Integrative characterization of oncogenesis and immune response in sarcoma

Julien Vibert

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Institut Curie

Integrative characterization of oncogenesis and immune response in sarcoma (Oncogenèse et infiltrat immunitaire dans les sarcomes)

Soutenue par

Julien VIBERT

Le 23 novembre 2021

Ecole doctorale n° 582

Cancérologie, biologie, médecine, santé

Spécialité

Biochimie et biologie moléculaire

Composition du jury :

Jessica, ZUCMAN-ROSSI PU-PH, Paris Descartes	*Présidente*
Natalie, JÄGER PhD, DKFZ Heidelberg	*Rapporteure*
Aurélien, DE REYNIES PU-PH, Université de Paris	*Rapporteur*
Eliezer, VAN ALLEN MD-PhD, Dana-Farber Cancer Institute	*Examinateur*
Franck, TIRODE PhD, Université Claude Bernard Lyon 1	*Examinateur*
Jean-Philippe, VERT PhD, Google Mines-ParisTech PSL	*Examinateur*
Joshua, WATERFALL PhD, Institut Curie PSL	*Directeur de thèse*
Olivier, DELATTRE MD-PhD, Institut Curie PSL	*Directeur de thèse*

institutCurie | PSL

学而不思则罔，思而不学则殆。

论语，为政第二 (15)

Learning without thinking is useless, thinking without learning is dangerous.

The Analects of Confucius, chapter II (15)

Apprendre sans réfléchir est vain, réfléchir sans apprendre est dangereux.

Entretiens de Confucius, chapitre II (15)

Summary

Sarcomas are cancers of mesenchymal origin that comprise more than a hundred different entities. They are mostly rare diseases that occur at all ages, including in children and young adolescents. Due to their rarity and diversity, diagnosis is often missed or delayed. Prognosis is generally poor in cases of advanced or metastatic disease and most treatment approaches rely on unspecific and highly toxic chemotherapy. There is thus an unmet need to improve the diagnosis of sarcomas and develop novel therapeutic approaches for these diseases.

RNA sequencing (RNA-seq) is a promising approach for the diagnosis of sarcomas, especially for translocation-related sarcomas that are characterized by chromosome translocations giving rise to fusion genes, such as *EWSR1-FLI1* in Ewing sarcoma. Using RNA-seq data of sarcomas of patients profiled at the Institut Curie, I explored the transcriptomic landscape of sarcomas and used machine learning and deep learning techniques to predict sarcoma type based on RNA-seq. This work led to the development of a tool currently in use at the Institut Curie to predict the origin of cancers of unknown primary and improve the diagnosis and prognosis of individual patients in clinical practice.

Immunotherapy has revolutionized cancer care for the last decade, however it has had only limited success in sarcomas, supposedly because they are not “immunogenic”. Indeed, most sarcomas, especially translocation-related ones, have a very low tumor mutational burden, which is believed to be the main driving force in the generation of tumor neoantigens recognized by the immune system. To gain further insight into the potential of immune response in sarcoma, I characterized the immune microenvironment and lymphocyte repertoires of multiple types of sarcomas using RNA-seq of tumor samples. While most of them were indeed poorly infiltrated by cells of the immune system, there were some exceptions to this rule suggesting that immunotherapy should be considered in some cases.

Another promising finding for immunotherapy of sarcomas was the identification of novel tumor-specific transcripts in multiple types of translocation-related sarcomas. These “neotranscripts” were driven by their characteristic oncogenic chimeric transcription factors such as *EWSR1-FLI1* in Ewing sarcoma; some of them were found to be translated by ribosomes into peptides. Therefore, these may represent a source of tumor-specific public neoantigens for immunotherapies of these translocation-related sarcomas.

To characterize in detail the immune microenvironment and oncogenic processes of specific sarcomas, single-cell RNA-seq was performed for some of them, notably dedifferentiated liposarcomas (DDLPS). It revealed higher infiltration by immune cells in the dedifferentiated compartment of the tumor, but with more exhausted and immunosuppressive phenotypes. It also allowed to characterize the oncogenic processes of DDLPS and notably the relationship between dedifferentiated and well-differentiated cells inside the same tumor.

Altogether, this work opens perspectives to improve diagnosis and develop immunotherapies for sarcomas by: 1) defining a global transcriptomic landscape of sarcoma types and their associated microenvironment; 2) identifying novel transcriptional processes in translocation-related sarcomas with potential for generation of neoantigens for immunotherapy; 3) characterizing at the single-cell level oncogenic processes and immune microenvironment of one type of sarcoma (DDLPS); 4) resulting in the development of a classifier tool for diagnostic prediction used in clinical practice at the Institut Curie.

Résumé (en français)

Les sarcomes sont des cancers d'origine mésenchymateuse qui comprennent plus d'une centaine d'entités. Ce sont pour la plupart des maladies rares qui peuvent survenir à tout âge, y compris pendant l'enfance et la jeune adolescence. En raison de leur rareté et diversité, le diagnostic en est souvent erroné ou retardé. Le pronostic est généralement sombre dans les formes avancées et métastatiques, et la plupart des traitements reposent actuellement sur des chimiothérapies non spécifiques et très toxiques. Il y a donc un besoin urgent d'améliorer le diagnostic des sarcomes et développer de nouvelles approches thérapeutiques pour ces cancers.

Le séquençage de l'ARN (RNA-seq) est une technique prometteuse pour le diagnostic des sarcomes, notamment dans le cas des sarcomes liés à des translocations qui sont caractérisés par des translocations chromosomiques à l'origine de gènes de fusion, par exemple *EWSR1-FLI1* dans le sarcome d'Ewing. A l'aide de la base de données du RNA-seq de sarcomes de patients de l'Institut Curie, j'ai exploré le paysage transcriptomique des ces cancers et utilisé des techniques d'apprentissage machine (machine learning) et d'apprentissage profond (deep learning) pour prédire le type de sarcome à l'aide du RNA-seq. Ce travail a ensuite permis le développement d'un outil actuellement utilisé à l'Institut Curie pour prédire la tumeur d'origine de cancers de primitif inconnu et ainsi améliorer le diagnostic et le pronostic de patients en pratique clinique courante.

Au cours de la dernière décennie, l'immunothérapie a été à l'origine d'une révolution dans le traitement de multiples cancers. Cependant, elle n'a eu qu'un succès très limité dans les sarcomes qui sont généralement considérés comme des tumeurs non « immunogéniques ». En effet, la plupart des sarcomes, notamment liés aux translocations, ont une charge mutationnelle très faible. Or ce dernier facteur est considéré comme l'un des principaux générateurs de néoantigènes tumoraux qui servent de cible au système immunitaire. Pour étudier plus en détail la possibilité d'une réponse immunitaire dans les sarcomes, j'ai caractérisé le microenvironnement tumoral immunitaire et les répertoires lymphocytaires dans de nombreux types de sarcomes à l'aide du RNA-seq d'échantillons tumoraux. Bien que la plupart sont effectivement peu infiltrés par des cellules du système immunitaire, il existe des exceptions qui font penser que l'immunothérapie pourrait être efficace dans certains cas.

Une autre piste prometteuse pour l'immunothérapie des sarcomes a été l'identification de nouveaux transcrits spécifiques dans de nombreux types de sarcomes liés à des translocations. Ces « néotranscrits » sont induits par le facteur de transcription oncogénique chimérique caractéristique de la tumeur, par exemple *EWSR1-FLI1* dans le sarcome d'Ewing. Certains d'entre eux sont traduits par les ribosomes en peptides. Ils représentent donc une source potentielle de néoantigènes publics spécifiques de la tumeur pour les approches d'immunothérapie dans les sarcomes liés à des translocations.

Pour caractériser en détail le microenvironnement immunitaire et les processus oncogéniques de sarcomes spécifiques, certains d'entre eux ont été étudiés par du RNA-seq à l'échelle unicellulaire (single-cell RNA-seq), notamment les liposarcomes dédifférenciés (DDLPS). Cette technique a mis en évidence une infiltration plus importante de cellules immunitaires dans le compartiment dédifférencié de la tumeur, ainsi qu'un phénotype « épuisé » (exhausted) et immunosuppresseur de ces cellules. Elle a aussi permis de caractériser les processus oncogéniques des DDLPS, notamment la relation entre les cellules bien différenciées et « dédifférenciées » au sein d'une même tumeur.

Au total, ce travail ouvre plusieurs perspectives pour l'amélioration du diagnostic et le développement d'immunothérapies pour les sarcomes, en : 1) définissant un paysage transcriptomique global des types de sarcomes et de leur microenvironnement immunitaire ; 2) identifiant de nouveaux mécanismes transcriptionnels dans les sarcomes liés à des translocations potentiellement à l'origine de néoantigènes pour l'immunothérapie ; 3) caractérisant à l'échelle unicellulaire les processus oncogéniques et le

microenvironnement immunitaire d'un type de sarcome (liposarcome dédifférencié) ; 4) mettant en place un outil d'aide au diagnostic actuellement utilisé en pratique clinique courante à l'Institut Curie.

Remerciements

Il est difficile de remercier assez toutes les personnes qui de près ou de loin ont contribué à l'aboutissement de ce travail de thèse, qui s'inscrit dans un long parcours de formation médicale et scientifique débuté il y a treize ans. Je vais néanmoins tenter de leur adresser ma plus sincère gratitude, tout en sachant que je vais inévitablement en oublier certains… Je leur prie d'avance de m'en excuser.

Olivier Delattre et Joshua Waterfall : vous êtes de formidables mentors et j'ai appris énormément à vos côtés, vous êtes pour moi de grandes sources d'inspiration !

Nadège Gruel : j'ai été ravi de travailler avec toi, et partager le même bureau que toi lorsque j'étais sur place à Curie (malheureusement trop peu souvent après mars 2020) !

Sarah Watson : tu as été à l'origine de ma rencontre avec Olivier et Josh, j'ai été extrêmement content de pouvoir travailler avec toi en recherche après avoir été ton interne, et surtout tu es pour moi un modèle à suivre dans ce parcours de combattant du médecin-chercheur !

Agathe Peltier, Céline Collin, Charlie Buchou, Margot Gautier, Floriane Petit, Olivier Saulnier, Anne-Laure Begue, Dorian Bochaton, Jérômine Vigneau, Laury Poulain : étudiants présents et passés de l'équipe, vous avez été de superbes compagnons !

Sandrine Grossetête, Véronique Hill, Maud Gautier, Amira Kramdi, Chloé Quignot, Olivier Mirabeau, Karine Laud-Duval, Caroline Louis, Sakina Zaidi, Didier Surdez, Calvin Rodrigues, Cécile Thirant, Isabelle Janoueix, Cécile Pierre-Eugène, Carole Drique… et tous les membres présents et passés de l'équipe Delattre : j'ai été ravi d'être accueilli parmi vous !

Gaëlle Pierron, Delphine Guillemot, Camille Benoist, Eléonore Frouin, Julien Masliah-Planchon, Virginie Bernard et toute l'équipe de l'Unité de Génétique Somatique : j'ai apprécié travailler avec vous au cours de ma thèse !

Paul Gueguen, Mathias Vandenbogaert, Wilfrid Richer, Benjamin Sadacca, Mercia Ngoma… et tous les membres présents et passés de l'équipe Waterfall : j'ai beaucoup apprécié les interactions et échanges très enrichissants sur tous les sujets de bioinformatique !

Yann Kieffer, Fatima Mechta-Grigoriou, Franck Bourdeaut, Manuel Rodrigues, Yago Arribas de Sandoval, Marc Deloger, Nicolas Servant, Pacôme Prompsy… et tous les membres présents et passés de l'unité U830 et de l'Institut Curie : j'ai été ravi de pouvoir échanger avec vous sur de multiples sujets et bénéficier de vos conseils !

Clémence Hénon, Léo Colmet-Daage, et toute l'équipe de Sophie Postel-Vinay à Gustave Roussy : j'ai beaucoup apprécié travailler avec vous sur les DSRCT !

Nicolas Perrin : j'ai beaucoup apprécié échanger avec toi, je suis très heureux qu'un mathématicien comme toi puisse contribuer à faire avancer la recherche en oncologie !

Louis Jacob et toute l'équipe de l'unité Covid de l'hôpital Raymond-Poincaré de Garches : j'ai été très heureux de travailler chez vous lors de la première vague Covid !

Tous les patients, leurs médecins et soignants, nos collaborateurs ainsi que les généreux donateurs qui ont permis que ces travaux soient possibles, notamment la Ligue contre le cancer et son comité départemental de la Savoie : un grand merci !

Et bien entendu je pense tout particulièrement à mes parents, mes petites sœurs Amélie et Roseline, et tout le reste de ma famille, ainsi que tous mes amis qui m'ont soutenu tout au long de cette aventure !

Enfin, je tiens à remercier les membres de mon jury de thèse qui ont gentiment accepté d'en faire partie : Jessica Zucman-Rossi, Natalie Jäger, Aurélien de Reyniès, Eliezer Van Allen, ainsi que Franck Tirode et Jean-Philippe Vert qui m'ont également accompagné au cours de ma thèse lors des comités intermédiaires. Merci !

Acknowledgments (in English)

It is difficult to be exhaustive in thanking all the people who have contributed in one way or another to the completion of this work, which is the result of a long journey of medical and scientific education that began thirteen years ago. I will nonetheless try to tell them here my most sincere gratitude, with the caveat that I will inevitably omit some people… I beg them to forgive me for this.

Olivier Delattre and Joshua Waterfall: you are wonderful mentors and I have learnt a lot from you, you are great sources of inspiration for me!

Nadège Gruel: I was very happy to work with you, and share the same room as you when I was working physically in Curie (unfortunately much too rarely after March 2020)!

Sarah Watson: you introduced me to Olivier and Josh, I was extremely happy to be able to work with you in research after having been your resident in the clinic, you are for me a role model to follow in this difficult journey of becoming a physician-researcher!

Agathe Peltier, Céline Collin, Charlie Buchou, Margot Gautier, Floriane Petit, Olivier Saulnier, Anne-Laure Begue, Dorian Bochaton, Jérômine Vigneau, Laury Poulain : current and former students of the team, you have been wonderful partners!

Sandrine Grossetête, Véronique Hill, Maud Gautier, Amira Kramdi, Olivier Mirabeau, Karine Laud-Duval, Caroline Louis, Sakina Zaidi, Didier Surdez, Cécile Thirant, Isabelle Janoueix, Cécile Pierre-Eugène, Carole Drique… and all current and former members of the Delattre team: I was delighted to be part of your team!

Gaëlle Pierron, Delphine Guillemot, Camille Benoist, Eléonore Frouin, Julien Masliah-Planchon, Virginie Bernard and all the team from the Unité de Génétique Somatique: I appreciated to work with you during my PhD!

Paul Gueguen, Mathias Vandenbogaert, Wilfrid Richer, Benjamin Sadacca, Mercia Ngoma… and all current and former members of the Waterfall team: I really appreciated all the rich interactions and discussions that we had about bioinformatics!

Yann Kieffer, Fatima Mechta-Grigoriou, Franck Bourdeaut, Manuel Rodrigues, Yago Arribas de Sandoval, Marc Deloger, Nicolas Servant, Pacôme Prompsy… and all current and former members of the Unit U830 and the Institut Curie: I was delighted to exchange with you on many subjects and receive your advice!

Clémence Hénon, Léo Colmet-Daage, and all members of the team of Sophie Postel-Vinay in Gustave Roussy: I appreciated a lot to work with you on DSRCT!

Nicolas Perrin: I really appreciated to exchange with you, I am very happy that a mathematician like you can contribute to research in oncology!

Louis Jacob and all the team of the Covid unit of the Raymond-Poincaré hospital in Garches: I was very happy to work with you during the “first wave” of Covid!

All patients, physicians and healthcare workers, collaborators and generous funders who made this work possible, especially the Ligue contre le cancer and its committee from Savoie: thank you!

I am of course very grateful to my parents, my little sisters Amélie and Roseline, and all my family and friends who have supported me during this adventure!

Finally, I would like to thank all members of the jury that have kindly accepted the invitation to be part of it: Jessica Zucman-Rossi, Natalie Jäger, Aurélien de Reyniès, Eliezer Van Allen, as well as Franck Tirode and Jean-Philippe Vert who were also present at the committee meetings during my PhD. Thank you!

Table of contents

General introduction

Dear reader, you are about to read a manuscript that is built on the results of three years of PhD work that I had the opportunity to realize at the Institut Curie under the supervision of Joshua Waterfall and Olivier Delattre. As you will see, this work is not constituted by one unique project progressing linearly from start to end. Indeed, I got involved in multiple projects during these three years that allowed me to tackle many conceptual and technical questions. Nonetheless, all these revolved around some main common threads that were designed to be woven together into a coherent whole, as I will show you in the following text.

The resulting "fabric" is therefore heterogeneous and multi-colored: while some projects were planned at the beginning of my PhD, others were the results of an unexpected biological observation or a pressing clinical case, some are not even mentioned in this manuscript. This diversity may be justified by the large number of datasets that I was fortunate to have access to, and the myriad possibilities of analyses that can nowadays be performed in bioinformatics. However, the main reason was probably my own tendency to embrace many projects at the same time. Indeed, I was eager to learn as much as possible in bioinformatics, analyze as many datasets as were available to me, and answer all biologically and clinically relevant questions that came to my mind.

This may be due to my medical background: I am trained as a medical oncologist and my bioinformatics background is therefore not as strong as many of my scientist colleagues. Back in 2012, I had nevertheless completed my Master's degree in systems biology and published a model of lymphocyte dynamics using ordinary differential equations[1]. However, computational biology and especially bioinformatics evolved rapidly before I came back into science for my PhD at the end of 2018. One landmark change was the advent of next-generation sequencing (NGS) technologies[2], which notably allowed the large-scale genomic and transcriptomic profiling of cancer samples. Another was the rapid development of machine learning and deep learning methods for so-called "artificial intelligence"[3]. Considering that my medical specialty is oncology, I was particularly excited by the potential of these novel techniques to better understand biological mechanisms of cancer and help to find some cures for patients. I therefore decided to realize a PhD in bioinformatics to learn this expertise and be able to leverage the potential of these data to address fundamental biological and clinical issues in oncology. To achieve this, I was very fortunate to work at the Institut Curie under the supervision of Joshua Waterfall and Olivier Delattre. Not only was I allowed to analyze rich resources of precious transcriptomic data of patients, but I was also able to participate in many exciting projects taking place here. This diversity of learning and practice was exactly what I desired for my PhD as a medical oncologist endeavoring to get a solid grasp of a large panel of computational methods for analysis of biomedical data.

Outline of the manuscript

I will now present the outline of this manuscript and the underlying common threads to draw a coherent "fabric" out of all the different parts. Since each project could also be narrated in a self-limited way, I have chosen to focus here on the big picture, and to introduce and discuss more thoroughly each subject inside its corresponding part.

One of the main threads of my work is the study of **sarcomas**. Sarcomas are cancers of mesenchymal origin, i.e. derived from bone and soft tissues such as muscle, fat, and cartilage. They form a vast group of heterogeneous malignant tumors comprising more than a hundred histologically distinct entities[4]. All types of sarcomas are mostly rare diseases (fewer than one case in a population of 2000 individuals in Europe).

Consequently, they are less well studied and have fewer therapeutic options: most patients with metastatic disease are treated by highly toxic chemotherapy, often without significant benefit[5].

Besides their histological classification by pathologists, sarcomas are classically divided into two groups based on their genomic alterations: either “complex” or “simple”[6]. Sarcomas with a complex genomic profile are characterized by a high number of genetic alterations including mutations, chromosome translocations, gains and deletions. They are mostly occurring in persons older than 60 years old. In contrast, sarcomas with a simple genomic profile are mostly driven by one unique genetic alteration which is in many cases a chromosome translocation resulting in a fusion gene that has oncogenic properties. These are called translocation-related sarcomas and are more often occurring in younger patients, including children and young adolescents[7,8]. One paradigmatic example of translocation-related sarcomas is Ewing sarcoma, which occurs in children and adolescents and is the result of a translocation between chromosomes 11 and 22 resulting in the fusion gene *EWSR1-FLI1* (*EWS-FLI1*)[9].

One critical aspect in the clinical care of patients with sarcomas is the accuracy of diagnosis. Since these are rare diseases, they are often confused with other more frequent cancers if doctors in charge are not familiar with sarcomas. This is the reason why all patients with suspected or confirmed sarcoma should be taken in charge by multidisciplinary reference centers: a national study in France showed improved survival with this measure[10]. However, accurate diagnosis remains a challenge even in specialized centers. As there are more than a hundred histological entities, and a rapidly growing number of characterized translocations for translocation-related sarcomas, there is a need for precise, rapid, and correct diagnosis that should not leave out even the rarest types of sarcomas from the differential diagnosis. One of the solutions could be the use of high-throughput molecular assays such as RNA sequencing (RNA-seq) to help in the classification and diagnosis of sarcomas. This is indeed what has been applied in practice since 2015 at the Institut Curie (see RNA-seq of sarcomas at the Institut Curie). As a result of this clinical sequencing effort, a rich resource of transcriptomic profiles of sarcomas has been constituted, and I had the opportunity to have access to it for my PhD. This was in fact the starting point of my thesis: to characterize the transcriptomic landscape of sarcomas using this database of clinical RNA-seq (see

Characterization of the transcriptomic landscape of sarcomas).

In addition to the characterization of tumor cells, RNA-seq realized on bulk clinical samples can also give insight into the cells that form the tumor microenvironment, notably cells of the immune system. Since immunotherapy has now become one of the main weapons of the oncologist with impressive results against many types of cancers, the characterization of the **tumor immune microenvironment** is of great interest for the design of immunotherapies[11]. This is another main thread in my PhD work. In sarcomas, only a few responses to immunotherapy have been observed[12], though the number of trials and treated patients has been much smaller than for other cancers. Most of the sarcomas that show promising responses are genomically complex, and it is assumed that genomically simple tumors such as translocation-related sarcomas are intrinsically less immunogenic because of a low tumor mutational burden. However, this deserves further study to evaluate the potential of immune response: using the database of Institut Curie, that contains a large number of translocation-related sarcomas, I was able to explore in more detail the tumor immune microenvironment of sarcomas (see **Characterization of the immune microenvironment of sarcomas**).

The immunogenicity of cancers is largely driven by the presence of tumor neoantigens that can be recognized by cells of the immune system such as cytotoxic lymphocytes[13]. Most tumor neoantigens are believed to be generated from DNA mutations in protein-coding genes leading to modified peptides presented at the surface of the tumor cell. This is the reason why most tumors with a low tumor mutational burden such as pediatric tumors and notably sarcomas are supposed to be non-immunogenic[14,15]. However, an unexpected observation in Ewing sarcoma at the beginning of my PhD led to the identification of a potential source of tumor-specific neoantigens for more than a dozen of translocation-related sarcomas and other cancers (see **Identification of novel transcripts in sarcoma**).

After having characterized the immune microenvironment by bulk RNA-seq and studied potential sources of immunogenicity of sarcomas, I had the opportunity to analyze more finely the immune infiltrate of some types of sarcomas at the single-cell level. This type of analysis is extremely rich and interesting and allowed to derive important insights into the oncogenesis of one specific type of sarcoma: dedifferentiated liposarcoma (see **Transcriptomic characterization of sarcomas and microenvironment at the single-cell level**).

From a biological perspective, the main threads of my work are thus sarcomas and the associated immune microenvironment. From an orthogonal point of view, all projects were involved with **RNA**: not only bulk RNA-seq which allowed broad characterization of the transcriptomic landscape of sarcomas, diagnosis prediction and study of the immune microenvironment, but also at the single-cell level once for the detailed study of cells of the immune system and for fine characterization of the oncogenesis of one specific type of sarcoma. While most of my analyses were focused on measures of reference transcripts, one of my projects also highlighted the richness of transcriptomic resources to potentially make biological “discoveries” inside these treasure troves of data. I also got the opportunity to analyze specifically the translation of RNA using the recent technique of ribosome profiling.

Finally, the computational thread of this work was the **analysis of high-dimensional biomedical data**, that allowed me to master multiple bioinformatics analyses including their use with high-performance computing (HPC) clusters, as well as machine learning and deep learning methods for characterization and processing of this large amount of data. I learnt to appreciate the potentials and pitfalls in the analysis of high-dimensional datasets. Single-cell RNA-seq particularly gave me the opportunity to explore critically the challenges of applying technically complicated procedures in complex biological settings, notably in the important open problem of data integration of different patients. Finally, I was fortunate to successfully apply an “artificial intelligence” to improve in clinical practice the diagnosis and prognosis of individual patients at the Institut Curie and beyond.

Introduction générale (en français)

Chère lectrice, cher lecteur, vous allez entamer la lecture d'un manuscrit qui est le fruit d'un travail de thèse de sciences qui a duré trois ans, que j'ai eu la chance d'effectuer à l'Institut Curie sous la direction de Joshua Waterfall et Olivier Delattre. Comme vous allez le constater, ce travail n'est pas constitué d'un seul projet ayant progressé linéairement du début à la fin de ma thèse. En effet, j'ai été impliqué durant ces trois années dans de nombreux projets, ce qui m'a permis de réfléchir à et travailler sur des questions conceptuellement et techniquement très variées. Cependant et malgré cet aspect bariolé, tous ces travaux ont été tissés à l'aide de fils conducteurs communs pour s'inscrire dans une étoffe de trame cohérente, comme je vais vous le montrer dans la suite de ce texte.

Le « tissu » final est donc hétérogène et multicolore : alors que certains projets étaient déjà conçus au début de ma thèse, d'autres ont vu le jour en cours, soit au décours d'une observation biologique inattendue, soit pour répondre au besoin d'un cas clinique urgent ; certains d'entre eux ne sont même pas mentionnés dans ce manuscrit. Cette diversité peut en partie être expliquée par le grand nombre de jeux de données auxquels j'ai eu la chance avoir accès, et la multitude d'analyses possibles grâce aux outils actuels de la bioinformatique. Cependant, la raison principale est probablement ma propre tendance à vouloir mener plusieurs projets simultanément. En effet, j'ai eu à cœur pendant ma thèse de sciences d'apprendre le plus possible de techniques de bioinformatique, analyser de multiples types de données, et tenter de répondre à toutes les questions pertinentes biologiquement et cliniquement qui se présentaient à mon esprit.

Cela est peut-être dû à ma formation de médecin : je suis interne en oncologie médicale et par conséquent mes compétences en bionformatique étaient bien inférieures à celles de mes collègues scientifiques au début de ma thèse. J'avais toutefois déjà obtenu un Master en biologie des systèmes en 2012 et publié un modèle de la dynamique des lymphocytes T au sein du thymus à l'aide d'équations différentielles ordinaires[1]. Cependant la biologie computationnelle et plus spécifiquement la bioinformatique ont évolué très rapidement avant mon retour en thèse de sciences fin 2018. Un changement majeur a été l'arrivée des techniques dites de séquençage de nouvelle génération (next-generation sequecing, NGS) [2], qui ont notamment permis le séquençage de génomes et transcriptomes de tumeurs à large échelle. Un autre a été le développement rapide des méthodes d'apprentissage machine (machine learning) et deep learning (apprentissage profond), communément regroupées sous le terme populaire d'« intelligence artificielle » [3]. Etant donné ma spécialisation en oncologie médicale, j'étais particulièrement intéressé par le potentiel de ces nouvelles techniques pour mieux comprendre les mécanismes biologiques du cancer et aider à trouver de nouveaux traitements pour les patients. C'est pourquoi j'ai décidé d'effectuer une thèse en bioinformatique pour acquérir cette expertise et les compétences nécessaires pour exploiter le riche potentiel de ces données à haut débit, afin de répondre à des questions fondamentales pour la biologie et la médecine en cancérologie. Pour atteindre cet objectif, j'ai eu beaucoup de chance de pouvoir travailler à l'Institut Curie sous la direction de Joshua Waterfall et Olivier Delattre. Non seulement j'ai pu avoir accès à de précieuses bases de données transcriptomiques de patients, j'ai également pu prendre part à une multitude de projets intéressants en cours au sein de l'Institut. Cette grande diversité d'apprentissage et de pratique correspondait parfaitement à ce que je recherchais en tant qu'oncologue médical souhaitant apprendre à maîtriser une large panoplie de méthodes computationnelles pour l'analyse de données biologiques et médicales.

Plan du manuscrit

Je vais maintenant exposer le plan de ce manuscrit et les fils conducteurs sous-tendant la trame d'ensemble des différentes parties de ma thèse. Etant donné que chaque projet pourrait être traité indépendamment, j'ai décidé dans cette introduction générale de présenter les grandes lignes de mon travail, avant d'introduire et discuter plus en détail chacun des sujets au sein de leurs parties respectives.

L'un des fils conducteurs de ma thèse est l'étude des **sarcomes**. Les sarcomes sont des cancers d'origine mésenchymateuse, c'est-à-dire dérivés des os et tissus mous tels que le muscle, la graisse ou le cartilage. Ils forment un vaste groupe hétérogène de tumeurs malignes comprenant plus d'une centaine d'entités histologiques différentes[4]. La majorité des types de sarcomes sont des maladies rares (moins d'un cas pour 2000 habitants en Europe). Par conséquent, ils sont moins bien connus et n'ont que peu d'options thérapeutiques : la plupart des patients avec une maladie métastatique sont traités avec des chimiothérapies très toxiques, souvent sans bénéfice clinique majeur[5].

Indépendamment de leur classification histologique par les anatomopathologistes, les sarcomes sont classiquement divisés en deux groupes en fonction de leur profil d'altérations génomiques dit « simple » ou « complexe » [6]. Les sarcomes avec un profil génomique complexe sont caractérisés par un grand nombre d'altérations génétiques comprenant notamment des mutations, translocations chromosomiques, gains et délétions. Ils surviennent principalement chez des patients adultes de plus de 60 ans. Au contraire, les sarcomes avec un profil génomique simple sont pour la plupart induits par une unique altération génétique ; celle-ci est dans la majorité des cas une translocation chromosomique donnant naissance à un gène de fusion codant pour une protéine chimérique avec des propriétés oncogéniques. Ces sarcomes sont dits liés à des translocations et touchent plus spécifiquement les patients jeunes, notamment les enfants et les jeunes adolescents[7,8]. Un des représentants les plus connus de cette classe de sarcomes liés à des translocations est le sarcome d'Ewing, qui survient chez les enfants et adolescents et est la conséquence d'une translocation entre les chromosomes 11 et 22 donnant naissance au gène de fusion *EWSR1-FLI1* (*EWS-FLI1*)[9].

Un aspect essentiel de la prise en charge des patients atteints de sarcomes est la précision du diagnostic. Comme ce sont des maladies rares, les sarcomes sont souvent confondus avec d'autres cancers plus fréquents si les médecins en charge du patient ne sont pas familiers avec ce type de diagnostic. C'est la raison pour laquelle il est obligatoire que tout patient avec suspicion ou confirmation d'un diagnostic de sarcome soit pris en charge dans un centre de référence multidisciplinaire : une étude nationale française a notamment montré un bénéfice en survie grâce à cette mesure[10]. Cependant, un diagnostic précis reste parfois difficile à obtenir y compris dans un centre spécialisé. Etant donné qu'il existe plus d'une centaine d'entités histologiques et par ailleurs un nombre croissant de translocations caractérisées pour les sarcomes liés à des translocations, il y a un besoin réel d'outils permettant un diagnostic précis, rapide et correct incluant même les sarcomes les plus rares dans le diagnostic différentiel. Une des solutions pourrait être l'utilisation de techniques de séquençage à haut débit comme le séquençage de l'ARN (RNA-seq) pour aider à la classification et au diagnostic des sarcomes. C'est exactement ce qui a été mis en place à l'Institut Curie depuis 2015 (voir RNA-seq of sarcomas at the Institut Curie). Grâce à cet usage en routine clinique du RNA-seq pour le diagnostic, une riche base de données de transcriptomes de sarcomes a pu voir le jour, et j'ai eu la chance d'y avoir accès pendant ma thèse. Cette base de données était d'ailleurs le point de départ pour l'objectif initial de mon travail de thèse : caractériser le paysage transcriptomique des sarcomes (voir

Characterization of the transcriptomic landscape of sarcomas).

Au-delà de la caractérisation des cellules tumorales, le RNA-seq effectué sur des prélèvements tumoraux en « bulk » permet également l'étude des cellules constituant le microenvironnement tumoral, notamment les cellules du système immunitaire. Etant donné que l'immunothérapie est maintenant devenue l'une des principales armes dans l'arsenal de l'oncologue médical, avec des résultats impressionnants dans de nombreux types de cancers, l'étude du **microenvironnement immunitaire de la tumeur** est absolument essentielle pour mieux utiliser ces immunothérapies[11]. Ceci est un autre fil conducteur de ma thèse. Dans les sarcomes, seules quelques rares réponses à l'immunothérapie ont pu être observées[12], bien que les nombres d'essais et de patients traités soient logiquement plus faibles que dans d'autres types de cancers plus fréquents. La plupart des sarcomes qui montrent des signes de réponse à l'immunothérapie sont à génomique complexe, et il est supposé que les sarcomes à génomique simple comme les sarcomes liés à des translocations sont intrinsèquement moins « immunogéniques » en raison d'une charge mutationnelle faible. Cependant, une étude plus poussée est nécessaire pour évaluer le réel potentiel d'une réponse à l'immunothérapie : grâce à la base de données transcriptomiques de l'Institut Curie qui contient notamment un grand nombre de sarcomes liés à des translocations, j'ai pu entreprendre une caractérisation plus en détail du microenvironnement immunitaire tumoral des sarcomes (voir **Characterization of the immune microenvironment of sarcomas**).

L'immunogénicité (capacité à induire une réponse immunitaire) des cancers est en grande partie liée à la présence de néoantigènes tumoraux qui peuvent être reconnus par des cellules du système immunitaire telles que les lymphocytes cytotoxiques[13]. Il est communément admis que la plupart des néoantigènes tumoraux sont créés suite à des mutations de l'ADN tumoral au niveau de gènes codant pour des protéines, ce qui résulte en la présentation de peptides modifiés à la surface des cellules tumorales. C'est la raison pour laquelle la majorité des tumeurs avec une charge mutationnelle faible, telles que les tumeurs pédiatriques et les sarcomes, sont considérées comme non-immunogéniques[14,15]. Pourtant, une observation inattendue dans le sarcome d'Ewing au début de ma thèse a conduit à l'identification d'une source alternative potentielle de néoantigènes spécifiques de la tumeur pour plus d'une douzaine de sarcomes et autres cancers liés à des translocations (voir **Identification of novel transcripts in sarcoma**).

Après avoir caractérisé le microenvironnement immunitaire grâce au RNA-seq en « bulk » et étudié des sources potentielles d'immunogénicité dans les sarcomes, j'ai eu la chance de pouvoir analyser plus finement l'infiltrat immunitaire de certains types de sarcomes à l'échelle unicellulaire. Cette analyse dite en « single-cell » est extrêmement riche et intéressante et a permis également d'obtenir des résultats importants concernant l'oncogenèse d'un type de sarcome en particulier : le liposarcome dédifférencié (voir **Transcriptomic characterization of sarcomas and microenvironment at the single-cell level**).

D'un point de vue biologique, les fils conducteurs de mon travail sont donc les sarcomes et leur microenvironnement immunitaire. Si l'on prend une perspective orthogonale, tous mes projets tournaient autour de l'**ARN** : non seulement le RNA-seq en « bulk » qui a permis la caractérisation large du paysage transcriptomique des sarcomes, la prédiction diagnostique et l'étude du microenvironnement immunitaire, mais aussi à l'échelle unicellulaire (« single-cell ») pour l'étude détaillée des cellules du système immunitaire et la caractérisation fine de l'oncogenèse d'un type de sarcome en particulier. Bien que la plupart des mes analyses aient été concentrées sur la mesure de transcrits de référence déjà connus, l'un de mes projets a également mis en lumière les potentielles « découvertes » (en l'occurrence, de nouveaux transcrits) qui peuvent être faites au sein de ces riches trésors que sont les bases de données de séquençage à haut débit. J'ai aussi eu l'occasion par ailleurs d'analyser plus spécifiquement la traduction des ARN à l'aide de la technique récente dite de Ribo-seq.

Enfin, le fil conducteur computationnel de mon travail a été l'**analyse de données biomédicales à haute dimension**, ce qui m'a permis d'apprendre à maîtriser de nombreux outils bioinformatiques, y compris à l'aide de clusters d'ordinateurs effectuant des calculs de haute performance, ainsi que des méthodes

d’apprentissage machine (machine learning) et apprentissage profond (deep learning) pour la caractérisation et l’analyse de cette grande quantité de données. J’ai appris à connaître le riche potentiel mais aussi les pièges éventuels de ces analyses de données à haute dimension. En particulier, le RNA-seq à échelle unicellulaire m’a donné l’occasion d’expérimenter moi-même les difficultés d’application de ces techniques compliquées à des données biologiques non moins complexes, notamment pour le problème épineux et non résolu de l’intégration de données provenant de plusieurs patients différents. Enfin, j’ai eu la chance de réussir à appliquer en pratique clinique un outil d’ « intelligence artificielle » pour améliorer en pratique clinique le diagnostic et le pronostic de patients à l’Institut Curie et au-delà.

RNA-seq of sarcomas at the Institut Curie

Introduction

Many sarcomas are characterized by chromosomal translocations[7,8] which give rise to fusion genes with driving oncogenic properties such as *EWSR1-FLI1* in Ewing sarcoma[9,16,17].

Diagnosis of these translocation-associated sarcomas is usually based on the detection of the characteristic gene fusion by Fluorescence In Situ Hybridization (FISH) or Reverse Transcriptase Polymerase Chain Reaction (RT-PCR). However, these diagnostic assays are supervised and call for a diagnostic hypothesis to be proposed by the clinician and pathologist. While this approach may be adapted for easy-to-diagnose cases, it can be complicated by the fact that 1) numerous types of sarcoma exhibit similar clinical and pathological characteristics, such as the groups of small round cell sarcomas[18] or Ewing sarcoma and Ewing-like tumors[19–21]; 2) some sarcomas may harbor previously uncharacterized molecular alterations such as novel chromosomal translocations. In these cases, FISH or RT-PCR is not appropriate to screen for all potential diagnoses, due to practical (scarcity of tumor tissue) and financial issues of performing multiple assays. It is also intrinsically unable to detect novel gene fusions.

In contrast, RNA sequencing (RNA-seq)[22] is a more recent assay, based on next-generation sequencing (NGS) of RNA, that can overcome these limits of FISH and RT-PCR for sarcoma diagnosis. Indeed, RNA-seq 1) is unsupervised and able to identify any known fusion gene giving rise to an expressed fusion transcript; 2) can moreover detect and characterize novel molecular alterations such as gene fusions.

Notwithstanding these advantages of RNA-seq as compared to FISH and RT-PCR, these last techniques are still considered the standard assays for diagnosis of sarcoma, mainly due to practical reasons: RNA-seq 1) is not available for all clinical laboratories; 2) requires frozen tissue (formalin-fixed and paraffin-embedded, FFPE tissue may be used but extracted RNA is of lower quality[23]); 3) is costly if compared to a single FISH or RT-PCR assay; 4) requires specific expertise, notably bioinformatics.

Clinical use of RNA-seq at the Institut Curie

However, some clinical reference centers do have the capacity to implement RNA-seq for diagnostic purposes: this is the case at the Institut Curie in Paris, which possesses in-house sequencing equipment and significant technical and bioinformatics expertise of RNA-seq. Starting from 2015 onwards, the Unité de Génétique Somatique (UGS) of Institut Curie, directed by Olivier Delattre and Gaëlle Pierron, has thus been a pioneer of using RNA-seq as a diagnostic procedure for patients presenting with a clinico-pathological suspicion of sarcoma, especially for children and young adolescents, from all over France. RNA-seq is particularly relevant for cases where clinico-pathological characteristics do not hint at a specific diagnosis straightforwardly, leaving the clinician and pathologist with a large array of differential diagnoses to consider. In these cases, performing multiple FISH or RT-PCR analyses would require a large amount of tissue material, which is not possible notably for pediatric patients, and even financially the cost would rapidly overcome that of a single RNA-seq analysis.

This use of RNA-seq as a diagnostic tool for sarcomas is unprecedented in the world: in addition to delivering cutting-edge diagnostic performance for clinical management of patients, including for all difficult cases from all over France, it has allowed the constitution of a large and unprecedented database of RNA-seq of patients with sarcomas, and the discovery of previously uncharacterized types of sarcoma associated to novel gene fusions[24,25]. Up to now, more than 2000 patients have been profiled in the UGS by RNA-seq for

diagnostic suspicion of sarcoma. In about 60% of cases, RNA-seq has confirmed the main diagnostic hypothesis of the clinician and pathologist, either by detecting the presence of the characteristic fusion gene (40%) or the absence of markers for differential diagnoses (20%). In another 20% of cases, some unexpected positive markers were detected and allowed a diagnosis not suspected by the clinician and pathologist. Finally in about 10% of cases, no diagnosis could be clearly confirmed but there were indications to explore further the case and clues to the potential diagnoses were delivered by RNA-seq. These real-world results showcase the feasibility of using RNA-seq in the diagnostic workflow of patients presenting with clinical suspicion of sarcoma, and its added value in helping to diagnose difficult cases for which usual testing would have been either unpractical (too many FISH or RT-PCR analyses needed to test all gene fusions suspected) or impossible (novel gene fusions).

Transcriptomic database of the Institut Curie

Independently of this clinical use of RNA-seq, which in practice is heavily biased towards detection of gene fusions associated to specific types of sarcomas, the database constituted by this sequencing effort represents a unique treasure trove for sarcoma research, as it contains genome-wide expression profiles of hundreds of patients diagnosed with a large array of types of sarcomas, including extremely rare ones.

Institut Curie is a reference center for sarcomas in France, especially for sarcomas of children and young adolescents. The age of patients profiled in the UGS by RNA-seq is thus biased towards the lower age spectrum: 66% were less than 25 years old at sampling time (Figure 1).

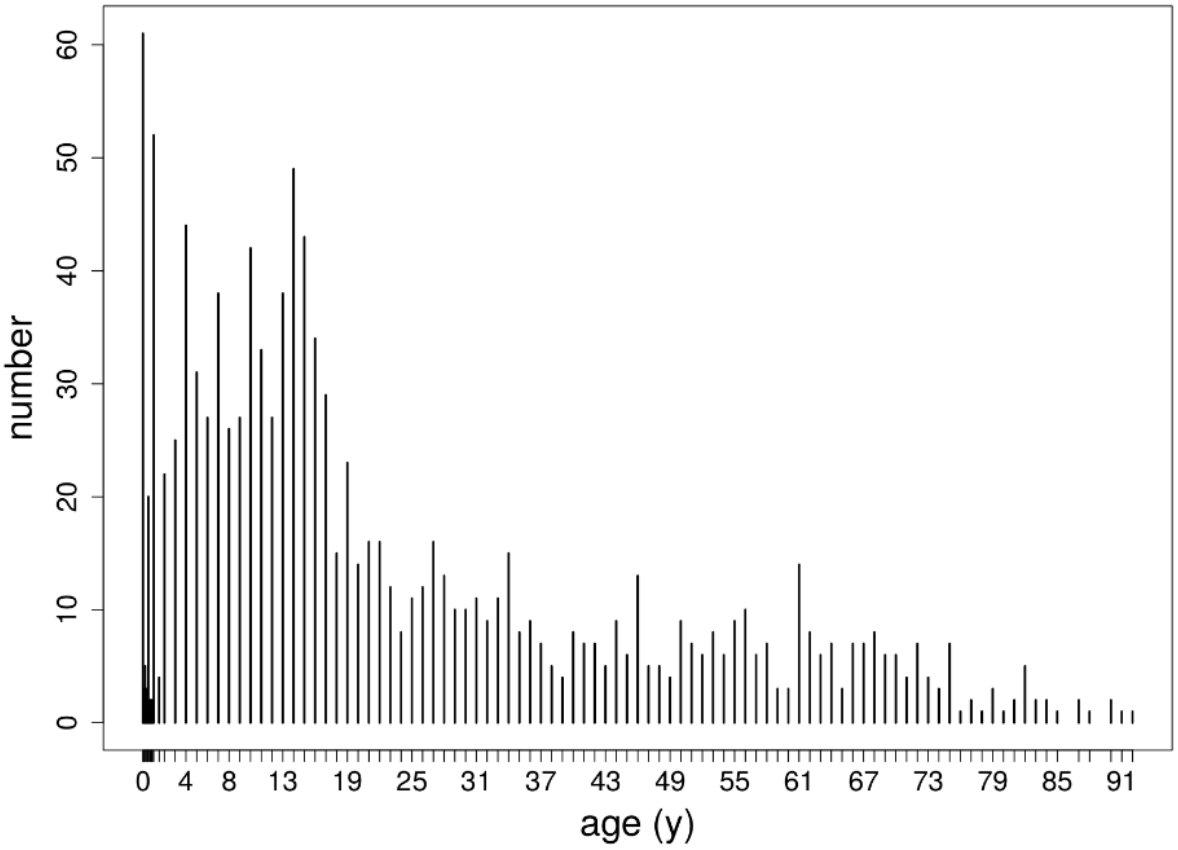

Figure 1: Age distribution of patients profiled by RNA-seq in the UGS.

Similarly, the distribution of final diagnoses (Figure 2) - after consideration of all evidence including RNA-seq - reflects a higher burden of sarcomas more prevalent in children and young adolescents, as well as diseases characterized by small round cells in pathology, for which the differential diagnosis is broad (Ewing sarcoma and Ewing-like tumors, desmoplastic small round cell tumor, neuroblastoma, lymphoma, and others) and RNA-seq particularly valuable.

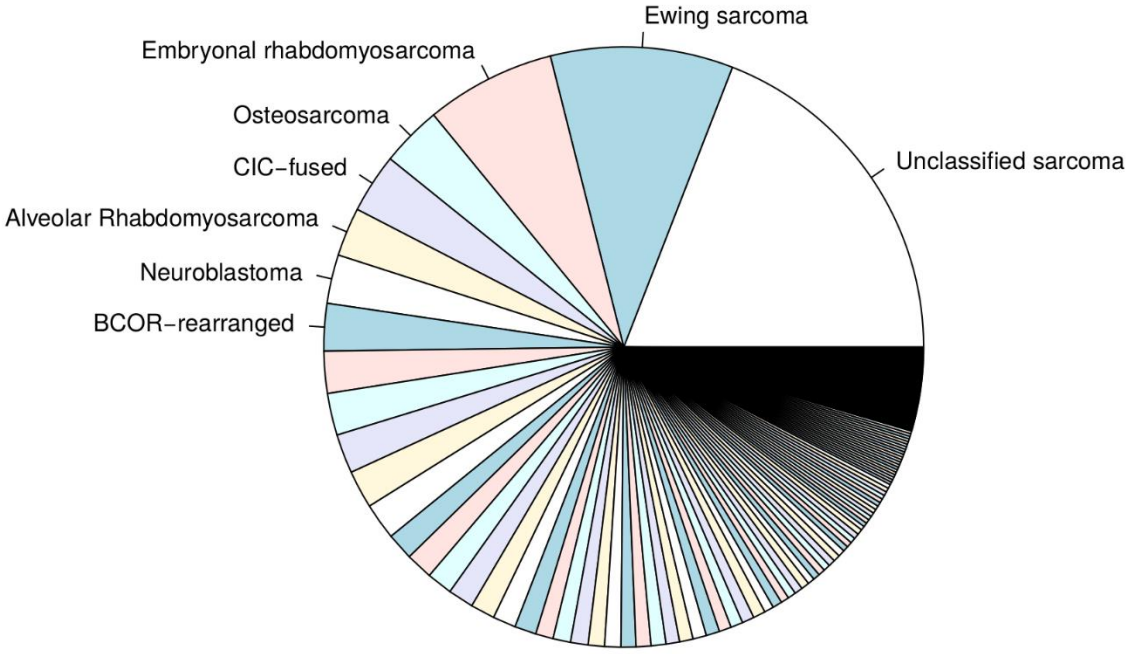

Figure 2: Diagnoses in the RNA-seq database of the UGS.

Notably, Ewing sarcoma is the most frequent type of sarcoma in the database, followed by other relatively frequent types such as alveolar and embryonal rhabdomyosarcoma, as well as osteosarcoma, and Ewing-like sarcomas (CIC-fused, BCOR-rearranged). Due to the prospective inclusion of patients with diagnostic suspicion of sarcoma for RNA-seq profiling, some final diagnoses are not sarcomas: after RNA-seq, some patients may be reclassified into differential diagnoses such as neuroblastoma (one of the most frequent non-sarcoma pediatric cancers in the database) and lymphoma. Additionally, there are a large number of pediatric brain tumors in the database, since Institut Curie is also a reference center for this type of cancers. Finally, the largest diagnostic category remains “Unclassified sarcoma”, showing the difficulty of making a diagnosis for a significant number of patients, even with the use of RNA-seq. This could be due to many reasons: either technical (tissue sample of low quality or low tumor cellularity, precluding a meaningful RNA-seq analysis) or biological (non-translocation-associated sarcoma without a specific expression profile, non-sarcoma cancer without identifiable characteristics in RNA-seq, or potentially novel uncharacterized type of sarcoma). Overall, there are more than 150 different diagnostic entities in the database, but most of them are extremely rare sarcoma types that are represented by less than five samples each.

The characterization of “unclassified sarcomas” by RNA-seq was precisely the object of some landmark studies realized in the team of Olivier Delattre with the use of this UGS database[24,25]. These showcased the ability of RNA-seq to single out specific groups of unclassified sarcomas and to discover novel gene fusions characterizing these, leading for instance to the identification of novel entities such as BCOR-rearranged sarcomas[24], and epithelioid and spindle-cell rhabdomyosarcomas characterized by EWSR1– or FUS–TFCP2 fusions[25].

This RNA-seq database was the starting point and main resource for many of my following projects, including the characterization of the immune microenvironment of sarcomas, their transcriptomic landscape and diagnosis prediction with machine learning, as well as the identification of novel transcripts in multiple translocation-related sarcomas.

Characterization of the immune microenvironment of sarcomas

Introduction

Immunotherapy has revolutionized care of many cancers in the last decade, however only a subset of patients usually responds to these treatments. There is thus a high need to determine the factors that determine response to immunotherapy[26]. One obvious requirement though is the ability of cells of the immune system to have access to the tumor in the first place. This has prompted a focus on the study of the tumor immune microenvironment of cancers[27] in order to determine the presence and quantity of specific cell populations that could influence and predict response to immunotherapy[28]. While definitive biomarkers have not yet been identified and these correlates of response vary according to cancer type, presence and high abundance of some cell types such as CD8+ T lymphocytes[29], B cells[30–32] and M1 macrophages[33], have notably been associated to response to immunotherapy in some cancers.

In the UGS, RNA-seq is performed on bulk tumor tissue samples and thus contains RNA not only from tumor cells, but also potentially from neighboring cells of the tumor microenvironment such as cells of the immune system, endothelial cells and fibroblasts. While it is not possible to assign each sequenced transcript to its parent cell among the millions that constitute the sample, a large number of bioinformatics methods have been developed to infer the characteristics of the tumor microenvironment from bulk RNA-seq. These methods notably estimate the presence and relative quantity of specific microenvironment cell populations, such as CIBERSORT[34] (and its more recent counterpart CIBERSORTx[35] which takes advantage of single-cell RNA-seq reference profiles), xCell[36], and MCP-counter[37]. This last method is one of the most popular in the literature and offers several advantages: 1) its underlying algorithm is simple to understand and interpret (it is based on the quantification of transcriptomic marker genes for specific microenvironment populations); 2) it has been properly validated in multiple settings, including in controlled *in vitro* cell mixture experiments and *ex vivo* immunohistochemistry, and in showing survival relevance in patients with some cancer types such as lung adenocarcinoma, colorectal and breast carcinomas; 3) it can be used to compare abundance of microenvironment populations between samples, as opposed to CIBERSORT (though this last method is more suited to compare quantity of different microenvironment populations within the same sample[38]).

