associés à l'émergence de TW. WT2 est un locus présent dans la région chromosomique 11p15 où une perte d'hétérozygotie conduit à une disomie uniparentale. Ceci entraîne un risque accru de TW dû à une surexpression de IGF2 qui est liée à la prolifération d'organes et de membres. Malgré ce portrait génétique, il n'existe pas de sous-groupes moléculaires qui permettent une classification des TW comme on observe chez d'autres cancers de l'enfant, notamment la leucémie.

Nous proposons de déterminer une stratification de sous-groupes groupes moléculaire à l'aide d'une signature moléculaire basée sur des profils d'expression qui améliorait la classification des TW dans leurs groupes de risque adéquat. Pour vérifier cette hypothèse, nous avons analysé le transcriptome d'une cohorte de 130 patients atteint d'une TW.

Grâce à un regroupement hiérarchique du profil d'expression des échantillons, nous avons identifié deux sous-groupes potentiels de TW. Un de ces sous-groupes est défini par une sous-expression de WT1 et une surexpression de gènes du début du développement musculaire chez les enfants et montrerait un risque plus fort de rechute. Ces résultats montrent que des outils de regroupement basés sur une signature moléculaire permettraient d'identifier des sous-groupes moléculaires chez les TW.

Mots-clés : tumeur de Wilms, cancer pédiatrique, rein, classification

Abstract

Wilms' tumor (TW) is a kidney cancer found mainly in children aged from 2 to 4 years old. It represents 90% of pediatric kidney cancers. The TW survival rate is over 90%. The latter is made possible by stratifying patients according to the risk of relapse. TW are classified into different risk groups according to stage, histology, tumor size and more recently the loss of heterozygosity at loci 1p and 16q.

Two regions are important in the genetics of WT, WT1 and WT2. WT1 is a gene that encodes an important transcription factor in different phases of renal development. WT1 abnormalities are found in some human syndromes such as WAGR and Denys-Drash which are associated with the emergence of TW. WT2 is a locus present in the chromosomal region 11p15 where a loss of heterozygosity leads to a uniparental disomy. This leads to an increased risk of WT due to an overexpression of IGF2 which is linked to the proliferation of organs and members. Despite this genetic portrait, there are no molecular subgroups which allow classification of TW as observed in other childhood cancers, including leukemia.

We propose to determine a stratification of WT using a molecular signature based on expression profiles in their proper risk group. To test this hypothesis, we analyzed the transcriptome of a cohort of 130 WT patients.

The hierarchical clustering of the of the sample's expressions profiles identified two potential WT subgroups. One of these subgroups can be described by a lower expression of WT1 and overexpression of genes for early muscle development in children and show a higher risk of relapse. These results show that clustering tools based on a molecular signature could allow treatment adjustment (i.e. precision medicine) and thus increase the survival rate.

Keywords : Wilm's tumor, pediatric cancer, kidney, classification.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

TW : tumeur de Wilms

WT1 : gène Wilms Tumor 1

WT2 : locus Wilms Tumor 2

RLE : Relative Log Expression

TARGET : Therapeutically Applicable Research To Generate Effective Treatments

VST : Transformation stabilisatrice de variance

LFC : Log2 Fold Change

À ma famille et mes amis

Remerciements

Dans un premier temps j'aimerais remercier le Dr Daniel Sinnett pour m'avoir accueilli dans son laboratoire afin d'effectuer cette maitrise. De plus, j'aimerais remercier Pascal St-Onge pour m'avoir guidé dans ma recherche. Un grand merci à toute l'équipe de recherche du laboratoire qui m'a conseillé lors des rencontres de laboratoire, et notamment toute l'équipe de bio-informatique. Je remercie aussi le Centre Hospitalier Universitaire Sainte Justine et l'Université de Montréal pour m'avoir procuré un environnement propice à toutes mes activités de recherche tout au long de ma maitrise.

Un grand merci tous mes amis proches, notamment Nicolas et Timothé qui ont toujours été là dans les moments les plus durs. Merci à mes colocataires Julie, Zacharie et Maxime et qui sont ma deuxième famille au Canada. Merci aussi à la Boucherie de Paris pour m'avoir nourri et à la Maisonnée pour m'avoir désaltéré tout au long de ce projet.

J'aimerais finir par remercier toute ma famille qui m'a supporté tout au long de mes études en me fournissant les outils nécessaires. Un grand merci à mon papa et ma maman qui m'ont fortement aidé lors la fin de ma maitrise.

Chapitre 1 – Introduction

Au Canada, 1500 nouveaux cas de cancers pédiatriques sont diagnostiqués chaque année. Les progrès des connaissances et les avances dans les traitements permettent de guérir environ 80% des cas [1]. Malgré tout, le cancer reste la principale cause de décès par maladie chez les enfants canadiens. Environ 20% des patients qui ne répondent pas aux traitements ou présentent des rechutes vont éventuellement succomber à leur maladie.

Les cancers pédiatriques sont différents de ceux survenant chez les adultes à la fois par leur type et leur comportement. Ils sont de types embryonnaires ou d'origine hématopoïétiques. Le type de cancer le plus fréquemment diagnostiqué chez les enfants est la leucémie (32 %), suivie des cancers du cerveau et du système nerveux (19 %) et des lymphomes (11 %).

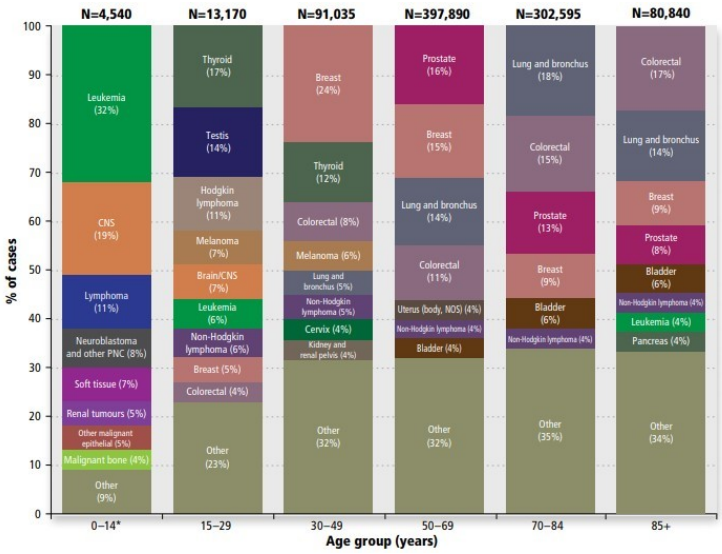

Figure 1. – Répartition des nouveaux cas de cancer selon le groupe d'âge, Canada, 2009-2013 [1].

Le traitement de ces patients est basé sur un processus de stratification associé à la notion de risque de récidive ou d'agressivité. Ces sous-groupes sont définis à l'aide d'une combinaison de caractéristiques cliniques, biologiques et moléculaires.

Alors que des sous-groupes moléculaires sont bien déterminés chez certains cancers, notamment les leucémies pédiatriques, j'ai concentré mon projet sur les tumeurs de Wilms, un cancer du rein qui n'a pas de sous-groupes moléculaires définis.

1.1 - Les tumeurs de Wilms

1.1.1 - La maladie

1.1.1.1 - Description

La tumeur de Wilms (TW) fascine les pathologistes depuis plus d'un siècle, car c'est l'un des exemples classiques de la façon qu'un cancer se développe au cours du développement de l'enfant. La TW, aussi connue sous le nom de néphroblastome, a classiquement une apparence triphasique, c'est-à-dire avec trois types de cellules : des cellules stromales, épithéliales et blastemales [2]. Cependant, la présence des trois types cellulaires n'est pas nécessaire pour diagnostiquer la tumeur de Wilms. Celle-ci est une masse au niveau du rein qui se développe à partir du tissu embryonnaire du rein. Ce tissu embryonnaire persiste lors du développement du rein, ce qui crée des précurseurs de tumeurs de Wilms qui sont retrouvés sous forme de résidus néphrogènes. Ceux-ci sont des fragments de tissu embryonnaire dans le rein qui sont retenus après la période de développement embryonnaire. Il existe deux types de résidus néphrogènes, les intra-lobulaires qui se retrouvent dans le lobe rénal, et les péri-lobulaires qui se retrouvent à l'extérieur du lobe rénal [3]. Comme illustré à la figure 2, ci-dessous, nous retrouvons des nids de tissus embryonnaires (taches noires dans le rein) lors du développement embryonnaire. Dans le cas d'un développement normal du rein, ces nids disparaissent. Lors d'une anomalie du développement causée par certaines mutations, nous observons une persistance de ces nids qui peuvent devenir par la suite des TW.

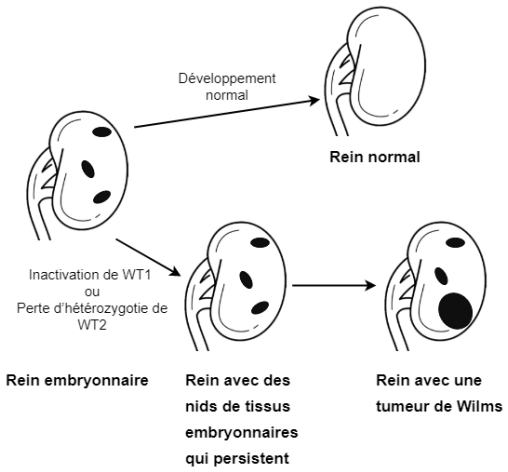

Figure 2. – Développement de la tumeur de Wilms due aux résidus néphrogènes

1.1.1.2 - Statistiques

La tumeur de Wilms est le cancer rénal le plus commun chez les enfants et les affecte surtout avant l'âge de cinq ans. C'est le sixième cancer pédiatrique en importance au Canada et représente 5% de tous les cancers pédiatriques et 95% des tumeurs du rein [4]. La tumeur de Wilms se présente également sous une forme héréditaire chez 5-7% des tumeurs rénales de l'enfant et peut alors se manifester sous une forme bilatérale, c'est-à-dire qu'il y a une tumeur sur chaque rein [2]. La survie des enfants atteints d'une TW est supérieure à 90%. Ce taux de survie dépend fortement du facteur de risque défini pour la TW diagnostiquée et peut tomber jusqu'à 56% de taux de survie pour les tumeurs les plus agressives qui représentes 2,3% des TW. Les nouvelles thérapies ont comme objectifs de maintenir et augmenter ce taux de guérison élevé tout en augmentant le taux de survie des sous-groupes à risques élevés.

Afin d'atteindre cet objectif d'obtenir de meilleurs taux de survies, nous devons trouver des solutions afin de mieux stratifier les tumeurs déterminant des facteurs de risques élevés. Ceci permettrait de cibler des traitements plus agressifs pour les TW ayant des taux de survie faible sans faire subir ces traitements aux patients ayant des tumeurs moins agressives et qui répondent bien aux traitements existants.

1.1.2 – Les classifications de tumeur de Wilms

La classification des TW est basée sur les caractéristiques suivantes afin de déterminer leurs facteurs de risques et donc de déterminer le traitement le plus adéquat pour un patient spécifique.

1.1.2.1 - Stadification

Les TW sont initialement classifiés selon un système de stade défini par le Children's Oncology Group [4]. Ces stades sont déterminés par la taille et l'avancement de la tumeur au niveau des organes externes. Ils existent cinq stades différents décrits dans le Tableau 1.

Tableau 1. – Stadification des tumeurs de Wilms [4]

Stade	Description
I	La tumeur est limitée au rein et peut être complètement retirée. La surface du rein reste intacte et la tumeur n'a pas été rompue avant et après le retrait de celle-ci.
II	La tumeur s'étend au-delà du rein, mais peut être complètement retirée. Il y a une extension régionale de la tumeur, pénétration de la surface extérieure du rein dans les tissus mous périrénaux.
III	Des résidus non hématogènes tumoraux confinés dans l'abdomen. Une ou plus de ces actions se produit : • Les ganglions lymphatiques sont impliqués dans la tumeur. • Il y a eu contamination du péritoine par la tumeur due à la biopsie, la rupture de la tumeur ou par la croissance de la tumeur. • Des implants sont retrouvés sur la surface du péritoine. • La tumeur s'étend au-delà des marges chirurgicales. • La tumeur n'est pas complètement résécable due à l'infiltration locale dans des structures vitales.
IV	Métastases hématogènes. Dépôts au-delà du stade III (poumons, foie, os et cerveau).
V	Atteinte rénale bilatérale. Le stade de chaque tumeur devrait être déterminé.

La stadification des TW n'est cependant pas la seule classification nécessaire pour déterminer les groupes de risques de ces tumeurs. Une classification histologique est aussi faite afin de mieux comprendre la tumeur.

1.1.2.2 - Classification histologique

La classification histologique sépare les tumeurs en deux grandes catégories : les tumeurs considérées anaplasiques et celles qui ne le sont pas [5]. La tumeur anaplasique présente de l'anaplasie chez leurs cellules, c'est-à-dire une perte anormale de certains caractères de la différenciation cellulaire, sans retour à l'état de cellule primaire. Les cellules anaplasiques ont généralement des noyaux hyperchromatiques avec un rapport de taille noyau:cytoplasme qui est proche de 1:1. Elles sont caractérisées également par une division cellulaire accrue. L'anaplasie est présente dans moins de 5% des TW.

Tableau 2. – Taux de survie par type histologique et stade de la tumeur avec la proportion de TW anaplasiques selon leurs stades [6].

Histologie	Stade	Survie sans rechute après 4 ans (%)	Survie globale après 4 ans (%)	Proportions par stades entre les histologies (%)
Favorable	I	91	96	21
	II	85	93	27,9
	III	84	89	28,5
	IV	75	81	12,5
Anaplasique	I	69	82	1,8
	II	82	81	1,7
	III	54	58	4,7
	IV	44	56	2.3

Comme nous pouvons le voir dans le tableau ci-dessus, le taux de rechute, et donc le taux de risque des tumeurs, est beaucoup plus fort chez les tumeurs de type histologique anaplasique. Ces taux de rechute sont aussi corrélés avec le stade de la tumeur.

1.1.2.3 - Perte d'hétérozygotie 1p et 16q

La perte d'hétérozygotie (ou LOH, *loss of heterozygosity*, en anglais) est la perte de matériel génétique qui peut être un allèle ou un locus spécifique due à une délétion importante ou à la perte d'une partie ou de la totalité d'un des chromosomes de la paire.

