University of Gdańsk

Uniwersytet Gdański,
Międzyuczelniany Wydział
Biotechnologii Uniwersytetu
Gdańskiego i Gdańskiego
Uniwersytetu Medycznego.

Międzyuczelniany Wydział Biotechnologii UG i GUMed

Intercollegiate Faculty of Biotechnology UG and MUG

Integration as a solution: A multi-omic approach to cancer diseases.

Marcos Yébenes Mayordomo

2022

University of Gdańsk

Uniwersytet Gdański,
Międzyuczelniany Wydział
Biotechnologii Uniwersytetu
Gdańskiego i Gdańskiego
Uniwersytetu Medycznego.

Rozprawa doktorska
Doctoral dissertation

Integration as a solution: A multi-omic approach to cancer diseases.

Integracja danych jako taktyka: multiomiczne podejście do chorób nowotworowych

mgr Marcos Yébenes Mayordomo

A Ph. D. dissertation in the field of Natural Sciences
specialization Biological Sciences presented
to The Scientific Council of Biological Sciences
University of Gdańsk

Promoter: Prof. Theodore R. Hupp
Assistant promoter: Dr. Javier A. Alfaro

Gdańsk, 2022

Aequam memento rebus in arduis servare mentem,
non secus in bonis

Acknowledgments

I would like to express my deepest appreciation to the International Center for Cancer Vaccine Science for granting me the opportunity to perform this collaborative Ph.D. in cancer bioinformatics and to the PL-Grid - CI TASK infrastructure for providing the resources that made the studies possible. I would like to extend my sincere thanks to Dr. Jakub Faktor and Graeme Grimes who helped me in the writing and development of the Sarcoma study.

I could not have undertaken this journey without the support of Prof. Ted Hupp and Dr. Javier Alfaro, who have helped me to overcome all the challenges during my studies and helped me grow as a scientist and as a person.

I am also grateful for the colleagues that shared the journey with me, although we have been losing contact, we stuck together since the beginning helping each other. Miko, without your help this thesis won't be possible, I won't be able to express in words my gratitude to you.

Me tomaré la libertad de escribir unas líneas en español para agradecer el apoyo de todos mis amigos que me han ayudado a mantener la cordura durante este viaje, ya sea compartiendo su misma experiencia de doctorado o ayudándome a desconectar del mío. Me gustaría dedicar también unas palabras de agradecimiento a mi familia y especialmente a mis padres, sin su ayuda y apoyo no podría haberme convertido en la persona que soy hoy, tanto personal como profesionalmente.

Por último, pero no por ello menos importante, quería agradecer de corazón el apoyo de una persona que ha compartido mis buenos y malos momentos en este viaje. Ilaria, este doctorado es tanto tuyo como mío. No solo me has ayudado a llevar esta carga, si no que has comprendido lo que un doctorado puede llegar a ser. Ver la dedicación que tienes hacia tu trabajo y el buen humor con el que lo afrontas día tras día es y será una inspiración que me da fuerzas para continuar. Te amo.

Index of contents

List of abbreviations

Next generation sequencing (NGS)

The Cancer Genome Atlas (TCGA)

International Cancer Genome Consortium (ICGC)

The Cancer Proteome Atlas (TCPA)

Post-transcriptional modifications (PTMs)

Whole-exome sequencing (WES)

Whole-genome sequencing (WGS)

Mass spectrometry (MS)

Oesophageal cancer clinical and molecular stratification (OCCAMS)

Esophageal adenocarcinoma (EAC)

Transcripts per million (TPM)

Filter Aided Sample Preparation (FASP)

Formic acid (FA)

Automated gain control (AGC)

False Discovery Rate (FDR)

Trimmed mean of M-values (TMM)

Tumor against normal esophagus (TvE)

Tumor against gastric tissue (TvG)

Immunohistochemistry (IHC)

Tissue microarrays (TMAs)

RNA binding motif (RBM)

Insulin-Like Growth Factor-Binding Protein 1 (IGF2BP1)

Glycoprotein A33 (GPA33)

Prothymosin alpha (PTMA)

Eukaryotic Translation Elongation Factors (EEF1)

Undifferentiated pleomorphic sarcoma (UPS)

Single nucleotide variants (SNV)

Variant allele frequency (VAF)

Ensembl Variant Effect Predictor (VEP)

Mutation Annotation Format (MAF)

Multidimensional scaling (MDS)

Filter-Aided Sample Preparation (FASP)

Phosphine hydrochloride (TCEP)

LC-MS acetonitrile (AcN)

Data dependent (DDA)

Data independent (DIA)

Reverse phase liquid chromatography (RPLC)

Single nucleotide polymorphisms (SNPs)

Insertions (INS)

Deletions (DEL)

Mutant-Allele Tumor Heterogeneity (MATH)

Median absolute deviation (MAD)

Tumor-infiltrating lymphocytes (TIL)

Mesenchymal progenitor cells (MPCs)

Abstract

Cancer is a disease of the genome. Tumor cells contain many genetic and epigenetic mutations affecting diverse genes involving relevant cellular processes, such as proliferation or the evasion of apoptosis. Recently, advances in genomics and transcriptomics research have achieved the discovery of biomarkers and therapeutic targets, positioning these methods as the main tools for cancer research. The complete multi-omic landscape of most cancer types is still unknown, a significant gap in understanding cancer as the past focus on genomics alone can't provide a full picture of the mechanisms inside the tumor cells. The integration of different molecular profiling technologies (multi-omics) is the central topic in the studies, presented as a tool to give a more complete molecular phenotype of the diseases.

The first study combines mass spectrometry along with genomics and transcriptomics from a cohort containing more than four hundred esophageal adenocarcinoma samples. The analysis of changes in expression between the RNA and the proteins provides the identification of tumor-specific genes involved in the disease, revealing deregulatory mechanisms in the tumor cells and creating new opportunities for the development of new therapies.

The second part of the thesis focuses on the study of undifferentiated pleomorphic sarcoma. The search of a common altered pathway is carried through the genomic characterization of twenty patient samples, the mutational landscape, and its heterogeneity. Although alterations shared between most of the patients were detected, and a possible therapy is suggested, the high variability between samples suggests that a patient-specific treatment might be the best approach. Therefore, a computational model was conceived to predict the immunological presentation of mutation-borne neoantigens. The model is based on proteogenomics integration and will aid the development of personalized therapies.

The last study presents a case report of Gorham-Stout disease, a rare syndrome characterized by the uncontrollable growth of vascular tissue and the consumption of the surrounding bone matrix. Through the integration of genomics and transcriptomics, new possible disease markers and improvements in the current therapies are revealed.

Streszczenie

Nowotwory nazywane są chorobami genomu. Komórki nowotworowe zawierają liczne mutacje genetyczne i epigenetyczne. Dotknięte nimi geny regulują procesy komórkowe takie jak proliferacja i apoptoza. Postęp w dziedzinie genomiki i transkryptomiki pozwolił na odkrycie istotnych biomarkerów oraz celów terapeutycznych w chorobach nowotworowych. Tym samym metody te stały się kluczowe dla badań nad nowotworami. Analiza multi-omiczna pozwala na pełniejsze opisanie krajobrazu poszczególnych nowotworów, jednak dla większości z nich tak zintegrowane dane nie są jeszcze dostępne. Genomika – do niedawna podstawowa metoda – nie jest w stanie zbadać i opisać wszystkich mechanizmów zachodzących w komórkach nowotworowych. Integracja różnych metod profilowania molekularnego (multi-omika) jest narzędziem pozwalającym na wyczerpujący opis fenotypów molekularnych choroby oraz tematem przewodnim tej pracy.

Pierwsze spośród prezentowanych badań łączy spektrometrię mas z genomiką i transkryptomiką, bazując na zestawie danych z ponad czterystu próbek gruczolakoraka przełyku. Analiza zmian ekspresji molekuł RNA oraz białek pozwoliła na ustalenie genów specyficznie aktywnych w nowotworze i ujawniła mechanizmy de-regulacji ekspresji, wspierając rozwój nowych metod leczenia.

Druga część pracy koncentruje się na analizie niezróżnicowanego mięsaka pleomorficznego. Dwadzieścia próbek pobranych od pacjentów zostało scharakteryzowanych w zakresie mutacji i ich heterogeniczności. Zaobserwowano zmiany współdzielone przez większość pacjentów, co pozwoliło na zaproponowanie odpowiedniej metody leczenia. Jednak wysoki poziom zmienności pomiędzy próbkami sugeruje, że terapia spersonalizowana może być najlepszym podejściem w tej chorobie. Stworzono model komputerowy pozwalający na przewidywanie immunologicznej prezentacji neoantygenów pochodzących ze zmutowanych genów. Model bazuje na integracji danych proteogenomicznych i wspomoże rozwój terapii spersonalizowanych.

Ostatnia składowa opisywanych badań to opis przypadku choroby Gorhama-Stouta. Jest to rzadki syndrom polegający na niekontrolowanym rozroście naczyń krwionośnych w kościach, prowadzącym do zaniku tkanki kostnej. Integracja badań genetycznych i transkryptomicznych pozwoliła na wykrycie potencjalnych markerów chorobowych i usprawnień leczenia.

1. Introduction

1.1 Biology and bioinformatics background of cancer research.

1.1.1 Biological characteristics of the tumor environment.

Cancer is the second leading cause of death worldwide, right after cardiovascular disease. Although cancer is always referred to as a disease, it is a collection of different diseases that share common hallmarks that lead to the development of the tumor[1,2]. The treatment for these diseases has an estimated cost of one trillion dollars annually[3], a value only surmounted by the tragic loss of life and morbidity of this disease.

Cancer is a disease of the genome. Tumor cells contain a large number of genetic and epigenetic mutations affecting diverse genes involving relevant cellular processes, such as proliferation or the evasion of apoptosis[4]. The hallmarks described for cancer can arise through the accumulation of genomic aberrations that lead to the modification more broadly of the molecular phenotype, as these aberrations progress along with the central dogma of information flow in the cell. As the disease develops, the tumor as a whole exhibits hallmarks[2] that include:

1. Sustaining proliferative signaling: The signaling pathways of cancer cells are altered by inducing a continuous signaling cascade of growth and replication.
2. Evasion of growth suppressors: Cancer cells grow fast and uncontrollably. One of the reasons for this phenomenon is the evasion of the cell growth control mechanisms.
3. Avoidance of immune destruction: Although cancer cells are detected by the immune system, another evasion event is the surpass of immune vigilance.
4. Enabling replicate immortality: The limits in cancer cells are breached, permitting the reversion to a pre-differentiated state and avoiding the cell cycle.

5. Tumor-promoting inflammation: The inflammatory mechanisms are used by cancer cells to promote their own growth and survival.
6. Activating invasion and metastasis: Cancer cells can expand to other locations and migrate to other tissues breaking away from the primary tumor.
7. Inducing angiogenesis: Blood vessel growth is stimulated in cancer tissue, conceding nutrients to the tumor tissue.
8. Genome instability and mutation: The DNA contained in tumor cells is highly altered, enhancing the other tumor characteristics and deactivating tumor-suppression mechanisms.
9. Resisting cell death: Cancer cells can avoid the programmed pathways of apoptosis that normally induce the death of damaged cells.
10. Deregulating cellular energetics: To accomplish the growth of tumor cells, the energy demand is increased when compared to normal cells. This requirement is usually satisfied by increasing aerobic glycolysis[5].

Tumor-like diseases share features with the cancer environment, like uncontrolled cell proliferation, alteration of the genome, or inflammation of the affected tissues, usually presenting the same type of lesions on the tissue. Due to the similar behavior, it is difficult to distinguish clinically the malignancy of the lesions[6–8]. The research on tumor-like diseases can be then compared to the study of cancer, applying the same techniques to reveal the details of the cellular processes affected.

A multi-omic approach will reveal further molecular features involving different hallmarks of cancer that would be missed without the integration of single-omic studies. We have applied a combination of genomics, transcriptomics, and proteomics (**Figure 1.1**) to achieve a better understanding of the tumor molecular phenotype.

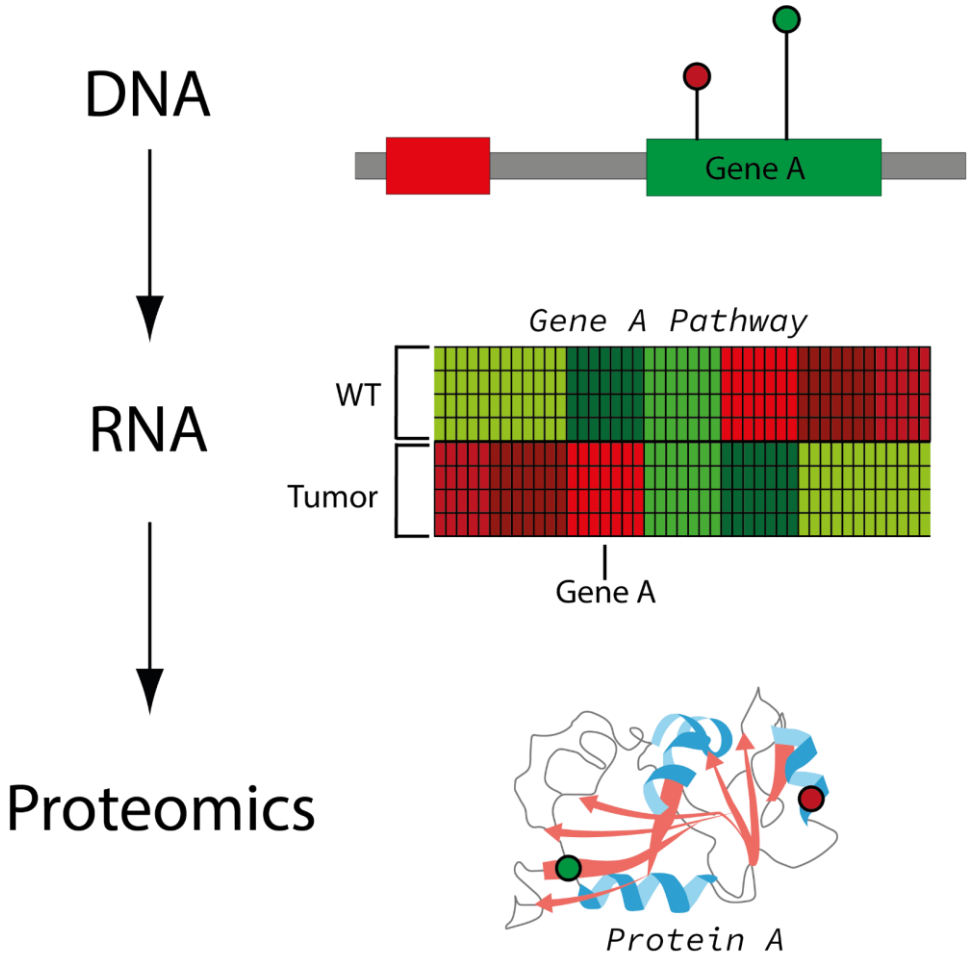


Figure 1.1. Multi-omic exploration allows the tracking of genomic mutations through the transcriptome and the expression profiles of its pathway until the altered protein product.

1.1.2 State of the art in cancer genomics studies.

Cancer is a disease of the genome, therefore the first line of analysis when performing cancer research has become the genomic analysis. The advances in cancer genomics have been made possible thanks to technological innovation in next-generation sequencing (NGS) techniques developed in the past 20 years and the careful gathering and exploitation of tumor biobanks rich with clinical data by clinicians. NGS has allowed the parallel sequencing of entire genomes and exomes as well as specific genomic regions, creating a unique opportunity to characterize cancer types by the parallel analysis of a large cohort of patients.

The evaluation of tumor mutations is a complex procedure that starts with obtaining samples from a specialist and its storage for future processing. Samples are then

collected in a laboratory and prepared for the desired analysis (DNA/RNA sequencing, immunohistochemistry, mass-spectrometry, etc.). Once completed, data analysis of the results is performed by bioinformaticians, reading, curating, and filtering the data to acquire the most relevant genomic/transcriptomic/proteomic information. The final results are then taken to clinical interpretation to investigate the relevance of the variants for the cancer type, taking diagnosis and possible treatments into consideration (**Figure 1.2**).

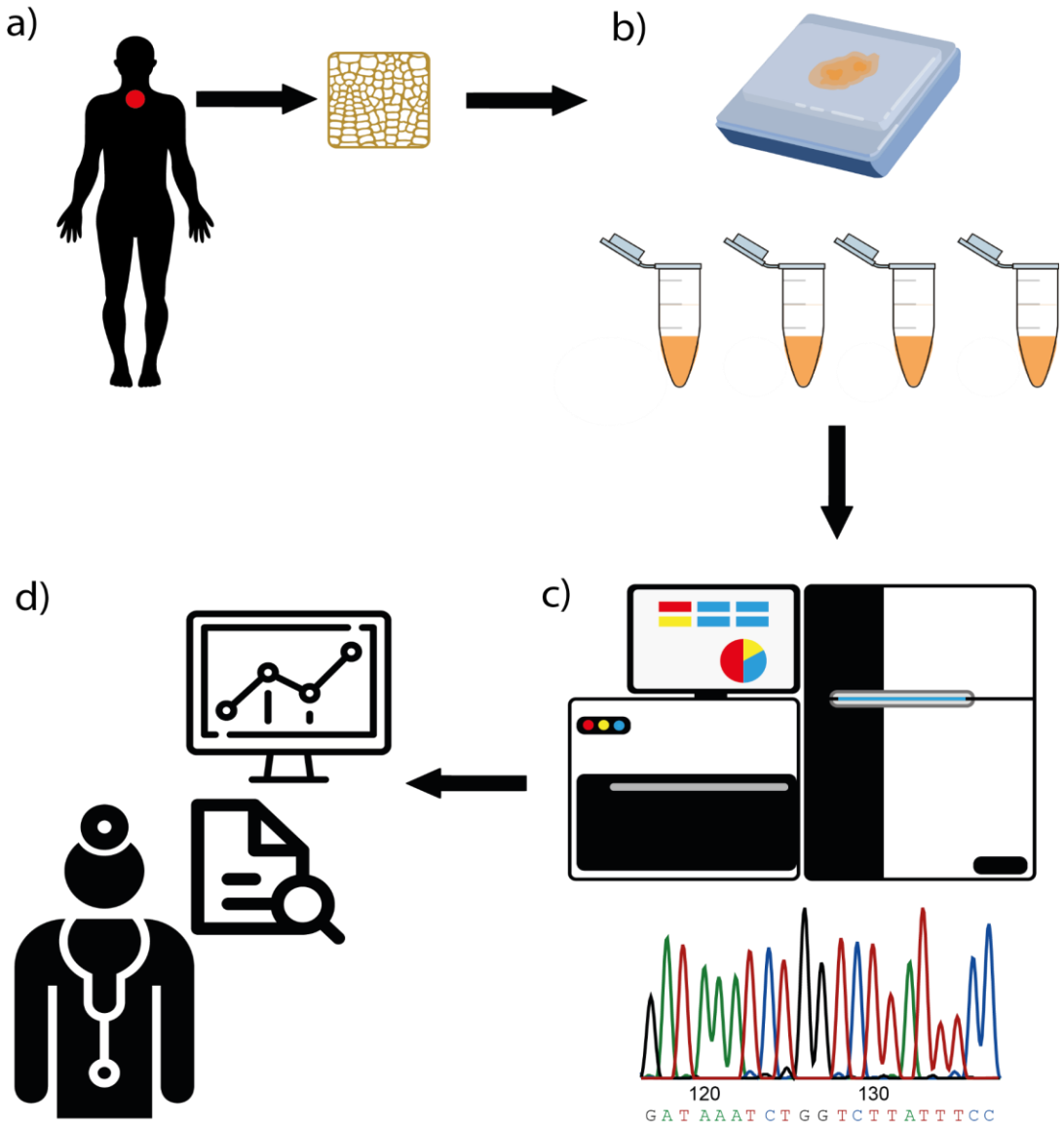


Figure 1.2. Workflow for NGS data processing and the evaluation cancer treatment. a) Tumor tissue is extracted from the patient. b) Extracted samples are processed and prepared for the NGS. c) Sequencing and mutation analysis is performed. c) Results are evaluated by an oncologist to determine their relevance and possible treatments.

The results of successful collaborative efforts determining the results of NGS studies have revealed the existence of a huge number of alterations that contribute to cancer formation and metastasis[9]. Mutations are then assessed as an impact on the population, studying the number of times an alteration is observed in a cohort (allele frequency) and performing survival analyses that allow determining the severity of the mutation[10]. The characterization of cancer genomes through population studies revealed relevant genes involved in cellular processes leading to the discovery of oncogenes in certain types of cancer, such as BRCA in breast and ovarian cancers[11]; BRAF in melanoma and colorectal cancers, or FLT3 in leukaemia[12].

Pan-cancer studies have revealed common genes altered across tumors, like the alteration of TP53, a gene coding for a tumor suppressor protein that is mutated in most types of cancers[13]. These studies elucidated if the role of mutated genes that were highly altered in a specific tumor were present in other tumors[14], allowing the discovery of altered genes[15] involved in fundamental cell processes such as DNA repair, cell cycle regulation, and cell adhesion[16].

The understanding and identification of gene mutations in cancer was propelled by the success of the Human Genome Project as well as the efforts of multiple consortia like the Cancer Genome Atlas[17] (TCGA), the Clinical Proteomics Tumour Analysis Consortium[18], and the International Cancer Genome Consortium[19] (ICGC). These platforms represent efforts to sequence tumors on mass and have contributed significantly to the understanding of the genetics behind cancer diseases. For the most part, cancer genomics has focused on the broad characterization of mutations and other such aberrations within large cancer cohorts generated by consortia like TCGA and ICGC. The methylation profile of DNA, along with these variants, have become the most frequent tools for DNA exploration in cancer research, providing a much-needed global perspective of nucleotide modification in this tissue.

1.1.3 Transcriptomics studies in cancer research.

Transcriptomics studies with their power to give a glimpse into the cellular states of the tumor are major contributors to the discovery of cancer biomarkers. As tumors accumulate aberrations in their genomes, the way that genes are transcribed to RNA changes with concomitant changes in gene expression indicative of an aberrant cellular state. Those changes can be used in clinical diagnosis, for estimating the stage of the tumor and therefore the prognosis of the patient or to distinguish benign tumors from malignant[20].

In transcriptomics, the discovery of biomarkers is usually achieved by the differential expression analysis of RNA sequencing data obtained from the comparison of tumor and normal tissue. RNA sequencing has substituted microarray technology, as its potential with a higher dynamic range and unlimited genome coverage has overcome the results used in the traditional technology[21]. The individual changes in each of the genes reveal the diversity between healthy and cancer cells. The gathering of this information into the affected metabolic pathways paints a portrait of the signaling cascade affected by the malignancy of the disease. Other RNA sequencing analyses in cancer research focus on the alterations of nucleotide sequences, like RNA editing[22] or variant calling of small RNA mutations[23]. However, the procedures of mutation calling in RNA sequencing have yet to be standardized[24], creating a lower performance and accuracy than DNA variant calling.

The evolution of transcriptomic technology has become a great tool for diagnosis and prognosis prediction in multiple cancer diseases, replacing imaging techniques and providing a large window of clinical applications[25]. On the other hand, the improvements in RNA sequencing have presented computational challenges in their analysis. The increase in read-length and the depth of RNA sequencing complicated the storage, transmission, quality control, and especially, the analysis of the data by adding a higher degree of complexity[26].

One of the past challenges in RNA sequencing has been the determination of tumor heterogeneity and the creation of an accurate interpretation of tumor diversity. The tumor extraction from some cancer patients itself presents a problem in the differentiation of

tumor tissue and its perfect separation from adjacent normal tissue. These cases also contribute to the complexity of the interpretation of the results in RNA sequencing. As a solution, single-cell RNA sequencing has achieved the interpretation of cancer heterogeneity, successfully revealing the tumor microenvironment and allowing the classification of different cancer cell types within a tumor sample[27,28].

The current challenges in RNA-sequencing are more focused on optimizing the implementation of transcriptomic studies, starting with the standardization of bioinformatic analysis and especially reducing the cost of the whole analysis, of which the approximate cost per patient is €7000[29].

1.1.4 Proteomic characterization of cancer disease.

Although the presence of mutations in genomics and transcriptomics studies adds complexity to the range of the studies, nucleic acids are mostly preserved making it easier to predict the protein translation[30]. Proteomic research, on the other hand, deals with a large dynamic range, generated by splice variants, RNA editing, and the possible combination of multiple post-translational modifications, such as glycosylation, phosphorylation, methylation, etc[31]. Regardless of the complexity of protein studies, proteomics has contributed to oncology by revealing potential drug targets and biomarkers.

The scientific community has tackled the problem of complexity by sharing in databases information about the protein expression and modifications in multiple tissues and diseases. Projects like the human protein atlas have created an interactive website that allows the exploration of protein expression, classifying them by tissue, cell type, or pathology[32]. Other projects have focused on cancer proteomics databases, like the cancer proteome atlas (TCPA)[33], in which current efforts in gathering protein information have obtained around 8000 samples in 32 different cancer types[34].

The analysis of proteomics data has a few similarities with transcriptomics. Protein intensities obtained from mass spectrometry analysis can be used to determine the differential expression between healthy and cancer tissue, as well as, to be used for pathway analysis. A further step is protein-protein interactions, a unique characteristic of

proteomics that can reveal insights into cancer biology, being tandem affinity purification, the most frequent method used[35]. Another similarity with RNA sequencing is the possibility to study variations. Post-transcriptional modifications (PTMs), in the proteomic case, are more challenging to be predicted computationally, however, with the standardization of laboratory techniques, the creation of databases similar to those observed in genomics can be achieved[36].

When proteomics is applied to cancer research, the profiling studies have the potential to discover molecular biomarkers for early diagnosis, prognosis markers, and possible therapeutic targets for cancer treatment[37]. One of the following studies has used mass spectrometry as the central technique to obtain a clear insight into the inner regulation of tumor cells correlating the proteomic information with genomics and transcriptomics.

1.2 A multi-omic perspective on cancer.

Multi-omics by definition, as well as, trans-omics, or pan-omics, combine two or more levels of information aboard this complexity[38]. The comprehensive study of complex diseases requires the interpretation of multiple levels of information such as the genome, transcriptome, or proteome[39], as it can provide a better understanding of the events and changes caused by a disease. The advances in high-throughput methodologies have created an overwhelming amount of single-omics data containing valuable information on the different levels of the cell environment. The increased dimensionality of the data carries with it two main challenges: an increased time of processing of samples, and the integration complexity of multiple layers of heterogeneous information[40].

We will review how genomics transcriptomics and proteomics integrate, explaining the different ways that they can be connected, using state-of-the-art techniques like DNA and RNA sequencing or mass spectrometry, and facing the multiple challenges of this discipline.

1.2.1 DNA and RNA sequencing combination.

Genomics and transcriptomics have been the most extended and developed single-omic analysis techniques, being whole-exome sequencing (WES), whole-genome sequencing (WGS), and RNA-sequencing the tip of the spear in the field[41,42]. The similarities in molecular composition, structure, and sequencing techniques between DNA and RNA sequencing allow much easier integration of the genomics and transcriptomics data.

The combination of these two levels of information empowers the analysis of single nucleotide variants (SNVs)[43]. Mutations caused in the DNA by a disease like cancer can be tracked down to RNA variations with the integration of sequencing techniques. The accuracy of the variant calling is therefore increased as mutations can be validated at the transcriptome level. Variants transcribed to RNA allow determining which mutations avoid the DNA repair mechanisms and will be later coded in the protein sequences, foreseeing the possible structural modification changes produced (**Figure 1.3**).

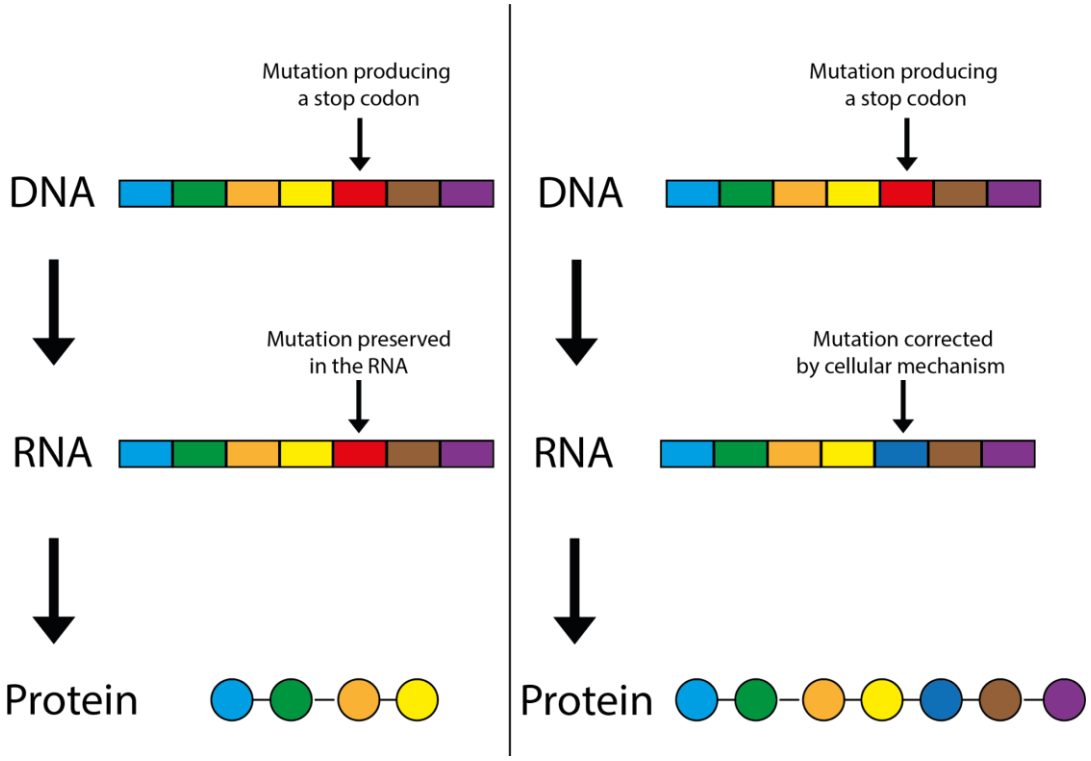


Figure 1.3. Example of a mutation detected in DNA that can be either preserved in the RNA and translated into a truncated protein or corrected by DNA repair mechanisms restoring the amino acid product and the full-length protein.

During the past years, cancer research has applied the combination of DNA and RNA variant calling to study the alterations in multiple patients, revealing information like early drivers of metastasis in breast cancer[44], new forms of prostate cancer[45], or specific gene isoforms in gastric cancer[46].

Another form of variant calling using genomics and transcriptomics that goes further than the detection of single nucleotide variants is the integrated call of gene fusions. Gene fusions are chromosomal rearrangements that result from a joint product of two previously separated genes[47]. The detection of the fused structural variants is usually performed by the analysis of DNA or RNA-sequencing reads, classifying the evidence reads into two types. Junction reads are those that contain nucleotide information from two different genes within the same fragment overlapping the fusion. On the other hand, in pair-end sequencing, spanning reads are read pairs mapping each to a different gene, surrounding the point of fusion[39] (**Figure 1.4**). While previous methods only achieved gene fusion calling using RNA-sequencing[39], novel bioinformatic methods have allowed the combination of both DNA and RNA sequencing mapped files to be checked for gene fusions, obtaining evidence reads on each level and therefore improving the confidence of the variant calling[48]. The advances of this technology have proven that gene fusion, as well as other structural variants, should be taken into consideration as a driver characteristic of diseases like cancer, where they have become a relevant biomarker for cancer diagnosis, prognosis, and therapeutic targets[49–52].