Cell composition of the sarcoma tumor microenvironment (MCP-counter)

MCP-counter has notably been used in a recent landmark study of the immune microenvironment in soft-tissue sarcomas[30]. This work clearly demonstrated the heterogeneity of the immune microenvironment within sarcoma types present in the TCGA[39], with five groups of patients including one associated with higher immune infiltration by B lymphocyes and tertiary lymphoid structures, showing improved survival and higher response rate to PD1 blockade immunotherapy.

However, this study was done exclusively in sarcoma types present in the TCGA, which is rather limited in number of diagnoses (mainly adult sarcomas: dedifferentiated liposarcomas, undifferentiated pleomorphic sarcomas, leiomyosarcomas) and for instance does not include any pediatric sarcomas. This prompted me to apply MCP-counter to our dataset of RNA-seq of sarcomas from the UGS.

To this end, I selected the samples from the UGS RNA-seq database fulfilling the following criteria: 1) final diagnosis of sarcoma, as validated by expert reviewing of all evidence including RNA-seq by a practicing sarcoma medical oncologist (Dr Sarah Watson from Institut Curie), to avoid mislabeling of samples; 2)

diagnostic category comprising at least four samples, to avoid diagnoses with too few samples. I also included some biologically sarcoma-related diagnoses such as translocation-associated carcinomas (midline carcinoma and TFE3-renal cell carcinoma) and non-malignant mesenchymal tumors (desmoid tumor and myoepithelioma). This amounted to 666 samples distributed into 34 different diagnoses (Table 1).

Table 1: Selected diagnoses for study

Diagnosis	Abbreviation	Number of samples
Alveolar rhabdomyosarcoma	aRMS	36
Alveolar soft part sarcoma	ASPS	10
Angiomatoid fibrous histiocytoma	AFH	7
Atypical teratoid rhabdoid tumor	ATRT	8
BCOR-rearranged sarcoma	BCOR	26
CIC-fused sarcoma	CIC	31
Clear cell sarcoma	CCS	6
Congenital fibrosarcoma	CFS	19
Desmoid tumor	Desmoid	39
Desmoplastic small round cell tumor	DSRCT	20
Embryonal rhabdomyosarcoma	eRMS	89
Ewing sarcoma	EwS	132
EWSR1-NFATC2 sarcoma	NFATC2	10
EWSR1-PATZ1 sarcoma	PATZ1	4
Extraskeletal myxoid chondrosarcoma	emCS	15
FET-TFCP2 epithelioid rhabdomyosarcoma	FET-TFCP2	4
Inflammatory myofibroblastic tumor	IMFT	11
Lipofibromatosis-like neural tumor	LFLNT	4
Liposarcoma_NOS	LPS	8
Low grade fibromyxoid sarcoma	LGFMS	8
Malignant peripheral nerve sheath tumor	MPNST	11
Malignant rhabdoid tumor	MRT	5
Mesenchymal chondrosarcoma	MCS	9
Midline carcinoma	Midline	4
Myoepithelioma	MYOEP	6
Myxoid liposarcoma	mLPS	29
NTRK-fused sarcoma	NTRK	5
Osteosarcoma	Osteo	42
Small cell carcinoma of the ovary-hypercalcemic type	SCCOHT	6
Solitary fibrous tumor	SFT	16
Synovial sarcoma	SS	25
TFE3 renal cell carcinoma	TFE3	9
Undifferentiated pleomorphic sarcoma	UPS	8
VGLL2-fused rhabdomyosarcoma	VGLL2	4

Method: RNA-seq reads were adapter-trimmed with Atropos (v1.1.21), aligned to the human reference genome (hg19) with GENCODE version 19 as the reference gene annotation with the use of STAR (v2.7.0e)[40] and quantified with the GeneCounts algorithm. Raw counts were normalized to transcripts per million (TPM)

and log2-transformed with pseudocount 1. MCPcounter (v1.1.0)[37] was used to calculate scores for ten immune and stromal cell populations with the option “HUGO_symbols”.

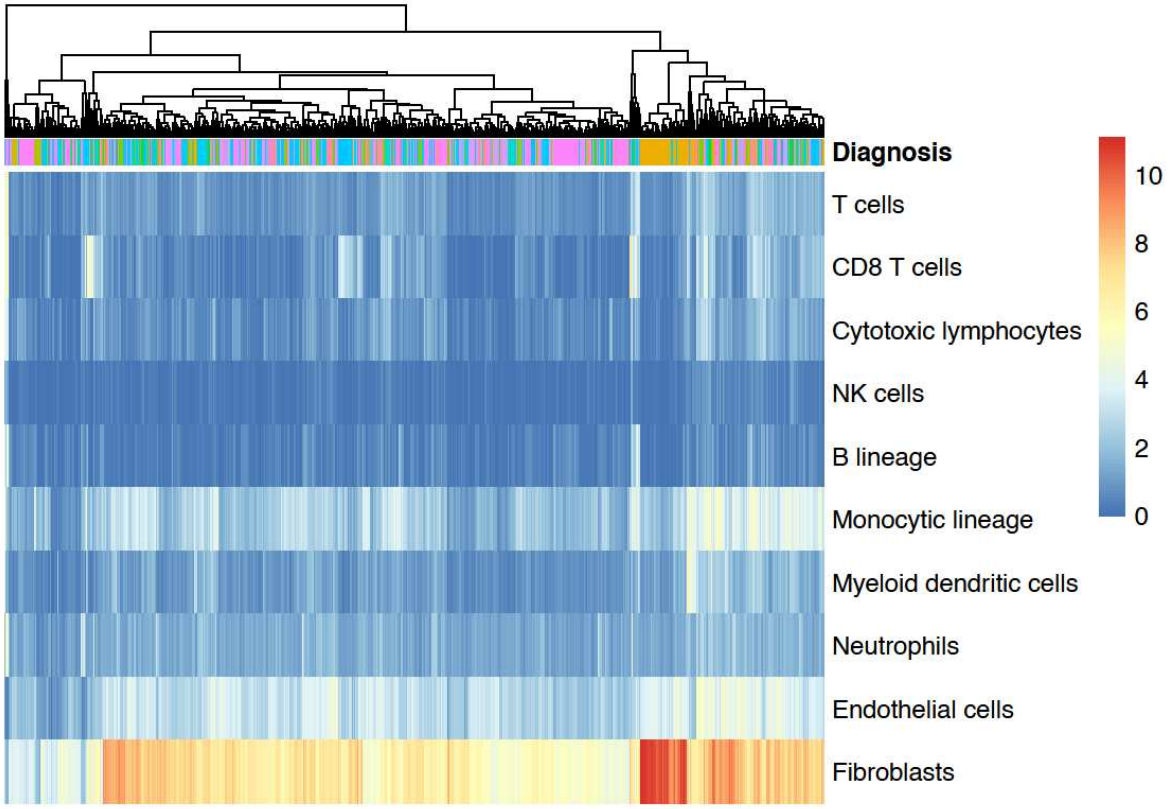

Figure 3: MCP-counter scores for all samples in Table 1. Hierarchical clustering by samples (columns) using R package pheatmap (v1.0.12).

Figure 3 shows a heatmap of MCP-counter scores of ten immune and stromal cell populations for all 666 samples studied, with hierarchical clustering of samples independently of diagnosis. Several observations can be made: 1) In contrast to the TCGA adult sarcoma cohort[30], there is no clear-cut separation of samples between immune-low and immune-high categories, especially concerning B and T lymphocytes. Except for some outliers, the MCP-counter scores of B and T lymphocytes as well as NK cells are globally low throughout the UGS cohort. 2) Cells such as monocytes and endothelial cells exhibit moderately higher scores but also show a global homogeneity overall. 3) The most striking separation concerns fibroblasts, for which a large part of samples have very large scores, while a minority (left of heatmap) display lower scores. These observations are consistent with what is already known of pediatric sarcomas: they are generally poorly infiltrated by immune cells especially from the adaptive immune system, probably due to factors such as low tumor mutational burden that render most pediatric tumors poorly immunogenic[14,15]. The intense fibroblastic signature is confounded by the mesenchymal origin of sarcomas, which intrinsically overexpress genes that are supposed to be fibroblast gene markers (such as collagen genes). This fibroblast-rich signal should therefore not be simply interpreted as an enrichment in cancer-associated fibroblasts in sarcomas. Conversely the minority of samples with lower fibroblast scores are the non-sarcoma cancers of the cohort; this lower score may also only reflect their non-mesenchymal origin as compared to sarcomas.

To further explore these results by diagnosis and facilitate visualization, I plotted a heatmap of MCP-counter median scores by diagnosis, scaled by row (cell population) in Figure 4.

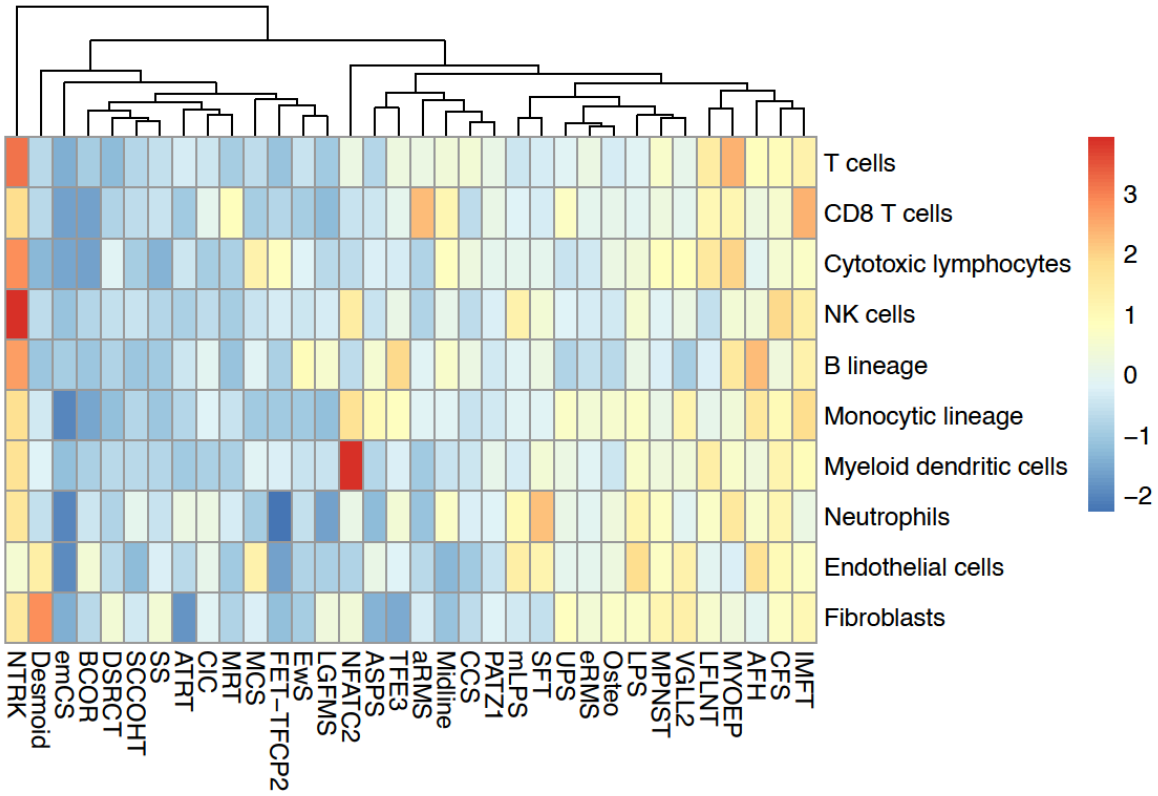

Figure 4: MCP-counter median scores by diagnosis, scaled by row (cell population). Hierarchical clustering by diagnoses (columns) using R package pheatmap (v1.0.12). Abbreviations are as in **Table 1**.

Once again, the overall picture is homogeneous: scaled scores tend to remain between -1 and 1 for all populations. Some outliers stand out nevertheless: NTRK-fused sarcoma (NTRK) has a relatively higher score for all immune cell populations, including cells from the adaptive immune system such as T and B lymphocytes; EWSR1-NFATC2 sarcoma (NFATC2) has a higher myeloid dendritic cell score. Desmoid tumors have the highest fibroblast score, which is consistent with its mesenchymal origin and its highly fibrous (desmoplastic) tissue component. Overall, the heatmap is quite homogeneous, though one could point out a subtle tendency of all cell population scores to be higher on the right part of the heatmap, as though there was a continuous gradient of increasingly higher scores from left to right: the most infiltrated diagnoses (on the rightmost part) are inflammatory myofibroblastic tumor (IMFT), congenital fibrosarcoma (CFS), angiomatoid fibrous histiocytoma (AFH). This makes sense especially for IMFT which is as its name implies an inflammatory tumor infiltrated by immune cells.

Relative homogeneity in median scores by diagnosis does not implicate absence of heterogeneity between tumors of the same diagnosis: to further explore intertumoral heterogeneity within the same diagnosis, I plotted heatmaps of individual sample MCP-counter scores for each diagnosis separately. Though most of the diagnoses also display homogeneity between individual samples (most of the thirty-four heatmaps are therefore not displayed in this document), some do exhibit a degree of intertumoral heterogeneity with regards to the microenvironment.

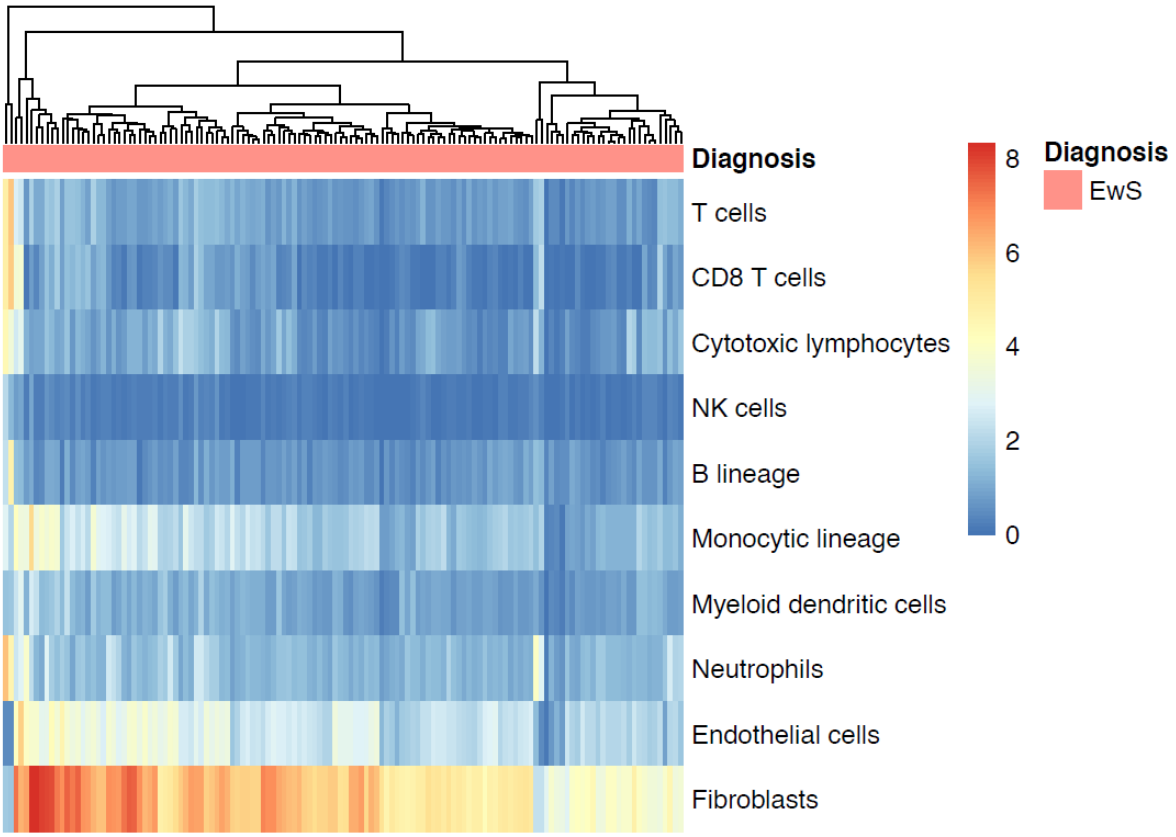

Figure 5: MCP-counter scores for Ewing sarcoma (EwS). Hierarchical clustering by diagnoses (columns) using R package pheatmap (v1.0.12).

For instance, Ewing sarcoma (EwS) has a microenvironment landscape which is globally homogeneous with low scores for most immune cell populations, however a few samples display significantly higher scores of immune cells, especially T cells and neutrophils (Figure 5). Unfortunately, the UGS database does not contain other characteristics than diagnosis, such as survival data, therefore it is not possible to correlate this interesting subset of EwS samples with clinical variables.

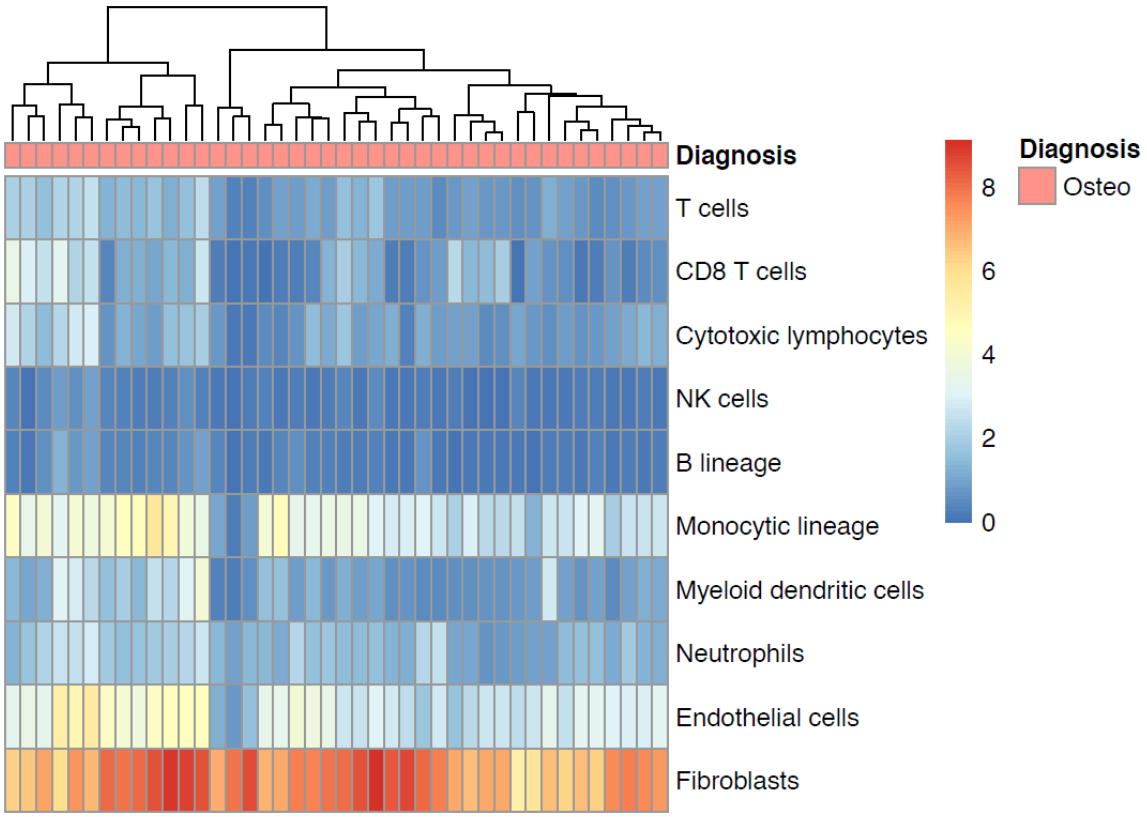

Figure 6: MCP-counter scores for osteosarcoma (Osteo). Hierarchical clustering by diagnoses (columns) using R package pheatmap (v1.0.12).

Osteosarcoma (Osteo) also displays globally low scores of immune cell populations, however there is a subset of samples (left of heatmap) that exhibits relatively higher scores of T lymphocytes (Figure 6).

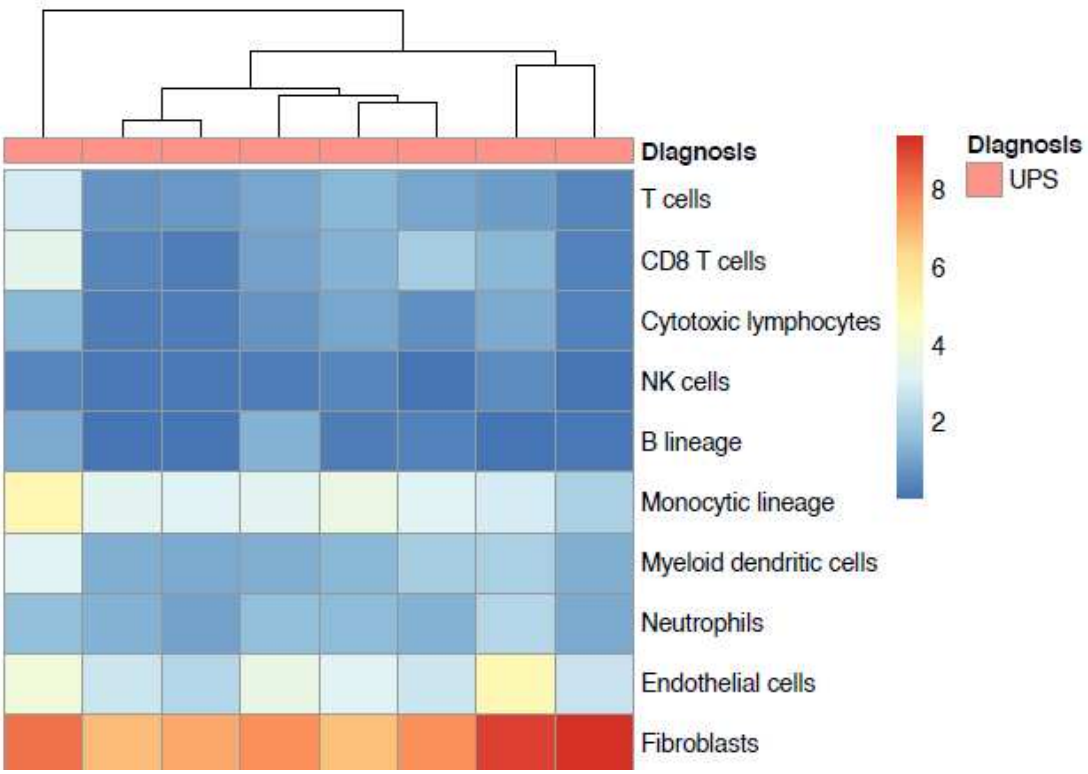

Figure 7: MCP-counter scores for undifferentiated pleomorphic sarcoma (UPS). Hierarchical clustering by diagnoses (columns) using R package pheatmap (v1.0.12).

Undifferentiated pleomorphic sarcoma (UPS) is one of the subtypes present in the TCGA study and it is interesting to see the heatmap for this diagnosis in our UGS cohort (Figure 7). Petitprez et al.[30] clearly demonstrated the existence of a highly infiltrated subgroup of UPS using MCP-counter and validated by other approaches: in our cohort there is one sample (left of heatmap) that displays higher scores of immune cell populations, which could be related to the observation of an infiltrated subgroup from the Petitprez study, though our sample numbers (n=8) preclude us from making definitive conclusions about the presence of an immune-rich group here. This low amount of samples is due to the difference between our cohort which is predominantly composed of pediatric sarcomas (UPS is an adult sarcoma), as opposed to the TCGA.

To conclude on this part, using MCP-counter on our UGS cohort has enabled us to get a complementary view of the sarcoma immune landscape in addition to the previous study on TCGA sarcomas[30]. Contrary to adult sarcomas in the TCGA which exhibit a subgroup of highly infiltrated tumors, the overall landscape of pediatric sarcomas is one of globally low infiltration by immune cells, especially from the adaptive immune system (B and T lymphocytes), though some outliers do stand out such as NTRK-fused sarcomas, inflammatory myofibroblastic tumors, and some Ewing sarcomas. This is consistent with the observation that pediatric tumors are overall poorly immunogenic, potentially because of their low tumor mutational burden and consequently low neoantigen generation potential[14,15]. While this also seems to corroborate the low response rates of sarcomas to immunotherapy compared to other cancers[12], this does not preclude the possibility of effective immunotherapy, especially for cases and sarcoma types with higher infiltration by cells of the immune system that probably deserve more exploration of the potential of immunotherapy.

Characterization of the infiltrating lymphocyte repertoires (MiXCR)

Methods such as MCP-counter infer global population abundances of immune and stromal cells in the tumor microenvironment from bulk RNA-seq, but they give no insight into the heterogeneity within these cell populations. This heterogeneity would however be especially interesting to study for cells of the adaptive immune system, which display responses to specific antigens and exhibit complex dynamics of cell proliferation and immune repertoire formation upon recognition of antigens[41]. Specifically, a high infiltration of T or B lymphocytes could be due to: 1) a polyclonal response to tumor antigens; 2) one or a few monoclonal expansions of T or B lymphocytes with a specific T-cell receptor (TCR) or B-cell receptor (BCR) recognizing one cognate tumor antigen. Knowing this information would surely be of relevance to the understanding of immune recognition of tumors by immune cells and the design of immunotherapies.

The so-called "repertoire" of T and B cells is the result of somatic recombination of V(D)J segments and accumulation of mutations in TCR and BCR: it is not encoded in the germline and cannot be recovered by standard alignment to the reference genome. State-of-the-art methods for characterizing this immune "repertoire" of T and B cells are based on specific next-generation sequencing of TCR and BCR in single-cell assays[42,43]. However, these methods are costly and are challenging to scale to large numbers of samples such as in the TCGA or UGS cohorts. Nonetheless, there are some methods that attempt to probe the immune repertoires of T and B cells from tumor bulk RNA-seq, such as MiXCR[44]. This method is based on the same principle of sequencing TCR and BCR as state-of-the-art methods, though its starting material is much less abundant since the TCR and BCR sequences are contained within a much higher number of non-TCR/BCR sequences. The objective of this tool is therefore to extract all sequences mapping to a TCR or BCR from all RNA-seq reads, and infer the specific clonotypes (i.e. one series/"clone" of T or B cells exhibiting the same TCR/BCR) present in the sample. For this, it uses specialized alignment and assembly techniques to recover the specific sequences of each clonotype. The output of this method for each sample is therefore composed of: 1) number of different clonotypes (measure of "clonality"); 2) CDR3 sequence of each clonotype; 3) number of cells per clonotype (measure of "expansion" of each clonotype). Moreover, this information is extracted for 1) B cells: IGH (immunoglobulin heavy chain), IGL/IGK (immunoglobulin light chains kappa/lambda); and 2) T cells: TRA (TCR alpha chain), TRB (TCR beta chain), TRG (TCR gamma chain) and TRD (TCR delta chain). I performed MiXCR on the same 666 samples used for the MCP-counter analysis (Table 1).

Method: RNA-seq reads were adapter-trimmed with Atropos (v1.1.21) and MiXCR (v3.0.5)[44] was used to extract clonotype information with options "—starting material rna – only-productive". Output tables from MiXCR were parsed to extract the following information for each sample: number of different clonotypes, maximum clonality (number of sequences for same clonotype), total number of sequences. To account for sequencing depth, these numbers were normalized to "per million RNA-seq reads", i.e. divided by total number of RNA-seq reads in the sample and multiplied by one million.

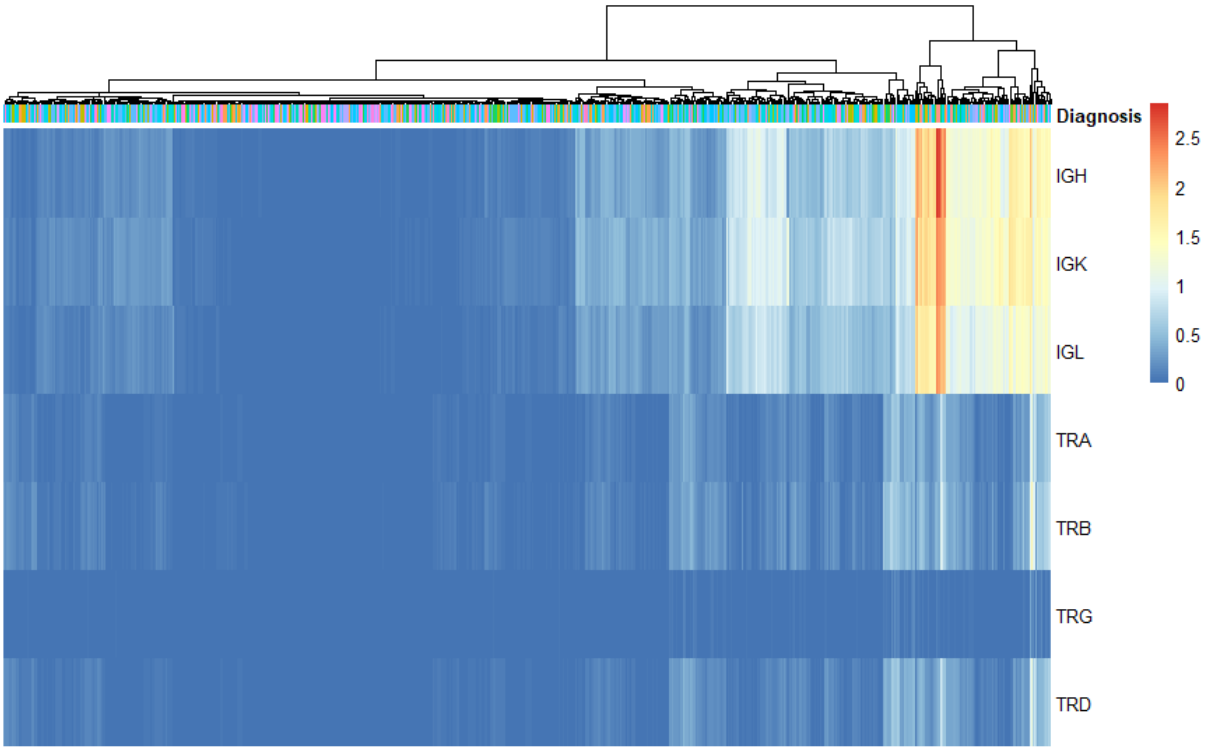

Figure 8: MiXCR number of clonotypes per sample. Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by samples (columns) using R package pheatmap (v1.0.12).

Figure 8 shows the normalized number of different clonotypes per sample retrieved by MiXCR for all 666 samples. Several observations can be made: 1) B clonotypes (IGH, IGH, IGL) tend to be more numerous than T clonotypes (TRA, TRB, TRG, TRD) in a given sample. This can be due to differential levels of infiltration but also to the process of somatic hypermutation, generating higher diversity specifically in B cells. 2) Overall, the number of different clonotypes is low across most samples (less than 10 for either BCR or TCR). 3) There is a minority of samples (right of heatmap) that displays higher number of clonotypes notably for B cells, and numbers of B and T clonotypes tend to be correlated.

To explore further these results by diagnosis and facilitate visualization, I plotted a heatmap of the median normalized number of different clonotypes by diagnosis in Figure 9.

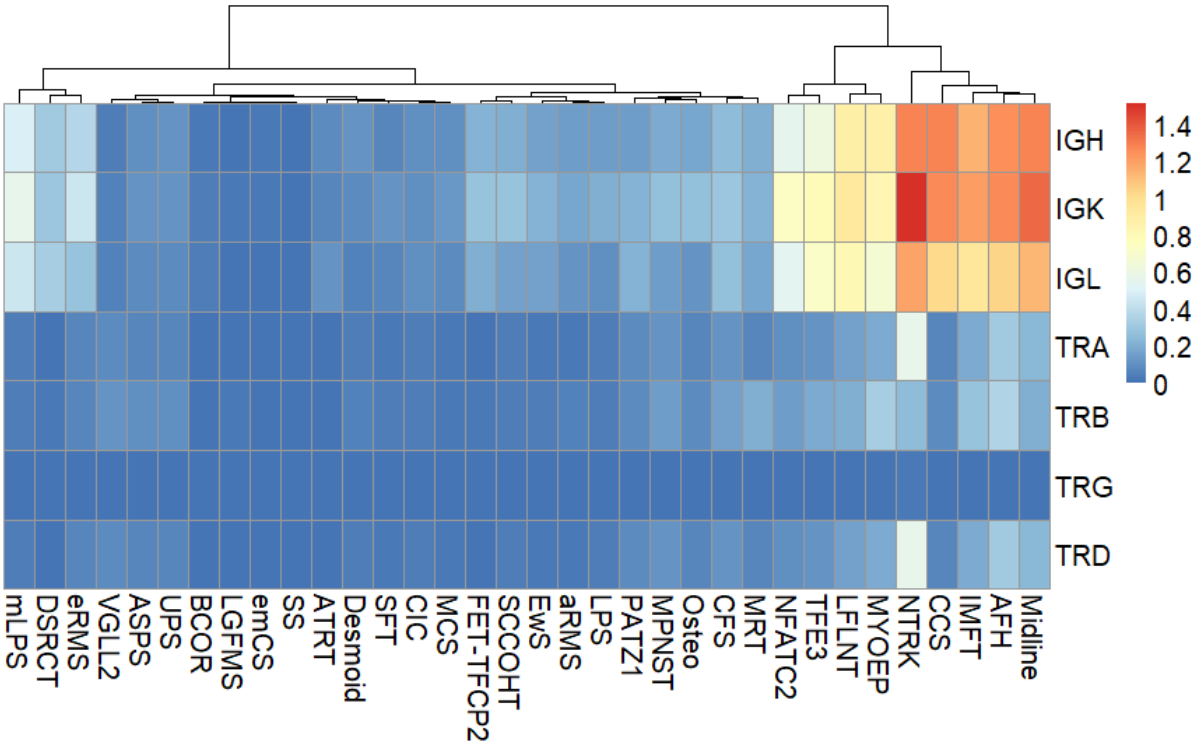

Figure 9: MiXCR median number of clonotypes per diagnosis. Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by diagnoses (columns) using R package pheatmap (v1.0.12).

The global picture is overall quite similar to the immune landscape depicted by MCP-counter: most diagnoses have low numbers of clonotypes, only a few have relatively higher values. It is interesting to note that diagnoses with higher numbers of clonotypes are generally those that have higher MCP-counter scores: for instance inflammatory myofibroblastic tumors (IMFT), NTRK-fused sarcomas (NTRK), angiomatoid fibrous histiocytoma (AFH). NTRK-fused sarcomas notably seem to have the highest number of T cell clonotypes.

The number of different clonotypes measures the diversity of the B and T cell infiltrate in tumors: it informs about the polyclonal or oligoclonal nature of the immune response. Another measure of the immune infiltrate is its global abundance, that can be estimated by the total number of BCR/TCR sequences retrieved by MiXCR, independently of their clonality. I therefore plotted a heatmap of the median total normalized number of sequences by diagnosis in Figure 10.

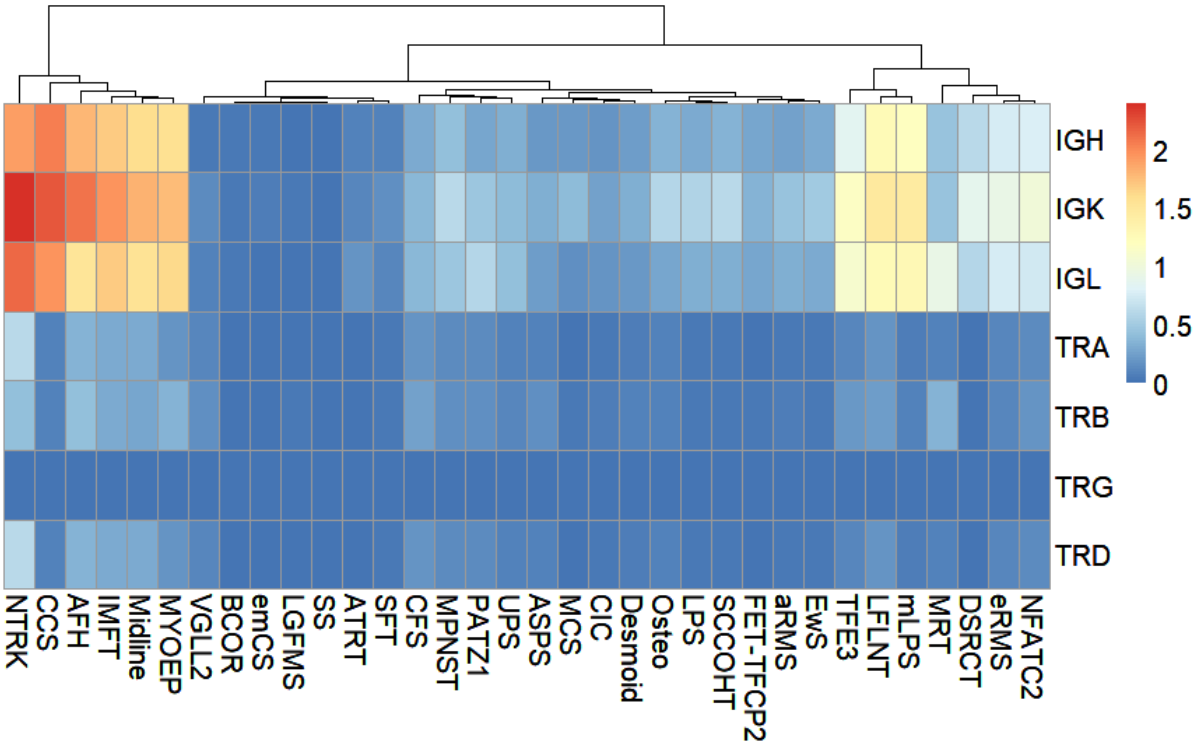

Figure 10: MiXCR median number of total BCR/TCR sequences per diagnosis. Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by diagnoses (columns) using R package pheatmap (v1.0.12).

This heatmap is overall very similar to Figure 9, perhaps unsurprisingly considering that the total number of sequences is the product of the number of different clonotypes and the (average) number of sequences per clonotype. This last factor is a measure of the expansion of clonotypes, i.e. the extent to which one specific clonotype is amplified due to antigenic positive selection. This can be approximately appreciated by plotting the median maximal clonality per diagnosis (Figure 11).

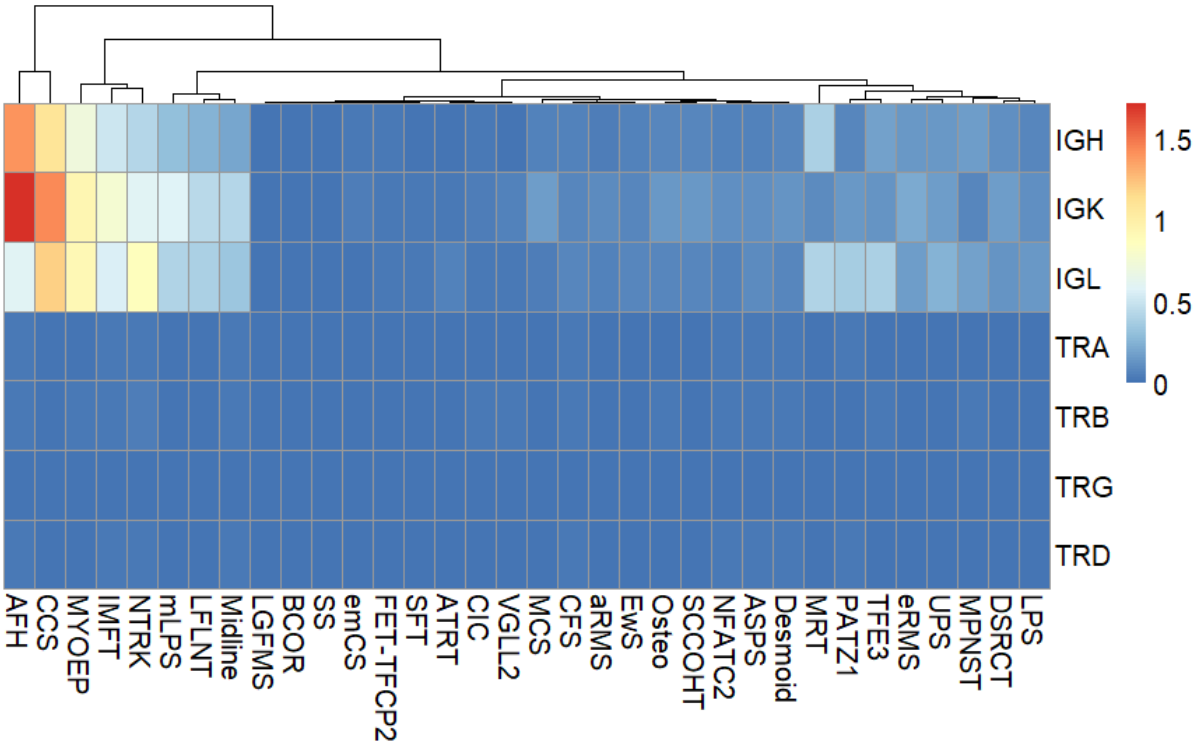

Figure 11: MiXCR median maximal clonality per diagnosis. Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by diagnoses (columns) using R package pheatmap (v1.0.12).

Interestingly, this heatmap looks also quite similar to Figure 9, meaning that diagnoses with higher numbers of clonotypes generally harbor more expanded clonotypes, which is also consistent with them having overall higher numbers of total TCR/BCR sequences (Figure 10). Though repertoire diversity and clonal expansion are not necessarily correlated (there could be a polyclonal infiltrate with no highly expanded clone, or inversely a high oligoclonal or monoclonal expansion), it seems that they generally are in sarcomas. One explanation could be that most sarcomas are poorly infiltrated, as shown by MCP-counter and MiXCR, and this accounts for both low diversity of repertoire and low clonal expansion; while in contrast the few diagnoses that have higher immune infiltration, exhibit statistically both higher diversity and higher clonal expansion than other sarcomas, even if this diversity or clonal expansion is generally quite moderate in comparison to other cancers.

To further explore intertumoral heterogeneity within the same diagnosis, I plotted heatmaps of MiXCR measures by individual sample for each diagnosis separately. As for the previous analyses, patterns of total number of BCR/TCR sequences and maximal clonality were quite similar to number of different clonotypes (all heatmaps not shown in this document), I will therefore only focus here on this last measure of repertoire diversity. While most diagnoses display relative homogeneity within all individual samples (most of the 34 heatmaps are therefore not displayed in this document), some subtypes of sarcoma exhibit a degree of intertumoral heterogeneity in terms of clonotype number, as detailed below.

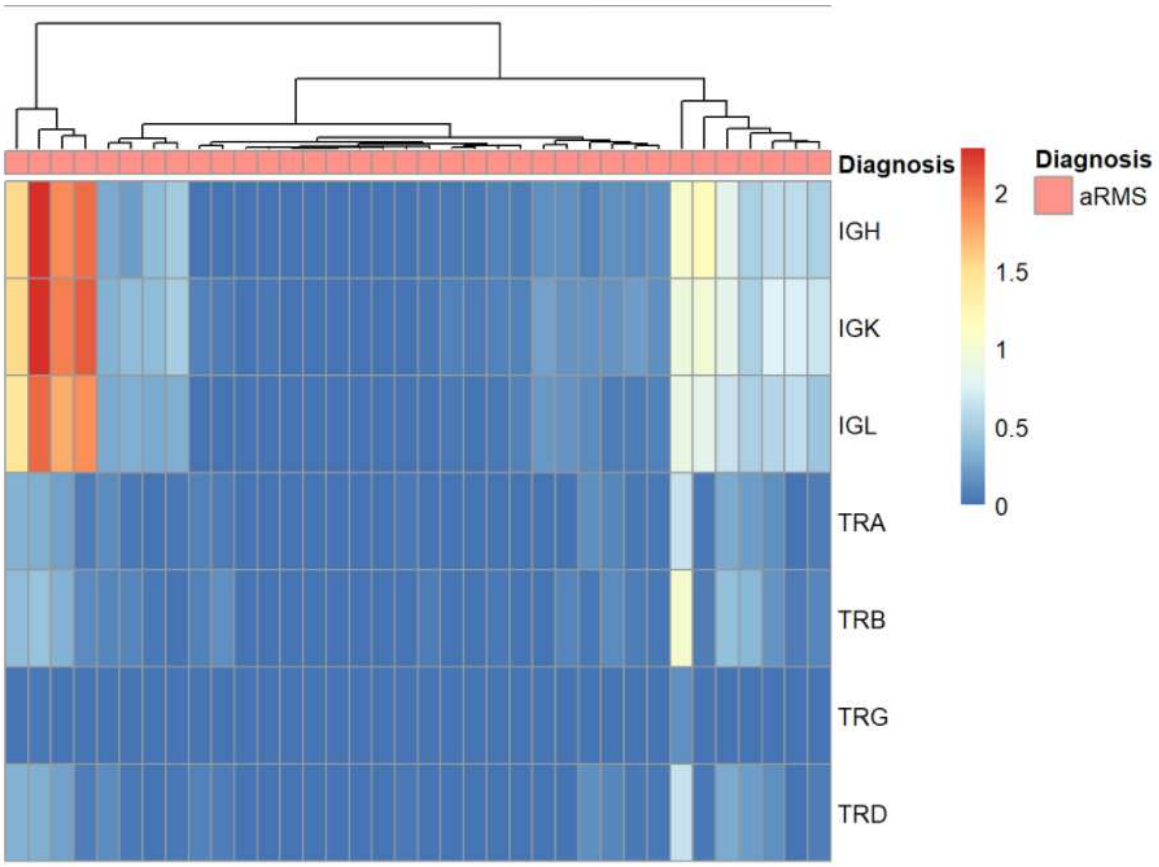

Figure 12: MiXCR number of clonotypes per sample in alveolar rhabdomyosarcoma (aRMS). Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by samples (columns) using R package pheatmap (v1.0.12).

For instance, alveolar rhabdomyosarcoma (aRMS) is generally lowly infiltrated by B and T cells, however some samples (left and right of heatmap) display rich repertoires especially of B cells (Figure 12). This is also the case for other lowly infiltrated sarcomas such as embryonal rhabdomyosarcoma (Figure 13) and synovial sarcoma (Figure 14).

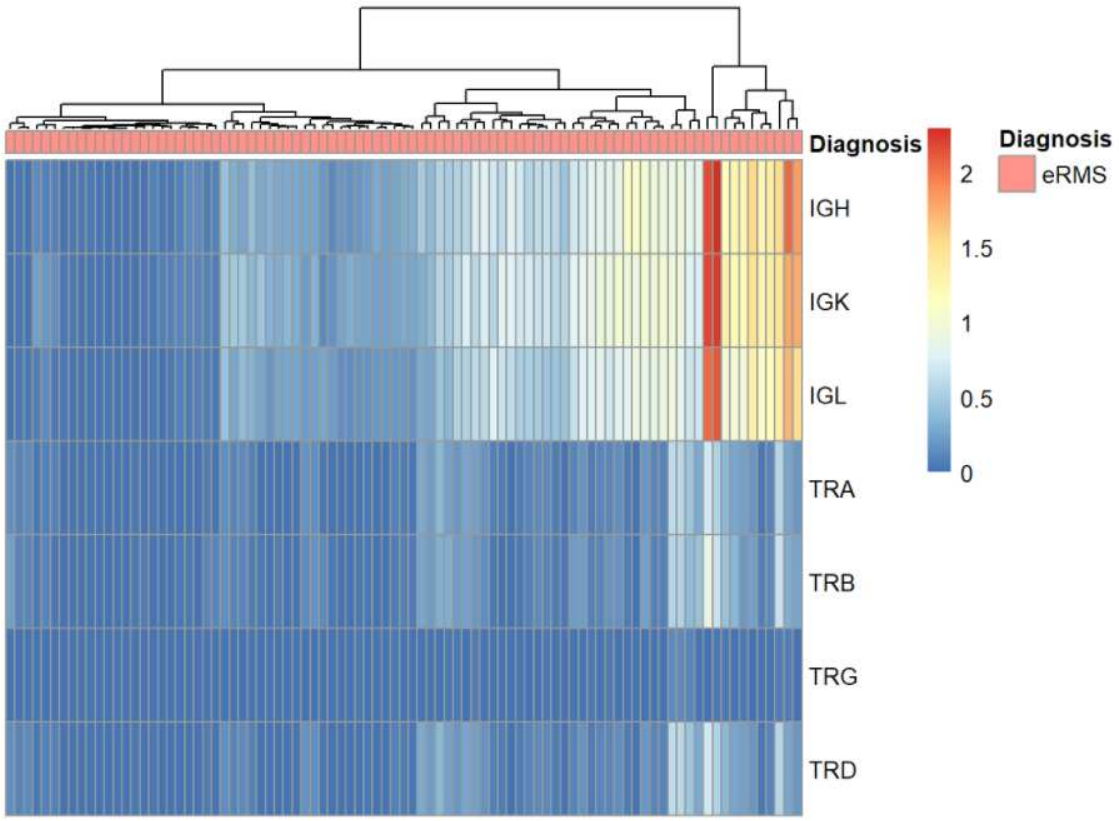

Figure 13: MiXCR number of clonotypes per sample in embryonal rhabdomyosarcoma (eRMS). Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by samples (columns) using R package pheatmap (v1.0.12).

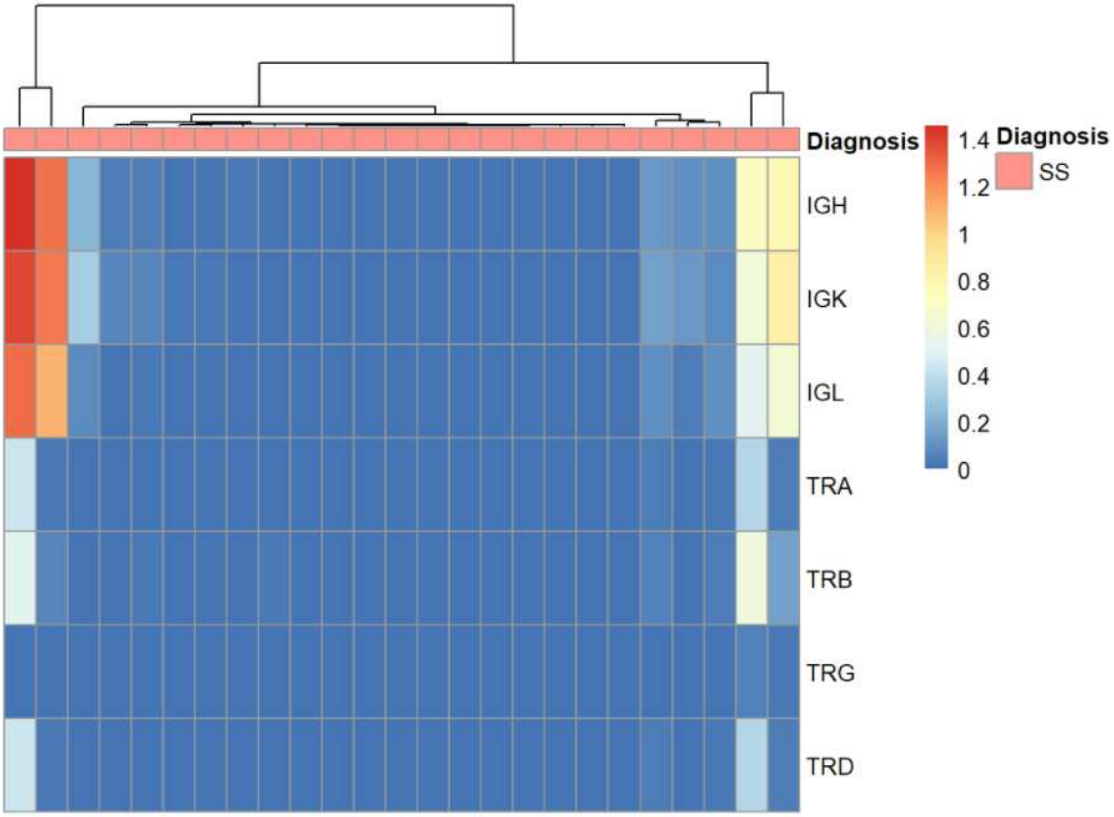

Figure 14: MiXCR number of clonotypes per sample in synovial sarcoma (SS). Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by samples (columns) using R package pheatmap (v1.0.12).