La perte d'hétérozygotie spécifique à la tumeur pour les chromosomes 1p et 16q identifie un sous-groupe de patients atteints d'une TW avec une histologie favorable, mais qui présentent un risque significativement accru de rechute et de décès. [7] Cette perte d'hétérozygotie doit donc être identifiée afin de pouvoir déterminer le groupe de risque de la tumeur.

Tableau 3. – Taux de tumeurs n'ayant pas fait de rechute selon la perte d'hétérozygotie (LOH) de 1p et 16q [8]

	LOH	Survie sans rechute après 4 ans (%)	Survie globale après 4 ans (%)	Proportions par stades (%)
Stade I et II avec une histologie favorable	Pas de LOH	91,2	98,4	77,3
	LOH 1p	80,4	91,2	6,2
	LOH 16q	82,5	98,1	11,8
	LOH 1p et 16q	74,9	90,5	4,7
Stade III et IV avec une histologie favorable	Pas de LOH	83,0	91,9	72,9
	LOH 1p	89,0	97,6	8,1
	LOH 16q	85,3	92,0	14,6
	LOH 1p et 16q	65,9	77,5	4,4

La perte d'hétérozygotie unique du locus 1p ou 16q montre un taux de rechute plus haut lorsque la TW est au stade I et II et un taux de rechute plus bas lorsque la TW est au stade III et IV. Cependant chez toutes les TW, une perte d'hétérozygotie des deux loci en même temps entraine une augmentation du taux de rechute des patients.

1.1.2.4- Groupes de risques des tumeurs de Wilms

Finalement, en prenant en compte du stade de la tumeur, de la classification histologique et de la perte d'hétérozygotie 1p et 16q de la tumeur, la classification des groupes de risque des tumeurs de Wilms suivante peut être effectuée. Ceci est fait grâce au diagramme de classification de la figure 3 ci-dessous.

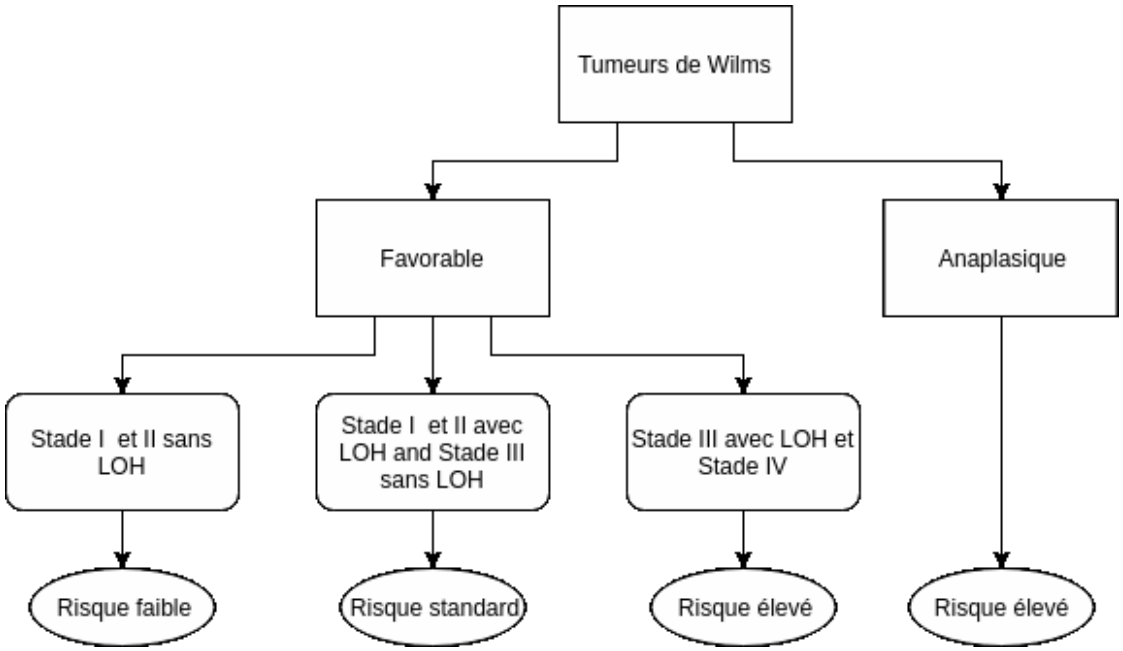

Figure 3. – Diagramme de la classification d'une tumeur de Wilms dans les groupes de risque [4]

Ces groupes de risques sont distribués en trois groupes : risque faible, risque standard et risque élevé. On peut attribuer des traitements spécifiques de plus en plus agressifs à chacun de ces groupes.

Nous voulons raffiner cette classification en essayant d'identifier et définir l'existence et le rôle de possibles sous-groupes moléculaires, et ceci en essayant de comprendre davantage la génétique derrière les TW.

1.1.3 - Génétique des tumeurs de Wilms

Les mécanismes génétiques et moléculaires derrière les TW ne sont pas entièrement compris, cependant il existe deux éléments importants dans ces mécanismes, le gène WT1 et le locus WT2.

1.1.3.1 - Gène Wilms Tumor 1 (WT1)

Le gène WT1 est localisé sur la région 11p13 du chromosome 11 et peut être muté dans la lignée germinale ou somatique dans environ 15% des cas de tumeurs de Wilms [9]. WT1 code pour un facteur de transcription important dans différentes phases du développement rénal. WT1 joue un rôle dans la croissance cellulaire du rein, notamment dans le processus de différenciation des cellules qui leur permet de remplir des fonctions spécifiques ainsi que dans le système d'apoptose des cellules du rein en développement. La protéine WT1 va réaliser ses fonctions grâce à une régulation de l'activité d'autres gènes en se liant directement à l'ADN. WT1 se comporte comme un gène suppresseur de tumeur, car les deux allèles doivent être supprimés ou inactivés afin d'aider la tumeur à se développer. Les tumeurs portant une mutation WT1 peuvent également présenter des mutations de gain de fonction dans le gène de la β-caténine (CTNNB1). L'accumulation de cette protéine provoque l'activation de la voie de signalisation canonique de Wnt qui a un impact sur la régulation de la transcription de gènes impliqués dans des processus cellulaires importants [10].

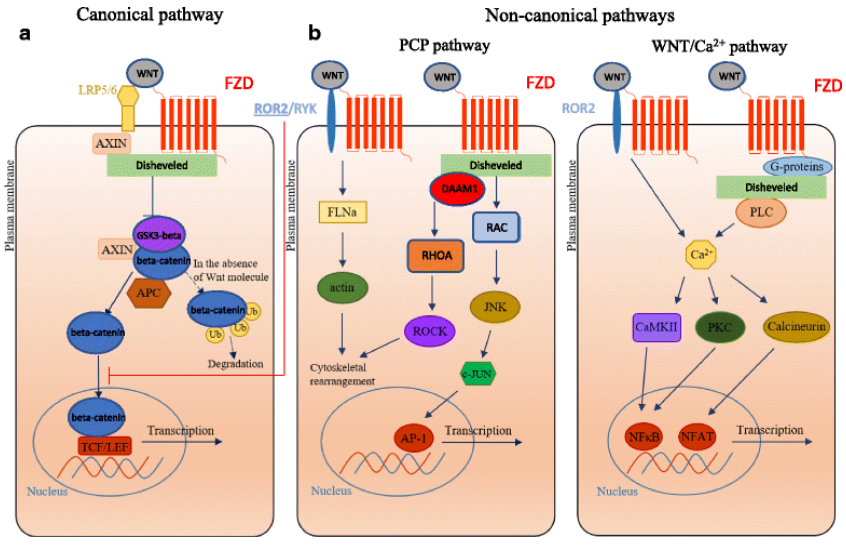

Figure 4. – Voies WNT canoniques (a) et non canoniques (b). [11]

On retrouve des anomalies de WT1 dans plusieurs syndromes humains. Le syndrome de WAGR (*Wilms' tumour, Aniridia, Genitourinary anomalies and Retardation*) entraîne une haplo-insuffisance de WT1 qui peut mener à une tumeur de Wilms dans 70% des cas ainsi qu'à une multitude d'anomalies gonadiques [12]. Des mutations dans le domaine des doigts de zinc de WT1 peuvent aussi conduire à un phénotype plus sévère associé au syndrome de Denys-Drash qui est également associé à de multiples anomalies du développement du système génito-urinaire. Ceci mène au développement de tumeurs de Wilms chez 90% des patients atteints du syndrome.

1.1.3.2 - Locus Wilms Tumor 2 (WT2)

WT2 est un locus trouvé dans la région chromosomique de 11p15. Cette région subit une perte d'hétérozygotie de l'allèle maternelle combinée à une duplication du 11p15 non perdue. La disomie uniparentale paternelle résultante engendre un risque accru de tumeurs de Wilms et d'anomalies du développement dues à une surexpression de IGF2 (*insulin-like growth factor 2*) [13]. La protéine d'IGF2 joue un rôle essentiel dans la croissance et le développement prénatal en favorisant la croissance et la prolifération de cellules dans de nombreux types de tissus. Seule la copie du père d'IGF2 est active ce qui est dû à l'empreinte génomique. Bien que le gène IGF2 soit hautement actif au cours du développement fœtal, il l'est beaucoup moins après la naissance.

1.1.3.3 - Tissus prédominants des tumeurs de Wilms

La forme la plus commune de TW est connue sous le nom de triphasique, comprenant des éléments majoritairement blastémales, avec des quelques éléments épithéliaux et stromal. Ces tumeurs présentent une architecture très similaire à celle du rein en développement.

Cependant, nous retrouvons des TW avec des tissus nettement prédominants. [14] La figure 5 illustre ces deux types de prédominance chez les TW et leurs développements. Tout d'abord, le développement normal des cellules du rein passe par une étape d'un signal d'induction qui permet le début de la prolifération des cellules souches du rein, puis une différentiation. Cependant des mutations chez WT1 ou WT2 vont bouleverser le développement des cellules rénales.

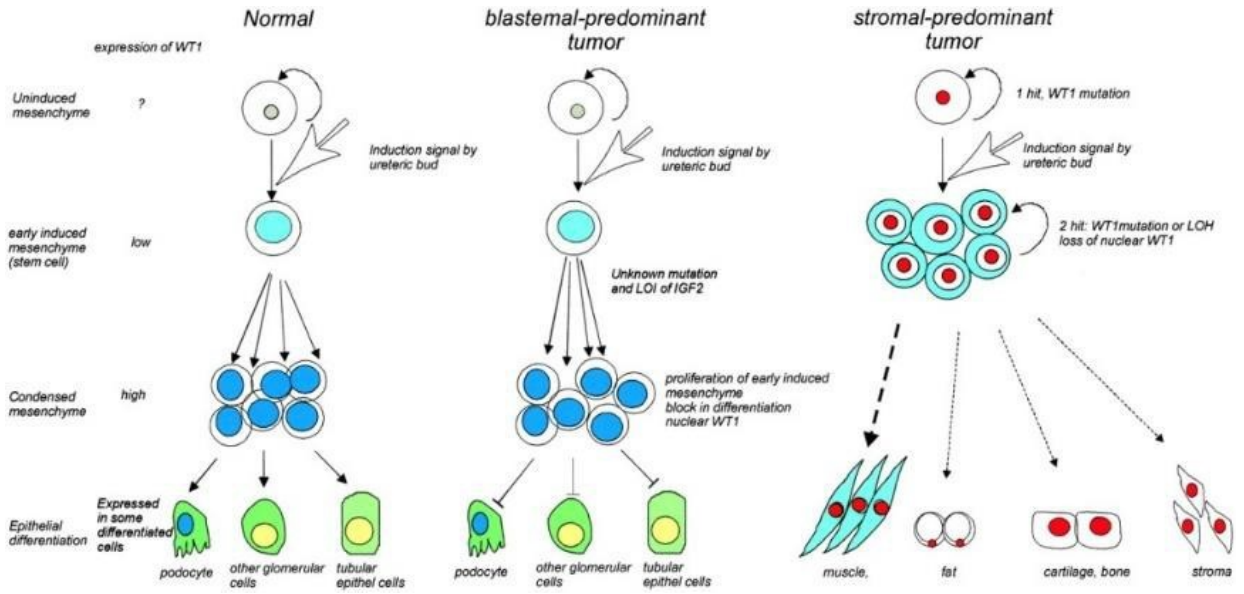

Figure 5. – Modèle pour le développement de deux sous-types différents de tumeurs de Wilms [14].

Les TW résultant de la mutation de WT1 sont principalement stromales et peuvent contenir des éléments de tissu hétérotypique, le plus souvent du muscle, mais plus rarement du cartilage, des os et de la graisse. Ceci serait dû à deux mutations de WT1 : la première mutation est soit dans la lignée germinale soit une mutation somatique très tôt dans le développement, et la deuxième mutation serait somatique en provoquant l'inactivation de WT1 dans les cellules souches lors du développement du rein. Ceci causerait une mauvaise différenciation de ces cellules en créant une tumeur à tissus hétérogènes. Comparée aux autres TW, cette forme qui peut être héréditaire ne répond pas bien à la chimiothérapie [14].

Les tumeurs contenant majoritairement des éléments blastémales seraient dues à la disomie uniparentale paternelle d'IGF2 et donc à WT2 [14]. Celle-ci engendre une surexpression de IGF2 responsable de la prolifération des cellules progénitrices et est suivie par un blocage de leurs différentiations en cellule épithéliale. De plus, ces tumeurs voient une forte présence des foyers néphrogéniques périlobaires considérés comme des lésions précurseurs aux TW et correspondent à un développement rénal avancé.

1.2 - Nouvelle approche de classification avec les méthodes de puces d'ADN

Jusqu'à récemment, la connaissance des bases génétiques de la TW était largement limitée aux aberrations discutées plus haut : mutations de WT1, anomalies de WT2 et mutations activatrices de Wnt impliquant CTNNB1. Cependant leurs implications dans les mécanismes génétiques et moléculaires qui sous-tendent ces associations ne sont pas entièrement comprises.

Une étude utilisant des données de puces d'ADN et des méthodes d'apprentissage non supervisées a essayé d'identifier des sous-groupes liés à l'expression génique [15]. Cette étude a été effectuée sur un groupe de 300 patients atteints de TW d'histologie favorable, la majorité ayant fait une rechute. Sur ces 300 patients, 224 échantillons ont été utilisés pour réaliser un contrôle de qualité.

Dans un premier temps, plusieurs analyses par regroupements hiérarchiques ont été réalisées. Une première analyse des 2000, 4000 et 10 000 gènes dont l'expression était la plus variable grâce à un regroupement hiérarchique non supervisée a été réalisée. Nous voyons le résultat du regroupement hiérarchique des 4000 gènes dans la figure 6a ci-dessous. On peut y observer deux sous-groupes S1, en bleu, et S2, en rouge, qui étaient stables dans chaque analyse, identifiant les mêmes échantillons dans chaque sous-ensemble.