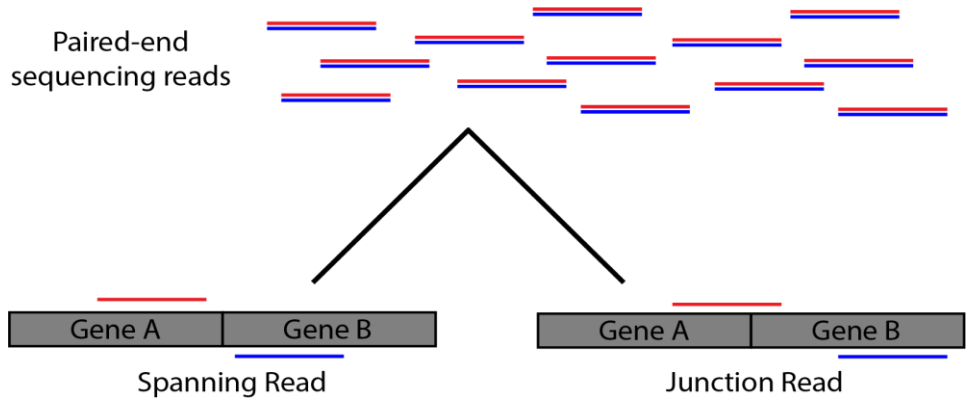


Figure 1.4. Gene fusion detection by DNA or RNA sequencing divides the evidence reads into spanning and junction reads, depending on if each read aligns to a gene or if one read has evidence of the fusion point respectively.

Besides the integration of DNA and RNA for variant calling, there are other forms of combining genomic and transcriptomic sequencing. An example of it, and one of the most common forms of analysis, is the assessment of the effect caused by mutations in gene expression. The exploration of the consequences caused by genetic mutations relies on the accuracy of variant calling in DNA sequencing and its combination with differential expression analysis in RNA sequencing. The bioinformatic process behind it is simple yet effective, having a large impact in cancer research where it has allowed the analysis of cancer-associated gene alterations[53–55]. The most recent studies have taken advantage of all the different combinatory levels mentioned in this section to obtain clinically actionable insights into diseases like lung cancer[56].

1.2.2 Changes in expression between RNA and proteomics.

The next two levels of integration are transcriptomics and proteomics. This combination gives a perspective of the molecular mechanisms of the cell as proteins can regulate metabolic activity and transcript expression[57]. In contrast to the sequencing techniques followed by genomics and transcriptomics, proteomics has multiple experimental procedures to determine protein abundances, such as gel electrophoresis or the wide range of mass spectrometry (MS). Although MS has been established as the most common technique for proteomic analysis, it still presents a huge variety in instrumentalization (LC-MS-MS, MALDI-TOF) and methodologies (labeling or label-free)[58]. To address the problem of multiple strategies in proteomics and transcriptomics a chapter will cover the current state of normalization methods and the combination used for this thesis.

Previous attempts of combining protein abundances and RNA intensities showed a limited correlation between the two levels of expression[59]. Multiple hypotheses have tried to explain the causes behind this phenomenon, like the different half-life between protein and mRNA inside the cell, the variability of mRNA expression levels during the cell cycle, or the experimental errors produced in the analysis procedure[60]. However, progress in the field of multi-omic analysis has been achieved, dealing with the complexity of the

correlation between proteomics and transcriptomics and obtaining valuable information from this type of analysis. In cancer research, multiple studies have proven the value of integrated transcriptomics and proteomics, revealing new insights into the field like potential targets or tumor signatures[61,62].

1.2.3 Proteogenomic integration analysis, current state, and limitations.

The first definition of proteogenomics was given in 2004 to describe a study where proteomic data was used to support genome information[63]. Since then, the term has evolved to a broader ambit where protein information is integrated with genomics and/or transcriptomics, not only providing validation but also improving the overall analysis[64]. The three different levels of information can be grouped into two categories depending on the starting point of the analysis:

1. Top to bottom proteogenomics: Following the events described in the central dogma of molecular biology, top to bottom proteogenomics uses genomics as the first approach followed by transcriptomics and proteomics. The flow of the analysis focuses on the study of changes caused by DNA alterations and their impact on the phenotype. The higher depth of sequencing achieved in GWS studies and its reliability create a perfect starting point for the analysis of diseases associated with or caused by genetic variants.

2. Bottom-up proteogenomics: The opposite case uses proteomics as the initial step of the analysis. Changes in the proteome are evaluated by genomics and transcriptomics interrogating their origin. This type of analysis is ideal when comparing two conditions, to obtain information on the expression changes between them and track down the causes of the alteration.

Independently of the directionality, proteogenomics studies provide a full insight into the events of the cell and its environment, revealing the connections between three different biological levels of data. The principle behind proteogenomics is to perform multiple layers of analysis within the same cohort of patients and take advantage of the correlation between the data types to obtain the biological characteristics of the subjects[65].

The current state of multi-omics can be easily observed in the scientific literature, where multiple studies use this combination to have a clearer picture of diseases like Mendelian disorders[66] and especially in cancer[67–70]. The advantages of the methodology have allowed the discovery of biomarkers and target candidates for new therapies in cancer, establishing this area of knowledge at the cutting edge of the field. Multiple tools have been developed to facilitate the development of multi-omic analysis, being categorized into tools that allow disease subtyping, tools that predict biomarkers or driver genes, and tools that provide an insight into the molecular biology inside a disease[39].

Besides the advantages of proteogenomics integration analysis, these studies are subjected to a meticulous design of the experiments and multiple considerations due to the number of limitations. The first obstacle in a proteogenomic analysis is the design of the experiment, having in mind the number of samples that will be processed, the budget is the first restriction that is faced. Although the cost of NGS techniques has been decreasing during the past years, a combinatory analysis of DNA sequencing, RNA sequencing, and mass spectrometry proteomics for multiple samples can be expensive, multiplying the cost not only for the number of patients but for the number of techniques.

A database approach can be used to perform a cost-effective analysis, but the variability of the methods used over time will add another limitation to the study. The usage of different techniques, the variance added between batches processed in multiple days, by different equipment, or by multiple analysts in different research centers are common phenomena in databases and online repositories[71,72]. This variance will be present in the raw data before the analysis, if the data obtained for the repository is already processed, variability will be added from the usage of different processing pipelines.

To perform a more robust analysis, it is recommended to create a unique pipeline for the analysis of the multiple samples proposed in the experimental design. This election is also limited, not only by the knowledge of prediction models and the selection of suitable software for the analysis but also by the computational power required to achieve the analysis. Once samples are processed, normalization methods have to be applied to remove the batch effects and consistent nomenclature has to be established across the different data types (ENSEMBL gene id, protein IDs, gene names...)[73].

Data interpretation from the proteogenomic analysis is also limited by other factors. The technical artifacts in single-omics analysis add up when scaling to multiple layers, making it difficult to distinguish changes coming from technical variance from those provoked by the disease[74]. The distinction of changes caused by environmental factors is to date one of the major unsolved challenges of multi-omics, besides efforts from the scientific community that has tried to tackle the problem[75,76].

Other limitations encountered in multi-omics originate from the restrictive choice of methods when performing an analysis. One example of this is portrayed in proteogenomic studies, where global MS is usually preferred for the exploration of proteomics. The analysis of mutations in proteomics will be then subject to a global technique where, although evidence of a mutation can be found, the reduced coverage of specific peptide fragments will decrease the possibilities of finding DNA and RNA variants in the proteome[77–79].

2. Aims and objectives.

The common thread of the research presented is the integration of multi-omic data to seek unexplored characteristics of three diseases (Esophageal adenocarcinoma, Undifferentiated pleomorphic sarcoma, and Gorham-Stout syndrome). With a multi-layer approach, the strategy develops around obtaining further insights into the mechanisms of each disease, revealing possible biomarkers and therapeutic targets that otherwise would be missed in single-omic studies.

Although the environment of the three diseases will be diverse due to the variability between cancer types, a common strategy for analysis reveals their characteristics. Our study through the analysis of mutations and other molecular profiling techniques follows the procedures of the latest advances in the field, observing how genomic mutations impact at the multi-omic level[80], the immune landscape of the tumor tissue and their surroundings[81], and the pathways affected by the modification of the internal mechanisms in tumor cells[82].

Genomics, transcriptomics, and mass spectrometry proteomics are the main -omics covered in the study, and for each one of them a shared processing pipeline needs to be developed to obtain data consistency across samples and/or projects. When the data per sample has been processed, either created by the pipelines developed or obtained from external sources, the challenge resides in merging all sample sources, considering all the technical biases that can be produced in the processing techniques, and integrating the multiple levels of information.

The specific aims for each study can be defined into:

1. Examine the changes in RNA/Protein expression in Esophageal adenocarcinoma and explore the effect they produce in the cell environment compared to normal tissue expression.
2. Characterize the genomic landscape of Undifferentiated pleomorphic sarcoma and determine new possible therapies.

3. Interrogate the events occurring in Gorham-Stout disease by the complete examination of genomics, transcriptomics, and proteomics and discover how the integration can reveal new insights into this rare disease.

The thesis has been developed as 3 different projects sharing common methodologies and presented in article format, of which one of them (the Gorham-Stout chapter) has been already published.

3. Esophageal adenocarcinoma.

3.1 Introduction.

Among esophageal cancers, esophageal adenocarcinoma is the predominant subtype in western countries, being the sixth-leading cause of tumor-caused death[83]. The high mortality cause of the disease is attributed to a poor prognosis, with a survival rate between 15-25% after five years[84]. During the past 50 years, esophageal adenocarcinoma incidence has increased drastically with a seven-fold change in the number of cases reported[85], and the growth expectations from 2015 to 2025 is an increase of 140%[86]. The mortality of esophageal adenocarcinoma combined with the increasing incidence of the disease has created the necessity to reveal the mechanisms inside EAC. Most of the current studies try to find early-detection biomarkers or discover new therapeutic targets to improve the 5-year prognosis.

Esophageal adenocarcinoma is characterized by a glandular differentiation that frequently originated from a precursor disease called Barrett's esophagus[87]. This disease creates a change in the esophageal tissue cells that develops them into mucosa cells similar to the ones found in the intestine[88]. Therefore, both EAC and Barrett's esophagus share risk factors like tobacco smoking and obesity[87], with gastroesophageal reflux being the most common cause of the disease[89]. There is also an important male predominance of the disease with a 7:1 ratio, although the causes of the gender-specificity remain unknown[90]. Another clinical characteristic that explains the high mortality of the disease is the common pattern of metastases to the lungs, liver, brain, and especially the lymph nodes[91].

Besides the removal of the tumorigenic tissue through surgical procedures and chemotherapy, EAC treatments have evolved over time to obtain the best long-term outcome for the patients. First, endoscopic therapy appeared as an option to surgical treatments, providing similar results to the previous procedures[92]. Current treatments focus on the use of neoadjuvant therapy, like metformin[93], to improve patient outcomes[94], while the most advanced techniques rely on the use of personalized therapy[95,96]. To perform

a successful evaluation of personalized therapy, a well-characterized landscape of the mutations in EAC tissue has to be provided as a starting point to biomarker discovery. Although genome sequencing studies of EAC have achieved to provide new insights into the disease[97] allowing the discovery of recurrent driver mutations[98] and copy number alterations[99], little to no efforts have characterized the landscape of the disease through a multi-omics integration[39].

Therefore, we have created a joint study with the Oesophageal cancer clinical and molecular stratification (OCCAMS) consortium to cover this lack of information in the literature with the expectation to reveal new therapeutic targets and biomarkers for esophageal adenocarcinoma. Our study gathered DNA-sequencing and RNA-sequencing data from over 300 EAC samples and combined it with mass spectrometry proteomics from a small subset of patients. The main focus of our study was to reveal the limitations of single-omic studies by evaluating the changes produced between RNA and protein expressions. Furthermore, we discovered that the origin of genes presenting drastic changes in expression between the two levels of information are not caused by somatic mutations but rather produced by other regulatory mechanisms of the tumor environment.

3.2 Materials and methods.

3.2.1 Sample collection.

Clinical samples from the esophageal adenocarcinoma, adjacent normal esophagus, and distal normal gastric tissue were collected from patients' biopsies. The tissue extracted was de-identified, frozen in liquid nitrogen, histopathologically reviewed by a pathologist, and stored at -80ºC until processing.

3.2.2 Whole-genome sequencing.

A total of 454 whole-genome sequencing EAC samples along with attached normal esophageal tissue were included in the study. The cohort is composed of samples previously published by Frankell et al.[39] added to 34 new WGS samples. FASTQ files were aligned to

the human reference genome (hg38) using the Burrows-Wheeler Aligner (bwa v0.7.17), Following GATK best practices pipeline, aligned bam files were marked for duplicate reads using Picard Tools (v2.18.23), thereafter, genomic variants were called per sample using GATK Mutect2 (v.4.1.3) and filtered against the variants detected in their respective paired normal esophageal tissue. Germline variants were labeled using the 1000 genomes project[100]. The analysis of the variant calling files across all patients was performed using custom scripts with Maftools in R (version 4.1.0) after the merge and annotation using VCF2MAF[101].

3.2.3 EAC RNA-sequencing data.

As well as in WGS, RNA-sequencing data was gathered with a joint effort with the OCCAMS consortium collecting a total of 281 EAC samples RNA-Seq, of which new 17 samples were added in our study, matching patients from the previously described WGS cohort. FASTQ files were mapped to the hg38 reference genome using STAR (v2.6.1). The resulting BAM files were quantified for transcript expression with RSEM (v1.3.3), obtaining normalized transcript per million (TPM). The remaining 264 EAC RNA-seq samples were obtained from Frankell et al.[39] in the form of expression files reporting FPKM values per gene and patient. The values were converted to TPM to standardize the quantification procedure.

3.2.4 Tissue processing for mass spectrometry.

Fragments of 20 to 30 mg of tissue were lysed in 8M urea buffer (8.5pH and 0.1 M Tris/HCl), combined with protease inhibitor (Merck, Darmstadt, DE), at approximately 5-cell-pellet volumes. Lysis was promoted through rapid freezing using liquid nitrogen, thawed, sonicated on ice, and centrifuged at 14,000 g/30 minutes/4℃. The supernatant was saved, and protein concentration was determined using RC-DC™ Protein Assay (BioRad, CA, USA). 100 µg of protein was loaded to Microcon, Ultracel-30 spin column with 30 kDa cutoff (Millipore, MA, USA) and digested using a Filter Aided Sample Preparation (FASP) protocol[102].

A urea lysis buffer was added to the column and centrifuged at 14,000 g/15 minutes/20°C. Protein reduction was performed using 100 mM Tris (2-carboxyethyl) phosphine hydrochloride (Sigma-Aldrich, MO, USA) in urea buffer for 30 minutes on a thermoblock set at 600 rpm/37°C. The column was again centrifuged at 14,000 g/15 minutes/20°C and free sulfhydryl groups were alkylated using 300 mM iodoacetamide in urea buffer (Sigma-Aldrich, MO, USA). Protein alkylation was held in the dark at 20ºC for 20 minutes followed by another centrifugation at 14,000 g/15 minutes/20°C.

After alkylation, 100 mM ammonium bicarbonate was added to the column, followed by centrifugation at 14,000 g/15 minutes/20°C. Trypsin (Promega, WI, USA) dissolved in 50 mM ammonium bicarbonate buffer was added at a 1:100 enzyme to protein ratio (w/w) and protein was digested overnight at 37°C.

3.2.5 TMT labeling and fractionation.

The peptide concentration was determined using Pierce™ Quantitative Colorimetric Peptide Assay (Thermo Fisher Scientific, MA, USA). An equal peptide concentration from each sample was transferred into the clean low peptide retention tubes. Samples were evaporated on the SpeedVac concentrator (Savant SPD121P, Thermo Scientific, MA, USA) until dry and dissolved in 50 µl 100 mM triethylammonium bicarbonate. TMT label tags were equilibrated to room temperature and resuspended in 41 µl of anhydrous acetonitrile, followed by 5 minutes of occasional vortexing and short spin. Half volume (20.5 µl) of the corresponding TMT label reagent (**Figure 3.1**) was added to each sample. Samples were incubated at room temperature for one hour. The reaction was quenched by 4 µl of 5% hydroxylamine for 15 minutes. Finally, all TMT labeled samples were pooled together into a clean low peptide retention tube and dried at room temperature in the SpeedVac concentrator. Pooled TMT samples intended for direct LC-MS analysis were dissolved in 0.1% (v/v) formic acid (FA) in LC-MS water and desalted using Micro SpinColumns C18 (Harvard Apparatus, MA, USA) packed with C18 sorbent followed by lyophilization in SpeedVac.concentrator. TMT-labeled samples intended for fractionation were dissolved in 300 µl of 0.1% (v/v) trifluoroacetic acid solution. Samples were

fractionated using Pierce™ High pH Reversed-Phase Peptide Fractionation Kit (Thermo Fisher Scientific, MA, USA) according to the manufacturer's instructions. Fractions were evaporated in the SpeedVac concentrator.

3.2.6 HPLC conditions and mobile phases.

The nano chromatographic separation was performed using a nano-RSLC UltiMate system (Thermo Fisher Scientific). For sample loading and desalting, a PepMap C18 Trap-column (Thermo Fisher Scientific) of 300 µm ID x 5 mm length, 5 µm particle size, and 100Å pore size was used. Nano HPLC Separation of labeled and digested proteins was performed using a C18 µPAC separation column (PharmaFluidics, Ghent, Belgium). The dimensions of the used µPAC column were: 5 µm pillar diameter, 2.5 µm inter-pillar distance, 18 µm pillar height; 315 µm bed channel width, and 200 cm column length. The pillars are superficially porous, end-capped with C18-chains.

The aqueous loading mobile phase contained 2% ACN,0.1% TFA, and 0.01% HFBA cooled to 3°C, as described earlier[103]. The mobile phase transferred the sample from the sample loop to the trap column by the loading pump at 30 µl/min to the trapping column, which was operated in the column oven at 50°C. The trapping time was set to 5 min, which was sufficient to load the sample to the trapping column and wash possible salts and contaminants. The cooled mobile phase enabled the trapping of the hydrophilic analytes at enhanced oven temperature. Peptides were separated and analyzed using positive nano HPLC-ESI-MS.

The gradient elution was performed by mixing two mobile phases composed of the following solvents:

- Mobile phase A (MPA): 95% H2O, 5% ACN, 0.1% FA;
- Mobile phase B (MPB): 95% ACN, 5% MeOH, 0.1% FA.

At 240 minutes, the trapping column was switched back into the separation flow and equilibrated for the following run.

The nano HPLC separation was performed on the µPAC separation column with a 2 m separation path. Due to the long separation path, the void volume of these columns

is higher in comparison to conventionally packed nano separation columns. Therefore, the flow rate of the nano HPLC pump was set to 800 nl/min for the first 10 minutes with a gradual lowering to 600 nl/min for separation purposes. The gradual increase of the flow rate at the end of the separation run to 800 nl/min is performed for speeding the equilibration of the separation column. Details on developing and testing different gradients for the 200-cm μPAC column are described by Tóth et al.[104]

Detection of eluting peptides is performed using both UV at 214 nm (3nl UV cell, ThermoFisher Scientific, Germering, Germany) and mass spectrometry (Q-Exactive Plus, ThermoFisher Scientific, Bremen, Germany).

3.2.7 Mass spectrometry.

Mass spectrometry analysis was performed using a Q-Exactive Orbitrap Plus equipped with the Flex nano-ESI source and stainless-steel needle (20 μm ID x 10 μm tip ID). The needle voltage was set to 3.1 kV, scan range was 200-2000 m/z. Full MS resolution was set to 70000, automated gain control (AGC) target to $3x10^6$, and maximum injection time was set to 50 ms. For MS/MS analysis, the mass resolution was set to 35.000, the AGC target to $1x10^5$, and the maximum injection time to 120 ms. The isolation width for MS/MS was set to m/z 1.5, and the top 15 ions were selected for fragmentation, single charged ions and ions bearing a charge higher than +7 were excluded from MS/MS. Dynamic exclusion time was set to 20 seconds.

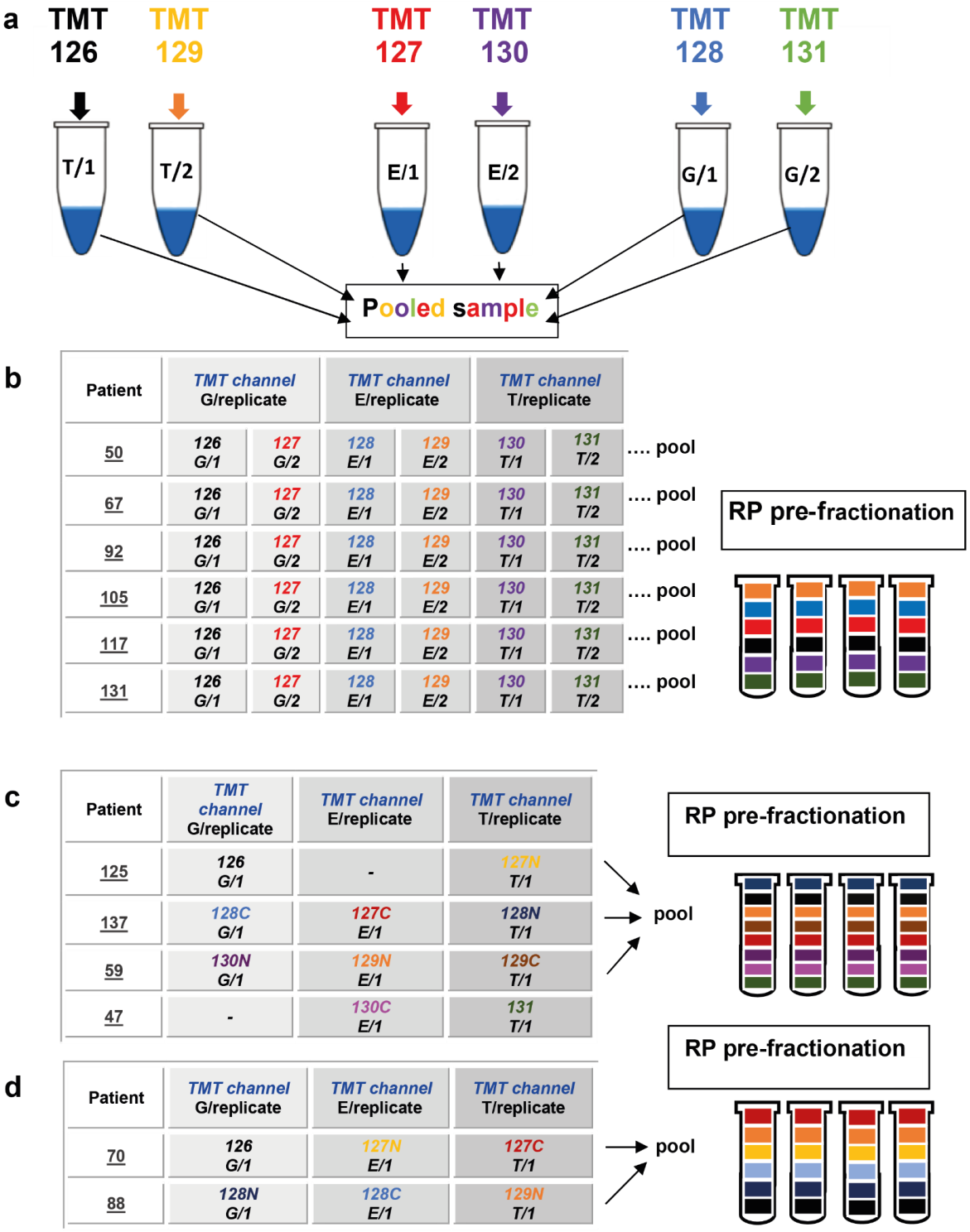


Figure 3.1. TMT sample preparation design and analysis. For each patient, a combination of EAC (T), paired normal esophageal (E), and gastric tissue (G) was processed for mass spectrometry analysis.

3.2.8 Database searching and peptide identification.

A total of 23 EAC patients were analyzed for TMT mass spectrometry, of which 7 were obtained from a previously published article[105] in the form of raw reporter ion data. ThermoFisher RAW files were converted using msconvert[106] to mzML files. Database searches were performed using MS-GF+[107] (v2019.04.18) against a custom database of Ensemble v94 proteins that were augmented with patient-specific variants arising from DNA-Seq[39]. Reverse decoy sequences and common contaminants were then added to the search database. Mass-spectrometry searching was conducted with the following parameters: 10 PPM precursor Tolerance, trypsin digestion with 2 missed cleavages, 6 to 40 peptide length, fixed Cysteine carbamidomethylation, and variable Methionine oxidation. A separate False Discovery Rate (FDR) was set to 1% at both the peptide and protein level (ENSP identifier) using Scavager[107] (v0.1.29). Reporter ion intensities were extracted from the mzML files using the pyopenms[108] package with a 0.01 Dalton tolerance. Peptide intensities were generated by summing the corresponding PSMs intensity values.

3.2.9 Quantitative analysis of proteomes.

Peptide expressions obtained in the previous step were filtered to those peptides mapping to a unique gene ENSG identifier and protein groups containing at least 2 peptides. Although data were obtained from different batches, once the samples were processed, the results were reported in a peptide level expression table with associated intensities for each of the samples.

3.2.10 Normal tissue proteomic and RNA sequencing data.

To facilitate comparison to normal healthy esophageal tissue, we incorporated a large study of normal transcript and protein level data into the study[109]. Furthermore, 32 different normal tissues with matched proteomic and transcriptomic expression, including esophageal, were obtained from the GTEx database[110,111].

3.2.11 Normalization of RNA-seq samples.

A crucial step before the comparison of different batches in experiments or data from multiple sources is the normalization. All the RNA sequencing files used in the study were analyzed by numerous methods using various sequencing companies, therefore requiring a normalization step to be able to compare their expression. We first merged all the expression values reported to the common unit of TPM and then applied a 2-step normalization method based on the trimmed mean of M-values (TMM) followed by quantile normalization (**Supplementary Figure 1**)[112].

3.2.12 Normalization of peptide intensities.

Mass spectrometry data contained a high variability due to the different sources of analysis (**Supplementary Figure 2a)**, thus requiring normalization to perform an effective analysis of the cohort. The normalization method consisted of three normalization steps, starting from reported peptide ion intensities we applied sample loading normalization, TMM, and quantile normalization (**Supplementary Figure 2b)**. Normalized intensities were then merged from peptide level to gene level using ENSG identifiers. To obtain the ENSGs values, peptides mapping to a common identifier were combined using the geometric mean for all technical replicates in each tissue (gastric, EAC, and normal esophageal) per single patient.

MS expression results from normal tissues obtained from external articles[109,110] were processed using the same approach. Being first filtered to peptides mapping to a unique ENSG and then merging peptides with common ENSG using the same geometric mean strategy.

3.2.13 Differential expression of protein intensities.

Differential expression analysis of tumor tissue was investigated using the mass spectrometry analysis of the 23 EAC samples compared to their matching normal esophageal and gastric tissue. To obtain a global expression per tissue across all technical and biological replicates in the 23 samples, we used a meta-analysis principle with a fixed-

effect model, where a weighted mean for each protein group was calculated according to the inverse of the variance in peptide expression[113]. For each of the patients, logarithmic fold change values were calculated using the ratios of protein expression from the tumor against normal esophagus and gastric tissue (TvE and TvG). Welch's modified t-test was used to test the hypothesis that relative protein expression in the TvE and TvG ratios was not different from the protein expression of the technical replicates (TvT, EvE, and GvG). P-values were corrected with the Benjamini-Yekutieli method for multiple hypothesis testing and the significance threshold was set to $p<0.05$.

3.2.14 Immunohistochemistry.

Immunohistochemistry (IHC) staining was performed at the histology research facility at the Institute of Genetics and Cancer, University of Edinburgh. The tissue microarray (TMAs) cores included patient-matched normal esophagus, normal stomach, esophageal adenocarcinoma, normal lymph nodes, and lymph node metastasis where present. TMA blocks were sectioned at a thickness of 5 μm and placed on positively charged slides (Thermo Fisher Scientific) to maximize tissue adherence. IHC was performed using BOND III autostainer with antibodies to *GPA33* (Abcam, ab108938, 1:250), (Sigma, HPA018858, 1:100) or *IGF2BP1* (Sigma, HPA002037, 1:500) incubated for 20 min at room temperature and detected using the Leica Bond Polymer Refine Detection kit (DS9800; Leica Biosystems), following the manufacturer's instructions. Sections representing normal colonic epithelium and normal tonsillar tissue were stained in parallel as positive and negative controls.

Assessment of the IHC intensity staining of TMA cores was performed by two expert histopathologists reaching consensus scores and these samples were graded 0–3 (0 = nil, 1 = weak, 2 = moderate, 3 = strong). Cores with significant artifacts (i.e., folded tissue) or loss of tissue material were excluded.

3.3 Results and discussion.

3.3.1 Differential expression analysis of EAC proteins.

To identify proteins with enriched EAC-specific expression, patient samples representing the primary tumor, adjacent normal esophagus, and stomach were collected from 23 patients. The combinatory analysis of all mass spectrometry samples determined a total of 5897 gene products quantified in at least one patient across the three different tissues (with a minimum of 2 peptides per ENSG group and 1% FDR at the PSM, peptide, and protein level). A comparison between the expression of these genes in each of the patient-matched tissues reveals proteins with enriched expression in EAC compared to surrounding normal tissues (**Figure 3.2**). A previous study identified and validated *EpCAM* as being highly expressed in EAC compared to surrounding normal tissues[105], an event confirmed again by our analysis. Furthermore, we predict similar patterns of enriched expression in EAC for many other novel genes.

Examining proteins quantified in at least 60% of patients and significantly upregulated in EAC compared to both normal esophagus and stomach, several further candidate EAC enriched protein groups were identified. Nineteen members of the RNA binding motif (RBM) protein family were identified, with 4 of them (RBM3, RBM6, RBM25, and RBMX) presenting enriched expression in EAC compared to both normal esophagus and normal stomach. RBM3, in particular, was 2-fold enriched when compared to both normal esophageal and normal gastric, being of particular interest as it has been previously related to cancer. Although RBM3 has been suggested as a potential tumor-suppressive gene whose lower expression is associated with tumor aggressiveness[114], other esophageal adenocarcinoma studies have correlated high expression of RBM3 with intestinal metaplasia[115].