There is generally good agreement between observations made by MiXCR and MCP-counter, as demonstrated for the heatmaps of Ewing sarcoma (Figure 15), osteosarcoma (Figure 16), and UPS (Figure 17).

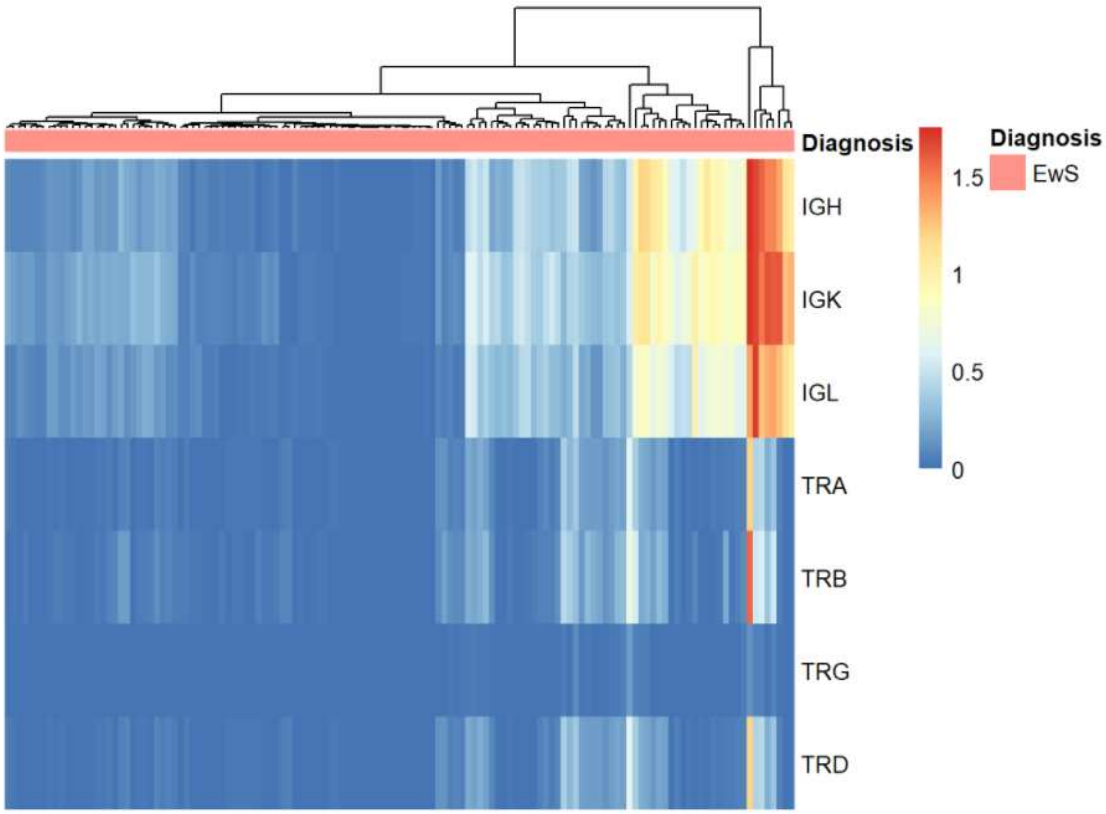

Figure 15: MiXCR number of clonotypes per sample in Ewing sarcoma (EwS). Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by samples (columns) using R package pheatmap (v1.0.12).

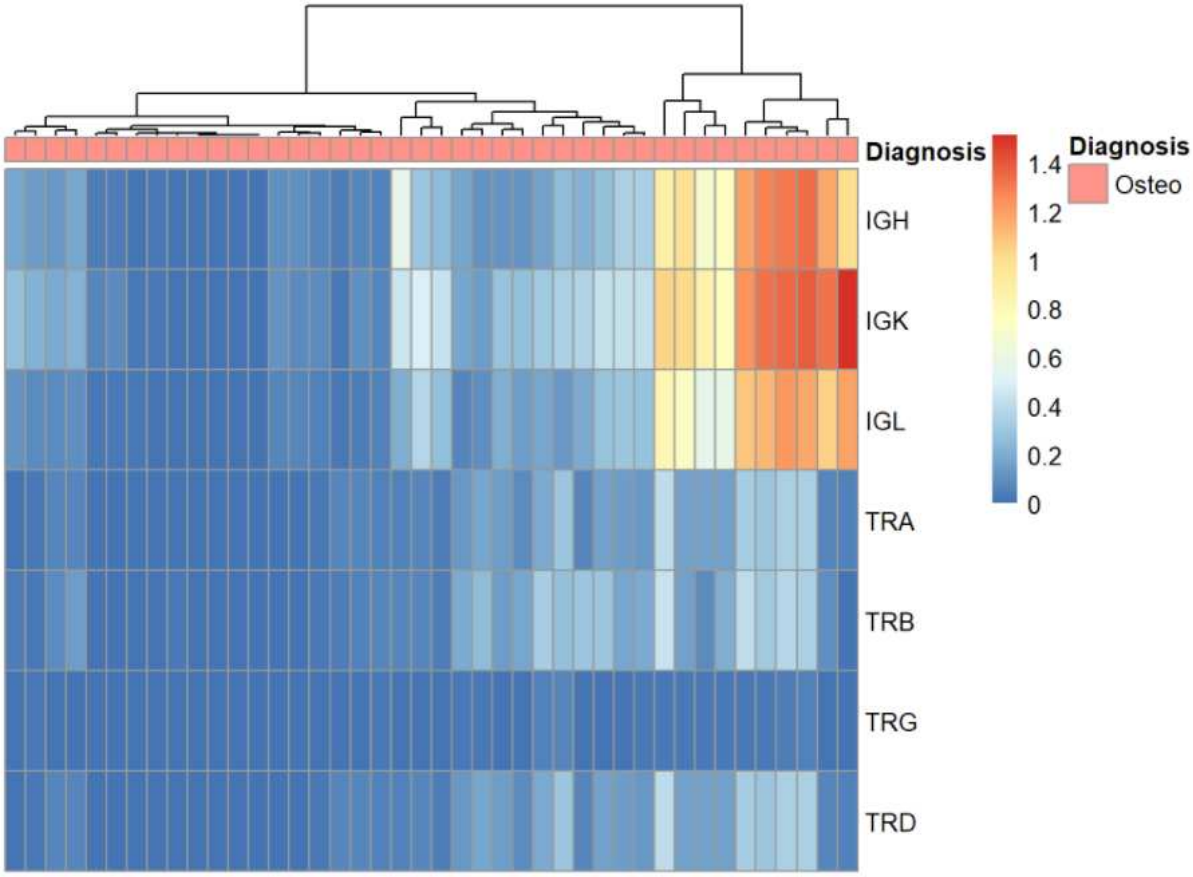

Figure 16: MiXCR number of clonotypes per sample in osteosarcoma (Osteo). Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by samples (columns) using R package pheatmap (v1.0.12).

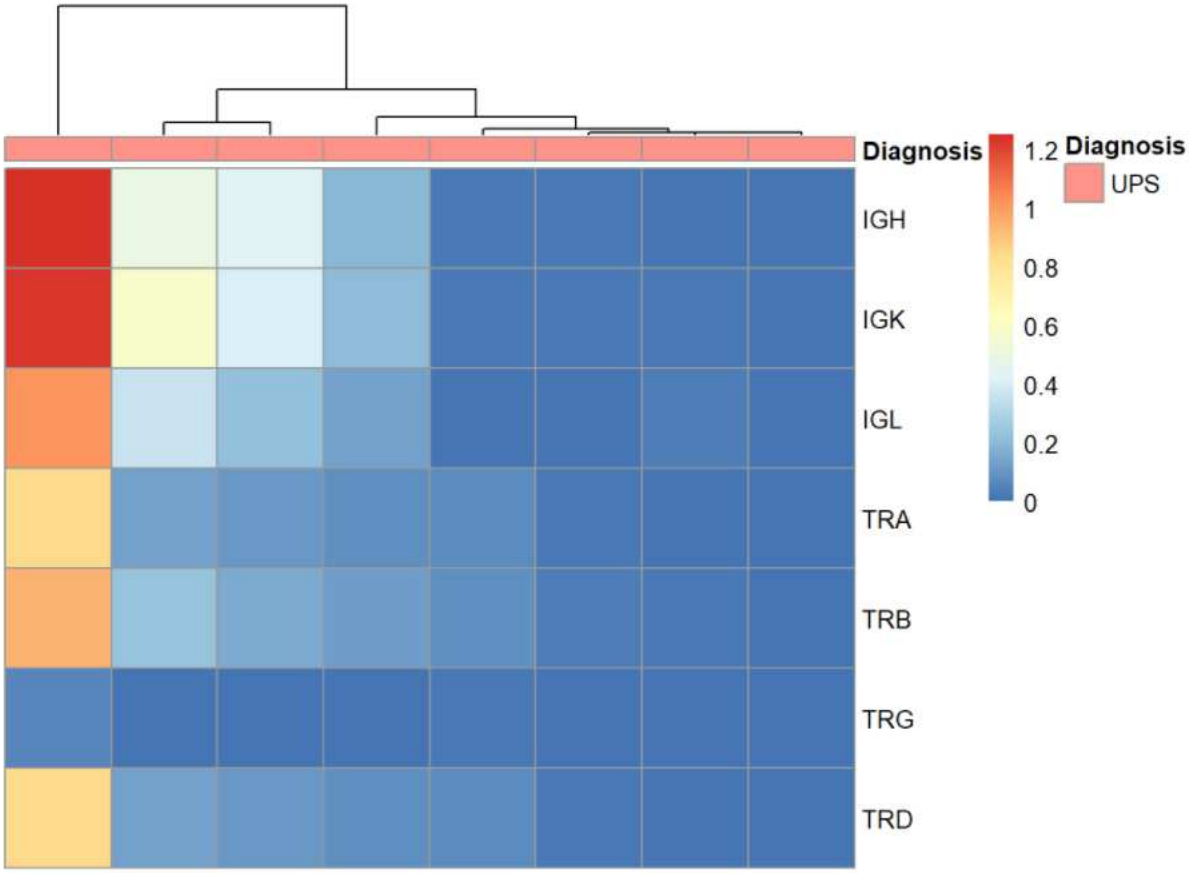

Figure 17: MiXCR number of clonotypes per sample in undifferentiated pleomorphic sarcoma (UPS). Values are normalized by million reads and log10-transformed with pseudocount 1. Hierarchical clustering by samples (columns) using R package pheatmap (v1.0.12).

In conclusion of this analysis by MiXCR of the UGS cohort, the global picture of the immune landscape that emerges is consistent with that already drawn by MCP-counter: pediatric sarcomas are generally lowly infiltrated by immune cells notably T and B lymphocytes, and they do not show high repertoire diversity or clonal expansion, except for a few diagnoses that are those already highlighted by MCP-counter, e.g. NTRK-fused sarcomas, inflammatory myofibroblastic tumors, angiomatoid fibrous histiocytomas. There exists however some degree of heterogeneity within some poorly infiltrated subtypes of sarcoma such as Ewing sarcoma, rhabdomyosarcomas and osteosarcoma, with a few samples showing higher lymphocyte infiltration, repertoire diversity and clonal expansion.

Discussion

We unfortunately do not have access to more information in individual samples to correlate these measures of immune infiltration to clinical variables such as treatment response and survival. It would be very interesting indeed to evaluate whether patients with higher infiltration by T or B cells, higher repertoire diversity or clonal expansion, are more responsive to treatments such as immunotherapies or show improved survival independently of administered therapy. In contrast to adult sarcomas studied in the TCGA and Petitprez et al. for which PD1-blockade immunotherapy has already been widely tested[12], pediatric sarcomas are still in need of further exploration concerning the potential efficacy of these immunotherapies. The relatively low number of immunotherapy trials has in part been due to the

widespread notion that pediatric sarcomas are poorly immunogenic as already discussed, and lowly infiltrated by lymphocytes as confirmed here in most samples. However, our observations that a subset of diagnoses and samples within otherwise poorly infiltrated subtypes of sarcoma can display higher values of immune infiltration do seem to warrant a more thorough exploration of the potential benefit of immunotherapies for these patients.

Characterization of the transcriptomic landscape of sarcomas

Introduction

Tumor bulk RNA-seq delivers insight into the tumor microenvironment as shown previously, but it also naturally allows genome-wide characterization of the "average" transcriptomic profile of tumor cells present in the sample. As discussed previously, our UGS cohort is unique in the sense that it is an unprecedented collection of RNA-seq of mostly rare diagnoses such as pediatric sarcomas. While the main subtypes of adult soft-tissue sarcoma have already been profiled by several high-throughput sequencing assays including RNA-seq, notably by the TCGA[39], there is to my knowledge no similar study yet involving RNA-seq of pediatric sarcomas.

One main theme in oncology which is particularly complex in sarcoma is the classification of disease: there are many subtypes of sarcoma as already discussed[4,45], diagnosis of which is based on multiple types of criteria (clinical, radiological, pathological, molecular). High-throughput molecular assays have in recent years been used to refine the classification of several cancers, as it allows finer characterization of the heterogeneity of some types of cancers that otherwise exhibit similar clinical or pathological characteristics[46]. By applying this strategy to sarcomas, one could hypothesize for instance that a molecular assay such as RNA-seq could 1) enable the delineation of a type of sarcoma into different molecular subtypes, or 2) demonstrate unexpected molecular similarity between different diseases. Indeed, there are already some examples of these occurring in the field of sarcomas. For the first case, a well-known example is the distinction of rhabdomyosarcomas between alveolar (characterized by a PAX-FOXO gene fusion) and embryonal (PAX-FOXO negative) rhabdomyosarcoma, or more recently the identification of many Ewing-like tumors with different translocations than Ewing sarcoma[20,21]. For the second case, one could mention the interesting observation that clear cell sarcoma and angiomatoid fibrous histiocytoma, two different subtypes of sarcoma as characterized by pathologists, nonetheless harbor the same characteristic molecular alteration (a EWSR1-CREB1/ATF1 gene fusion).

Another related question that could be addressed by a molecular high-throughput assay such as RNA-seq is the diagnosis of an individual patient: given a tumor bulk expression profile, what is the correct diagnosis? This is closely related indeed to the previous issue of classification of sarcomas, as the diagnosis could itself be potentially improved with a refined molecular picture of the classification of sarcomas.

Unsurprisingly, these questions have already been widely addressed in oncology, notably by a series of landmark studies using DNA methylation arrays to characterize the molecular classification of some types of cancer including sarcomas, and to help in the diagnosis of individual patients. For this, techniques of machine learning (often popularly referred as "artificial intelligence") have been of paramount importance. One high-profile study used DNA methylation arrays to classify brain tumors[47] and a random forest machine learning algorithm to predict the diagnosis of new patients. A recent work from the same team in Heidelberg used the same technique for sarcomas[48]. These studies demonstrated capability of molecular assays to define a refined classification of cancers and of machine learning for automatic prediction of diagnosis based on these assays. The advantages of using DNA methylation for these studies were both 1) technical: standardized arrays are available and relatively easy to perform; and 2) biological: DNA methylation is generally well conserved in tumor cells as compared to the cell-of-origin, allowing a biologically relevant classification mirrored by the different cells-of-origin of different types of cancer[49,50]. However, the results of DNA methylation arrays are difficult to interpret in terms of functionality: the classification in these studies depends on methylation "probes" that have no straightforward biological meaning. This does not help interpretability of the classification and diagnosis prediction; it also reinforces the "black-box" nature of the machine learning algorithm.

In contrast, RNA-seq is more suited for interpretability, as it measures expression levels of mostly well-characterized genes that can also be organized into biologically relevant functional sets and pathways. It also delivers a high amount of information that is comparable in number of dimensions to DNA methylation arrays, and its result could therefore be another molecular assay to be used for the classification and diagnosis prediction problems. Biologically, while the transcriptomic profile may generally be less related to the cell-of-origin than DNA methylation, it nonetheless reflects the various oncogenic processes at play in the tumor cells and can thus potentially be used to finely characterize different subtypes of cancer[51].

Transcriptomic landscape of the Institut Curie dataset

As discussed previously, RNA-seq was already in use at the Institut Curie to help in the diagnosis of pediatric sarcomas since 2015, and while the most important information from this assay for diagnosis was the presence of characteristic fusion transcripts, the genome-wide expression profile was also being used as a tentative tool to help in the diagnosis of individual patients. Specifically, a simple procedure of hierarchical clustering based on the transcriptomic profile was performed and the diagnosis of individual patients could be estimated by observing their position within the clustering.

To go further in this direction of transcriptomic characterization of sarcomas, I started by visualizing in two dimensions the overall picture of the sarcoma transcriptomic landscape in our UGS cohort. For this, I once again used the 666 samples described in Table 1. To enable visualization of high-dimensional data, there are several techniques of dimensionality reduction that attempt to encode the maximum amount of the original high-dimensional information into a space of reduced number of dimensions. One well-known and often used dimensional reduction technique is principal component analysis (PCA), a linear technique that attempts to find in the data the most representative principal components, i.e. linear combinations of all original features[52]. On the other hand, other techniques can make use of non-linear relationships in between the features to better capture the overall structure of the data, especially when transforming it into spaces easily visualizable by humans, e.g. in two dimensions. Indeed, PCA is not as suited for this visualization task since it often needs many more than two PCs to capture a sizable fraction of the total variance present in a dataset. Powerful non-linear dimensional reduction techniques for visualization, especially popular in single-cell analysis in biology, are t-distributed stochastic neighbor embedding (t-SNE)[53] and Uniform Manifold Approximation and Projection (UMAP)[54]. Both methods are well suited to represent in lower dimensional embeddings the original local neighborhoods of individual points, but UMAP seems to be better at conserving global distances found in the original space[55]. Here, I used UMAP reduction of the original high-dimensional transcriptomic space into two dimensions to visualize it. Note that in all following analyses, I first reduced the original space (> 50 000 dimensions, i.e. all GENCODE v19 entries in the RNA-seq count matrix) to the 5000 most variable features in order to facilitate computation and discard non-variable features that may only reflect technical noise.

Method: RNA-seq reads were adapter-trimmed with Atropos (v1.1.21), aligned to the human reference genome (hg19) with GENCODE version 19 as the reference gene annotation with the use of STAR (v2.7.0e)[40] and quantified with the GeneCounts algorithm. Raw counts were normalized to transcripts per million (TPM) and log-transformed with pseudocount 1. The 5000 most variable features after a variance-stabilizing transform were selected with the package Seurat (v3.1.4) using FindVariableFeatures (option “vst”) and values were scaled using ScaleData. RunUMAP was then used to calculate the UMAP coordinates for all samples and plotted using DimPlot.

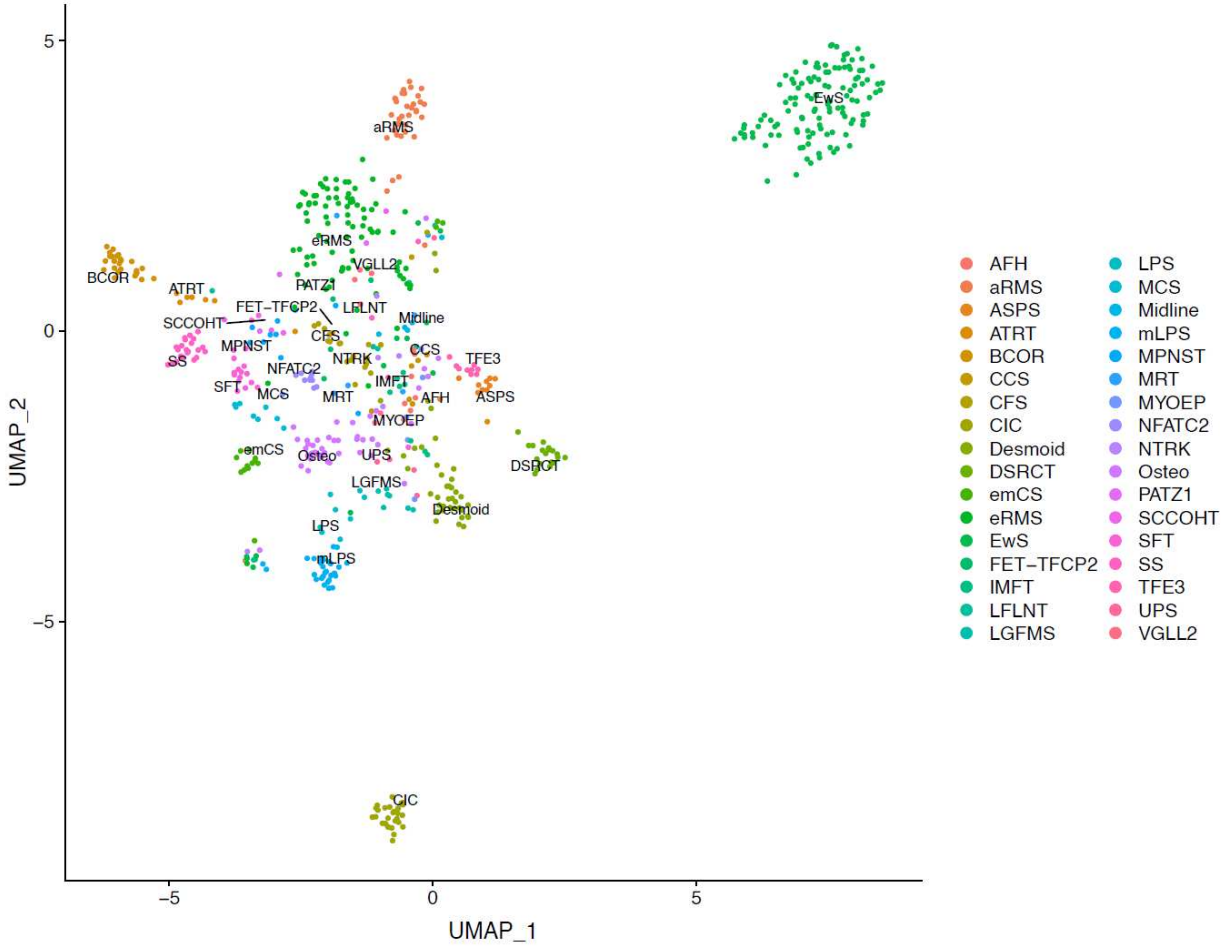

Figure 18: UMAP of sarcomas using 5000 most variable features.

The resulting UMAP (Figure 18) on all variable features already delivers several insights: 1) Most diagnoses can be readily identified in one specific location of the UMAP, though many have overlapping samples with other diagnoses, especially in the middle-left concentration of samples. 2) Some sarcoma subtypes are clearly separated from others, such as Ewing sarcoma (EwS), CIC-fused sarcoma (CIC), and to a lesser degree BCOR-rearranged sarcomas (BCOR) and alveolar rhabdomyosarcomas (aRMS). These are all translocation-related sarcomas characterized by a specific fusion gene and reflect the specific transcriptomic profiles of these subtypes induced by the oncogenic chimeric transcription factor deriving from the chromosome translocation. Other diagnoses tend to cluster together in the middle-left of the UMAP, though some are more distinct at the periphery, notably other translocation-related sarcomas such as desmoplastic small round cell tumors (DSRCT) and myxoid liposarcomas (mLPS).

This first UMAP is insightful but has some limits: it is calculated from 5000 features and therefore does not give an easy way of interpreting it, also it is not straightforward to use as a diagnostic tool as such. One way to enhance interpretability and facilitate diagnosis prediction would be to further reduce the space of features and "encode" the high-dimensional transcriptomic profile in a more human-interpretable format. A thought experiment helps to conceptualize the following ideal tool: a dimensional reduction technique that extracts from the original space of features, one unique "meta-feature" very specifically associated to each diagnosis, which would give in this space of ideal meta-features a very clear separation between all diagnoses, as well as a very straightforward way of classifying new samples just by calculating each of these discriminating meta-features. Moreover, this tool would potentially be interpretable if each meta-feature could be decomposed in its contributing genes or functional pathways. However, reality is more complex,

and it is not obvious that we can in any way clearly delimit specific diagnoses without any overlap between them. For instance, some diseases might be variants on a continuous spectrum of the same diagnostic entity and intrinsically non-separable in transcriptomic space.

Nonetheless, I continued in this direction and tried to further meaningfully reduce dimensions of the dataset by using PCA, one of the most widely used, well-known and best performing linear dimensional reduction techniques for high-dimensional data. In this approach, I reduced the data to 50 dimensions (arbitrary "round" number, chosen to accommodate the number of diagnoses: 34) using the 50 first principal components, and once again used UMAP to visualize the resulting 50-dimensional transcriptomic space.

Method: Same as for the previous UMAP, except for the following steps after the scaling of data: PCA was run on the 5000 most variable features using RunPCA (Seurat), and RunUMAP was then used to calculate the UMAP coordinates based on 50 PCs for all samples.

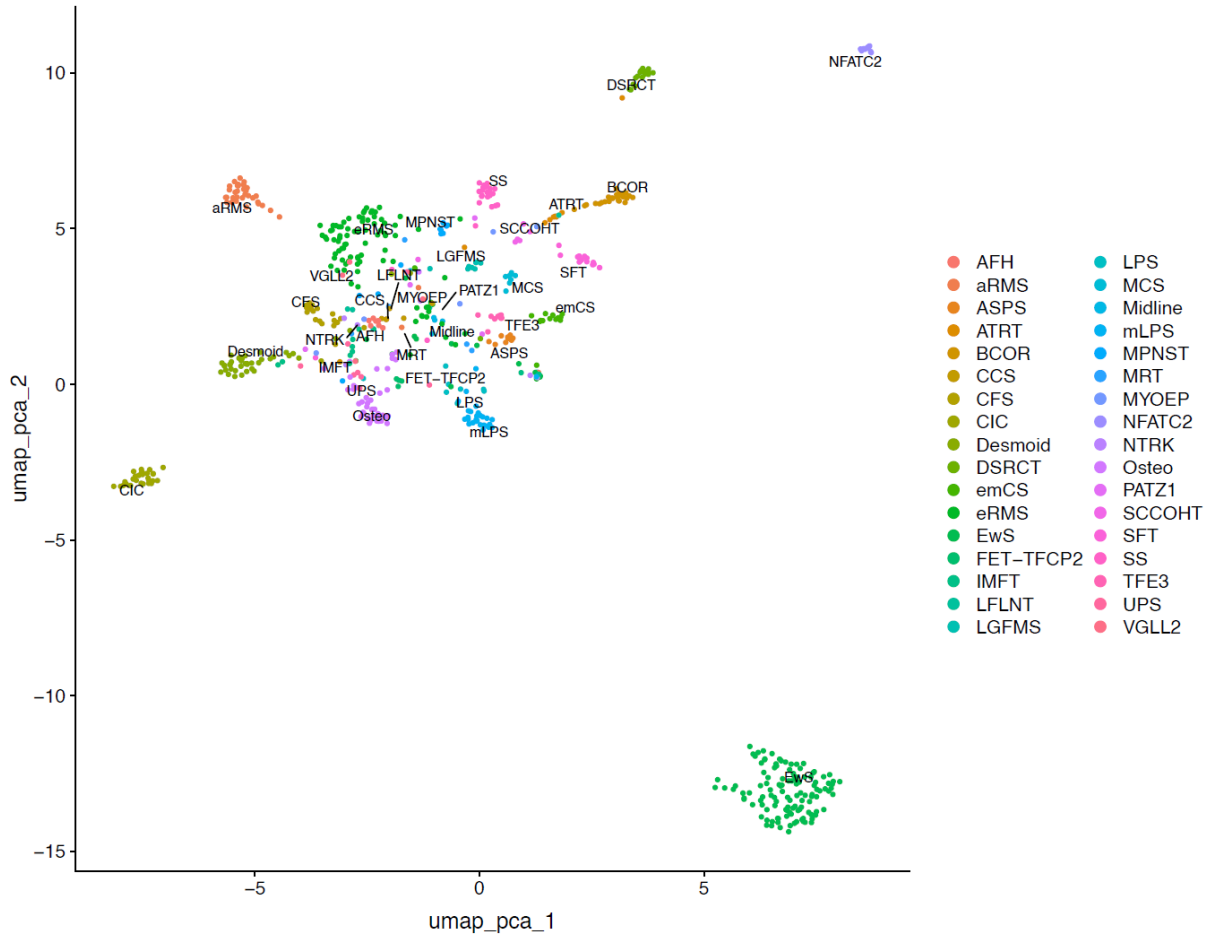

Figure 19: UMAP of sarcomas using 50 principal components calculated by PCA.

The resulting UMAP (Figure 19) has many similar characteristics with the previous one, however several diagnoses tend to be more clearly separated, especially DSRCT and notably EWSR1-NFATC2 sarcomas (NFATC2), which were not distinct in the previous UMAP. This shows that at least for some diagnoses, reducing the transcriptomic space with PCA is also likely to improve classification and diagnosis prediction.

The variational autoencoder (VAE)

Considering these preliminary results, I could have chosen to use PCA to further explore classification, diagnosis prediction and interpretability for this dataset. However, I instead investigated other dimensional reduction techniques that were likely to be better-performing and potentially more biologically interpretable than PCA. Specifically, PCA is a linear dimensional reduction technique, but most biological phenomena are the result of complex, often non-linear relations between interacting biological factors[56] such as gene expression. I therefore chose to test a non-linear dimensional reduction technique on this dataset. Recently, methods of machine learning and deep learning have been used with success to extract meaningful non-linear features from high-dimensional datasets[3]. Specifically, a method from the field of deep learning called the variational autoencoder (VAE)[57] has been used successfully for this purpose. As with many methods of deep learning, it has first been used for image data: for instance, the VAE is able to extract from pictures of faces some meaningful non-linear features such as gender, age, skin color, even smile and glasses-wearing. In contrast, linear techniques such as PCA are unable to extract this kind of meaningful information, since these extracted features are intrinsically non-linear in the space of original features (pixels).

The VAE is based on an architecture of two neural networks: the "encoder" and "decoder" neural networks. To train these, the first objective of the VAE is to accurately "reconstruct" a high-dimensional dataset from a corresponding low-dimensional "encoded" representation. To achieve this optimally, the "encoder" network is trained to learn a low-dimensional representation, while the "decoder" network is concurrently trained to reconstruct in the most accurate way the original high-dimensional dataset from this low-dimensional representation. At the end of training, the optimal neural networks have thus learnt a good low-dimensional "encoding" of the original high-dimensional dataset. Using these two neural networks it is then possible to "encode" new data into this optimal low-dimensional representation. In most cases, one is not necessarily interested in the best reconstruction of an original dataset (an exception being image compression), but rather in the low-dimensional encoded space which allows not only manipulation of the dataset in low dimensions without too much loss of information, but also potentially a more meaningful and interpretable representation of the data. Indeed, each of the low-dimensional features of the encoded space is a non-linear combination of original features and could potentially be a high-level meaningful feature, such as smile or gender in face pictures. This requirement is present in the objective of the VAE, as reflected in the loss function used during its training, which is composed of a first part that incentivizes accurate reconstruction, but also a second "penalty" term (or regularization term in machine learning jargon) that favors mathematically "smooth" and regular low-dimensional spaces that have nice properties for subsequent manipulation and interpretation. This second "aesthetic" term is also motivated by the fact that in domains such as imaging and biology, it may be assumed that most of the dimensions in the original space are not independent and can be reduced to a much smaller number of meaningful factors. Constraining the encoded space to be more "regular" also attenuates the risk of overfitting, which is unavoidable for highly flexible machine learning procedures such as neural networks, that can accommodate and learn any high-dimensional pattern if they are allowed to explore all the possibilities offered by the fitting procedure[58]. In a sense, we impose on the VAE a preconceived notion of what an optimal encoded space should be: this is why the VAE has some connections with so-called Bayesian machine learning[57], i.e. it uses *prior* knowledge of the best parameter space to learn the data.

In a sense, this procedure of encoding with a VAE is similar to what is accomplished by classical techniques of dimensional reduction that search for the "best" combinations of original features: for instance PCA is finding the best set of linear combination of features to "encode" the original data into a low-dimensional PC space. One could therefore think of the VAE as a kind of "non-linear" PCA, where each "non-linear principal PC" is potentially more meaningful than a linear PC since it can accommodate non-linear relationships in the original dataset.

I became very interested by this new technique due to its theory but also because some authors had recently started to apply it to biology: notably a paper[59] that encoded RNA-seq data from the TCGA into a low-dimensional space, with some of the encoded features being associated to biologically meaningful variables such as gender or immune infiltration. This seemed very promising, and I therefore wanted to try to encode the sarcoma RNA-seq dataset with a VAE into a potentially biologically meaningful low-dimensional encoded space: not only may I then be able to extract biologically meaningful features, but also it would potentially improve and facilitate the classification and diagnosis prediction tasks, as a dimensionality reduction technique such as PCA had already hinted at. In analogy to the PCA approach, I trained a VAE to find a 50-dimensional encoded space. A schematic of the VAE is in Figure 20.

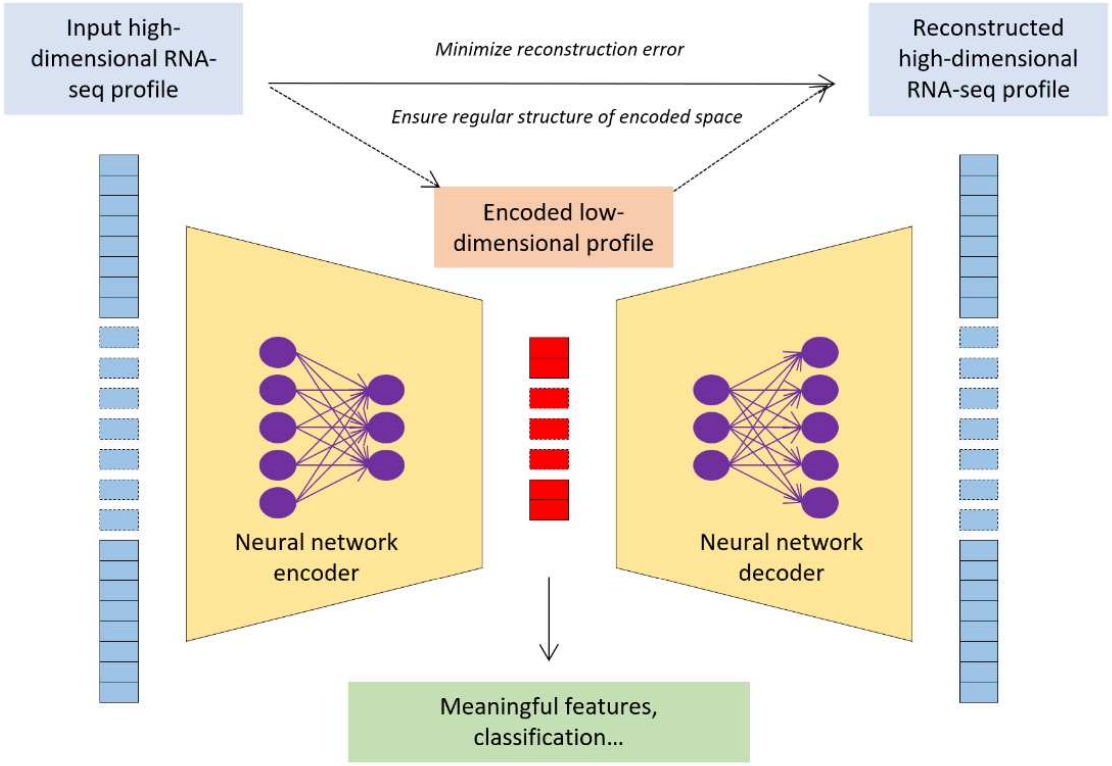

Figure 20: Schematic of the VAE for RNA-seq.

Method: The VAE was trained on the same 5000 most variable features as previously. The encoder neural network was fully connected and one layer deep, with an encoding intermediate layer of 50 neurons, and the decoder network was also fully connected and one layer deep. The latent space was therefore 50-dimensional. Input features were scaled between 0 and 1 before training (divided by maximum value of the corresponding feature). The VAE was implemented and trained with Keras v2.2.4 (TensorFlow v1.14.0), optimized with Adam, batch-normalized. Activation was relu (rectified linear unit) for the encoding layer and sigmoid for the decoding layer. Learning rate was 0.0005 and the model was trained for 50 epochs with no evidence of overfitting.

To visualize the encoded 50-dimensional space after training, I once again used UMAP dimensionality reduction in two dimensions (Figure 21).

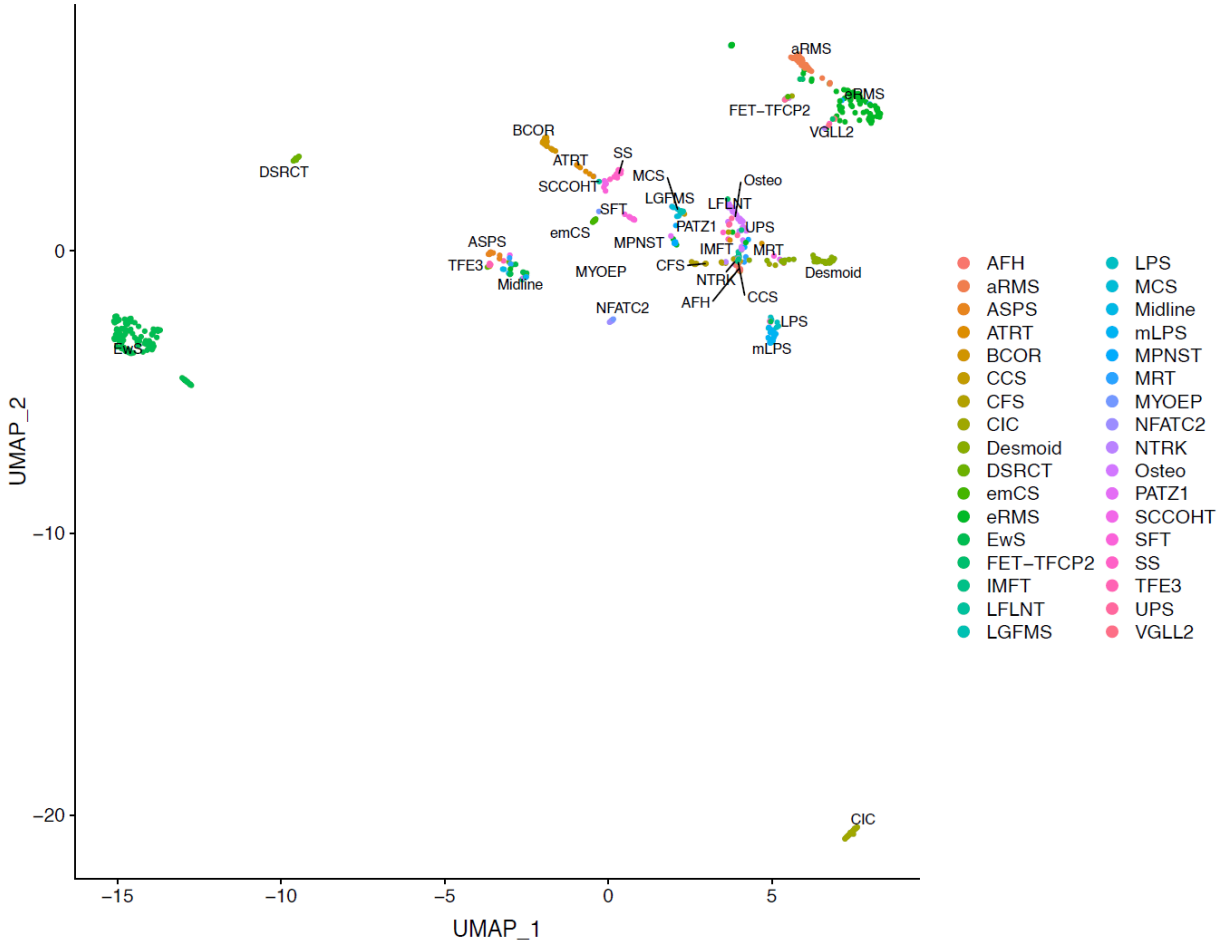

Figure 21: UMAP of sarcomas using the 50 encoded features of the VAE.

The VAE-encoded space as visualized by UMAP displays interesting properties already found in the two previous UMAPs (all variable features and PCA): some diagnoses are clearly distinct from others, such as Ewing sarcoma (EwS), CIC-fused sarcomas (CIC). However, in contrast to PCA and especially the non-reduced space, several diagnoses clearly stand out as distinct this time: desmoplastic small round cell tumors (DSRCT), EWSR1-NFATC2 sarcomas (NFATC2), BCOR-rearranged sarcomas (BCOR), myxoid liposarcomas (mLPS), desmoid tumors. It is interesting to note that at the top right corner, all rhabdomyosarcomas (alveolar and embryonal) and related diagnoses (VGLL2-fused and FET-TFCP2) are closely connected but still distinct. This is also the case for instance between alveolar soft part sarcoma (ASPS) and TFE3-renal cell carcinoma (TFE3) which are two distinct pathologic diseases but share the same fusion gene. The rest of the diagnoses, which are mostly the non-translocation-related sarcomas, as previously tend to cluster together and overlap in the middle, though an underlying structure can still be seen. Overall, even though UMAP is only a qualitative way of comparing between methods, it seems that encoding by the VAE allows a clearer separation of diagnoses than using no dimensional reduction or PCA.

This first evaluation of the VAE was encouraging for the tasks of sarcoma classification and diagnosis: indeed clearly separating different entities in the VAE space is a prerequisite for both tasks, and above all this is accomplished using “only” 50 dimensions as opposed to the original > 50,000, therefore allowing easier manipulation and interpretation of the data. Nonetheless, I did not venture much further for the classification of sarcomas, due to the following reasons considering the UMAP in Figure 21: the separate entities corresponded overall to the previous diagnostic labeling, and apart from confirming that some sarcomas, especially translocation-related sarcomas, have very distinct transcriptomic profiles from the rest,

I did not see either novel clusters of unknown significance or conversely separation of the same diagnostic label into distinct sub-entities. One important insight to me though was the observation that some sarcomas, especially non-translocation-related ones, tend to cluster together and overlap, hinting that these may be transcriptomically very similar and even impossible to distinguish. This fact is corroborated in the clinic where sarcoma diagnosis is particularly difficult because of the overlapping nature of the presentations of different diseases. In fact, one could potentially ask the relevance of distinguishing these diseases, as they may for some be different states of a same continuous spectrum of disease. For instance, it is known that undifferentiated pleomorphic sarcoma can take differentiation characteristics of bone, cartilage or muscle, and the potential plasticity of these tumors could account for the fact that they have a common transcriptomic backbone that can evolve between different forms of the same disease. One is even tempted to make a parallel with normal development, where stem cells can take on characteristics of different tissues, and specifically in the case of "mesenchymopoesis", a mesenchymal stem cell can become an adipocyte, a muscle cell, a chondrocyte or an osteocyte[60].

RNA-seq for diagnostic prediction

For diagnosis prediction, a first approach takes advantage of the visualization by UMAP: a new sample is simply projected on the two-dimensional representation, since the UMAP transformation is learned once based on the original features and can then be applied to any sample in the original feature space. Since the "best" UMAP seems to be the one after encoding by VAE, this is the approach that I used for a first guess at diagnosis prediction: encoding a new RNA-seq sample into the low-dimensional VAE feature space, and then projecting this encoding on the UMAP. Unsurprisingly, this basic method works well for "easy" cases, i.e. diagnoses localizing in a distinct part of the UMAP, such as Ewing sarcoma and most other translocation-related sarcomas. For other cases, when the sample ends up in the middle within the overlapping diagnoses, it is less straightforward to predict the correct diagnosis, though the underlying structure often makes for a good guess. In fact, I have implemented an algorithm of this sort that has now been included in the diagnostic pipeline of the UGS for nearly two years, where eight new patients are processed each week, and this tool is an addition to the other tests for helping with the diagnosis. However, the added value of this prediction method is often modest: the "easy" cases are also easy for standard techniques, especially when RNA-seq is used to search for translocations, while difficult cases tend to go into the middle of the UMAP, thus not necessarily improving the diagnostic hypotheses.

Before definitely downplaying the added value of this method, I wanted to test a more systematic method of prediction than simple visualization of the UMAP. Indeed, UMAP visualization is constrained to show information in only two dimensions and can thus potentially hide important insights from the 50-dimensional original encoded space. I therefore used a random forest[61], a classical high-performing machine learning technique which is a tree-based classifier, based on the 50 features encoded by the VAE.

Method: The RF classifier was trained using the RandomForest package in R, with 5000 trees, mtry= 10, using as input the 666 samples encoded in 50 features by the VAE, with "Diagnosis" as label.

A confusion matrix of this RF classifier is revealing (Figure 22).

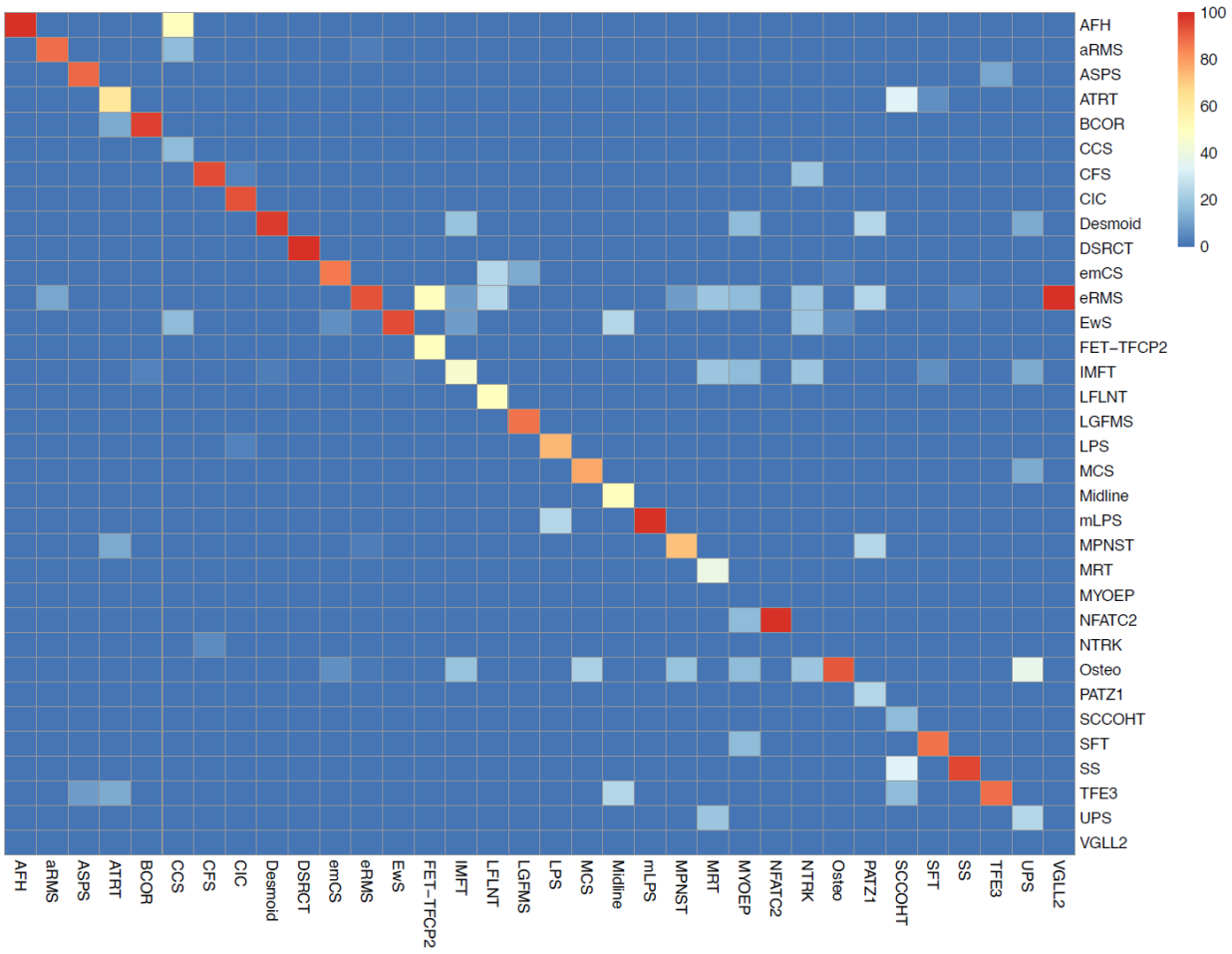

Figure 22: Confusion matrix of a sarcoma random forest classifier based on 50 VAE features. Values are in percentage. Rows correspond to the predicted diagnoses and columns to the true diagnoses.

Diagnoses which are "easy" to predict, i.e. have near to 100% correct prediction on the diagonal of the confusion matrix, are translocation-related sarcomas such as Ewing sarcoma (EwS), CIC-fused sarcomas (CIC) or desmoplastic small round cell tumors (DSRCT). In contrast, some diagnoses tend to be confused with one another, corresponding to overlapping diagnoses on the UMAP such as osteosarcoma (Osteo) and undifferentiated pleomorphic sarcoma (UPS). For most cases, the RF classifier is indeed equivalent to a "visual" prediction by UMAP.

As for interpretability, I first visualized a heatmap of the 50 VAE encodings for all diagnoses (Figure 23).

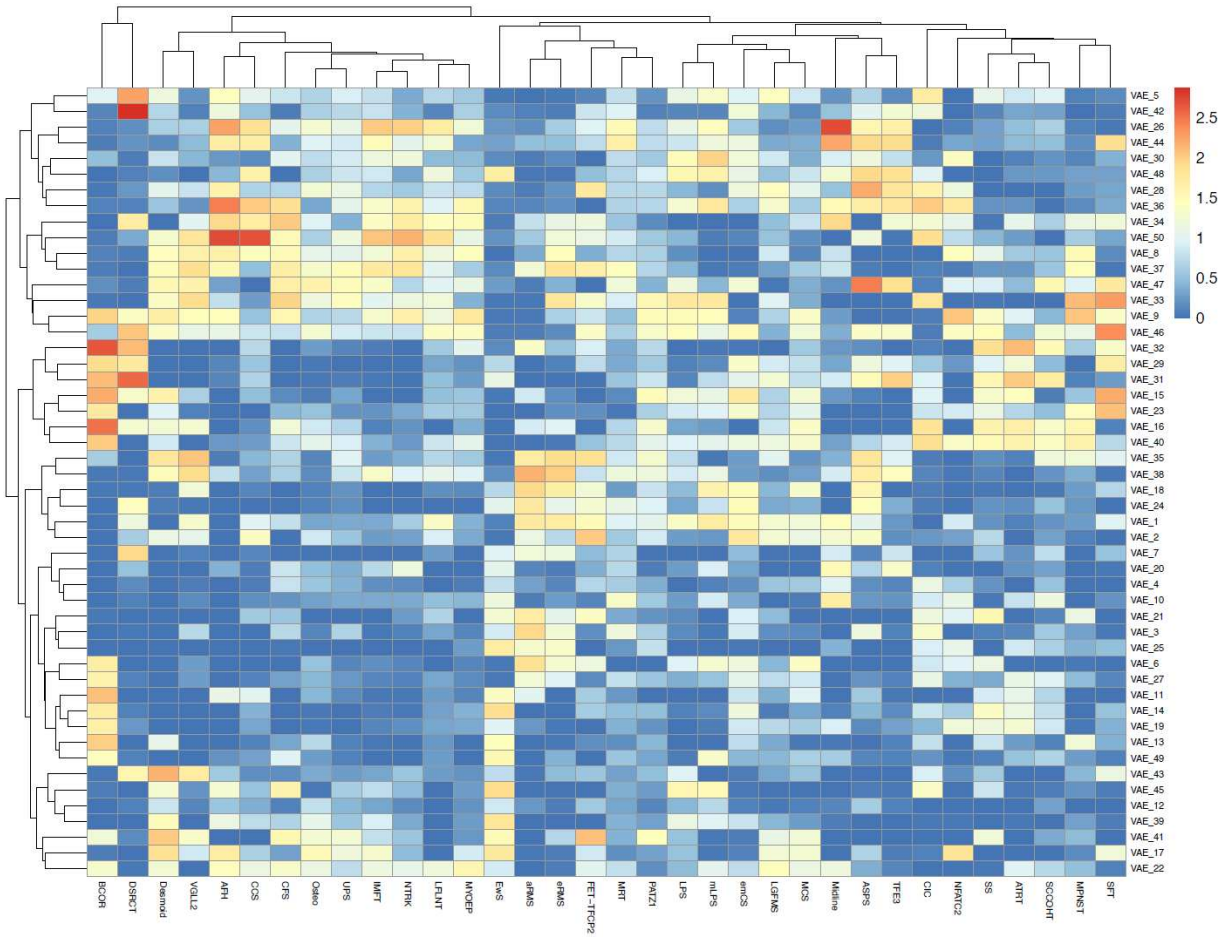

Figure 23: Heatmap of mean VAE feature values for sarcomas by diagnosis.