Une nouvelle analyse a été effectuée par regroupement hiérarchique en utilisant les 54 gènes qui avaient un coefficient de corrélation de Pearson supérieur à 0,60 ou inférieur à -0,60 pour les deux allèles WT1 disponibles. Cette analyse représentée dans la figure 6b identifie deux sous-groupes supplémentaires S3 et S4, indiquée en vert et violet respectivement. Les gènes associés avec la différentiation musculaire sont identifiés sur cette figure.

Finalement, la figure 6c montre une dernière analyse par regroupement hiérarchique des gènes cibles de la voie de signalisation Wnt définis par la liste de Stanford. On y retrouve notamment le regroupement des sous-groupes S1 et S2, avec S1 qui présente une sous-expression des cibles Wnt et S2 une surexpression de ces cibles.

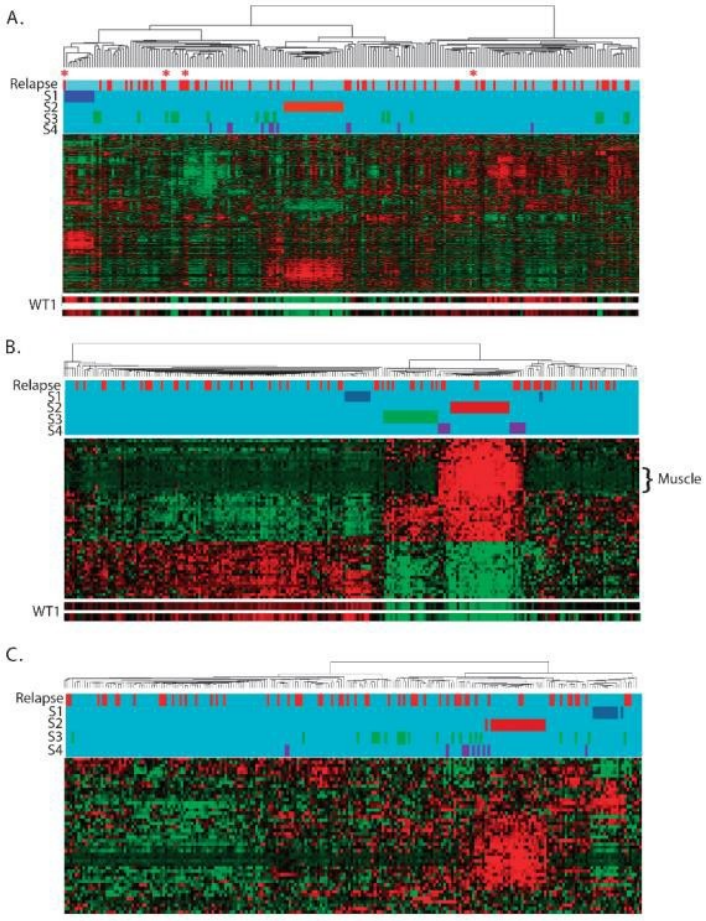

Figure 6. – Analyses par regroupement hiérarchique de TW présentée par l'étude Gadd, S., et al., 2012 [15]

Ces résultats illustrent le potentiel de regrouper certaines TW en divers sous-groupes à l'aide d'un profil d'expression génique similaire. L'étude a été poursuivie avec une analyse de la différence de taux d'expression génique entre les différents sous-groupes identifiés. Cette analyse a été réalisée avec l'outil GSEA qui a permis de caractériser les profils d'expression de gènes en utilisant des listes de gènes fournies. La figure 7 illustre les modèles d'expression des gènes permettant de distinguer les sous-groupes.

Figure 7. – Expression de gènes sélectionnés par l'étude Gadd, S., et al., 2012 [15] selon leurs sous-groupes. L'axe des y représente le logarithme du niveau d'expression des gènes sélectionnés. L'axe des x reflète un numéro arbitraire de la tumeur permettant de regrouper les tumeurs du même sous-groupe ensemble.

Par ces gènes, on note la présence de IGF2 qui est associé à WT2 ainsi que WIF1 et KLK6 pour lesquels la surexpression est liée à la perte de WT1. On observe aussi des gènes reliés au début du développement musculaire, plus particulièrement MYH3, TTN et ACTA1.

Cette étude montre qu'une meilleure compréhension du profil d'expression génique des TW peut mener à une classification des sous-groupes moléculaires. Ces derniers pourraient permettre une meilleure stratification des groupes de risques afin de proposer une thérapie optimale en fonction des anomalies caractéristiques au sous-groupe.

1.3 La Transcriptomique

La transcriptomique est le domaine d'étude portant sur le transcriptome. Le transcriptome représente l'ensemble des transcrits d'ARN dans une cellule pour un stade de développement ou d'une condition physiologique spécifique. L'étude du transcriptome est importante afin de mieux comprendre une maladie et les mécanismes moléculaires sous-jacents. L'interprétation des données transcriptomiques permet de déterminer le profil d'expression des gènes, et de quantifier leurs niveaux d'expression au cours du développement [16].

Au cours des années, de multiples technologies ont été développées afin d'étudier le transcriptome. Dans le cadre de mon mémoire, nous allons nous concentrer sur le séquençage à haut débit du transcriptome des ARNs (RNA-seq).

La technique de RNA-Seq permet d'examiner la quantité et la séquence des ARNs présents dans une cellule. Cette technique est possible grâce aux technologies de séquençages de nouvelle génération telle qu'Illumina [17]. Plusieurs étapes sont nécessaires lors d'une étude type du transcriptome par ARN-Seq.

Dans un premier temps, l'ARN doit être isolé et évalué quantitativement et qualitativement afin d'être transformé en librairies prêtes pour le séquençage [18]. Ces librairies sont composées de fragments d'ADNc auxquels des adaptateurs de séquençage sont fixés, soit à une des extrémités pour avoir un séquençage *single-end* ou soit aux deux extrémités pour un séquençage *paired-end*.

Le séquençage à haut débit est ensuite effectué sur chaque fragment à partir de l'extrémité à laquelle l'adaptateur est fixé ou à partir des deux extrémités lors d'un séquençage *paired-end*. Les lectures sont généralement comprises entre 30 et 400 paires de bases selon la technologie de séquençage utilisée.

Les données des lectures séquencées sont fournies dans des fichiers au format FASTQ qui sont généralement contaminés par des artéfacts ou des erreurs de séquençage qui pourraient conduire à une mauvaise interprétation des résultats [19]. Un contrôle de qualité est nécessaire afin de déterminer un score de qualité. Ce score permet de supprimer ou couper les lectures de

mauvaise qualité. Les séquences brutes nettoyées peuvent ensuite être alignées sur un génome de référence pour déterminer leurs localisations dans le génome grâce à des outils d'alignement [20]. L'alignement permet de réaliser une quantification des transcrits afin d'obtenir le niveau d'expression de chaque localisation génique couvert par les transcrits séquencés [21]. Finalement les profils d'expression d'échantillons représentant des conditions différentes peuvent être étudiés en comparant leurs niveaux d'expression dans le génome [22].

1.4 L'apprentissage machine en bio-informatique

L'apprentissage machine en bio-informatique est utilisé afin de permettre l'extraction d'information dans les données biologiques dont la quantité croît exponentiellement [23]. L'apprentissage machine est une application de l'intelligence artificielle qui utilise des techniques computationnelles et statistiques afin d'établir un modèle à partir de données préexistantes pour permettre de faire émerger des prédictions ou des décisions. L'application de l'apprentissage machine en bio-informatique pose des problèmes qui n'apparaissent pas lors de son utilisation classique dans des domaines tels que le traitement de langage naturel ou la classification d'images. Ces problèmes proviennent notamment du fait que pour les applications classiques de l'apprentissage machines les résultats peuvent facilement être vérifiés.

Les types d'algorithmes d'apprentissage automatiques diffèrent par leurs approches. Les deux grandes catégories d'algorithmes d'apprentissage sont les apprentissages non supervisés et supervisés.

1.4.1 Apprentissage non supervisé

L'apprentissage non supervisé prend des jeux de données qui ne contiennent pas d'étiquette sur le résultat attendu. L'algorithme va identifier les points communs des données et réagit en fonction de la présence ou de l'absence de tels points communs dans chaque nouvel élément de données. L'algorithme essaye de découvrir un modèle à partir des données d'entrée. De multiples algorithmes d'apprentissages non supervisés sont utilisés dans le domaine de l'analyse de données biologiques tels que le regroupement hiérarchique ou les réductions de dimensionnalité comme l'analyse en composantes principales (PCA).

Le regroupement hiérarchique est une approche qui applique une fonction de distance entre tous les points et qui forme des groupes selon la distance entre chaque point [24]. Le résultat final est un ensemble de groupes de points regroupés selon leurs distances directes entre eux. La séquence d'étapes permettant de créer les différents groupes peut être mise en œuvre selon une approche agglomérative pour produire la hiérarchie imbriquée des groupes [25]. Celle-ci commence par chaque point appartenant à leur propre groupe. À chaque étape, les deux groupes les plus similaires sont réunis jusqu'à ce que, après (N - 1) étapes, avec N le nombre d'observation, toutes les observations appartiennent à un seul groupe de taille N. La similarité entre deux groupes est mesurée à l'aide d'une fonction de distance qui est définie par l'algorithme de regroupement hiérarchique utilisé. Ensuite, une fonction de liaison est utilisée pour étendre cette notion de distance à des paires de groupes. Le résultat d'un regroupement hiérarchique est généralement observé sous la forme d'un dendrogramme [26].

La PCA fait partie des algorithmes d'apprentissage non supervisé de la famille de réduction de dimensionnalité [27]. Les algorithmes de réduction de dimensionnalité ont pour but d'extraire les caractéristiques les plus influentes du jeu de données. Comme son nom l'indique, ceci est fait en réduisant le nombre de dimensions des données d'entrée en limitant la perte d'information que cela produit. La PCA va permettre une réduction de dimensionnalité en transformant l'ensemble de données en un nouvel ensemble de variables appelé composantes principales (PC) qui ne sont pas corrélées et qui sont ordonnées de telle manière que les premières PC conservent l'essentiel de la variation présente dans les données d'entrées du jeu de donnée d'origine.

1.4.2 Apprentissage supervisé

L'apprentissage supervisé permet de déterminer un modèle de prédiction à partir d'un ensemble de données d'entrée auquel la valeur de sortie est connue. Le modèle est entrainé en comparant la valeur prédite à la vraie valeur de sortie via une fonction de coût prédéterminée et va modifier le modèle selon les erreurs de prédictions. La généralisation du modèle peut être évaluée grâce à un jeu de données externe, généralement nommée *test*.

1.5 Problématique et hypothèse

La stratification des cancers dans leurs différents groupes de risque est une étape importante afin de déterminer le traitement du patient. Cette stratification permet d'adapter le choix du traitement, avec un programme intensifié chez les patients avec une probabilité de survie plus faible, tout en épargnant les patients avec un faible risque de la toxicité de certains traitements. Pour les tumeurs de Wilms, cette stratification des groupes de risques est faite avec le stade de la tumeur et son type histologique et la perte d'hétérozygotie de 1p et 16q. Cependant, nous ne retrouvons pas de stratification en sous-groupes moléculaires comme on le retrouve pour d'autres cancers pédiatriques comme les ALL. Nous proposons donc d'identifier ces sous-groupes moléculaires avec l'hypothèse que l'intégration de données de séquençage de nouvelle génération, notamment la RNA-Seq, couplée avec l'application d'algorithme d'apprentissage machine non supervisée sur ces données permettrait d'identifier des sous-groupes moléculaires chez les tumeurs de Wilms permettant de reclassifier dans des groupes de risque plus adéquats.

L'objectif principal de cette étude est d'explorer les données d'ARN-Seq afin de trouver une signature d'expression permettant d'identifier des sous-groupes moléculaires chez les TW. Nos objectifs spécifiques sont donc de :

- Elaborer une méthode pour identifier différents profils d'expression dans nos échantillons de TW en utilisant les données d'ARN-Seq
- Regrouper les échantillons présentant des profils d'expression similaire en utilisant des algorithmes d'apprentissage non supervisé
- Caractériser les sous-groupes potentiels identifiés par analyse d'expression différentielle.

Chapitre 2 – Méthodologie

2.1. Les données TARGET

2.1.1 TARGET

Nous utilisons dans cette étude des données provenant d'une initiative de la National Cancer Institute nommée TARGET (Therapeutically Applicable Research To Generate Effective Treatments). TARGET utilise une caractérisation moléculaire afin de déterminer des changements génétiques qui entraînent l'initiation et la progression des cancers de l'enfant difficiles à traiter. TARGET met ces données générées à la disposition des chercheurs afin que de nouvelles stratégies de traitement plus efficaces puissent être développées et appliquées.

Nous avons récupéré les données de 130 échantillons de TW contenant à la fois des données d'ARN-Seq et les caractéristiques de la TW correspondante.

2.1.2 Caractéristiques des données

Les caractéristiques de chaque échantillon de TW permettent de voir les stades, les pertes d'hétérozygotie de 1p et 16q et la classification histologique de chaque échantillon. La distribution de ces caractéristiques est illustrée à la figure 10. Nous voyons une forte quantité d'échantillons de stades II et III comprenant 53 et 45 échantillons, respectivement. Le reste est distribué à travers les stades I avec 17 échantillons, IV avec 12 échantillons et finalement le stade V avec seulement 3 échantillons. Nous avons seulement 9 des 130 échantillons avec des pertes d'hétérozygotie 1p et 16q (Figure 8A). La classification histologique a une forte représentation de TW anaplasiques avec 31% des échantillons alors que nous retrouvons normalement que 5% de TW anaplasiques dans une distribution normale des TW. Ces caractéristiques nous permettent de déterminer le groupe de risque de chaque échantillon qui n'est pas explicitement représenté dans les métadonnées en suivant l'information du diagramme de la figure 3. Nous avons donc 55 échantillons dans le groupe de risque élevé, 29 de risque standard et 46 de risque faible (Figure 8B).