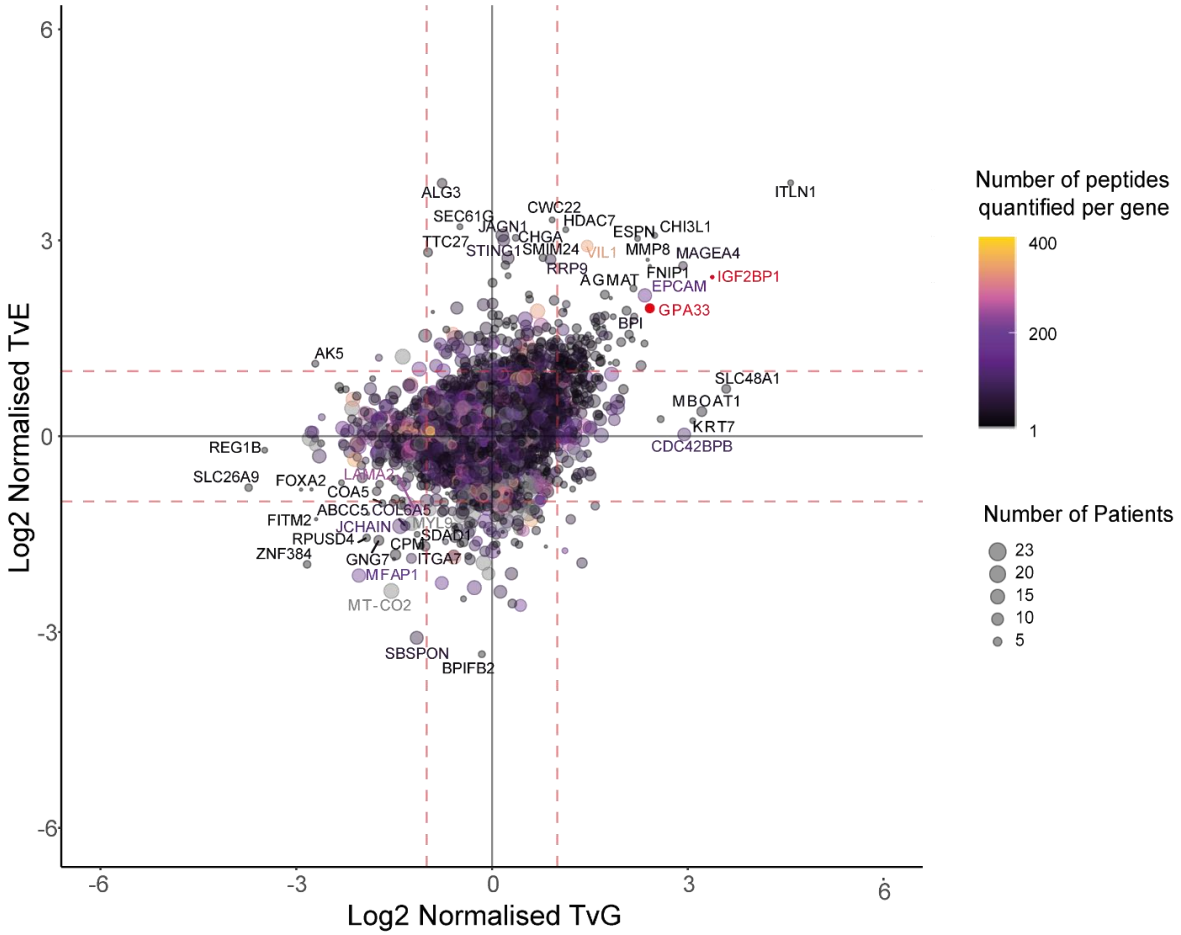


Figure 3.2. The landscape of protein expression in EAC compared to matched normal esophageal and normal gastric tissue in 23 patients. Relative expression of 5507 genes expressed across all tissues in more than one patient. The size of each point indicates the number of patients in which the protein has been quantified and the color represents the number of peptides quantified per protein.

Several cancer/testis antigens were identified as EAC-enriched including melanoma-associated antigen family members *MAGEA4*, *MAGEA10*, *MAGED2*, and *MAGEB2*. This group of genes has a well-established role in other cancers and *MAGEA4* has been previously demonstrated to be overexpressed in esophageal cancer[116]. Due to its tumor-specific expression, *MAGEA4* is a compelling target for immunotherapeutic approaches. Clinical trials are already underway for *MAGEA4*-directed adoptive T-cell therapies for patients with *MAGEA4* expressing esophageal cancer[117,118].

Further cancer/testis antigen, *IGF2BP1*, was identified as EAC-enriched in 10% of our cohort. There is little published literature on the role of this protein in EAC and we, therefore, sought to validate its expression by immunohistochemistry.

3.3.2 Validation of EAC-specific proteins by IHC analysis.

Protein expression was determined in tissue microarrays comprising patient-matched primary esophageal adenocarcinoma, lymph node metastases, uninvolved lymph nodes, normal gastric, and normal esophageal tissue samples, from a cohort of 115 patients in whom 75% had surgery with no prior oncological treatment [105].

Insulin-Like Growth Factor-Binding Protein 1 (*IGF2BP1*) is a member of a conserved family of proteins that have a role in embryogenesis and tissue development and its expression was detected in normal and tumor tissues. A recent study reported higher expression of *IGF2BP1* in colonic tumor tissue relative to their normal counterparts in a set of 13 paired samples[119].

IHC staining showed high levels of *IGF2BP1* protein expression in 12% of the EAC samples whilst such expression was absent in the normal esophageal squamous tissue (**Figure 3.3a**). Furthermore, five EAC cases showed positive *IGF2BP1* whilst their patient-matched normal squamous tissue was negative (**Figure 3.3b, C-50, C-25, C-26, C-33, and C-86**). These observations support the hypothesis presented in the proteomics analyses. Moreover, we found high IGF2BP1 protein expression in cancerous cells in 19% of the metastatic lymph nodes (**Figure 3.3a**). Although around a quarter of uninvolved lymph nodes (25%) and normal gastric mucosa (23%) cases had high *IGF2BP1* protein expression, the staining was confined mainly to a few scattered lymphocytes (T and B cells) (**Supplementary Figure 3b**) or within the gastric glands (**Supplementary Figure 3c**), suggesting that IGF2BP1 tends to be from tumor cells (tumor-specific). However, more than two-thirds (72%) of the EAC cases showed negative protein expression for IGF2BP1 by IHC. Overall, the data suggest that IGF2BP1 protein can be considered to be a moderately tumor-specific biomarker that is also expressed in some normal lymphocytes making it less likely to be useful in a clinical context for identifying EAC metastases.

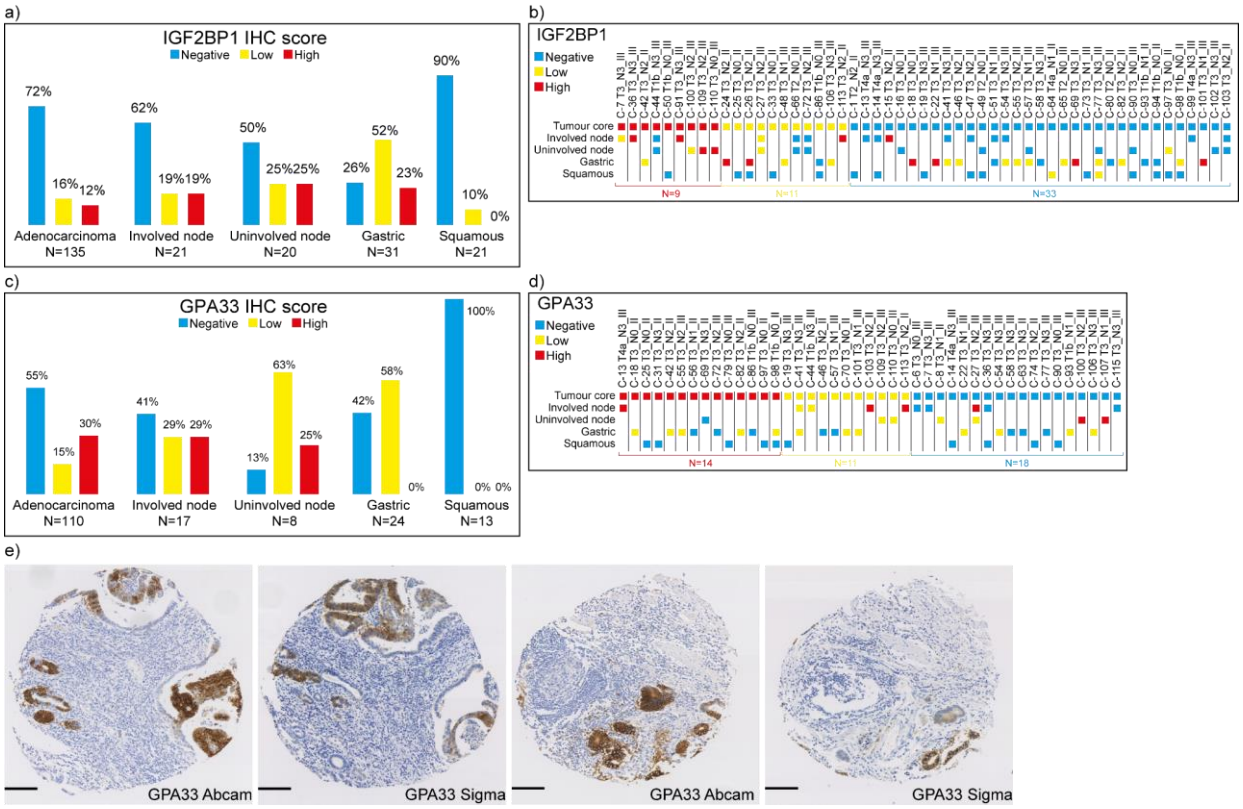


Figure 3.3. Validation of candidate tumor-specific proteins (IGF2BP1 and GPA33) in tissue microarrays contains patient-matched primary esophageal adenocarcinoma (or tumor core), involved lymph node (metastasis), uninvolved lymph nodes, normal gastric, and normal esophageal squamous samples. **a-b) Immunohistochemistry (IHC) scores for IGF2BP1 protein expression.** a) Histoscores according to tissue type. b) A total of 53 patient-matched samples which includes tumor cores and/or involved and uninvolved lymph nodes along with normal gastric and normal esophageal squamous tissues. **c-d) Assessment of GPA33 protein expression.** c) IHC scores according to tissue type. d) A total of 43 patient-matched samples which includes tumor cores and/or involved and uninvolved lymph nodes along with normal gastric and normal esophageal squamous tissues. e) Representative IHC core images showing similar staining patterns of two different anti-GPA33 antibodies as indicated in each one.

Glycoprotein A33 (*GPA33*) is a cell surface antigen and a member of the immunoglobulin superfamily with a suggested role in cell-cell adhesion[120]. *GPA33* is widely expressed in the intestinal mucosa, gastric cancer tissues, and in > 95% of colonic cancers, and resulting from these findings, anti-*GPA33* antibodies were developed as immunotherapy for gastric and colorectal cancers[121–123]. Clinical trials of monoclonal anti-*GPA33* showed good safety and tolerability[124], while other clinical trials investigating novel anti-*GPA33* antibodies in colonic cancers are still reporting their results (NCT02248805).

Analyzing *GPA33* IHC staining showed high levels of GPA33 protein expression in around a third of the EAC samples (30%) (**Figure 3.3c**) which is consistent with our proteomics data that showed increased expression in 8 of the 23 patient samples. In

contrast, only 2 cases across all of the normal tissues (normal lymph node, normal gastric mucosa, and normal esophageal squamous epithelium) showed similarly high *GPA33* expression (**Figure 3.3c**). Remarkably, of the 43 multi-sampled patients we analyzed, there were 14 (33%) showed high GPA33 protein levels in the primary esophageal adenocarcinoma with low or negative expression in metastatic tumor deposits (**Figure 3.3d, red labeled**). Of these 14, just one case showed high expression in metastatic cancer in lymph nodes (**Figure 3.3d, C-13**). On the other hand, there were only 2 cases across all of the normal tissues (normal lymph node, normal gastric mucosa, and normal esophageal squamous epithelium) that showed such high expression of *GPA33* with negative *GPA33* staining in the primary EAC (**Figure 3.3d, C-100, and C-107**). Of note, we further stained a sub-population of the patient cohort (n=62) with another anti-*GPA33* antibody (Sigma) (**Figure 3.3e**) and we found a highly significant correlation (Rho=0.822, p-value <0.001) between the two different antibodies IHC staining patterns, indicating that *GPA33* is a consistent EAC tumor biomarker. Overall, this analysis suggests that *GPA33* IHC detection shows high tumor specificity and consistency.

3.3.3 Protein to RNA expression in matched proteogenomics samples.

We measured the correlation of expression between proteomics and mRNA levels in a subset of seven patients (**Supplementary Figure 4a-g**). The patients whose tumor biopsy passed the quality threshold were designated for matched DNA and RNA sequencing along with MS proteomics. The selected "matched patients" offer an opportunity to explore the event of decoupling RNA and protein abundance to regulate expression levels by exploring protein-to-RNA abundance in EAC.

A total of 5,531 genes were commonly detected in both the RNA and protein abundances across all patients. The number of correlated genes per patient was highly variable, indicating that when an increased number of genes were detected shown the correlation trend improved in comparison with those samples containing fewer genes. In general, and following the principles of the central dogma of molecular biology, high

transcript expression values were leading to increased protein abundances (**Supplementary Figure 4h**).

Among the high expression proteins, we found tumor-specific candidates detected in previous studies, like *GAPDH*[125] or prothymosin alpha (*PTMA*)[126], as well as protein families like the Eukaryotic Translation Elongation Factors (*EEF1*)[127] or the keratin gene family (*KRT15*)[128]. The tumor-specific gene expression detected in our results matches previous hypotheses in the literature, indicating the continued applicability of our research.

3.3.4 Disproportionate protein to RNA expression in EAC.

The combined RNA and protein approach herein provides an opportunity to understand the dysregulation of protein to RNA abundance in EAC, which could lead to the discovery of new mechanisms of tumor suppression or oncogene over-expression previously undocumented. Our interests focused on identifying genes with unbalanced levels of expression, either presenting low protein intensities with uncharacteristically high levels of RNA expression or the opposite case. These groups of genes might not drag the attention when performing single-omics by themselves and therefore might be missed by individual analysis.

Among the genes with unbalanced levels of protein to RNA expression, *RHNO1*, *CHFR*, and *CENPE* stand out as proteins present in high protein abundance despite low transcript levels (**Supplementary Figure 4h**). *RHNO1* has been recently reported as a prognostic marker in colorectal cancer[129], renal, and liver cancer, where the high expression of the gene ends in an unfavorable prognosis for the patient[130]. Therefore, *RHNO1* could be a possible prognostic marker for esophageal adenocarcinoma contributing to the survival analysis of EAC patients. Similarly, *CHFR*, a gene that is known to be downregulated or silenced in esophageal adenocarcinoma at the RNA level[131,132] stands out, again having extreme protein abundance despite low levels of RNA. Another gene following this pattern of expression is *CENPE*, a gene associated with other types of adenocarcinoma[133] and prognostic potential at the genomic level for esophageal adenocarcinoma[134]. Our hypothesis for genes with the mentioned characteristics aims for a potential inside mechanism of the

diseases that escapes the RNA regulation and compensates it with an overexpression of the protein.

On the other hand, we have explored disproportionately expressed genes with high levels of RNA and medium to low protein abundances. Most of them have been previously reported to be involved in other cancer diseases, either as oncogenic, inducing tumor cell proliferation, or causing poor prognosis (*ACTG1*[135,136], *PPIA*[137], *GNAS*[138], *BTF3*[139], or *YBX1*[140,141]). A pattern of expression was also observed for mitochondrial genes (*MT-CO3*, *MT-ND3*, or *MT-ND4*) that, although might not have the same impact as the oncogenic genes, might indicate an affected pathway in EAC.

3.3.5 Matched patients as an adequate representative of global EAC Protein to RNA expression changes.

We have examined how the integration of protein and RNA expression has followed the general trends of correlation between the two levels of expression in our matched esophageal patients. To evaluate the representativeness of our matched patients with a more general cohort of samples we combined the totality of the RNA and MS samples described in our study as global analysis of the Protein to RNA expression changes in EAC. Genes presenting changes between the Protein and RNA expression in the seven matched EAC patients (**Figure 3.4a**) were labeled and used as guidance for the analysis, comparing how their expression changed across the global tumor samples. The RNA and protein correlation in the matched patients generalized with equal distribution in the larger unmatched cohort containing a global analysis of all the available transcriptomics and proteomics samples from EAC patients (**Figure 3.4b**).

Genes presenting increased RNA-seq and decreased protein expression remained in the same Protein to RNA ratio levels when comparing the two EAC cohorts, as well as most genes with high protein expression and low RNA-seq intensities (**Figure 3.4c**). In general, genes presenting discrepancies in the Protein to RNA abundances showed a low coefficient of variation, indicating consistency across all the EAC samples independently of the

matched or unmatched origin. Besides this, a small proportion of genes with high protein abundances and low RNA-seq expression do not follow the same trend, tending to have changes in protein expression when comparing the general cohort to the matched dataset. The changes in the expression originated at the protein level are accompanied by a high coefficient of variation between samples. Genes like NUBP1 or AKR1C1, presented heterogeneous abundances across patients, suggesting high inter-tumor heterogeneity.

3.3.6 Protein to RNA expression changes in normal esophageal tissue.

The changes in expression between protein and RNA abundances and how they are regulated as part of normal tissue homeostasis are poorly described in the literature[142], as well as in esophageal cancer[143]. However, both protein degradation and synthesis pathways can be tightly regulated[144]. The former through post-translational modifications of proteins targeting them towards proteasomal degradation and the latter arises through tight post-transcriptional regulation of the epitranscriptome which decorrelates RNA and protein abundances. All aspects of gene expression are dysregulated in cancer, for example through mutation, copy number variation, or epigenetic modification. Dysregulation of carefully controlled protein to RNA ratios through post-translational or post-transcriptional pathways is likely another exploitable mechanism.

There have been two significant efforts to study RNA and protein abundances in normal tissue: GTEx[110,111] and Kuster et al.[109]. To compare how the dysregulated genes found in EAC relate to normal esophageal samples, we have used these studies as a backdrop for the study of Protein to RNA ratios against our matched and general cohort of EAC patients.

The combined study of Protein to RNA expression of all datasets (matched patients, unmatched patients, GTEx data, and Kuster et al. data), correlated 2977 genes found in common across all studies (**Figure 3.4**). An analysis of the correlated Protein and RNA expressions between the previously explored EAC (**Figures 3.4a and 3.4b**) and the normal tissues from the Kuster et al. article and the GTEx database (**Figures 3.4d and 3.4e**) shows a clear shift in both expression levels for most of the dysregulated candidate genes.

The characteristic changes in expression point to some dysregulated genes in EAC as new possible tumor-specific candidates presenting Protein to RNA abundances that differed between normal tissues and the EAC tumors (**Figure 3.4f**). Some of the candidate genes (KRT5, ANXA1, SPRR3, S100A8, or S100A9) showed a decreased RNA expression in both of the EAC datasets when compared to normal esophageal tissue while staying at the same protein expression levels.

In contrast, other dysregulated genes become unlinked from the correlation when comparing normal esophageal expression against EAC tissue, increasing the protein abundances while presenting similar transcript expression levels. TAOK2, MAPKAPK3, or HIGD2A are perfect examples of tumor-specific candidate genes with those characteristics. The increased activity at the protein level in EAC tissue could be indicative of a way to regulate tumor oncogenic pathways through Protein to RNA abundance dysregulation, raising the expression of proteins associated with cancer hallmarks. For example, TAOK2 and MAPKAPK3 would be upregulated kinases with links to the p38/MAPK activation pathway.

3.3.7 Tissue specificity of Protein to RNA expressions changes.

Further study of tissue specificity was performed by using a combinatory examination of the Protein to RNA expression between the two esophageal adenocarcinoma cohorts previously described and 32 other normal tissues available from the literature[110,111]. Candidate genes with high variation between protein and RNA abundances in matched EAC samples were compared, for protein to RNA ratios, against the global EAC cohort and the collection of normal tissues. To further interrogate the origin of EAC and its relationship to Barrett's esophagus, marker genes from undifferentiated Barrett's tissue published by Nowicki-Osuch et al.[145] were labeled and included in the examination. Among all these candidates, only genes presenting differences in the protein to RNA ratios between EAC and other normal tissues were selected (**Figure 3.5**).

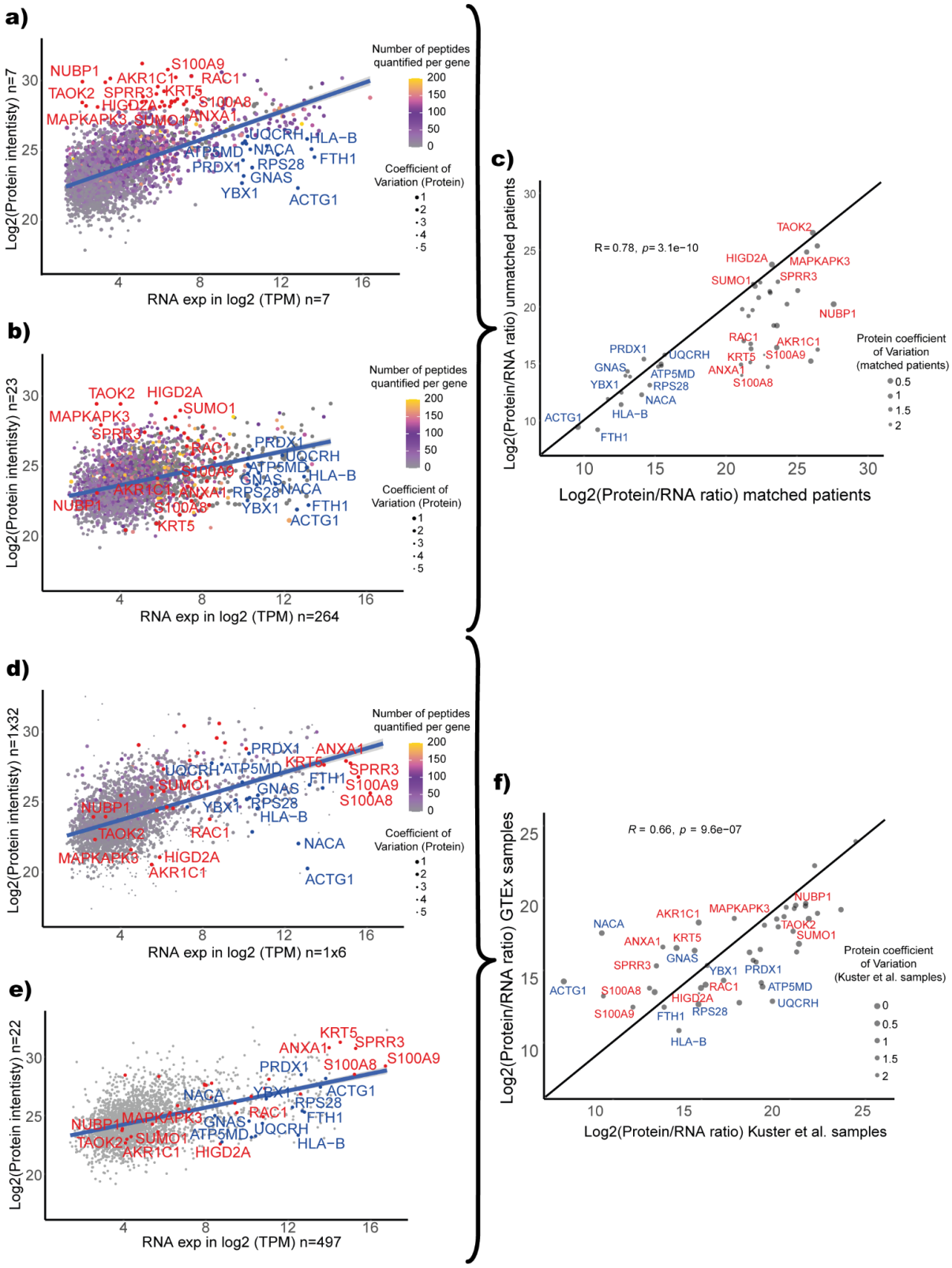


Figure 3.4. Protein to RNA correlation of 2977 genes common in EAC and normal esophageal. (a) Correlation of RNA and Protein expression from 7 EAC patients with matched proteogenomics. (b) Correlation of RNA and Protein expression from all EAC samples (23 proteomics and 264 transcriptomics) (c) Correlation of Protein to RNA ratios of selected candidate genes in both EAC cohorts. (d) Correlation of RNA and Protein expression from Kuster et al. normal esophageal tissue. (e) Correlation of RNA and Protein expression of esophageal mucosa tissue from the GTEx database. (f) Correlation of Protein to RNA ratios in both normal esophageal studies for the selected candidate genes.

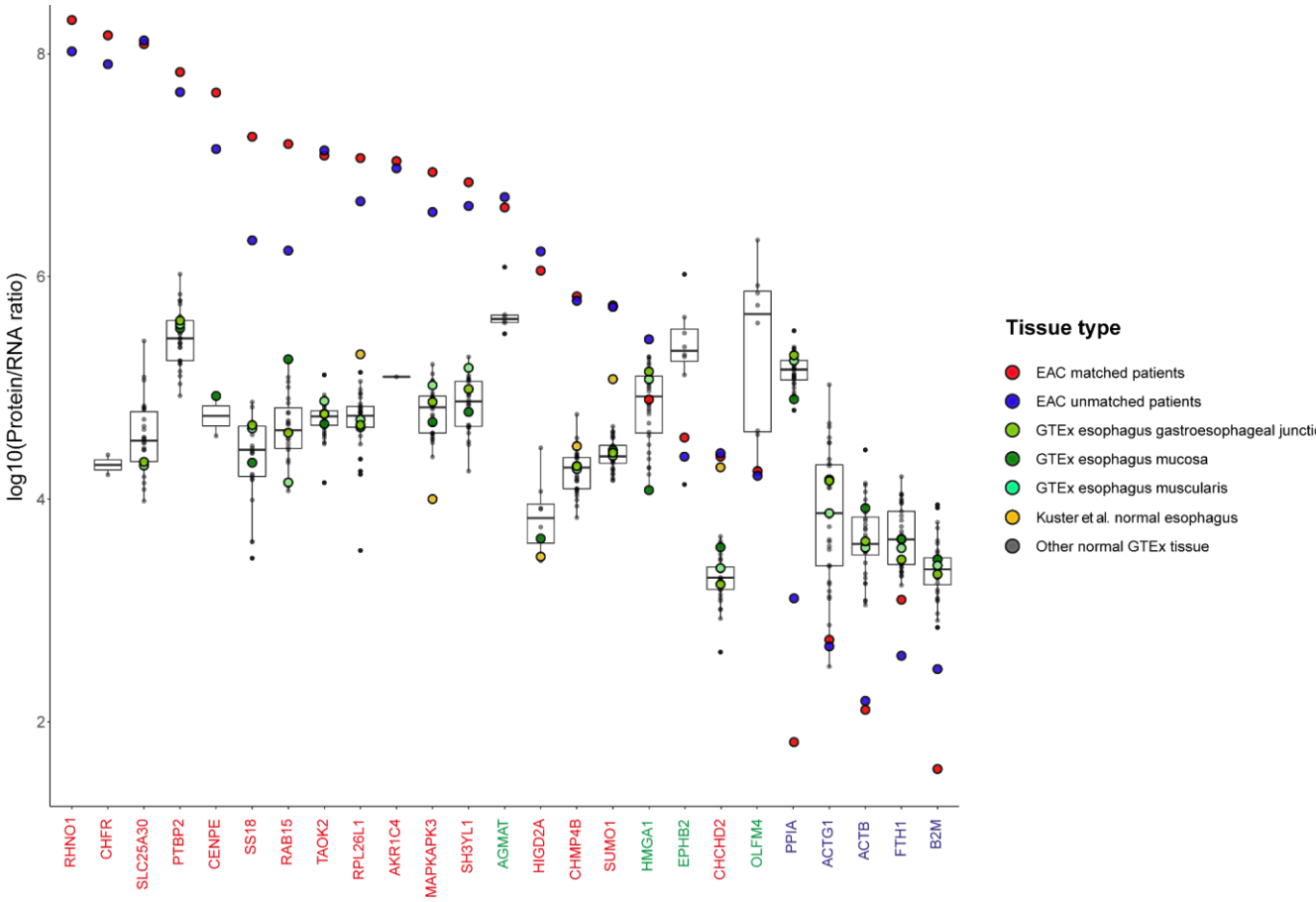


Figure 3.5. Scatter plot of the candidate genes presenting a disproportionate Protein to RNA ratio, either with high protein and low RNA expression (red) or the opposite (blue). Genes labeled in green were marked as undifferentiated Barrett's esophagus. For each gene, we compared ratios from EAC-matched patients (red dots) and from the global EAC cohort (blue dots) against a variety of normal (distribution plot) containing esophagus samples from Kuster et al. (yellow dots), three different esophageal tissues from the GTEx database (light green, green, and turquoise dots), and other normal GTEx tissues (gray dots).

Candidate genes like RHNO1, CHFR, or AKR1C4 were detected in EAC while not being reported in normal esophageal or other normal tissues from the GTEx database. The protein to RNA ratio for cases like RHNO1 is complemented with low expression at the RNA level, where a comparison between EAC and normal tissues shows a decreased expression in the tumor sample. Despite the low mRNA expression. protein expression was only detected in tumor samples, producing a high Protein RNA ratio. The results suggest that genes containing low RNA expression and high protein intensities have an inner mechanism that produces this dysregulation event, being for some of them, specific to esophageal adenocarcinoma.

Most of the candidates reported (**Figure 3.5**) were found in normal esophageal or other normal tissue in the GTEx database, allowing a comparison of the results. Genes whose protein to RNA ratio in EAC differs drastically from normal samples were especially of interest as they suggest a variation in either the protein, RNA, or both levels of expression. Inside this group, genes containing a low ratio discrepancy between matched proteogenomic samples and the global EAC cohort revealed strong tumor-specific candidates. Genes within those characteristics are MAPKAPK3 and TAOK, members of the p38 MAPK signaling pathway which is known to be dysregulated in cancer by showing a variety of post-transcriptional modifications which are currently used as an attractive target for tumor treatments[146,147].

The MAPK signaling pathway seems to be heavily affected in EAC. In addition to TAOK and MAPKAPK3, AGMAT has been recently reported to promote tumorigenesis by using this signaling cascade. Agmatinase was labeled as one of the genes from Barrett's undifferentiated tissue[145] that presented a significant difference between EAC patients and other normal tissues. Furthermore, AGMAT was previously reported in the differential expression analysis of EAC proteins as one of the genes with increased fold-change values in tumors when compared to matched normal esophagus and gastric tissue (**Figure 3.2**). The combination of the analysis suggests AGMAT as one of the oncogenic drivers of EAC, presenting an increased protein product while keeping the RNA expression levels similar to normal tissue.

Genes showing a high difference in the protein to RNA expression ratios between EAC samples and normal tissues open the opportunity for new therapeutic targets like AGMAT and the other members of the p38 MAPK pathway. Another example can be found in SLC25A30 and PTBP2, the genes presenting the higher difference in the protein to RNA ratio. Both of these genes are implicated in mitochondrial functions and were identified in cancer as possible biomarkers correlating to metastasis and poor prognosis of the patients[148,149], making them good candidates for early diagnosis and therapeutic targets.

3.3.8 Mutation analysis of candidate genes.

Although usually the expression changes observed in cancer are attributed to aberrations in the genome or transcriptome, we hypothesized that dysregulation between the RNA and Protein intensities are caused by other oncogenic mechanisms using the machinery of the cell. To check the veracity of our hypothesis, we checked the mutations detected on the previously described candidate genes using both patient-specific data from our matched EAC patients and a larger cohort of DNA sequencing samples[39]. The seven patients from the matched cohort showed a lack of mutations in the DNA for all of the candidate genes. On the other hand, a total of 144 mutations were detected in the global cohort of 454 EAC patients across all candidate genes. Further analysis revealed that all the mutations were detected in only 26 patients, where some of the candidate genes were not mutated in the whole dataset. The most commonly mutated candidate genes across the cohort affected only 3% of the population (**Figure 3.6**), entrenching the theory that gene alterations are not the driver of the expression changes.