This is far from an “ideal” heatmap which would have shown features lighting up in only one or a few specific diagnoses: most features have non-zero values in more than one diagnosis, which is not inconsistent with the hypothesis that some diagnoses may be viewed as different states of a single continuous spectrum in transcriptomic space. There are however some features more specifically associated to only one or a few diagnoses, and without going into the detail of all features I present here some examples.

To interpret a feature, we must remember that each one is the result of a non-linear combination of the initial transcriptomic features, and the associated weights in the decoder network of the VAE give us an idea of the genes most contributing to this feature. I therefore extracted and explored the genes with highest “weights” for each feature. As a first example, consider VAE feature number one (VAE_1): its value is plotted in all samples grouped by diagnosis in Figure 24.

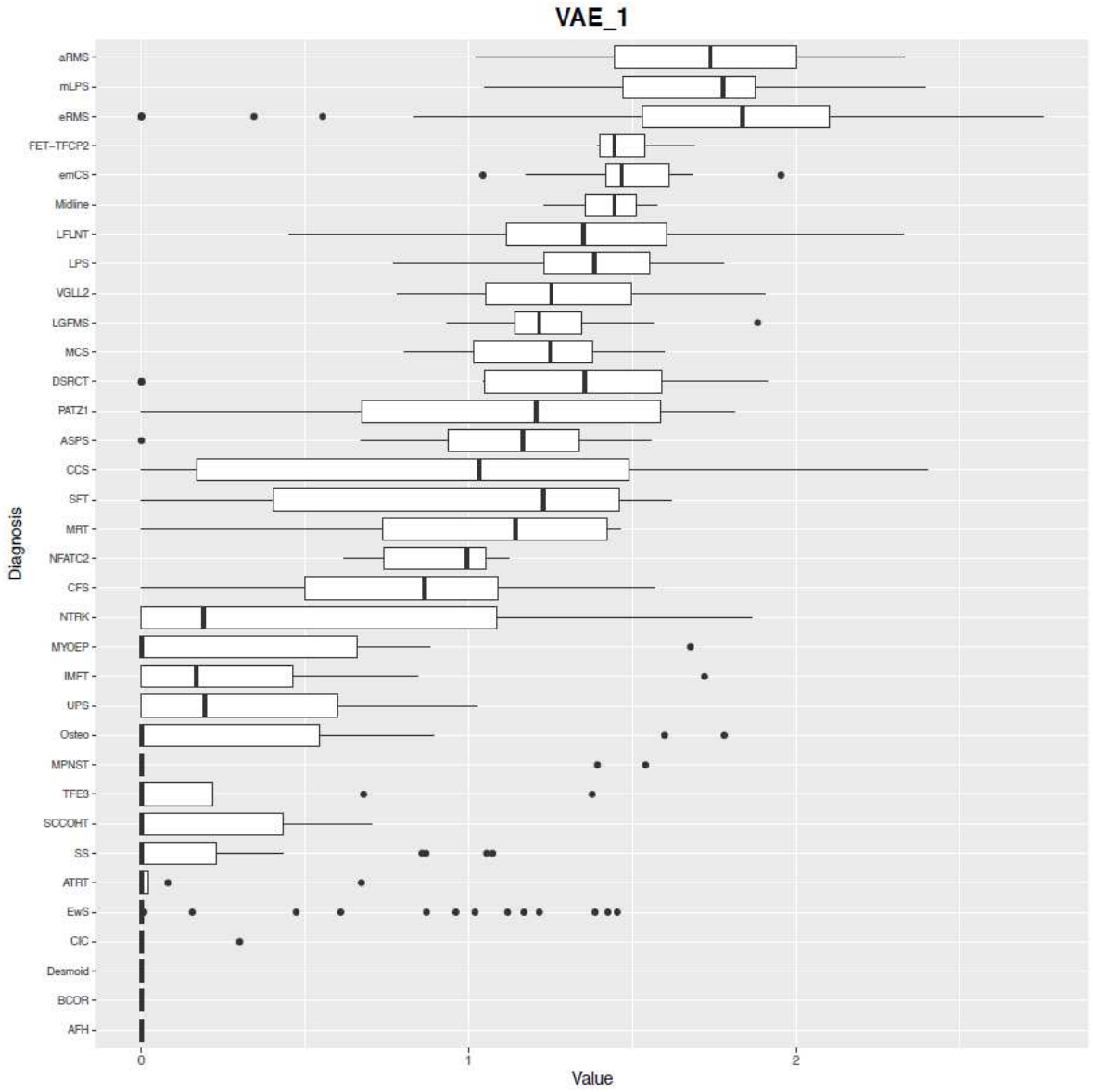

Figure 24: VAE feature 1 (VAE_1) values in all samples grouped by diagnosis.

This feature is especially expressed in rhabdomyosarcomas and related diagnoses (aRMS, eRMS, FET-TFCP2, VGLL2). Its top 20 highest-weight genes are: *SHD, SGCA, IGF2, CHRND, MIR483, MYL4, RAPSN, IGF2-AS, RP1-302G2.5, DLK1, RP11-2E11.9, VGLL2, KREMEN2, C20orf166-AS1, PLA2G16, CHRNG, MYOG, ITIH5, NNAT, MYOD1*. It is notable that there are multiple genes related to muscle tissue (*MYL4, MYOG, MYOD1*), as well as the eponymous gene *VGLL2*.

As another example, take feature number 42 which on the heatmap of mean encodings (Figure 23) is especially high in desmoplastic small round cell tumors (DSRCT), as shown in Figure 25.

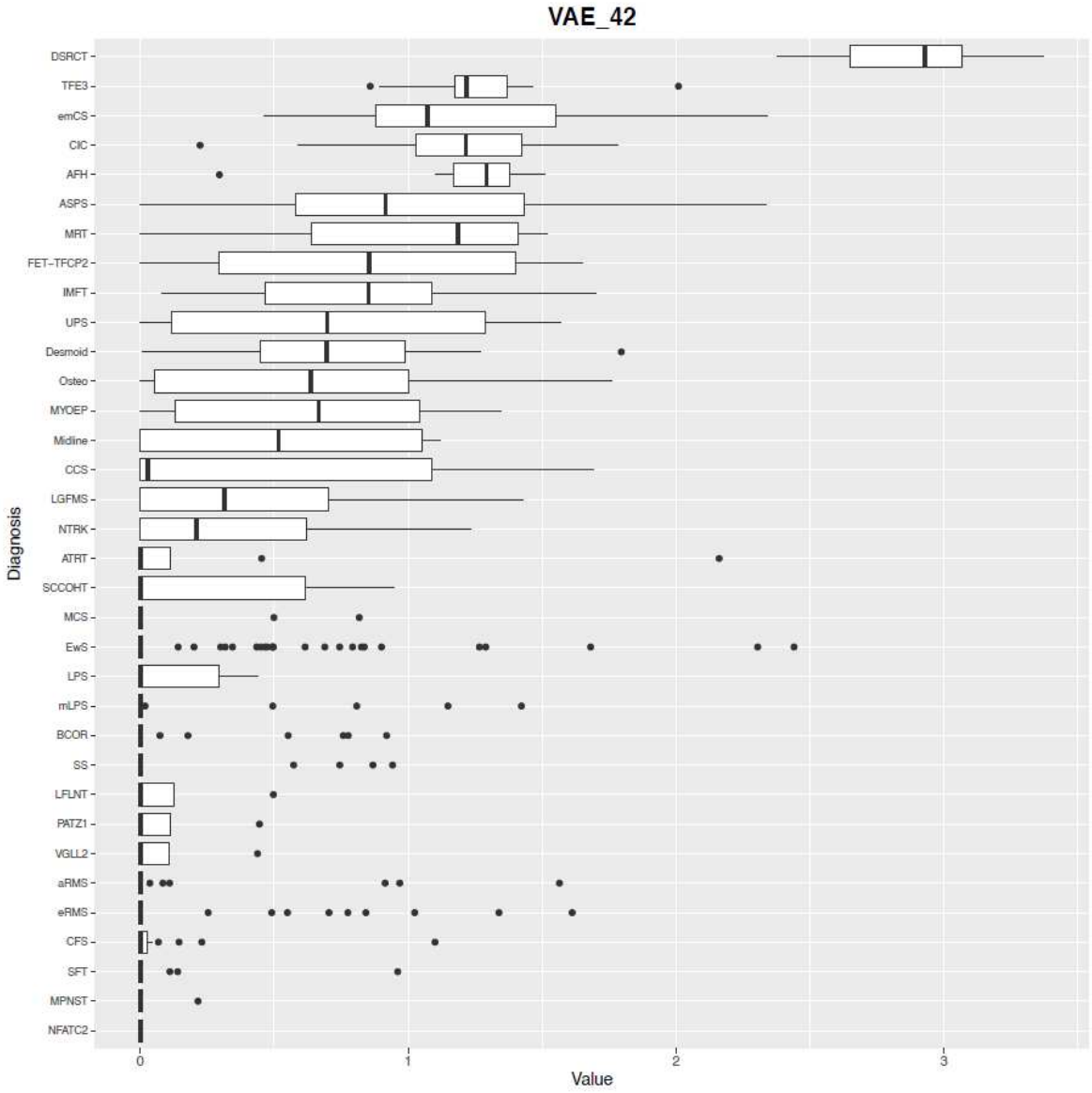

Figure 25: VAE feature 42 (VAE_42) values in all samples grouped by diagnosis.

Its top 20 genes are: *PHLDA2, GJB2, THBS2, CHI3L1, KIF26B, F2RL1, BAI1, FOXC2, MEIS3, AHNAK2, SHOX2, GAL, MT1E, MT1F, AP000688.8, ISM1, COL8A1, SCG5, LAMP5, ALOX5*. Three of these genes at least (*GJB2, CHI3L1, GAL*) are known to be overexpressed in DSRCT.

As a final representative example, consider VAE number 26 which shows its highest mean value in midline carcinoma (Figure 23), as shown in Figure 26.

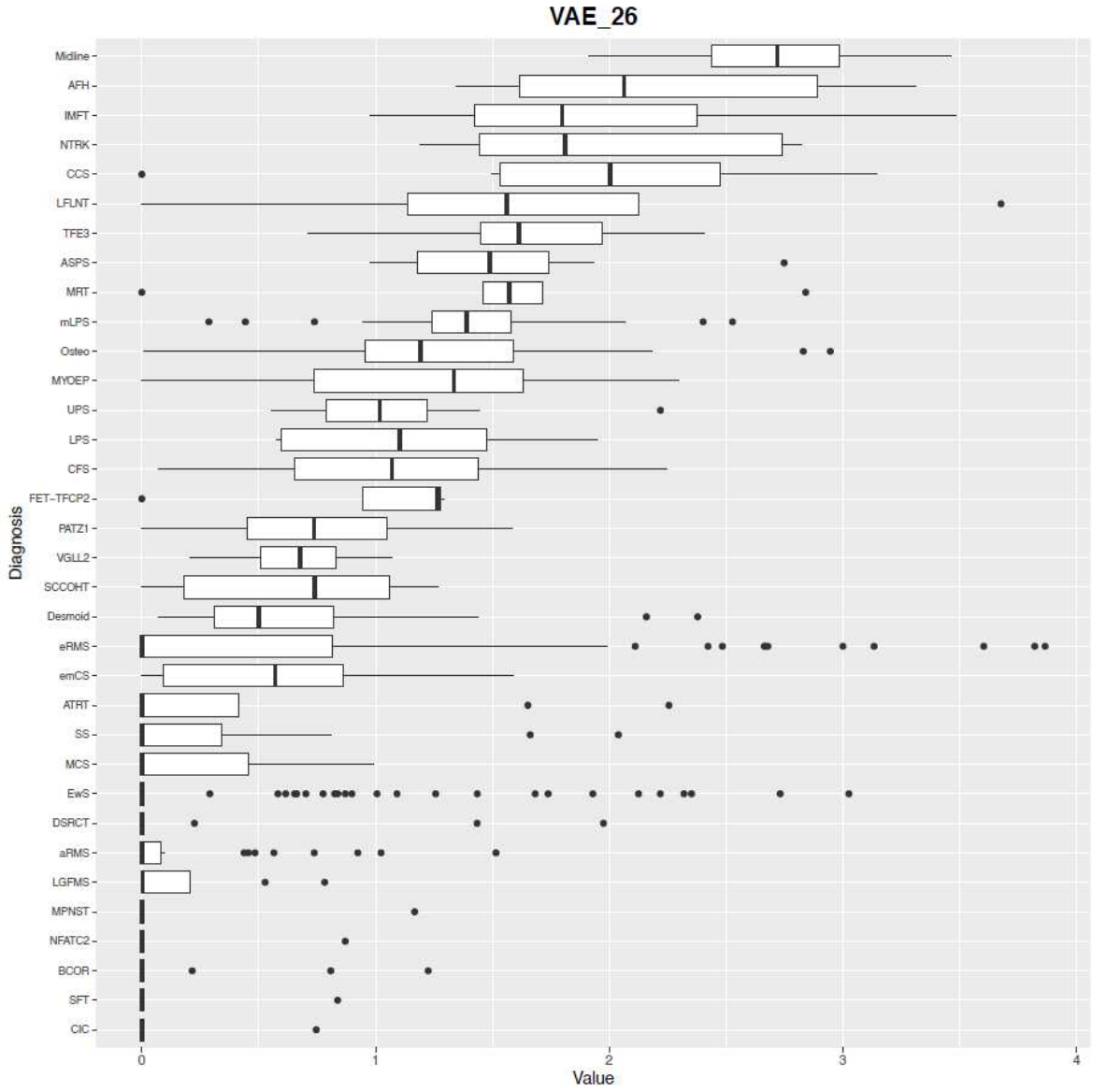

Figure 26: VAE feature 26 (VAE_26) values in all samples grouped by diagnosis.

Its top 20 genes are: *IGKV3-20, IGKV1-5, IGLV3-21, IGKV1-27, IGHV1-69, C15orf48, IGHA1, S100A9, IGHV3-21, LAMC2, IGLC2, IGHV1-18, IGHV4-34, IGLC3, IGKV3-15, IGKV3D-20, MGST1, S100A8, IGKV3-11, IGHM*. These genes are almost all coding for immunoglobulin heavy or light chains, and this VAE feature is probably capturing a measure of immune infiltration, especially from B cells. It is moreover interesting to see a good correlation between the distribution of this feature and the results from MCP-counter (Figure 4) and MiXCR (Figure 9), as for instance the top 5 diagnoses for this feature are also the most infiltrated with B cells according to these other tools: midline carcinoma, angiomatoid fibrous histiocytoma, inflammatory myofibroblastic tumor, NTRK-fused sarcoma and clear cell sarcoma.

Overall, without detailing all 50 VAE features, we can observe that some features correspond to sets of genes with biologically meaningful interpretation. However, the interpretation is limited by the fact that the relationship between genes in a feature is non-linear, thus the weights of a neural network are not easily interpretable. Nonetheless, there is still value in highlighting important genes, especially for genes that are not known to be involved in disease, since some of these high-weight genes may have biological

significance for the diagnoses in which the feature is enriched. In total, the VAE is relatively interpretable since one sample is represented by "only" 50 features, each of which can be assigned specific biological significance through its high-weight genes. Also, the VAE potentially allows novel biological insights by extracting biologically meaningful features and highlighting specific genes in these.

Extension of diagnostic tool to cancers of unknown primary

As discussed previously, the VAE can potentially be used to improve diagnosis prediction, I therefore implemented this method as an additional tool to the standard diagnosis of patients profiled in RNA-seq at Institut Curie. However, clinical practice soon brought forth a challenge for the initial tool and motivated the finding of new ideas to improve it. Indeed, a diagnostic prediction tool of this kind is based on the "learning" of previously known diagnoses, however the method cannot predict an entity which it has never "seen". Thus, the prediction can only be accurate if the correct diagnosis is already present in the training dataset. But in clinical practice, the patient that is addressed for RNA-seq profiling in Institut Curie is by definition a case for which there is diagnostic doubt. Though we have high clinical suspicion of sarcoma, it is not rare to discover that the tumor is in fact another type of cancer: carcinoma, lymphoma or neuroblastoma for instance (this is why there are a significant number of non-sarcomas in the UGS database, cf Figure 2). For the tool initially implemented for the UGS, this issue was addressed in the first place by including in the training dataset not only the sarcomas, but the whole set of diagnoses present in the database.

Nonetheless, the diagnoses in the UGS still exhibit a bias towards particular pathologies: sarcomas, small round cell tumors and pediatric tumors in general. There are virtually no cases of adult carcinomas for instance, which are much more frequent cancers in the general population. Soon after implementation of my first tool for the clinical diagnostic workflow of the UGS, one clinical case of a patient illustrated this issue and prompted the development of the next phase of the project: a 30-year old male patient was addressed to Institut Curie (Dr Sarah Watson) for suspicion of bone sarcoma after having been evaluated in multiple other clinical centers in Paris. He presented with rapid degradation of performance status, an abdominal mass of unknown origin as well as diffuse bone and sub-diaphragmatic lymph node metastases. Pathological review of the tumor biopsy could not determine the origin of his cancer. Based on the aggressive nature of the cancer and the abdominal mass, chemotherapy for pancreatic cancer was proposed by the tumor board in Institut Curie. Before the beginning of chemotherapy, Dr Watson had the idea to address this patient to the UGS for RNA-seq profiling. However, in light of the clinical history and considering that the sample ended up in the "middle" of the UMAP as though it did not belong to any of the learned entities, we suspected that the correct diagnosis was not present in the UGS database used for training.

It was a fortunate coincidence that I had just finished at the time a months-long work of reprocessing raw RNA-seq data of diverse public projects including The Cancer Genome Atlas (TCGA, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga), Genotype-Tissue Expression (GTEx, https://gtexportal.org/home/), Human Protein Atlas (HPA)[62] using the same bioinformatics pipeline, an effort designed to answer a different question, as discussed in the following section on Ewing sarcoma neotranscripts (see **Identification of novel transcripts in sarcoma**). I had thus at hand RNA-seq training data not only for UGS diagnoses but also for all other major cancer types present in the TCGA, as well as normal tissues from TCGA, GTEx and HPA, all of them having been processed with the same bioinformatics pipeline (> 20,000 samples in total). I therefore took advantage of all this processed data to design a new "classifier" based on a VAE: for training I used all cancer samples from the TCGA, and also included all normal tissues from TCGA, GTEx and HPA, to let the VAE "learn" normal tissue profiles and account for potential normal cell contamination of tumor samples. Since the number of diagnoses was close

to 100, I chose this number of encodings for the training of the VAE. After visualization of the encoded space by UMAP and projection of the unknown sample on the plot, the diagnosis was clear: this patient clustered perfectly within the kidney cancers.

The case was then reviewed by a kidney cancer specialist, who confirmed that this diagnosis was indeed very likely, considering that the abdominal mass was possibly developed from an embryological remnant of developing kidney tissue, thus explaining the absence of tumoral mass in the kidneys: this situation is rare but is known to occur in some patients[63]. The patient was thus treated as for kidney cancer: he was included in a clinical trial evaluating an anti-PD1 immune checkpoint inhibitor in combination with an anti-angiogenic tyrosine kinase inhibitor. First evaluation at three months showed a complete response, and according to the latest report, the patient is still progression-free. Altogether, this story showcased the potential of RNA-seq to help with the diagnosis of difficult cases and orient treatment with success.

This prompted us with Dr Watson to launch a clinical study of the potential utility of RNA-seq to determine the tumor of origin of cancers of unknown primary (CUP), a well-defined clinical entity that comprises up to 5% of metastatic cancers[64].

As this project was the object of a peer-reviewed publication[65], I reproduce it in Annex 1 without developing this theme further in the manuscript.

Identification of novel transcripts in sarcoma

Introduction

As discussed previously, the immune landscape of pediatric sarcomas is relatively poor compared to other more immunogenic tumors such as melanoma or non-small cell lung cancer[27]. This is likely due to many factors including low tumor mutational burden leading to fewer neoantigens[14,15]. However, as shown previously, some pediatric sarcomas do have higher immune infiltration, and sometimes present isolated cases of response to immunotherapy, even in genomically "calm" tumors[66]. Therefore, it may not be excluded that some tumor antigens could be recognized in these tumors by the immune system, independently of the generation of neoantigens by classical processes such as DNA mutations[13].

This part of my thesis is precisely about the characterization of a potential novel source of neoantigens for some pediatric sarcomas. It is fair to say that I had absolutely not expected this to become the main part of my thesis when I began my PhD, but serendipity is often generous in science, and I welcomed it gratefully.

Identification of novel transcripts in Ewing sarcoma

When I began my PhD in November 2018, a PhD student (Olivier Saulnier) and a Master's student (Jérômine Vigneau) from our team were just beginning to unravel an interesting observation they had made in a pioneering experiment. They had used long-read RNA sequencing (PacBio technology), a technique that enables sequencing of reads of up to 10 kilobases (compared to reads of about 75-150 nucleotides in short-read RNA-seq), to characterize the long-read transcriptome of a Ewing sarcoma cell line (A673). Interestingly, apart from reads that aligned to known human transcripts annotated in the RefSeq reference transcriptome, they found that 145 high-quality sequences did not align to any known human gene, but rather to intergenic regions. They called these sequences "NA" (non-annotated) sequences. They explored each of these sequences manually by visualizing them in a genome browser (Integrative Genomics Viewer, IGV, https://software.broadinstitute.org/software/igv/). They discarded 15 of these as having low-read support, and flagged 50 of them as errors of automatic annotation (i.e. sequences that did in fact correspond to known human genes). However, there remained 80 sequences corresponding to potentially "novel" transcripts originating from completely unannotated intergenic regions of the human genome. They hypothesized that some of these novel transcripts might be related to the oncogenic chimeric transcription factor of Ewing sarcoma: EWS-FLI1[9]. Using short-read RNA-seq in the same cell line (A673) with controlled knockdown of EWS-FLI1 by a doxycycline-inducible short hairpin RNA (shRNA) system, they found that 42 of these transcripts were downregulated following the knockdown of EWS-FLI1. It is known that EWS-FLI1 binds the genome at sites called GGAA microsatellites[67,68]: they found that 20 of these 42 transcripts had a GGAA microsatellite just upstream of their transcription start site (TSS). The binding of EWS-FLI1 at these sites, as well as presence of chromatin activation marks (H3K27ac and H3K4me3), were confirmed for 18 of these novel transcripts by ChIP-seq (Chromatin ImmunoPrecipitation sequencing) in the same cell line. Finally, they concentrated on a set of 4 novel transcripts that were also found to be expressed in human tumor RNA-seq data. This filtering procedure is summarized in Figure 27.

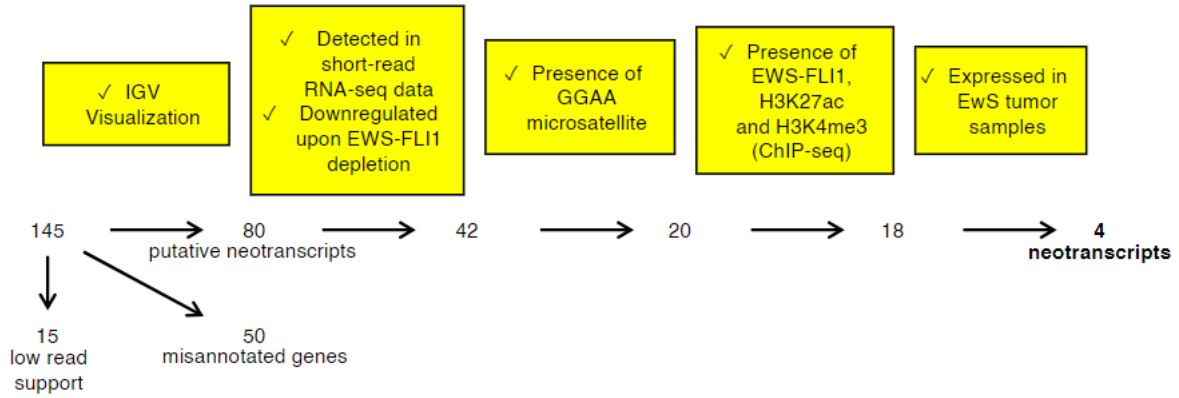

Figure 27: Filtering procedure from 145 PacBio NA sequences to 4 novel transcripts.

A characteristic genomic profile of one of these four transcripts is reproduced in Figure 28: this transcript is composed of four exons in the forward orientation, there is a GGAA microsatellite near its transcription start site (TSS), where there is evidence in ChIP-seq of binding of EWS-FLI1, as well as presence of H3K27ac and H3K4me3 chromatin activation marks; these ChIP-seq signals disappear in the doxycycline-induced EWS-FLI1 depletion state (+DOX). As for RNA-seq, the transcript is expressed both in the A673 cell line and human tumors, and its expression disappears in the cell line after EWS-FLI1 knockdown (+DOX).

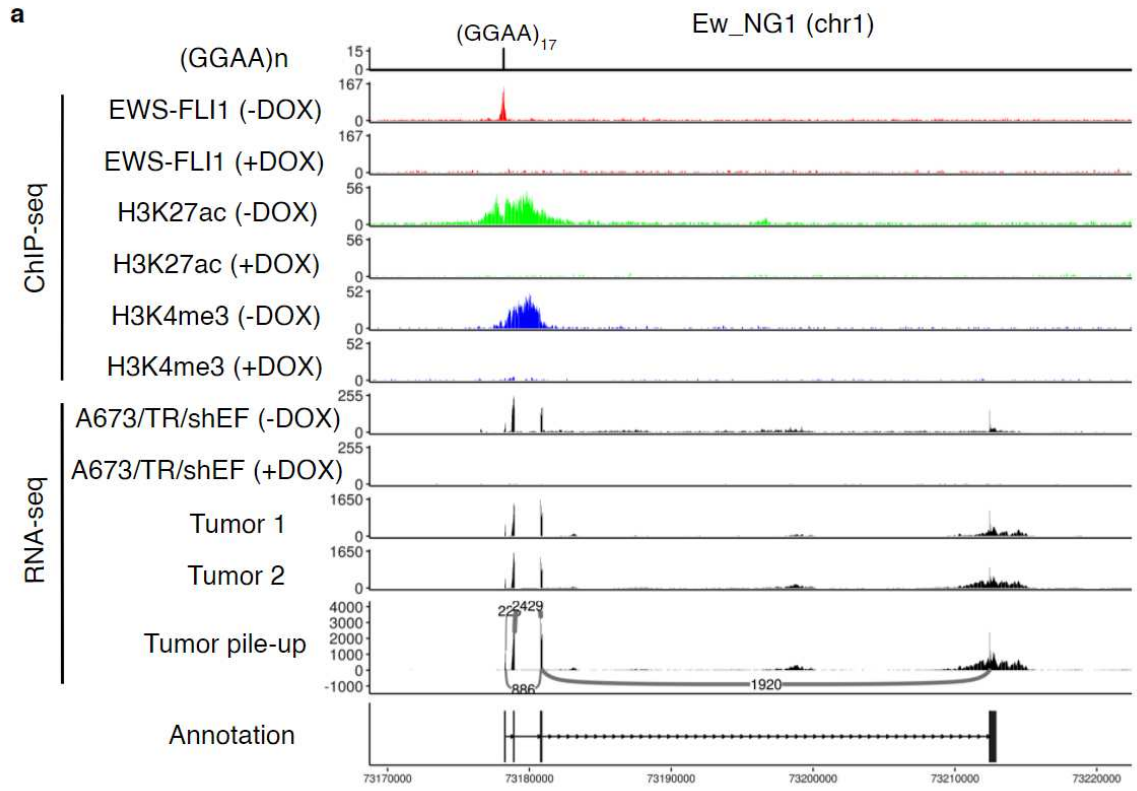

Figure 28: Genomic view of a novel transcript in Ewing sarcoma (Ew_NG1).

In addition to these characteristics, further experimental evidence showed the dependence of these novel transcripts on EWS-FLI1: quantitative reverse transcription PCR (RT-qPCR) showed downregulation of their

expression in the A673 cell line after knockdown of EWS-FLI1 (+DOX), while on the contrary their expression was induced in mesenchymal stem cells (the putative cells-of-origin of EwS) transfected with EWS-FLI1, which showed that EWS-FLI1 was both necessary and sufficient for their expression.

At this point of the project, I had just begun my PhD, and one important question was still remaining: was the expression of these novel transcripts really EwS-specific, as could be hypothesized based on this mechanical understanding of their expression? To answer this question and confirm the hypothesis that these novel transcripts are direct products of the EwS-specific oncogenic chimeric transcription factor EWS-FLI1, I took advantage of large public gene expression datasets available for analysis: specifically, I searched for the presence of these novel transcripts in more than 20,000 RNA-seq samples, including all primary tumors (n=10,201) and juxta-tumor normal tissues (n=746) in TCGA, as well as all samples in GTEx (n=9,659) and the Human Protein Atlas (n=200) that comprise the vast majority of normal tissues. I also searched for their presence in all annotated samples of the UGS dataset.

Method: FASTQ files from TCGA, GTEx and HPA (http://www.proteinatlas.org) were downloaded from their respective platforms and aligned to the hg19 genome assembly using STAR (v2.7.0e). The GTF file used for alignment and quantification of gene expression was based on evidence-based annotation of the human genome (GRCh37), version 19 (Ensembl 74) provided by GENCODE, to which was added the annotation of the four novel transcripts in GTF format. Gene expression was quantified using the GeneCounts procedure from STAR. Raw counts were then normalized to Transcripts Per Million (TPM). This analysis was performed on the Curie HPC cluster using NextFlow, and raw FASTQ files were deleted after analysis to avoid storing large data files on the cluster.

The results confirmed what we expected: expression of these four novel transcripts is extremely specific to EwS, one example of a transcript is in Figure 29.

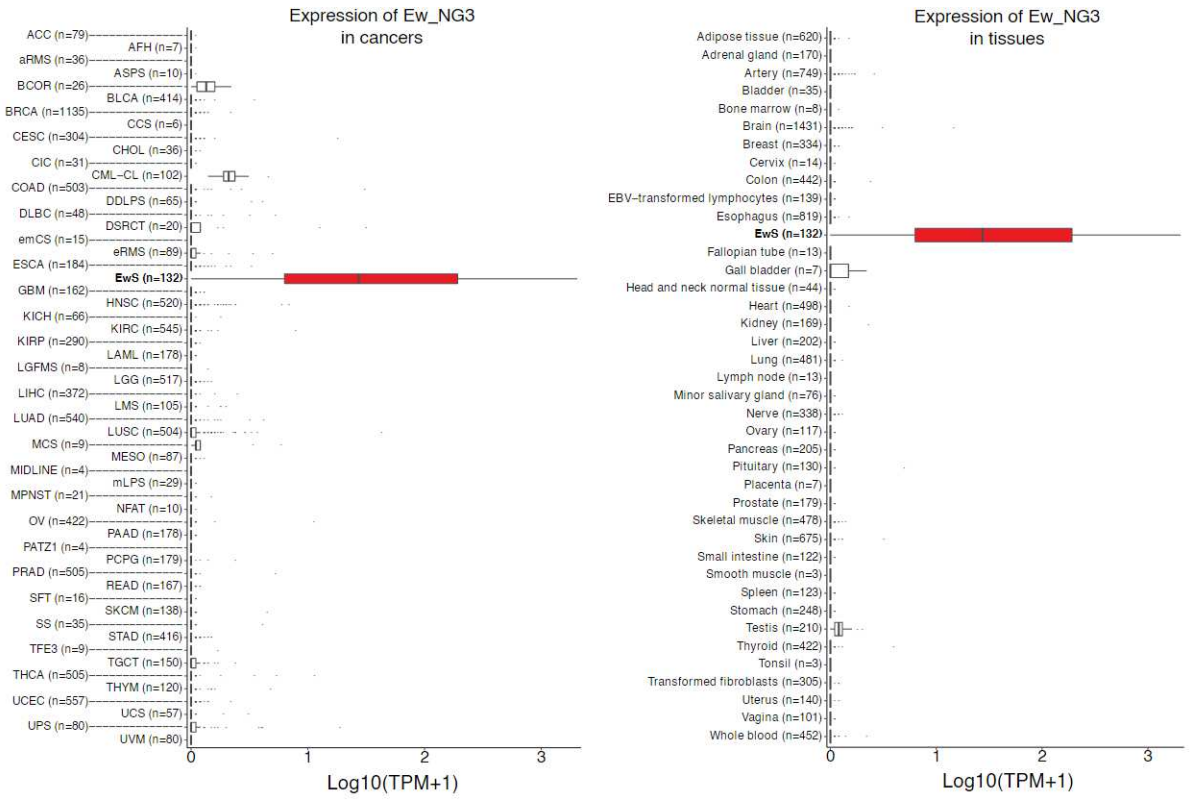

Figure 29: Expression level of a EwS-specific novel transcript (Ew_NG3) across cancer types and normal tissues. Abbreviations are as in TCGA and in the submitted manuscript.

There are some non-EwS samples that show non-zero expression levels of these novel transcripts, however these are very low levels compared to EwS and should probably be assigned to "background noise" expression. For this specific transcript though (Ew_NG3), there is low but consistent expression in a chronic myelogenous leukemia-derived cell line (CML-CL), BCOR-rearranged sarcoma (BCOR) and testis. This search for the expression of these novel transcripts across a large panel of other cancers and normal tissues was not only important to confirm the hypothesis that they are induced by EWS-FLI1, but also because we had already considered their potentially promising translational relevance as tumor-specific markers, for diagnostic and even therapeutic purposes. Indeed, one holy grail of cancer treatment research is to find molecular targets that are tumor-specific and absent in all non-tumor cells. This is especially true of therapies harnessing the immune system ("immunotherapies") that rely on the ability of the adaptive immune system to recognize an antigen on the surface of a tumor cell[11]. The ideal tumor target for immunotherapy is therefore a tumor-specific antigen not expressed in normal cells to avoid toxicity in healthy tissues. This is the reason why a large part of immunotherapy target research has concentrated on tumor-specific neoantigens, i.e. antigens that are novel and not expressed in a normal context. Up to now, the main source of neoantigens that has been explored is their generation by DNA mutations in protein-coding genes[13,69]. Due to the mainly "private" nature of most mutations (i.e. they are not shared across patients), these neoantigens are mostly patient-specific, and it is rare to characterize a neoantigen shared across multiple patients, e.g. through a common mutation[70]. These so-called "public" neoantigens[71] would however be more attractive for clinical practice, in the sense that a therapy targeting them could be delivered "off-the-shelf" for many patients, without all the cumbersome process of identifying neoantigens and designing specific therapies for each patient.

The identification of tumor-specific novel transcripts that are directly induced by the oncogenic driver event in EwS is therefore very interesting in this regard: these novel transcripts are only found in EwS (which is corroborated by the mechanistic insights as well as the demonstration of their absence in non-EwS samples) and thus could potentially be a source of tumor-specific neoantigens, that are importantly "public" since they are shared across patients of the same disease. To come back to the observation that some novel transcripts seem to be expressed at low levels in other cancer types and testis, this is currently unexplained but it would not compromise a therapeutic strategy targeting them, since activity against other cancer types is not a problem in practice, while testis is an immune sanctuary and is therefore not subject to immunotherapy toxicity.

However, there is a significant gap to be filled before demonstrating that these novel transcripts can generate neoantigens and be of relevance for immunotherapy. Antigens are small peptides that must be presented by the MHC (Major Histocompatibility Complex) to be recognized by immune cells. In any case, an RNA transcript is not by itself an immune target and must be translated to be recognized as such. This question will be explored subsequently in the manuscript.

To come back to these initial results, we identified four novel transcripts originating from otherwise silent genomic regions in EwS, that are directly induced by the oncogenic chimeric transcription factor EWS-FLI1, and that are not found in non-EwS cancers and normal tissues. We decided to name these novel transcripts "neotranscripts", and to name their corresponding genomic loci as "neogenes" (NG). These four initial neogenes were thus named *Ew_NG1-4*.

Neotranscript discovery using short-read RNA-seq

Ew_NG1-4 were discovered using long-read RNA sequencing in a EwS cell line; however this technique is costly and sequencing depth is not as high as in short-read RNA-seq. We hypothesized that more

neotranscripts could be present in EwS but had not been detected in the initial experiment. Instead of using long-read RNA sequencing, which is moreover technically difficult to perform on clinical samples, we wondered if we could search for neotranscripts in short-read RNA-seq data, of which we have a large number of samples including from tumors of patients.

To this effect I started exploring ways of finding novel transcripts from short-read RNA-seq data. Initially I focused on identifying specific k-mers in RNA-seq samples, notably using a method named DE-kupl[72]. This method is designed to find differentially expressed k-mers in a set of samples as opposed to a set of reference samples, which it then uses to reconstruct approximately a differentially expressed transcript sequence with help from alignment to the reference genome. To test this method, I tried to "re-discover" *Ew_NG1-4* with its help. When I used this method on a set of RNA-seq samples from the A673 cell line as compared to the same cell line with EWS-FLI1 knockdown, the results were both promising and disappointing: for all neotranscripts the method could definitely pinpoint a differentially expressed transcript in the correct genomic region, however the reconstructed transcript sequence was very imprecise and did not allow correct annotation of neotranscripts, unless manual annotation was performed with the help of visualization of RNA-seq alignments in a genome browser.

I therefore turned to more straightforward and precise methods of transcript reconstruction. There are many computational methods designed to assemble the exact sequence of transcripts from RNA-seq reads, either completely *de novo*[73] or using help from alignment to a reference genome ("genome-guided transcript assembly"[74]). The former approach is logically more prone to making errors and is generally used when no reference genome is available; in the case of human RNA-seq the best-performing methods are the genome-guided ones. Multiple methods have been developed for genome-guided transcript assembly, some of the most popular being Cufflinks[75] and StringTie[76]. I chose to use a more recent tool named Scallop[77], a genome-guided transcript assembly method based on phase-preserving graph decomposition. During benchmarking, it notably reconstructed 34.5% more transcripts than StringTie, especially lowly expressed ones (67.5% more than StringTie). It also has the advantage of not requiring excessive computational resources and memory. I therefore ran Scallop on an RNA-seq sample of a human EwS tumor to assemble all expressed transcripts, independently of their being or not part of the reference transcriptome. I then only kept the assembled transcripts without any overlap with a known annotation in GENCODE.

Method: Paired-end FASTQ files were first aligned to the hg19 human reference genome using STAR (v2.7.0e). Then Scallop (v0.10.4) was used on the resulting BAM file with default parameters to assemble all expressed transcript sequences. To conserve only unannotated transcripts, I used Gffcompare[78] to compare the Scallop output GTF file with the reference GENCODE v19 GTF file, and conserved only transcripts labeled by Gffcompare as « u » (unknown, intergenic), « y » (contains a reference within its introns) and « x » (exonic overlap on the opposite strand). Finally, to remove lowly expressed transcripts and decrease the rate of false positives, I removed all transcripts with coverage less than 10 as output by Scallop.

In total, Scallop assembled 70,061 transcripts from the EwS tumor RNA-seq sample, of which I conserved 1,069 which did not overlap any known annotation and were of sufficient coverage. My objective was to identify recurrent and EwS-specific novel transcripts: I therefore designed the following strategy to rapidly identify the best candidates among these 1,069 potentially novel transcripts, considering that most of the false positives or "private" (not shared across patients) novel transcripts would be filtered out by this procedure.

Method: I applied a first filter based on high and tumor-specific expression as compared to a limited set of other tumors: for this I quantified the expression of these 1,069 transcripts in samples of eight different sarcoma types including EwS (three samples of each tumor type) by re-aligning each sample with STAR and quantifying expression using the GeneCounts procedure with the GENCODE v19 reference GTF file to which

were added the sequences of the 1,069 candidate neotranscripts. Raw counts were converted to transcripts per million (TPM) before the filtering process. To retain only tumor-specific and highly expressed candidates, I selected transcripts with 1) mean expression in EwS of more than 10 TPM; 2) log-fold change of mean expression in samples of other diagnoses versus mean expression in EwS of less than -2; 3) mean expression in samples of other diagnoses of less than 3 TPM; 4) maximum expression in samples of other diagnoses of less than 15 TPM.

This resulted in the selection of 114 candidate neotranscripts. To conserve only highly, recurrently expressed and above all tumor-specific neotranscripts, I then undertook a similar approach as for the initial four neotranscripts, i.e. searching for their expression in a large panel of other cancer types and normal tissues. However, as the initial search for the four neotranscripts, which was based on the totality of samples of TCGA, GTEx, HPA (n>20,000), was very time-consuming to repeat, I instead chose to limit the search to 50 samples of each cancer and normal tissue type.

Method: I quantified expression of these 114 candidate neotranscripts in all tumor samples from the UGS, all cancer types in TCGA (either all samples from one type, or only 50 samples if number of samples exceeded 50), all normal tissue samples in TCGA, all normal tissue types in GTEx (either all samples from one type, or 50 samples if number of samples exceeded 50) and all normal tissue samples from the Human Protein Atlas. Every sample was re-aligned with STAR and expression quantified by the GeneCounts procedure with the use of a GTF file including GENCODE v19 and the candidate neotranscripts. Raw counts were converted to TPM before filtering. To retain tumor-specific candidates with a relatively high expression level in EwS (to account for potentially lower tumor content in some samples, I lowered the first threshold as compared to the first filter) and near-zero expression in other cancers and normal tissues, I selected transcripts with 1) mean expression in EwS of more than 7.5 TPM; 2) log-fold change of mean expression in other samples versus mean expression in EwS of less than -3; 3) mean expression in other samples of less than 2 TPM; 4) 99 % quantile of expression in other samples of less than 10 TPM; 5) maximum mean expression in another cancer or normal tissue of less than 10 TPM (excluding testis and placenta).

I noted during this procedure that some neotranscripts could be moderately expressed (most less than 10 TPM) in germinal tissues (testis and placenta), reflecting known higher transcriptomic diversity and exclusivity there (e.g., for cancer-testis antigens[79]), and therefore allowed the few genes (less than 1.5% of neotranscripts in this study) expressed in these tissues at more than 10 TPM to pass filter 5) nonetheless.

This pipeline (summarized schematically in Figure 30) resulted in a total of 61 EwS-specific neotranscripts, corresponding to 25 different genomic loci or EwS-specific neogenes = “Ew_NGs” (one “neogene” can have multiple alternative “neotranscripts” due to alternative splicing).

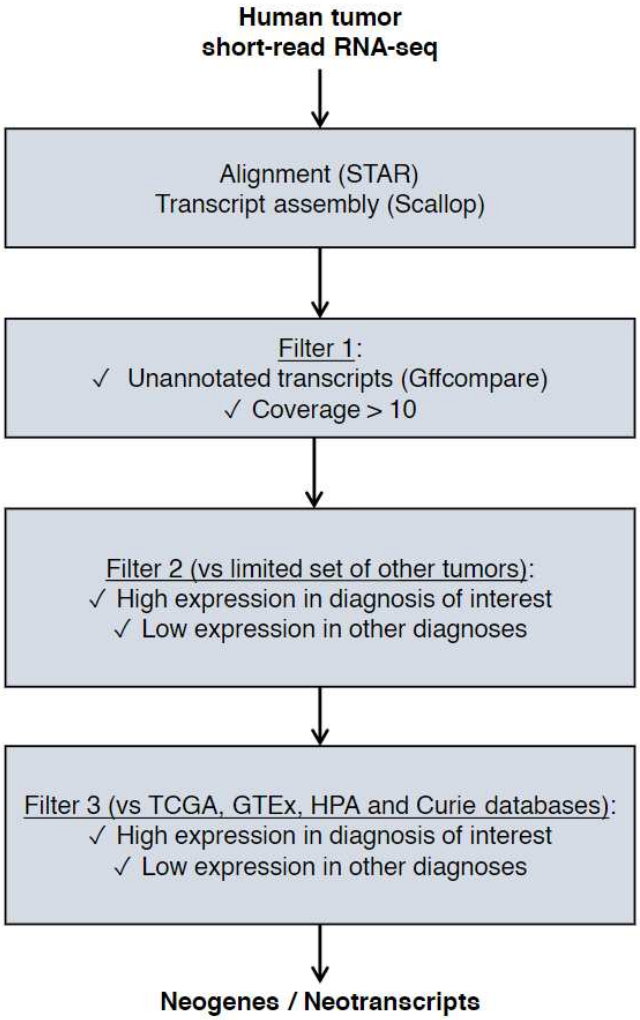

Figure 30: Schematic diagram of the neotranscript discovery pipeline from short-read RNA-seq data.

Interestingly, three out of the four initial neogenes were retrieved within these 25 Ew_NGs, moreover Scallop allowed some splice variants of these three neogenes to be identified. It was intriguing though that one previously characterized neogene (*Ew_NG3*) had not been identified by Scallop, since it seemed to be very highly and recurrently expressed in EwS (Figure 29). To understand why, I searched manually for *Ew_NG3* at the different steps of the pipeline: it turned out that a corresponding transcript had accurately been assembled by Scallop but had been filtered out at the initial step of discarding transcripts overlapping any known annotation. Indeed, the reason for this discrepancy between its identification from the PacBio experiment as opposed to Scallop rapidly became clear: this transcript had been filtered out in my pipeline because it corresponded to an annotated transcript in GENCODE (AC073135.3), however Olivier Saulnier had used the RefSeq reference transcriptome to filter out known transcripts from the PacBio experiment. This transcript was only annotated in GENCODE but not in RefSeq, which contains fewer transcripts, notably fewer non-coding RNAs. Nonetheless, it appeared that *Ew_NG3*, which is very specific to EwS and induced by EWS-FLI1, was already "known" and not novel. I therefore searched for the origin of this annotation and found the explanation: this transcript annotation in GENCODE had been derived from a series of five expression sequence tags (ESTs) exclusively from EwS cell lines[80], and no relevant information was available for this transcript in any database. Considering this and although this neotranscript was already "known", we conserved this neogene in our list and thus had in the end a set of 26 Ew_NGs corresponding to 62 EwS-

specific neotranscripts. Their expression profile across cancer types and normal tissues is represented in Figure 31.

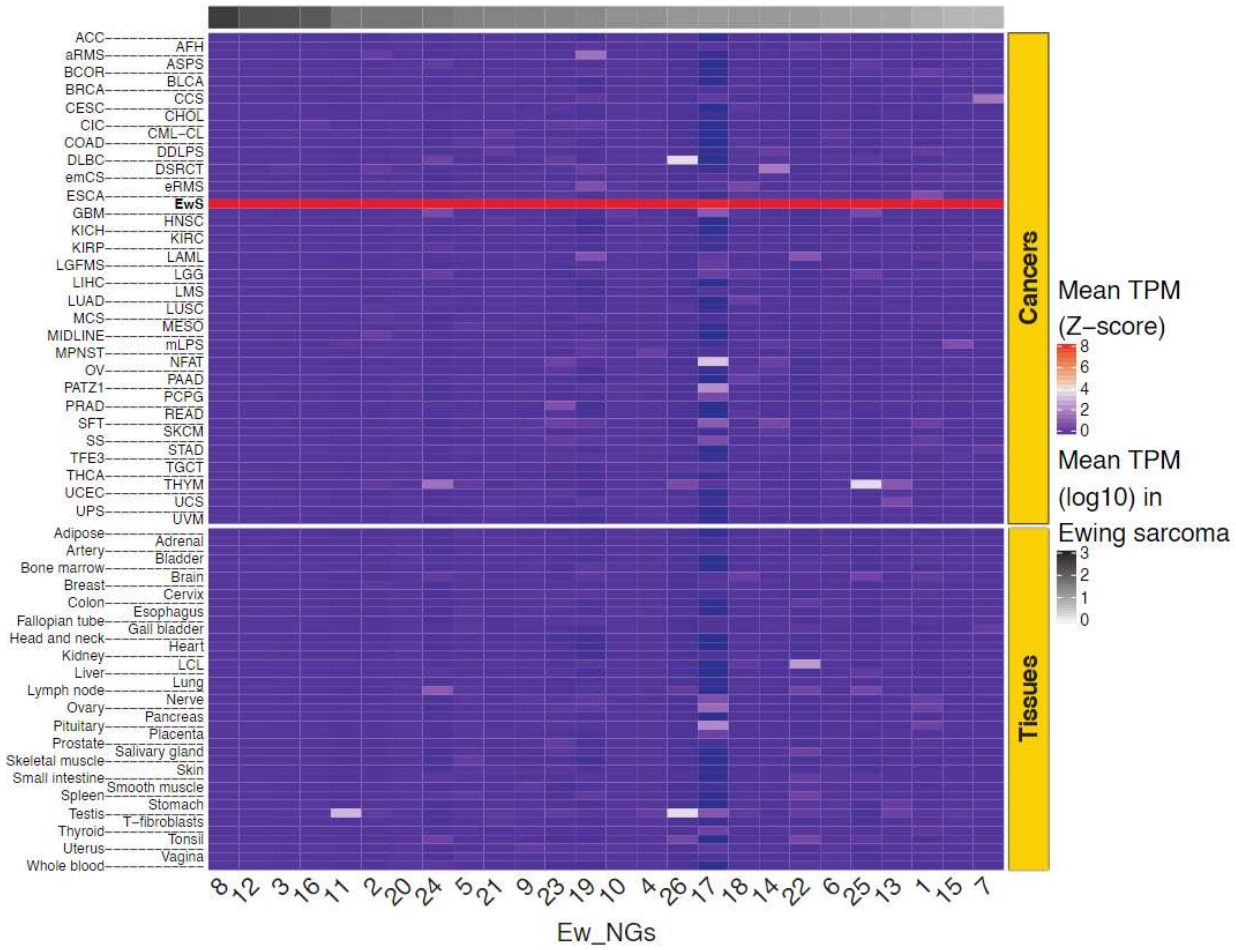

Figure 31: Expression levels across cancers and normal tissues of Ew_NGs. Abbreviations are as in TCGA and in the submitted manuscript.

Considering this set of 26 Ew_NGs, an important question was the same as for the four initial ones: could they be directly induced and regulated by the oncogenic chimeric transcription factor EWS-FLI1? To answer this, I analyzed RNA-seq data - already available in the lab - of nine EwS cell lines where EWS-FLI1 expression had been downregulated by short hairpin (sh-) or small interfering (si-) RNA. I also searched for expression of Ew_NGs in mesenchymal stem cells transfected with EWS-FLI1. Results are in Figure 32.