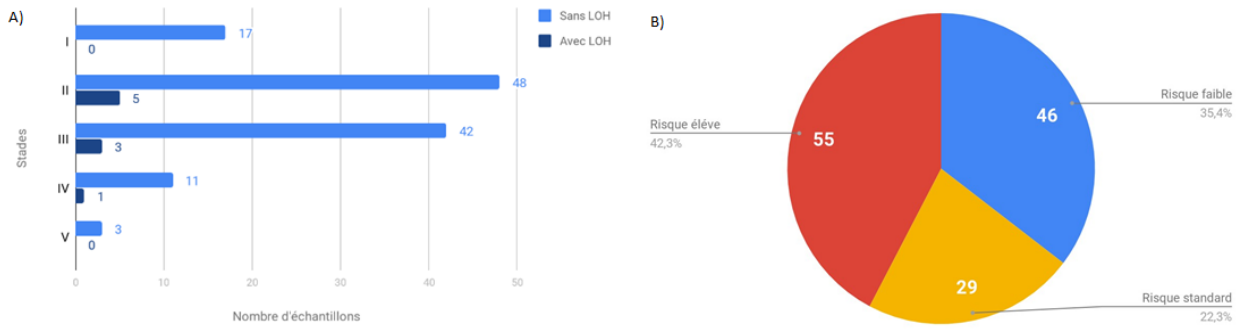

Figure 8. – A) Stadifications avec les pertes d'hétérozygotie 1p et 16q et B) groupes de risques des 130 échantillons de TW

2.2 Méthodes d'analyses

2.2.1 Comptes bruts

Chaque échantillon est tout d'abord soumis au pipeline standard utilisé dans le laboratoire de Dr Sinnett. L'alignement des transcrits de séquences est fait sur le génome de références GRCh37 avec l'aligneur STAR. La quantification des lectures est effectuée grâce à l'outil HTSeq.

Finalement les matrices de comptes bruts générées pour chaque échantillon sont ensuite fusionnées dans une unique matrice. C'est cette matrice avec l'identification des échantillons comme colonnes et le nom gènes en lignes remplie par comptes bruts qui sera utilisée comme donnée d'entrées pour notre analyse RNA-Seq.

2.2.2 Analyse des données RNA-Seq avec DESeq2

L'outil DESeq2 est utilisé pour la normalisation des comptes bruts grâce à la normalisation d'expression logarithmique relative (RLE) ainsi que la transformation stabilisatrice de variance (VST). DESeq2 est aussi utilisé pour faire l'analyse d'expression différentielle et de faire la sélection des gènes les plus différentiellement exprimés [28]. Finalement nous utilisons l'algorithme de regroupement hiérarchique de l'outil *pheatmap* avec une visualisation du profil d'expression par carte thermique.

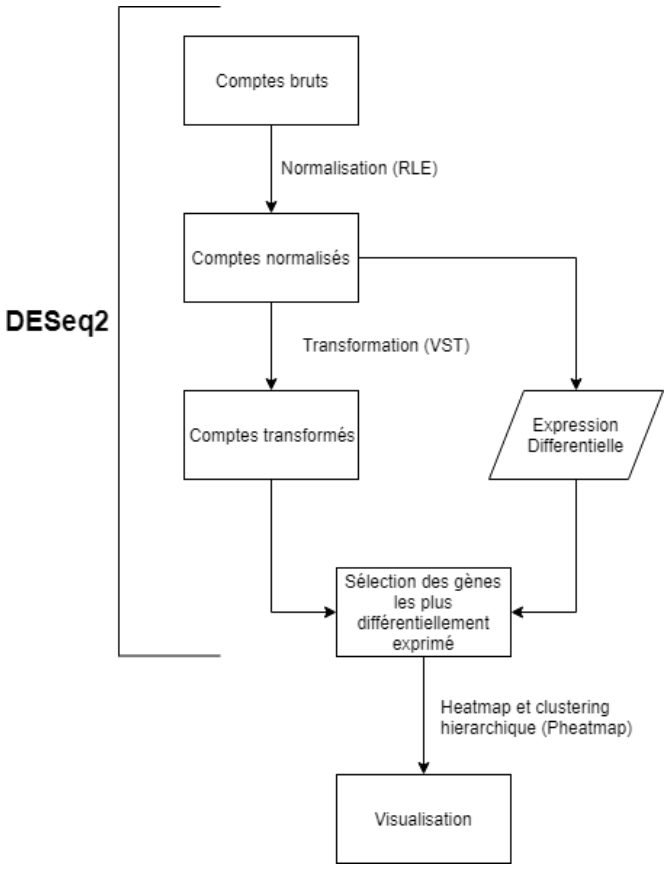

Figure 9. – Pipeline DESeq2 de l'analyse de données RNA-Seq.

2.2.2.1 Normalisation RLE (*Relative Log Expression*)

La normalisation que nous allons utiliser est la normalisation RLE fournie par l'outil DESeq2. RLE permet une normalisation en calculant un facteur de taille pour chaque échantillon permettant de rectifier les différences de profondeur de séquençage entre les échantillons. Le calcul de ces facteurs de taille est fait grâce à la méthode du ratio médian décrite par l'équation ci-dessous [29]:

$$\hat{s}_j = \underset{i}{\text{median}} \frac{K_{ij}}{(\prod_{v=1}^{m} K_{iv})^{1/m}}$$

Où nous avons la matrice de taille n × m, avec les comptes k_{ij}, où i = 1, ..., n indexe les gènes et j = 1, ..., m indexe les échantillons. S_j est le facteur de taille calculé. Selon l'article, le dénominateur de cette équation peut être interprété comme un pseudo-échantillon de référence obtenu en prenant la moyenne géométrique entre tous les échantillons.

Les comptes normalisés peuvent donc finalement être obtenus en divisant les comptes bruts d'un échantillon par son facteur de taille. La normalisation est directement appliquée aux comptes bruts grâce à la fonction *DESeq*. Cette fonction effectue la normalisation en calculant les facteurs de taille avec la fonction *estimationSizeFactors*.

2.2.2.2 Transformation stabilisatrice de variance (VST)

Lors de l'application d'un algorithme d'apprentissage non supervisé tel que le regroupement hiérarchique, celui-ci fonctionne mieux pour des données homoscédastiques. Les données sont dites homoscédastique lorsque la variance attendue est approximativement la même pour différentes valeurs moyennes. Les données RNA-Seq sont hétéroscédasticités par nature, car lorsque nous avons des gènes hautement exprimés, la moyenne du nombre de lectures augmente ainsi que sa variance. Par exemple, si l'on effectue le regroupement hiérarchique directement sur une matrice de comptages de lectures normalisés par facteur de taille comme RLE, le résultat ne dépend généralement que des quelques gènes les plus fortement exprimés, car ils présentent les plus grandes différences absolues entre les échantillons. Comme solution, DESeq2 propose des transformations pour les données de comptage qui stabilisent la variance à travers la moyenne tel que la transformation stabilisatrice de variance (VST).

La VST est donc effectuée sur les comptes normalisés seulement lorsque ceux-ci vont être utilisés par un algorithme d'apprentissage non supervisé. La transformation est faite grâce à la fonction *varianceStabilizingTransformation* incluse dans le l'outil de DESeq2 qui calcule la VST à partir des estimations de dispersions obtenues grâce à la fonction *estimationDispersions* et transforme les comptes normalisés.

2.2.2.3 Expression différentielle

Afin d'estimer la variabilité moléculaire des échantillons de TW dans les sous-groupes potentiels, nous effectuons une analyse d'expression différentielle entre ces sous-groupes. Des approches statistiques approfondies ont été développées pour tester l'expression différentielle avec des données de puces à ADN, où les intensités des sondes à travers les réplicas peuvent être approximées par une distribution normale. En principe ces approches sont également applicables aux données d'ARN- Seq. Cependant des études ont démontré que l'hypothèse de Poisson ne

permet pas de prendre en compte la variabilité biologique, ce qui entraîne des taux de faux positifs élevés dus à une sous-estimation de l'erreur d'échantillonnage [30]. En effet, il a été démontré que les modèles de distribution binomiale négativent qui tiennent compte de la surdispersion, c'est-à-dire une variation plus grande que prévu par les variables aléatoires de Poisson, correspondent le mieux à la distribution du nombre de lectures sur les réplicas biologiques [31]. Cette approche est utilisée dans l'outil DESeq2 pour l'analyse d'ARN-Seq [32].

La distribution binomiale négative détermine ses paramètres uniquement avec la moyenne μ du nombre de lectures pour un gène donné et sa variance σ2. Cependant, le nombre d'échantillons ou de réplicas dans les ensembles de données est souvent trop petit pour estimer la moyenne et la variance de manière fiable pour chaque gène.

DESeq2 va estimer une moyenne grâce au facteur de taille calculé lors de la normalisation et à la valeur moyenne du nombre de lectures par gène selon la condition qui correspond au sous-groupe potentiel auquel les échantillons appartiennent. La variance est estimée avec la moyenne précédemment définie et la variance du nombre de lectures à travers tous les échantillons qui appartiennent à la même condition [29]. DESeq2 estime que la quantité de données disponibles dans des expériences typiques d'ARN-Seq est suffisamment importante pour permettre une estimation locale suffisamment précise de la variance. L'expression différentielle est faite grâce à DESeq2 et à la fonction globale *DESeq* de l'outil. La différence d'expression d'un gène entre les deux conditions est calculée par le logarithme du ratio des deux taux d'expression, cette valeur est généralement référencée par l'acronyme LFC pour *Log2Fold-Change*. Un LFC positive représente une surexpression et un LFC négative une sous-expression.

Un test de rapport de vraisemblance (LRT pour *likelihood ratio* en anglais) est effectué afin de calculer les p-value et p-value ajustées.

Le LRT est effectué en estimant deux modèles et en comparant l'ajustement d'un modèle à l'ajustement de l'autre. Dans le cas de notre étude, ceci serait le modèle avec les sous-groupes potentiels et le modèle sans sous-groupes potentiels. Le LRT va comparer les probabilités logarithmiques des deux modèles : si cette différence est statistiquement significative, alors le modèle le moins restrictif (celui avec le plus de variables) est censé mieux ajuster les données que

le modèle plus restrictif. Lorsque nous obtenons les vraisemblances logarithmiques des modèles, le LRT est assez facile à calculer. La formule de la statistique de LRT est :

$$LR = -2ln\left(\frac{L(m_1)}{L(m_2)}\right) = 2\big(loglik(m_2) - loglik(m_1)\big)$$

Où $L(m_x)$ indique la probabilité du modèle. $loglik(m_x)$ est donc la vraisemblance logarithmique des modèles. *m1* est le modèle le plus restrictif et *m2* est le modèle le moins restrictif [33].

Le test LRT est effectué grâce à l'outil *DESeq2* qui l'implémente directement, et le résultat est reporté dans un objet DESeqDataSet, avec toutes les valeurs de LFC, p-value et p-value ajustées pour chaque gène, qui est accessible grâce à la fonction *results*.

2.2.2.4 Sélection des gènes

Afin d'essayer d'identifier des profils d'expressions différents à travers les échantillons de TW nous effectuons une première sélection des gènes par rapport à la variance du taux d'expression génique entre les échantillons. Les 500 gènes montrant les plus fortes variances sont sélectionnés. Pour ce faire, nous calculons la variance de chaque ligne de notre matrice de comptes normalisés par DESeq2. Le nombre de 500 gènes est arbitrairement déterminé et l'impact de ce choix est exploré en appliquant la même approche pour une sélection de 50, 100, 1000 et 5000 gènes.

Une deuxième sélection est effectuée après l'analyse d'expression différentielle. Cette fois-ci, nous sélectionnons les 50 gènes les plus différentiellement exprimés entre les deux sous-groupes à l'étude. Cette sélection est faite grâce aux résultats de l'analyse d'expression différentielle qui nous permet de classer les gènes en ordre décroissant des valeurs absolues de LFC.

2.2.3 Regroupement hiérarchique et visualisation

Le regroupement hiérarchique est performé par la fonction *hclust* incluse dans l'outil *pheatmap*. Ce regroupement hiérarchique utilise l'algorithme UPGMA (*unweighted pair group method with arithmetic mean*).

UPGMA utilise la méthode agglomérative avec une fonction de distance qui calcule la moyenne de la distance entre chaque point du groupe et tous les points d'un autre groupe [34]. Prenons les groupes A et B de taille respective |A| et |B|, l'équation suivante montre la fonction de distance utilisée par UPGMA :

$$dist(A,B) = \frac{1}{|A| \times |B|} \sum_{x \in A} \sum_{y \in B} dist(x,y)$$

La visualisation du résultat du regroupement hiérarchique est faite grâce au l'outil *pheatmap* de R grâce à un dendrogramme ainsi qu'une carte thermique qui permet de visualiser le patron d'expression à travers les échantillons. Un regroupement hiérarchique est aussi effectué sur les gènes afin de les regrouper selon leurs taux d'expression simplement pour avoir une meilleure visualisation du patron d'expression.

Chapitre 3 – Résultats

3.1 Un potentiel sous-groupe

3.1.1 Observation d'un potentiel sous-groupe grâce aux algorithmes apprentissages non-supervisés

La première étape de notre analyse pour essayer d'identifier un sous-groupe moléculaire dans nos échantillons de TW et ceci grâce à une signature génique. Nous effectuons cette étape grâce à l'utilisation d'algorithmes d'apprentissages non supervisés, notamment le regroupement hiérarchique.

Nous devons d'abord faire une sélection des gènes afin de pouvoir retirer une signature génique. Nous avons choisi une approche qui sélectionne les gènes qui ont une forte variabilité d'expression à travers les 130 échantillons de TW. Ceci est effectué en calculant la variance du nombre de comptes normalisés par RLE et transformés par VST et en classant cette variance en ordre décroissant. Nous sélectionnons ensuite les 500 gènes, un nombre arbitrairement déterminé, qui ont les plus fortes variances.

La sélection des gènes faite, nous appliquons le regroupement hiérarchique grâce à la méthode de l'algorithme UPGMA. La visualisation de ce résultat est réalisée grâce à une carte thermique du taux d'expression pour les gènes sélectionnés, et un dendrogramme associé pour le regroupement hiérarchique des échantillons de TW (Figure 10).

Nous décidons aussi d'explorer l'impact du choix arbitraire du nombre de 500 gènes en effectuant la même analyse sur les 50, 100, 1000 et 5000 gènes les plus variables (Figure 11).