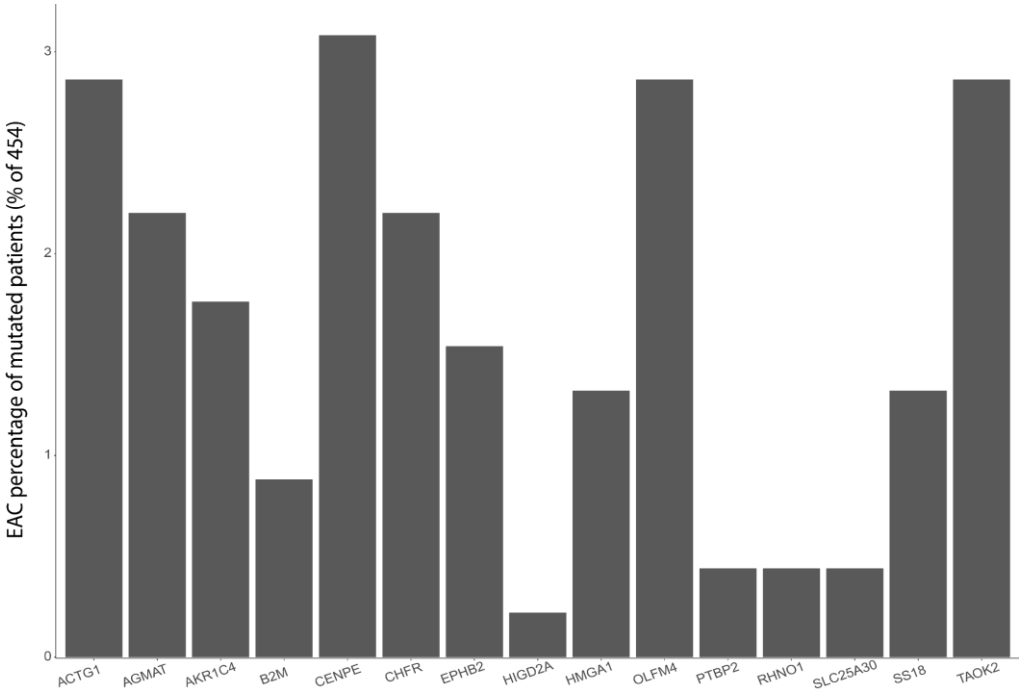


Figure 3.6. Mutational bar plot of candidate genes presenting Protein to RNA dysregulated ratios. The cohort is composed of 454 whole-genome sequencing EAC samples. The candidate genes present a low mutation percentage, not reaching even a 3% of mutated patients for most of the genes. The results suggest the alterations observed between the specific Protein to RNA expression changes are not caused by genomic alterations.

3.4 Conclusion.

The combination of mass spectrometry and RNA differential expressions, the correlation of RNA and protein abundances, and the further exploration of the genes involved by DNA mutation analysis has created a complete proteogenomics insight into the cellular environment in esophageal adenocarcinoma.

The study of tumor-specific proteins in EAC matched previously published results[150], showing EPCAM as one of the most differentiated proteins expressed in EAC when compared to surrounding normal tissues, showing this difference for 80% of the 23 patients. Besides being presented in a fewer number of patients, GPA33 and IGF2BP1 showed similar tumor-specific protein expression. The specificity of these two genes combined with the high expression levels, arise them as possible protein biomarkers for EAC.

PTBP2, CHCHD2, and CHFR were proteins detected whose expression was previously reported in other cancer tissues[149,151,152], being possible candidates for metastasis and poor prognosis in esophageal patients. Many of the EAC-specific proteins detected are associated with intestinal tissue expression or goblet cell phenotype (INTL1, VIL1, OLFM4, REG4, ANXA13) which may reflect the origin of EAC from glandular metaplasia of the esophagus.

The combined analysis of protein intensities and RNA expression in EAC patients revealed a small number of genes containing discordance between the two levels of expression. Genes showing a disassociation in protein and RNA expression presented concordance when the samples for the study were increased in numbers, showing that the small cohort of matched patients is a good representation of the molecular events inside the EAC tissue. When the dysregulated genes in EAC tissue were compared to normal esophageal, it was observed that a few candidate genes shifted to more proportionate ratios of expression. Genes in the esophageal normal tissue that showed proportionate protein and RNA levels, but presented a change in expression in EAC, are tumor-specific candidates possibly modified by the inner regulatory mechanism of the disease.

The comparison of genes presenting disproportionate ratios of expression in EAC against the library of normal tissues showed two pathways strongly affected by candidates presenting differences in expression ratios between tumor and non-tumor samples. The first pathway is related to mitochondrial functions that seem to be altered by genes like SLC25A30 and PTBP2, known to promote the proliferation of kidney cancer[148] and glioma[153] respectively. Both of these genes are known to be involved in auto-regulatory mechanisms by splicing factors[151] or long non-coding RNA[154], which strengthens our hypothesis that the origin of Protein to RNA expression changes derives from alternative mechanisms rather than gene mutations. Other genes presenting high differences in Protein to RNA ratios between tumor and normal samples are TAOK, MAPKAPK3, and AGMAT, clear contributors to the modification of the p38/MAPK signaling cascade. AGMAT also showed a high tumor-specificity in the differential expression analysis of protein analytes, raising the gene as a possible biomarker for the disease.

Overall, the study of esophageal adenocarcinoma has provided a complete proteogenomic exploration combining DNA and RNA sequencing with mass spectrometry proteomics. The unique nature of the analysis has revealed new insights into the mechanisms of the diseases, indicating possible gene markers and altered pathways for the development of new therapies.

4. Undifferentiated pleomorphic sarcoma.

4.1 Introduction

Sarcomas are relatively uncommon cancers representing approximately 0.5-1% of the human cancers[155]. The complexity of sarcomas is defined by the multiple subtypes that can emerge from diverse tissues such as blood vessels, nerves, muscle, fat, ligaments, tendons, and joints[156], creating over fifty genetically distinct subtypes. Among all the sarcoma subtypes, the most common in adults is pleomorphic high-grade soft-tissue sarcoma (UPS)[157]. UPS is defined as a tumor of mesenchymal origin with no identifiable line of differentiation. The so-called "pleomorphic" appearance may be the de-differentiation endpoint of sarcomas in any of the sub-groups that arise from different tissue types such as muscle, fat, or cartilage[158].

Limited by the relative rarity, the small market entailed, and the diversity of the sarcoma subtypes, few novel treatments have been able to improve patient outcomes. The genetic heterogeneity reduces the likelihood of finding "common" sarcoma-type-specific drugs that target rare driver mutations. Current standard treatments generally include doxorubicin alone or in combination with other drugs[159]. However, the survival of patients with metastatic cancer treated by such anti-cancer therapies is just over one year[160].

Although previous genetic analysis using targeted RNA-seq has failed to reveal any "highly penetrant" tumor-specific aberrations in UPS[161], more recent studies based on genome-wide approaches have shown recurrent mutations in TP53, RB1, and CDKN2A[162,163]. Furthermore, certain hallmarks of cancer have been described in soft tissue sarcomas, such as the activation of replicative senescence through telomerase mutations in TERT ATRX and DAXX or the avoidance of the immune system by the upregulation of immune checkpoints of PD-L1 and CTLA4[164].

We define the expressed oncogenic mutational landscape of UPS using a proteogenomics approach and identify potential options to inform future therapeutic strategies. This involves (i) Next-generation exome sequencing of 20 UPS sarcoma DNAs

and their matched normal exomes to define somatic mutations including single nucleotide variants (SNV), small insertions or deletions (INDELS), and copy number variants (CNVs); (ii) Deep sequencing of the T-cell receptor variable region in a subset of these specimens to detect the presence of physiological somatic variation in the t-cell receptors, and to use this to quantify the immune infiltrate clonality; and (iii) tumor-specific RNA-seq and mass spectrometric based-proteomics on a subset of cases to identify potential patient-specific neoantigens. Together these approaches have allowed us to compile a map of the expressed oncogenic landscape of UPS and to propose therapeutic strategies for improved treatment of patients with UPS.

4.2 Materials and methods

4.2.1 Sequencing and processing of DNA.

To identify the most-common genetic alterations that define high-grade undifferentiated pleomorphic sarcoma (UPS), we performed whole-exome sequencing on twenty patients containing tumor-normal tissue pairs. One tumor sample was divided in half to measure intra-tumor heterogeneity providing a total of 21 tumor samples from twenty patients. For each of the histological types of high-grade soft tissue sarcomas, histologically normal adjacent tissue was chosen to define the non-tumor germline. We used normal tissue distant from the tumor in the same surgical specimen to provide the patient-specific control reference DNA database.

Exome Sequencing was performed using Agilent V5+UTR Exome Capture Kit (75Mb) and 100bp paired-end reads were acquired until completing coverage of 100x for tumor and 30x for normal tissue samples. Paired de-multiplexed fastq files were generated using CASAVA software (Illumina) and initial quality control was performed using MultiQC[165]. Paired de-multiplexed fastq files were aligned to the GRCh38 human genome reference assembly with BWA 0.7.9a[166]. Duplicate reads were marked with Picard MarkDuplicates 1.102 (http://broadinstitute.github.io/picard). Following GATK best practices for calling somatic variants, single nucleotide variants and indels were identified

using MuTect2 [167]over the target capture regions (Agilent SureSelect All Exon v5+UTR, including 100 bp padding). Variants were filtered using SNPsift for a variant allele frequency (VAF) greater than 0.03 and a tumor alternative read depth of greater than 5. Additionally, the variant allele must have been observed more than once on both strands. Functional consequences of Somatic Variants were predicted using Ensembl Variant Effect Predictor (VEP) and converted into Mutation Annotation Format (MAF) format using the vcf2maf tool (https://github.com/mskcc/vcf2maf, version 1.6.10).

4.2.2 Copy number variants.

We next evaluated whether UPS displays common genomic regions of amplification or deletion that would identify genetic drivers of this cancer type. Tumor and normal variants identified by the GATK Mutect2 were filtered by removing indels and keeping SNVs with a minimum coverage of 15 and Non-Reference (ALT).

4.2.3 RNA sequencing.

Paired de-multiplexed fastq files from RNA-Seq libraries were mapped to the hg38 reference genome using STAR (v2.6.1). The resulting BAM files were quantified for transcript expression with RSEM (v1.3.3), obtaining normalized transcript per million (TPM) per sample. Differential expression analysis of tab-delimited text files was performed in the R computing environment (version 4.1.0 for Windows). Transcript-wise mapping matrices were created to summarize the mapping results in columns of TPM values per sample. Further analysis was performed in the R package Limma (release 3.14[168]) using the multidimensional scaling (MDS) tool for explorative analysis and the different available statistical tools for the identification of differentially affected genes.

4.2.4 Sample preparation for SWATH-MS.

Approximately, 5 µg of sarcoma tissue were lysed in 300 µl of Urea buffer (8 M Urea, 0.1 M Tris/HCl, pH 8.5) to retrieve tissue protein lysates. The protein concentration in tissue protein lysates was determined using micro-BCA (Thermo, MA, USA) according

to the manufacturer's instructions. Sample preparation for bottom-up mass spectrometry was performed on 10 kDa filters (Millipore, MA, USA, P/N: MRCPRT010) following a modified protocol inspired by Filter-Aided Sample Preparation (FASP) (Wisniewski et al., 2009). Briefly, a volume of protein extract corresponding to 100 µg of protein was mixed with 200 µl of Urea buffer. Filters were then centrifuged at 14,000 g for 15 min at 20°C. Proteins were reduced by 100 µl of Urea buffer with 0.1 M Tris(2-carboxyethyl) phosphine hydrochloride (TCEP) added to the filter followed by centrifugation at 14,000 g for 15 min at 20°C. Protein alkylation was performed by 100 µl Urea buffer with 50 mM iodoacetamide. After incubation on the filter in a thermomixer at 600 rpm for 30 min at 37°C, the liquid was removed from the filter by centrifugation at 14,000 g for 15 min at 20°C. Urea buffer was exchanged for ammonium bicarbonate buffer by adding 100 µl of 0.1 M ammonium bicarbonate and then centrifuged at 14,000 g for 15 min at 20°C. The step was repeated twice. Following, 100 µl of ammonium bicarbonate buffer (50 mM ammonium bicarbonate in water) was added to the filter together with 3.3 µl of 1 µg/µl trypsin in water (Promega, WI, USA). Tryptic protein digestion was held overnight at 37°C in the incubator. Filters were placed to clean tubes and centrifuged at 14,000 g for 15 min at 20°C to retrieve tryptic peptides.

4.2.5 Peptide desalting.

Wash the C18 column (Micro Spin, MA, USA, Harvard apparatus) with 200 µl of 0.1% (v/v) formic acid (FA) in LC-MS acetonitrile (AcN), then centrifuge at 300 g for 3 min at room temperature. Repeat the step twice. Hydrate the C18 column with 200 µl of 0.1% (v/v) formic acid (FA) in LC-MS water then centrifuge at 700 g for 2 min at room temperature. Let the C18 column incubate with 200 µl of 0.1% (v/v) formic acid (FA) in LC-MS water for 15 min then centrifuged at 700 g for 2 min at room temperature. Load the sample to the column and then centrifuge at 700 g for 2 min at room temperature. Three times wash the sample with 200 µl of 0.1% (v/v) formic acid (FA) in LC-MS water then centrifuge at 700 g for 2 min at room temperature. Replace the collecting tube with a new one. Elute the desalted peptides with an increasing percentage of ACN as follows. First

elution: 200 mL 50% (v/v) ACN with 0.1% (v/v) FA, second elution: 200 mL 80% (v/v) ACN with 0.1% (v/v) FA and third elution 200 mL 100% (v/v) ACN with 0.1% (v/v) FA, each elution followed by centrifugation at 700 g for 2 min at room temperature. Eluates were pooled and evaporated in Speed-Vac to the dryness.

4.2.6 Mass spectrometry.

Sarcoma tryptic peptide samples were measured using data dependent (DDA) and data independent (DIA) mass spectrometry acquisition under identical liquid chromatography conditions into 8 fractions on TripleTOF5600+.

4.2.7 Sample dissolving and liquid chromatography separation.

Tryptic digests were dissolved in 100 μl of loading buffer (5 % acetonitrile (ACN), 0.05 % TFA in the water) and analyzed by a pipeline inspired by Dias et al. (10.1016/j.isci.2021.102878). Samples were thoroughly vortexed and 5 min sonicated on the sonication bath. Nanodrop (Thermo, MA, USA) was used to measure peptide concentration and approximately 2 μg of peptides were loaded onto a chromatographic column. Eksigent nanoLC 400 (SCIEX, Canada) coupled to a TripleTOF 5600+ mass spectrometer (SCIEX, Canada) was used to separate the peptides using reverse phase chromatography (RPLC). Peptides were loaded on a cartridge trap column (300 μm i.d. × 5 mm) packed with C18 PepMap100 sorbent with 5 μm particle size (Thermo Scientific, MA, USA) using a 5 μl/min flow of loading buffer. Peptides were eluted to a capillary emitter column (75 μm i.d. × 250 mm, fused-silica) (New Objective, MA, USA) in-house packed with ProntoSIL C18 AQ 3 μm beads (Bischoff Analysentechnik GmbH, Germany). Peptides were separated using a linear gradient of mobile phase A (0.1 % (v/v) formic acid (FA) in water) and mobile phase B (0.1 % (v/v) FA in ACN) with a constant flow of 300 nl/min. Peptide separation started at 5% mobile phase B followed by its linear increase up to 40% B in 120 min. Separated peptides were ionized in a nano-electrospray ion source with 2.65 kV at the capillary emitter.

4.2.8 SWATH acquisition.

Measurement of each sarcoma tissue peptide digest was repeated three times to yield three technical replicates. SWATH-MS data acquisition was operated in high sensitivity positive mode. Precursor masses falling within a precursor range of 400 Da up to 1200 Da were included in the experiment. The selected precursor range was divided into 67 precursor SWATH windows with a constant width of 12 Da. Overlaps between each consecutive SWATH windows were set to 1 Da. MS/MS spectra were scanned from 360 to 1360 Da. MS/MS signal from each SWATH window was accumulated for 50.9 msec. Precursor ions were fragmented using Rolling collision energy with a 15-mV collision energy spread setting.

4.2.9 Spectral library generation.

The spectral library measurement and data analysis were inspired as described in Herranz et al. (PMID: 30951861) and Faktor et al. (DOI: 10.14735/amko20164S54). Data-dependent mass spectrometry was used to generate spectral library files from each of the sarcoma tissue samples. DDA method operated in a positive mode with a precursor range from 400 Da up to 1250 Da. MS/MS range was set from 200 Da up to 1600 Da. A method was set to fragment the top 20 most intense precursor ions excluding them once measured for 12 sec. The resulting DDA method had a 2.3 sec cycle time.

DDA data were searched against a Homo sapiens search database (Uniprot entries) concatenated with a decoy database in ProteinPilot 4.5 software (AB-SCIEX, Canada). Search engine settings were as follows: enzyme-trypsin, fixed modifications – carbamidomethyl, none of the variable modifications were emphasized. MS and MS/MS mass tolerances were set to predefined search settings for TripleTOF 5600+. The FDR calculation was performed in the decoy database. A spectral library was generated from the results. GROUP file in SWATH™ Acquisition MicroApp 1.0 a plugin for PeakView 1.2.0.3. (AB-SCIEX). The spectral library was built from proteins identified at FDR<1 %. Up to 4 no modified and no miscleaved peptides with at least 99% peptide confidence per protein were included in a spectral library, and the rest of the peptides were suspended from

the experiment. A maximum of 6 transitions per peptide were included in the spectral library.

4.2.10. Quantitative SWATH-MS data extraction and statistical analysis.

Protein quantitation was done via the AB-SCIEX SWATH data analysis pipeline. Briefly, SWATH™ Acquisition MicroApp 1.0 plugin running under PeakView 1.2.0.3 (AB-SCIEX) extracted SWATH data for transitions listed in the spectral library. MS/MS signal of product ions was extracted ± 4 min to the left and right from the peptide retention time indexed in the spectral library. Product ion extraction mass accuracy was set to ± 0.05 Da around the expected m/z. Product ion peak areas per each transition were integrated into extracted product ion chromatograms. MarkerView 1.2.1.1. (AB-SCIEX) software summarised peptide and protein intensities. Protein intensities were normalized on total ion current (Sum of all protein intensity). Statistical analysis was performed using a pairwise t-test in MarkerView software. The confidence levels of detected relative differences (foldchanges) in a protein expression level among compared conditions are expressed as p-value. The false rate discovery rate of the assay was cured by a recalculation of p-values to adjusted p-values using Benjamini-Hochberg correction.

4.3 Results and discussion.

4.3.1 The landscape of cancer-specific single nucleotide variants in UPS

We evaluated the mutational spectrum in UPS to determine whether commonly mutated genes could be identified and whether additional therapeutic approaches could be derived from genomic data. A summary of the variant type, variant classification, and the number of variants per sample is shown in **Figure 4.1**. The signature C>T dominates in the single nucleotide variants (**Figure 4.1C**). The number of non-synonymous tumor-specific mutations per sample at the cut-offs used ranges from approximately 20 through to over 200 with a median of 43.5 (**Figure 4.1D**). There are likely more non-synonymous mutations within the tumor, especially if it is heterogeneous. However, with the depth

(coverage) in the sequencing reads (100x), we focus here only on the dominating high confidence single nucleotide variants. The most frequently mutated genes are highlighted as a function of the number of patient samples with mutations (**Figure 4.1F**). These include, of known therapeutic interest, p53, ATRX, and ELMO1.

The most commonly mutated gene with single nucleotide substitutions was p53 occurring in 6 patients (**Figure 4.2**) along with deletion of 17p.1 in other cancers highlighting p53 as a commonly mutated target (**Figure 4.3**). The second most commonly mutated gene was ATRX (**Figure 4.2**), occurring in six patients. The position of mutations and the coding change is shown in **Figure 4.4.** The mutations reside within the kinase domain. As frameshifts are observed, the data suggest that the mutation might create a loss of function or dominant-negative effector, presumably impacting protein-protein interactions in the ATRX protein life-cycle. ATRX is a relatively large gene and the significance of these is often minimized due to the increased mutation frequency as a function of size. Nevertheless, ATRX function as a chromatin-modifying protein could be analyzed in the future as a potential target pathway to evaluate in stratified UPS patients. One prior study highlighted the co-mutation of p53 and ATRX genes in pediatric adrenocortical tumours[169]. However, in our UPS samples, the mutation of ATRX and p53 appears mutually exclusive (**Figure 4.2**). The remaining commonly mutated genes occurred in 2-3 patients out of twenty thus reflecting the usual frequency of mutation (e.g., 5-15%) of genes in cancer cohorts. The only known, possibly "activating" oncogenic mutation with a potential druggable pathway is ELMO1 (**Figure 4.2**), which is mutated in 3 out of twenty patients. Gain-of-function mutations in ELMO1 occur in 6% of patients with Oesophageal Adenocarcinoma and are thought to activate RAC1-dependent migration or invasion[98]. Indeed, ELMO1 is implicated in metastasis in rhabdomyosarcoma[170]. These data together provide three potential therapeutic target pathways; p53, ATRX, and ELMO (RAC). The low penetrance of commonly mutated genes has been noted in many other cancer types raising the problem of how cancer genetics can find common solutions across many patients for therapeutics based on cancer gene sequencing.

4.3.2 Inter-tumor heterogeneity of Mutational Signatures in UPS.

We observed four different UPS-signature mutational patterns within the twenty patients (**Figure 4.5A**). Individual UPS-signature mapped against all 20 cancers is highlighted in **Figure 4.5B**. The UPS-signature class 1, predominating in C>T mutations, is associated with aging processes due to DNA polymerase errors accumulated at repeated cell divisions and is one of the most frequent signatures observed across many cancer types. Tumors 94 and 55 have more than half of their mutation pattern composed of UPS-signature 1 (**Figure 4.5B**). While UPS signatures 2 and 3 also predominate in C>T mutations they also contain C>A and T>C mutations respectively, relatively frequent as well in many cancer types. Tumors 100364, 141343, 090244, 97a, and 141430 have more than half of their mutation pattern comprised by UPS-signature 2 (**Figure 4.5B**), whilst tumors 66, 070052, 080258, 100378, 080107, 100297, have more than half of their mutation pattern comprised by UPS-signature 3. UPS-signature 4 corresponds to the less frequently observed T>G mutation signature. UPS signature 4 dominates in tumors 74, 60, 59, and 84 that have more than half of their mutation pattern composed of the T/G subclass (**Figure 4.5B**). These data indicate that most UPS samples have all four signatures present within each sample, with the penetration variable so that, for example, tumor 100297 shows almost all mutations can be attributed to UPS-Signature 3, whilst 84 is dominated by UPS-Signature 4 and tumor 55 by UPS-Signature 1 (**Figure 4.5B**).

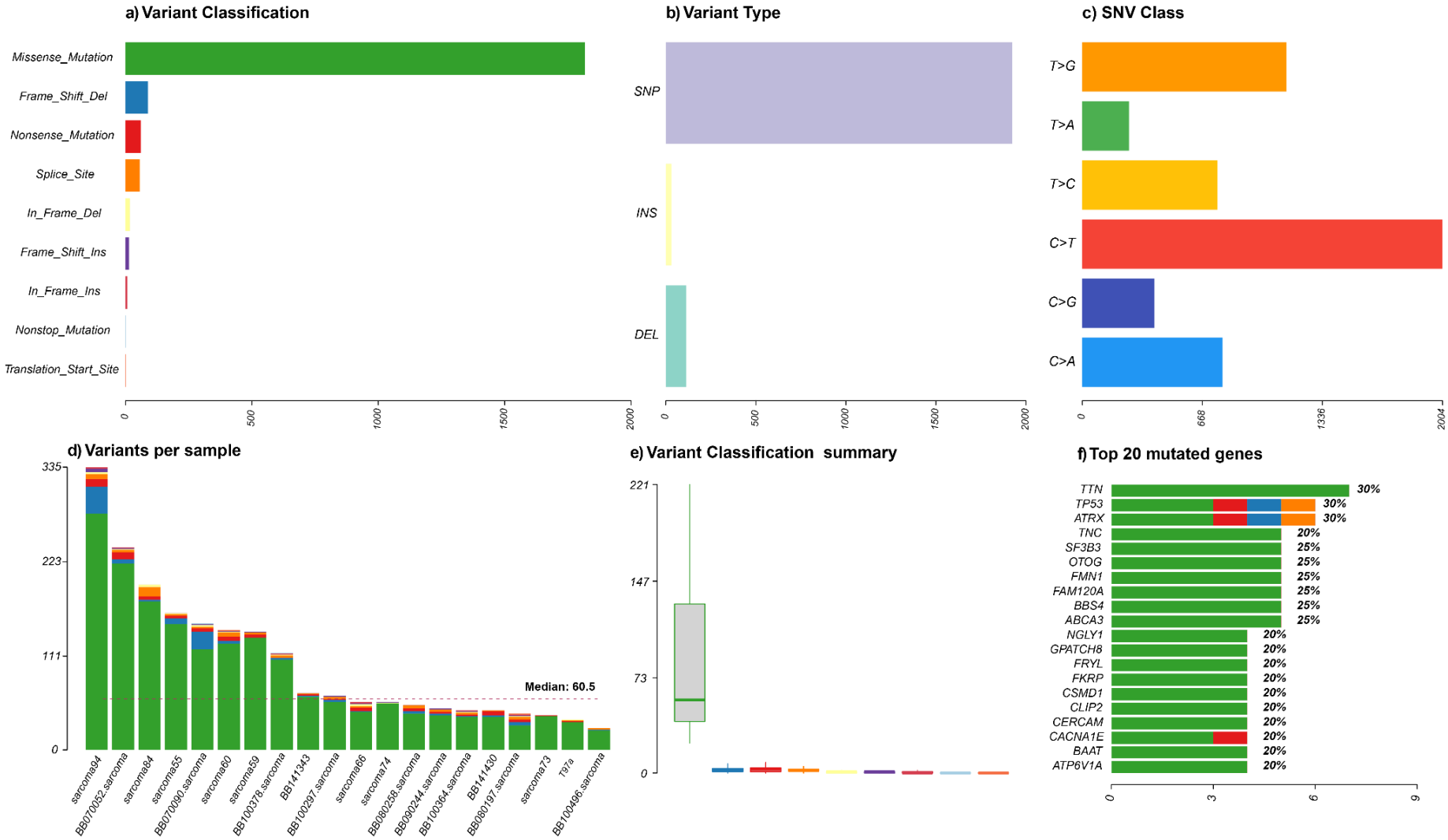


Figure 4.1. WGS variant analysis of 21 UPS samples. **a)** Variant classification of the mutations detected. **b)** Types of variants classified in single nucleotide polymorphisms (SNP), insertion (INS), and deletions (DEL). **c)** Cumulative bar plot of nucleotide substitutions. **d)** Number of variants detected per sample. **e)** Distribution of variant types across all samples. **f)** Top 20 mutated genes across all samples.

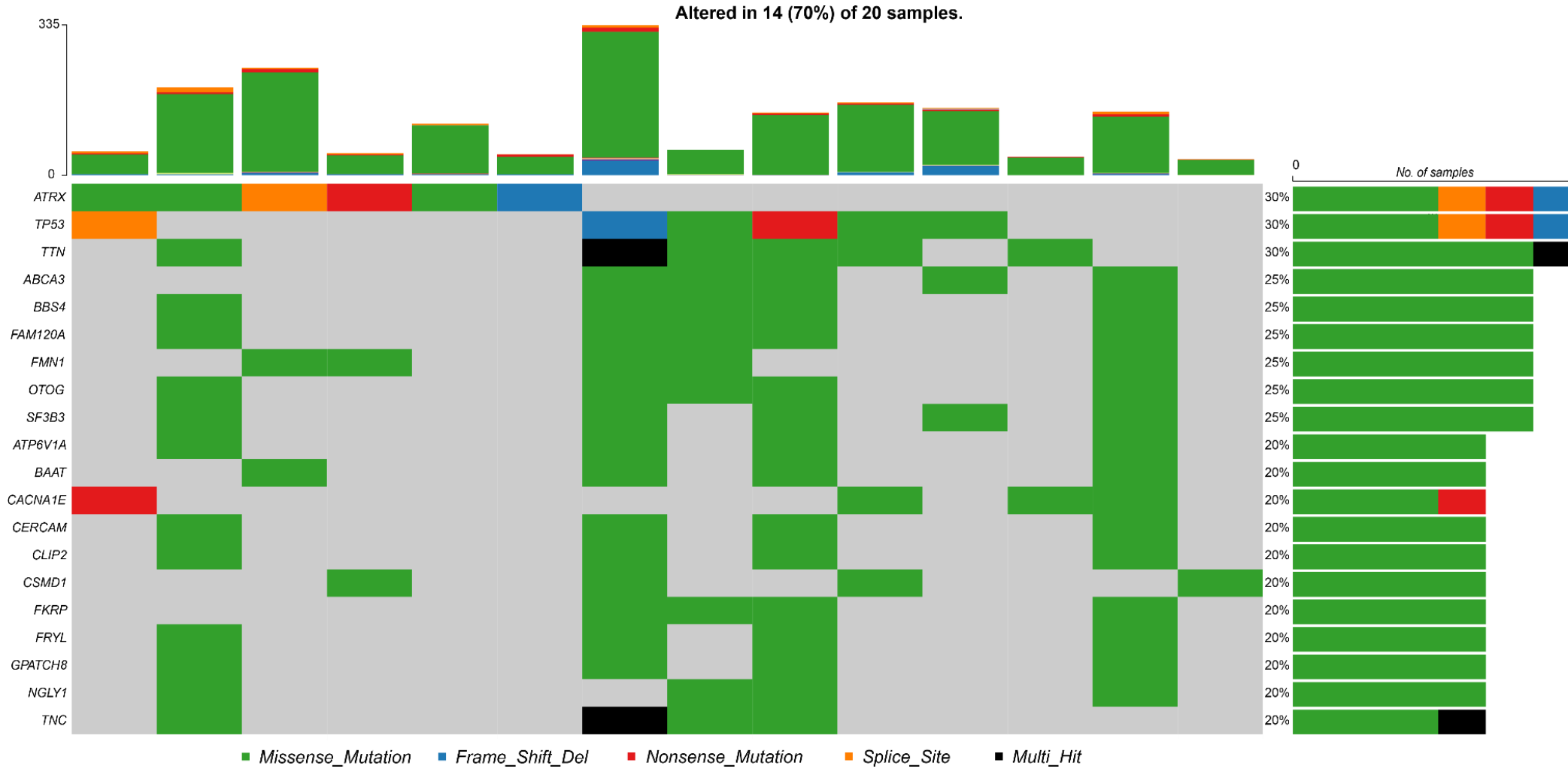


Figure 4.2. Oncoplot of the top 20 mutated genes across all samples. The heatmap represents the type of variant detected in each sample for each of the genes with a bar plot on the top showing the number of total mutations detected in the sample. A cumulative bar plot shows the total number of variants detected across all samples with the type of variant identifier on the right side.