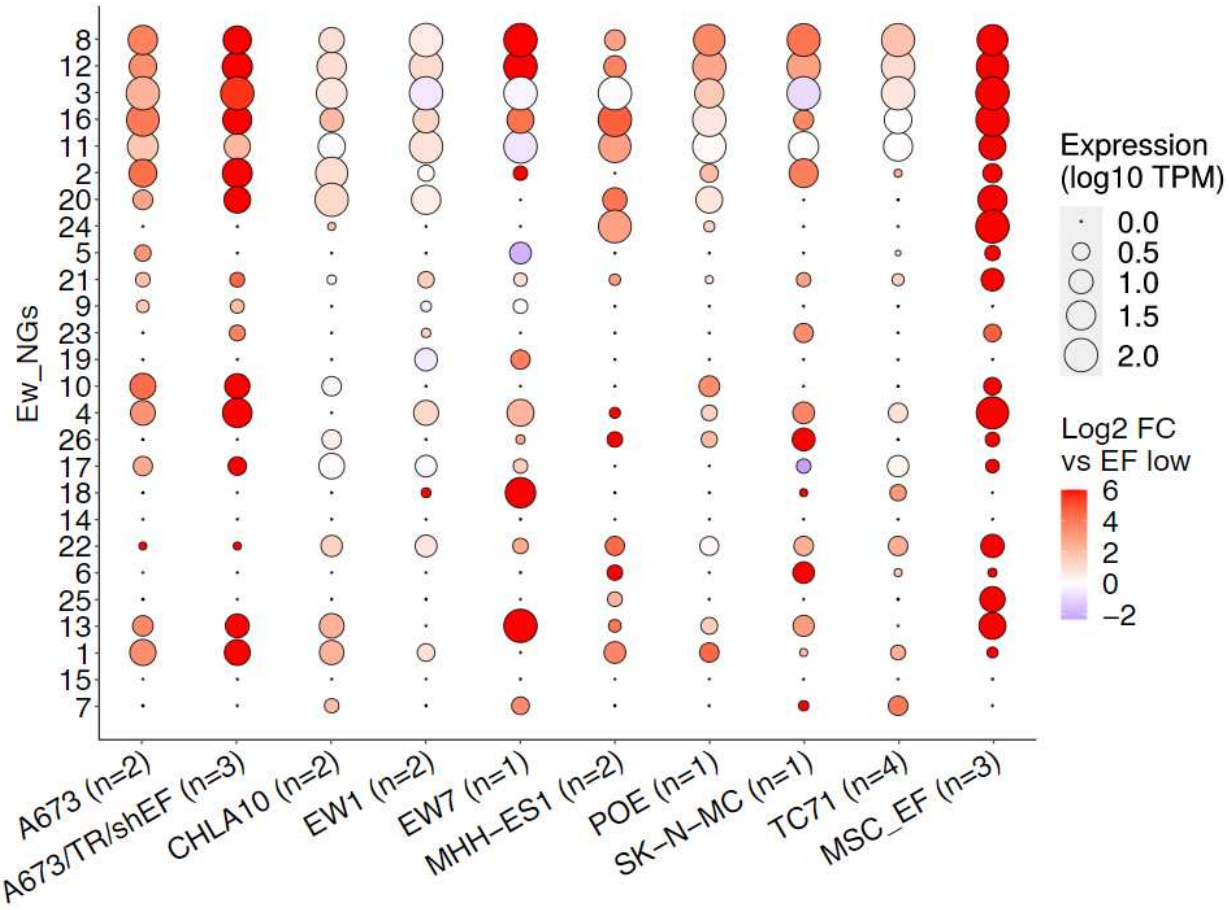

Figure 32: Ew_NG expression in EwS cell lines and MSCs with EWS-FLI1 modulation. Dot size shows mean expression level in EwS cell lines and MSCs transformed by EWS-FLI1 (capped at 100). Color represents log2 fold-change (capped at 6) as compared respectively to EWS-FLI1 knocked-down conditions and parental MSCs (EF low).

This showed that a large part - but not all - of Ew_NGs were downregulated with knockdown of EWS-FLI1 in EwS cell lines and were induced in MSCs expressing EWS-FLI1 (all Ew_NGs had zero expression level in MSCs without EWS-FLI1). As for the four initial Ew_NGs, I examined manually their genomic context in a genome browser with attending functional data including ChIP-seq for EWS-FLI1, H3K27ac and H3K4me3, and found that a large number of them had profiles similar to the initial ones, i.e. binding of EWS-FLI1 on a GGAA microsatellite just near their TSS, and presence of chromatin activation marks H3K27ac and H3K4me3. With the help of a bioinformatician in the team (Maud Gautier), we formally analyzed the distance between TSS and nearest EWS-FLI1-bound GGAA microsatellite for all Ew_NGs and known transcripts in GENCODE, to show the enrichment of EWS-FLI1-bound GGAA microsatellites near the TSS of Ew_NGs (Figure 33).

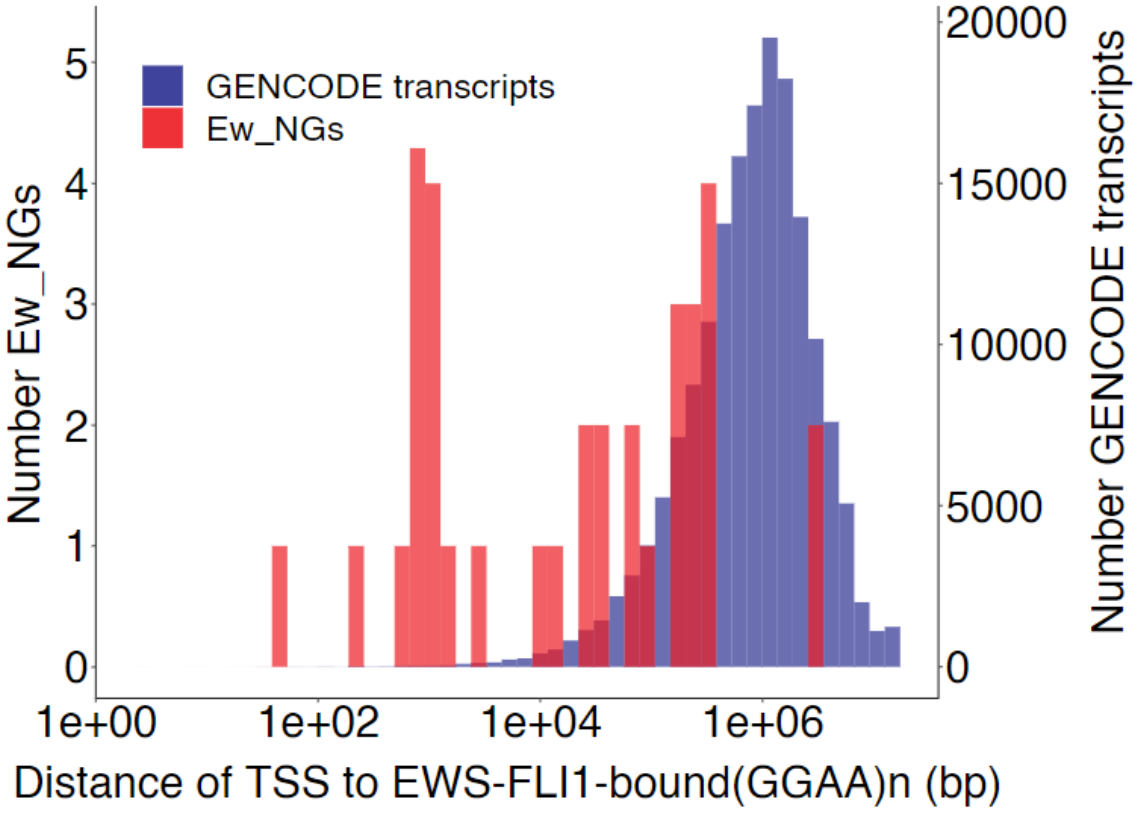

Figure 33: Distance from TSS to nearest EWS-FLI1-bound GGAA microsatellite for Ew_NGs and GENCODE transcripts.

For three Ew_NGs, no EWS-FLI1-bound GGAA microsatellite was found near the TSS: using analysis of H3K27ac HiCHIP (a protein-centric chromatin conformation mapping method[81]) data from the lab by a bioinformatician in the team (Véronique Hill), we could show that they were instead present within so-called enhancer-promoter chains regulated by EWS-FLI1[82]. An example (*Ew_NG17*) is represented in Figure 34.

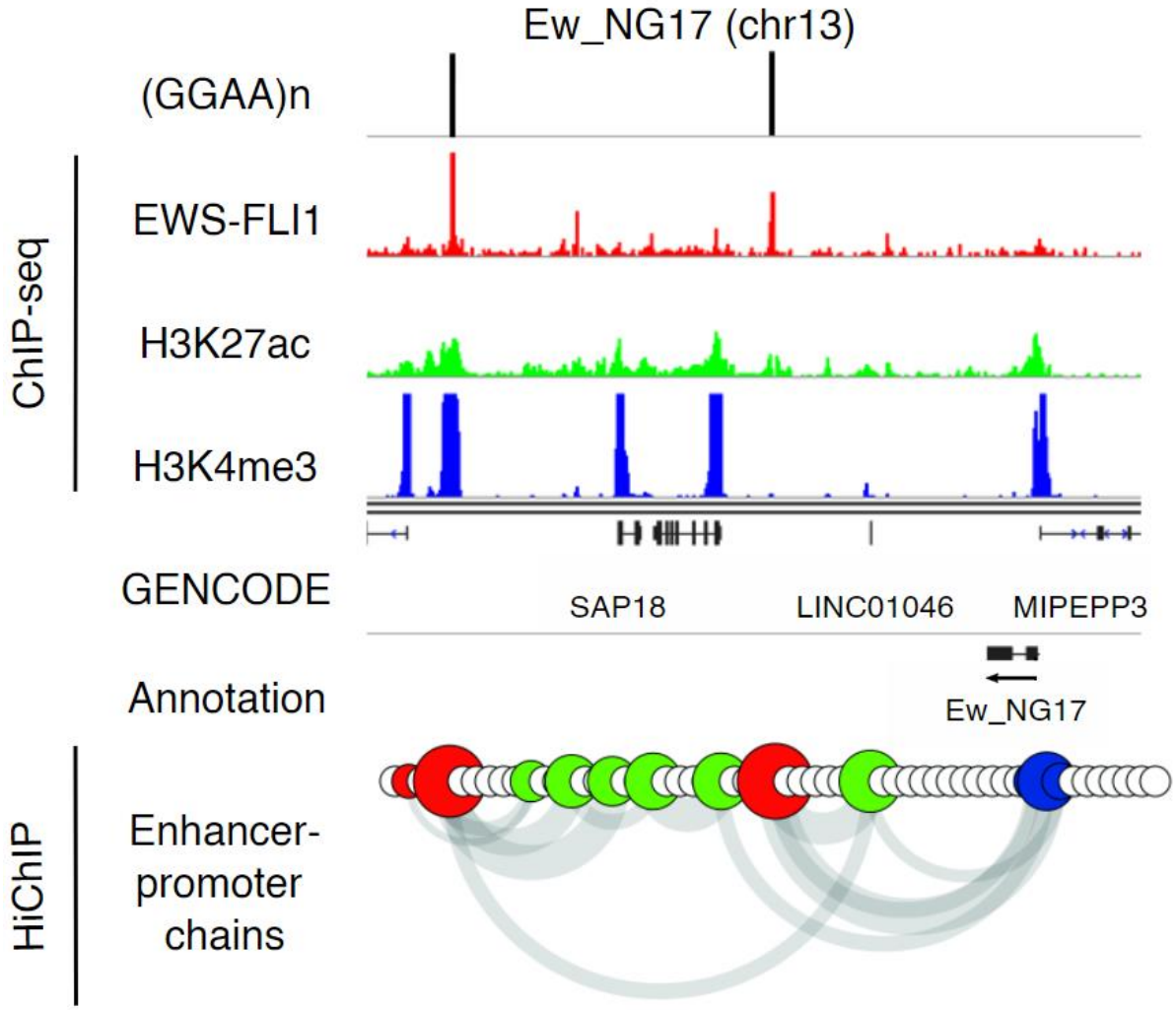

Figure 34: Example of a Ew_NG regulated within an EWS-FLI1 enhancer-promoter chain (Ew_NG17).

These results strongly suggested that most Ew_NGs are directly dependent on EWS-FLI1 binding for their induction. To further demonstrate this fact, we used CRISPR interference (CRISPRi) targeting DNA sequences flanking GGAA microsatellites upstream of six Ew_NGs to prevent EWS-FLI1 from binding to them: this abrogated expression of Ew_NGs (experiments done by Céline Collin, a PhD student in the team). Interestingly, we also used CRISPRi to prevent binding of EWS-FLI1 to the two enhancers upstream of *Ew_NG17*: targeting of both enhancers was synergistic for the downregulation of this NG expression. These results are in Figure 35.

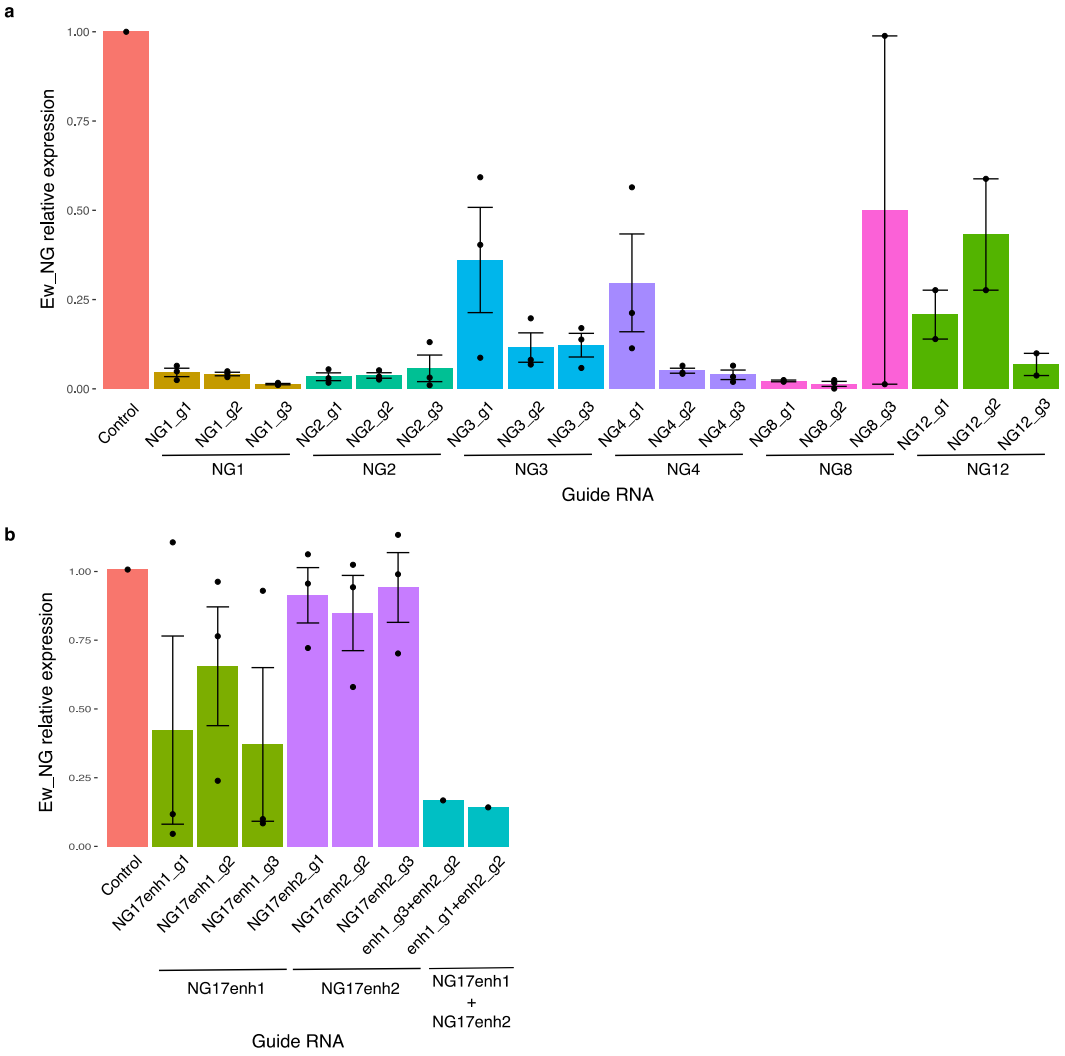

Figure 35: Ew_NGs are downregulated by CRISPRi targeting of sequences close to GGAA microsatellites. Barplots show Ew_NG relative expression measured by qPCR after CRISPRi of upstream GGAA microsatellite for **a**, Ew_NGs with GGAA microsatellite near the promoter (three guide RNAs each), **b**, Ew_NG17 regulated within two enhancer-promoter chains (three guide RNAs for each enhancer GGAA microsatellite targeted individually and targeting of both enhancers enh1 and enh2). Barplots show mean ± s.e.m. of individual replicates (dots).

Finally, to demonstrate that Ew_NGs are specifically expressed by EwS cells and can be detected in single-cell RNA-seq (scRNA-seq) data, I reprocessed some single-cell RNA-seq data of an EwS tumor, along with a non-EwS tumor (desmoplastic small round cell tumor, DSRCT), to quantify expression of Ew_NGs in single cells.

Method: To quantify expression of neotranscripts in scRNA-seq data, a custom transcriptome was produced by appending sequences of the neotranscripts to the reference transcriptome and running Cellranger count on scRNA-seq FASTQ files with this custom index. Counts for neotranscripts were log-normalized and the average log-normalized expression level was plotted with FeaturePlot.

The following UMAP plot shows that Ew_NGs are expressed specifically in EwS cells but not in cells of the microenvironment nor in DSRCT cells (Figure 36).

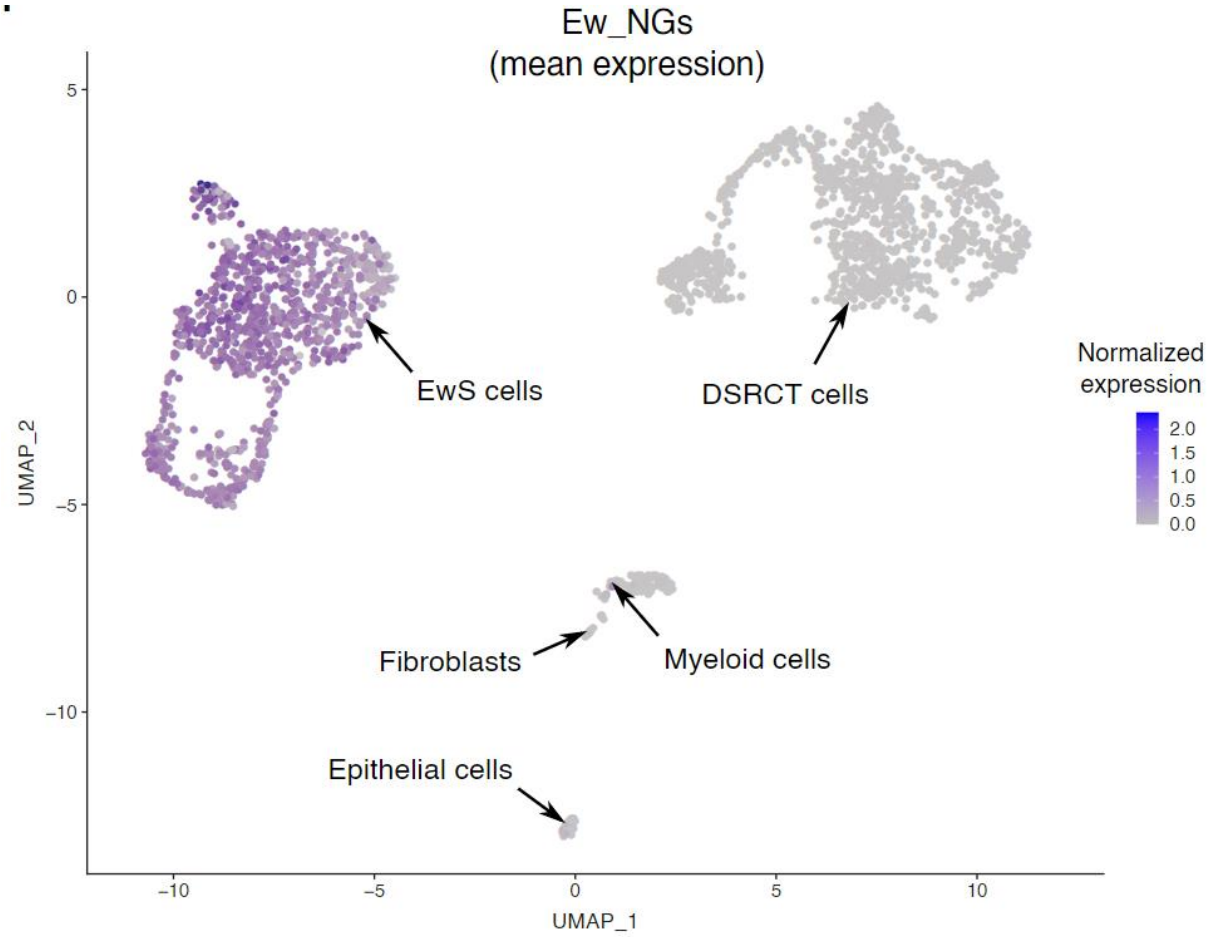

Figure 36: UMAP plot of scRNA-seq of EwS and DSRCT tumor samples showing mean expression level of Ew_NGs.

Altogether, we have shown the existence of a set of EwS-specific neotranscripts, of which a large part are directly induced by the binding of EWS-FLI1 to GGAA microsatellites in otherwise silent intergenic regions. Based on our knowledge of the mechanisms of action of EWS-FLI1, we proposed a mechanistic model to explain this phenomenon: EWS-FLI1 binding sites such as GGAA microsatellites are in closed chromatin conformation in non-EwS cells. Upon binding of EWS-FLI1, these sequences are transformed into neo-enhancers able to activate neighboring known target genes[83,84] but also to induce transcription of Ew_NGs (Figure 37).

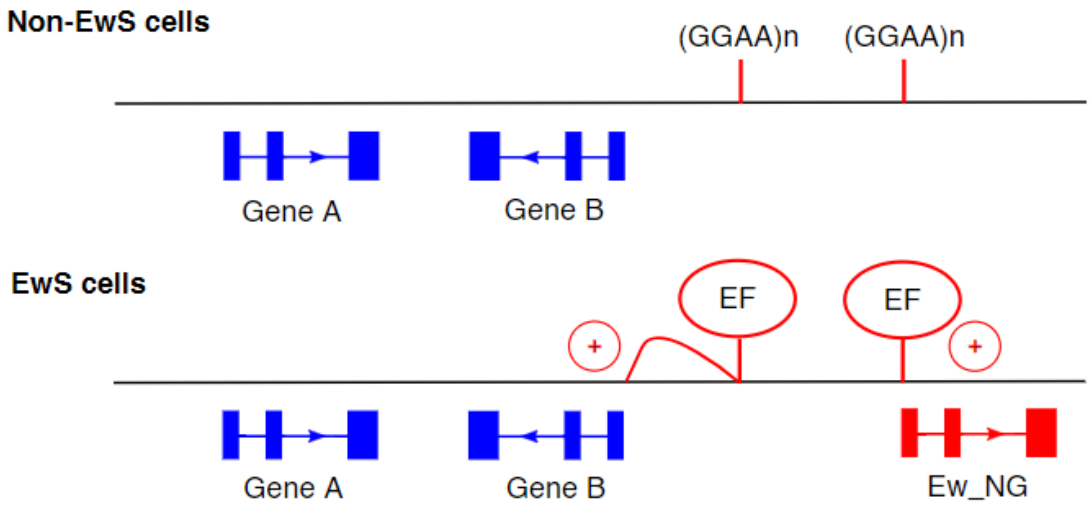

Figure 37: Proposed model for EWS-FLI1 induction of Ew_NGs. EF: EWS-FLI1.

Extension to other sarcomas

EWS-FLI1 is an oncogenic chimeric transcription factor (OCTF) resulting from a translocation between chromosomes 11 and 22 that characterizes EwS[9]. Other sarcoma types are defined by different chromosome translocations giving rise to specific OCTFs[7,8]. Some non-sarcoma cancers are also characterized by OCTFs (Table 2).

Table 2: List of OCTF-driven cancers (non-exhaustive).

Cancer type	Abbreviation	OCTF
Angiomatoid fibrous histiocytoma	AFH	EWSR1-ATF1/CREB1
Alveolar rhabdomyosarcoma	aRMS	PAX3/PAX7-FOXO1
Alveolar soft part sarcoma	ASPS	ASPSCR1-TFE3
BCOR-rearranged sarcoma	BCOR	BCOR-CCNB3
Clear cell sarcoma	CCS	EWSR1-ATF1/CREB1
CIC-fused sarcoma	CIC	CIC-DUX4/NUTM1
Desmoplastic small round cell tumor	DSRCT	EWSR1-WT1
Extraskeletal myxoid chondrosarcoma	emCS	EWSR1-NR4A3
Ewing sarcoma	EwS	EWSR1-FLI1/ERG
Low-grade fibromyxoid sarcoma	LGFMS	FUS-CREB3L2
Mesenchymal chondrosarcoma	MCS	HEY-NCOA2
Midline carcinoma	MIDLINE	BRD-NUT
Myxoid liposarcoma	mLPS	FUS-DDIT3
EWSR1-NFATC2 sarcoma	NFAT	EWSR1-NFATC2
EWSR1-PATZ1 sarcoma	PATZ1	EWSR1-PATZ1
Solitary fibrous tumor	SFT	NAB2-STAT6
Synovial sarcoma	SS	SS18-SSX1/SSX2
TFE3-translocated renal cell carcinoma	TFE3	ASPSCR1-TFE3

Transcription factors bind specific sequences across the genome to induce transcription of downstream genes. By definition, OCTFs are novel transcription factors that do not exist physiologically and are the result of a gene fusion in a cancer cell. Due to the modification of their structure as compared to the wild-type transcription factor which constitutes only one part of the gene fusion, they may bind different sites of the genome and/or recruit different co-factors, thus acquiring oncogenic properties. For instance, EWS-FLI1 binds GGAA microsatellites whereas wild-type FLI1 recognizes a single GGAA canonical motif[67,68]. While most of the binding sites for other OCTF-driven cancers have not been characterized yet, we hypothesized that other OCTFs could also bind specific regions of the genome not bound by the wild-type transcription factor, including otherwise silent regions of the genome, and thus give rise to other OCTF-driven tumor-specific neotranscripts.

As a proof-of-concept that more neogenes could be found in another OCTF-driven cancer, we focused on desmoplastic small round cell tumor (DSRCT), which is an aggressive soft tissue sarcoma mainly occurring in young male adults and is driven by EWS-WT1[85]. Using the same neotranscript discovery pipeline as for EwS (Figure 30), I found 37 DSRCT-specific neogenes (DSRCT_NGs) corresponding to 105 neotranscripts (Figure 38).

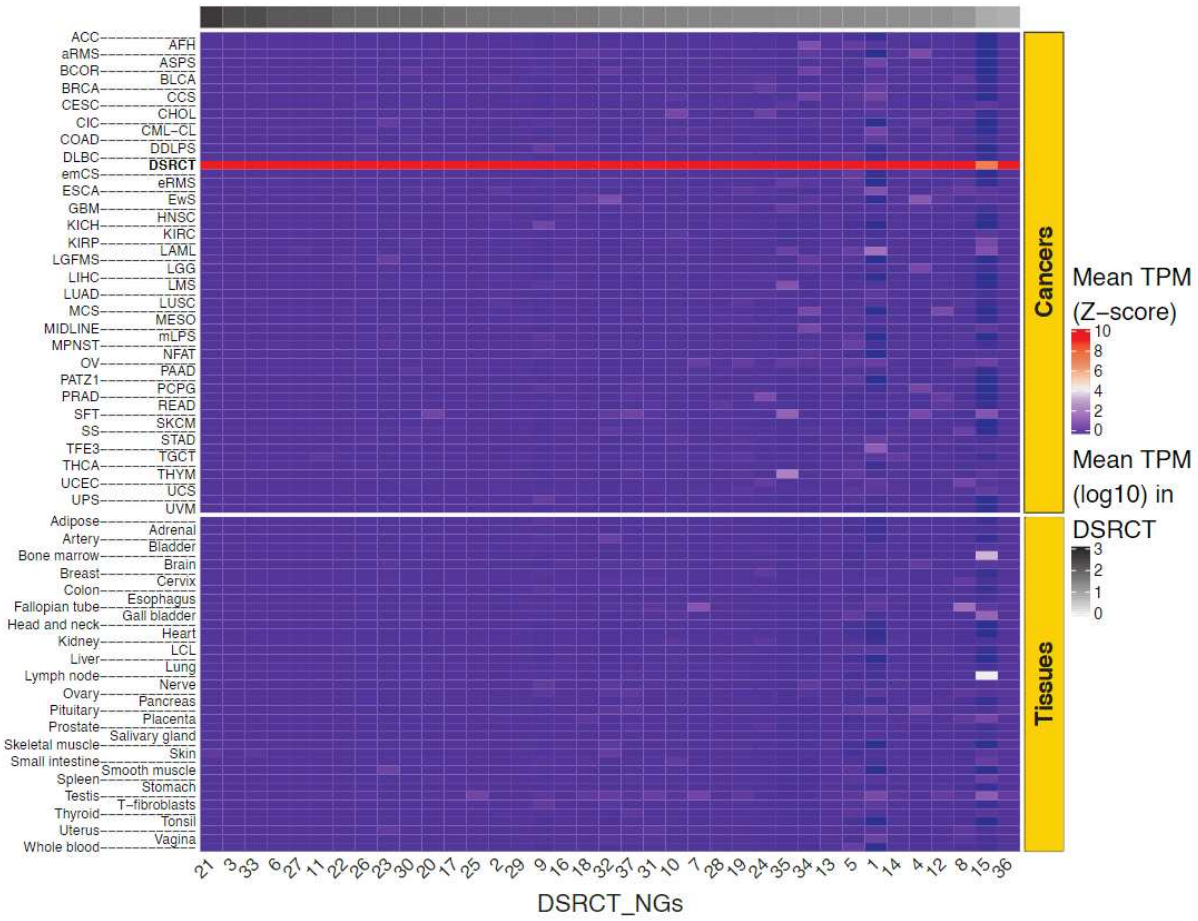

Figure 38: Expression levels of 37 DSRCT_NGs identified from short-read RNA-seq data. Abbreviations are as in TCGA and in the submitted manuscript.

To unravel the potential role of EWS-WT1 in the induction of these DSRCT_NGs, I took advantage of published RNA-seq data of two DSRCT cell lines with siRNA-induced inactivation of EWS-WT1[86], in which I

quantified expression of DSRCT_NGs. Most of DSRCT_NGs showed downregulation with inactivation of EWS-WT1 (Figure 39).

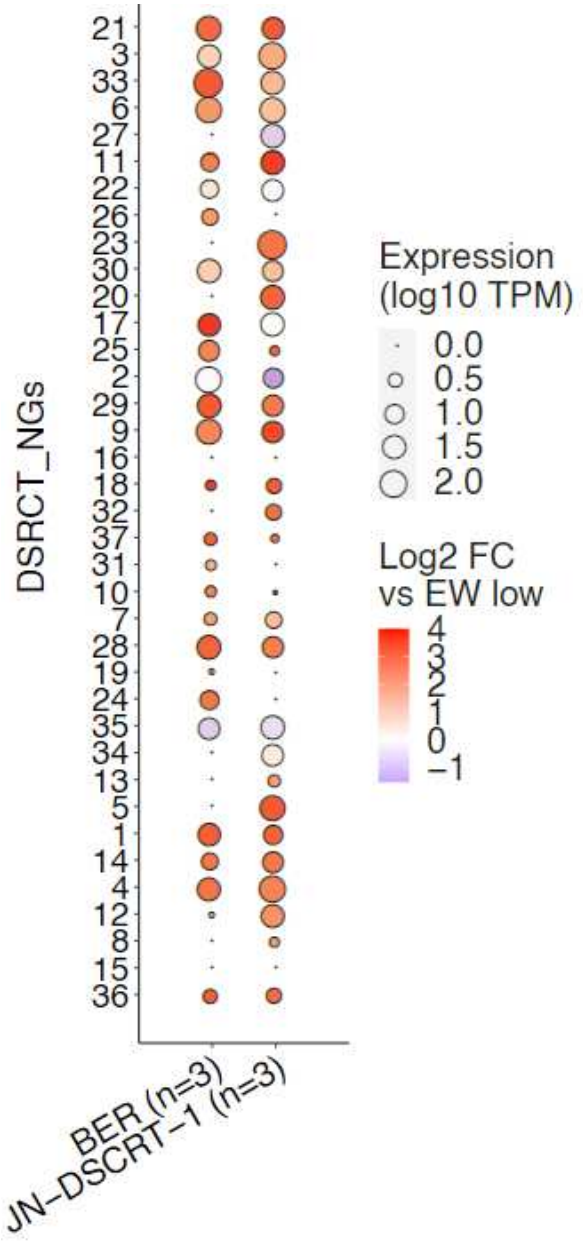

Figure 39: DSRCT_NGs are expressed in DSRCT cell lines and downregulated in EWS-WT1-low (EW low) conditions.

I also used published ChIP-seq data of a cell line (JN-DSRCT-1) to show with the help of Maud Gautier that EWS-WT1 was binding near the TSS of a large part of these DSRCT_NGs (Figure 40).

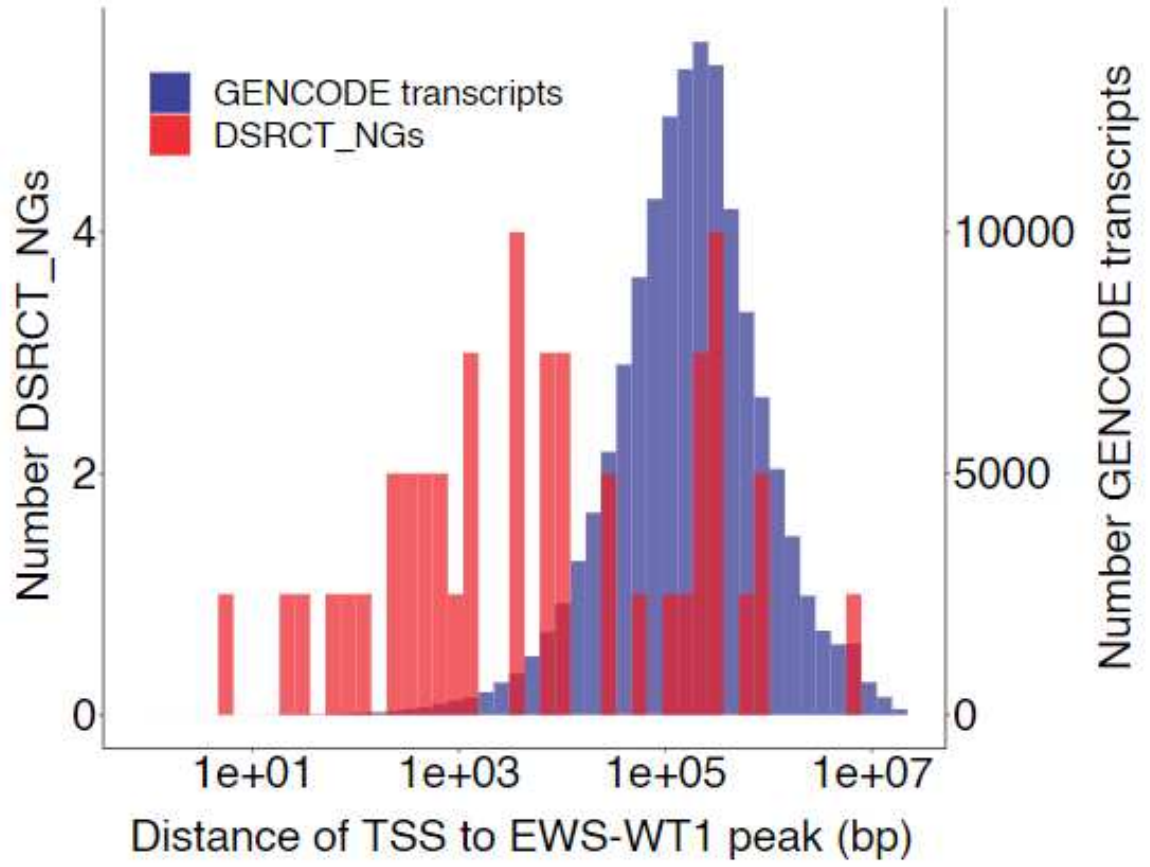

Figure 40: Distance from TSS to nearest EWS-WT1 peak for DSRCT_NGs and GENCODE transcripts.

Finally, I quantified DSRCT_NGs in single-cell RNA-seq data of a DSRCT tumor and a EwS tumor to show specificity of expression in DSRCT tumor cells (Figure 41).

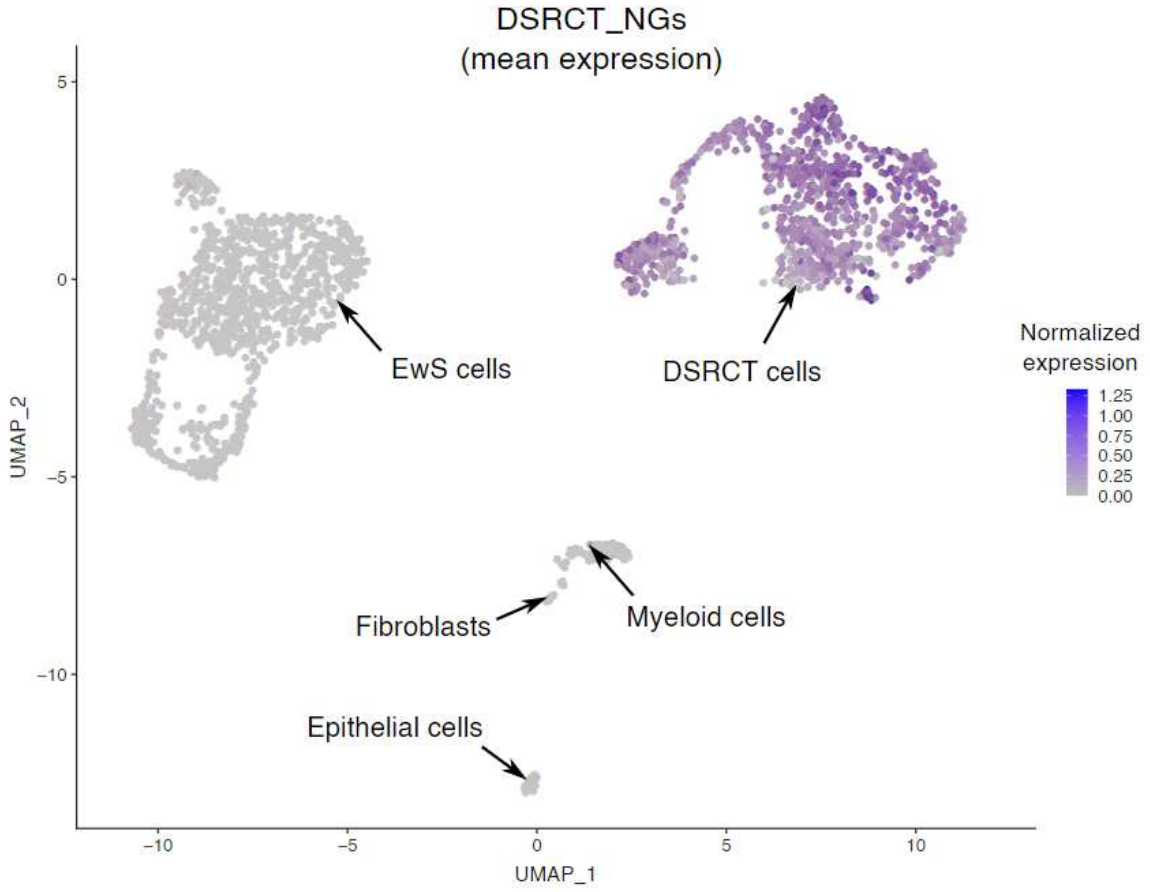

Figure 41: UMAP plot of scRNA-seq of EwS and DSRCT tumor samples showing mean expression level of DSRCT_NGs.

Altogether, these data showed that as in EwS, a set of DSRCT-specific neotranscripts could be identified and were for a large part of them directly driven by EWS-WT1.

Considering that two different OCTFs led to induction of specific neotranscripts, we hypothesized that the same phenomenon might be observed in other OCTF-driven cancers. I took advantage of our collection of RNA-seq in the UGS to run my neotranscript discovery pipeline (Figure 30) in the sixteen other OCTF-driven cancers listed in Table 2. Overall, I identified 398 neogenes corresponding to 807 neotranscripts across all cancer types (Figure 42).

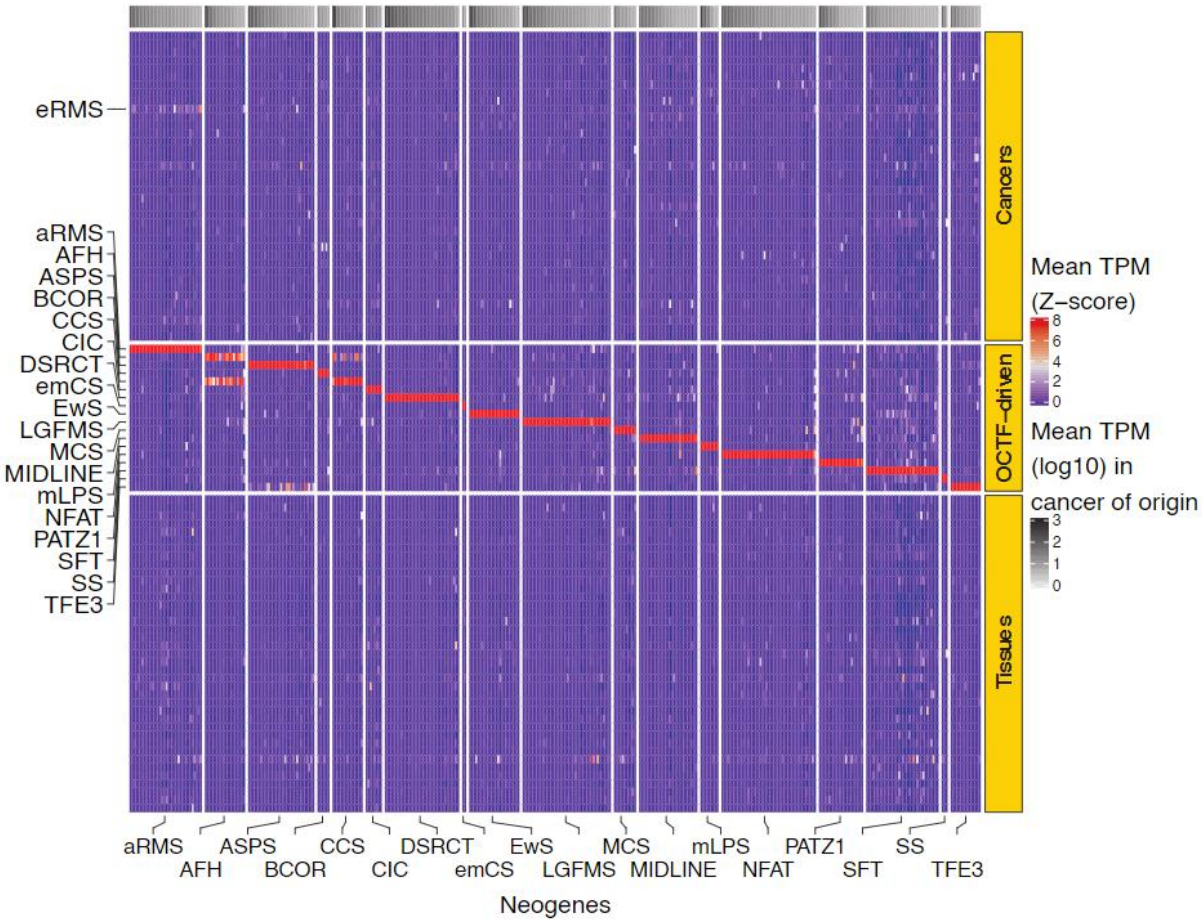

Figure 42: Expression levels of all neogenes identified across eighteen OCTF-driven cancers. Abbreviations are as in Table 2.

This demonstrated that tumor-specific neotranscripts could be found in all OCTF-driven cancers studied. However, contrary to EwS and DSRCT, there were no functional data available in the other tumor types to determine the potential role of the OCTF in the expression of their respective NGs. Nonetheless, some observations that I made while running the neotranscript discovery pipeline were in favor of this potential role: when quantifying expression of candidate tumor-specific neotranscripts across the range of cancers and normal tissues, I noticed that sometimes interesting candidates were also expressed in a second type of tumor. Specifically, some candidate NGs from angiomatoid fibrous histiocytomas (AFH) were also expressed in clear cell sarcoma (CCS), while candidate NGs from CCS could be expressed in AFH (see columns “AFH” and “CCS” in Figure 42). The same phenomenon was taking place between NGs of alveolar soft part sarcoma (ASPS) and TFE3-translocated renal cell carcinoma (TFE3). This intriguing observation was less surprising in the light of the potential mechanism for the induction of NGs: AFH and CCS on one side, and ASPS and TFE3 on the other side, while being distinct pathological entities, share the same OCTF (Table 2). Obviously, this is not as strong an argument as the functional data presented for EwS and DSRCT, nevertheless this suggests that some of the NGs identified here are also directly induced by a specific OCTF, in line with our proposed model of the induction of novel transcripts by the specific binding of an OCTF to silent regions in the genome.

These results demonstrated that OCTFs, which are completely novel and aberrant transcription factors not present physiologically, have the neomorphic ability to bind specific regions of the genome that are normally “silent”, probably as a consequence of their unique structural properties. This binding leads to a cascade of epigenetic events that induces robust transcription of novel spliced, polyadenylated transcripts

(neotranscripts), a phenomenon which is reminiscent of enhancer RNA generation at enhancer sites such as GGAA microsatellites for EwS[87], though enhancer RNAs are usually short and bidirectionally transcribed[88]. As these neotranscripts are products of novel transcription factors that have presumably not previously existed during evolution, it is not surprising to observe that their sequences do not show high scores of evolutionary conservation as compared to protein-coding genes (Figure 43).

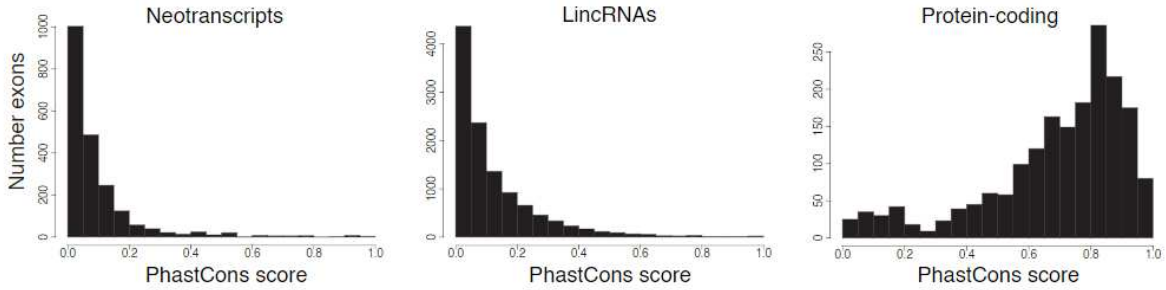

Figure 43: Sequence conservation scores for neotranscripts as compared to protein-coding transcripts and lincRNAs.

Instead, neotranscripts have sequence conservation scores more similar to lincRNAs and may *a priori* be considered as such, as they do not show signs of evolutionary selection to encode stable proteins. However a large number of lincRNAs are known to have functional roles including in cancer development and progression[89]; it thus remains to be seen if these neotranscripts are of functional significance. From an evolutionary standpoint, these neotranscripts seem to be "by-products" of completely novel transcription factors, and considering that they arise from otherwise silent genomic regions with non-conserved sequences, it is probably reasonable to assume that they may not have any essential function. However, biological sequences evolve through acquired mutations and it is tempting to think that this phenomenon may help to address the unresolved question of *de novo* gene birth[90]. Indeed, completely novel genes can arise in evolution from otherwise silent regions of the genome, and can acquire specific functions over the course of evolution: it is probable that they do not exhibit functional activity at their birth, but may progressively acquire functionality through further genetic mutations. In total, it is tempting to hypothesize that the modification of transcription factors could be a general gene-forming mechanism in evolution, by inducing novel binding events in the genome and subsequent opening of chromatin to make possible the appearance of novel units of transcription that may become over the course of evolution *bona fide* functional and even protein-coding genes.

Translation of neotranscripts (Ribo-seq)

We have shown that tumor-specific neotranscripts are expressed in OCTF-driven cancers and should probably be considered *a priori* as lincRNAs without any functional role or protein-coding potential. In fact, many cancer-specific lincRNAs have already been identified in other tumor types[91]. However, it has recently been shown that lincRNAs - notwithstanding their eponymal "non-coding" status - can in some cases be translated by ribosomes and give rise to peptides coded by open reading frames (ORFs). These ORFs are usually shorter than for protein-coding transcripts, nonetheless they can be a source of peptides recognized by the immune system as neoantigens[92–97]. As discussed previously, the neotranscripts we have found could be - if translated - a very important source of tumor-specific neoantigens absent from normal tissues, shared across patients of the same tumor type ("public"), with significant translational relevance as potential targets of immunotherapies. In order to demonstrate that neotranscripts can indeed be translated as other

lincRNAs, we focused on EwS-specific neotranscripts and addressed this question using state-of-the-art techniques.

To be translated, transcripts have to be bound by ribosomes: a technique called ribosome profiling, or Ribo-seq, is specifically designed to detect the RNA molecules that are bound by ribosomes in a cell (Ribosome-Protected Fragments, RPFs)[98]. The principle of this technique is to block active translation with inhibitors such as cycloheximide, then to degrade all RNA not protected by ribosomes and sequence the remaining RPFs by next-generation sequencing[99,100]. This experiment thus allows the profiling of "ribosome footprints", which represent the presence of ribosomes across actively translated transcripts. Due to the three-nucleotide periodicity of ribosome translation that proceeds one codon after another, the specific frame of translation can be inferred from this profile, allowing determination of the open reading frame of the transcript[101]. This technique has notably been used to show that, in addition to coding sequences (CDS) of protein-coding transcripts, other so-called "untranslated" parts of these transcripts - such as 5'UTRs and 3'UTRs - can also be bound by ribosomes and give rise to small ORFs[102]. Similarly, it has allowed to discover that many "non-coding" RNAs are in reality translated[103].

However, Ribo-seq is a non-trivial technique: there is no gold-standard procedure, many protocols exist and results may vary according to experimental conditions. In fact, only a few specialized teams use it in routine; they usually have their own adapted protocol and custom bioinformatic tools. We had previously no experience of this technique in the team and had to ask for help from Olivier Namy's group in the Institute for Integrative Biology of the Cell (I2BC) for our first experiments with Ribo-seq. Céline Collin (a PhD student in the team) thus profiled two EwS cell lines with the help of colleagues from Orsay. However, as I performed quality control of the results, the data was not showing the expected profiles. This highlighted the experimental difficulty of Ribo-seq, even using a protocol from an experienced team. Fortunately, we had also in parallel sent the same samples to a company specialized in Ribo-seq (Ribomaps Ltd, https://ribomaps.com); their experiments were successful and I therefore focused on analyzing this Ribo-seq dataset.

There were six samples in total: two EwS cell lines (A673 and EW7) in three replicates each. The sequencing depth was of 40 to 95 million raw reads per sample. Ribo-seq analysis has many specificities in contrast to other NGS assays: the most important reasons for this being that RPFs normally 1) are of very short length (29 nucleotides in humans); 2) accumulate at the position of the translation start codon; 3) display three-nucleotide periodicity, with the majority of reads positioned in-frame relative to the translated ORF. Therefore, specific bioinformatics tools have been developed in recent years to analyze Ribo-seq experiments[101]; however no gold-standard tool exists and there are multiple "home-made" methods of analysis (for instance our colleague team in Orsay are developing their own tool). After discussion with experts from Orsay and exploration of many existing tools, I decided to focus on a set of complementary packages that performs comprehensive Ribo-seq analysis and is currently being used by many teams: RiboseQC[104] and ORFquant[105]. The first package is designed to perform all Ribo-seq quality control analyses and extract ribosome footprints, while the second one is more specifically aimed at determining the sequences of open reading frames (and the specific isoform being translated for alternatively spliced transcripts).