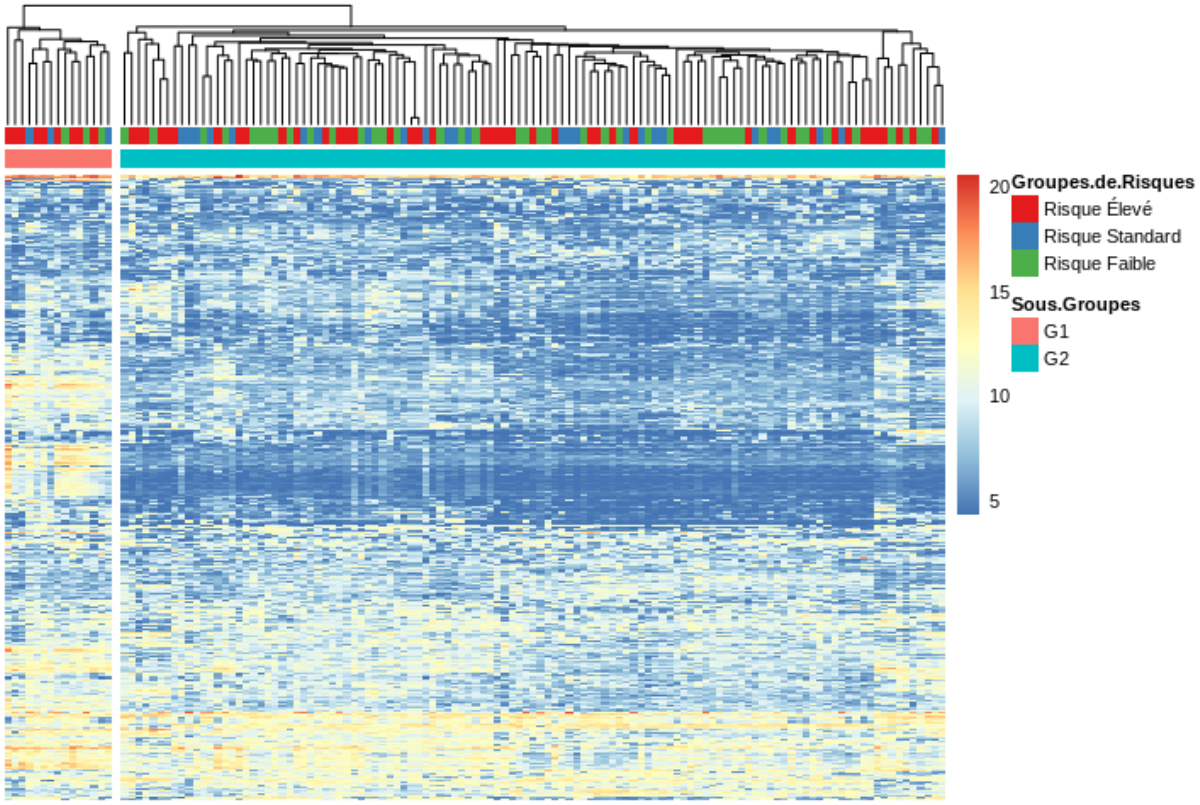

Figure 10. – Regroupement hiérarchique des 130 échantillons de TW selon les 500 gènes qui ont le taux d'expression plus variable à travers tous les échantillons. Le regroupement hiérarchique est réalisé par l'algorithme UPGMA implémenté par la fonction *hclust* de R avec le paramètre *average* pour la méthode. Le regroupement hiérarchique est accompagné d'une carte thermique du nombre de comptes normalisés par RLE et transformés avec une VST créée par l'outil *pheatmap*. La barre de température de la légende représente le nombre de comptes.

Nous observons dans cette première analyse deux sous-groupes distincts G1 et G2. Le sous-groupe G1 étant le plus petit contient 15 échantillons de TW. Ce sous-groupe inclut 9 TW à haut risque, 3 TW à risque standard et 3 TW à risque faible. G1 montre aussi un profil d'expression visuellement distinct comparé à G2.

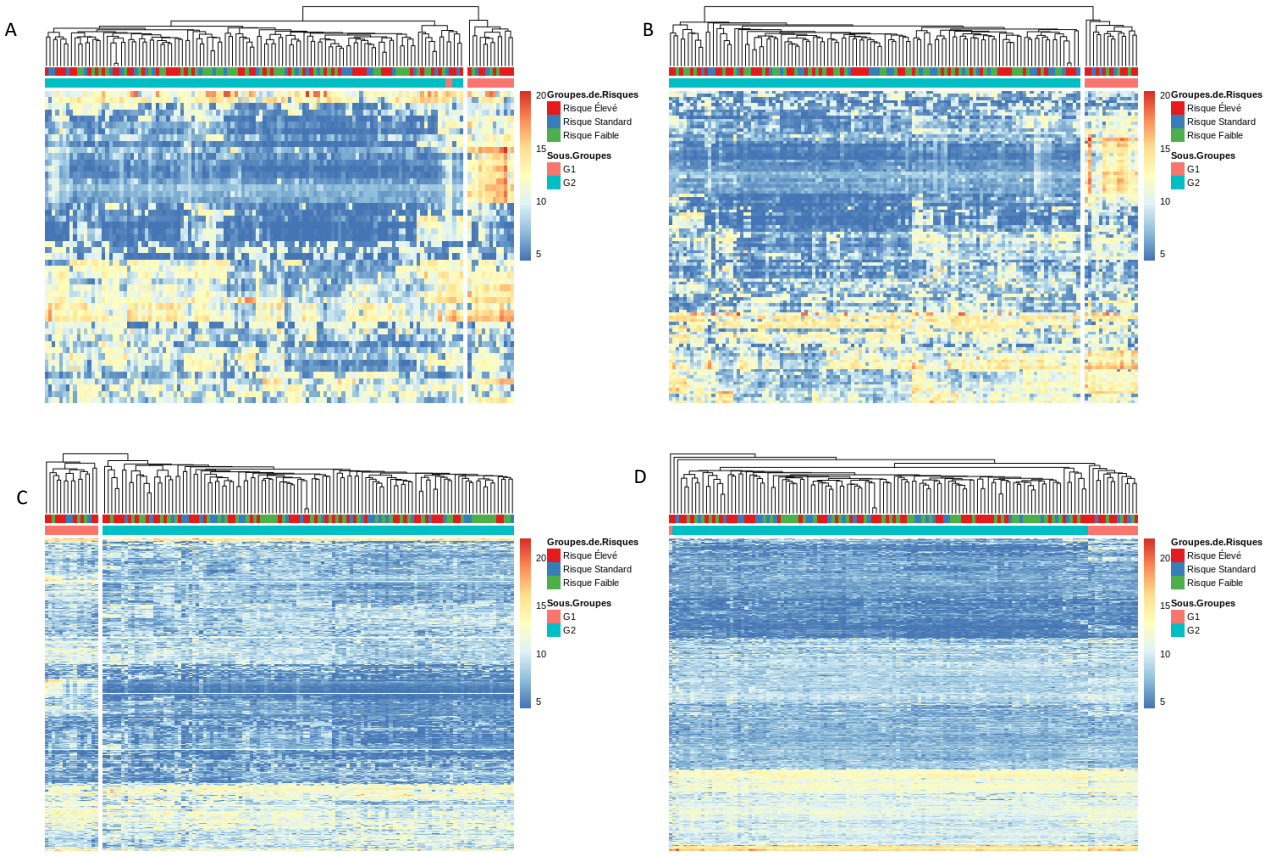

Figure 11. – Regroupement hiérarchique des 130 échantillons de TW selon les A) 50, B) 100, C) 1000 et D) 5000 gènes qui ont le taux d'expression plus variable à travers tous les échantillons. Le regroupement hiérarchique est réalisé par l'algorithme UPGMA implémenté par la fonction *hclust* de R avec le paramètre *average* pour la méthode. Le regroupement hiérarchique est accompagné d'une carte thermique du nombre de comptes normalisés par RLE et transformés avec une VST créée par l'outil *pheatmap*. La barre de température de la légende représente le nombre de comptes.

Cette figure représente la même analyse que celle réalisée pour la figure 10 avec un nombre de gènes sélectionné différent. On observe un bon regroupement lors de la sélection de 100 (figure 11B) et 1000 (figure 11C) gènes pour les TW du groupe G1 identifiés précédemment. Nous avons deux échantillons que ne se groupent pas avec le reste des échantillons de G1 lorsque nous avons une sélection de seulement 50 gènes (figure 11A). Le regroupement hiérarchique des 5000 gènes (figure 11D) ne permet pas de pouvoir découper l'arbre de façon à ce que tous les échantillons de G1 se trouvent dans le même groupe même s'ils semblent être dans un sous-groupe similaire.

Afin de confirmer ces résultats pour l'identification d'un sous-groupe G1 différent de G2, nous avons appliqué un autre algorithme d'apprentissage non supervisé. Nous avons utilisé une réduction de dimension par PCA (Figure 12).

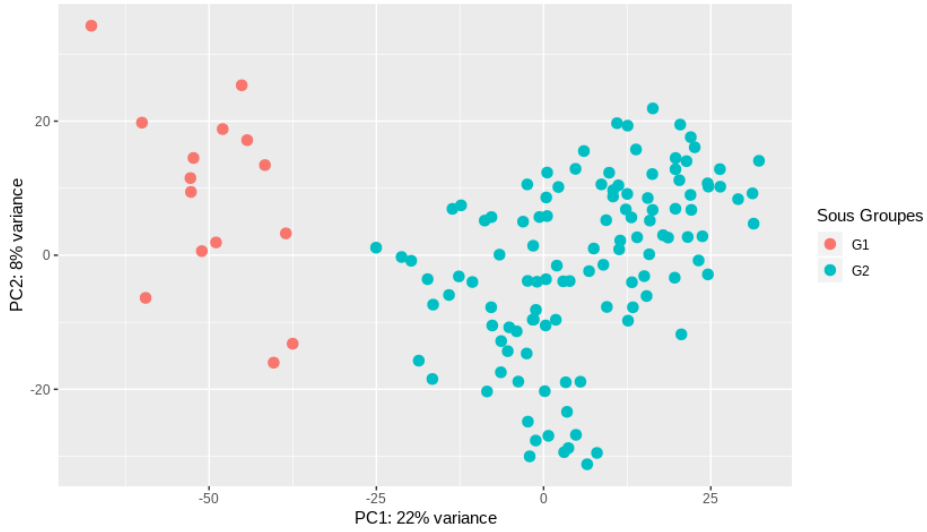

Figure 12. – Réduction de dimensionnalité par PCA et représentation graphique en 2 dimensions grâce aux deux premières PCs. Implémentation faite par l'outil DESeq2 avec la fonction *plotPCA* utilisant les comptes transformés par une VST.

On observe la séparation des deux sous-groupes, G1 et G2 précédemment identifiés, qui pourraient donc être de potentiels sous-groupes moléculaires de TW.

Afin d'étudier plus en profondeur la différence moléculaire de ces deux sous-groupes, nous allons ensuite effectuer une analyse par expression différentielle.

3.1.2 Expression différentielle entre les deux sous-groupes G1 et G2

L'expression différentielle permet de trouver les gènes qui ont un taux d'expression différent entre deux sous-groupes. Dans le contexte de nos deux sous-groupes G1 et G2, nous avons effectué l'analyse d'expression différentielle en comparant les taux d'expression génique de G1 comparés à G2. Ceci est réalisé grâce à un test de rapport de vraisemblance qui est implémenté dans l'outil *DESeq2*. Le test calcule les LFC et valeurs P pour chaque gène.

Dans un premier temps, nous pouvons sélectionner les gènes les plus différentiellement exprimés en classant ces gènes en fonction de la valeur absolue des LFC en ordre décroissant. Nous avons décidé de sélectionner les 50 gènes les plus différentiellement exprimés afin d'essayer de caractériser le sous-groupe G1. Nous pouvons observer ces gènes avec leur LFC et valeurs P dans le tableau 4 ci-dessous.

Dans un second temps, nous voulons observer l'effet de ces 50 gènes sur le regroupement des groupes G1 et G2. Pour ce faire, nous effectuons la même analyse de regroupement hiérarchique que dans la partie 3.1.1, mais cette fois-ci avec les 50 gènes sélectionnés. Ceci nous permet aussi de visualiser le taux d'expression de ces gènes à travers les échantillons des deux sous-groupes. (Figure 13).

Tableau 4. – LFC et Valeurs P pour les 50 gènes les plus différentiellement exprimés. Ces valeurs ont été calculées selon par un test LRT implémenté par l'outil de *DESeq2*.

NOM DU GENE	*LFC*	*VALEUR P*	*VALEUR P AJUSTEE*
MYOZ2	10.45238506	2.91E-119	3.96E-116
XIRP2	10.42226053	8.95E-70	4.86E-67
MYBPC1	10.37874815	1.82E-74	1.17E-71
MYOT	10.18381402	2.14E-64	8.99E-62
MYL1	10.04226253	2.80E-34	5.52E-32
MYL2	9.979898466	2.03E-31	3.62E-29
SMYD1	9.535749138	1.18E-49	3.60E-47
CAV3	9.529454754	9.38E-47	2.67E-44
HSPB3	9.487806873	6.49E-50	2.01E-47
FBP2	9.436188596	1.63E-73	9.90E-71
MYBPH	9.384658124	8.24E-70	4.55E-67
PPDPFL	9.353966168	1.27E-23	1.56E-21
LANCL1-AS1	9.333959142	1.35E-66	6.37E-64
CACNG1	9.286362788	4.26E-131	7.93E-128
APOBEC2	9.240329274	3.18E-48	9.27E-46
MRLN	9.086303065	3.41E-79	2.51E-76
LRTM1	9.057376124	1.74E-37	4.03E-35
MIR133A1HG	9.03363911	3.56E-87	3.07E-84
MYLPF	8.899506311	1.20E-191	1.06E-187
CKM	8.888613746	5.23E-141	2.05E-137
TTN	8.880693277	7.13E-200	1.26E-195
MYF6	8.794950291	1.32E-67	6.37E-65
C1ORF105	8.783840763	3.98E-139	1.30E-135
ACTA1	8.778377614	5.19E-93	5.09E-90
AC124301.1	8.596322758	1.16E-23	1.43E-21
MLIP	8.56085154	1.13E-76	7.70E-74
AL451062.1	8.530794707	1.87E-14	1.19E-12
FOXD3	8.528705967	1.61E-35	3.36E-33
CASQ2	8.511174147	1.57E-95	1.63E-92
ACTC1	8.491345303	1.39E-72	8.19E-70
MYF5	8.478269095	9.57E-24	1.19E-21
VGLL2	8.455446415	7.48E-68	3.72E-65
MYL4	8.385766786	4.28E-153	2.16E-149
AC087672.1	8.359201166	1.23E-45	3.46E-43
TECRL	8.349858671	1.21E-31	2.20E-29
COX6A2	8.346215635	1.63E-81	1.28E-78
RAPSN	8.346125509	4.03E-139	1.30E-135
LMOD3	8.344679616	2.30E-131	4.51E-128
NMRK2	8.315723356	2.39E-76	1.59E-73
PRKAG3	8.29555172	1.63E-53	5.49E-51
ABRA	8.276590285	2.21E-49	6.61E-47
MYMK	8.220894557	3.05E-71	1.74E-68
MYH8	8.191868394	1.09E-64	4.66E-62
TNNC2	8.109394155	2.04E-130	3.61E-127
MYH3	8.104041903	9.71E-256	3.43E-251
PPP1R3A	8.065777055	1.50E-21	1.64E-19
LINC01405	8.051660393	2.90E-50	9.06E-48
UNC45B	8.049334923	1.66E-138	4.88E-135
MYH7	8.029718139	2.75E-119	3.88E-116
ACTN2	8.027397536	8.76E-150	3.87E-146