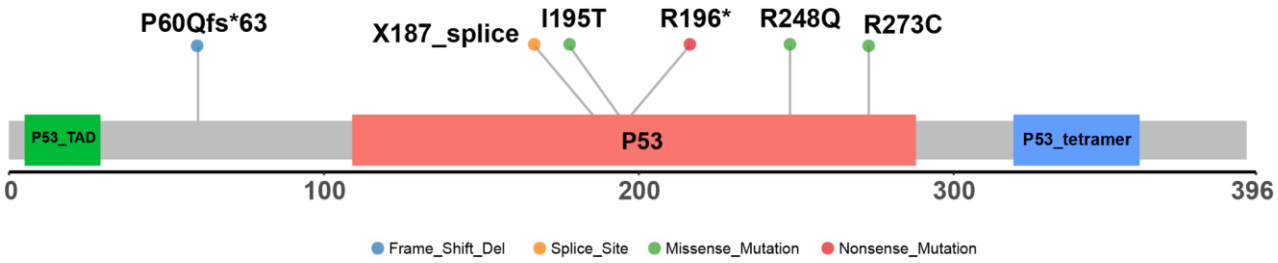


Figure 4.3. Lollipop plot of TP53 showing the SNPs detected across all patients, the position of the mutation, and the type of variant caused.

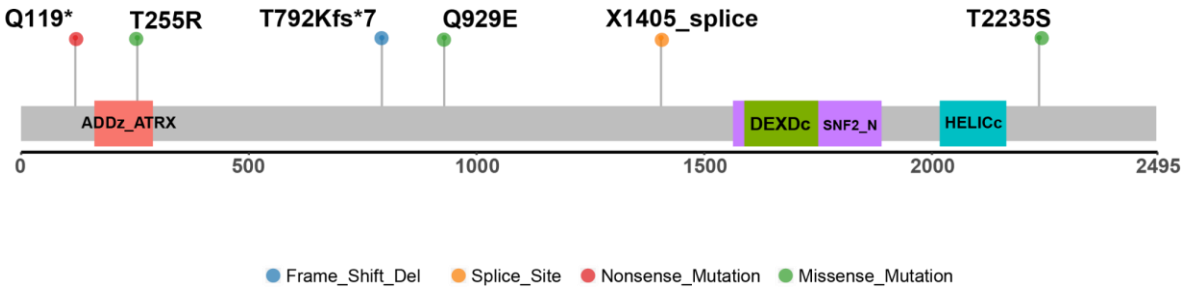


Figure 4.4. Lollipop plot of ATRX showing the SNPs detected across all patients, the position of the mutation, and the type of variant caused.

4.3.3 Intra-tumor heterogeneity of somatic mutations in UPS

Detailed analysis of tumor mutations through ultra "deep" sequencing at high read-depth to detect rare mutations, has revealed a striking level of heterogeneity in tumors[171]. Such tumors are thought to include sub-clones that can evade standard therapeutics[172]. Genomically distinct subpopulations of cells in a cancer biopsy can show differences in mutant variants as defined by the fraction of DNA sequencing reads that harbor a mutated allele[173]. To assess the tumor heterogeneity in all twenty tumors, MATH (Mutant-Allele Tumor Heterogeneity) scores were calculated[174]. The MATH score is calculated as 100x median absolute deviation (MAD)/median of the variant allele frequencies and describes the ratio of the width of the data to the center of the distribution among tumor-specific mutated loci. A homogenous tumor will have a narrower distribution of mutant-allele fractions amongst loci, centered at a lower fraction, than a heterogeneous tumor. Thus, a wider width of distribution will define enhanced diversity among loci arising from more heterogeneous cell populations.

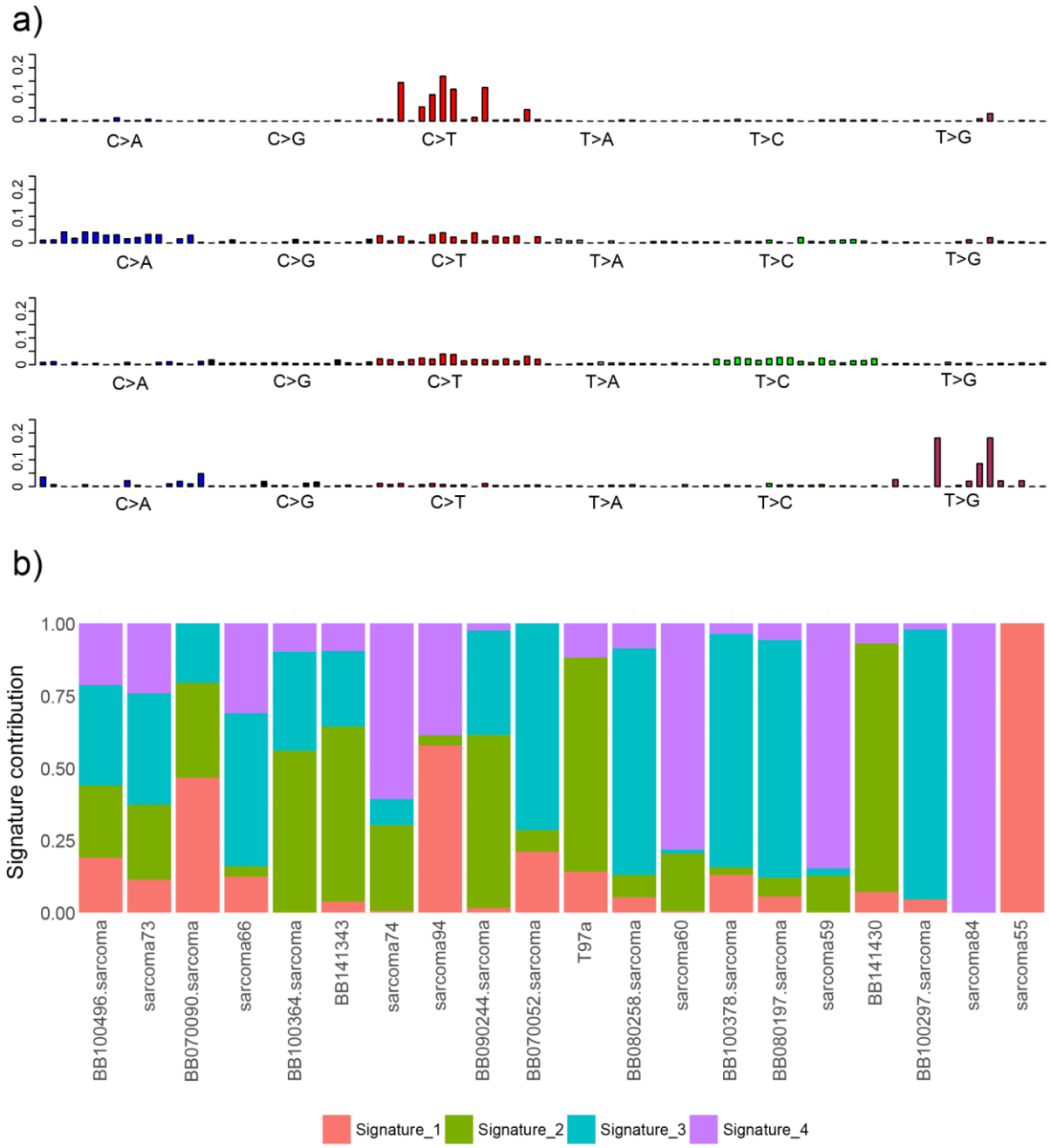


Figure 4.5. a) Probability bar plot of mutational signatures in UPS. The X-axis represents the type of nucleotide substitution for each of the samples represented as a bar plot where the Y-axis is the probability of that mutation to occur. Four main mutational signatures were observed in the 20 UPS patients. **b)** Cumulative bar plot of the contribution of each mutational signature to each one of the UPS samples.

The variant allele frequencies (alternate depth/total depth) of all twenty tumor samples are plotted in **Figure 4.6A**. When applied to the exome sequencing data of the twenty UPS tumor/normal pairs (**Figure 4.6B**), the MATH score average was 50.80 (ranging from 22.57 to 92.99). Example plots of density vs variant allele frequency for samples 55 or 94 are shown in **Supplementary Figures 5A and 5B**. Tumor 55 is unique

amongst the 20 tumor samples in containing a high frequency of variants approaching 1.0 suggestive of a tumor with a high degree of haploidy (**Supplementary Figure 5A**). Near-haploidy states have been observed previously in UPS[175] and leukemia[176]. Tumor 94 highlights a more representative MATH profile in which there is an enrichment in different mutant allele fractions below 0.5 suggestive of distinct, dominating cell populations (heterogeneity) (**Supplementary Figure 5B**). The enrichment of one mutant allele fraction density approaching 0.5 in Tumor 94 is suggestive of shared variants within cell populations of the heterogeneous tumor. Finally, although intratumor heterogeneity and mutation rate are different concepts, it is possible that tumors with high mutation rates would simply have greater intratumor heterogeneity. This was not the case. Altogether, the data suggest a high degree of intra-tumor tumor heterogeneity in UPS.

4.3.4 Copy number alterations in UPS reveal high-frequency dual loss of RB1 and p53 loci.

We next evaluated whether UPS displays common genomic copy number variations that would identify genetic drivers of this cancer type. The most significantly amplified region (in 32% of samples) was on chromosome 1q21.2. There are 14 genes in this cluster including Histones and 2 additional genes embedded within this region, BOLA1, and FCGR1A. Carcinosarcomas were previously observed to amplify 1q22, which contains the HIST2H gene cluster[177]. Quantitative SWATH proteomics of tumor 55 (**Supplementary File 1**) reveals a highly significant elevation of production of Histone H1 paralogues, followed by Histone 2 paralogues. These data nevertheless suggest a common genetic event is the amplification of the Histone H2 loci.

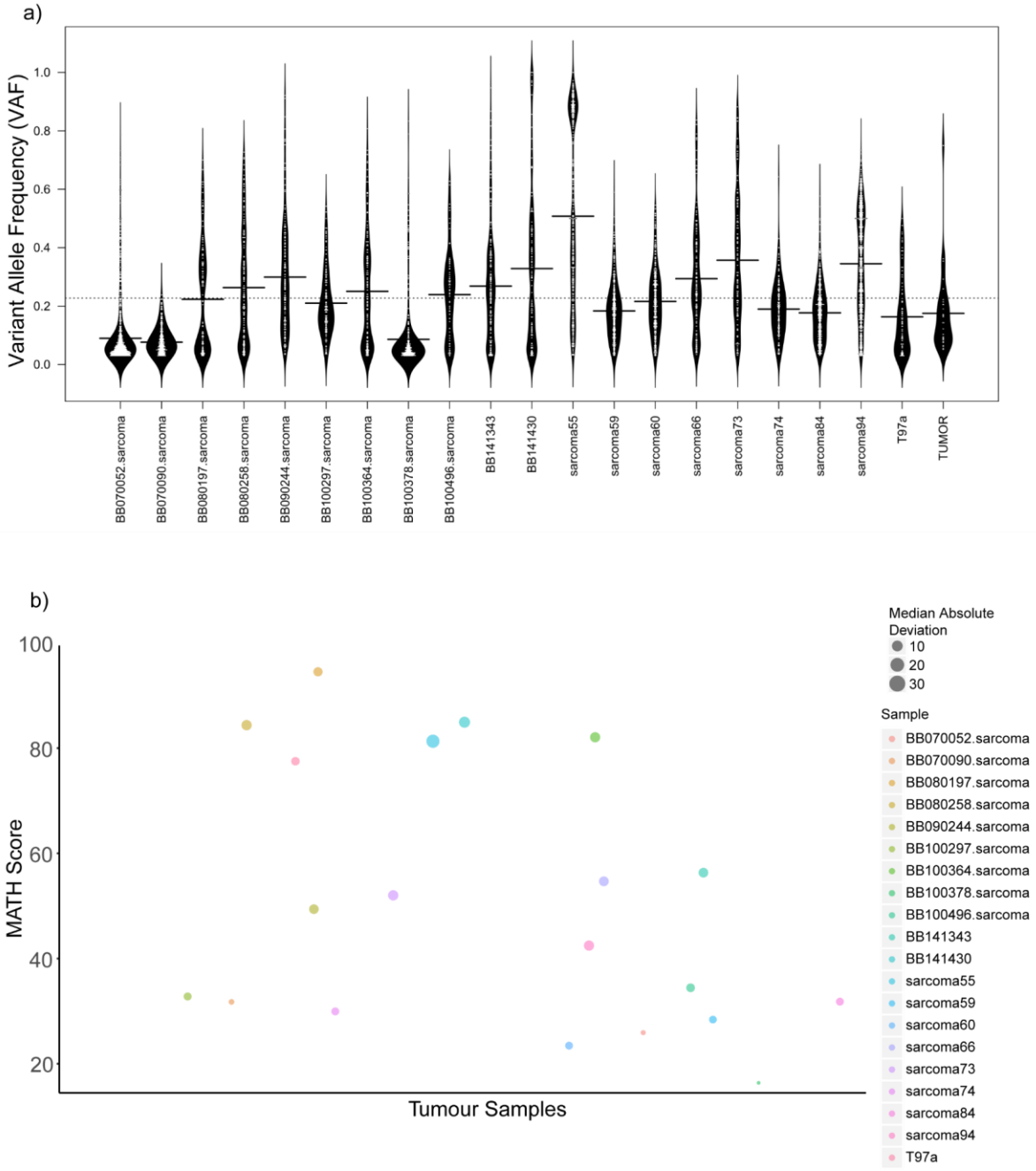


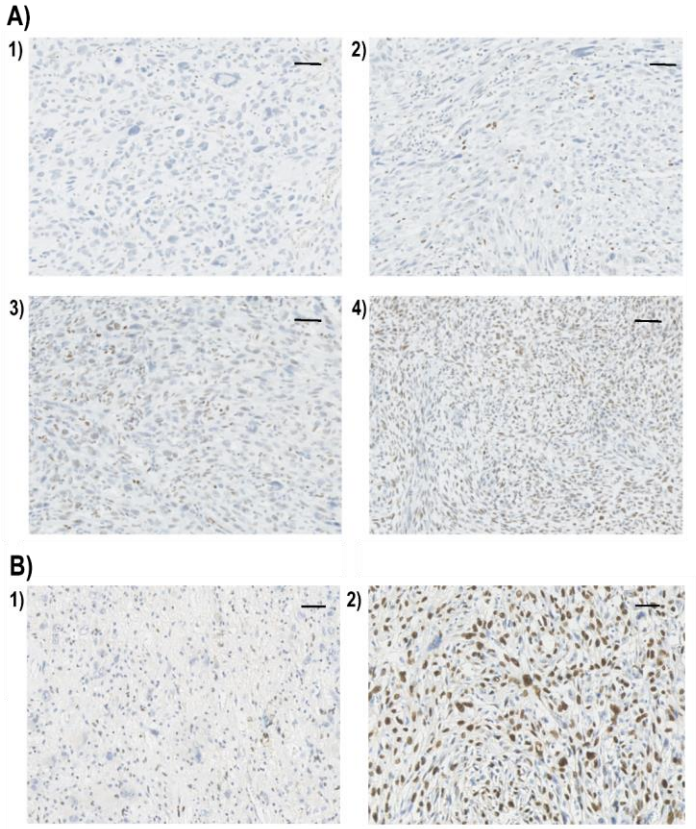
Figure 4.6. a) Violin plot showing the distribution of Variant Allele frequencies for each of the UPS samples. **b)** Mutant-Allele Tumor Heterogeneity (MATH) score for each of the UPS samples. The MATH score represents the intra-tumor heterogeneity of each of the patients.

BOLA1 is a mitochondrial protein that regulates the mitochondrial thiol redox potential[178] and its amplification might protect against mitochondrial damage permitting cancer cell survival. FCGR1A is a high-affinity Fc-γ receptor whose suppression by an IgG4 blockade impacts tumor immunity in melanoma[179]. The two most significant regions of deletion were mapped to 13q14.2 (84% of samples) and 17q13.1 (47% of cancers). 13q14.2

contains 3 genes in peak including the tumor suppressor protein RB1. 17p13.1 contains one gene in the peak, the tumor suppressor p53. Thus, we identified a striking dual mutation of the p53 and RB regions in UPS. This analysis provides one future therapeutic strategy for treating UPS patients; that is developing Rb/p53-mesenchymal cell models and drug screening assays that target cells with dual inactivation of both p53 and Rb.

4.3.5 Targeting mutated p53 cells as a potential therapeutic approach in UPS.

We explored the top ten genes whose deletion or mutation map to cancer-associated genes (**Supplementary Figure 6**). Focusing on RB1 and p53, the data suggest that loss of both p53 and RB1 loci form very common genetic events in UPS, occurring jointly in twelve out of twenty patients. Immunohistochemical analysis of formalin-fixed tissue samples confirms that sarcoma 55 and 74 (containing missense mutations in the p53 gene) result in elevated nuclear staining of p53 protein which is an indicator of p53 gene mutation (**Figure 4.7**). Clues into the significance of this dual mutation of these dominant tumor suppressor genes come from mouse transgenic data showing that deletion of p53 and RB1 accelerated sarcoma development faster than only p53 gene deletion[180]. As p53 deletion, but not RB deletion, stimulates sarcoma development in mice, we can hypothesize that loss of RB within a p53 mutated UPS presumably impacts the de-differentiated phenotype of UPS. At the same time, loss of p53 impacts genetic instability that accelerates mutagenesis and chromosomal rearrangements driving UPS development. This provides two independent roles for p53 and RB in sarcoma genesis. These data are consistent with a study showing that SV40 large T antigen, which can bind both p53[181,182] and Rb1[183], can drive the transformation of mesenchymal stem cells into a UPS-like cellular phenotype[184].



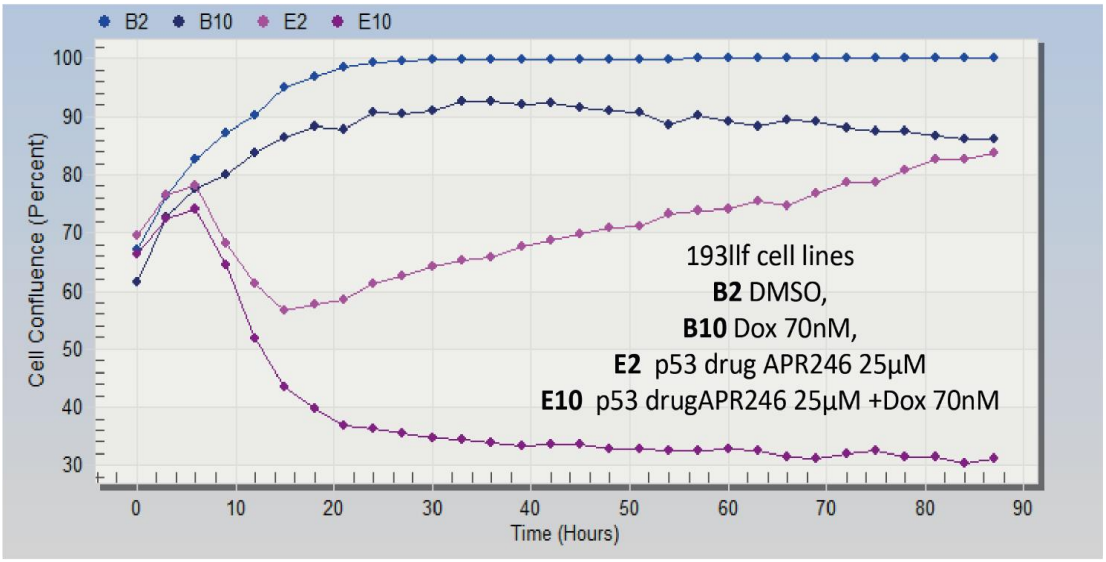
	Rb Score	p53 Score
sarcoma73	0	0
sarcoma60	++	0
sarcoma94	+++	0
sarcoma66	0	0
sarcoma59	0	0
sarcoma55	+	++++
sarcoma74	0	++++
sarcoma84	+++	0
78	0	++++
97	0	0
55	+	++++
88	0	+
13805	+	+

Where 0 is no staining of tumour cells
\+ <5%
++ 5-49%
+++ 50-90%
++++ 90-100%

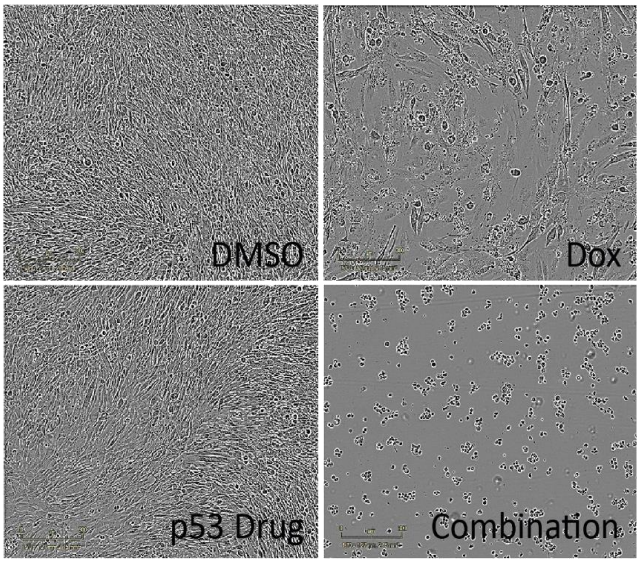
Figure 4.7. Representative immunohistochemistry staining shows the differences between **A)** RB1 in 4 UPS patients, and **B)** High and low degrees of P53 staining in two UPS patients. The immunohistochemistry degree of staining was collected for the different UPS samples in the right table.

The high-frequency loss of both of these tumor suppressor genes suggests that a common therapeutic strategy to impact UPS patients would be to identify drugs that can kill tumor cells with loss of p53 and/or pRB1. A compelling new drug lead named APR-246 has been identified that kills tumor cells with mutant p53 has been tested in a variety of preclinical models and is undergoing clinical trials in human patients[185,186]. As such, we tested whether a novel angiosarcoma mouse model driven by mutant p53 would serve as a model system to evaluate whether APR-246 can enhance the killing of sarcoma cells containing mutated p53. Although the derived mutant p53-murine sarcoma cell line 193llf is partially resistant to doxorubicin and APR-246, the combined treatment effectively kills the sarcoma cells as demonstrated using xCELLigence real-time growth assays, cell density, and Alamar-blue (**Figure 4.8a-c, respectively**).

a)



b)



c)

Doxorubicin (µM)

APR - 246 (µM)	0	0.078	0.156	0.3125	0.625	1.25	2.5	5	10
0	100	95.72	83.1	75.82	87.03	47.66	28.36	20.64	11.32
6.25	100	85.4	74.77	68.33	47.56	26.44	43.34	20.13	10.51
12.5	79.73	68.33	60.64	57.4	30.19	21.68	19	13.71	4.25
25	4.27	4.17	4.29	4.14	4.18	4.18	4.25	4.24	4.33
50	4.18	4.12	4.1	4.1	4.18	4.27	4.14	4.23	4.25
100	4.28	4.23	4.25	4.22	4.2	4.21	4.31	4.27	4.3

No growth inhibition Partial growth inhibition Maximal growth inhibition

Figure 4.8. a) Response of the mutant p53 murine-derived cell line (193llf) against 4 different treatments. **b)** Microscopy image of cell survival against different treatments. **c)** Growth inhibition results from testing different concentrations from the combination of APR-246 and Doxorubicin.

Together, these data suggest one obvious therapeutic strategy against UPS using APR-246 or equivalent pharmacological agents that selectively kill cancer without functional wt-p53. Identifying such drugs of course has been a long-term challenge since p53 mutation was first recognized as a common oncogenic mechanism across many cancer types. Developing synthetic lethal screening strategies in cancer cells that lack both p53 and Rb might accelerate drug discovery in UPS. Such cell models might be possible considering human MSCs can be transformed into UPS-like cells using Large-T-antigen[184]. It remains to be seen whether dual knock-out of RB and p53 can produce the same transformation of a mesenchymal cell.

4.3.6 Heterogeneity of infiltrating immune cells in UPS

One emerging approach to exploit personalized patient cancer genomic sequencing is to develop neoantigen vaccines that stimulate pre-existing tumor-specific T-cells recognizing mutated neuropeptides[187–190]. As our data suggest significant inter-tumor heterogeneity could be observed within each UPS sample, then vaccination strategies would be required that can capture individualized neoantigen burden accurately. The data thus highlight the need for more accurate tools to define potential neoantigen burden to fully target the tumor using neoantigen stratagies[187–190]. Another approach would be to utilize monoclonal antibodies targeting immune checkpoint receptors to achieve an auto-immune reaction against cancer-specific targets, and this approach has had great success in melanoma amongst other diseases[191]. Immunotherapy has the potential advantage that it can target multiple identifiable cancer-specific epitopes in a single tumor, and potentially also provide a bystander effect against adjacent cancer cells, thus overcoming some of the challenges of tumor heterogeneity.

To begin to examine the immune-cancer synapse as a possible therapeutic target in individual UPS patients, we analyzed the extent of immune infiltrate and in particular the extent of T-cell clonality, reasoning that oligoclonality, as opposed to polyclonality, might be indicative of tumor neoantigen-specific T-cell clonal expansion. To determine the repertoire and degree of tumor-infiltrating lymphocytes (TIL) the CDR3 regions of the

TCR beta regions were sequenced in 8 tumor samples. Samples were analyzed by high-throughput sequencing of the TCR β CDR3 region using the ImmunoSEQ immune profiling system at the survey level (Adaptive Biotechnologies, Seattle, WA). ImmunoSEQ data were exported from Adaptive Biotechnologies and imported into the Bioconductor package LymphoSeq (version 1.0.2). Only TCR β CDR3 that produced productive sequences were included for analysis. The relative degree of clonality is represented by a Lorenz curve drawn such that the x-axis represents the cumulative percentage of unique sequences and the y-axis represents the cumulative percentage of reads (**Figure 4.9**). A line passing through the origin with a slope of 1 reflects equal frequencies of all clones. The Gini coefficient is the ratio of the area between the line of equality and the observed Lorenz curve over the total area under the line of equality. Both Gini coefficient and clonality are reported on a scale from 0 to 1 where 0 indicates all sequences have the same frequency and 1 indicates the repertoire is dominated by a single sequence. Tumor samples 94 and 59 have the highest clonality of T-cells, with tumor sample 60 having the least.

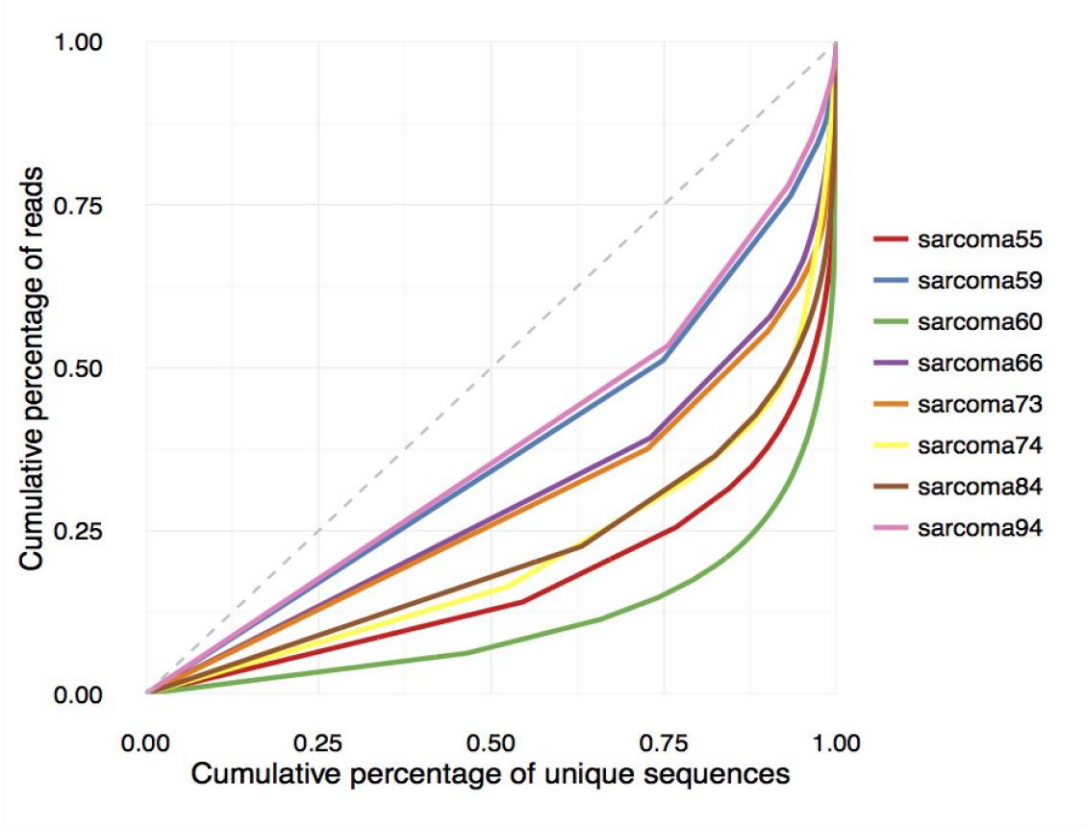


Figure 4.9. Lorenz curve plotting the cumulative sequences on the y-axis against the cumulative percentage of unique reads on the x-axis. Higher clonality is represented as the closeness of the curves to the dotted correlation line.

The repertoire diversity was further highlighted by visualizing whether there are any common CDR sequences amongst the samples. The number of overlapping productive amino acid sequences between samples was visualized with a UpSetPlot generated using the R package UpSetR (**Figure 4.10**). For instance, sarcoma 74 displayed 35 sequences in common with Sarcoma 60. The majority of shared CDR sequences between tumors were in the range of 1-7 (**Figure 4.10**). Together the data suggest a relatively high degree of individuality in immune cells with tumor-specific recognition in each patient's tumor. T-cell receptor deep sequencing indicated the presence of dominant clonal T-cell infiltrates in the range associated with other cancers responsive to immunotherapy, suggesting an alternative immunotherapeutic strategy, perhaps enhanced by the development of patient and cancer-specific vaccines[192].