Before using these specific Ribo-seq analysis tools, there are necessary steps of data pre-processing that are common to other NGS analyses: notably adapter-trimming and alignment of reads to the genome. As discussed previously, there are no gold-standard analysis pipelines for Ribo-seq, including for these early steps. After exploring many protocols used in the literature, I settled for the following consensus pipeline implemented with commonly used tools. Some protocols contain a specific step of discarding ribosomal RNA - which is expected to be present in samples - by aligning to ribosomal genes. I chose the alternative option of directly mapping to the whole human genome without this step. To avoid spurious mapping of ribosomal RNA, I used one of the best-performing mappers (STAR[40]) and specified very stringent criteria as opposed to classical RNA-seq, in order to only conserve reads mapped with very high confidence:

specifically I discarded all multi-mapping reads, and all reads with more than two mismatches during alignment to the genome. The outline of the whole pipeline is in Figure 44.

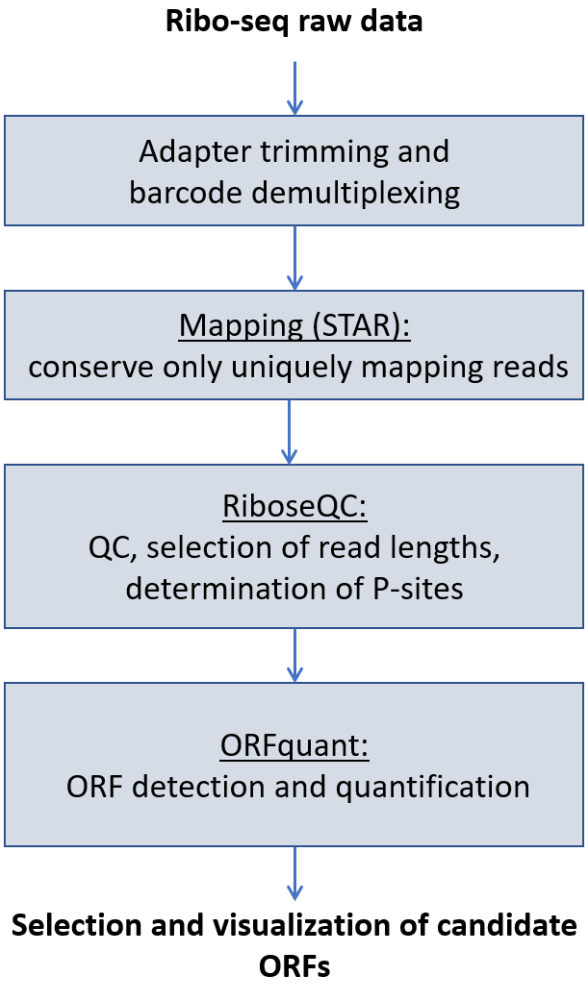

Figure 44: Pipeline for Ribo-seq analysis.

Method: Adapters were trimmed using Trim Galore! (v.0.6.5). Ribo-seq reads were mapped with STAR (v2.7.0e) with options --outFilterMultimapNmax 1 --outFilterMismatchNmax 2 to conserve only uniquely mapping reads with a maximum of two mismatches, using a GTF file containing GENCODE v19 reference transcripts to which was added the annotation of the Ew_NGs. Ribo-seq quality control analyses were performed with RiboseQC (v0.99.0) using default parameters, after which P-site positions and number of reads mapping to Ew_NGs (raw and TPM) were extracted. ORF predictions were then performed with ORFquant (v1.02.0) using default parameters on RiboseQC output data.

A crucial part of Ribo-seq analysis is quality control: since all RNA molecules in the sample are sequenced at the end of the experimental protocol, we have to make sure that the experiment went well and that we are indeed looking at RPFs. Indeed, a "failed" experiment without successful isolation of RPFs would still result in the sequencing of millions of transcripts present in the sample, that could also map in majority to protein-coding genes but would not be proper RPFs. The importance of this step is compounded by the fact that Ribo-seq is experimentally challenging as previously discussed, and "failed" experiments happen more

often than for other well-established techniques such as classical RNA-seq (this was indeed the case for our first experiments in the lab). To select RPFs present in the data, we must first discard all RNA molecules that are potentially contaminating, notably all fragments that do not have the required length of 29 nucleotides. For this, all reads that were aligned uniquely to the genome are filtered based on their length; by plotting the length distribution of reads, ideal Ribo-seq data should display a peak at 29 nucleotides, more or less a few nucleotides to account for experimental variability. A second quality control measure is the distribution of aligned reads onto coding sequences (CDS) and other sequences: we expect a major part of RPFs to align to CDS. This second control is necessary but not specific for Ribo-seq (it would also be expected of classical mRNA-seq). A third control is then applied that is this time truly specific to Ribo-seq: by plotting the position of reads relatively to the start codon and known ORF of protein-coding transcripts, we expect to observe accumulation of reads at the start codon and three-nucleotide periodicity with the majority of reads positioned in-frame relatively to the ORF ("frame 0", as opposed to "frame 1" and "frame 2").

I used RiboseQC to perform all these quality control analyses. The read length and location distributions for our six samples are represented in Figure 45.

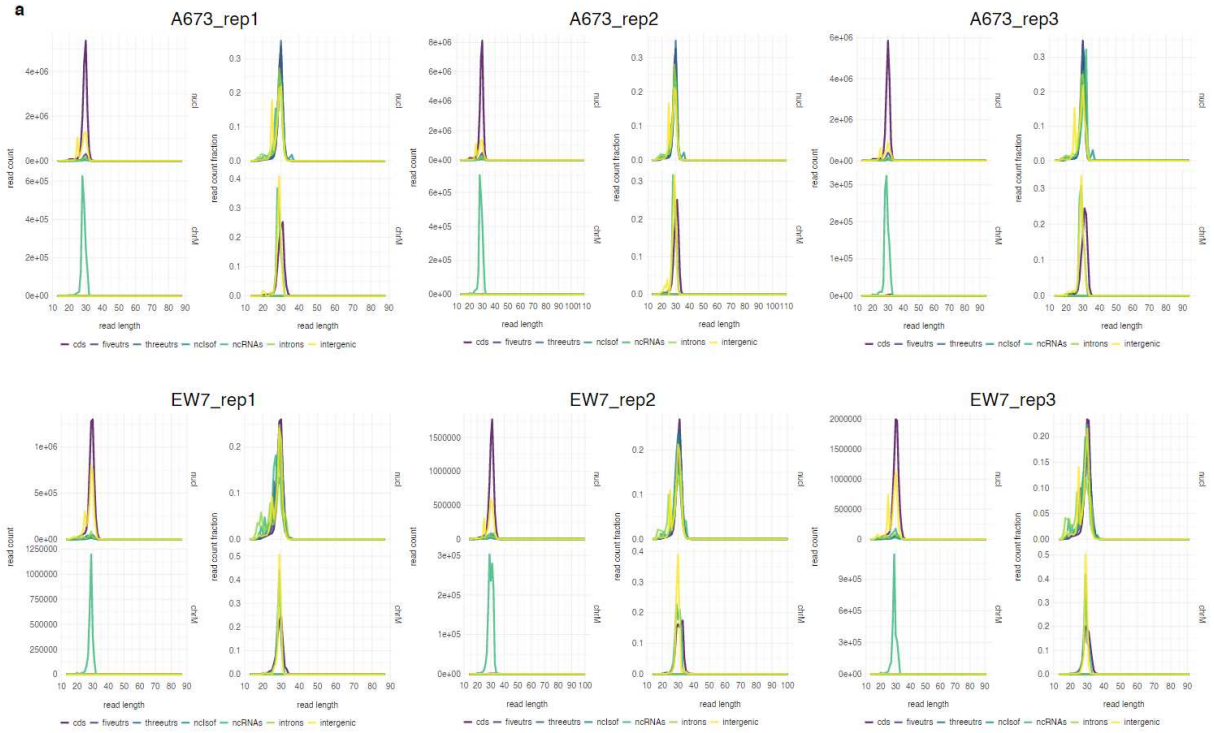

Figure 45: Ribo-seq read length and location distribution.

As expected, the read length distributions are peaking at 29 nucleotides and the location distributions show that most of the reads are mapping to CDS (purple). These profiles are comparable to high-quality Ribo-seq datasets in the literature. To explore the quality control of nucleotide periodicity, each read must be assigned a frame among the three possible ones. For this, one does not use directly the 5' end of the read, because the ribosome P-site - that corresponds to the translating position of the ribosome - is positioned at an offset (12 nucleotides usually) downstream of the 5' end. When plotting the P-site profiles (ribosome footprints) of a "metagene", i.e. all genes collapsed to the same abscissa, we expect to see: 1) most of the P-sites at the starting codon "ATG"; 2) most reads in the "frame 0" that is the reading frame corresponding to the known ORF. P-site profiles for our six samples are represented in Figure 46.

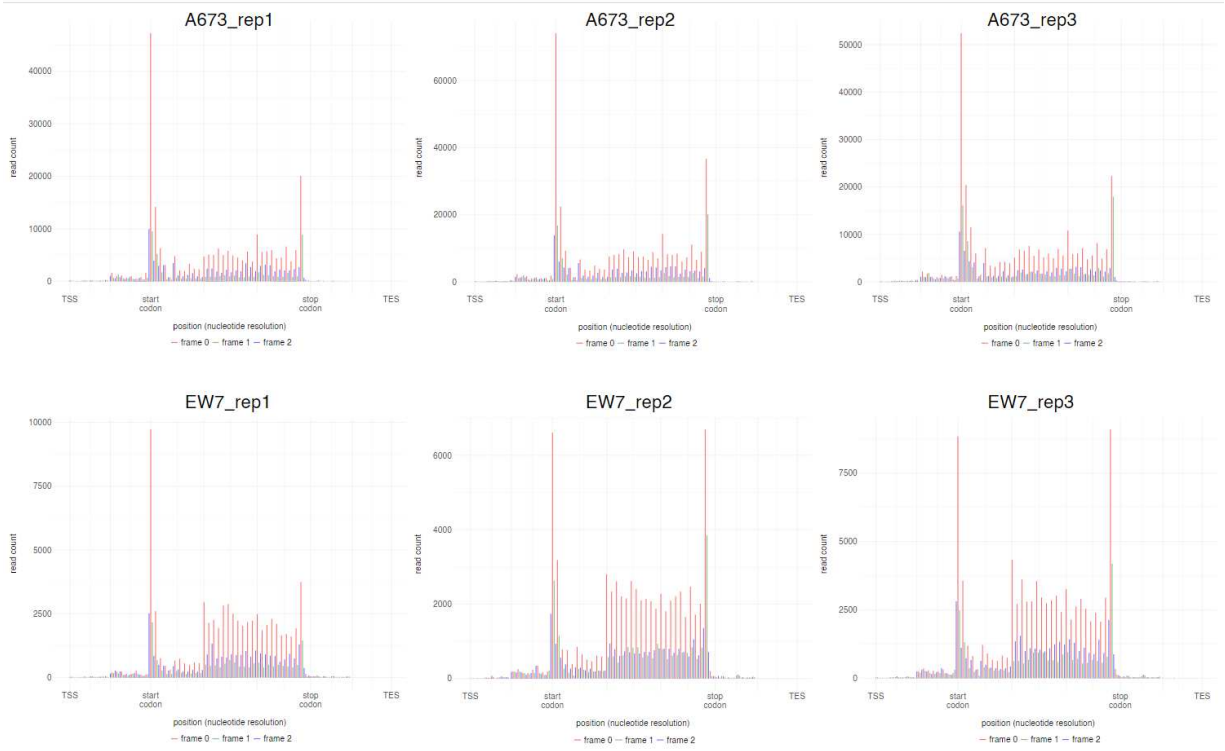

Figure 46: P-site profiles colored by frame. TES: transcription end site.

The P-site profiles each display the expected three-nucleotide periodicity and accumulation at the starting codon of translation. After all these quality controls, we can confirm that we are indeed analyzing genuine RPFs in these six samples, in contrast to the first experiments done in our team that did not pass these quality criteria (data not shown in this manuscript).

Once we were confident of the quality of this data, we could ask the question of whether some neotranscripts are bound by ribosomes and potentially translated into peptides. After discarding reads outside the expected length distribution by RiboseQC, I quantified the RPFs mapping to the genomic loci of neotranscripts. As a control for these analyses, I used high-quality Ribo-seq data of two non-EwS cell lines (K562, HepG2) from the literature[105] which I processed using exactly the same pipeline. After normalization in TPM by RiboseQC, sixteen Ew_NGs showed a level of associated RPFs of more than 0.1 TPM. Their expression values in EwS and non-EwS cell lines are represented in Figure 47.

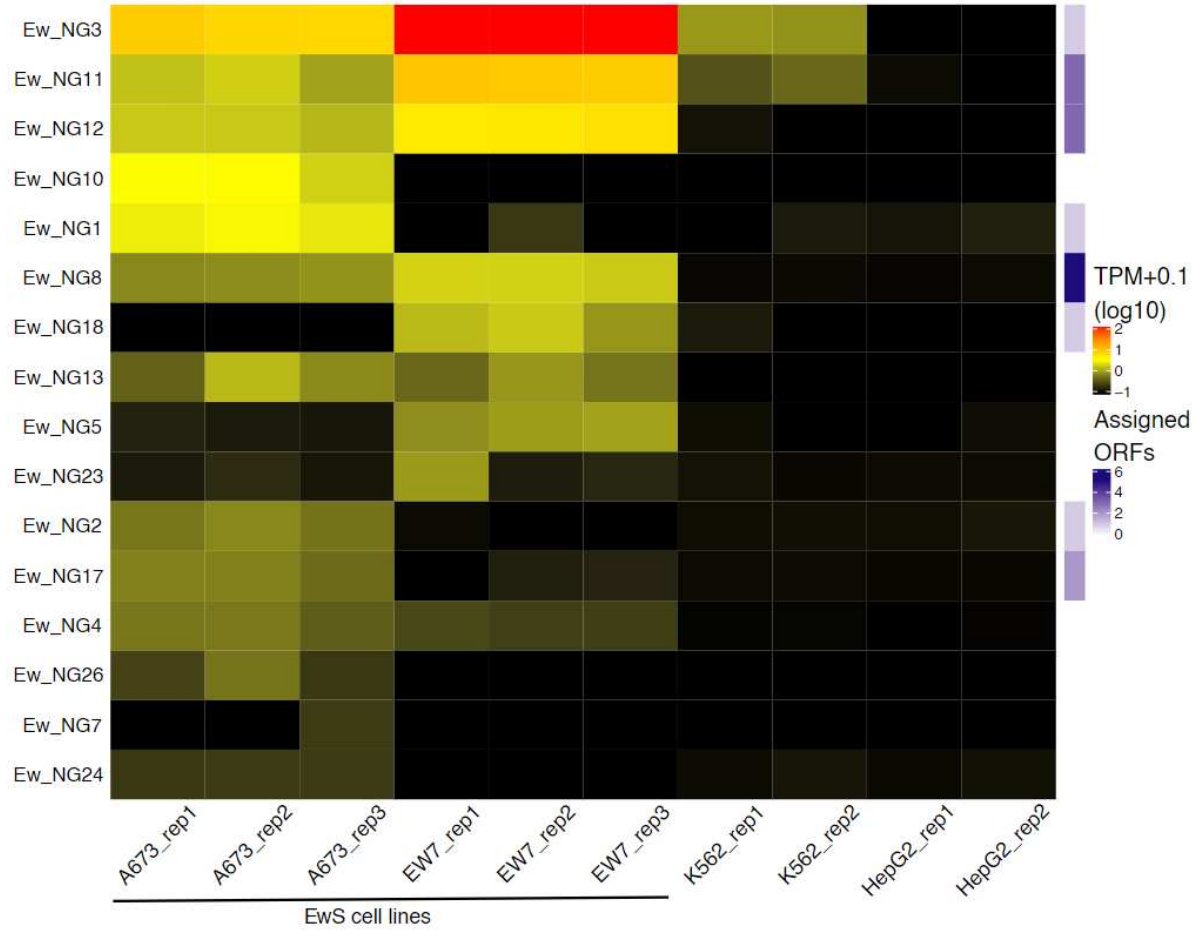

Figure 47: Heatmap of ribosome protected fragments (RPFs) mapping to Ew_NGs in EwS and non-EwS cell lines. Levels are in log10(TPM+0.1). Ew_NGs are ordered from top to bottom by maximum RPF levels. Right heatmap reports number of computationally predicted ORFs in corresponding Ew_NG. rep: replicate.

Based on the periodicity of RPFs mapping to Ew_NGs, the software ORFquant was also able to infer the sequence of seventeen ORFs for six of these NGs (the number of inferred ORFs for each NG is represented in Figure 47).

Overall, the amount of RPFs associated to Ew_NGs is low (except for *Ew_NG3*, which is also the most abundant NG in RNA-seq) as compared to protein-coding transcripts, which can show levels of over a thousand TPM for highly translated genes. However, this is not surprising considering that translation of lincRNAs is usually much less prevalent in the cell. Though numbers of RPFs are not high enough to allow confident ORF sequence prediction in most Ew_NGs, these results show that a large part of them can be bound by ribosomes and potentially translated. While this phenomenon is mostly present at low levels, it does not necessarily compromise their ability to be recognized as antigens[96]. Also, it is possible that increasing the depth of sequencing could have allowed the detection of more RPFs associated to these and additional Ew_NGs, as well as more computationally predicted ORF sequences.

Finally, another argument that reinforced the confidence in these Ribo-seq results came from a completely orthogonal approach to detect translation of neotranscripts, as described next.

Translation of neotranscripts into peptides (Proteomics)

Ribo-seq aims at detecting ribosome-bound transcripts that are potentially translated, while proteomics looks directly at the peptide level: this set of techniques uses mass spectrometry (MS) to detect all proteins in a sample. Briefly, proteins are digested by trypsin and resulting tryptic peptides are identified based on their characteristics of mass, charge, and polarity[106]. To explore the potential translation of neotranscripts into peptides, we used mass spectrometry in EwS cell lines. Floriane Petit (a post-doctoral student in the lab) profiled ten EwS cell lines in five replicates each with the help of the Institut Curie mass spectrometry facility. MS has the particularity that it detects tryptic peptides according to their mass to charge ratio, but the experimentally obtained MS "spectra" have to be compared to reference spectra derived from known peptides potentially present in the sample. In a way, this is analogous to genome mapping of RNA-seq which uses the reference transcriptome to infer the identity of the read sequences. Thus, MS is a reference-based search for peptides, and one cannot detect a peptide that is not in the reference used for searching. Usually, one searches against reference MS spectra of peptides derived from known proteins, such as the set of human proteins in Uniprot. However, for our purpose to search for peptides translated from neotranscripts, I had first to computationally predict all ORF sequences that could potentially be generated from neotranscripts. Using the software ORFfinder, I derived all possible ORFs from neotranscript sequences in all three reading frames, starting either at a canonical start codon "ATG" or a non-canonical one. This "neopeptide" list was then fused with the Uniprot database, and the search was performed against this merged database in all samples processed by the MS platform. The bioinformatic processing was performed by dedicated bioinformaticians in the MS facility and Olivier Ayrault's team in Institut Curie (Jacob Torrejon Diaz).

Method: ORFs of Ew_NGs were computationally predicted with ORFfinder (v.0.4.3; options: minimal length=75 nucleotides; start codon: any) to constitute a database of potential neopeptides (Fasta file). For identification, the data were searched against the Homo Sapiens (UP000005640_9606) UniProt database, this neopeptide database and a database of the common contaminants using Sequest HT through Proteome Discoverer (version 2.4). Enzyme specificity was set to trypsin and a maximum of two miss cleavages sites were allowed.

Overall, 247 neopeptides were identified in EwS samples, including 65 that were present in at least three out of five replicates of all ten cell lines and that could be precisely quantified. These included three tryptic peptides predicted to be derived from the same ORF that was identified in Ribo-seq for *Ew_NG3*, the most highly expressed NG in RNA-seq and Ribo-seq (Figure 48). To confirm specificity of this finding, we searched for neopeptides in eight non-EwS (medulloblastoma) samples processed with a comparable protocol (data from Olivier Ayrault's team in Curie) and found zero EwS neopeptide in this dataset.

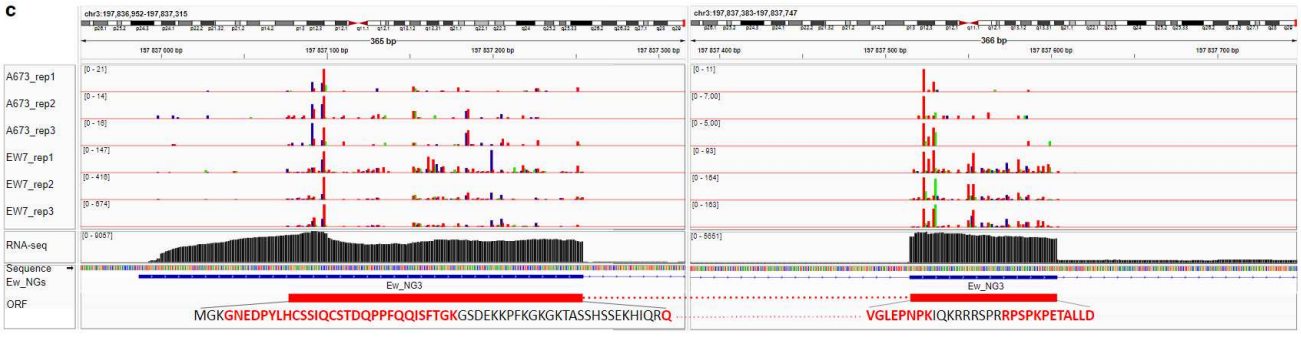

Figure 48: Genomic view of first two exons of Ew_NG3 and derived ORF predicted by Ribo-seq and detected in mass spectrometry (NG3:0:356). From top to bottom: Ribo-seq P-sites for EwS cell lines, colored by frame (0: red; 1: green; 2:blue); RNA-seq in A673; nucleotide sequence; transcript annotation; predicted ORF. Peptides highlighted in red are detected in mass spectrometry.

Altogether, while quantification levels are - as for Ribo-seq - moderate compared to known proteins, proteomics confirms that neotranscripts can be translated into peptides. Moreover, finding evidence for the same translated ORF in Ribo-seq and MS is very significant and adds confidence to these results. Since Ribo-seq and MS are both techniques more suited to detect products of highly translated protein-coding genes, it is a very promising result that we can already find signal for the translation of neotranscripts in this data, considering that other neotranscripts may also be translated at lower levels and may not have been detected by these techniques.

As this work on neotranscripts has been submitted to a peer-reviewed journal, I reproduce the manuscript in Annex 2.

From neotranscripts… to neoantigens?

Since neotranscripts can be translated into peptides - at least in EwS -, it is legitimate to further ask whether these "neopeptides" may be presented by the MHC complex at the cell surface and recognized by the immune system as neoantigens. As discussed previously, this would be of high translational significance for the design of immunotherapies, not only in EwS but also in all other cancers harboring potentially translated neotranscripts.

This is the reason why several experiments are currently undergoing in our team to characterize: 1) the ability of the MHC complex to present neopeptides to the immune system, and the ability of T lymphocytes to recognize them and mount an immune response following this recognition (MHC tetramer assays); 2) the potential presence of naïve T lymphocytes able to recognize and respond to such neopeptides in blood of healthy patients; 3) the potential presence of memory T lymphocytes that recognize neopeptides in blood of EwS patients, as a proof of their previous encounter with neopeptides. Using a commonly used software to predict peptide affinity for binding to the MHC-I complex (NetMHCpan 4.1[107]), I estimated that, in EwS only, there were 4355 potential peptides predicted to be strongly bound to the MHC-I HLA-A2 allele (the most widespread allele in the European population). We cannot of course test all these peptides; the Ribo-seq and proteomics data have enabled us to choose the best candidates for these immunology experiments, including the ORF from *Ew_NG3* that was detected independently both in Ribo-seq and MS experiments.

If we find convincing evidence that neopeptides derived from neotranscripts can indeed be recognized as neoantigens and stimulate an immune response against tumor cells in lymphocytes, these could be used as targets for immunotherapies such as cancer vaccines[108,109], adoptive cell therapy with chimeric antigen receptor (CAR)- or TCR-T cells[110] or bispecific antibodies[111], with potential application in multiple types of sarcomas.

Transcriptomic characterization of sarcomas and microenvironment at the single-cell level

Introduction

Up to now I have focused on RNA sequencing of bulk tumor samples for characterization of tumor cells and their microenvironment, however it is inherently impossible to assign measures obtained in bulk samples to a specific cell population, let alone a single cell. Multiple methods have thus been developed to computationally "deconvolve" the signal obtained in bulk experiments into the contributions of different cell populations composing the bulk sample, such as MCP-counter[37] discussed previously. While these methods certainly perform well in many cases, with rigorous benchmarking performed in studies involving controlled experiments of mixing pure cell populations *in vitro* or *in silico*[112], it is not possible to evaluate their results in comparison to "gold-standard" references when considering tumor samples.

In recent years, this technical difficulty has been successfully overcome with the advent of single-cell technologies that allow the isolation of single cells and identification of their material (DNA[113], RNA[114], even proteins[115] and multiple modalities in the same cell[116,117]). Most of these techniques rely on ingenious barcoding systems to conserve the information of the cell-of-origin of each single measurement[118,119]. The most advanced of single-cell technologies is single-cell RNA-seq (scRNA-seq): multiple experimental protocols are available, with varying throughput in numbers of cells, depth of sequencing per cell, or read length (full-length[120,121] or 3'/5'-end). One of the most widely used methods for scRNA-seq is the Chromium platform developed by 10x Genomics (https://www.10xgenomics.com/), which is based on droplet isolation of single cells, barcoding of single cells and high-throughput sequencing of the 3'-end of transcripts.

Single-cell RNA-seq of sarcomas at the Institut Curie

Single-cell RNA-seq has already widely been applied to the study of tumor samples[122–126], to characterize the different populations of tumor cells and the tumor microenvironment at single-cell resolution. While some protocols can be applied on frozen tissue samples, the best experimental quality is obtained with rapidly processed fresh tumor tissue[127]: this may constitute an important limiting factor for the study of clinical samples of patients. However, we have the advantage in Institut Curie of being a reference center for sarcomas; we are especially home to one of the most active surgery departments for soft-tissue sarcoma in France. In collaboration with our surgeons (Sylvie Bonvalot, Dimitri Tzanis), we have thus been able to launch a project named "SingleSARC" (led by Sarah Watson; experiments performed by Nadège Gruel) to characterize by scRNA-seq some types of sarcomas that are resected at the Institut Curie, in order to better characterize their tumor microenvironment and decipher the different tumor cell populations.

Since scRNA-seq is costly and requires fresh tissue, it is not currently possible to profile large numbers of patients by this technique. Moreover, sarcomas are relatively rare diseases and surgical samples are precious. To account for these limiting factors, we decided to concentrate our attention on a type of adult soft-tissue sarcoma named dedifferentiated liposarcoma (DDLPS).

Dedifferentiated liposarcoma (DDLPS)

This is one of the most frequent sarcomas in adults. As its name implies, this tumor has some characteristics of, and is thought to be derived from, normal adipose tissue. Indeed, so-called "well-differentiated" liposarcoma is composed of tumoral adipocytes with a macroscopic appearance of fat tissue. DDLPS is a subtype of liposarcoma[128], which further includes other subtypes such as pure well-differentiated liposarcoma (WDLPS), pleomorphic liposarcoma and myxoid liposarcoma. While this last subtype is a translocation-related sarcoma (it was studied in the neotranscripts project), DDLPS and WDLPS are molecularly characterized by a driver genomic alteration - chromosome 12q amplification - with overexpression of corresponding genes in the amplicon, notably *MDM2* and *CDK4*.

WDLPS is thus composed of tumoral "adipocytes" with chromosome 12q amplification; it is a slow-growing tumor that is usually located in the abdominal cavity and can measure up to tens of centimeters at diagnosis, since its slow-growing pace and location do not usually cause any symptoms apart from progressive abdominal diameter growth – often confused with physiologic weight gain. This is a relatively benign tumor, which does not metastasize and can be cured by surgery[129].

However, an intriguing phenomenon sometimes occurs in WDLPS: the appearance of one or multiple areas of undifferentiated tumor cells also displaying the 12q amplification, classically designated as "dedifferentiated" contingents inside the initially pure WDLPS. As soon as this "dedifferentiated compartment" appears, the tumor is classified as a "dedifferentiated liposarcoma" (DDLPS). Thus, there is a very close relationship between these two subtypes of tumors, since DDLPS seems to arise on a background of WDLPS. In fact, the sharing of a same molecular driver alteration (12q amplification) and their invariable temporal succession suggest that WDLPS and DDLPS are probably more two different stages of a same disease process than two separate entities. The term "dedifferentiation" precisely assumes that the undifferentiated cells of the dedifferentiated compartment are originating from the well-differentiated tumoral adipocytes, however this has never been demonstrated.

Finally, DDLPS is also one of the few sarcoma types to have shown promising responses to anti-PD1 immunotherapy[12]. Altogether, adding to the fact that DDLPS is one of the most operated sarcomas in Institut Curie, we decided to focus on DDLPS for our scRNA-seq studies, to explore the tumor microenvironment in light of the potential responses to immunotherapy, and to decipher the molecular underpinnings of the relationship between well-differentiated (WD) and undifferentiated/"dedifferentiated" (DD) cells, notably the putative dedifferentiation of WD cells into DD cells.

Single-cell RNA-seq of DDLPS

The experimental protocol was as followed: as soon as a DDLPS tumor was resected by the surgeon, it was sent to the pathologist who selected two samples for scRNA-seq: one from the WD compartment and another from the DD compartment. These samples were then processed by Nadège Gruel with cell dissociation and scRNA-seq using the 10x Genomics Chromium platform (chemistry version 3). In total, we profiled more than ten DDLPS tumors in this way. However, due to different technical issues (notably pathologists eventually refuting the diagnosis of DDLPS, or scRNA-seq failure on one or two samples of the same tumor), we only focused in the end on four different tumors (eight samples) that were of high quality and came from pathologically confirmed DDLPS. I performed all bioinformatics analyses of the scRNA-seq data generated by this project.

In parallel with the development of experimental protocols for scRNA-seq, a myriad of methods have been developed for the analysis of the high-dimensional data that result from these experiments[130,131]. As a simple calculation shows, while one sample of bulk RNA-seq can be represented by a vector of tens of thousands of entries, one single dataset of scRNA-seq is composed of a set of thousands of such vectors

(one per individual cell). This multiplication of data points has catalyzed the development of novel methods of analysis; it has also stimulated the widespread adoption in modern computational biology of techniques from machine learning and even deep learning, that can only be used with sufficiently large high-dimensional datasets[3,132].

This richness of data comes with many opportunities but also many potential pitfalls. This is moreover complicated by the high sparsity of the single-cell data: due to the necessarily incomplete sequencing of all transcripts in each single cell, especially in low sequencing depth protocols such as 10x, most of the entries are equal to zero ("dropouts"[133]) in the resulting count matrix.

With these caveats in mind, many aspects of cancer biology can nonetheless be explored and precisely characterized with scRNA-seq: 1) the composition at single-cell level of bulk tumor tissues; 2) the study of transcriptomic profiles of different cell types and states, of tumor cells and the microenvironment; 3) the relationship between cell types and states such as differentiation "trajectories"; 4) the gene regulatory networks within, and cellular interactions between cells in the tumor sample.

I will now detail the bioinformatics analyses that I performed on these DDLPS scRNA-seq samples.

Analysis of single-cell RNA-seq

As there are literally hundreds of computational methods available to perform analyses in scRNA-seq (stating that currently about one new tool is published every day should not be too far from reality[134]), I inevitably tested a lot of them during my thesis, but I chose those that I used in the end based on the following criteria: open-source, easy-to-use, widely used tools with good documentation available and a broad user community. I only used less well-known and more specialized tools in cases of more complicated or field-specific analyses (such as gene regulatory network inference), I also chose an alternative tool for some analyses if the commonly used tool showed obvious limitations.

Since we used the 10x Genomics platform to perform experiments, I chose to use the dedicated Cell Ranger software to preprocess the raw data, as advised by 10x Genomics and performed by most users of this platform. This preprocessing step is designed to generate the count matrix for a scRNA-seq sample, basically by mapping reads to the 3'-end of reference transcripts, and using the barcode associated to each read to assign it to its single cell of origin in the experiment. An "UMI" (unique molecular identifier) tag attached to each read moreover allows to identify PCR duplicates and thus avoids multiple quantification of the same read. To be more specific, a read is not assigned to a single "cell" but to a single "droplet", since this technology is based on isolating single cells inside droplets before sequencing their RNA. While one droplet is ideally only containing one single cell, it is possible and indeed happens that one droplet contains in fact more than one cell (so-called "doublets" for two cells in the same droplet), or no cell at all ("empty droplets"). In these cases, the specific barcode assigned to the droplet will be associated to either two or more different cells, or no cell at all ("empty droplets" contain only "ambient" RNA that may have been enclosed inside the droplet during the experiment). A first step in the processing of scRNA-seq data is thus to discard these barcodes corresponding to "empty droplets": they contain a significantly lower number of reads per barcode, and are discarded by a dedicated well-performing package called "EmptyDrops"[135] that is implemented inside the Cell Ranger command-line tool that I used.

Method: Single cell RNA-seq raw base call (BCL) files were demultiplexed and converted into FASTQ files by using the 10X Genomics Cell Ranger pipeline (v3.0.2) "mkfastq" command. FASTQ files were then processed with the Cell Ranger "count" command to perform quality control, barcode processing, and single-cell gene counting. Sequencing reads were aligned to the GRCh38 human reference genome (v3.0.0 Cell Ranger index).

After this first step of preprocessing, the dataset consists of a count matrix in which each line is a gene (also called a "feature") and each column is - normally - a single cell (more rigorously it is a single droplet, so it could still be a "doublet", whereas "empty droplets" have normally been filtered out in the previous step). Before further downstream analysis, there are some quality controls that have to be performed, in order for instance to filter out problematic low-quality cells, or filter out genes not expressed in the sample to facilitate computation. I chose to perform all these downstream analyses with Seurat (https://satijalab.org/seurat/), which is currently the most widely used package in the community for this type of analyses.

First, one usually wants to discard genes (features) that are very lowly or not at all expressed in the dataset, in order to reduce the number of lines of the - already huge - count matrix. This is usually done by fixing a minimal threshold for the number of cells expressing one feature, for instance one may discard all features that are not expressed in more than 3 cells in the dataset (there are usually thousands of cells in one experiment). Then, it is common practice to filter out cells that do not show proper quality criteria ("low-quality cells"), such as cells that have very low RNA content or few expressed features. For instance, a commonly used filter is to discard all cells that have less than 200 expressed features (an expressed feature has non-zero expression). This criteria of number of expressed features is sometimes also used to discard cells which display a high value for this parameter, the assumption being that "cells" with too many expressed genes may correspond to "doublets". However, the number of expressed genes is known to be highly variable between different cell types: there is thus a risk of discarding some cell populations that truly display an "outlier" distribution of higher number of expressed genes[130]. To account for this, I chose not to use this filter based on the higher end of number of expressed genes. Finally, it is known that "dying" cells - including cells damaged by the experimental protocol - release a lot of mitochondrial transcripts in the droplet, so it is also common practice to discard all cells showing a higher proportion of mitochondrial reads. However, this proportion can also vary with the underlying biology of the cell[130], for instance cells relying on a high amount of oxidative phosphorylation for energetic purposes (e.g. cardiomyocytes) display a very high proportion of mitochondrial reads and may be filtered out if not accounting for this biological variability. As we did not know *a priori* what proportion of mitochondrial reads was to be considered normal in DDLPS samples, and considering that this parameter may vary according to underlying biology, I chose to fix a filtering threshold based on visual appreciation of the distribution of mitochondrial read proportion in all cells of each experiment. Outlier cells with proportion of mitochondrial reads higher than this threshold were then discarded.

As a more general comment on quality control analyses and filtering procedures in scRNA-seq, there are no agreed gold-standard criteria for these[130], since every experiment can be different in terms of protocol, cell populations and other technical or biological factors. In fact, one constant in computational analysis of biological data is the need to "correct" for technical artefacts of the experimental assay, without "erasing" true underlying biological variability that may itself masquerade as a technical artefact. Indeed, unraveling technical from biological variability is often challenging and sometimes impossible. This issue of trade-off between technical "correction" and "conservation" of biological variability is particularly exacerbated in scRNA-seq, as will be discussed later for data integration. In the case of filtering, one must find the correct balance between throwing out real technical artefacts (low-quality cells and doublets) and true biologically outlier cells (for instance cells with intrinsically lower or higher RNA content). This explains why filtering procedures used in the literature are often dependent on results of quality control analyses and vary according to the specific experiment analyzed; they are data-driven and adaptive procedures. Unfortunately, it does not facilitate comparison and reproducibility of analyses, so I tried as much as possible to conserve the same preprocessing steps for all my analyses after assessment of quality control analyses, and to use commonly used criteria from the literature. I also chose to follow this rule, which is of course debatable and reflects my own "philosophical" stance for this: I preferred not to apply too stringent criteria for filtering cells out, in order to be able to recover potentially interesting rare populations that may

be of biological significance, with the risk of keeping truly technical outlier cells (for instance “doublets” or lower-quality cells), since these problematic cells can generally be detected and accounted for in the downstream steps. To say it differently using the previously discussed framework of “technical correction/biological conservation” trade-off, I favored conservation of biological variability at the expense of technical correction, assuming that technical artefacts can still be figured out at a later point in the analysis, while true biological signal cannot be recovered after having been erased during preprocessing.

Method: Downstream analysis was conducted using Seurat (https://satijalab.org/seurat/) (v3.1.4) in R (v3.5.1). Cells with fewer than 200 features, and features expressed in less than 3 cells, were filtered out (standard default filters in Seurat). Cells with a high proportion of mitochondrial reads (threshold varying between 15 and 30% based on the distribution of mitochondrial reads in each sample) were filtered out. Both sample (WD and DD) count matrices were merged with the function “merge”.

A specific step in the analysis of DDLPS samples is due to the fact that we have each time profiled two samples from the same tumor, so that there are two count matrices per tumor. Since both samples come from the same tumor of the same patient, were resected and processed at the same time under the same experimental conditions, the so-called “batch effect”[136] between different scRNA-seq samples is minimal; this is the reason why I chose to perform all downstream analyses on a “merged” matrix for each patient (using the “merge” function in Seurat) which is simply the concatenation of the two count matrices without any mathematical transformation of the counts to account for a potential batch effect (as opposed to data integration discussed in a next section).

After having filtered out lowly expressed genes and low-quality cells from the count matrix, a usual downstream step is “normalization” to account for variable sequencing depth in each single cell; one common method is the “log-normalization” of counts (essentially a normalization by the total number of reads in each cell and a log-transformation with a pseudocount of 1 to avoid negative infinite values). Another usual step is designed to further reduce the dimension of the matrix, specifically by discarding genes with supposedly no relevant biological information that may increase technical noise in the analysis and overburden the computation. To do this, we retain only the most variable genes – assuming these are the ones that are most likely to contain biological “information” - in the matrix based on their variance across all cells, after correcting for mean-variance bias with a variance-stabilizing transform.

While log-normalization has been the most widely used normalization method from the beginning of scRNA-seq data analysis, it has recently been shown to have limitations[137,138], especially because of the “pseudocount” that is added to the majority of zero counts present in the matrix so as to avoid negative infinity log-normalized values. Indeed, this may lead to skewed results in downstream analyses, notably falsely called differentially expressed genes due to sequencing depth variability. To avoid these pitfalls, multiple alternative methods have been developed, including GLM-PCA[139] and the sctransform[140] method of Seurat. This last method is based on a regularized negative binomial regression of counts and was shown in benchmarking to avoid the main pitfalls of log-normalization; it is currently one of the most widely used methods in the literature. I therefore chose to use this method for normalization, which is implemented in the “sctransform” function of Seurat. This function also selects the 3000 most variable features and scales the data for downstream processing.

Method: Normalization was performed using the Seurat function sctransform (v0.2.0).

We thus end up with a matrix containing normalized counts of 3,000 most variable features for filtered cells (usually between 3,000 and 10,000 per experiment). Though 3,000 is much fewer than the initial > 20,000 features, it is still too large a number for most methods of data analysis, which have difficulty to scale and avoid the curse of dimensionality at such a high number of dimensions[141]. Usually, the matrix is therefore once again reduced, often with a commonly used linear dimensional reduction technique: PCA (described in the first part of my thesis). This allows one to drastically reduce the number of features in the matrix by decomposing the transcriptomic space into major axes of variation that are linear combinations of the initial

features (principal components, PCs). While PCA can calculate as many PCs as there are of features (3,000 in this case), a choice has to be made for the number of PCs to be kept for downstream analysis; keeping too few may erase important biological signal, while retaining too many may in the contrary decrease the signal-to-noise ratio. There are a number of more or less sophisticated methods to choose an "optimal" number of PCs, for instance the "JackStraw" procedure in Seurat which evaluates statistically the significance of each PC using permutations of the data[118]. In contrast, a commonly used and less time-consuming method is the "elbow" rule, which is based on the plot of the percentage of variance explained by each PC ranked in decreasing order of value: the observer chooses the number of PCs at the point of the graph which makes an "elbow", i.e. where the percentage of variance explained by adding more PCs drops down and levels off at near-zero values. This heuristic method is unfortunately subject to bias and variability between users. To keep in line with my previously detailed "philosophical" stance of analysis, I chose to keep the first 50 PCs for each analysis, since this number is generally higher than for the "elbow" method, and therefore minimizes the risk of losing rare populations. Conversely, it is not too large and avoids the accumulation of unwanted noise as well as the curse of dimensionality. Finally, this number may be arbitrary but it is the same in all my analyses; it is also within the range of recommended number of PCs to use with the sctransform method.

The rest of the analysis is thus performed on a reduced matrix of only 50 dimensions, which is well suited to most mathematical tools and avoids the curse of dimensionality. One common and useful step is designed to visualize the global structure of the data: for this we can use non-linear dimensional reduction methods such as t-SNE[53] and UMAP[54] to project the 50-dimensional PCA space into human-friendly two-dimensional representations.

Method: UMAP was performed with the function "RunUMAP" on 50 principal components after "RunPCA".

As an example, the UMAP representation of the scRNA-seq data of a DDLPS patient is plotted in Figure 49, in which each sample origin (WD and DD) is colored differently.

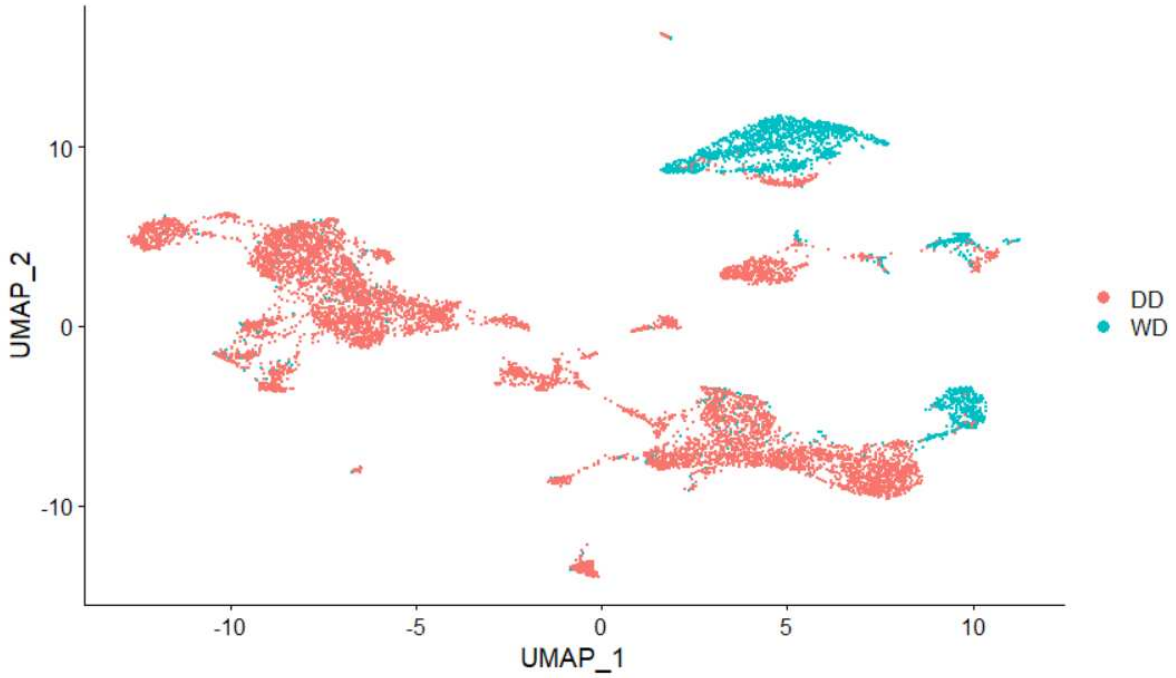

Figure 49: UMAP plot of a DDLPS tumor, colored by sample origin (WD and DD).

Regarding this first UMAP plot we can make several observations: 1) There are multiple distinct "clusters" of cells; 2) the distribution of these clusters between WD and DD samples is different, with some clusters present in both samples, whereas others are predominantly found in only one compartment of the tumor.

To formally segregate cells into so-called "clusters" that should represent groups of cells with a homogeneous transcriptomic profile, many clustering methods can be used for single-cell analysis. The most commonly used is based on so-called "graph clustering", i.e. inferring clusters from the graph of nearest neighbors in the transcriptomic space (which is usually reduced to the PCA space to avoid the curse of dimensionality, i.e. all distances are nearly equal when the number of dimensions of the space is too large). The algorithm used by default in Seurat is the Louvain clustering algorithm[142]. One important parameter when performing clustering is the "resolution": as its name implies it controls the "granularity" of the clusters that are defined by the algorithm. It is in some way related to the height of the horizontal line that can be drawn on a hierarchical clustering plot to determine the number of clusters. The number of clusters is higher (respectively lower) by increasing (respectively decreasing) this parameter; it is a supervised parameter that has to be chosen by the user. To avoid too much bias (there is once again no gold-standard for the "optimal" resolution), I consistently used 0.6 as the resolution value, which is a standard choice in many contexts, and only changed this parameter if downstream biological interpretation of clusters justified either higher (to refine the clustering of some cell populations) or lower (to merge two similar cell populations) resolution.

Method: The nearest-neighbor graph was calculated using the "FindNeighbors" function in the PCA 50-dimensional space. Then clusters were inferred using the "FindClusters" function with resolution 0.6.

The clusters found in the same tumor presented above are plotted in Figure 50.

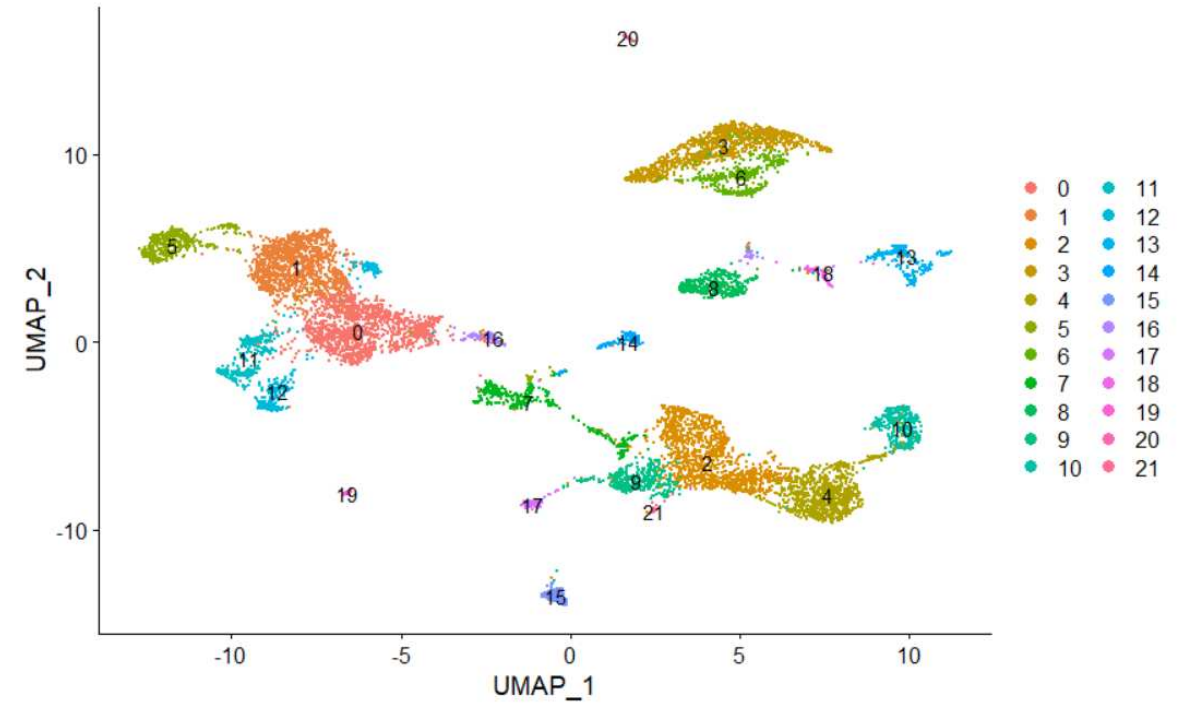

Figure 50: Clusters found in a DDLPS tumor.

Biological interpretation of results

Using this analysis pipeline, there are thus 22 different clusters in this DDLPS tumor. The next step in the analysis is to assign a biological identity to each one of these clusters. For this, there are multiple ways to proceed, and often they are complementary. On one side, one can use prior biological knowledge to "guess" the identity of the different cell populations in the sample: for instance, by plotting on the UMAP the expression of some known "marker" genes of pre-defined cell populations (e.g. *CD3G* and *CD3D* for T lymphocytes), and thus searching for them in a supervised way. On the other side, one can first calculate for each cluster the differentially expressed genes as compared to other cells in the sample, in order to derive a list of "marker genes" that may inform on the identity of the cluster. Finally, one can use automatic methods that have been developed recently[143–145]: they generally take advantage of large reference datasets containing transcriptomic profiles (single-cell or bulk) of known cell populations, and by comparing either marker genes or whole transcriptomes, try to infer the identity of each single cell (or cluster) in the data by finding the "nearest neighbor" in the reference cell populations.

In our case, we can already make good guesses at the cell populations present in the sample: notably tumor cells, which in DDLPS show overexpression of the genes *MDM2* and *CDK4* due to the canonical 12q amplification. Indeed, we rapidly identify their presence on the UMAP by plotting the normalized expression value of these genes (Figure 51). A complementary criterion to confirm the identity of DDLPS tumor cells is their genomic profile, which can also be inferred from scRNA-seq (see Inference of copy number alterations at the single-cell level).

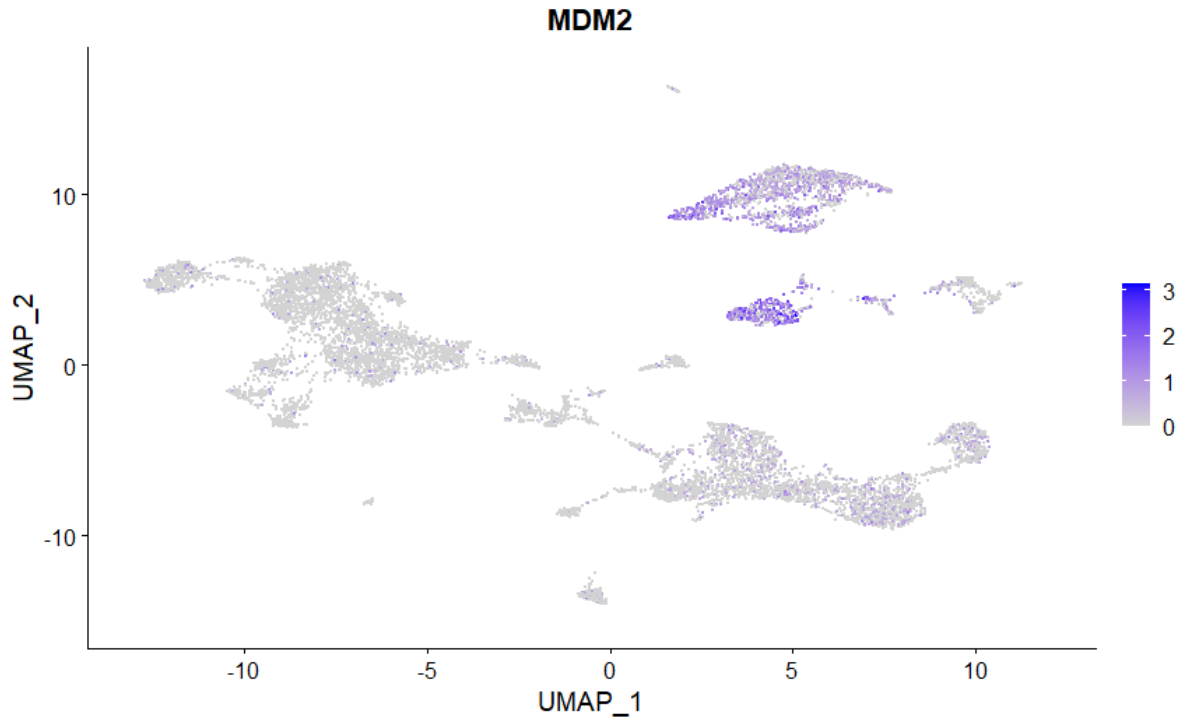

Figure 51: MDM2 expression in the DDLPS tumor.