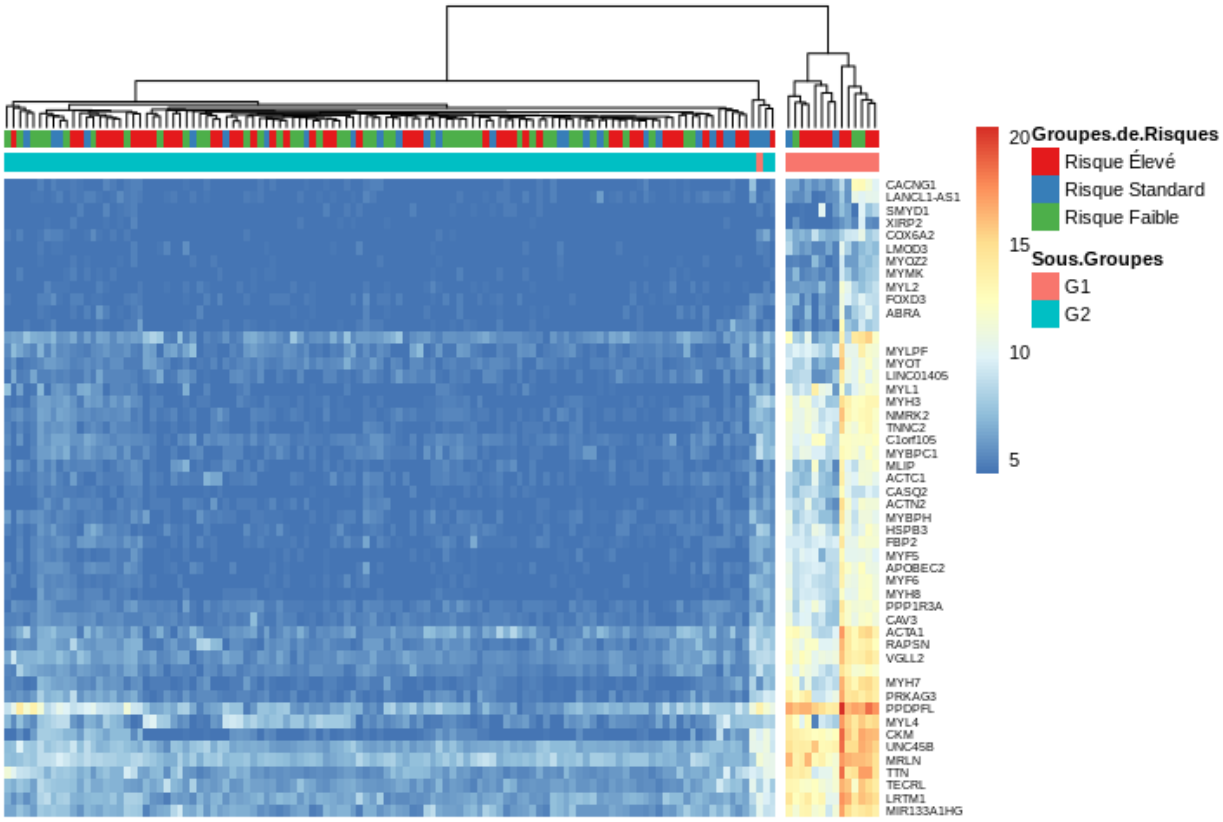

Figure 13. – Regroupement hiérarchique des 130 échantillons de TW selon les 50 gènes les plus différentiellement exprimés. L'expression différentielle a été effectuée par test LRT implémentée dans l'outil *DESeq2*. La sélection des gènes est faite grâce à leur classement par rapport aux LFC calculés. Le regroupement hiérarchique est réalisé par l'algorithme UPGMA implémenté par la fonction *hclust* de R avec le paramètre *average* pour la méthode. Le regroupement hiérarchique est accompagné d'une carte thermique du nombre de comptes normalisés par RLE et transformés avec une VST créée par l'outil *pheatmap*. La barre de température de la légende représente le nombre de comptes.

On observe dans cette figure un regroupement des deux sous-groupes attendus à l'exception d'un échantillon de G1 qui se retrouve dans G2.

Dans le groupe G1, nous pouvons aussi voir la formation de ce qui ressemble à deux petits sous-groupes avec des taux d'expression moins élevés dans un que dans l'autre. Cependant ces deux sous sous-groupes ne semblent pas correspondre particulièrement à un groupe de risque spécifique.

Il semble également apparaître que les gènes sélectionnés comme étant les plus différentiellement exprimés seraient surexprimés au sein du groupe G1 comparé au groupe G2.

Afin de confirmer cette observation et de vérifier que ceci n'est pas dû à une erreur dans l'analyse, nous avons représenté tous les LFC dans un graphique selon la moyenne des comptes normalisés par gènes (Figure 14).

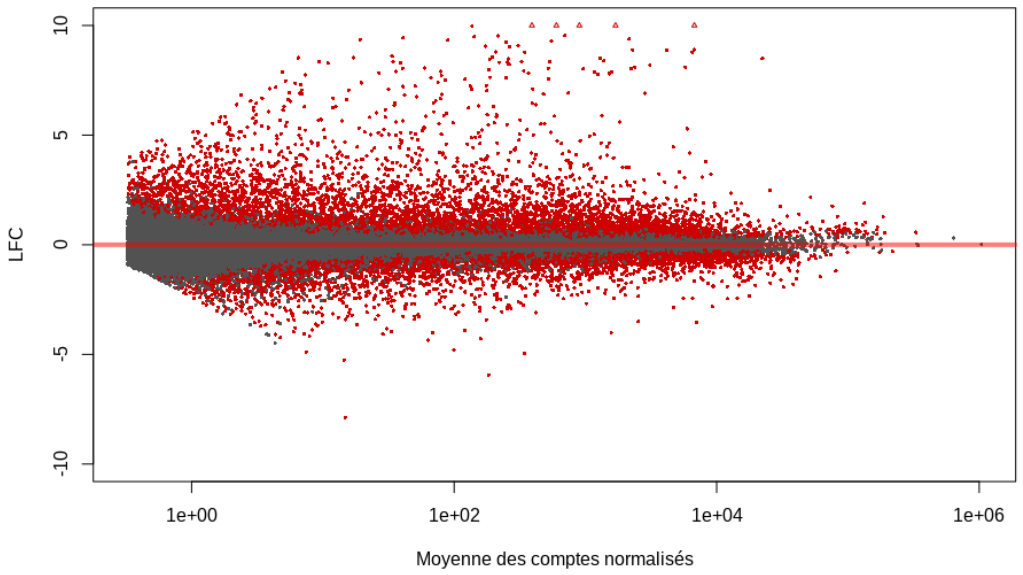

Figure 14. – Représentation du LFC par rapport à la moyenne des comptes normalisés montrant une forte présence de gènes surexprimés par rapport aux gènes sous-exprimés. Graphique effectuer grâce à la fonction *plotMA* incluse dans l'outil *DESeq2*. Les LFC sont calculés grâce au test LRT implémenté dans l'outil *DESeq2*.

On peut voir que nous avons une forte distribution de gènes qui aurait une surexpression significative comparée à un plus bas nombre de gènes avec une sous-expression significative. On peut aussi voir que nous avons beaucoup de gènes avec une forte surexpression, ce qui explique pourquoi nous obtenons seulement des gènes surexprimés lors de notre sélection des gènes les plus différentiellement exprimés.

Nous avons décidé d'aussi étudier le taux d'expression des gènes WT1 et WT2 entre les deux sous-groupes G1 et G2. Pour ceci, nous avons effectué une représentation du nombre de comptes normalisés dans les échantillons de TW par rapport à leur appartenance à un sous-groupe.

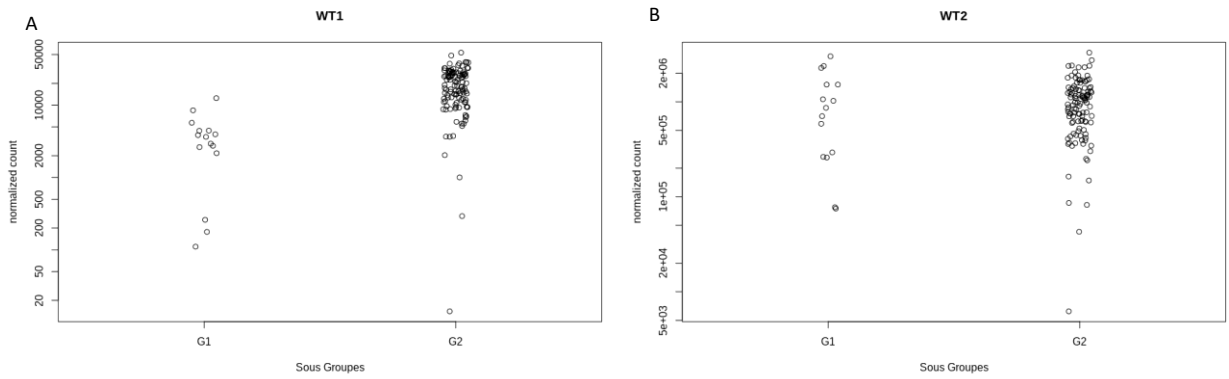

Figure 15. – Représentation graphique du taux d'expression pour les gènes A) WT1 et B) WT2 chez les différents échantillons de TW selon leur sous-groupe. Les figures sont faites grâce à la fonction *plotCounts* de DESeq2 avec les comptes normalisés par RLE.

Ces graphiques montrent une sous-expression de WT1 (Figure 15A) dans le groupe G1 et une expression similaire de WT2 (Figure 15B).

Finalement nous avons voulu regarder les processus biologique associés au gènes présent dans la liste du tableau 4. Une grande majorité des processus biologique était relies aux tissus musculaires (voir annexe 1). Les processus biologiques qui porte au développement et différentiation des tissus musculaire sont répertorier dans le tableau 5 ci-dessous. Nous avons aussi représenté le taux d'expression des gènes MYH3 et TTN dans la figure 16 ci-dessous où on observe une forte surexpression de ces deux gènes dans le groupe G1.

Tableau 5. – Processus biologiques de différenciation et développement de tissus musculaires auxquels sont associés les 50 gènes les plus différentiellement exprimés entre G1 et G2. Analyse faite avec l'outil *DAVID* [35] et les termes de *Gene Ontology*.

NOMS DES GENES	PROCESSUS BIOLOGIQUES ASSOCIES
CASQ2 **MYH7** **TTN**	Développement du tissu musculaire squelettique
SMYD1 **MYH5** **MYH6**	Régulation positive de la différenciation des myoblastes
CAV3 **MYHF5** **MYH3** **UUNC45B**	Développement des organes musculaires
SMYD1 **MYF5** **MYF6**	Différenciation des cellules musculaires squelettiques
MYF5 **MYF6**	Régulation positive de la différenciation des cellules musculaires

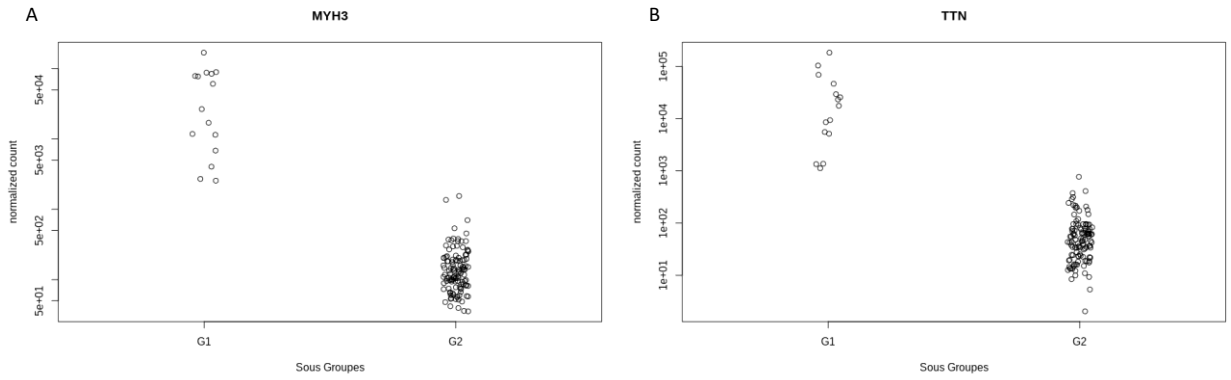

Figure 16. – Représentation graphique du taux d'expression pour les gènes A) MYH3 et B) TTN chez les différents échantillons de TW selon leur sous-groupe. Les figures sont faites grâce à la fonction *plotCounts* de DESeq2 avec les comptes normalisés par RLE.

3.2 Exploration du sous-groupe G2

Le groupe G2 regroupant 115 échantillons, nous avons voulu étudier si la même analyse que dans la partie 3.1.1 en retirant les échantillons du groupe G1 mènerait à l'identification de sous-groupes à l'intérieur de G2. Nous pouvons voir les résultats de cette analyse dans la figure 17 et figure 18 ci-dessous.