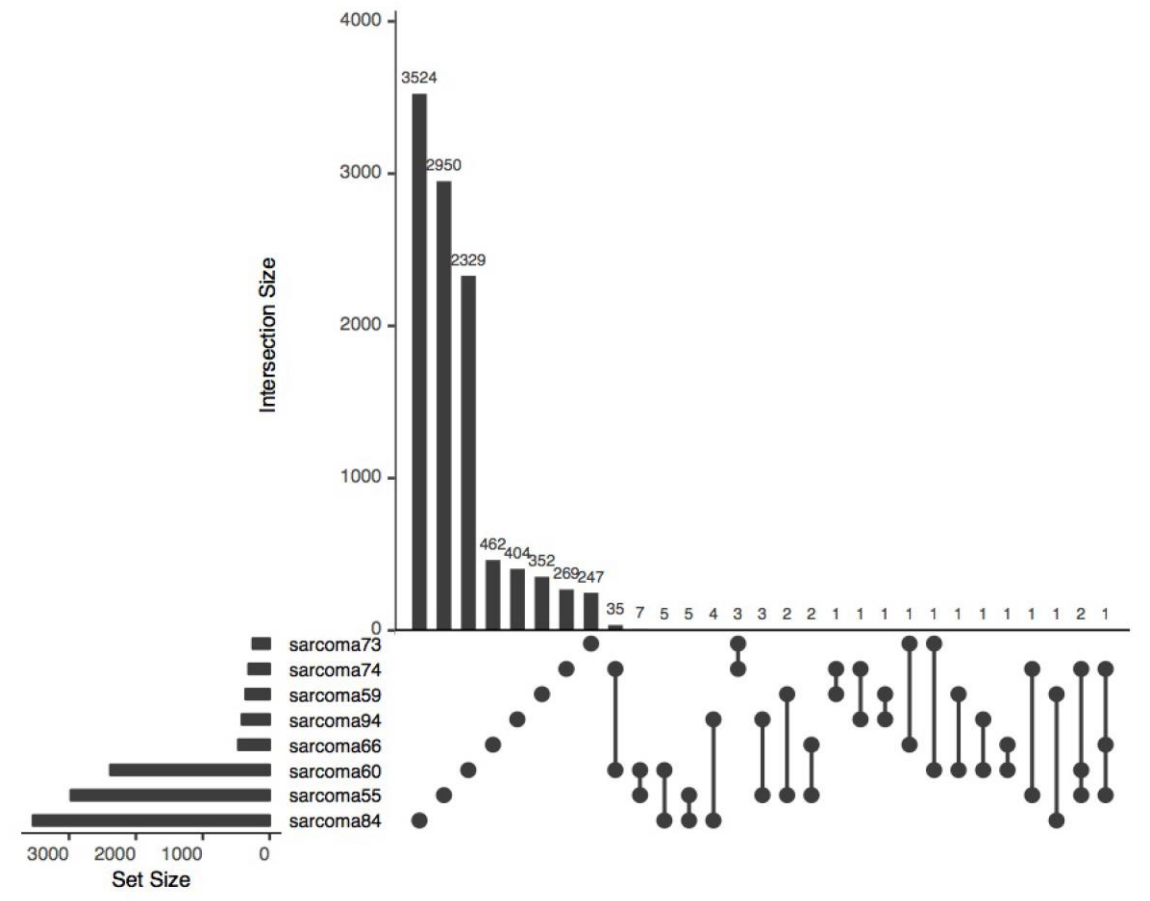


Figure 4.10. Correlation of the complementary-determining region 3 between each of the 8 UPS samples sequenced. The first 8 bars represent the amino acid size of that region while the rest indicates the number of common regions between patients.

4.3.7 Towards personalized proteogenomics in UPS

The high levels of immune cell clonality infiltrating UPS suggest that personalized neoantigens might be of use as personalized therapeutics. Towards this end, we aimed to define a proteogenomics methodology for defining potential neoantigens in UPS. Most neoantigen discovery methods use DNA-seq as a source of potentially mutated peptides. We used proteogenomics to include the subset of mutated genes that have detected peptides by shotgun mass spectrometry of the tumor biopsy. The tumor sample 55 versus normal tissue had over 1,743 proteins identified. This includes the previously identified "oncogenic" antigen overexpressed in many sarcoma subtypes; CLIC1[193](**Supplementary File 1**). Of these differentially quantified proteins, 30 were linked to mutated genes. This reduces the list of mutated genes identified as a source of neoantigens by over 85%. That is, we rule out a mutated gene if the levels of its protein are not detectable.

We next determined whether we could identify high-affinity MHC Class I peptides using MHCNET4.0, an algorithm used to predict high-affinity peptide binders[194]. The HLA class of tumor 55 defined using http://nagasakilab.csml.org/hla includes: C*07:02:01:03, B*07:02:01, A*03:01:01:01, C*04:01:01:01, B*35:01:01:02, and A*29:02:01:02. When these alleles are filtered through MHCNET4.0 using the mutated, trimmed peptide library derived from the SNVs of tumor 55, then high-affinity neoantigens can be defined. Using SNVs defined using DNA-seq filtered using a SWATH-MS dataset, then there were potential neoantigens defined including CADM1, IDH3G, SSCPDH, PLEC, and PDCD6. The additional use of the IDA peptide library derived from the tumor increased the potential neoantigens by one, PABPC1.

The use of RNA-seq as a source of expressed genes, filtered using the SWATH-MS library, resulted in the identification of 12 potential mutated proteins (**Supplementary Table 1**). The number with predicted MHC Class I binders included CAMD1, PDCD6, IDH3G, HUWE1, and MTCH2. The inclusion of the peptide library derived from Tumor 55 increased by 6 the number of potential mutated proteins. Of these, there were four with potential neoantigens including HLA-A, HLA-DRB1, IGHG4, and IGH3G. Interestingly, the MHC Class II protein HLA-DRB5 had 5 tumor-specific mutations from amino acids

40-55 residing in the peptide-binding pocket. PDCD6 has a peptide that when trimmed to a 9-mer is predicted to bind with high affinity to B*35:01:01:02 and C*07:02:01:03 (**Supplementary Figure 6A**). Mutated PDCD6 mRNA was detected by shotgun RNAseq in tumor 55 (**Supplementary Figure 6B**). We, therefore, focused on determining whether mutated peptides can be detected by mass spectrometry for PDCD6; high-confidence mutated peptide was detectable (**Supplementary Figure 6C**) providing evidence for the concept that the genomic sequencing of UPS can be used to identify mutated proteins expressed in the tumor.

4.4 Conclusion

In the case of UPS, our data suggest that (i) the cancers are heterogeneously mutated: (ii) the most common genetic alteration based on chromosomal analysis is loss of both RB and p53 tumor suppressor gene pathways; and (iii) the vast majority of mutated genes are largely patient specific. Although common mutations were found highlighting the dual inactivation of the p53 and RB pathway, the data also suggest that more personalized strategies using next-generation DNA sequencing will be one of the best approaches for "rare" human cancers such as UPS. These personalized strategies could include the use of patient-specific cancer neoantigens as vaccine therapeutics.

Next-generation sequencing of sarcoma subtypes is emerging and is informing on cancer-specific mechanistic driver events. The vast majority of synovial sarcomas have a unique chromosomal translocation, t(X;18)[195] although druggable targets are not apparent from the targeted sequencing of this tumor class. Genome sequencing of chondroblastoma highlighted the very high frequency activating mutation at Histone $H3.3^{K36M}$ that impacts on altered expression of cancer-associated genes[196] and differentiation of mesenchymal progenitor cells (MPCs)[197]. Pulmonary sarcomatoid carcinoma exome sequencing has revealed over 20% of tumors display exon14 skipping in the met receptor gene that provides a druggable oncogenic driver event[198]. Targeted sequencing of Desmoid type fibromatosis has identified CTNNB1or APC pathway mutation in over 90% of patient samples[199] and suggests specific genetic pathways for developing targeted therapeutics. Genome

sequencing of Angiosarcomas identified truncating mutations in the PTPRB phosphatase gene in 26% of patients and 9% of cancers harboring possible activating mutations (R707Q) in the plcg1 tyrosine kinase gene[200]. Together, these data highlight the power of next-generation cancer genome sequencing to annotate the genetic blueprint of a pathologically defined sarcoma subtype and to often provide focused therapeutic strategies.

UPS is the most common adult sarcoma. Our previous proteomics analysis in UPS identified a highly expressed oncogenic target exemplified by CLIC1[193]. The CLIC1 target was also highly expressed in many different types of sarcomas, in essence producing a common therapeutic option for many sarcoma subtypes[193]. However, CLIC1 is not necessarily a highly druggable target. As such, we initiated a genomics study to complement the proteomics to determine whether additional knowledge can be acquired on how to develop therapeutic targets for UPS. Whole exome sequencing of cancer genomes derived from twenty UPS patients identified key options to analyze in the future as a point of focus; (i) Can Rb/p53 hypomorphic cancers provide therapeutic options; ii) does the ATRX interaction landscape and/or kinase activity provide therapeutic options; and iii) can exploitation of cancer genome and proteomic technologies (proteogenomics) identify personalized vaccine candidates for neoantigen therapies in UPS.

Compared to the previous study of esophageal adenocarcinoma, the multi-omics approach developed in UPS has a more genome-centric approach. The study of DNA mutations followed by the development of personalized strategies based on shotgun RNA sequencing and mass spectrometry proteomics has proven to be an informative way of proteogenomic integration.

5. Gorham-Stout disease (published article).

Although the first patient presenting symptoms, like the disappearance of bone tissue of the humerus, was reported in 1838[201] it wasn't until 1955 that Gorham and Stout presented 8 cases of the syndrome with details of their clinical characteristics[202]. The patients suffered from massive osteolysis and lymphangiogenesis where the formation of vascular tissue was promoting the destruction of the surrounding bone. The study of GSD over the years gave further details of the clinical characteristics, revealing the formation of abnormal lymphatic vessels in the bone tissue[203], and the involvement of the immune system in osteoclast activation, either through T-lymphocites[39], leukocytes, and dendritic cells[204] or through macrophages[39]. Regardless of these efforts, the inside mechanisms of regulation and molecular triggers that cause GSD are still unknown along with the exact pathophysiology of the disease[205].

The incidence of Gorham-Stout syndrome remains unclear. The disease doesn't seem to present any gender preference[206,207], besides being the majority of the reported cases in men[208]. The range of age varies from 1 to 70 years, although there is a tendency to affect children and young adults under the age of 40[209]. Jaw, shoulder, pelvis, spine, and skull are the most commonly affected tissues of the disease, with the possibility of affecting multiple parts of the body at the same time[210].

The rareness of Gorham-Stout syndrome has created a spread distribution in the number of cases detected through the ages, which has provoked the use of different therapeutic strategies against the disease. The use of multiple strategies has produced a lack of consensus, with current therapies covering a broad range of options[211]. Most of the current procedures converge in the starting point of the treatment, initiating with a surgical resection followed by radiotherapy. Recent advances have shown successful management of the disease when the use of radiotherapy was complemented with chemical therapies like zoledronic acid, vitamin D and propranolol[212], or bisphosphonate[213,214]. The most novel approach to the treatment of the disease is the use of sirolimus (rapamycin), an oral *mTOR* inhibitor that focuses on the *PI3K-Akt-mTOR* signaling pathway which affects

angiogenesis and lymphangiogenesis. Although most of the patients were very responsive to this treatment the safety of Sirolimus remains unknown[215].

Another effect of the low incidence of Gorham-Stout is that the studies are restricted to case reports over time, making it difficult to perform a global study of the disease in multiple patients. During the past years, the few published studies focusing on genomic approaches to the syndrome have discovered mutations in *TNFRSF11A* and *TREM2* as possible drivers of the disease[213], as well as shared mutations between cancer and Gorham-Stout in *KRAS*[216,217].

To fill the void of a multi-omic approach in the literature we have developed a study of a 45-year-old female patient with marked bone loss of the left humerus associated with vascular proliferation, diagnosed with Gorham-Stout disease. By performing DNA and RNA sequencing combined with antibody-based staining of the immune infiltrate in the Gorham-Stout tissue we provided a unique insight into the disease that has revealed major chromosomal rearrangements that could be drivers of the disease. Further exploration revealed the success behind current treatments focusing on the *PI3K-Akt-mTOR* pathway, as it has been characterized as the most affected signaling cascade in Gorham-Stout tissue. Furthermore, a combined analysis of immunohistochemistry and RNA-sequencing data have confirmed previous findings on M2 macrophage infiltration in the GSD lesional tissue[39].

Yébenes Mayordomo *et al.*
BMC Medical Genomics (2022) 15:128
https://doi.org/10.1186/s12920-022-01277-x

BMC Medical Genomics

CASE REPORT Open Access

Gorham-Stout case report: a multi-omic analysis reveals recurrent fusions as new potential drivers of the disease

Marcos Yébenes Mayordomo[1*], Sofian Al Shboul[3], Maria Gómez-Herranz[1,2], Asim Azfer[2], Alison Meynert[4], Donald Salter[2], Larry Hayward[2], Anca Oniscu[5], James T. Patton[6], Ted Hupp[2], Mark J. Arends[2] and Javier Antonio Alfaro[1*]

Abstract

Background: Gorham-Stout disease is a rare condition characterized by vascular proliferation and the massive destruction of bone tissue. With less than 400 cases in the literature of Gorham-Stout syndrome, we performed a unique study combining whole-genome sequencing and RNA-Seq to probe the genomic features and differentially expressed pathways of a presented case, revealing new possible drivers and biomarkers of the disease.

Case presentation: We present a case report of a white 45-year-old female patient with marked bone loss of the left humerus associated with vascular proliferation, diagnosed with Gorham-Stout disease. The analysis of whole-genome sequencing showed a dominance of large structural DNA rearrangements. Particularly, rearrangements in chromosomes seven, twelve, and twenty could contribute to the development of the disease, especially a gene fusion involving *ATG101* that could affect macroautophagy. The study of RNA-sequencing data from the patient uncovered the *PI3K/AKT/mTOR* pathway as the most affected signaling cascade in the Gorham-Stout lesional tissue. Furthermore, M2 macrophage infiltration was detected using immunohistochemical staining and confirmed by deconvolution of the RNA-seq expression data.

Conclusions: The way that DNA and RNA aberrations lead to Gorham-Stout disease is poorly understood due to the limited number of studies focusing on this rare disease. Our study provides the first glimpse into this facet of the disease, exposing new possible therapeutic targets and facilitating the clinicopathological diagnosis of Gorham-Stout disease.

Keywords: Gorham-Stout, Genomics, Transcriptomics, Autophagy, Fusions, Mutations, *PI3K*, *AKT*, *mTOR*, Case report

Background

Gorham-Stout disease (GSD) (OMIM #123880) or vanishing bone disease is an extremely rare illness that causes a proliferation of lymphatic vascular channels inducing massive osteolysis and bone loss. Less than 400 cases have been reported since the disease was first described in 1955 [1], leading to challenges in diagnosing and treating the disease. Although it can affect any part of the skeleton, the shoulders and pelvis are the most commonly affected areas [2].

The disorder can be diagnosed at any age but is generally present in patients between 13 and 30 years of age with no sex or ethnic predisposition [3]. The alterations in bone resorption may be one of the reasons why young

*Correspondence: MARCOS.YEBENES@GMAIL.COM; JAVIER.ALFARO@PROTEOGENOMICS.CA

[1] International Center for Cancer Vaccine Science (ICCVS), University of Gdansk, Gdańsk, Poland
Full list of author information is available at the end of the article

adult patients are the main group affected by the disease [4], as resorption is part of normal bone growth. Main therapies consist of surgical excision and reconstruction of bone tissue, radiotherapy, and sirolimus (rapamycin) treatment. Rapamycin treatment represents a novel approach whose safety and efficacy remain unclear [5].

Recent research investigated mutations in this disorder, focusing on 50 cancer-related genes and revealing a somatic activating mutation in *KRAS* causing a gain of function [6]. This variant has been previously reported in other cancer studies [7, 8]. It is known to promote cell growth by activating the *RAS/MAPK* and *PI3K/AKT* signaling pathways relevant to lymphatic vascular growth and angiogenesis [9]. Concurrent with this study, *Nassim Homayun-Sepehr et al* [10] presented a model where a different mutation affecting *KRAS* also activates the development of lymphatic vessels in bone through the same signaling cascade.

The aim of our study is to provide new insights into the genomics and transcriptomics characteristics of the Gorham-Stout disease by performing the first multi-omics exploration of whole-genome sequencing data and RNA-sequencing data in a Gorham-Stout patient.

Case presentation

The 45-year-old white female patient presented with left arm vascular proliferative disease within and around the humerus bone resulting in a pathological fracture. A clinical diagnosis of Gorham-Stout disease was made following pathological and radiological investigations of the lesion at this site (Additional file 1: Fig. S1). Macroscopic examination showed an abnormality of the upper arm muscles which appeared vascular and spongiotic, soft in some areas and fibrotic in others. The cortical bone of the humerus was thin around the fracture site and the bone marrow appeared, similar to the soft tissue, vascular with some large cysts and hemorrhage noted. Tissue blocks of all the abnormal areas from the soft tissue and humeral bone were sampled for histological examination and further investigations.

Histological sections of the affected regions confirmed indeed an abnormal vascular proliferation dissecting through fibroadipose connective tissue, skeletal muscle, and bone. Despite its dissecting nature, the vascular proliferation was composed of thin-walled vascular spaces and papillary projections lined by cytologically bland endothelial cells. The site of the fracture showed reparative fibrotic changes, with extensive granulation tissue and fibrosis. The dissecting vascular proliferation also affected the bone and was associated with cystic changes within the bone. The bone marrow showed vascular and fibrotic changes with some granulation tissue. Occasional discrete granulomas were scattered throughout the lesion. These changes are all consistent with the clinical diagnosis of Gorham-Stout disease or vanishing bone disease. Past medical history includes a previous diagnosis 1–2 years earlier of a benign vascular proliferation/vascular malformation involving both bone and soft tissue of the left humerus compatible with Gorham-Stout disease. The expert pathological review from the Department of Musculoskeletal Pathology, Royal Orthopaedic Hospital NHS Foundation Trust, Stanmore, was in agreement.

Genomic exploration of Gorham-Stout (GS) lesional tissue

Small DNA variants were called using whole-genome sequencing data of GS vascular proliferation tissue and surrounding normal tissue extracted from the same patient (Additional file 2: Methods). Although the clinical characteristics of GS can be similar to those manifested in Ewing's sarcoma [11], the genomic profile differs from most cancer-like diseases, as it doesn't seem to be mutation-driven. A total of 643 mutations in 233 genes were found in GS lesional tissue when compared to the adjacent normal. Of those mutated genes, neither *TP53, RB1, CDKN2A* nor any of the known sarcoma biomarkers [12] were reported to contain any small mutations.

The classification of variants (Fig. 1a) showed that, although most of the variants code for missense mutations, a considerable number of insertions and deletions were found. Among the top mutated genes (Fig. 1b), most of the genes reported contained insertions, deletions, or splice site changes. Some known cancer-related genes like the mucin family (*MUC3A* and *MUC12*) in colorectal cancer, or *ZNF703* in breast cancer, seemed to be mutated. Other genes like *CST5* and *UNC5B*, related by previous studies to the *P53* pathway [13, 14], were also involved in other signaling pathways that are manifested in GS, like endochondral ossification and autophagy respectively. Other relevant gene families affected are those from *TNFRSF10A* and *ANKRD36* genes, previously identified as mutated in a scapular lesion affected with Gorham-Stout disease [3].

Structural variants and gene fusions

We noticed a high proportion of genes with insertions and deletions displayed in the top 20 mutated genes. This motivated an analysis of larger indels and structural variants to explore the possibility that those alterations might have been caused by major chromosomal events. (Fig. 1c). Matching our previous discoveries, a high number of structural variants were identified, especially in chromosomes seven, twelve, and twenty. Around 1000 structural variants were found, showing duplications, deletions, and tandem repetitions. The most frequent structural events detected were chromosome translocations, suggesting that gene fusion

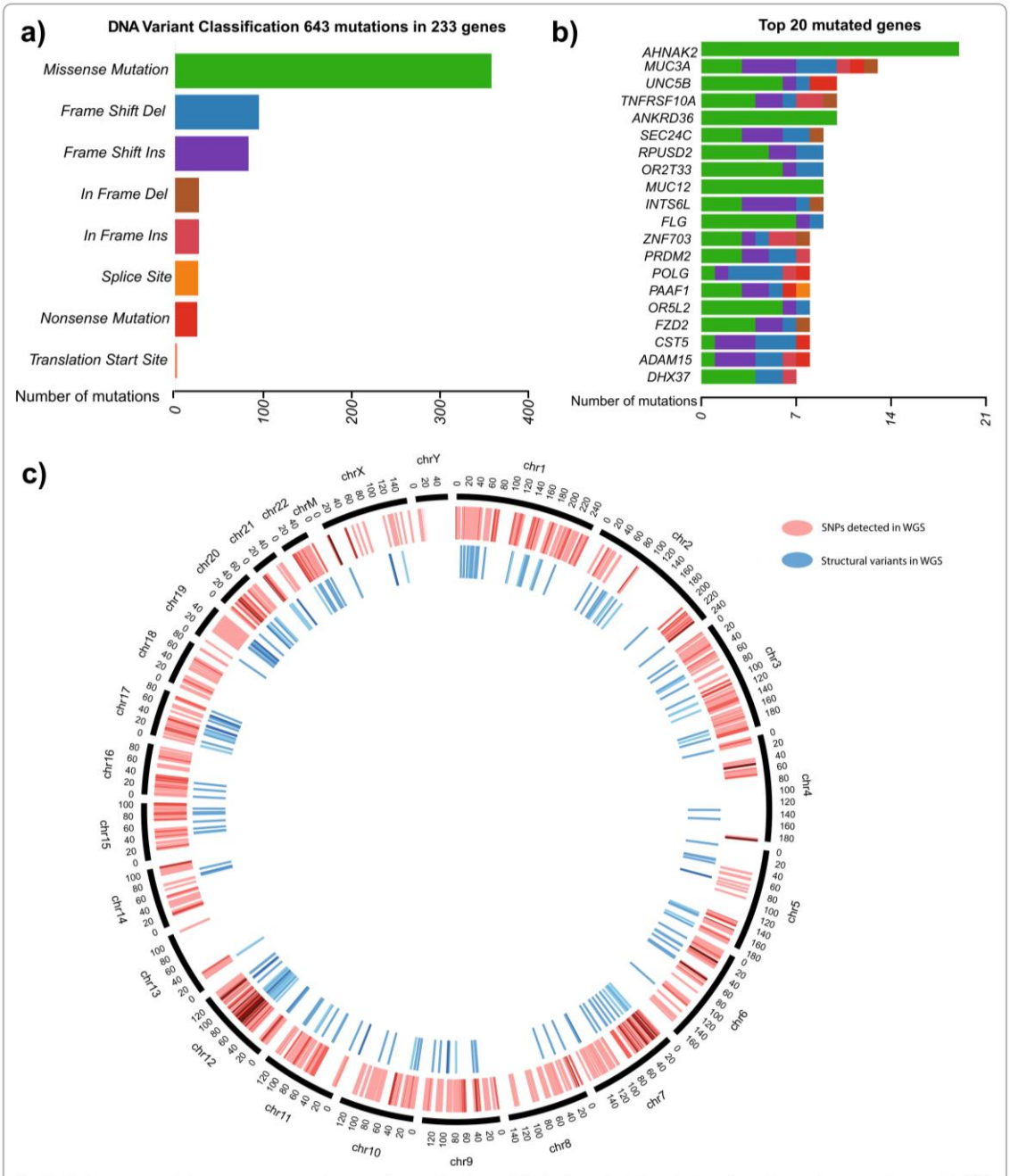


Fig. 1 Gorham stout whole-genome sequencing mutation exploration. **a** 643 single nucleotide variants and small mutations were detected in 233 genes in the GS tissue, **b** Top 20 genes containing the most number of mutations in the GS tissue. **c** Circos plot comparing the number of small mutations (red) and structural variants (blue)

variants could be a major event for GS disease. To further explore these rearrangements, we combined DNA and RNA gene fusion calls for the case (Fig. 2a). The fusions were categorized in different tiers depending on the evidence of the mutation at both the DNA and RNA level based on the genes surrounding the fusion or reads containing the translocation (Additional file 3: Table S1). Although most of the fusions reported were intrachromosomal events, chromosomes twelve, seven, and twenty shared a relevant number of interchromosomal mutations.

Based on the evidence of the gene fusions and their biological relevance, seven gene fusions were selected for RT-PCR validation (Additional file 4: Fig. S2). The fusion of *ATG101* and *SLC4A8* in chromosome 12 (Fig. 2b) involves the autophagy-related protein part of the macroautophagy signaling pathway [15]. The other validated fusion (Fig. 2b) involves sarcoglycan delta (*SGCD*), a gene

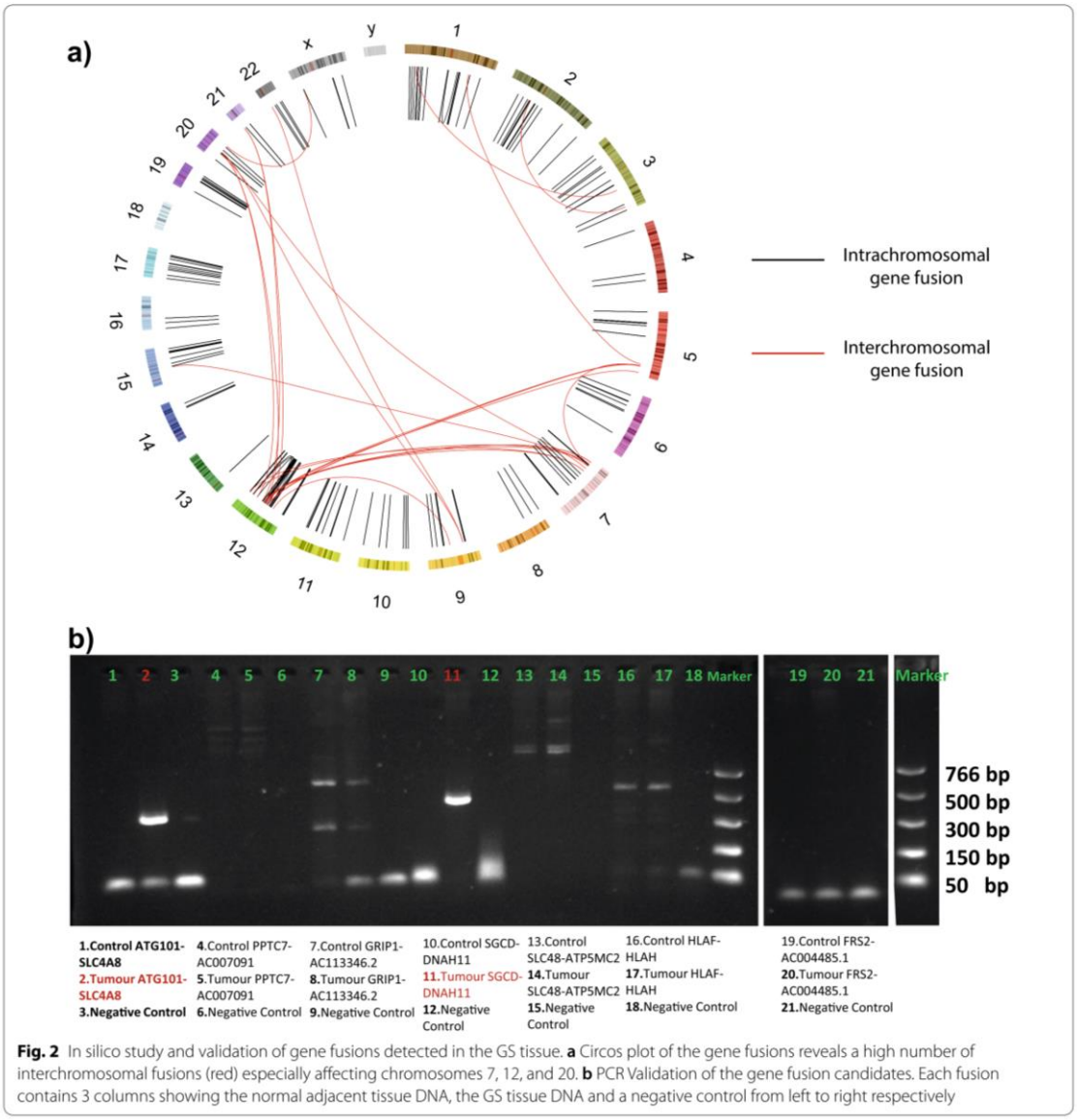


Fig. 2 In silico study and validation of gene fusions detected in the GS tissue. **a** Circos plot of the gene fusions reveals a high number of interchromosomal fusions (red) especially affecting chromosomes 7, 12, and 20. **b** PCR Validation of the gene fusion candidates. Each fusion contains 3 columns showing the normal adjacent tissue DNA, the GS tissue DNA and a negative control from left to right respectively

known to cause muscular dystrophy in mammals when deleted or mutated [16–19], an event that could lead to vascular malformations as reported in the GS lesions.

Gene expression in Gorham-Stout disease

The differential expression analysis of the technical replicates from Gorham-Stout lesional tissue and attached normal revealed that a high proportion of genes (36.6% out of the 58,884 total) were either up-regulated or down-regulated (Fig. 3a). Gene set enrichment analysis showed statistically significant differences in pathways like lymphangiogenesis (Additional file 5: Table S2) and osteolysis (Additional file 5: Table S3), suspected of being drivers of the pathological characteristics and potential targets of drug treatments for the disease [20] (Fig. 3b). Inside these pathways, gene families like *VEGF* or *NOTCH* were detected to change drastically in expression from normal to Gorham-Stout lesional tissue (Additional file 7: Fig. S3).

One of the main pathways affected by changes in expression is the phosphatidylinositol 3-kinase (*PI3K*), involved in the proliferation, growth, and regulatory processes of the cell. In our study, we detected that the expression of *PI3K* is considerably downregulated in GS lesions when compared to normal tissue, therefore the phosphorylation of *PIP2* to *PIP3* by this gene will be decreased in the disease. This event is confirmed by the up-regulation of PTEN which regulates *PI3K* by the dephosphorylation of the *PIP3* product [21].

The changes in expression of the *PI3K* and *PTEN* pathway (Fig. 4a–b) are reminiscent of other cancers, where the pathway is deactivated or mutated affecting regulation of mTOR [22]. The expression of *PI3K, AKT,* and *mTOR* in the Gorham-Stout lesion was decreased, while *PTEN* expression was higher when compared to the attached normal. The alterations suggest the promotion of irregular endothelial cell growth and angiogenesis through *VEGFA* and *VEGFB* via the *VEGFR1-PI3K-AKT* signaling pathway [23, 24].

Another expression event that may be related to cancer is the activation of the *NF-kB* signaling pathway, which is down-regulated in Gorham-Stout lesional tissue when compared to matched normal. The levels of expression of *NF-kB* and *IKK* in the normal adjacent tissue could contribute to inflammation and macrophage activation [25, 26].

Although the expression changes previously mentioned are occurring in normal adjacent tissue, there are events in GS tissue that share similarities with cancer. One of them is the high expression of *MDM2*, a p53-specific E3 ubiquitin ligase, which leads to the degradation of the *p53* tumor suppressor protein [27]. All the events linked to the *PI3K* signaling pathway paint a picture of the possible inner mechanisms in Gorham-Stout and surrounding tissue providing unique insights into the disease that could lead to the development of new therapeutic strategies targeting the mentioned pathways.

Immune infiltrates in Gorham-Stout lesional tissue

Immunohistochemical and H&E staining was carried out using formalin-fixed, paraffin-embedded (FFPE) sections to investigate the immune system response of a single case. Evaluation of the immune cell infiltration was assessed by immunohistochemistry with five immune cell markers: $CD3^+$ T cells, $CD4^+$ T cells, $CD8^+$ T cells, $CD20^+$ B cells, and $CD163^+$ M2 macrophages (Fig. 5A–F). Stained Gorham-Stout disease slides were scanned using a Hamamatsu NanoZoomer XR slide scanner at × 40 magnification. Digital images were analyzed using QuPath (version 0.2.0-m7) to quantify the positive staining, validated by manual counting in selected areas that showed a highly significant correlation (Table 1).

The Gorham-Stout disease specimen showed significantly higher $CD163^+$ M2 macrophage infiltration, compared with other examined immune cell markers (Fig. 5E and Table 1). This was consistent with the RNA data that revealed increased infiltration of M2 macrophages in Gorham-Stout disease compared with normal samples (Fig. 5G). In contrast, $CD4^+$ T cell staining was particularly low with only 0.06% positive staining (Fig. 5B and Table 1). Moderate staining was found for the other 4 lymphocytic cell markers: $CD3^+$ T cells (8.22%), $CD8^+$ T cells (5.93%), and $CD20^+$ B cells (4.86%).