Besides tumor cells, we also expect cells of the tumor microenvironment – cells of the immune system, endothelial cells, pericytes, etc - to be present in this whole tumor sample, since we did not perform any experimental selection of cells before scRNA-seq (in contrast, many studies focus on specific cell populations by isolating them using e.g. flow cytometry before performing scRNA-seq). Using manual exploration (plotting of marker genes, curation of differentially expressed genes) and helped by an automatic annotation procedure (SingleR[143]), I annotated the different clusters as in Figure 52.

Method: Lists of marker genes (differentially expressed genes versus all other clusters) for each cluster were generated with the function "FindAllMarkers" using the default Wilcoxon's test. SingleR (v0.2.2) was used to calculate the most probable cell identity for each cell with use of Human Primary Cell Atlas bulk RNA-seq as reference data.

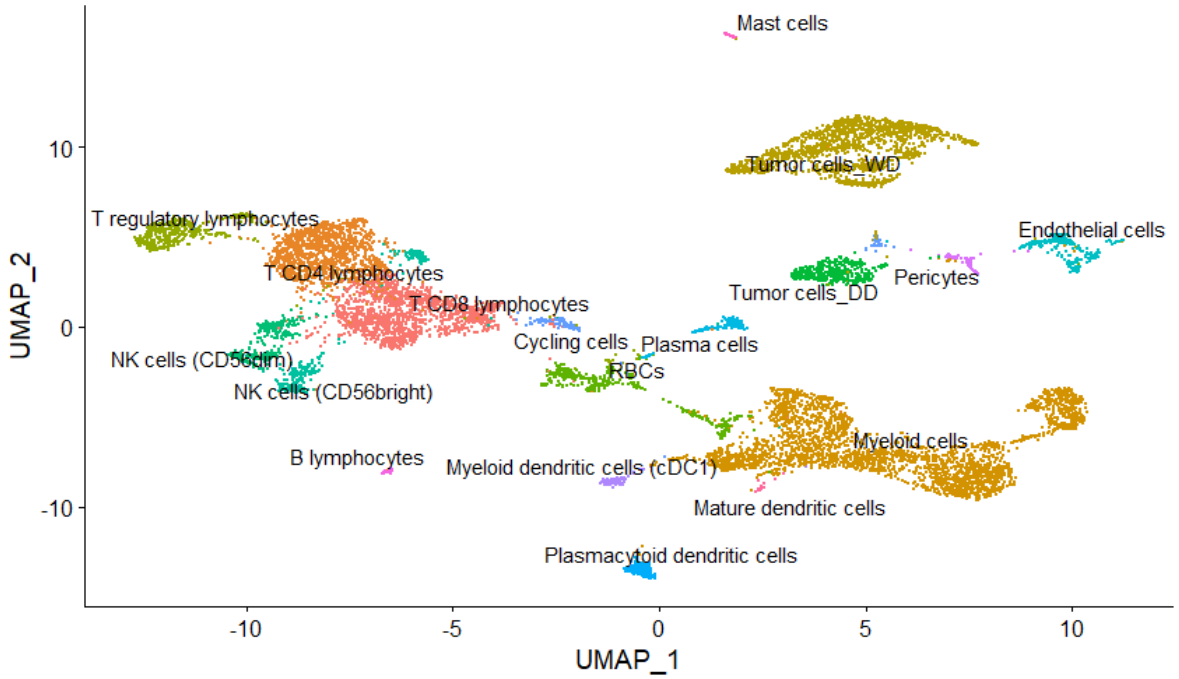

Figure 52: Cluster annotations for the DDLPS tumor. RBCs: red blood cells.

By comparing with Figure 49, we can observe that: 1) tumor cells are separated into two distinct groups of cells (clusters 3-6 on one side and cluster 8 on the other); 2) tumor cells originating from the WD compartment are all part of one of these groups (clusters 3-6), while those from the DD compartment constitute all cells of cluster 8, but can also be present in cluster 6; 3) most cells of the microenvironment, especially cells of the immune system, are for the major part localized in the DD compartment of the tumor.

Inference of copy number alterations at the single-cell level

As mentioned previously, DDLPS tumor cells carry a 12q amplification, and it is possible using dedicated tools to infer computationally a single-cell copy-number profile from gene expression (scRNA-seq) data. Several methods exist, including HoneyBadger[146] which is based on an Hidden Markow Model (HMM), and inferCNV (https://github.com/broadinstitute/inferCNV), which is one of the most commonly used and that I chose to perform. Basically, inferCNV assigns copy-number alterations (such as gains or deletions) by comparing the combined expression of sliding windows of a hundred genes across all chromosomes between cells of interest and "reference" cells that are supposed to have a normal "flat" genomic profile: global overexpression (underexpression respectively) of a gene window leads the algorithm to infer a gain (deletion respectively) at this genomic location. For this tumor, I used T lymphocytes from the same sample as reference cells to infer the copy-number profile of the tumor cell clusters. The result of inferCNV for the same DDLPS tumor is displayed in Figure 53.

Method: inferCNV (v0.8.2) was run on the filtered count matrix with default parameters, using normal stromal cells (T lymphocytes here) as reference cells and tumor clusters as query cells for the algorithm.

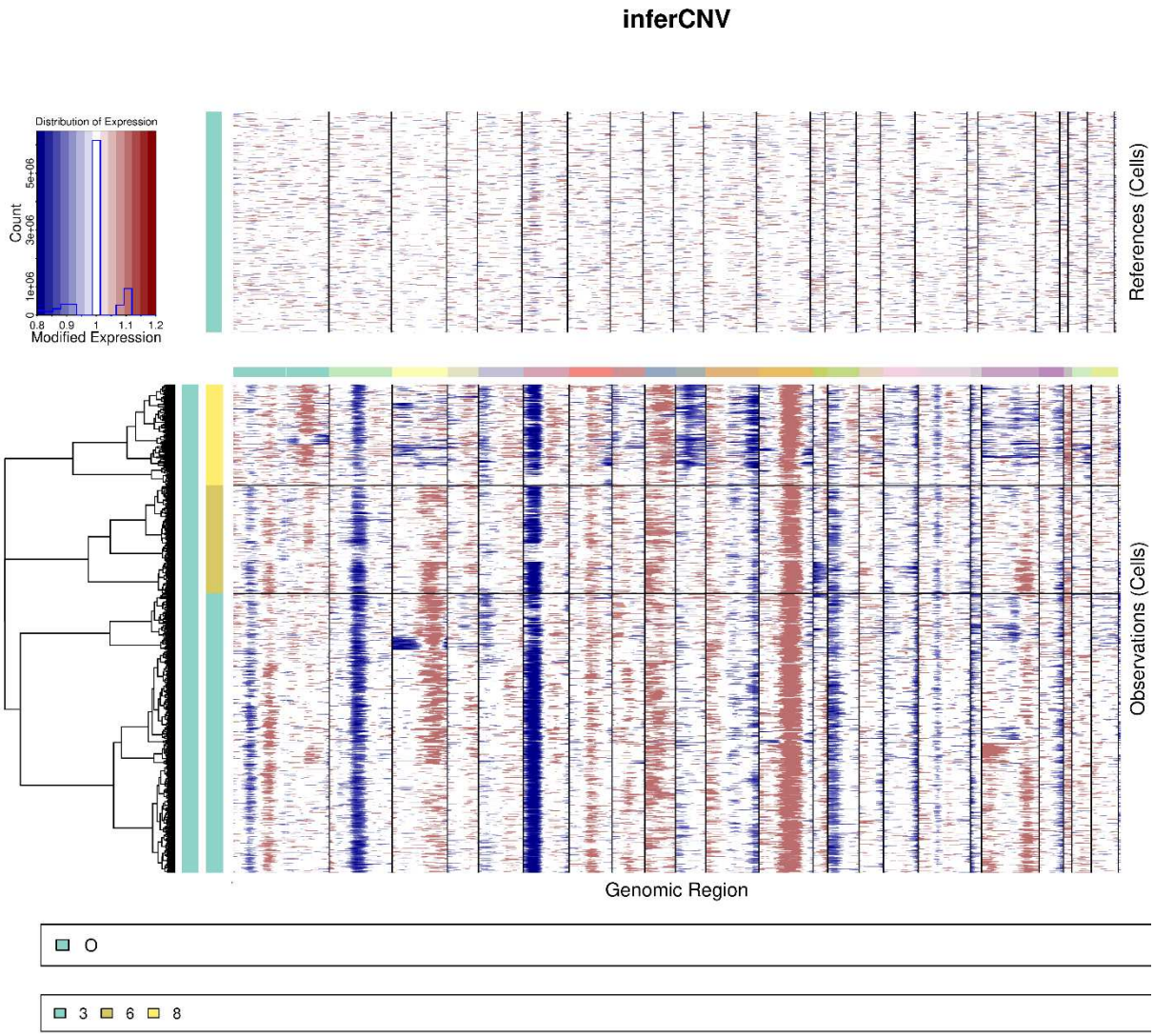

Figure 53: Inferred CNV profile of DDLPS tumor cells. Clusters are numbered as in Figure 50**.** Chromosomes are ordered from 1 to 22, X and M.

From the inferred CNV profile, we can clearly see the 12q amplification that is present in tumor cells (clusters 3, 6, 8). The 6p loss is an artefact due to the overexpression of the HLA locus at this site in reference cells (lymphocytes). We also observe other copy-number alterations that are either shared by all tumor cells (clonal) or private to some cells only (sub-clonal). The main observation here - and in other patients (data not shown) - is that, except for some shared copy-number variations (CNVs) such as the ubiquitous 12q amplification, WD and DD cells (here, clusters 3 and 6 versus cluster 8) generally harbor specific CNVs. This is in favor of the existence of different genomic profiles for these two types of transcriptomically distinct cells, that may thus constitute different clonal populations that have diverged at an early time point from a common precursor cell displaying the driver event of 12q amplification. This would corroborate studies using bulk whole-exome sequencing of WD and DD compartments of the same tumor, which favor a model of early divergence from a common precursor, with specific genetic alterations in each compartment[147–149]. However, we have to keep in mind that CNVs inferred from scRNA-seq may not be true genomic copy-number alterations but rather the reflection of epigenetically coordinated overexpression (or underexpression) of specific genomic regions as compared to the cells used as reference (e.g. the artefactual 6p loss mentioned previously).

Analyses of individual patients

I will not detail in this manuscript all analyses of the four DDLPS tumors profiled for this study, since the main observations were largely similar in all samples. There are invariably two transcriptomically distinct groups of tumor cells, one exclusively composed of cells from the anatomical DD compartment, that I called "DD cells" (in a transcriptomic sense), and one composed of cells both from the anatomical WD (in majority) and also sometimes the DD compartment, which I called "WD cells" (in a transcriptomic sense). The DD cells have expression profiles and marker genes reminiscent of fibroblasts, whereas WD cells display a pre-adipocytic profile (in fact, they cluster closely with normal pre-adipocytes that are found in two of the tumors, the main difference being the absence of the 12q amplification and associated MDM2 overexpression in normal pre-adipocytes). In a couple of samples, we can also observe a little cluster of cells that have a mature adipocytic profile, but do not seem to harbor 12q amplification (though there are too few cells to be confident about the result of inferCNV). This confirms that adipocytes - whether normal or tumoral - are technically difficult to capture by droplet-based scRNA-seq due to their large size; they may however be profiled more easily by single-nucleus RNA-seq after disruption of the cell membrane[150].

Besides these common observations, I noticed an interesting variation in the DD cells of the fourth patient (patient ID 0504583): next to the "classical" fibroblast-like cells, there were two clusters with differing characteristics. One of them expressed desmin (*DES*) - a classical marker of smooth muscle cells - as a top marker gene, while the other expressed in abundance multiple keratins (*KRT*), similarly to a classical keratinocyte (Figure 54: "Tumor cells_DD_DES" and "Tumor cells_DD_KRT"). These two clusters were tumoral as inferred from their CNV profile but displayed some subclonal alterations as compared to the other fibroblast-like cells (Figure 55). This observation is consistent with what is known from the pathology of DDLPS: so-called "dedifferentiated" cells are classically spindle-cell shaped like fibroblasts, however there is sometimes the presence of smooth-muscle-like or keratinizing compartments. This may be due to the fact that "dedifferentiated" cells are probably phenotypically closer to the normal stem cell of adipogenesis and thus display more propensity to acquire alternative differentiation characteristics. Finally, we can also observe in this patient that one of the clusters besides all the other WD tumor cell clusters is in fact non-tumoral, since it does not display the 12q amplification (Figure 55), and corresponds to normal pre-adipocytes, confirming the close proximity of WD tumor cells to normal adipocyte progenitors.

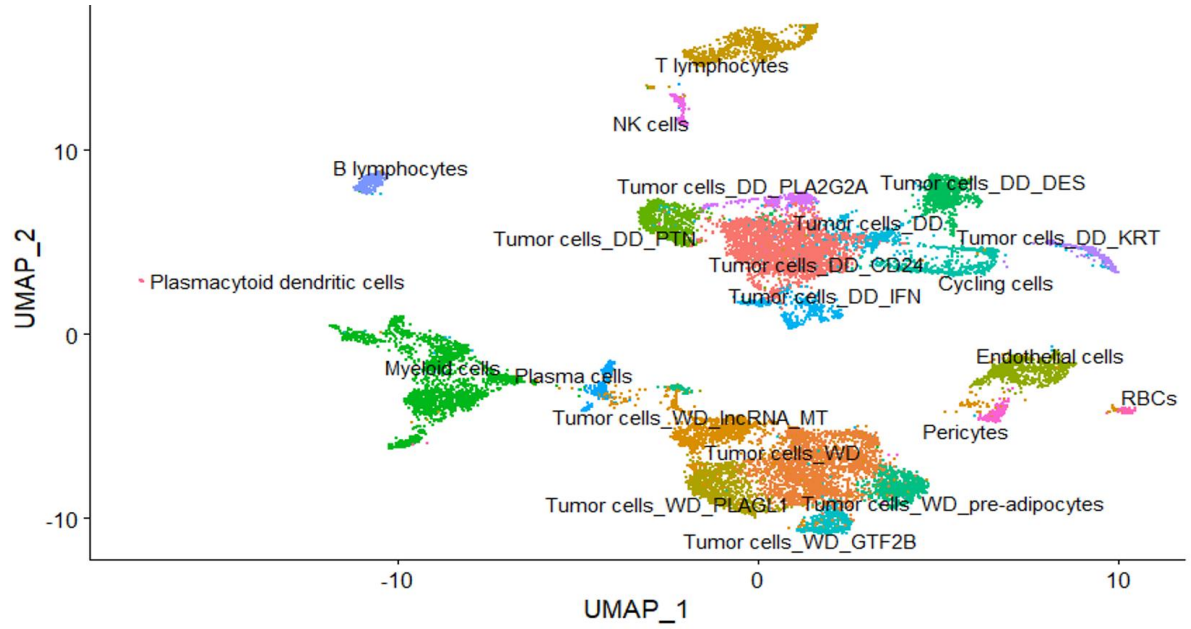

Figure 54: Annotated clusters in another DDLPS tumor. RBCs: red blood cells.

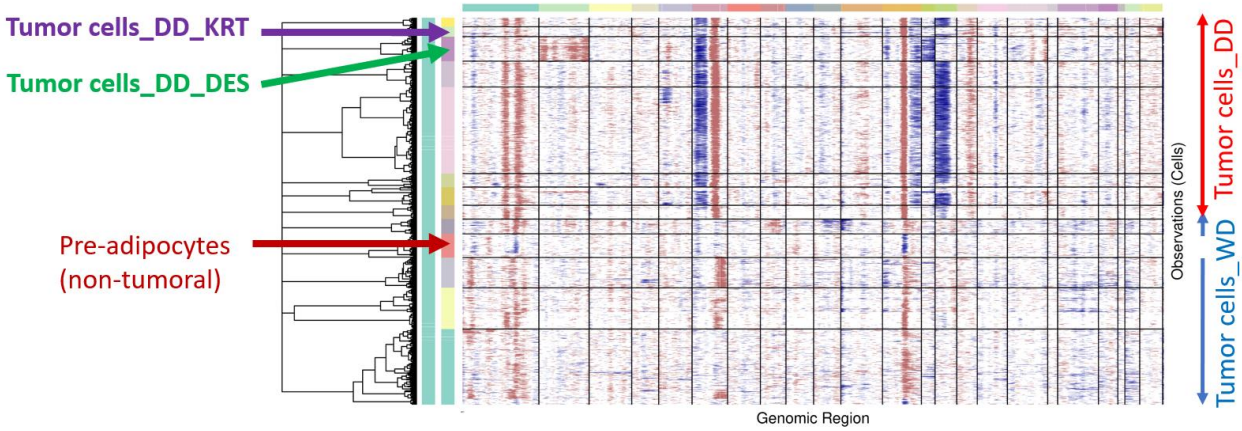

Figure 55: Inferred CNV profiles of the DDLPS tumor cells in Figure 54. Endothelial cells and pericytes were used as reference cells. A cluster within the WD tumor cells is not displaying the 12q amplification and corresponds to non-tumoral pre-adipocytes.

For the tumor microenvironment, all patients show the same trends: cells of the immune system are found predominantly in the DD compartment, and differential expression analysis reveals a more "exhausted" phenotype of T lymphocytes[151], as well as an enrichment in "M2" (anti-inflammatory, pro-tumoral) macrophages[152] in the DD compartment as compared to the WD compartment.

Data integration

All these analyses were performed in individual tumors, however it would be interesting to analyze in a joint manner all patients, in order to facilitate the assessment of common characteristics of all patients as well as inter-patient heterogeneity. This would also increase statistical power by augmenting the number

of cells in common cell populations, potentially even allowing some clusters of rare populations to be defined. This "integration" of scRNA-seq datasets is not a trivial issue and can rarely be performed satisfactorily by a simple "merge" of count matrices (one counter-example is the integration by "merge" of the WD and DD samples from the same patient that I detailed previously, in the virtual absence of "batch effect"). Indeed, scRNA-seq is highly sensitive to the so-called "batch effect", which is a term encompassing all technical covariates that may influence the generation of the final count matrix, independently of real "biological" signal. Multiple methods have been developed to address this data integration challenge[131,153] and evaluate batch correction[154,155] in scRNA-seq; the objective is to correct for technical "batch effect" and conserve biological variation. Once again, we can think of this in terms of the previously discussed general trade-off between "technical correction" and "biological conservation" of signal. However, this issue is much more complicated for scRNA-seq than in classical "batch-correction" for bulk RNA-seq, which can often be addressed with linear models[156]. In scRNA-seq, batch correction requires in most cases non-linear methods, in order to account for technical variation at the level of genes but also of single cells, that depending on their cell type may be more or less affected by technical covariates between experiments. This is the reason why the best-performing methods in the literature are complex non-linear methods[155,157–160] that generally try first to identify common cells between experiments (using for instance "mutual nearest neighbors"[161] in transcriptomic space) and use this information to correct the count matrix in a cell-specific way, in order to align together cells that are inferred to be similar across datasets.

One very important assumption of these methods though is that there exist some common cell types between the datasets to be integrated: indeed they will then try to identify these, before using them as "anchors" (to talk with the terminology of Seurat) in order to align different datasets onto the same transcriptomic space (for instance using canonical correlation analysis, CCA)[162]. In the case of normal cells - which are generally similar from one person to another - these methods perform correctly: e.g. cells of the immune system are commonly used for benchmarking and relatively easily integrated using these methods[155]. However, this assumption is not so obvious in the case of tumor samples. Indeed, tumors from different patients, even if they are from the same tumor type, often contain specific alterations that may confer them patient-specific transcriptomic profiles: in fact, it is generally the case that simply "merging" scRNA-seq datasets of tumors from different patients shows clustering of microenvironment cell populations by cell population (independently of patient origin), while tumor cells cluster separately by patient[123,125] (cf below in DDLPS). These methods of integration can thus reasonably be used only if we assume that at least some of the tumor cells are similar between different patients, so that the method can identify these similar cells and use them to align datasets and enable integrated analysis. If this assumption turns out to be false, there is a risk that the method falsely "aligns" datasets that are not truly similar. Indeed, since the method is defining "common" cells ("anchors" in Seurat) as "nearest" cells - in relative distance - between datasets, these "common" cells may in fact be biologically different (in absolute transcriptomic distance) but still be aligned together since they are - relatively - nearer each other when compared to other cells in the data. In other terms, there is a risk of "over-correction": correction of technical batch effect ("alignment" of datasets) is also erasing true biological variability (grouping different cells in "common"): one has once again to find the right balance between these two factors.

To perform integrated analysis of all DDLPS patients, I decided to use several of the currently best-performing and most widely used methods, in order to compare different approaches and assess the resulting integrations; I assumed that each method had downsides, but using several may allow them to compensate for each other's failings and add to the robustness of results if they were reproducible across methods. Indeed, I was particularly keen at avoiding "over-correction", so as not to erase biological variability that might be present between patients. To this end, I tried first to be very thorough in the exploration of the individual datasets, in order to have good prior knowledge of what "should" be a reasonable integration. To be more specific, a good annotation of the individual datasets allows one to evaluate in a critical way how the integration procedure performs in "gluing" together identical populations

and leaving apart rare and patient-specific populations. This last point is particularly critical to avoid "over-correction", where non-overlapping cell types are "glued" together. This way of proceeding may seem circular because the "reference" is manual annotation of clusters in individual datasets, however since there is no gold-standard of integration especially for tumor cells[163], I thought it was a reasonable way to evaluate integration. In these datasets, I notably focused on some patient-specific clusters: for instance the keratin-high and desmin-high DD cells of the fourth patient, as well as rare cell populations from the microenvironment, such as some neurons and testis cells that were found in one patient (DDLPS_1819409), whose tumor had developed inside the testis.

I compared four methods of data integration: the first one was not strictly speaking "integration" but simple "merging" of datasets, in order to have an idea of the intensity of the "batch effect" between patients, notably between cells of the microenvironment that are not supposed to be very different between patients, as opposed to tumor cells that usually cluster by patient[123,125]. The second method was Harmony[158], a PCA-based approach that learns a simple linear cell-specific adjustment function. The third was the method developed by the Seurat team using identification of "anchors" and alignment of transcriptomic space using CCA[157]. The fourth was reciprocal PCA (RPCA)[164], a method also developed by the Seurat team, that was specifically designed to attenuate the "over-correction" known to occur sometimes with the classical Seurat integration method: in essence it is the same procedure beginning with identification of "anchors", but CCA is replaced by reciprocal PCA that represents a more conservative approach where cells in different biological states are less likely to "align" after integration.

Methods: 1) Merge: All individual filtered count matrices were merged with the function "merge". As "sctransform" has not been validated for the merging of samples, I used the standard normalization workflow in Seurat, i.e. log-normalization with "NormalizeData", selection of most variable features by "FindVariableFeatures" (method "vst", variance-stabilizing transform), scaling of data with "ScaleData". PCA was then applied with "RunPCA", and UMAP was calculated by "RunUMAP" based on 50 PCs. 2) Harmony: As Harmony has not been designed to work with sctransform, I used the standard normalization workflow of Seurat as detailed in 1) and used Harmony (v1.0) on the first 50 PCs of the resulting merged Seurat object. UMAP was then calculated by "RunUMAP" based on 50 Harmony components. 3) Seurat integration with anchors: using the individual Seurat objects preprocessed with "sctransform", integration features were selected with "SelectIntegrationFeatures", preparation of integration was done with "PrepSCTIntegration", anchors were found with "FindIntegrationAnchors" (dims = 1:50, normalization method = "sct"), and data was integrated using "IntegrateData" (dims = 1:50, normalization method = "sct"). PCA was then performed on the integrated matrix with "RunPCA", and UMAP was calculated by "RunUMAP" based on 50 PCs. 4) Reciprocal PCA (RPCA): the workflow is the same as for 3), except for the function "FindIntegrationAnchors" which is run with reduction = "rpca".

To visually assess the different integration results in this manuscript, I will plot below for each method the resulting integrated UMAP split by individual patient and annotated by cluster names of the individual datasets. In this way, one can see if cells from different patients characterized by the same original cluster annotations are localized together in the integrated UMAP. I apologize for the crowded nature of some cluster annotations, but I will try to guide the reader and convey the main messages of each plot in the text.

Figure 56 shows the "merge" of all patients. Unsurprisingly, tumor cells cluster by patient and do not overlap in the UMAP. However, cells of the microenvironment cluster by cell type, which is reassuring for the confidence in the original individual annotations and also a sign that the "batch effect" is probably not very important between these four experiments.

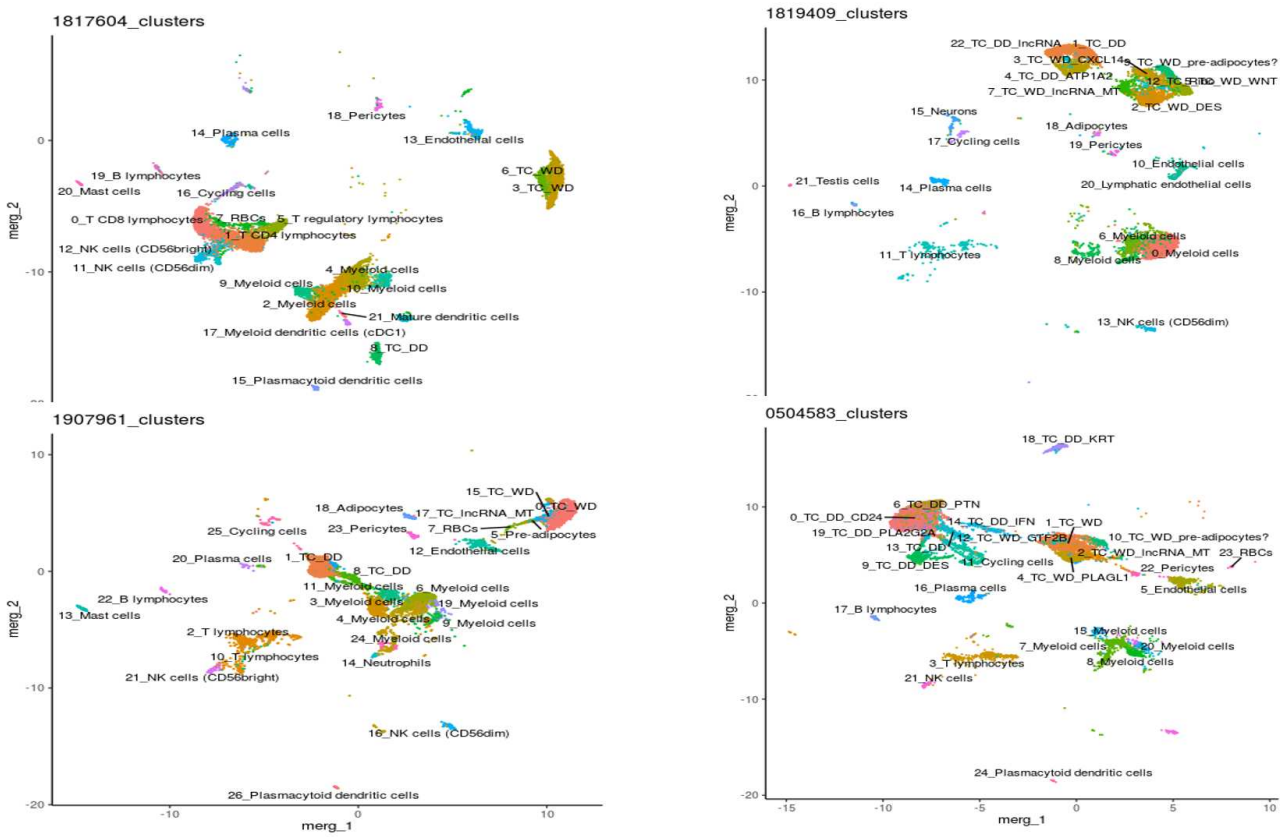

Figure 56: Merge of all four DDLPS patients (Patient IDs: 1817604, 1819409, 1907961, 0504583). Integrated UMAP is split by patient, annotations are for clusters from each individual patient (preceded by their cluster number annotation, a suffix is added to tumor cell clusters that indicates WD/DD classification and marker gene(s)). TC: tumor cells; RBCs: red blood cells.

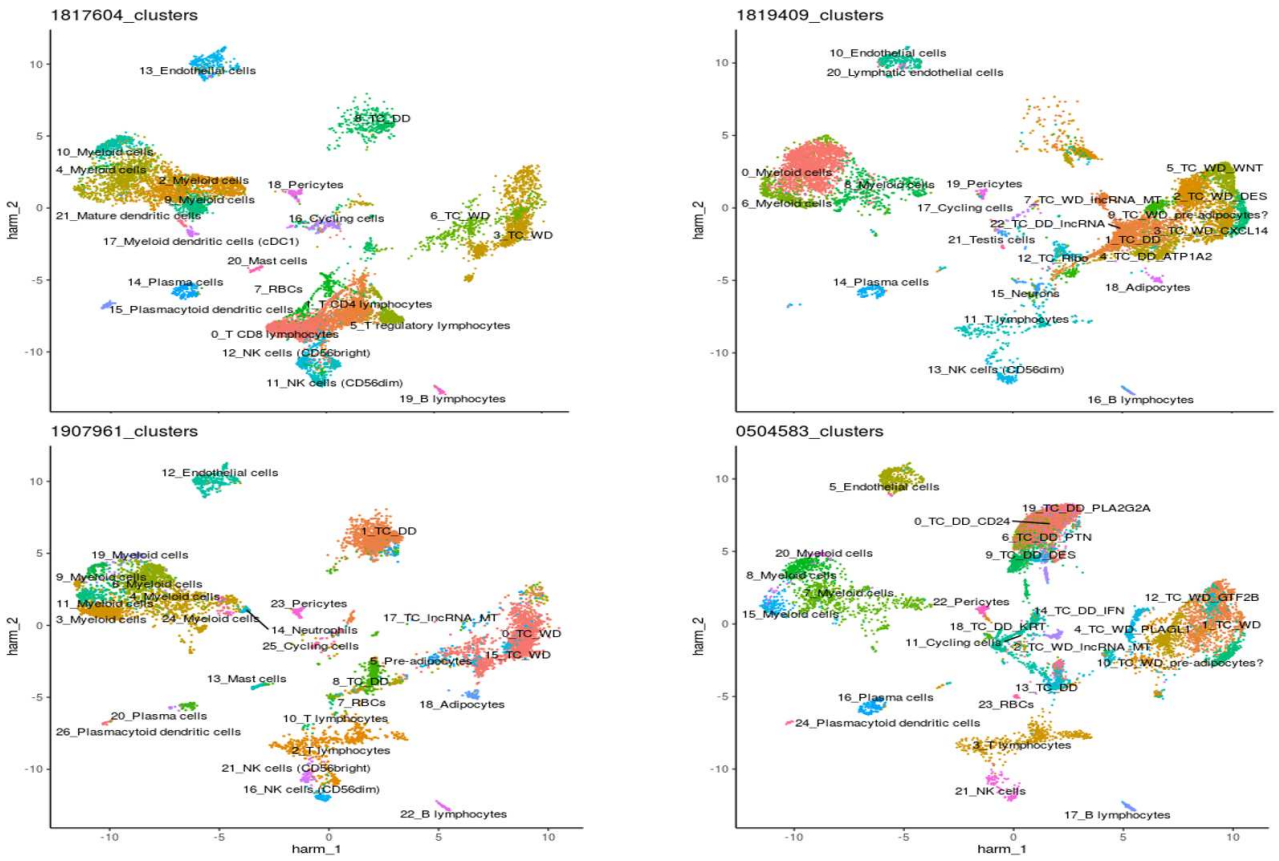

Figure 57: Harmony integration of all four DDLPS patients (Patient IDs: 1817604, 1819409, 1907961, 0504583). Integrated UMAP is split by patient, annotations are for clusters from each individual patient (preceded by their cluster number annotation, a suffix is added to tumor cell clusters that indicates WD/DD classification and marker gene(s)). TC: tumor cells; RBCs: red blood cells.

Figure 57 shows the integration by Harmony of all patients: tumor cells are overlapping and do not cluster by patient anymore, while cells of the microenvironment also cluster by cell type. Overall, the integration seems to have been performed correctly by the algorithm. However, close observation suggests some caveats: 1) Rare cell types from the second patient (1819409, top-right) including neurons (cluster 15) and testis cells (cluster 21) do not form distinct clusters and overlap with other unrelated cell types; 2) In this same patient (1819409), the DD tumor cell (TC_DD) clusters (clusters 1, 4, 22) do not cluster with the TC_DD clusters of the other three patients (that together form a distinct TC_DD cluster in the top-middle of the UMAP) but are closer to the center of the UMAP, near the WD tumor cell clusters (TC_WD); 3) There seems to be an intriguing phenomenon of "centripetal attraction" that I also noticed when using Harmony with other datasets (data not shown): cells that are *a priori* more "difficult" to integrate by the algorithm tend to be plotted in the middle, for instance here neurons and testis cells that are rare populations only present in one patient. A consequence of this phenomenon is the appearance in the UMAP of central connections ("bridges") between completely different cell populations, for instance here between tumor cells and T lymphocytes. Other clusters that seem "difficult" to integrate and get "attracted" to the center of the UMAP inside this "middle cloud" are some tumor cell clusters (cluster 8_TC_DD in patient 1907961, cluster 13_TC_DD in patient 0504583) and notably the rare cluster of keratin-high DD cells in the fourth patient 0504583 (cluster 18_TC_DD_KRT).

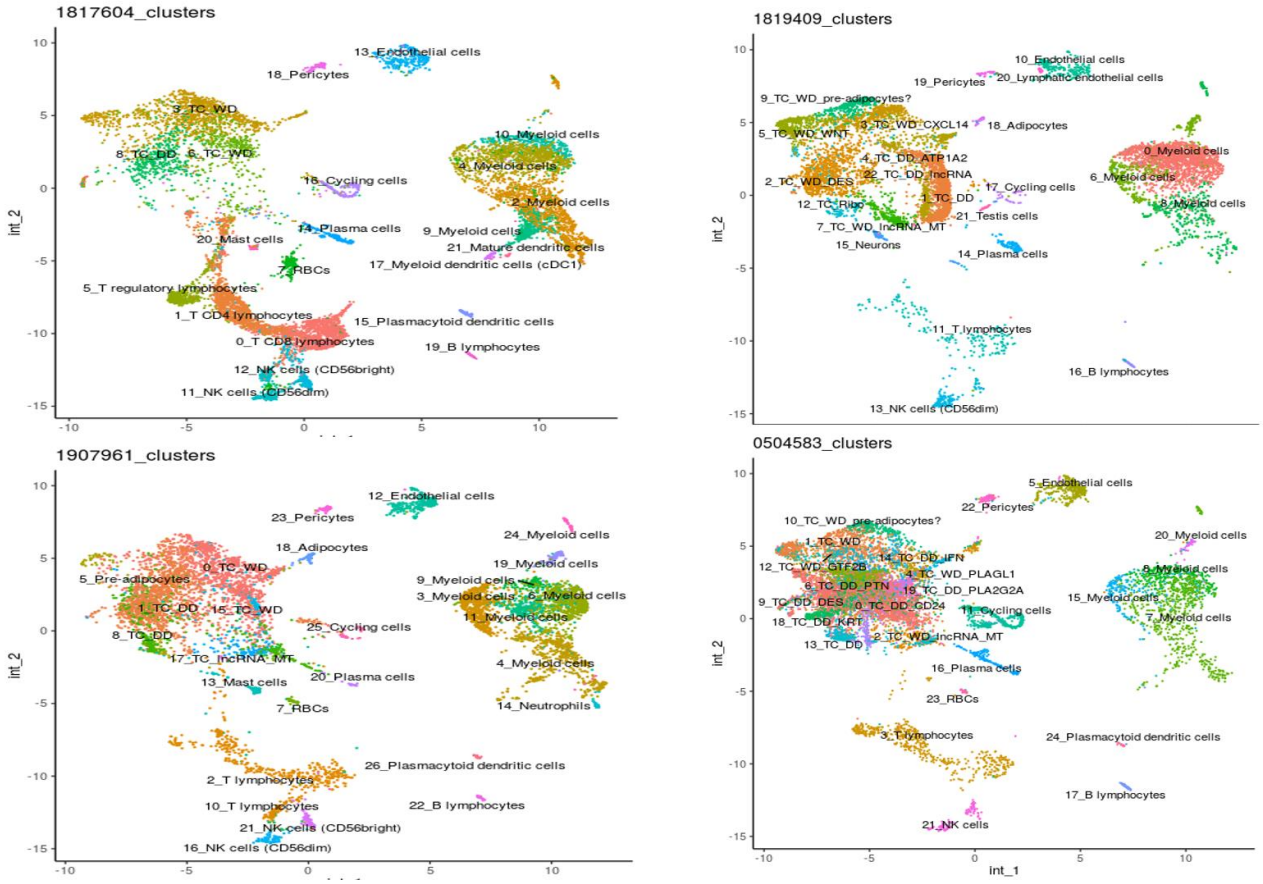

Figure 58: Seurat integration with anchors of all four DDLPS patients (Patient IDs: 1817604, 1819409, 1907961, 0504583). Integrated UMAP is split by patient, annotations are for clusters from each individual patient (preceded by their cluster number annotation, a suffix is added to tumor cell clusters that indicates WD/DD classification and marker gene(s)). TC: tumor cells; RBCs: red blood cells.

Figure 58 shows the integration using the Seurat method with anchors and CCA. Once again, the integration seems to be coherent with overall overlap of tumor cells and microenvironment by cell type. In comparison to Harmony: 1) Rare cell types in the second patient 1819409 - neurons and testis cells - are this time localized distinctly from the other cell types and are accurately represented as patient-specific clusters. This is a good indication that for this dataset Seurat seems to be better than Harmony in the conservation of biological information as opposed to technical "over-correction"; 2) The intriguing observation in Harmony about DD cells of the second patient 1819409 (they were clustering separately from other patients and were closer to WD cells) is not reproduced here. Using this method, all tumor cells cluster together at the left of the UMAP, with an approximate "gradient" from WD (top) to DD (bottom) that seems to be similar in all patients. My impression though is that tumor cells are more tightly clumped together and intra-tumoral heterogeneity - notably between WD and DD clusters - is less clear than in Harmony. This may be a sign of "over-correction", considering also that the two patient-specific clusters (18_TC_DD_keratin-high and 9_TC_DD_desmin-high) in the fourth patient 0504583 are not clearly distinct from other tumor cells; 3) There is no more "centripetal attraction" of some "difficult"-to-classify cells to the center of the UMAP, so there is no intriguing "bridge" between tumor cells and lymphocytes as in Harmony.

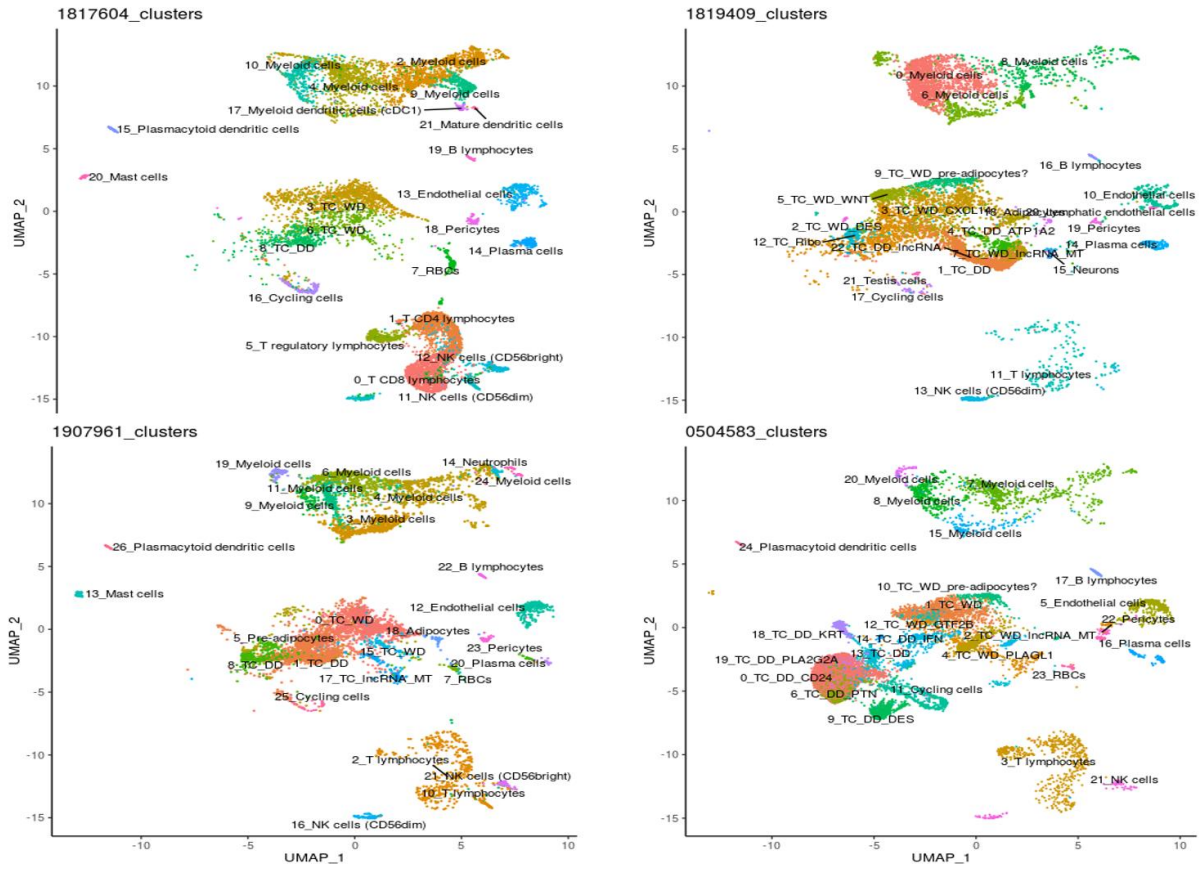

Figure 59: Integration using reciprocal PCA (Seurat RPCA) of all four DDLPS patients (Patient IDs: 1817604, 1819409, 1907961, 0504583). Integrated UMAP is split by patient, annotations are for clusters from each individual patient (preceded by their cluster number annotation, a suffix is added to tumor cell clusters that indicates WD/DD classification and marker gene(s)). TC: tumor cells; RBCs: red blood cells.

Finally, Figure 59 shows the integration using the reciprocal PCA method (RPCA) of Seurat. The overall picture is satisfying as well, with clustering of tumor cells and microenvironment by cell type. In comparison to the previous methods: 1) Rare cell types (neurons and testis cells from the second patient) form distinct clusters as in the Seurat integration with CCA; 2) Tumor cells are more clearly separated into WD cells (top-right) and DD cells (bottom-left); there is also evidence of patient heterogeneity, notably DD cells from the second patient 1819409 are more to the right of the UMAP, while those from the fourth patient 0504583 are at the extreme bottom-left. Interestingly, the two patient-specific DD clusters of patient 0504583 (18_TC_DD_keratin-high and 9_TC_DD_desmin-high) form "satellite clusters" clearly distinct from the other DD cells in this integration. 3) There is still no "centripetal attraction" nor "bridge" between tumor cells and lymphocytes in contrast to Harmony.

Overall, the main conclusions from this comparison of integration methods are: 1) Simple "merge" is sufficient to correctly cluster the cells of the microenvironment by cell type, however all tumor cells cluster by patient if no further computational integration is performed. One could argue that this method is the closest to "biological reality", with conservation of "real" inter-patient tumoral heterogeneity. We cannot exclude some "batch effect" nonetheless, and since all patients *a priori* display the same disease, we can suppose that computational integration of tumor cells is still possible and justified. 2) The three computational integration methods (Harmony, Seurat CCA and Seurat RPCA) perform well overall to cluster tumor and microenvironment cells together by cell type: differences between methods are rather subtle in fact if we only concentrate on the big picture and particularly the main cell populations of the microenvironment. 3) For finer detail though, Harmony seems to struggle for some cell types, notably rare

and patient-specific cell populations. It also shows a tendency to accumulate these “difficult”-to-integrate cells in the middle of the UMAP, thus creating “bridges” or a “middle cloud” between non-similar cell types. 4) In contrast, both methods from Seurat overcome this difficulty: they conserve rare cell types as distinct clusters and do not create a “bridge” between non-similar cell types. However, Seurat integration with anchors and CCA tends to clump together all tumor cells, whereas RPCA manages to conserve a degree of inter-patient tumoral heterogeneity which is consistent with the analyses of individual samples, notably for patient-specific tumor clusters.

Altogether, while this comparison of integration methods is surely not exhaustive in terms of methods and parameters - the definition of the “best” integration method may moreover vary with other evaluation criteria -, it seems that Seurat RPCA is in this case the best method to perform integration in terms of the trade-off between “technical over-correction” (present for rare cell types in Harmony and tumor cells in Seurat CCA) and “biological conservation” of variability (“merge” also conserves biological variability but retains technical batch effect). More generally, I think that whichever integration method is used, this balance between batch effect correction and conservation of biological variability should always be weighed with caution. This is complicated by the fact that there is no gold-standard in cancer studies to correctly rate the “best” integration for tumor cells, since it is not generally known what is supposed to be the real “biological” heterogeneity between different tumors, in contrast to simple “technical” heterogeneity of batch effect. One strategy to alleviate this issue would be to minimize the batch effect by experimental design, for instance by processing all patients at the same time; however this is not possible if we are profiling fresh tumor tissue coming directly from the operating room (freezing samples and processing them altogether later in time could be an alternative though).

In practice, the strategy that I would recommend to perform computational integration of tumor cells is to first thoroughly analyze individually each patient, in order to have a good *a priori* idea of the similarities and specificities between patients. If no similarities can be found between tumor cells of different patients at this stage, it would probably be better not to proceed with integration, since as discussed previously all integration methods rest on the assumption that common cells exist between datasets, so the results could be misleading by “forcing” alignment of unrelated tumor cells of different patients. If integration is considered possible and justified, I would recommend trying at least two methods in order to evaluate the consistency of the results. Indeed, one observation that is not reproduced by other integration methods should be subject to caution (cf the DD cells of the second patient 1819409 that were clustering near WD cells when using Harmony: due to some specific characteristics they were probably “difficult” to integrate and placed in the middle of the UMAP by Harmony). To compare between different integration results, it is important to focus attention on rare populations or patient-specific clusters that are more subject to technical over-correction. In the end, the “best” integration should be coherent with the analyses of individual patients and conserve biological specificities such as rare populations or patient-specific clusters. Of course, one could also favor one integration method based on the objective of the study: if it is only designed to find global similarities between patients (such as between DD and WD cells of different patients in DDLPS), a method which slightly “over-corrects” will not necessarily be a problem. However, if the objective is to study in detail inter-patient heterogeneity and find subtle differences or patient-specific clusters, using a method that better conserves biological variation may be the best solution.

In the end, the comparison of integration methods also reminds us that all these methods are computational transformations of the data and may show great variability according to parameters and datasets. Therefore, one should not make hasty conclusions based on one unique analysis. For instance here, the integration by Harmony had localized DD cells from the second patient next to WD cells: without performing the other integrations - or knowing from the individual analysis that DD cells from this patient were clearly not similar to WD cells -, we could have falsely concluded that DD cells from this patient were in fact WD cells, and that this patient was not a “true” DDLPS (it was of course a DDLPS, as confirmed by pathology). The reason for this confusion by Harmony seems to be that DD cells of the second patient are

slightly different from the other patients (some marker genes are indeed specific to this patient) and thus should be integrated in a specific location – as done by Seurat RPCA -, however Harmony struggles to conserve this patient specificity and - as for other rare populations - instead localizes these cells in the middle of the UMAP, next to the WD cells in this case.

Immune microenvironment

For the study of the immune microenvironment of DDLPS, integration allows joint analysis of all patients, but as shown previously a simple "merge" could have been sufficient for this, and in the end the main observations are similar to the analyses of individual patients. Altogether, scRNA-seq has allowed deciphering of the composition of the different cell populations present in the tumor, including differing proportions and differential gene expression between the WD and DD compartments. The main observation is that most cells of the immune system are preferentially infiltrating the DD compartment. Also, T lymphocytes and macrophages display more exhausted and immunosuppressive phenotypes, respectively, in this compartment. This is interesting since it may reflect modified crosstalk between cells of the immune system and DD tumor cells, as compared to WD cells. On one hand DD cells may attract more immune cells through chemokine and cytokine signaling, or by increased neoantigen burden due to the presence of more genetic alterations. On the other hand, DD cells may also be able to "blunt" the immune system by inducing exhaustion of T lymphocytes and conversion of macrophages towards more immunosuppressive phenotypes. Since it is the DD compartment that represents the aggressive side of the tumor and is monitored for treatment response in clinical trials, it is likely that the responses seen with immunotherapy in some DDLPS patients[12] were related to the presence of this immune infiltrate which may have been reactivated by the immune checkpoint inhibitor, despite pre-existing exhaustion. However, this is mainly speculation and precise immune correlates of response to immunotherapy in DDLPS are unfortunately not possible to deduce from these experiments since these patients were not exposed to immunotherapy. There are nevertheless a number of studies in the literature that have started exploring this fascinating question of identifying the immune markers that will predict response to immunotherapy by using scRNA-seq (often with associated single-cell TCR sequencing) of tumors[124,165–168].

Relationship between WD and DD cells

During my PhD, I performed many other analyses of scRNA-seq data on DDLPS samples as well as other types of sarcomas, such as undifferentiated pleomorphic sarcoma, desmoplastic small round cell tumor, and Ewing sarcoma. However, I will not detail all of these in this manuscript, but instead focus on some complementary interesting analyses that were performed in DDLPS, notably to try to answer the initial question of the relationship between the WD and DD compartments of this tumor.

As shown previously, we can identify two different types of tumor cells in each patient, that I named WD and DD cells, but in a transcriptomic sense. This must be differentiated from the terms "WD" and "DD" used to define the anatomical compartments that are defined macroscopically by the pathologist and that are used to identify the origin of the two samples profiled by scRNA-seq. We observe in all our patients that: the DD anatomical compartment is composed of transcriptomically DD cells in majority but can also contain some transcriptomically WD cells, while the WD anatomical compartment only contains transcriptomically WD cells. If we correlate this with the clinical observation that the DD compartment invariably appears on a background of WD tumor and often forms "nodules" inside the WD compartment, we may propose the following conceptual model to explain this: the transcriptomically WD cells may be the earliest ones present

in the tumor and form the background of the whole tumor, while transcriptomically DD cells appear only afterwards in localizations which are already occupied by WD cells. Therefore, a pathologist that cuts tissue from the WD compartment will retrieve only transcriptomically WD cells in the sample, while cutting tissue from the DD compartment may retrieve not only transcriptomically DD cells but also "background" WD cells.