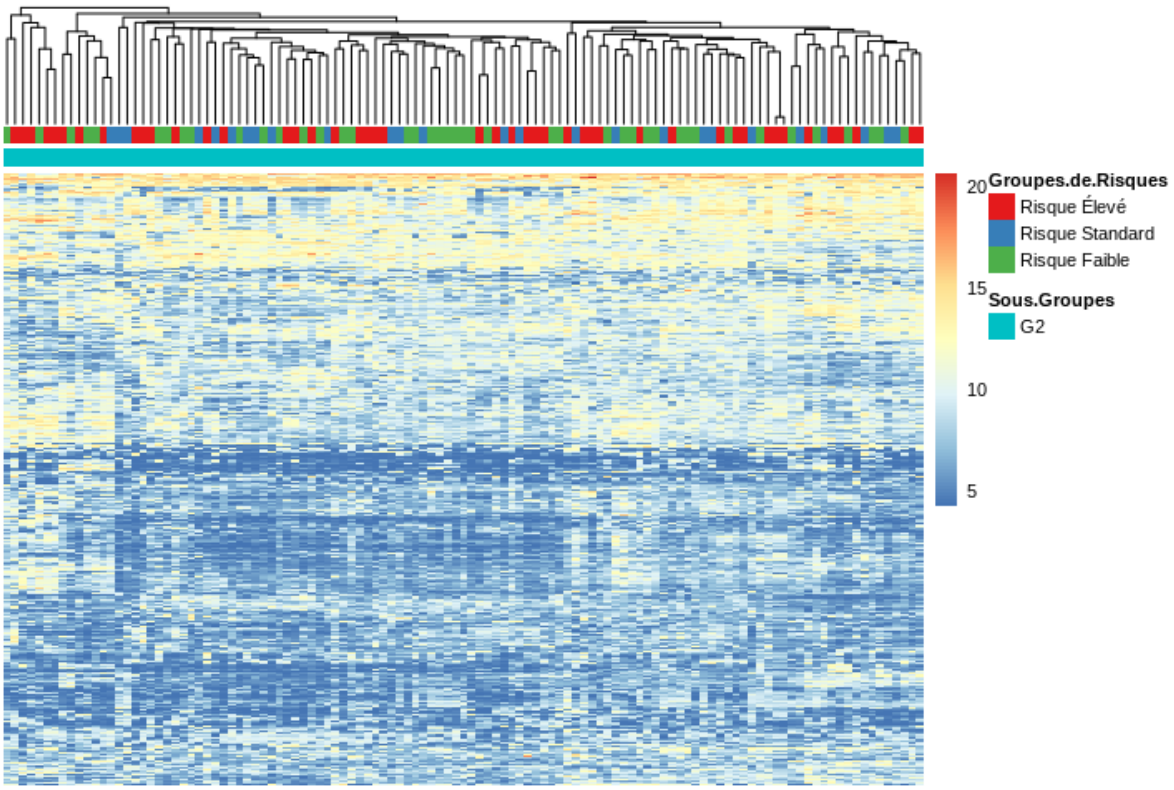

Figure 17. – Regroupement hiérarchique des 130 échantillons de TW selon les 500 gènes qui ont le taux d'expression plus variable à travers tous les échantillons. Le regroupement hiérarchique est fait par l'algorithme UPGMA implémenté par la fonction *hclust* de R avec le paramètre *average* pour la méthode. Le regroupement hiérarchique est accompagné d'une carte thermique du nombre de comptes normalisés par RLE et transformés avec une VST créée par l'outil *pheatmap*. La barre de température de la légende représente le nombre de comptes.

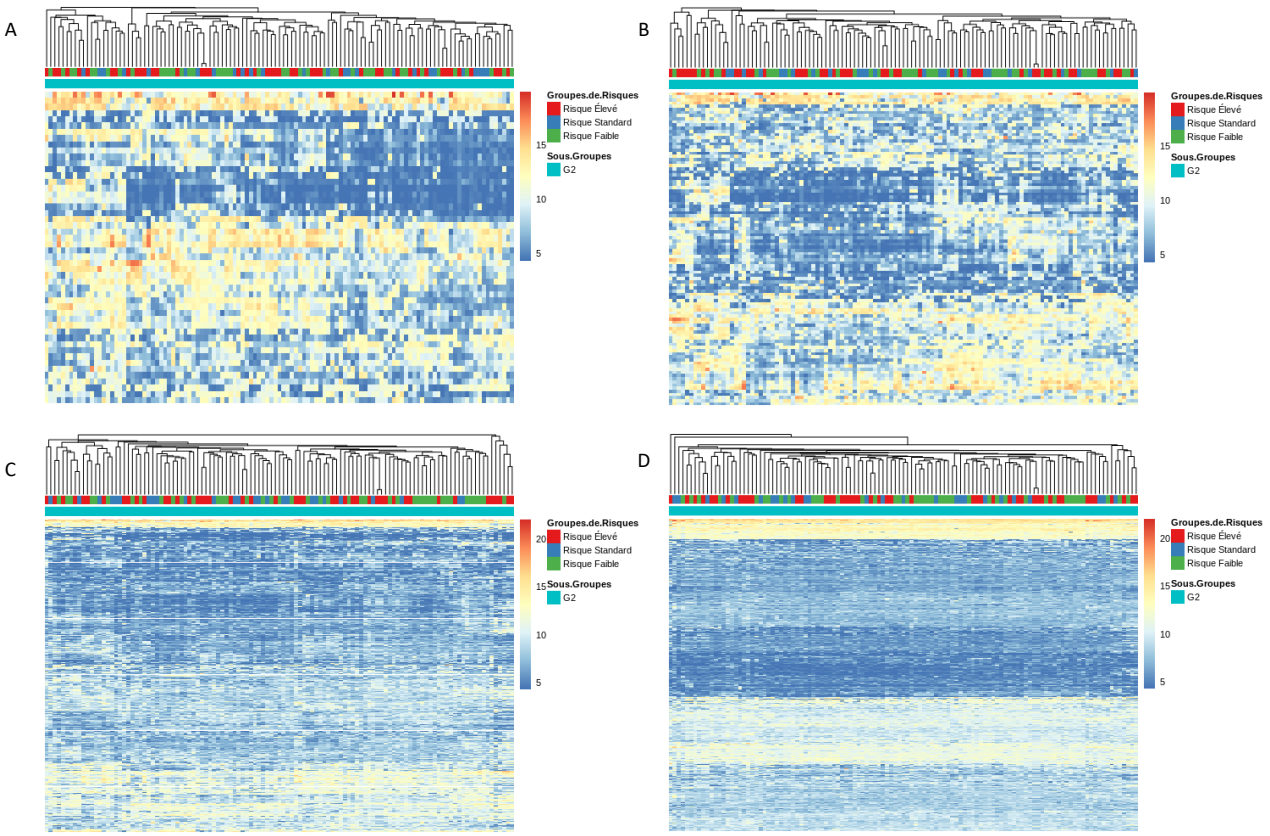

Figure 18. – Regroupement hiérarchique des 130 échantillons de TW selon les A) 50, B) 100, C) 1000 et D) 5000 gènes qui ont le taux d'expression plus variable à travers tous les échantillons. Le regroupement hiérarchique est fait par l'algorithme UPGMA implémenté par la fonction *hclust* de R avec le paramètre *average* pour la méthode. Le regroupement hiérarchique est accompagné d'une carte thermique du nombre de comptes normalisés par RLE et transformés avec une VST créée par l'outil *pheatmap*. La barre de température de la légende représente le nombre de comptes.

Nous n'observons pas de sous-groupes évidents sur ces deux figures. Le regroupement hiérarchique ne permet pas de découper le groupe G2 en plus petits groupes et le profil d'expression semble être homogène à travers le groupe G2.

Chapitre 4 - Discussion

4.1 Une approche d'identification de sous-groupes moléculaires

4.1.1 Identification de sous-groupes potentiels

Le premier objectif de cette étude était d'élaborer une méthode permettant de trouver des profils d'expression génique qui permettrait d'identifier des sous-groupes moléculaires potentiels.

En utilisant les données de RNA-Seq provenant de 130 échantillons de la cohorte, nous avons décidé de faire une sélection de gènes qui pourrait permettre d'observer différents profils d'expression. Cette sélection est faite avec les gènes qui montraient une forte variance du nombre de comptes à travers les échantillons. Cette méthode permet de rapidement observer le taux d'expression de gènes qui sont susceptibles de différentier le profil d'expression de chaque échantillon. Cependant, pour ce faire nous avons dû déterminer un nombre de gènes qui pourrait correctement décrire le profil d'expression de ces échantillons. Un nombre trop bas pourrait exclure certains gènes clés à la définition du profil d'expression, et un nombre trop grand pourrait noyer le signal de ces gènes clés. Nous observons ces deux extrêmes dans les figures 10A) et 10B) respectivement. Lorsque nous avons un petit nombre de gènes sélectionnés, deux échantillons sont exclus des sous-groupes. Dans le cas d'une sélection trop grande, nous observons une homogénéisation du profil d'expression des échantillons. La décision de prendre les 500 gènes présentant les plus fortes variations du nombre de comptes a donc été faite, car ceci semble être un juste milieu entre les deux extrêmes et permettre la mise en évidence la plus claire des deux sous-groupes.

L'objectif subséquent était de déterminer un regroupement des échantillons montrant un profil similaire grâce aux différents profils d'expressions des échantillons. Nous avons approché cet objectif en utilisant des algorithmes d'apprentissage non supervisé, notamment le regroupement hiérarchique. Un des défis rencontrés lorsqu'on utilise le regroupement hiérarchique est la sélection du niveau auquel on va couper le dendrogramme afin de former des sous-groupes viables [36]. Ceci est résolu en couplant le résultat du regroupement hiérarchique avec une carte

thermique du taux d'expression ce qui permet de visuellement confirmer un bon découpage des sous-groupes selon un profil d'expression distinct entre eux.

Le résultat du regroupement hiérarchique des échantillons par rapport aux taux d'expression des 500 gènes les plus variables est montré dans la figure 9. Nous pouvons y observer la division du dendrogramme résultant du regroupement hiérarchique en deux sous-groupes présentant des profils d'expression visuellement différents. Nous procédons donc en nommant ces deux sous-groupes potentiels G1 et G2.

Afin d'essayer de confirmer la validité de la création de ces deux sous-groupes, nous procédons à l'utilisation d'un autre type d'apprentissage non supervisé, la réduction de dimensionnalité, en appliquant la méthode de PCA. On observe la bonne séparation des deux sous-groupes dans l'espace de 2 dimensions créées par les deux premières dans la figure 11. Cette approche permet d'avoir une redondance du résultat de regroupement, mais aussi de confirmer le bon découpage du dendrogramme du regroupement hiérarchique autrement qu'avec la visualisation par carte thermique du profil d'expression.

Notre approche d'identification permet donc d'obtenir des sous-groupes potentiels de TW uniquement grâce aux profils d'expressions des échantillons de ces TW sans les données des caractéristiques externes. Cependant, il faut par la suite pouvoir caractériser ces sous-groupes afin de comprendre pourquoi ils ont un profil d'expression différent.

4.1.2 Caractérisation des sous-groupes potentiels

L'objectif de caractériser les sous-groupes potentiels identifiés par la méthode décrite précédemment est fait grâce à l'analyse par expression différentielle des différents sous-groupes. Pour cela, nous avons utilisé l'outil DESeq2 afin de comparer l'expression entre les deux sous-groupes G1 et G2. Une sélection des 50 gènes les plus différentiellement exprimés est effectuée afin de caractériser le sous-groupe G1.

Dans un premier temps, nous avons voulu visualiser le taux d'expression de ces gènes à travers nos échantillons de TW, et on remarque que tous ces gènes sont surexprimés dans le sous-groupe G1. Ceci est collaboré par la figure 13 qui montre une forte représentation de gènes surexprimés

avec de grand LFC. On remarque dans la liste des 50 gènes les plus différentiellement exprimés, décrite dans le tableau 4, la présence des gènes importants relatifs au début du développement musculaire tel que MYH3 et TTN dont la présence était rapportée dans le l'étude décrite dans la partie 1.2 [15].

Cette dernière observation nous a menés à regarder les processus biologiques reliés au développement et à la différentiation de cellule musculaire associés aux gènes que nous décrivons dans le tableau 4. Ce résultat est décrit dans le tableau 5 et montre en effet des processus tels que la régulation positive de cellules musculaire et des myoblastes, qui sont des cellules embryonnaires précurseurs aux cellules musculaires.

Ces analyses nous permettent de proposer un profil des caractéristiques du sous-groupe G1. Les TW du sous-groupe G1 seraient des TW avec une prédominance de tissus stromals comprenant une majorité de tissus musculaires comme décrit dans la partie 1.1.3.3 [14]. Cette hypothèse est soutenue par les évidences de surexpression de gènes liés à des processus impliquant les tissus musculaires. La sous-expression de WT1 observée dans la figure 14A expliquerait aussi cette caractéristique avec les TW à prédominance de tissus stromals caractérisées par la perte de WT1 expliquant leur mauvaise différentiation (voir partie 1.1.3.3). Les tumeurs présentant une prédominance de tissus stromals peuvent être héréditaires et ne répondent pas bien à la chimiothérapie par comparaison aux autres TW [14]. Ceci, couplé avec la forte représentation de TW du groupe de risque élevé pourrait mettre en évidence un sous-groupe qui correspondrait à un risque accru de rechute En effet, 9 échantillons sur les 15 présents dans le sous-groupe G1 correspondent à une TW à risque élevé. Ceci représente 60% des échantillons comparé à la proportion de 42% des TW à risque élevé dans notre ensemble de données (voir figure 8B). Ces caractéristiques du sous-groupe G1 présentent aussi de fortes ressemblances aux caractéristiques du groupe S2 de l'étude présentée dans la partie 1.2 [15]. Les échantillons de TW utilisés dans cette étude proviennent majoritairement de patients ayant fait une rechute, apportant ainsi une autre évidence au fait que le sous-groupe G1 pourrait correspondre des TW à risque de rechute accru.

4.2 Méthodes alternatives pour l'analyse d'expression différentielle.

Dans le cadre de cette étude, nous avons utilisé l'outil de DESeq2 pour réaliser la grande majorité des étapes nécessaires pour l'analyse par expression différentielle. Cependant il existe d'autres outils qui permettent l'analyse d'expression différentielle des données de RNA-Seq telle que edgeR [37] et limma-voom [38]. EdgeR et DESeq2 sont similaires, car elles utilisent toutes les deux une approche par distribution binomiale négative pour modéliser la différence d'expression génique. La différence majeure entre les deux outils est leurs approches pour normaliser les comptes bruts. EdgeR utilise la méthode TMM (*Trimmed Mean of M-value*) qui sélectionne un échantillon de référence qui va être utilisé par la suite afin de calculer tous les facteurs de normalisation pour tous les autres échantillons [39]. Limma-voom utilise une approche complètement différente pour déterminer la différence d'expression génique : l'outil applique des méthodes statistiques développées pour les puces à ADN sur les données RNAseq.

L'outil DESeq2 a été choisi par commodité. Cependant, une étude comparant les différents outils d'analyse différentielle a conclu que edgeR et limma-voom surpasse les autres méthodes dont la méthode de DESeq2 [40].

Dans cette optique une étude de comparaison des résultats obtenus dans le cadre de ce mémoire et les résultats d'une recréation de l'approche décrite ici avec les outils edgeR et limma-voom pourrait être envisagée.