Discussion and conclusions

Large chromosomal events were detected in chromosomes seven, twelve, and twenty, where gene fusions were the dominant event. The gene fusion of *ATG101* and *SLC4A8* (Additional file 8: Fig. S4) involved a binding protein (*ATG101*) essential for macroautophagy [15, 28], which could affect the macrophag e signaling pathway. This event has to be confirmed in other Gorham-Stout patients, as the study of the structural variants and gene fusions is a novel insight in this field. Evidence for this fusion to be pathological was strengthened upon further investigation by ensuring adherence to the ACMG standards and guidelines for the interpretation of sequence variants [29] (Additional file 2: Methods).

The small mutations found in this case did not match any of the previously mutated genes found in the literature, although genes were belonging to the same families and/or affected the same pathways previously found in neoplasms. The tumor necrosis factor receptor *TNFRSF10A* was detected in our study among the top five mutated genes showing multiple deletions, insertions, and point mutations. *TNFRSF11A*, a member of

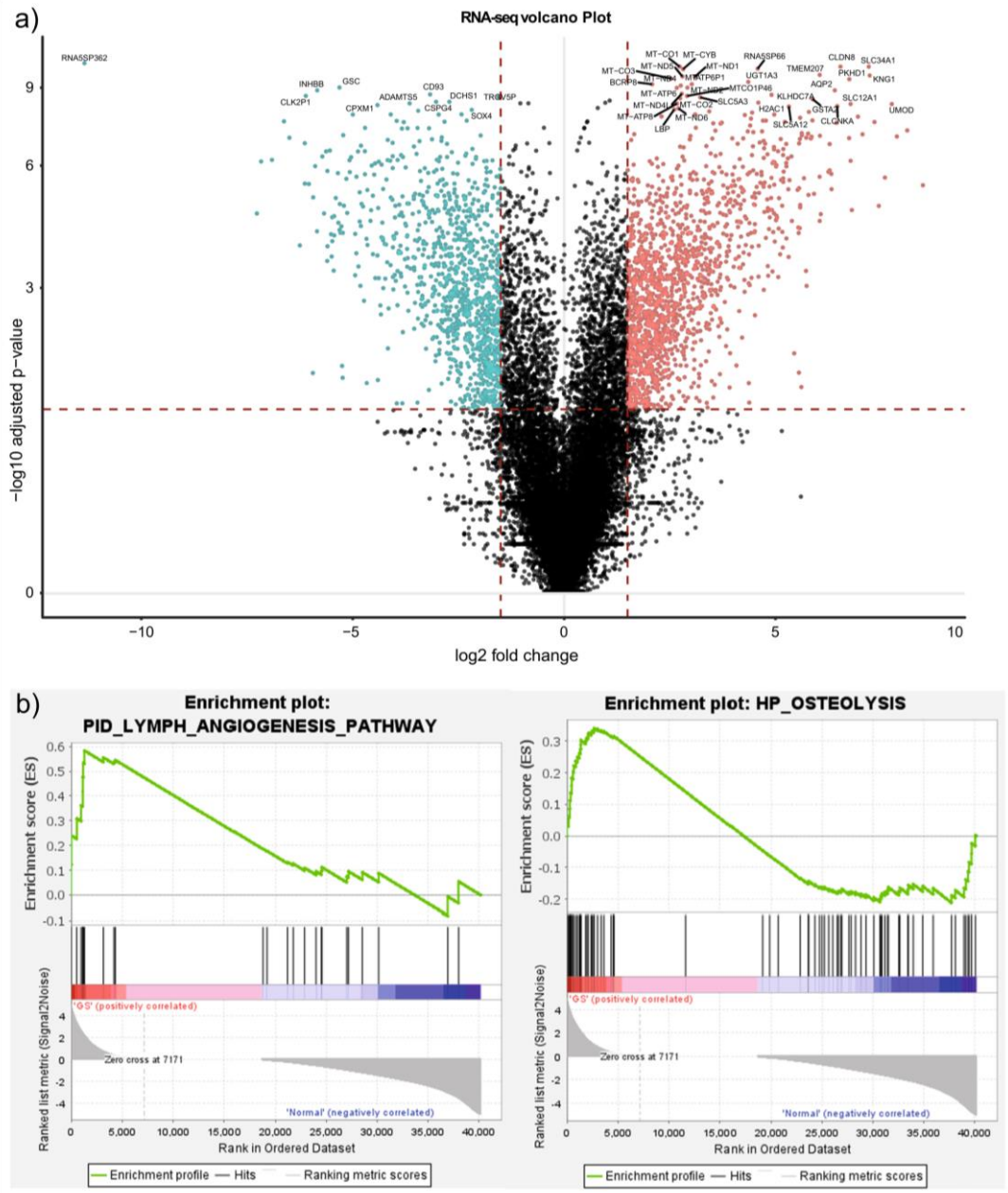


Fig. 3 RNA expression analysis between Gorham-Stout tissue and adjacent normal. **a** Volcano plot of the RNA-seq differential expression analysis showing a large number of genes up and down-regulated in GS tissue. **b** Results of the gene set enrichment analysis (GSEA) selecting the lymph-angiogenesis (VEGF) and osteolysis (NOTCH) pathways. A heavily weighted distribution in multiple genes was found in both pathways, with a high number of genes overexpressed in either GS tissue or the attached normal

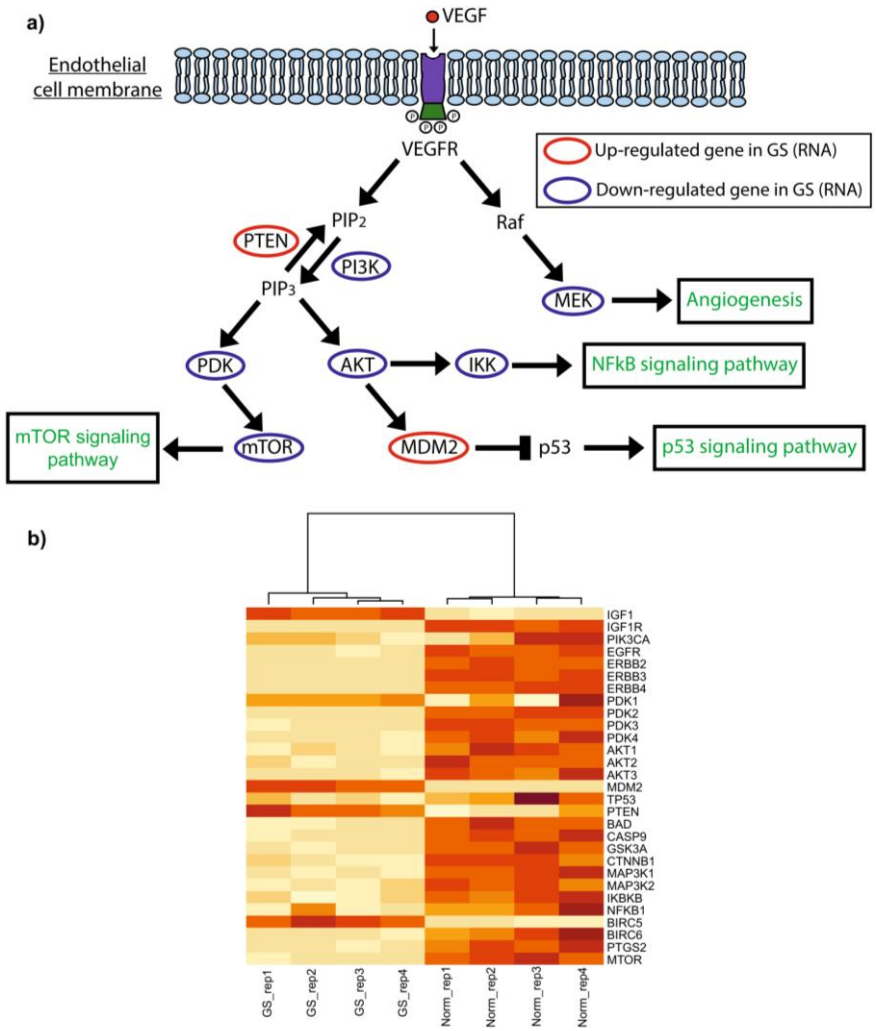


Fig. 4 Alterations in the PI3K/AKT signaling pathway in GS tissue. **a** Brief representation of gene members in the PI3K/AKT pathway. Most genes showed changes in expression when comparing adjacent normal to GS tissue affecting other metabolic pathways. **b** Heatmap representation of the gene members affected by RNA-seq expression changes between normal and GS tissue. PTEN and MDM2 showed over-expression in GS tissue while most of the other affected genes appear down-regulated when compared to adjacent normal

the same family of genes, was reported in a previous case report of a Gorham-Stout patient [3] and linked to muscular dystrophy and osteolysis [30, 31]. Another example of mutated gene families affecting the same pathway is the missense mutation found in *PIK3AP1* (c.1139A > T), which belongs to the *PTEN/PI3K/AKT* signaling cascade. Genes belonging to this family, like *PIK3CA*, are known to cause lymphatic and vascular overgrowth disorders [32] while others like *PTEN* have been reported as mutated in Gorham-Stout disease patients [33].

The alterations of the *PTEN/PI3K/AKT* signaling cascade were not only observed by mutations of some of the members of the pathway but also reported as gene expression changes in the RNA sequencing data. We observed that most of the genes involved in the signaling cascade were either up-regulated or down-regulated when compared to normal adjacent tissue. The

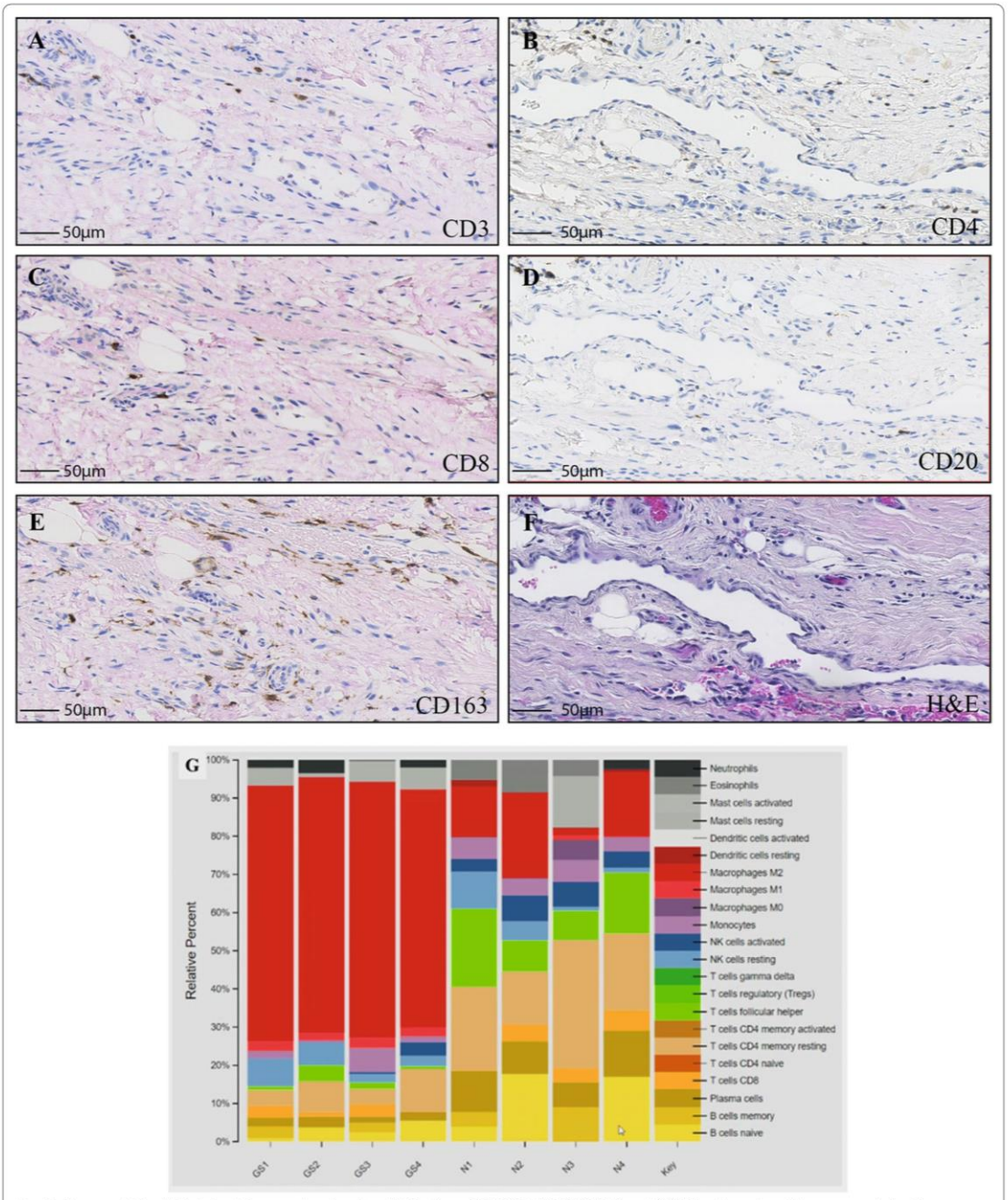


Fig. 5 Representative IHC stained images showing the distribution of CD3, CD4, CD8, CD20, and CD163 cell markers. The representative images exhibit the immunohistochemical features of infiltrating immune cells: **a** $CD3^+$ T cells, **b** $CD4^+$ T cells, **c** $CD8^+$ T cells, **d** $CD20^+$ B cells, **e** $CD163^+$ M2 macrophages, and **f** H&E staining. Scale bars show 50 μm. **g** RNA comparison of a variety of immune cell types between Gorham-Stout disease and normal specimens. The analysis was conducted with 4 technical replicates

Table 1 Summary of the percentage of cells stained for CD3, CD4, CD8, CD20, and CD163 cell markers within the Gorham-Stout disease sample

Sample	Positive (%)	Rho	P value
CD3	8.22	0.997	< 0.0001
CD4	0.06	0.986	< 0.0001
CD8	5.93	0.991	< 0.0001
CD20	4.86	0.995	< 0.0001
CD163	20.01	0.996	< 0.0001

The biopsy sections were stained with the following cell markers: CD3, CD4, CD8, CD20, and CD163. Positively stained cells (expressed as a percentage of total cells) were automatically counted using QuPath (version 0.2.0-m7). The methodology was verified by comparing manual counting with QuPath counting in 0.2 mm. [44–46] areas selected randomly across the different sections. Pearson's correlation (Rho) and P values were calculated

modifications of the PI3K pathway are known to cause lymphatic malformations [9, 34] and during recent years have become the main target for inhibitor therapies designed to decrease *VEGF* secretion and angiogenesis [35–38]. Although the *PI3K* pathway was already known to be affected in Gorham-Stout disease, as well as other lymphatic malformations, our study is the first to profile Gorham-Stout lesions by RNA-Seq analysis and this has demonstrated new possible candidates for therapy like the targeting of *MDM2-p53* already developed for cancer therapy [39, 40].

Besides angiogenesis and osteolysis, another characteristic of GS disease is osteoclast formation. Previous studies have suggested this event is stimulated by macrophage secretion of *TNFα* and *IL-6* [41] and linked it to the clinical characteristics of the disease [42]. Our study has shown that M2 macrophages tend to infiltrate the Gorham-Stout vascular proliferation tissue, while other immune cells appear to be less frequent. The results match previous findings in the literature where CD163 staining was also performed [6], as well as the mostly negative staining for other immune cells [43].

We have presented a detailed molecular investigation of a single patient with Gorham-Stout disease. Whole-genome sequencing data of the Gorham-Stout vascular proliferation lesion revealed that the main driver of the genomic events appears to be large structural alterations, though single nucleotide variants and small mutations were also present. The transcriptomics showed changes in expression between the normal and the Gorham-Stout tissue, involving the osteolysis and angiogenesis pathways. The alteration of the *PI3K/AKT/mTOR* pathway along with the macrophage infiltration in the Gorham-Stout tissue are congruent with emerging trends in this disease. As with any rare disease, the inclusion of further GSD patients into the future with a combined genomic and transcriptomic profile could confirm the insights we have revealed on the mechanisms of the disease.

Abbreviations

GSD: Gorham-Stout disease; GS: Gorham-Stout; WGS: Whole-genome sequencing; BWA: Burrow-Wheeler aligner; Hg38: Homo sapiens genome assembly GRCh38; GATK: Genome Analysis Toolkit; SNV: Single nucleotide variants; Indel: Insertions and deletions; RSEM: RNA-seq by expectation maximization; TPM: Transcripts per million; DEA: Differential expression analysis; IHC: Immunohistochemistry; H&E: Hematoxylin and eosin; FFPE: Formalin-fixed, paraffin-embedded; SGCD: Sarcoglycan delta; PI3K: Phosphatidylinositol 3-kinase.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-022-01277-x.

Additional file1: Figure S1. Pathological fracture of left humerus caused by Gorham-Stout disease in a 45-year-old white female patient. Sequential radiographs over a 1-year period show gradual disappearance of the proximal humerus.

Additional file 2. Supplementary Methods. Detailed information about sample processing for DNA, RNA sequencing and IHC analysis.

Additional file3: Table S1. Table of gene fusions detected and selected for validation including chromosome information and the number of reads in DNA and RNA level.

Additional file4: Figure S2. Gel electrophoresis containing control, Gorham-Stout tissue, and negative control of the gene fusion candidates. In the second gel only one set of amplified products was used in the final photograph as 19.Control FRS2-AC004485.1, 20.GS-FRS2-AC004485.1 21.Negative PCR Control (FRS2-AC004485.1). The wells 19, 20, and 21 were cropped in the final photograph including the PCR marker. A quick load PCR marker was used from NEB #N0475.

Additional file5: Table S2. Table results of the gene set expression analysis for lymphangiogenesis showing the affected genes in the RNA-seq expression and their contribution to the pathway.

Additional file6: Table S3. Table results of the gene set expression analysis for osteolysis showing the affected genes in the RNA-seq expression and their contribution to the pathway.

Additional file7: Figure S3. Barplot of RNA-seq expression (TPM) of VEGFC/D and NOTCH 2/3/4 genes in Gorham-Stout tissue compared with adjacent normal.

Additional file8: Figure S4. Gene fusion between ATG101 and SLC4A8 detected in Gorham-Stout patient. RNA evidence reads: 118 encompasing and 138 spanning reads. DNA evidence reads: 411 encompassing and 139 spanning reads in Gorham-Stout tissue. 0 reads found in normal tissue.

Acknowledgements

The study was supported by the project 'International Centre for Cancer Vaccine Science' that is carried out within the International Agendas Programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. The authors would like to thank the PL-Grid Infrastructure and CI-TASK, Poland for providing their hardware and software resources.

Author contributions

MYM, JAA, TH, and DS conceived of, initiated, coordinated, and supervised the project. LH, AO and JP were involved in acquisition and processing of the samples as well as the interpretation of the data throughout the project. MYM, AM, and JAA were involved in the genomics analysis and interpretation in the project. The IHC analysis was performed by SAS, MG, and MJA. The first draft of the manuscript was written by MYM, SAS, MG, AA, TH, MJA, and JAA

Subsequently, the manuscript was substantially revised and approved by all authors. All authors read and approved the final manuscript.

Funding
The study was supported by the project "International Centre for Cancer Vaccine Science" that is carried out within the International Agendas Programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. The funding body was involved in the design of the study, analysis, interpretation of data, and in writing this manuscript.

Availability of data and materials
The raw datasets generated and analyzed during the current study are not publicly available in order to protect participant confidentiality. The datasets obtained during the current study are available from the corresponding author if the requirements are reasonable.

Declarations

Ethics approval and consent to participate
The study was conducted according to the guidelines of general approval for use of surgically obtained tissue and approved by NHS Lothian NRS BioResource and the Public Health Office. The Lothian NRS BioResource is a HRA approved research tissue bank (REC Ref: 20/ES/0061, previously 15/ES/0094). This approval was given by East of Scotland Research Ethics Service REC 1. As part of this approval, the REC Committee confirmed that the favorable ethical opinion also applies to all research projects conducted in the UK using tissue or data supplied by the tissue bank, subject to the submission of an approved application to the tissue bank.

Consent for publication
Written informed consent for publication of their clinical details and genomic data was obtained from the patient. A copy of the consent form is available upon request from the corresponding authors.

Competing interests
All authors declare no competing interest in this manuscript.

Author details
[1]International Center for Cancer Vaccine Science (ICCVS), University of Gdansk, Gdańsk, Poland. [2]Edinburgh Pathology, Institute of Genetics and Cancer (IGC), University of Edinburgh, Edinburgh, Scotland. [3]Department of Basic Medical Sciences, Faculty of Medicine, The Hashemite University, Zarqa, Jordan. [4]MRC Human Genetics Unit, MRC Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, Scotland. [5]Department of Pathology, Royal Infirmary of Edinburgh, Edinburgh, Scotland. [6]Department of Orthopaedic Surgery, Royal Infirmary of Edinburgh, Edinburgh, Scotland.

Received: 25 February 2022 Accepted: 23 May 2022
Published online: 06 June 2022

References

1. Gorham LW, Stout AP. Massive osteolysis (acute spontaneous absorption of bone, phantom bone, disappearing bone); its relation to hemangiomatosis. J Bone Joint Surg Am. 1955;37-A:985–1004.
2. Patel DV. Gorham's disease or massive osteolysis. Clin Med Res. 2005;3:65–74.
3. Li M, et al. Successful management of Gorham-Stout disease in scapula and ribs: a case report and literature review. Orthop Surg. 2018;10:276–80.
4. Yerganyan VV, Body JJ, De Saint-Aubain N, Gebhart M. Gorham-Stout disease of the proximal fibula treated with radiotherapy and zoledronic acid. J Bone Oncol. 2015;4:42–6.
5. Liang Y, et al. Gorham-Stout disease successfully treated with sirolimus (rapamycin): a case report and review of the literature. BMC Musculoskelet Disord. 2020;21:577.
6. Nozawa A, et al. A somatic activating KRAS variant identified in an affected lesion of a patient with Gorham-Stout disease. J Hum Genet. 2020;65:995–1001.
7. Aoki Y, Niihori T, Inoue S, Matsubara Y. Recent advances in RASopathies. J Hum Genet. 2016;61:33–9.
8. Muñoz-Maldonado C, Zimmer Y, Medová M. A comparative analysis of individual ras mutations in cancer biology. Front Oncol. 2019. https://doi.org/10.3389/fonc.2019.01088.
9. Nguyen H-L, Boon LM, Vikkula M. Vascular anomalies caused by abnormal signaling within endothelial cells: targets for novel therapies. Semin Interv Radiol. 2017;34:233–8.
10. Homayun-Sepehr N, et al. *KRAS*-driven model of Gorham-Stout disease effectively treated with trametinib. JCI Insight. 2021. https://doi.org/10.1172/jci.insight.149831.
11. Zheng C, et al. Gorham-Stout disease of the malleolus: a rare case report. BMC Musculoskelet Disord. 2019;21:3.
12. Steele CD, et al. Undifferentiated sarcomas develop through distinct evolutionary pathways. Cancer Cell. 2019;35:441-456.e8.
13. Hünten S, Hermeking H. p53 directly activates cystatin D/CST5 to mediate mesenchymal-epithelial transition: a possible link to tumor suppression by vitamin D3. Oncotarget. 2015;6:15842–56.
14. Yang Z, et al. UNC5B promotes vascular endothelial cell senescence via the ROS-mediated P53 pathway. Oxid Med Cell Longev. 2021;2021:5546711.
15. Mercer CA, Kaliappan A, Dennis PB. A novel, human Atg13 binding protein, Atg101, interacts with ULK1 and is essential for macroautophagy. Autophagy. 2009;5:649–62.
16. Durbeej M, Campbell KP. Muscular dystrophies involving the dystrophin-glycoprotein complex: an overview of current mouse models. Curr Opin Genet Dev. 2002;12:349–61.
17. Cox ML, et al. Exome sequencing reveals independent SGCD deletions causing limb girdle muscular dystrophy in Boston terriers. Skelet Muscle. 2017;7:15.
18. Townsend D. Finding the sweet spot: assembly and glycosylation of the dystrophin-associated glycoprotein complex. Anat Rec. 2014;2007(297):1694–705.
19. Younus M, et al. SGCD homozygous nonsense mutation (p.Arg97(*)) causing limb-girdle muscular dystrophy type 2F (LGMD2F) in a consanguineous family: a case report. Front Genet. 2018;9:727.
20. Mavrogenis AF, Zambirinis CP, Dimitriadis PA, Tsakanikas A, Papagelopoulos PJ. Gorham-stout disease. J Surg Orthop Adv. 2010;19:85–90.
21. Carracedo A, Pandolfi PP. The PTEN–PI3K pathway: of feedbacks and cross-talks. Oncogene. 2008;27:5527–41.
22. Chalhoub N, Baker SJ. PTEN and the PI3-kinase pathway in cancer. Annu Rev Pathol. 2009;4:127–50.
23. Karar J, Maity A. PI3K/AKT/mTOR pathway in angiogenesis. Front Mol Neurosci. 2011;4:51.
24. Pandey AK, et al. Mechanisms of VEGF (vascular endothelial growth factor) inhibitor-associated hypertension and vascular disease. Hypertension. 2018. https://doi.org/10.1161/HYPERTENSIONAHA.117.10271.
25. Liu T, Zhang L, Joo D, Sun S-C. NF-κB signaling in inflammation. Signal Transduct Target Ther. 2017;2:1–9.
26. Timmer AM, Nizet V. IKKβ/NF-κB and the miscreant macrophage. J Exp Med. 2008;205:1255–9.
27. Moll UM, Petrenko O. The MDM2-p53 interaction. Mol Cancer Res. 2003;1:1001–8.
28. Hosokawa N, et al. Atg101, a novel mammalian autophagy protein interacting with Atg13. Autophagy. 2009;5:973–9.
29. Richards S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405–23.
30. Passarelli C, et al. Tumor necrosis factor receptor SF10A (TNFRSF10A) SNPs correlate with corticosteroid response in duchenne muscular dystrophy. Front Genet. 2020;11:605.
31. Hughes AE, et al. Mutations in TNFRSF11A, affecting the signal peptide of RANK, cause familial expansile osteolysis. Nat Genet. 2000;24:45–8.
32. Luks VL, et al. Lymphatic and other vascular malformative/overgrowth disorders are caused by somatic mutations in PIK3CA. J Pediatr. 2015;166(1048–1054):e1-5.
33. Hopman SMJ, et al. PTEN hamartoma tumor syndrome and Gorham-Stout phenomenon. Am J Med Genet A. 2012;158A:1719–23.
34. Ozeki M, Fukao T. Generalized lymphatic anomaly and gorham-stout disease: overview and recent insights. Adv Wound Care. 2019;8:230–45.

35. Mathew M, Goyal A, Khan A, Yuen T. Drugs for rare diseases of bone. In: Zaidi M, editor. Encyclopedia of bone biology. Cambridge: Academic Press; 2020. p. 711–22. https://doi.org/10.1016/B978-0-12-801238-3.62273-0.
36. Rössler J, Saueressig U, Kayser G, von Winterfeld M, Klement GL. Personalized therapy for generalized lymphatic anomaly/gorham-stout disease with a combination of Sunitinib and Taxol. J Pediatr Hematol Oncol. 2015;37: e481.
37. Hammer F, et al. Gorham-Stout disease-stabilization during bisphosphonate treatment. J Bone Miner Res Off J Am Soc Bone Miner Res. 2005;20:350–3.
38. Nozawa A, et al. Gorham-stout disease of the skull base with hearing loss: dramatic recovery and antiangiogenic therapy. Pediatr Blood Cancer. 2016;63:931–4.
39. Shangary S, Wang S. Targeting the MDM2-p53 interaction for cancer therapy. Clin Cancer Res Off J Am Assoc Cancer Res. 2008;14:5318–24.
40. Chène P. Inhibiting the p53–MDM2 interaction: an important target for cancer therapy. Nat Rev Cancer. 2003;3:102–9.
41. Faruqi T, et al. Molecular, phenotypic aspects and therapeutic horizons of rare genetic bone disorders. BioMed Res Int. 2014;2014: e670842.
42. Yeter HH. Gorham-Stout disease or new entity on the basis of vasculopathy. Alex J Med. 2017;53:193–6.
43. Colucci S, et al. Gorham-stout syndrome: a monocyte-mediated cytokine propelled disease. J Bone Miner Res. 2006;21:207–18.
44. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, McQuaid S, Gray RT, Murray LJ, Coleman HG, James JA, Salto-Tellez M, Hamilton PW. QuPath: Open source software for digital pathology image analysis. Sci Rep. 2017. https://doi.org/10.1038/s41598-017-17204-5.
45. Berben L, Wildiers H, Marcelis L, Antoranz A, Bosisio F, Hatse S, Floris G. Computerised scoring protocol for identification and quantification of different immune cell populations in breast tumour regions by the use of QuPath software. Histopathology. 2020;77(1):79–91. https://doi.org/10.1111/his.14108.
46. Al Shboul S, Curran OE, Alfaro JA, Lickiss F, Nita E, Kowalski J, Naji F, Nenutil R, Ball KL, Krejcir R, Vojtesek B, Hupp TR, Brennan PM. Kinomics platform using GBM tissue identifies BTK as being associated with higher patient survival. Life Sci Alliance. 2021;4(12):e202101054. https://doi.org/10.26508/lsa.202101054.

Publisher's Note

6. Conclusion

Multi-omic research has proved to be one of the most useful methods to characterize complex diseases like cancer, presenting a technical challenge as well when dealing with the complexity of data combination.

Chapter 3 focuses on esophageal adenocarcinoma, demonstrating the challenges that must be overcome to integrate proteomics with other -omic datasets. In a maturing field, multiple mass spectrometry methods exist to both generate and process mass-spectrometry data. Normalization across all samples in the dataset layers allowed us to develop an integrated analysis of the changes between RNA and protein gene expressions. The integration of RNA and protein intensities is one of the unexplored analyses for most tissues and diseases due to the challenge to correlate these two layers of the central dogma. In our analysis, we were able to compare multiple normal tissues (including esophageal) with the integrated multi-omic analysis of esophageal adenocarcinoma patients. As a result, we identified genes that would have been missed through single-omic studies, containing dysregulated abundances between the proteome and transcriptome. We hypothesized that oncogenic genes with low RNA and high protein expression evade the regulatory processes of the cell through tumor evasion mechanisms, causing the high protein expression of these genes and therefore the malignancy in the tissue.

While the esophageal adenocarcinoma study utilized proteomics as the main source of information to develop a multi-omics approach, the study of the landscape in undifferentiated pleomorphic sarcoma used whole-exome sequencing as guidance. The exploration of DNA sequencing (Chapter 4.3.1) showed shared mutations across the UPS patients where a mutually exclusive mutation of TP53 and ATRX was to be the most common event dominant. Further exploration of the copy number alteration revealed the deletion of DNA areas coding for the tumor suppressor genes TP53 and RB1 for most of the patients (Chapter 4.3.4). Based on this discovery we assessed the presence/absence of both genes by IHC staining of the UPS samples and developed a combined therapeutic strategy to reduce the spread of the disease (Chapter 4.3.5). Lastly, due to the heterogeneity of the mutational landscape in UPS (Chapter 4.3.2-4.3.3), we proposed personalized

neoantigen therapies as a possible approach to tackle the disease. By integrating both DNA and RNA sequencing with mass spectrometry proteomics we obtained possible targets for the development of new therapies (Chapter 4.3.7). Altogether the study of UPS has relied heavily on the genomics exploration, showing that DNA-sequencing can be used to predict mutated proteins uniquely expressed by sarcoma cells and having supporting data from the transcriptomics and proteomics analyses.