Concerning the relationship of transcriptomically WD and DD cells, apart from their differing marker genes and genomic profiles as shown previously, we also asked the question of the potential existence of a "trajectory" between these two types of cells, since the pre-existing clinical hypothesis is that DD cells are the result of a dedifferentiation of WD cells. This type of "trajectory" analysis is particularly suited for analyzing the cellular states of normal development, where one stem cell gives rise to progressively more differentiated cell types and can be represented by a continuous spectrum of transcriptomic change across "trajectories" between cell types visible on a UMAP. Many methods[169–173] have taken advantage of scRNA-seq to infer "trajectories" between different cell states or clusters inside a single-timepoint sample, without any temporal information about the system. To achieve this, these methods make the following assumptions: cells that are currently differentiating or traveling along a trajectory will be sampled at all steps of this trajectory and will display a continuum of gene expression profile between the start and end states of this trajectory. They then apply mathematical constructions such as graphs, principal curves or trees to infer the trajectories in transcriptomic space (which is often reduced to the PCA or even UMAP space). However, these methods only allow the inference of potential trajectories between different cell states or clusters, without giving any information on the direction of these trajectories. In other words, we cannot deduce from them the start and end points of the trajectory. To achieve this, specific methods use the ratio of spliced and unspliced transcripts in scRNA-seq to infer the near-future expression state of each gene in each cell: either increasing expression (higher number of unspliced reads) or decreasing expression (higher number of spliced reads). Based on this information and considering the differential expression of genes along the trajectory, these methods can then infer the direction of the trajectory. This concept is termed "RNA velocity", the main methods for its computation are Velocyto[174] and scVelo[175].

When implementing these methods for our DDLPS samples, I did not find any trajectory between DD and WD cells, and RNA velocity was also uninformative. This was in fact already expected, as WD and DD cells invariably form distinct clusters in the UMAPs of all patients: since trajectories are inferred from transcriptomic proximity, it is unsurprising that the algorithms do not find trajectories between clusters that are localized separately in a UMAP.

Discussion

Altogether, these analyses tend to favor a model of DDLPS in which it is formed of two different clonal populations of the same disease: they both share the 12q amplification, but have probably diverged early from a common precursor, with resulting differences in genomic and transcriptomic profiles, and no active "trajectory" between these two types of cells. While this does not rule out that a transcriptomically WD cell may be the cell-of-origin of a clone that subsequently gives rise to the DD cells (dedifferentiation model), we are tempted to favor a model inspired from hematologic malignancies such as chronic myeloid leukemia (CML)[176]. In this tumoral proliferation of mature myeloid cells, which all display the Philadelphia chromosome encoding the *BCR-ABL1* fusion gene, all cells retain the ability to differentiate into mature myeloid cells. However, during "acutization" of the leukemia, one of these cells loses the ability to differentiate properly and gives rise to a less differentiated tumor cell with enhanced proliferation rate, resulting in the transformation to a much more aggressive disease, acute myeloid leukemia. We hypothesize that in pre-existing well-differentiated liposarcoma, WD tumoral cells retain the ability to differentiate into tumoral adipocytes (we indeed find in our data that they display a transcriptomic profile close to pre-adipocytes or adipocyte normal progenitors, and pathology shows large tumoral adipocytes).

However, at one point one of these WD cells loses this differentiation capacity and gives rise to a DD clone which is undifferentiated and much more aggressive, resulting in the appearance of dedifferentiated liposarcoma that thus contains two genomically divergent clones.

We do not know the specific molecular events that are responsible for this divergence between WD and DD cells, but some clues emerge from scRNA-seq: notably, using a method to infer gene regulatory networks known as "regulons" in scRNA-seq data[177], we have been able to show that some important transcription factors for WD cells as opposed to DD cells are *MYC*, *FOS* and *JUN*. Experiments are currently undergoing to try to demonstrate the functional role of these genes in DDLPS. In parallel, a manuscript is being prepared for submission to a peer-reviewed journal in the next weeks.

General conclusion :

Dear reader, thank you for having read this manuscript to the end! I hope that the picture that I drew from my thesis was not too heterogeneous and that you still managed to follow the underlying threads of all this work.

I have already discussed most of the results inside each of the dedicated parts of this manuscript, however I would like to finish here with some more general comments and perspectives.

Unfortunately for a clinician like me, the main biological results of this work are still descriptive and do not impact clinical practice at least in the short term. For instance, the analysis of the transcriptomic landscape of sarcomas is interesting but only paints a broad picture of the biological processes at play in sarcoma. Classification was not overhauled by RNA-seq, diagnosis prediction was effective but difficult in many cases. I could also have analyzed in more depth the transcriptomic changes of specific types of sarcomas.

However, I believe that molecular high-throughput assays have the potential to guide and significantly improve the clinical management of cancer patients in the near future, especially with the development of techniques such as machine learning to analyze them. The work presented here on classification of cancers of unknown primary is an example of this, where the characterization by RNA-seq allows to correctly predict the tissue of origin in most cases. While this is a small study that does not prove a gain in overall survival for patients, I expect this kind of tool to be more broadly used in the clinic as soon as they are shown to be effective in larger studies. Indeed, there are already some molecular assays that can be used in practice, for instance to guide the clinician in the prescription of adjuvant chemotherapy for localized breast cancer[178]. Conceptually, molecular data can simply be considered as an additional piece of information available to the clinician to help in the elaboration of a diagnosis, prognosis, or therapeutic strategy. While some of these techniques may still be costly and unavailable in some places, we can make a parallel with the use in medicine of - what are now considered standard - biological, pathological, or radiological exams. In their early days, these were also considered as highly advanced and costly technologies. However they were broadly adopted as soon as they showed their significant impact on the clinical management of patients, since this stimulated the investment in technologies to allow their high-scale deployment and decrease in cost[179,180]. Concerning molecular assays, I have no doubt that they allow us to gain precious information about specific characteristics of a patient's cancer which cannot be captured by standard biological, pathological or radiological exams of the tumor. The main difficulty with high-throughput assays is the very large amount of information delivered, which is enormous and cannot be interpreted as such by a bioinformatician, let alone a clinician. There are so many dimensions in the data that it is difficult to "make sense" of what is the important signal, and what is to be considered as "noise". The main challenge is thus to "translate" high-dimensional molecular data into human-understandable information that can guide the clinician. Fortunately, the disciplines of statistics and specifically machine learning have progressed rapidly in recent years and continue to deliver at a high pace novel ways of analyzing this high-dimensional data to extract from it the information relevant to the clinician. Moreover, the development of analysis methods will undoubtedly be facilitated by the positive spiral of increasing use of these assays and demonstration of their clinical benefit. This could indeed lead to the constitution of larger databases of molecular data that would not only allow studies to better understand molecular correlates of cancer diagnosis, prognosis and treatment efficacy, but also contribute enough training data to refine the machine learning algorithms used to analyze them. In the end, I expect that this piece of information could be used in complement with the other data (clinical, biological, pathological, radiological) to determine a precise diagnosis, estimate the prognosis and define the best therapeutic strategy. Since cancer is inherently a complex disease with specific characteristics for each patient, I believe the concept of "precision medicine"[181] is justified and molecular assays are bound to contribute to this personalized characterization of cancer. Finally, the continued improvement in technology could also lead to the use of more precise and sensitive assays at a

lower cost and using less tissue material; this could then allow for instance the better characterization of cancer samples in space (different locations in the same patient), time (before, during and after treatment) and in multiple modalities. In the case of sarcomas, one potential way of refining classification would be for example to integrate different modalities that have already proved of value individually, such as RNA-seq and DNA methylation[48].

To come back to my work, the analysis of the immune microenvironment was broad and only descriptive across sarcomas. The main conclusions were expected: paucity of immune infiltration with only some outliers. It is difficult to draw conclusions on the reasons behind higher immune infiltration in some subtypes without further experiments. Analysis by single-cell RNA-seq is much more precise but still quite descriptive and cannot be extended to large numbers of patients as bulk RNA-seq. From a clinical point of view, it is difficult to infer correlates of response to immunotherapy since none of the patients analyzed either by bulk or single-cell RNA-seq were treated by immunotherapy. One can only hope that future studies will allow to answer to these questions more precisely.

However, in this case also I expect that significant advances in the near future could rapidly lead to more direct impact on cancer patients. While immunotherapy is already well integrated in clinical practice and now represents one of the main weapons of the medical oncologist[11], as well as the bulk of current therapies in human trials[182], the main issue remains that only a subset of patients will respond to it, and it is still very difficult to predict which patient will at the individual level[26]. This is one of the most pressing clinical challenges currently in oncology, because these therapies are highly effective in some cases but will lead to ineffective treatment, delaying of effective therapy or even "hyperprogression" in some patients[183]. Based on the mechanical understanding of immunotherapy, it seems that the correlates of response should be found within factors such as the specific therapy used, the characteristics of the tumor, as well as the status of the immune system of the patient[184], which can depend on both host and environmental factors such as gut microbiota[185]. An algorithm to predict response to immunotherapy would probably rely on a combination of all of these. The aspect that was more specifically addressed in my thesis was the immune system, i.e. the presence and abundance of specific types of immune cells such as lymphocytes in the tumor microenvironment. However, we expect and know that there is no simple relationship between this and response to immunotherapy: high infiltration by lymphocytes for instance may be correlated to higher response rates, but it is not a perfect correlation and this measure cannot be used as a reliable predictor of response in the clinic[186]. Some related measures of immune infiltration have been proposed as candidates such as the pathologically measured "Immunoscore"[187,188], while other biomarkers have taken advantage of some potentially immunogenic characteristics of tumors such as mismatch repair deficiency[189], high tumor mutational burden[190,191] or PD-L1 expression level[192]. One aspect though that is the most interesting to me is the focus on the specific antigens that are recognized by the immune system, since this opens a window into the mechanistic understanding of the response to immunotherapy, which is currently in most cases only limited to statistical concepts such as "high tumor mutational burden". Indeed, it is possible that response to immunotherapy is in a large way dependent on some specific mechanistic event, such as the "right" antigen(s) being recognized by the "right" immune cell in the "right" immune microenvironment. This could account for the intrinsic uncertainty of relying on statistical "bulk" measures such as tumor mutational burden or global lymphocyte infiltration to predict response to immunotherapy. However, this deep mechanical understanding at the antigenic level is still very imprecise. Even if we can accurately determine the sequences of TCRs and BCRs, it is exquisitely difficult to know which antigens they recognize[193]. Using my study on MiXCR as an example, it is very interesting to see clonal expansions of some TCRs or BCRs in some samples, however we cannot know what is the corresponding antigen that has driven their expansion. We can expect though that this understanding could probably help us in deciphering if, how, and why an immunotherapy is going to work. While this is still an open problem, there are signs that it is not intractable: high-throughput methods to characterize TCRs and BCRs on one side[194], and antigens on the other side, are allowing the accumulation of data to help in figuring out and predicting

correspondences between antigens and BCRs/TCRs[195,196]. The latest advances in predictions of protein structures achieved by deep learning (such as "Alphafold"[197]) also point to the possibility of using three-dimensional information to better understand this complex problem of the recognition of an antigen (in the form of a peptide/MHC complex) by a TCR or BCR. While in hematology the therapeutic principle of using immune cells to target an antigen expressed by tumor cells (such as CD19) has been demonstrated to be highly effective (for instance using modified T lymphocytes such as CAR-T cells[198,199]), it is realistic to assume that, in solid tumors also, characterizing the correct "target" antigens for each cancer, but in a personalized way[69,200], could lead to the design of a corresponding immunotherapy such as a vaccine, CAR-T cell or bispecific antibody. In addition to the design of these specific immunotherapies, better understanding of the immune response would also allow the enhancement of existing unspecific immunotherapies such as immune checkpoint inhibitors. First, we could probably better predict response with information about neoantigens present in the tumor, lymphocyte repertoire and immune status of the patient, gut microbiota or other features. For non-responding patients, interventions could then be tailored to either stimulate the expression of neoantigens, generate immune cells recognizing them, and/or boosting those that are present but held in check by some immune checkpoint. In the end, these conceptual considerations show the necessary limits of the present study that only reports global measures of immune cell abundance or measures of clone without other information. Further insights will be needed to understand and improve the clinical response of patients to immunotherapy, but these goals seem possible to attain in the near future.

The characterization of neoantigens was suitably the subject matter of another part of my work. The discovery of novel transcripts driven by oncogenic chimeric transcription factors was totally unexpected and highlights the potential phenomena that may still be "hidden" inside NGS data. Of course, we are far from having proved that these are sources of neoantigens that could be targeted with success using immunotherapies, however it is a very promising avenue of research and opens perspectives for other cancers that may also harbor novel transcripts induced by aberrant transcription factors.

This potential mechanism of neoantigen generation is very interesting because it would lead to tumor-specific public antigens[71]. Most currently characterized neoantigens are derived from mutational processes specific to each patient and are thus called "private" antigens. Designing personalized immunotherapies to target them is time- and resource- consuming. In contrast, some antigens derived from normal cells can be shared by tumor cells across patients but will lead to toxicities in normal tissues[201]. The ideal targets are therefore tumor-specific, not expressed in normal cells, clonally expressed in all tumor cells and shared by many patients – tumor-specific public antigens. In this case, it would be possible to design immunotherapies available "off-the-shelf" to treat many patients, which could in turn lead to more rapid and less costly treatments. In addition, resistance to immunotherapies is often driven by the loss of the antigen by the tumor[202]. Since neotranscripts are direct downstream consequences of the oncogenic driver event, they would not easily be "lost" by the tumor without it rewiring a critical circuit of its oncogenic process. Independently of their translational relevance, what is also fascinating to me from this study is the ability of modified transcription factors to induce totally novel transcription units – genes-. This leads us to speculate if this could not be one of the processes at play behind the appearance of novel genes "de novo" during evolution[90]: transcription factors that are randomly modified could lead to their acquiring novel binding properties, thus targeting them to silent regions in the genome and allowing the transcription of novel transcripts, that may then through genetic drift acquire functional properties and even be translated into novel proteins.

One significant advance in the understanding of the immune microenvironment of cancers has been the advent of single-cell molecular assays, of which single-cell RNA-seq is the most developed in terms of experimental techniques and bioinformatics analyses. While previous assays on bulk tumor samples had to estimate the different immune populations using algorithms such as MCP-counter, single-cell RNA-seq has allowed to directly characterize the specific transcriptome of each single cell and the precise composition

of the immune microenvironment in one sample. Indeed, we were able to quantify the different proportions of immune cells inside dedifferentiated liposarcoma and showed that the dedifferentiated compartment is more highly infiltrated by lymphocytes and macrophages. We also could characterize their phenotype as more "exhausted" and immunosuppressive. This is a significant improvement as compared to bulk RNA-seq; however practical considerations - of cost notably - preclude a comparatively large study across numerous tumors. For the study of the adaptive immune system in particular, the possibility to sequence TCRs and BCRs at the single-cell level is also a significant advance as compared to global results obtained from bulk RNA-seq using methods such as MiXCR. I did not get the opportunity to analyze this sort of single-cell data, however these assays have the potential to greatly enhance our understanding of the immune response to cancer. As discussed previously, this precise characterization could eventually pinpoint the combination of specific TCR/BCR and corresponding antigen responsible for response to immunotherapy. Indeed, several studies have already tried to profile this information in patients before and after treatment with immunotherapy; this is surely one of the exciting potential uses of this kind of technology for personalized immunotherapy[166].

Concerning the mechanisms of "oncogenesis" of sarcomas, single-cell RNA-seq also opens many more perspectives than bulk RNA-seq: I could have explored more in depth some types of sarcomas in bulk RNA-seq to look for functional information on specific genes or pathways contributing to disease, however being able to characterize the transcriptome at the single-cell level was a lot more instructive to characterize molecular forces at play in DDLPS. Only a few papers have been published (Ewing sarcoma[203], synovial sarcoma[204], osteosarcoma[205]) but we can expect more single-cell studies in sarcoma to be conducted in the near future to enable better understanding of the oncogenesis of these cancers.

Sarcomas are indeed one of the best models to study oncogenic processes, since many of them - especially translocation-related sarcomas – are the consequences of one unique driver event that leads to oncogenic transformation[7,8]. In a way, this process is "simple" and facilitates the deciphering of oncogenesis. Therefore, the possibilities offered by scRNA-seq of depicting the heterogeneity within the same tumor is even more interesting for sarcomas, since it can pinpoint phenomena of plasticity that are independent of genetic alterations[206]. Obviously, obtaining complementary information from epigenetics would be needed to explain this, but this is now possible using single-cell assays of chromatin accessibility (assay for transposase-accessible chromatin: scATAC-seq)[207,208] or chromatin-binding proteins (scChIP-seq[209], and single-cell Cleavage Under Targets and Tagmentation: scCUT&Tag)[210]. I did not get the opportunity to analyze these other modalities, but I expect them to be very important for our understanding of the plasticity of cancer in its adaptation to a hostile microenvironment, interactions with normal cells and resistance to therapy. One important issue that could be addressed also is the question of the cell of origin[211,212], since the epigenetic and transcriptomic profiles might conserve a signature of this cell. This is a fundamental question in oncology since oncogenic processes should probably occur in one specific cell of origin to become cancerous. Sarcoma, due to its relatively simple oncogenic process, is thus a key model to understand this interaction between normal cellular programs and development of cancer. In parallel with hematology, where each leukemia or lymphoma subtype can in a way be related to a specific cell of origin in the hematopoiesis tree[213], one could contemplate the prospect of mapping each sarcoma to its cell of origin on a developmental manifold of "mesenchymopoiesis". Another exciting avenue is the ability to profile multiple modalities in the same single cell[116,117,214] - including genomic alterations, transcriptomic perturbations and the underlying epigenetic changes - opening the way for a more comprehensive picture of the oncogenic processes at play at the single-cell level. This could also lead to better understanding of cells of the microenvironment such as immune cells; for instance they are probably driven by epigenetic and programs that determine their transcriptome, function and possibly response to immunotherapy[215]. In the case of lymphocytes, the prospect of studying clonal dynamics is also enabled by the complementary assays of single-cell TCR/BCR profiling. One very exciting avenue for single-cell studies is the ability to gather information about space: so-called "spatial transcriptomics" are developing at a great pace and several

technologies are already available[216], though most are still not technically attaining single-cell resolution. Classical scRNA-seq involves dissociation of the cancer tissue, but spatial organization of cells surely plays a major role in the oncogenic process of the tumor and its interactions with the microenvironment. For instance, receptor-ligand interactions necessitate physical proximity to take place: computational algorithms can try to predict ligand-receptor interactions from classical scRNA-seq data[217,218], however these predictions would be much more confident with spatial information to corroborate them. Another issue that could also be addressed is the relative position between a cancer cell expressing one specific antigen relatively to the immune cell recognizing it; it is expected that this could possibly influence response to immunotherapy[184]. Indeed, spatial restriction could be one way for tumor cells to evade the immune system. This could also complement previous insights about the potential dynamic processes at play at the interface between cancer cells and immune cells: proliferation of clonotypes targeting specific antigens in the tumor, leading to selection pressures and so-called "immunoediting" with consequent loss of the antigen by the tumor cells[219]. Finally, one obvious extension to this deep characterization would be longitudinal profiling across time, but tissue and money are limited; for now we can only dream of a complete characterization of the dynamics in space and time of oncogenic processes inside cancer cells, their interactions with the environment and response to treatment.

From a computational point of view, the analysis of single-cell RNA-seq in particular gave me the opportunity to experience both the enormous potential and dangerous pitfalls of the analysis of high-dimensional data in biology and medicine. On one side, this technology gives an unprecedentedly detailed snapshot of biology and cancer; on the other hand this huge amount of data needs highly rigorous statistical tools and critical analysis to avoid falling into inconsistent results. Indeed the more data there is, the more likely we are of "falling" into false positive results. The crucial issue of data integration for cancer samples, which remains an open problem, allowed me to appreciate critically the power of computational algorithms and the risk of their "over-use", notably when considered in the more general framework of computational methods that have to find the right balance between technical correction of batch effect (or smoothing of the data for analysis purposes) as opposed to conservation of biological signal (and possibly associated noise or residual batch effect that may disturb the analysis). One may even think that there is probably no gold-standard method attainable for this issue and that the optimal balance depends upon specific questions and experiments at hand.

Single-cell RNA-seq is one of the domains where machine learning and deep learning methods have been the most widely imported into biology, due to the high amount of data generated by these single-cell assays[132]. This allowed me to learn and manipulate multiple methods of classical machine learning but also deep learning, which are indeed very powerful both for supervised (e.g. diagnosis prediction) and unsupervised analysis (e.g. encoding the transcriptomic landscape into meaningful low dimensions). More than ever computational methods have the power to analyze, interpret and even make "discoveries" for biology. Moreover, they can be used in the clinic to improve diagnosis and prognosis of individual patients, as was the case for our "classifier" of cancers of unknown primary.

Genomic data is currently the main training data for machine learning algorithms in medical oncology (radiotherapy in contrast is taking advantage of large amounts of radiological data, for instance with so-called "radiomics"[220]), due to its high-dimensional throughput. As a clinician, I am also very interested in developing these tools to analyze other clinical data such as patient symptoms, bio-variables, and EHR (electronic health record) data[221]. I believe that the main challenge for using these types of data for now is that they are less well-structured than genomic data; however I am sure that we could use them with great power once we address this issue. This could for instance be improved by standardizing EHRs and other clinical variables, as well as sensitizing clinicians to the importance of recording well-structured data and simplifying the processes for doing so. The challenge for machine learning is usually not the power of the algorithms, but the quality and quantity of the data that is used to train them: if we progress in this respect, I have no doubt that machine learning could one day be - not a replacement but - a very powerful helping

hand for the clinician. Finally, this emphasizes the importance of training clinicians to become familiar with concepts of machine learning, so that these tools are not blindly but correctly used by them for the benefit of the patient.

Considering the huge amount of biomedical data currently generated, I am confident that judicious use of computational methods will result in even more exciting developments for the fundamental understanding of the biology of cancer and the care of patients. In addition to my clinical duties as a medical oncologist, I will definitely continue to learn and work in this direction to try to contribute to these objectives.

Thank you very much for reading me!

References :

1. Vibert, J. & Thomas-Vaslin, V. Modelling T cell proliferation: Dynamics heterogeneity depending on cell differentiation, age, and genetic background. *PLoS Comput. Biol.* **13**, e1005417 (2017).
2. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
3. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).
4. WHO Classification of Tumours Editorial Board. *Soft Tissue and Bone Tumours*. (2020).
5. Ratan, R. & Patel, S. R. Chemotherapy for soft tissue sarcoma. *Cancer* **122**, 2952–2960 (2016).
6. Dufresne, A., Brahmi, M., Karanian, M. & Blay, J.-Y. Using biology to guide the treatment of sarcomas and aggressive connective-tissue tumours. *Nat. Rev. Clin. Oncol.* **15**, 443–458 (2018).
7. Mertens, F. *et al.* Translocation-related sarcomas. *Semin. Oncol.* **36**, 312–323 (2009).
8. Perry, J. A., Seong, B. K. A. & Stegmaier, K. Biology and Therapy of Dominant Fusion Oncoproteins Involving Transcription Factor and Chromatin Regulators in Sarcomas. *Annu. Rev. Cancer Biol.* **3**, 299–321 (2019).
9. Delattre, O. *et al.* Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature* **359**, 162–165 (1992).
10. Blay, J.-Y. *et al.* Surgery in reference centers improves survival of sarcoma patients: a nationwide study. *Ann. Oncol.* **30**, 1143–1153 (2019).
11. Waldman, A. D., Fritz, J. M. & Lenardo, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat. Rev. Immunol.* **20**, 651–668 (2020).
12. Tawbi, H. A. *et al.* Pembrolizumab in advanced soft-tissue sarcoma and bone sarcoma (SARC028): a multicentre, two-cohort, single-arm, open-label, phase 2 trial. *Lancet Oncol.* **18**, 1493–1501 (2017).
13. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
14. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

15. Majzner, R. G., Heitzeneder, S. & Mackall, C. L. Harnessing the Immunotherapy Revolution for the Treatment of Childhood Cancers. *Cancer Cell* **31**, 476–485 (2017).

16. Grünewald, T. G. P. *et al.* Ewing sarcoma. *Nat. Rev. Dis. Primer* **4**, 1–22 (2018).

17. Riggi, N., Suvà, M. L. & Stamenkovic, I. Ewing's Sarcoma. *N. Engl. J. Med.* **384**, 154–164 (2021).

18. Lessnick, S. L. *et al.* Small round cell sarcomas. *Semin. Oncol.* **36**, 338–346 (2009).

19. Delattre, O. *et al.* The Ewing family of tumors--a subgroup of small-round-cell tumors defined by specific chimeric transcripts. *N. Engl. J. Med.* **331**, 294–299 (1994).

20. Renzi, S., Anderson, N. D., Light, N. & Gupta, A. Ewing-like sarcoma: An emerging family of round cell sarcomas. *J. Cell. Physiol.* **234**, 7999–8007 (2019).

21. Sbaraglia, M., Righi, A., Gambarotti, M. & Dei Tos, A. P. Ewing sarcoma and Ewing-like tumors. *Virchows Arch. Int. J. Pathol.* **476**, 109–119 (2020).

22. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).

23. Jones, W. *et al.* Deleterious effects of formalin-fixation and delays to fixation on RNA and miRNA-Seq profiles. *Sci. Rep.* **9**, 6980 (2019).

24. Pierron, G. *et al.* A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion. *Nat. Genet.* **44**, 461–466 (2012).

25. Watson, S. *et al.* Transcriptomic definition of molecular subgroups of small round cell sarcomas. *J. Pathol.* **245**, 29–40 (2018).

26. Pilard, C. *et al.* Cancer immunotherapy: it's time to better predict patients' response. *Br. J. Cancer* 1–12 (2021) doi:10.1038/s41416-021-01413-x.

27. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).

28. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **24**, 541–550 (2018).

29. Tumeh, P. C. *et al.* PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).

30. Petitprez, F. *et al.* B cells are associated with survival and immunotherapy response in sarcoma. *Nature* **577**, 556–560 (2020).

31. Helmink, B. A. *et al.* B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* **577**, 549–555 (2020).

32. Cabrita, R. *et al.* Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* **577**, 561–565 (2020).

33. Liu, R. *et al.* Influence of Tumor Immune Infiltration on Immune Checkpoint Inhibitor Therapeutic Efficacy: A Computational Retrospective Study. *Front. Immunol.* **12**, 2397 (2021).

34. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol. Biol. Clifton NJ* **1711**, 243–259 (2018).

35. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

36. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).

37. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).

38. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

39. Abeshouse, A. *et al.* Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* **171**, 950-965.e28 (2017).

40. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).

41. Chaplin, D. D. Overview of the Immune Response. *J. Allergy Clin. Immunol.* **125**, S3-23 (2010).

42. Pai, J. A. & Satpathy, A. T. High-throughput and single-cell T cell receptor sequencing technologies. *Nat. Methods* 1–12 (2021) doi:10.1038/s41592-021-01201-8.

43. Goldstein, L. D. *et al.* Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Commun. Biol.* **2**, 1–10 (2019).

44. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).

45. Kallen, M. E. & Hornick, J. L. The 2020 WHO Classification: What's New in Soft Tissue Tumor Pathology? *Am. J. Surg. Pathol.* **45**, e1 (2021).

46. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).

47. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).

48. Koelsche, C. *et al.* Sarcoma classification by DNA methylation profiling. *Nat. Commun.* **12**, 498 (2021).

49. Bernstein, B. E., Meissner, A. & Lander, E. S. The Mammalian Epigenome. *Cell* **128**, 669–681 (2007).

50. Bormann, F. *et al.* Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal Carcinogenesis. *Cell Rep.* **23**, 3407–3418 (2018).

51. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

52. Hastie, T., Tibshirani, R. & Friedman, J. Linear Methods for Regression. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds. Hastie, T., Tibshirani, R. & Friedman, J.) 43–99 (Springer, 2009). doi:10.1007/978-0-387-84858-7_3.

53. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

54. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).

55. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).

56. Janson, N. B. Non-linear dynamics of biological systems. *Contemp. Phys.* **53**, 137–168 (2012).

57. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *ArXiv13126114 Cs Stat* (2014).

58. Hastie, T., Tibshirani, R. & Friedman, J. Neural Networks. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds. Hastie, T., Tibshirani, R. & Friedman, J.) 389–416 (Springer, 2009). doi:10.1007/978-0-387-84858-7_11.

59. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **23**, 80–91 (2018).

60. Pittenger, M. F. *et al.* Mesenchymal stem cell perspective: cell biology to clinical progress. *Npj Regen. Med.* **4**, 1–15 (2019).

61. Hastie, T., Tibshirani, R. & Friedman, J. Random Forests. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds. Hastie, T., Tibshirani, R. & Friedman, J.) 587–604 (Springer, 2009). doi:10.1007/978-0-387-84858-7_15.

62. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, (2015).

63. Costantino, C., Thomas, G. V., Ryan, C., Coakley, F. V. & Troxell, M. L. Metastatic renal cell carcinoma without evidence of a renal primary. *Int. Urol. Nephrol.* **48**, 73–77 (2016).

64. Pavlidis, N. & Fizazi, K. Carcinoma of unknown primary (CUP). *Crit. Rev. Oncol. Hematol.* **69**, 271–278 (2009).

65. Vibert, J. *et al.* Identification of Tissue of Origin and Guided Therapeutic Applications in Cancers of Unknown Primary Using Deep Learning and RNA Sequencing (TransCUPtomics). *J. Mol. Diagn. JMD* S1525-1578(21)00215–4 (2021) doi:10.1016/j.jmoldx.2021.07.009.

66. Hutzen, B. *et al.* Immunotherapies for pediatric cancer: current landscape and future perspectives. *Cancer Metastasis Rev.* **38**, 573–594 (2019).

67. Gangwal, K. *et al.* Microsatellites as EWS/FLI response elements in Ewing's sarcoma. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10149–10154 (2008).

68. Guillon, N. *et al.* The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. *PloS One* **4**, e4932 (2009).

69. Blass, E. & Ott, P. A. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nat. Rev. Clin. Oncol.* **18**, 215–229 (2021).

70. Hsiue, E. H.-C. *et al.* Targeting a neoantigen derived from a common TP53 mutation. *Science* **371**, (2021).

71. Pearlman, A. H. *et al.* Targeting public neoantigens for cancer immunotherapy. *Nat. Cancer* **2**, 487–497 (2021).

72. Audoux, J. *et al.* DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *Genome Biol.* **18**, 243 (2017).

73. Hölzer, M. & Marz, M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience* **8**, (2019).

74. Florea, L. D. & Salzberg, S. L. Genome-Guided Transcriptome Assembly in the Age of Next-Generation Sequencing. *IEEEACM Trans. Comput. Biol. Bioinforma. IEEE ACM* **10**, 1234–1240 (2013).

75. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

76. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

77. Shao, M. & Kingsford, C. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* **35**, 1167–1169 (2017).

78. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **9**, 304 (2020).

79. Gjerstorff, M. F., Andersen, M. H. & Ditzel, H. J. Oncogenic cancer/testis antigens: prime candidates for immunotherapy. *Oncotarget* **6**, 15772–15787 (2015).

80. Kimura, K. *et al.* Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16**, 55–65 (2006).

81. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).

82. Surdez, D. *et al.* STAG2 mutations alter CTCF-anchored loop extrusion, reduce cis-regulatory interactions and EWSR1-FLI1 activity in Ewing sarcoma. *Cancer Cell* (2021) doi:10.1016/j.ccell.2021.04.001.

83. Tomazou, E. M. *et al.* Epigenome mapping reveals distinct modes of gene regulation and widespread enhancer reprogramming by the oncogenic fusion protein EWS-FLI1. *Cell Rep.* **10**, 1082–1095 (2015).

84. Riggi, N. *et al.* EWS-FLI1 utilizes divergent chromatin remodeling mechanisms to directly activate or repress enhancer elements in Ewing sarcoma. *Cancer Cell* **26**, 668–681 (2014).

85. Dufresne, A. *et al.* Desmoplastic Small Round Cell Tumor: Current Management and Recent Findings. *Sarcoma* **2012**, e714986 (2012).

86. Gedminas, J. M. *et al.* Desmoplastic small round cell tumor is dependent on the EWS-WT1 transcription factor. *Oncogenesis* **9**, 1–8 (2020).

87. Boulay, G. *et al.* Epigenome editing of microsatellite repeats defines tumor-specific enhancer functions and dependencies. *Genes Dev.* **32**, 1008–1019 (2018).

88. Sartorelli, V. & Lauberth, S. M. Enhancer RNAs are an important regulatory layer of the epigenome. *Nat. Struct. Mol. Biol.* **27**, 521–528 (2020).

89. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).

90. Oss, S. B. V. & Carvunis, A.-R. De novo gene birth. *PLOS Genet.* **15**, e1008160 (2019).

91. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).

92. Ruiz Cuevas, M. V. *et al.* Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep.* **34**, 108815 (2021).

93. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).

94. Laumont, C. M. *et al.* Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **7**, 10238 (2016).

95. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).

96. Ouspenskaia, T. *et al.* Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer. *bioRxiv* 2020.02.12.945840 (2020) doi:10.1101/2020.02.12.945840.

97. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).

98. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).

99. Brar, G. A. & Weissman, J. S. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* **16**, 651–664 (2015).

100. Ingolia, N. T. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165**, 22–33 (2016).

101. Calviello, L. & Ohler, U. Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends Genet.* **33**, 728–744 (2017).

102. Miettinen, T. P. & Björklund, M. Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3' untranslated regions. *Nucleic Acids Res.* **43**, 1019–1034 (2015).

103. Bazin, J. *et al.* Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. *Proc. Natl. Acad. Sci.* **114**, E10018–E10027 (2017).

104. Calviello, L., Sydow, D., Harnett, D. & Ohler, U. Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data. *bioRxiv* 601468 (2019) doi:10.1101/601468.

105. Calviello, L., Hirsekorn, A. & Ohler, U. Quantification of translation uncovers the functions of the alternative transcriptome. *Nat. Struct. Mol. Biol.* **27**, 717–725 (2020).

106. Noor, Z., Ahn, S. B., Baker, M. S., Ranganathan, S. & Mohamedali, A. Mass spectrometry–based protein identification in proteomics—a review. *Brief. Bioinform.* **22**, 1620–1638 (2021).

107. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).

108. Saxena, M., van der Burg, S. H., Melief, C. J. M. & Bhardwaj, N. Therapeutic cancer vaccines. *Nat. Rev. Cancer* **21**, 360–378 (2021).

109. Van Hoecke, L. *et al.* mRNA in cancer immunotherapy: beyond a source of antigen. *Mol. Cancer* **20**, 48 (2021).

110. Feins, S., Kong, W., Williams, E. F., Milone, M. C. & Fraietta, J. A. An introduction to chimeric antigen receptor (CAR) T-cell immunotherapy for human cancer. *Am. J. Hematol.* **94**, S3–S9 (2019).

111. Zhukovsky, E. A., Morse, R. J. & Maus, M. V. Bispecific antibodies and CARs: generalized immunotherapeutics harnessing T cell redirection. *Curr. Opin. Immunol.* **40**, 24–35 (2016).

112. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).

113. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).

114. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).

115. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

116. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).

117. Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* **10**, e63632 (2021).

118. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

119. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).

120. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

121. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).

122. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

123. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

124. Bi, K. *et al.* Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. *Cancer Cell* **39**, 649-661.e5 (2021).

125. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).

126. Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11**, 2285 (2020).

127. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).

128. Keung, E. Z. & Somaiah, N. Overview of liposarcomas and their genomic landscape. *J. Transl. Genet. Genomics* **3**, (2019).

129. Thway, K. Well-differentiated liposarcoma and dedifferentiated liposarcoma: An updated review. *Semin. Diagn. Pathol.* **36**, 112–121 (2019).

130. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

131. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).

132. Raimundo, F., Meng-Papaxanthos, L., Vallot, C. & Vert, J.-P. Machine learning for single-cell genomics data analysis. *Curr. Opin. Syst. Biol.* **26**, 64–71 (2021).

133. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 1169 (2020).

134. Zappia, L. & Theis, F. J. *Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape*. 2021.08.13.456196 https://www.biorxiv.org/content/10.1101/2021.08.13.456196v1 (2021) doi:10.1101/2021.08.13.456196.

135. Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).

136. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).

137. Lun, A. Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. *bioRxiv* 404962 (2018) doi:10.1101/404962.

138. Booeshaghi, A. S. & Pachter, L. Normalization of single-cell RNA-seq counts by log(x + 1)† or log(1 + x)†. *Bioinformatics* **37**, 2223–2224 (2021).

139. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).

140. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

141. Hastie, T., Tibshirani, R. & Friedman, J. Overview of Supervised Learning. in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (eds. Hastie, T., Tibshirani, R. & Friedman, J.) 9–41 (Springer, 2009). doi:10.1007/978-0-387-84858-7_2.

142. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

143. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).

144. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).

145. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).

146. Fan, J. *et al.* Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* gr.228080.117 (2018) doi:10.1101/gr.228080.117.

147. Amin-Mansour, A. *et al.* Genomic Evolutionary Patterns of Leiomyosarcoma and Liposarcoma. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **25**, 5135–5142 (2019).

148. Beird, H. C. *et al.* Genomic profiling of dedifferentiated liposarcoma compared to matched well-differentiated liposarcoma reveals higher genomic complexity and a common origin. *Mol. Case Stud.* **4**, a002386 (2018).

149. Hirata, M. *et al.* Integrated exome and RNA sequencing of dedifferentiated liposarcoma. *Nat. Commun.* **10**, 5683 (2019).

150. Sun, W. *et al.* snRNA-seq reveals a subpopulation of adipocytes that regulates thermogenesis. *Nature* **587**, 98–102 (2020).

151. Davoodzadeh Gholami, M. *et al.* Exhaustion of T lymphocytes in the tumor microenvironment: Significance and effective mechanisms. *Cell. Immunol.* **322**, 1–14 (2017).

152. Najafi, M. *et al.* Macrophage polarity in cancer: A review. *J. Cell. Biochem.* **120**, 2756–2765 (2019).

153. Forcato, M., Romano, O. & Bicciato, S. Computational methods for the integrative analysis of single-cell data. *Brief. Bioinform.* **22**, (2021).

154. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).

155. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).

156. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

157. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

158. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

159. Barkas, N. *et al.* Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).

160. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).

161. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

162. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

163. Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp. Mol. Med.* **52**, 1452–1465 (2020).

164. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

165. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **175**, 998-1013.e20 (2018).

166. Yost, K. E. *et al.* Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med.* **25**, 1251–1259 (2019).

167. Yofe, I., Dahan, R. & Amit, I. Single-cell genomic approaches for developing the next generation of immunotherapies. *Nat. Med.* **26**, 171–177 (2020).

168. Davis-Marcisak, E. F. *et al.* From bench to bedside: Single-cell analysis for cancer immunotherapy. *Cancer Cell* **39**, 1062–1080 (2021).

169. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

170. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

171. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

172. Chen, H. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).

173. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

174. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

175. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).

176. Quintás-Cardama, A. & Cortes, J. Molecular biology of bcr-abl1–positive chronic myeloid leukemia. *Blood* **113**, 1619–1630 (2009).

177. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

178. Sparano, J. A. *et al.* Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *N. Engl. J. Med.* **379**, 111–121 (2018).

179. Rubin, G. D. Computed Tomography: Revolutionizing the Practice of Medicine for 40 Years. *Radiology* **273**, S45–S74 (2014).

180. Race, G. J., Tillery, G. W. & Dysert, P. A. A history of pathology and laboratory medicine at Baylor University Medical Center. *Proc. Bayl. Univ. Med. Cent.* **17**, 42–55 (2004).

181. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).

182. Upadhaya, S., Hubbard-Lucey, V. M. & Yu, J. X. Immuno-oncology drug development forges on despite COVID-19. *Nat. Rev. Drug Discov.* **19**, 751–752 (2020).

183. Adashek, J. J. *et al.* Hyperprogression and Immunotherapy: Fact, Fiction, or Alternative Fact? *Trends Cancer* **6**, 181–191 (2020).

184. Chen, D. S. & Mellman, I. Elements of cancer immunity and the cancer-immune set point. *Nature* **541**, 321–330 (2017).

185. Zitvogel, L., Ma, Y., Raoult, D., Kroemer, G. & Gajewski, T. F. The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies. *Science* **359**, 1366–1370 (2018).

186. Paijens, S. T., Vledder, A., de Bruyn, M. & Nijman, H. W. Tumor-infiltrating lymphocytes in the immunotherapy era. *Cell. Mol. Immunol.* **18**, 842–859 (2021).

187. Galon, J. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313**, 1960–1964 (2006).

188. El Sissy, C. *et al.* Therapeutic Implications of the Immunoscore in Patients with Colorectal Cancer. *Cancers* **13**, 1281 (2021).

189. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).

190. Goodman, A. M. *et al.* Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol. Cancer Ther.* **16**, 2598–2608 (2017).

191. Jardim, D. L., Goodman, A., Gagliato, D. de M. & Kurzrock, R. The Challenges of Tumor Mutational Burden as an Immunotherapy Biomarker. *Cancer Cell* **39**, 154–173 (2021).

192. Davis, A. A. & Patel, V. G. The role of PD-L1 expression as a predictive biomarker: an analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors. *J. Immunother. Cancer* **7**, 278 (2019).

193. Joglekar, A. V. & Li, G. T cell antigen discovery. *Nat. Methods* **18**, 873–880 (2021).

194. Chen, S.-Y., Yue, T., Lei, Q. & Guo, A.-Y. TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. *Nucleic Acids Res.* **49**, D468–D474 (2021).

195. Zvyagin, I. V., Tsvetkov, V. O., Chudakov, D. M. & Shugay, M. An overview of immunoinformatics approaches and databases linking T cell receptor repertoires to their antigen specificity. *Immunogenetics* **72**, 77–84 (2020).

196. Fischer, D. S., Wu, Y., Schubert, B. & Theis, F. J. Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Mol. Syst. Biol.* **16**, e9416 (2020).

197. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 1–7 (2021) doi:10.1038/s41586-021-03819-2.

198. Maude, S. L. *et al.* Tisagenlecleucel in Children and Young Adults with B-Cell Lymphoblastic Leukemia. *N. Engl. J. Med.* **378**, 439–448 (2018).

199. Schuster, S. J. *et al.* Tisagenlecleucel in Adult Relapsed or Refractory Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* **380**, 45–56 (2019).

200. Yamamoto, T. N., Kishton, R. J. & Restifo, N. P. Developing neoantigen-targeted T cell–based treatments for solid tumors. *Nat. Med.* **25**, 1488–1499 (2019).

201. Berner, F. *et al.* Association of Checkpoint Inhibitor–Induced Toxic Effects With Shared Cancer and Tissue Antigens in Non–Small Cell Lung Cancer. *JAMA Oncol.* **5**, 1–6 (2019).

202. Anagnostou, V. *et al.* Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. *Cancer Discov.* **7**, 264–276 (2017).

203. Aynaud, M.-M. *et al.* Transcriptional Programs Define Intratumoral Heterogeneity of Ewing Sarcoma at Single-Cell Resolution. *Cell Rep.* **30**, 1767-1779.e6 (2020).

204. Jerby-Arnon, L. *et al.* Opposing immune and genetic mechanisms shape oncogenic programs in synovial sarcoma. *Nat. Med.* **27**, 289–300 (2021).

205. Zhou, Y. *et al.* Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nat. Commun.* **11**, 6322 (2020).

206. Yuan, S., Norgard, R. J. & Stanger, B. Z. Cellular Plasticity in Cancer. *Cancer Discov.* **9**, 837–851 (2019).

207. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

208. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).

209. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066 (2019).

210. Wu, S. J. *et al.* Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nat. Biotechnol.* **39**, 819–824 (2021).

211. Visvader, J. E. Cells of origin in cancer. *Nature* **469**, 314–322 (2011).

212. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).

213. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018).

214. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).

215. Villanueva, L., Álvarez-Errico, D. & Esteller, M. The Contribution of Epigenetics to Cancer Immunotherapy. *Trends Immunol.* **41**, 676–691 (2020).

216. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).

217. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

218. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).

219. Rosenthal, R. *et al.* Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).

220. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).

221. Morin, O. *et al.* An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. *Nat. Cancer* **2**, 709–722 (2021).

Annex 1

Vibert, J. *et al.* Identification of Tissue of Origin and Guided Therapeutic Applications in Cancers of Unknown Primary Using Deep Learning and RNA Sequencing (TransCUPtomics). *J. Mol. Diagn. JMD* S1525-1578(21)00215–4 (2021) doi:10.1016/j.jmoldx.2021.07.009.

Annex 2

Vibert, J. *et al.* Oncogenic chimeric transcription factors drive recurrent patterns of tumor-specific transcription, processing, and translation of silent genomic regions. (submitted manuscript)

RÉSUMÉ

Les sarcomes sont des cancers d'origine mésenchymateuse qui comprennent plus d'une centaine d'entités. Ce sont pour la plupart des maladies rares qui peuvent survenir à tout âge, y compris pendant l'enfance et la jeune adolescence. En raison de leur rareté et diversité, le diagnostic en est souvent erroné ou retardé. Le pronostic est généralement sombre dans les formes avancées et métastatiques, et la plupart des traitements reposent actuellement sur des chimiothérapies non spécifiques et très toxiques. Il y a donc un besoin urgent d'améliorer le diagnostic des sarcomes et développer de nouvelles approches thérapeutiques pour ces cancers.

Le séquençage de l'ARN (RNA-seq) est une technique prometteuse pour le diagnostic des sarcomes, notamment dans le cas des sarcomes liés à des translocations qui sont caractérisés par des translocations chromosomiques à l'origine de gènes de fusion, par exemple *EWSR1-FLI1* dans le sarcome d'Ewing. A l'aide de la base de données du RNA-seq de sarcomes de patients de l'Institut Curie, j'ai exploré le paysage transcriptomique des ces cancers et utilisé des techniques d'apprentissage machine (machine learning) et d'apprentissage profond (deep learning) pour prédire le type de sarcome à l'aide du RNA-seq. Ce travail a ensuite permis le développement d'un outil actuellement utilisé à l'Institut Curie pour prédire la tumeur d'origine de cancers de primitif inconnu et ainsi améliorer le diagnostic et le pronostic de patients en pratique clinique courante.

Au cours de la dernière décennie, l'immunothérapie a été à l'origine d'une révolution dans le traitement de multiples cancers. Cependant, elle n'a eu qu'un succès très limité dans les sarcomes qui sont généralement considérés comme des tumeurs non « immunogéniques ». En effet, la plupart des sarcomes, notamment liés aux translocations, ont une charge mutationnelle très faible. Or ce dernier facteur est considéré comme l'un des principaux générateurs de néoantigènes tumoraux qui servent de cible au système immunitaire. Pour étudier plus en détail la possibilité d'une réponse immunitaire dans les sarcomes, j'ai caractérisé le microenvironnement tumoral immunitaire et les répertoires lymphocytaires dans de nombreux types de sarcomes à l'aide du RNA-seq d'échantillons tumoraux. Bien que la plupart sont effectivement peu infiltrés par des cellules du système immunitaire, il existe des exceptions qui font penser que l'immunothérapie pourrait être efficace dans certains cas.

Une autre piste prometteuse pour l'immunothérapie des sarcomes a été l'identification de nouveaux transcrits spécifiques dans de nombreux types de sarcomes liés à des translocations. Ces « néotranscrits » sont induits par le facteur de transcription oncogénique chimérique caractéristique de la tumeur, par exemple *EWSR1-FLI1* dans le sarcome d'Ewing. Certains d'entre eux sont traduits par les ribosomes en peptides. Ils représentent donc une source potentielle de néoantigènes publics spécifiques de la tumeur pour les approches d'immunothérapie dans les sarcomes liés à des translocations.

Pour caractériser en détail le microenvironnement immunitaire et les processus oncogéniques de sarcomes spécifiques, certains d'entre eux ont été étudiés par du RNA-seq à l'échelle unicellulaire (single-cell RNA-seq), notamment les liposarcomes dédifférenciés (DDLPS). Cette technique a mis en évidence une infiltration plus importante de cellules immunitaires dans le compartiment dédifférencié de la tumeur, ainsi qu'un phénotype « épuisé » (exhausted) et immunosuppresseur de ces cellules. Elle a aussi permis de caractériser les processus oncogéniques des DDLPS, notamment la relation entre les cellules bien différenciées et « dédifférenciées » au sein d'une même tumeur.

Au total, ce travail ouvre plusieurs perspectives pour l'amélioration du diagnostic et le développement d'immunothérapies pour les sarcomes, en : 1) définissant un paysage transcriptomique global des types de sarcomes et de leur microenvironnement immunitaire ; 2) identifiant de nouveaux mécanismes transcriptionnels dans les sarcomes liés à des translocations potentiellement à l'origine de néoantigènes pour l'immunothérapie ; 3) caractérisant à l'échelle unicellulaire les processus oncogéniques et le microenvironnement immunitaire d'un type de sarcome (liposarcome dédifférencié) ; 4) mettant en place un outil d'aide au diagnostic actuellement utilisé en pratique clinique courante à l'Institut Curie.

MOTS CLÉS

Sarcomes ; immunothérapie ; bioinformatique ; ARN ; single-cell

ABSTRACT

Sarcomas are cancers of mesenchymal origin that comprise more than a hundred different entities. They are mostly rare diseases that occur at all ages, including in children and young adolescents. Due to their rarity and diversity, diagnosis is often missed or delayed. Prognosis is generally poor in cases of advanced or metastatic disease and most treatment approaches rely on unspecific and highly toxic chemotherapy. There is thus an unmet need to improve the diagnosis of sarcomas and develop novel therapeutic approaches for these diseases.

RNA sequencing (RNA-seq) is a promising approach for the diagnosis of sarcomas, especially for translocation-related sarcomas that are characterized by chromosome translocations giving rise to fusion genes, such as *EWSR1-FLI1* in Ewing sarcoma. Using RNA-seq data of sarcomas of patients profiled at the Institut Curie, I explored the transcriptomic landscape of sarcomas and used machine learning and deep learning techniques to predict sarcoma type based on RNA-seq. This work led to the development of a tool currently in use at the Institut Curie to predict the origin of cancers of unknown primary and improve the diagnosis and prognosis of individual patients in clinical practice.

Immunotherapy has revolutionized cancer care for the last decade, however it has had only limited success in sarcomas, supposedly because they are not "immunogenic". Indeed, most sarcomas, especially translocation-related ones, have a very low tumor mutational burden, which is believed to be the main driving force in the generation of tumor neoantigens recognized by the immune system. To gain further insight into the potential of immune response in sarcoma, I characterized the immune microenvironment and lymphocyte repertoires of multiple types of sarcomas using RNA-seq of tumor samples. While most of them were indeed poorly infiltrated by cells of the immune system, there were some exceptions to this rule suggesting that immunotherapy should be considered in some cases.

Another promising finding for immunotherapy of sarcomas was the identification of novel tumor-specific transcripts in multiple types of translocation-related sarcomas. These "neotranscripts" were driven by their characteristic oncogenic chimeric transcription factors such as *EWSR1-FLI1* in Ewing sarcoma; some of them were found to be translated by ribosomes into peptides. Therefore, these may represent a source of tumor-specific public neoantigens for immunotherapies of these translocation-related sarcomas.

To characterize in detail the immune microenvironment and oncogenic processes of specific sarcomas, single-cell RNA-seq was performed for some of them, notably dedifferentiated liposarcomas (DDLPS). It revealed higher infiltration by immune cells in the dedifferentiated compartment of the tumor, but with more exhausted and immunosuppressive phenotypes. It also allowed to characterize the oncogenic processes of DDLPS and notably the relationship between dedifferentiated and well-differentiated cells inside the same tumor.

Altogether, this work opens perspectives to improve diagnosis and develop immunotherapies for sarcomas by: 1) defining a global transcriptomic landscape of sarcoma types and their associated microenvironment; 2) identifying novel transcriptional processes in translocation-related sarcomas with potential for generation of neoantigens for immunotherapy; 3) characterizing at the single-cell level oncogenic processes and immune microenvironment of one type of sarcoma (DDLPS); 4) resulting in the development of a classifier tool for diagnostic prediction used in clinical practice at the Institut Curie.

KEYWORDS

Sarcomas; immunotherapy; bioinformatics; RNA; single-cell