4.3 Limitations

La plus grande limitation de mon approche est la dépendance de celle-ci à l'identification de profils d'expression grâce à la variance du nombre des comptes par gènes à travers les échantillons. En effet lorsque nous ne pouvons pas identifier ces profils d'expression, comme on peut le remarquer avec l'application de l'approche sur le groupe G2, nous ne pouvons pas procéder à l'extrapolation et la catégorisation de sous-groupes.

Cette limitation est amplifiée par l'utilisation d'une exploration de potentiels sous-groupes par regroupement hiérarchique. L'utilisation d'une plus grande variété d'algorithmes de regroupement basé sur l'apprentissage machine avec une sélection de gènes plus permissifs pourrait atténuer ce problème. On pourrait, par exemple, utiliser des algorithmes de réduction de dimensionnalité plus sophistiquée telle que t-SNE [41] ou même des réseaux de neurones artificiels en utilisant un auto-encodeur [42].

Le t-stochastic neighborhood embedding t-SNE, est une méthode non-linéaire de réduction de dimensionnalité. T-SNE minimise la divergence de Kullback-Leibler entre deux distributions, c'est-à-dire une minimisation de la dissimilarité entre les deux distributions. Une distribution qui mesure les similitudes par paires d'échantillons d'entrée en haute dimension et une autre distribution de loi de Student qui mesure les similitudes par paires d'échantillons correspondants dans un espace de 2 ou 3 dimensions [43].

Les auto-encodeurs sont une architecture de réseau de neurones artificiels, qui tente de reconstruire les données d'entrée avec un encodage sur une couche de neurones cachée de plus petite taille que les données d'entrée. Il optimise cet encodage en comparant le résultat de la reconstruction avec les données d'entrée [44].

Cependant, dans notre cas, ce manque d'identification de profil d'expression par la sélection des gènes montrant la plus forte variance pourrait aussi être dû à l'homogénéité du profil d'expression des TW. Ceci pourrait être étudié en appliquant l'approche à d'autres types de cancers.

4.4 Conclusion et perspective

Pour conclure, notre étude décrit une approche utilisant des données d'ARN-Seq et des outils d'apprentissage machine non superviser qui permet d'identifier des potentiels sous-groupes moléculaires en utilisant uniquement le profil d'expression des échantillons. Nous avons réussi à identifier et à caractériser un sous-groupe de TW nommé G1 avec des indications que ce sous-groupe pourrait être associé avec un risque accru de rechute.

Cependant, l'importance de ce sous-groupe reste à être démontrée. En effet, ce profil reste préliminaire et des analyses supplémentaires doivent être réalisées afin de confirmer ses caractéristiques. Par exemple, des analyses sur les types de cellules retrouvées dans les échantillons de G1 pourraient être faites pour confirmer la présence de cellules de type musculaire. De plus une étude de survie sans rechute et globales sur des TW qui correspondent aux caractéristiques de G1 permettrait de déterminer le taux de risque que présente ce sous-groupe comparé aux autres TW.

Références bibliographiques

1. Society, Canadian Cancer Society's Advisory Committee on Cancer Statistics., *Canadian Cancer Statistics 2017.* 2017.
2. Charlton, J., et al., *Bilateral Wilms tumour: a review of clinical and molecular features.* Expert Rev Mol Med, 2017. **19**: p. e8.
3. Beckwith, J.B., *Nephrogenic rests and the pathogenesis of Wilms tumor: developmental and clinical considerations.* Am J Med Genet, 1998. **79**(4): p. 268-73.
4. Davidoff, A.M., *Wilms' tumor.* Curr Opin Pediatr, 2009. **21**(3): p. 357-64.
5. Dome, J.S., E.J. Perlman, and N. Graf, *Risk stratification for wilms tumor: current approach and future directions.* Am Soc Clin Oncol Educ Book, 2014: p. 215-23.
6. Dome, J.S., et al., *Treatment of anaplastic histology Wilms' tumor: results from the fifth National Wilms' Tumor Study.* J Clin Oncol, 2006. **24**(15): p. 2352-8.
7. Grundy, P.E., et al., *Loss of heterozygosity for chromosomes 16q and 1p in Wilms' tumors predicts an adverse outcome.* Cancer Res, 1994. **54**(9): p. 2331-3.
8. Grundy, P.E., et al., *Loss of heterozygosity for chromosomes 1p and 16q is an adverse prognostic factor in favorable-histology Wilms tumor: a report from the National Wilms Tumor Study Group.* J Clin Oncol, 2005. **23**(29): p. 7312-21.
9. Hastie, N.D., *Wilms' tumour 1 (WT1) in development, homeostasis and disease.* Development, 2017. **144**(16): p. 2862-2872.
10. Carraro, D.M., R.F. Ramalho, and M. Maschietto, *Gene Expression in Wilms Tumor: Disturbance of the Wnt Signaling Pathway and MicroRNA Biogenesis*, in *Wilms Tumor*, M.M. van den Heuvel-Eibrink, Editor. 2016: Brisbane (AU).
11. Rapp, J., et al., *WNT signaling - lung cancer is no exception.* Respir Res, 2017. **18**(1): p. 167.
12. Dome, J.S. and M.J. Coppes, *Recent advances in Wilms tumor genetics.* Curr Opin Pediatr, 2002. **14**(1): p. 5-11.
13. Dao, D., et al., *Multipoint analysis of human chromosome 11p15/mouse distal chromosome 7: inclusion of H19/IGF2 in the minimal WT2 region, gene specificity of H19*

silencing in Wilms' tumorigenesis and methylation hyper-dependence of H19 imprinting. Hum Mol Genet, 1999. **8**(7): p. 1337-52.

14. Schumacher, V., et al., *Two molecular subgroups of Wilms' tumors with or without WT1 mutations.* Clin Cancer Res, 2003. **9**(6): p. 2005-14.
15. Gadd, S., et al., *Clinically relevant subsets identified by gene expression patterns support a revised ontogenic model of Wilms tumor: a Children's Oncology Group Study.* Neoplasia, 2012. **14**(8): p. 742-56.
16. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
17. Buermans, H.P. and J.T. den Dunnen, *Next generation sequencing technology: Advances and applications.* Biochim Biophys Acta, 2014. **1842**(10): p. 1932-1941.
18. Kukurba, K.R. and S.B. Montgomery, *RNA Sequencing and Analysis.* Cold Spring Harb Protoc, 2015. **2015**(11): p. 951-69.
19. Schmieder, R. and R. Edwards, *Quality control and preprocessing of metagenomic datasets.* Bioinformatics, 2011. **27**(6): p. 863-4.
20. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
21. Chandramohan, R., et al., *Benchmarking RNA-Seq quantification tools.* Annu Int Conf IEEE Eng Med Biol Soc, 2013. **2013**: p. 647-50.
22. Oshlack, A., M.D. Robinson, and M.D. Young, *From RNA-seq reads to differential expression results.* Genome Biol, 2010. **11**(12): p. 220.
23. Cook, C.E., et al., *The European Bioinformatics Institute in 2016: Data growth and integration.* Nucleic Acids Res, 2016. **44**(D1): p. D20-6.
24. Kimes, P.K., et al., *Statistical significance for hierarchical clustering.* Biometrics, 2017. **73**(3): p. 811-821.
25. Johnson, S.C., *Hierarchical clustering schemes.* Psychometrika, 1967. **32**(3): p. 241-54.
26. Fionn Murtagh, P.C., *Methods of Hierarchical Clustering.* 2011.
27. Ringner, M., *What is principal component analysis?* Nat Biotechnol, 2008. **26**(3): p. 303-4.

28. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
29. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
30. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biol, 2010. **11**(3): p. R25.
31. Wu, H., C. Wang, and Z. Wu, *A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data.* Biostatistics, 2013. **14**(2): p. 232-43.
32. Anders, S., et al., *Count-based differential expression analysis of RNA sequencing data using R and Bioconductor.* Nat Protoc, 2013. **8**(9): p. 1765-86.
33. Fox, J., *Applied regression analysis and generalized linear models*. 2016, Thousand Oaks, CA: SAGE.
34. Hua, G.J., et al., *MGUPGMA: A Fast UPGMA Algorithm With Multiple Graphics Processing Units Using NCCL.* Evol Bioinform Online, 2017. **13**: p. 1176934317734220.
35. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
36. Wild, D.J. and C.J. Blankley, *Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering.* J Chem Inf Comput Sci, 2000. **40**(1): p. 155-62.
37. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
38. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biol, 2014. **15**(2): p. R29.
39. Maza, E., *In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design.* Front Genet, 2016. **7**: p. 164.
40. Soneson, C. and M. Delorenzi, *A comparison of methods for differential expression analysis of RNA-seq data.* BMC Bioinformatics, 2013. **14**: p. 91.

41. Laurens van der Maaten, G.H., *Visualizing Data using t-SNE.* Journal of Machine Learning Research, 2008.
42. Urda, D., et al. *Deep Learning to Analyze RNA-Seq Gene Expression Data*. 2017. Cham: Springer International Publishing.
43. Li, W., et al., *Application of t-SNE to human genetic data.* J Bioinform Comput Biol, 2017. **15**(4): p. 1750017.
44. Xie, R., et al., *A deep auto-encoder model for gene expression prediction.* BMC Genomics, 2017. **18**(Suppl 9): p. 845.

Annexes

Category	Term	Count	%	PValue	Genes
GOTERM_BP_DIRECT	muscle filament sliding	12	26.66667	2.58E-21	MYL4, ACTA1, MYBPC1, ACTC1, MYH3, ACTN2, MYL1, MYL2, TNNC2, MYH8, MYH7, TTN
GOTERM_BP_DIRECT	muscle contraction	10	22.22222	2.22E-12	MYLPF, ACTA1, CAV3, MYL1, MYOT, MYH8, CACNG1, LMOD3, MYH7, TTN
GOTERM_BP_DIRECT	cardiac muscle contraction	7	15.55556	8.00E-10	MYL4, ACTC1, MYL1, MYL2, CASQ2, MYH7, TTN
GOTERM_BP_DIRECT	skeletal muscle contraction	4	8.888889	2.27E-05	MYH3, TNNC2, MYH8, MYH7
GOTERM_BP_DIRECT	sarcomere organization	4	8.888889	4.06E-05	MYH3, ACTN2, CASQ2, TTN
GOTERM_BP_DIRECT	skeletal muscle thin filament assembly	3	6.666667	5.23E-05	ACTA1, ACTC1, TTN
GOTERM_BP_DIRECT	negative regulation of potassium ion transmembrane transporter activity	3	6.666667	1.10E-04	CAV3, ACTN2, CASQ2
GOTERM_BP_DIRECT	skeletal muscle tissue development	4	8.888889	2.23E-04	MYLPF, MYF6, VGLL2, MYF5
GOTERM_BP_DIRECT	cardiac myofibril assembly	3	6.666667	3.42E-04	ACTC1, MYL2, TTN
GOTERM_BP_DIRECT	cardiac muscle tissue morphogenesis	3	6.666667	3.42E-04	ACTC1, XIRP2, TTN
GOTERM_BP_DIRECT	striated muscle contraction	3	6.666667	4.03E-04	CASQ2, MYH7, TTN
GOTERM_BP_DIRECT	regulation of the force of heart contraction	3	6.666667	8.77E-04	MYL4, MYL2, MYH7
GOTERM_BP_DIRECT	positive regulation of myoblast differentiation	3	6.666667	0.001073	MYF6, SMYD1, MYF5
GOTERM_BP_DIRECT	muscle organ development	4	8.888889	0.001146	UNC45B, MYH3, CAV3, MYF5
GOTERM_BP_DIRECT	ATP metabolic process	3	6.666667	0.002495	MYH3, MYH8, MYH7
GOTERM_BP_DIRECT	regulation of heart rate	3	6.666667	0.002652	CAV3, CASQ2, MYH7
GOTERM_BP_DIRECT	skeletal muscle cell differentiation	3	6.666667	0.00577	MYF6, SMYD1, MYF5
GOTERM_BP_DIRECT	detection of muscle stretch	2	4.444444	0.009259	CAV3, TTN
GOTERM_BP_DIRECT	muscle tissue morphogenesis	2	4.444444	0.01156	MYF6, MYF5
GOTERM_BP_DIRECT	mesenchyme migration	2	4.444444	0.01156	ACTA1, ACTC1
GOTERM_BP_DIRECT	muscle cell fate commitment	2	4.444444	0.013857	MYF6, MYF5
GOTERM_BP_DIRECT	cardiac muscle cell development	2	4.444444	0.016148	CAV3, ACTN2
GOTERM_BP_DIRECT	positive regulation of skeletal muscle fiber development	2	4.444444	0.018434	MYF6, MYF5
GOTERM_BP_DIRECT	positive regulation of myotube differentiation	2	4.444444	0.020715	CAV3, SMYD1
GOTERM_BP_DIRECT	regulation of striated muscle contraction	2	4.444444	0.020715	MYL2, MYBPH
GOTERM_BP_DIRECT	negative regulation of potassium ion transport	2	4.444444	0.02299	ACTN2, CASQ2
GOTERM_BP_DIRECT	negative regulation of protein localization to cell surface	2	4.444444	0.02299	CAV3, ACTN2
GOTERM_BP_DIRECT	heart contraction	2	4.444444	0.029786	ACTC1, MYL2
GOTERM_BP_DIRECT	myofibril assembly	2	4.444444	0.029786	MYOZ2, LMOD3
GOTERM_BP_DIRECT	actin filament-based movement	2	4.444444	0.038776	ACTC1, MYH3
GOTERM_BP_DIRECT	positive regulation of myoblast fusion	2	4.444444	0.038776	MYF6, MYF5
GOTERM_BP_DIRECT	positive regulation of muscle cell differentiation	2	4.444444	0.054313	MYF6, MYF5
GOTERM_BP_DIRECT	ventricular cardiac muscle tissue morphogenesis	2	4.444444	0.056513	MYL2, MYH7
GOTERM_BP_DIRECT	heart development	3	6.666667	0.067313	MYL2, XIRP2, SMYD1
GOTERM_BP_DIRECT	somitogenesis	2	4.444444	0.086788	MYF6, MYF5
GOTERM_BP_DIRECT	establishment of protein localization to plasma membrane	2	4.444444	0.093151	CAV3, ACTN2

Annexe 1 – Résultat de l'analyse des processus biologiques associés aux 50 gènes les plus différentiellement exprimées entre les sous-groupes G1 et G2. Analyse faites avec l'outil DAVID.