In the last study, we performed a unique investigation of the mechanisms of Gorham-Stout disease. For the first time reported in the literature, we have provided DNA and RNA sequencing information of a Gorham-Stout patient with attached normal tissue. Furthermore, the combination of the two levels of information has shown hints at the mutational profile of the disease, revealing structural events as possible drivers of the syndrome. The success of the study relies on the top to bottom multi-omic integration, starting at the genomics level, and going through the transcriptomics until its final validation. Besides this, the major limitation is caused by the absence of multiple patients due to the rareness of the disease and the difficulty of publicly sharing anonymous data from the literature.

With a growing influence on the current studies in the field, multi-omics analyses are a useful resource to further interrogate the origin of a disease. One of the most important considerations in these procedures is the design of the experiment. A careful and consistent analysis of the samples will solve most of the complexity encountered when integrating multiple layers of information. Computationally, heterogeneity in data due to sample preparation can be handled using normalization, but this experimental heterogeneity can sometimes be difficult to overcome resulting in reduced sensitivity of the analysis.

Overall, we have studied, developed, and implemented the most up-to-date best practices for a multi-omic approach to three different diseases. We have used this integrated strategy to discover new possible causes of Gorham-Stout syndrome, explored new therapy strategies in undifferentiated pleomorphic sarcoma, and revealed biomarkers that could lead to possible therapeutic targets in esophageal adenocarcinoma.

Furthermore, it has demonstrated the functionality of multi-omic integration and its applicability, dealing with the complexity that resides in the multiple cases presented.

7. Bibliography

1. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
3. Pudata, V. & v, S. V. A Short Note on Cancer. *J. Carcinog. Mutagen.* **02**, (2012).
4. Teer, J. K. An improved understanding of cancer genomics through massively parallel sequencing. *Transl. Cancer Res.* **3**, 243–259 (2014).
5. Dang, C. V. Links between metabolism and cancer. *Genes Dev.* **26**, 877–890 (2012).
6. Greenspan, A. *Differential diagnosis of tumors and tumor-like lesions of bones and joints*. (Thieme, 2000).
7. EDMONDSON, H. A. DIFFERENTIAL DIAGNOSIS OF TUMORS AND TUMOR-LIKE LESIONS OF LIVER IN INFANCY AND CHILDHOOD. *AMA J. Dis. Child.* **91**, 168–186 (1956).
8. Tamboli, P., Ro, J. Y., Amin, M. B., Ligato, S. & Ayala, A. G. Benign tumors and tumor-like lesions of the adult kidney. Part II: Benign mesenchymal and mixed neoplasms, and tumor-like lesions. *Adv. Anat. Pathol.* **7**, 47–66 (2000).
9. Inazawa, J., Miki, Y. & Nakamura, Y. Integrative cancer genomics in the era of precision cancer medicine. *J. Hum. Genet.* **66**, 843 (2021).
10. Mendiratta, G. *et al.* Cancer gene mutation frequencies for the U.S. population. *Nat. Commun.* **12**, 5961 (2021).
11. Maxwell, K. N. *et al.* BRCA locus-specific loss of heterozygosity in germline BRCA1 and BRCA2 carriers. *Nat. Commun.* **8**, 319 (2017).

12. Levis, M. FLT3 mutations in acute myeloid leukemia: what is the best approach in 2013? *Hematol. Educ. Program Am. Soc. Hematol. Am. Soc. Hematol. Educ. Program* **2013**, 220–226 (2013).

13. Soussi, T. & Wiman, K. G. TP53: an oncogene in disguise. *Cell Death Differ.* **22**, 1239–1249 (2015).

14. Nagy, Á., Munkácsy, G. & Győrffy, B. Pancancer survival analysis of cancer hallmark genes. *Sci. Rep.* **11**, 6047 (2021).

15. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

16. Tiong, K.-L. & Yeang, C.-H. Explaining cancer type specific mutations with transcriptomic and epigenomic features in normal tissues. *Sci. Rep.* **8**, 11456 (2018).

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

18. Rudnick, P. A. *et al.* A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. *J. Proteome Res.* **15**, 1023–1032 (2016).

19. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database J. Biol. Databases Curation* **2011**, bar026 (2011).

20. Henry, N. L. & Hayes, D. F. Cancer biomarkers. *Mol. Oncol.* **6**, 140–146 (2012).

21. Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C. & Chinnaiyan, A. M. RNA-Seq Accurately Identifies Cancer Biomarker Signatures to Distinguish Tissue of Origin.

Neoplasia N. Y. N **16**, 918–927 (2014).

22. Kung, C.-P., Maggi, L. B. & Weber, J. D. The Role of RNA Editing in Cancer Development and Metabolic Disorders. *Front. Endocrinol.* **9**, 762 (2018).

23. Deelen, P. *et al.* Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* **7**, 30 (2015).

24. Jehl, F. *et al.* RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. *Front. Genet.* **12**, 1104 (2021).

25. Sager, M. *et al.* Transcriptomics in cancer diagnostics: developments in technology, clinical research and commercialization. *Expert Rev. Mol. Diagn.* **15**, 1589–1603 (2015).

26. Hong, M. *et al.* RNA sequencing: new technologies and applications in cancer research. *J. Hematol. Oncol.J Hematol Oncol* **13**, 166 (2020).

27. Bagnoli, J. W., Wange, L. E., Janjic, A. & Enard, W. Studying Cancer Heterogeneity by Single-Cell RNA Sequencing. in *Lymphoma: Methods and Protocols* (ed. Küppers, R.) 305–319 (Springer, 2019). doi:10.1007/978-1-4939-9151-8_14.

28. Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat. Commun.* **12**, 2540 (2021).

29. Tsimberidou, A. M., Fountzilas, E., Bleris, L. & Kurzrock, R. Transcriptomics and solid tumors: The next frontier in precision cancer medicine. *Semin. Cancer Biol.*

(2020) doi:10.1016/j.semcancer.2020.09.007.

30. Omenn, G. S. *et al.* Research on the Human Proteome Reaches a Major Milestone: >90% of Predicted Human Proteins Now Credibly Detected, According to the HUPO Human Proteome Project. *J. Proteome Res.* **19**, 4735–4746 (2020).

31. KOLCH, W., MISCHAK, H. & PITT, A. R. The molecular make-up of a tumour: proteomics in cancer research. *Clin. Sci.* **108**, 369–383 (2005).

32. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

33. Li, J. *et al.* TCPA: a resource for cancer functional proteomics data. *Nat. Methods* **10**, 1046–1047 (2013).

34. Chen, M.-J. M. *et al.* TCPA v3.0: An Integrative Platform to Explore the Pan-Cancer Analysis of Functional Proteomic Data. *Mol. Cell. Proteomics MCP* **18**, S15–S25 (2019).

35. Adelmant, G., Garg, B. K., Tavares, M., Card, J. D. & Marto, J. A. Tandem Affinity Purification and Mass Spectrometry (TAP-MS) for the Analysis of Protein Complexes. *Curr. Protoc. Protein Sci.* **96**, e84 (2019).

36. Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **21**, 255–261 (2003).

37. Kwon, Y. W. *et al.* Application of Proteomics in Cancer: Recent Trends and Approaches for Biomarkers Discovery. *Front. Med.* **8**, 1644 (2021).

38. Krassowski, M., Das, V., Sahu, S. K. & Misra, B. B. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Front.*

Genet. **11**, 1598 (2020).

39. Frankell, A. M. *et al.* The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nat. Genet.* **51**, 506–516 (2019).

40. Mirza, B. *et al.* Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes* **10**, 87 (2019).

41. Goldfeder, R. L. *et al.* Medical implications of technical accuracy in genome sequencing. *Genome Med.* **8**, 24 (2016).

42. Su, Z. *et al.* A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).

43. Wilkerson, M. D. *et al.* Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res.* **42**, e107 (2014).

44. Siegel, M. B. *et al.* Integrated RNA and DNA sequencing reveals early drivers of metastatic breast cancer. *J. Clin. Invest.* **128**, 1371–1383 (2018).

45. Wu, C. *et al.* Integrated genome and transcriptome sequencing identifies a novel form of hybrid and aggressive prostate cancer. *J. Pathol.* **227**, 53–61 (2012).

46. Liu, J. *et al.* Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. *Nat. Commun.* **5**, 3830 (2014).

47. Wu, H., Li, X. & Li, H. Gene fusions and chimeric RNAs, and their implications in cancer. *Genes Dis.* **6**, 385–390 (2019).

48. Zhang, J. & Maher, C. A. Gene Fusion Discovery with INTEGRATE. *Methods*

Mol. Biol. Clifton NJ **2079**, 41–68 (2020).

49. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).

50. Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. *Nat. Commun.* **5**, 4846 (2014).

51. Schram, A. M., Chang, M. T., Jonsson, P. & Drilon, A. Fusions in solid tumours: diagnostic strategies, targeted therapy, and acquired resistance. *Nat. Rev. Clin. Oncol.* **14**, 735–748 (2017).

52. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371–381 (2015).

53. Gerstung, M. *et al.* Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.* **6**, 5901 (2015).

54. Jia, P. & Zhao, Z. Impacts of somatic mutations on gene expression: an association perspective. *Brief. Bioinform.* **18**, 413–425 (2017).

55. Calabrese, C. *et al.* Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).

56. Haynes, B. C. *et al.* An Integrated Next-Generation Sequencing System for Analyzing DNA Mutations, Gene Fusions, and RNA Expression in Lung Cancer. *Transl. Oncol.* **12**, 836–845 (2019).

57. Wang, X. *et al.* Integrated analysis of transcriptomic and proteomic data from tree peony (P. ostii) seeds reveals key developmental stages and candidate genes related to oil biosynthesis and fatty acid metabolism. *Hortic. Res.* **6**, 1–19 (2019).

58. Aslam, B., Basit, M., Nisar, M. A., Khurshid, M. & Rasool, M. H. Proteomics: Technologies and Their Applications. *J. Chromatogr. Sci.* **55**, 182–196 (2017).

59. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **4**, 117 (2003).

60. Haider, S. & Pal, R. Integrated Analysis of Transcriptomic and Proteomic Data. *Curr. Genomics* **14**, 91–110 (2013).

61. Yang, W. *et al.* Integrating proteomics and transcriptomics for the identification of potential targets in early colorectal cancer. *Int. J. Oncol.* **55**, 439–450 (2019).

62. Syafruddin, S. E., Nazarie, W. F. W. M., Moidu, N. A., Soon, B. H. & Mohtar, M. A. Integration of RNA-Seq and proteomics data identifies glioblastoma multiforme surfaceome signature. *BMC Cancer* **21**, 850 (2021).

63. Jaffe, J. D., Berg, H. C. & Church, G. M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4**, 59–77 (2004).

64. Nesvizhskii, A. I. Proteogenomics: concepts, applications, and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014).

65. Song, M. *et al.* A Review of Integrative Imputation for Multi-Omics Datasets. *Front. Genet.* **11**, 1215 (2020).

66. Kopajtich, R. *et al. Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders*. 2021.03.09.21253187 https://www.medrxiv.org/content/10.1101/2021.03.09.21253187v1 (2021) doi:10.1101/2021.03.09.21253187.

67. Chakraborty, S., Hosen, M. I., Ahmed, M. & Shekhar, H. U. Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. *BioMed Res. Int.* **2018**, 9836256 (2018).

68. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).

69. Ren, S. *et al.* Integration of Metabolomics and Transcriptomics Reveals Major Metabolic Pathways and Potential Biomarker Involved in Prostate Cancer. *Mol. Cell. Proteomics MCP* **15**, 154–163 (2016).

70. Dou, Y. *et al.* Proteogenomic Characterization of Endometrial Carcinoma. *Cell* **180**, 729-748.e26 (2020).

71. Lauss, M. *et al.* Monitoring of Technical Variation in Quantitative High-Throughput Datasets. *Cancer Inform.* **12**, 193–201 (2013).

72. Wang, X., Qiao, J. & Wang, R. Exploration and validation of a novel prognostic signature based on comprehensive bioinformatics analysis in hepatocellular carcinoma. *Biosci. Rep.* **40**, BSR20203263 (2020).

73. Manzoni, C. *et al.* Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.* **19**, 286–302 (2018).

74. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).

75. Wild, C. P. Complementing the genome with an 'exposome': the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev.*

Oncol. **14**, 1847–1850 (2005).

76. Du, Y., Fan, K., Lu, X. & Wu, C. Integrating Multi–Omics Data for Gene-Environment Interactions. *BioTech* **10**, 3 (2021).

77. Ruggles, K. V. *et al.* An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer *. *Mol. Cell. Proteomics* **15**, 1060–1071 (2016).

78. Alfaro, J. A. *et al.* Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines. *Genome Med.* **9**, 62 (2017).

79. Lin, T.-T. *et al.* Mass spectrometry-based targeted proteomics for analysis of protein mutations. *Mass Spectrom. Rev.* e21741 (2021) doi:10.1002/mas.21741.

80. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173**, 305-320.e10 (2018).

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

82. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-337.e10 (2018).

83. Then, E. O. *et al.* Esophageal Cancer: An Updated Surveillance Epidemiology and End Results Database Analysis. *World J. Oncol.* **11**, 55–64 (2020).

84. Domper Arnal, M. J., Ferrández Arenas, Á. & Lanas Arbeloa, Á. Esophageal cancer: Risk factors, screening and endoscopic treatment in Western and Eastern countries. *World J. Gastroenterol.* **21**, 7933–7943 (2015).

85. Pera, M., Manterola, C., Vidal, O. & Grande, L. Epidemiology of esophageal

adenocarcinoma. *J. Surg. Oncol.* **92**, 151–159 (2005).

86. Napier, K. J., Scheerer, M. & Misra, S. Esophageal cancer: A Review of epidemiology, pathogenesis, staging workup and treatment modalities. *World J. Gastrointest. Oncol.* **6**, 112–120 (2014).

87. Jain, S. & Dhingra, S. Pathology of esophageal cancer and Barrett's esophagus. *Ann. Cardiothorac. Surg.* **6**, 99–109 (2017).

88. Shaheen, N. & Ransohoff, D. F. Gastroesophageal Reflux, Barrett Esophagus, and Esophageal CancerScientific Review. *JAMA* **287**, 1972–1981 (2002).

89. Naini, B. V., Souza, R. F. & Odze, R. D. Barrett's Esophagus: A Comprehensive and Contemporary Review for Pathologists. *Am. J. Surg. Pathol.* **40**, e45–e66 (2016).

90. Lagergren, J. Adenocarcinoma of oesophagus: what exactly is the size of the problem and who is at risk? *Gut* **54**, i1–i5 (2005).

91. Liu, L. *et al.* Significance of the Depth of Tumor Invasion and Lymph Node Metastasis in Superficially Invasive (T1) Esophageal Adenocarcinoma. *Am. J. Surg. Pathol.* **29**, 1079–1085 (2005).

92. Prasad, G. A. *et al.* Endoscopic and Surgical Treatment of Mucosal (T1a) Esophageal Adenocarcinoma in Barrett's Esophagus. *Gastroenterology* **137**, 815–823 (2009).

93. Skinner, H. D. *et al.* Metformin use and improved response to therapy in esophageal adenocarcinoma. *Acta Oncol.* **52**, 1002–1009 (2013).

94. Walsh, T. N., Grennell, M., Mansoor, S. & Kelly, A. Neoadjuvant treatment of advanced stage esophageal adenocarcinoma increases survival*. *Dis. Esophagus* **15**, 121–

124 (2002).

95. Talukdar, F. R. *et al.* Molecular landscape of esophageal cancer: implications for early detection and personalized therapy. *Ann. N. Y. Acad. Sci.* **1434**, 342–359 (2018).

96. Derouet, M. F. *et al.* Towards personalized induction therapy for esophageal adenocarcinoma: organoids derived from endoscopic biopsy recapitulate the pre-treatment tumor. *Sci. Rep.* **10**, 14514 (2020).

97. Ross-Innes, C. S. *et al.* Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat. Genet.* **47**, 1038–1046 (2015).

98. Dulak, A. M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478–486 (2013).

99. Wang, X. *et al.* Copy number alterations detected by whole-exome and whole-genome sequencing of esophageal adenocarcinoma. *Hum. Genomics* **9**, 22 (2015).

100. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

101. Kandoth, C. *mskcc/vcf2maf*. (Memorial Sloan Kettering, 2020).

102. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).

103. Schöbinger, M., Klein, O.-J. & Mitulović, G. Low-Temperature Mobile Phase for Peptide Trapping at Elevated Separation Temperature Prior to Nano RP-HPLC-MS/MS. *Separations* **3**, 6 (2016).

104. Tóth, G., Panić-Janković, T. & Mitulović, G. Pillar array columns for peptide separations in nanoscale reversed-phase chromatography. *J. Chromatogr. A* **1603**, 426–432 (2019).

105. O'Neill, J. R. *et al.* Quantitative shotgun proteomics unveils candidate novel esophageal adenocarcinoma (EAC)-specific proteins. *Mol. Cell. Proteomics* (2017) doi:10.1074/mcp.M116.065078.

106. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).

107. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* **5**, (2014).

108. Röst, H. L., Schmitt, U., Aebersold, R. & Malmström, L. pyOpenMS: A Python-based interface to the OpenMS mass-spectrometry algorithm library. *PROTEOMICS* **14**, 74–77 (2014).

109. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, (2019).

110. Jiang, L. *et al.* A Quantitative Proteome Map of the Human Body. *Cell* **183**, 269-283.e19 (2020).

111. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

112. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

113. *Introduction to meta-analysis*. (John Wiley & Sons, 2009).

114. Grupp, K. *et al.* Reduced RBM3 expression is associated with aggressive tumor features in esophageal cancer but not significantly linked to patient outcome. *BMC Cancer* **18**, 1106 (2018).

115. Jonsson, L. *et al.* High expression of RNA-binding motif protein 3 in esophageal and gastric adenocarcinoma correlates with intestinal metaplasia-associated tumours and independently predicts a reduced risk of recurrence and death. *Biomark. Res.* **2**, 11 (2014).

116. Tang, W.-W., Liu, Z.-H., Yang, T.-X., Wang, H.-J. & Cao, X.-F. Upregulation of MAGEA4 correlates with poor prognosis in patients with early stage of esophageal squamous cell carcinoma. *OncoTargets Ther.* **9**, 4289–4293 (2016).

117. Zhang, Y., Zhang, Y. & Zhang, L. Expression of cancer–testis antigens in esophageal cancer and their progress in immunotherapy. *J. Cancer Res. Clin. Oncol.* **145**, 281–291 (2019).

118. Adaptimmune. *A Phase 2 Open-Label Clinical Trial of ADP-A2M4CD8 in Subjects With Advanced Esophageal or Esophagogastric Junction Cancers*. https://clinicaltrials.gov/ct2/show/NCT04752358 (2021).

119. Chen, H.-M. *et al.* Insulin-Like Growth Factor 2 mRNA-Binding Protein 1 (IGF2BP1) Is a Prognostic Biomarker and Associated with Chemotherapy Responsiveness in Colorectal Cancer. *Int. J. Mol. Sci.* **22**, 6940 (2021).

120. Heath, J. K. *et al.* The human A33 antigen is a transmembrane glycoprotein and a novel member of the immunoglobulin superfamily. *Proc. Natl. Acad. Sci. U. S. A.* **94**,

469–474 (1997).

121. Garinchesa, P. *et al.* Organ-specific expression of the colon cancer antigen A33, a cell surface target for antibody-based therapy. *Int. J. Oncol.* **9**, 465–471 (1996).

122. Wu, Z., Guo, H.-F., Xu, H. & Cheung, N.-K. V. Development of a Tetravalent Anti-GPA33/Anti-CD3 Bispecific Antibody for Colorectal Cancers. *Mol. Cancer Ther.* **17**, 2164–2175 (2018).

123. Moore, P. A. *et al.* Development of MGD007, a gpA33 x CD3-Bispecific DART Protein for T-Cell Immunotherapy of Metastatic Colorectal Cancer. *Mol. Cancer Ther.* **17**, 1761–1772 (2018).

124. Infante, J. R. *et al.* Safety, pharmacokinetics and pharmacodynamics of the anti-A33 fully-human monoclonal antibody, KRN330, in patients with advanced colorectal cancer. *Eur. J. Cancer Oxf. Engl. 1990* **49**, 1169–1175 (2013).

125. Hao, L. *et al.* Elevated GAPDH expression is associated with the proliferation and invasion of lung and esophageal squamous cell carcinomas. *Proteomics* **15**, 3087–3100 (2015).

126. Zhu, Y. *et al.* Identification of prothymosin alpha (PTMA) as a biomarker for esophageal squamous cell carcinoma (ESCC) by label-free quantitative proteomics and Quantitative Dot Blot (QDB). *Clin. Proteomics* **16**, 12 (2019).

127. Veremieva, M., Khoruzhenko, A., Zaicev, S., Negrutskii, B. & El'skaya, A. Unbalanced expression of the translation complex eEF1 subunits in human cardioesophageal carcinoma. *Eur. J. Clin. Invest.* **41**, 269–276 (2011).

128. Lin, J. *et al.* KRT 15 as a prognostic biomarker is highly expressed in esophageal

carcinoma. *Future Oncol.* **16**, 1903–1909 (2020).

129. Yang, F. *et al.* Identification of a Five-Gene Prognostic Model and Its Potential Drug Repurposing in Colorectal Cancer Based on TCGA, GTEx and GEO Databases. *Front. Genet.* **11**, (2021).

130. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, (2017).

131. Shibata, Y. *et al.* Chfr expression is downregulated by CpG island hypermethylation in esophageal cancer. *Carcinogenesis* **23**, 1695–1699 (2002).

132. Soutto, M. *et al.* Epigenetic and genetic silencing of CHFR in esophageal adenocarcinomas. *Cancer* **116**, 4033–4042 (2010).

133. Shan, L. *et al.* CENPE promotes lung adenocarcinoma proliferation and is directly regulated by FOXM1. *Int. J. Oncol.* **55**, 257–266 (2019).

134. Zhu, X. *et al.* CENPE expression is associated with its DNA methylation status in esophageal adenocarcinoma and independently predicts unfavorable overall survival. *PLOS ONE* **14**, e0207341 (2019).

135. Dong, X., Han, Y., Sun, Z. & Xu, J. Actin Gamma 1, a new skin cancer pathogenic gene, identified by the biological feature-based classification. *J. Cell. Biochem.* **119**, 1406–1419 (2018).

136. Gao, B., Li, S., Tan, Z., Ma, L. & Liu, J. ACTG1 and TLR3 are biomarkers for alcohol-associated hepatocellular carcinoma. *Oncol. Lett.* **17**, 1714–1722 (2019).

137. Wang, S., Li, M., Xing, L. & Yu, J. High expression level of peptidylprolyl isomerase A is correlated with poor prognosis of liver hepatocellular carcinoma. *Oncol. Lett.* **18**, 4691–4702 (2019).

138. Jin, X. *et al.* Elevated expression of GNAS promotes breast cancer cell proliferation and migration via the PI3K/AKT/Snail1/E-cadherin axis. *Clin. Transl. Oncol. Off. Publ. Fed. Span. Oncol. Soc. Natl. Cancer Inst. Mex.* **21**, 1207–1219 (2019).

139. Zhang, Y. *et al.* BTF3 confers oncogenic activity in prostate cancer through transcriptional upregulation of Replication Factor C. *Cell Death Dis.* **12**, 1–15 (2021).

140. Xu, L., Li, H., Wu, L. & Huang, S. YBX1 promotes tumor growth by elevating glycolysis in human bladder cancer. *Oncotarget* **8**, 65946–65956 (2017).

141. Kuwano, M., Shibata, T., Watari, K. & Ono, M. Oncogenic Y-box binding protein-1 as an effective therapeutic target in drug-resistant cancer. *Cancer Sci.* **110**, 1536–1543 (2019).

142. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644 (2020).

143. Kosti, I., Jain, N., Aran, D., Butte, A. J. & Sirota, M. Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci. Rep.* **6**, 24799 (2016).

144. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).

145. Nowicki-Osuch, K. *et al.* Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition. *Science* **373**, 760–767 (2021).

146. Koul, H. K., Pal, M. & Koul, S. Role of p38 MAP Kinase Signal Transduction in Solid Tumors. *Genes Cancer* **4**, 342–359 (2013).

147. Zou, X. & Blank, M. Targeting p38 MAP kinase signaling in cancer through post-translational modifications. *Cancer Lett.* **384**, 19–26 (2017).

148. Rochette, L. *et al.* Mitochondrial SLC25 Carriers: Novel Targets for Cancer Therapy. *Molecules* **25**, 2417 (2020).

149. Gundamaraju, R., Lu, W. & Manikam, R. CHCHD2: The Power House's Potential Prognostic Factor for Cancer? *Front. Cell Dev. Biol.* **0**, (2021).

150. O'Neill, J. R. *et al.* Quantitative Shotgun Proteomics Unveils Candidate Novel Esophageal Adenocarcinoma (EAC)-specific Proteins. *Mol. Cell. Proteomics MCP* **16**, 1138–1150 (2017).

151. Guo, J., Jia, J. & Jia, R. PTBP1 and PTBP2 impaired autoregulation of SRSF3 in cancer cells. *Sci. Rep.* **5**, 14548 (2015).

152. Sun, Z., Liu, J., Jing, H., Dong, S.-X. & Wu, J. The diagnostic and prognostic value of CHFR hypermethylation in colorectal cancer, a meta-analysis and literature review. *Oncotarget* **8**, 89142–89148 (2017).

153. Cheung, H. C. *et al.* Splicing factors PTBP1 and PTBP2 promote proliferation and migration of glioma cell lines. *Brain* **132**, 2277–2288 (2009).

154. Nolte, W., Weikard, R., Albrecht, E., Hammon, H. M. & Kühn, C. Metabogenomic analysis to functionally annotate the regulatory role of long non-coding RNAs in the liver of cows with different nutrient partitioning phenotype. *Genomics* **114**, 202–214 (2022).

155. D'Angelo, S. P., Tap, W. D., Schwartz, G. K. & Carvajal, R. D. Sarcoma immunotherapy: past approaches and future directions. *Sarcoma* **2014**, 391967 (2014).

156. Klug, L. R. & Heinrich, M. C. PDGFRA Antibody for Soft Tissue Sarcoma. *Cell* **168**, 555 (2017).

157. Taylor, B. S. *et al.* Advances in sarcoma genomics and new therapeutic targets. *Nat. Rev. Cancer* **11**, 541–557 (2011).

158. Roland, C. L. *et al.* Analysis of Clinical and Molecular Factors Impacting Oncologic Outcomes in Undifferentiated Pleomorphic Sarcoma. *Ann. Surg. Oncol.* **23**, 2220–2228 (2016).

159. Judson, I. *et al.* Doxorubicin alone versus intensified doxorubicin plus ifosfamide for first-line treatment of advanced or metastatic soft-tissue sarcoma: a randomised controlled phase 3 trial. *Lancet Oncol.* **15**, 415–423 (2014).

160. Vitfell-Rasmussen, J. *et al.* A Phase I/II Clinical Trial of Belinostat (PXD101) in Combination with Doxorubicin in Patients with Soft Tissue Sarcomas. *Sarcoma* **2016**, e2090271 (2016).

161. Hofvander, J. *et al.* Recurrent PRDM10 Gene Fusions in Undifferentiated Pleomorphic Sarcoma. *Clin. Cancer Res.* **21**, 864–869 (2015).

162. Steele, C. D. *et al.* Undifferentiated Sarcomas Develop through Distinct Evolutionary Pathways. *Cancer Cell* **35**, 441-456.e8 (2019).

163. Li, G. Z. *et al.* Rb and p53-Deficient Myxofibrosarcoma and Undifferentiated Pleomorphic Sarcoma Require Skp2 for Survival. *Cancer Res.* **80**, 2461–2471 (2020).

164. Hames-Fathi, S., Nottley, S. W. G. & Pillay, N. Unravelling undifferentiated soft tissue sarcomas: insights from genomics. *Histopathology* **80**, 109–121 (2022).

165. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

166. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

167. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. 861054 (2019) doi:10.1101/861054.

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

169. Pinto, E. M. *et al.* Genomic landscape of paediatric adrenocortical tumours. *Nat. Commun.* **6**, 6302 (2015).

170. Rapa, E., Hill, S. K., Morten, K. J., Potter, M. & Mitchell, C. The over-expression of cell migratory genes in alveolar rhabdomyosarcoma could contribute to metastatic spread. *Clin. Exp. Metastasis* **29**, 419–429 (2012).

171. Griffith, M. *et al.* Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1**, 210–223 (2015).

172. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).

173. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).

174. Mroz, E. A. & Rocco, J. W. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol.* **49**, 211–215 (2013).

175. Aspberg, F. *et al.* Near-haploidy in two malignant fibrous histiocytomas. *Cancer Genet. Cytogenet.* **79**, 119–122 (1995).

176. Chen, C. *et al.* Next-generation-sequencing-based risk stratification and identification of new genes involved in structural and sequence variations in near haploid lymphoblastic leukemia. *Genes. Chromosomes Cancer* **52**, 564–579 (2013).

177. Zhao, S. *et al.* Mutational landscape of uterine and ovarian carcinosarcomas implicates histone genes in epithelial-mesenchymal transition. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 12238–12243 (2016).

178. Willems, P. *et al.* BOLA1 is an aerobic protein that prevents mitochondrial morphology changes induced by glutathione depletion. *Antioxid. Redox Signal.* **18**, 129–138 (2013).

179. Karagiannis, P. *et al.* IgG4 subclass antibodies impair antitumor immunity in melanoma. *J. Clin. Invest.* **123**, 1457–1474 (2013).

180. Lin, P. P. *et al.* Targeted mutation of p53 and Rb in mesenchymal cells of the limb bud produces sarcomas in mice. *Carcinogenesis* **30**, 1789–1795 (2009).

181. Lane, D. P. & Crawford, L. V. T antigen is bound to a host protein in SV40-transformed cells. *Nature* **278**, 261–263 (1979).

182. Linzer, D. I. & Levine, A. J. Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. *Cell* **17**, 43–52 (1979).

183. DeCaprio, J. A. *et al.* SV40 large tumor antigen forms a specific complex with the product of the retinoblastoma susceptibility gene. *Cell* **54**, 275–283 (1988).

184. Matushansky, I. *et al.* Derivation of sarcomas from mesenchymal stem cells via inactivation of the Wnt pathway. *J. Clin. Invest.* **117**, 3248–3257 (2007).

185. Fransson, Å. *et al.* Strong synergy with APR-246 and DNA-damaging drugs in primary cancer cells from patients with TP53 mutant High-Grade Serous ovarian cancer. *J. Ovarian Res.* **9**, 27 (2016).

186. Zawacka-Pankau, J. & Selivanova, G. Pharmacological reactivation of p53 as a strategy to treat cancer. *J. Intern. Med.* **277**, 248–259 (2015).

187. Deng, X. & Nakamura, Y. Cancer Precision Medicine: From Cancer Screening to Drug Selection and Personalized Immunotherapy. *Trends Pharmacol. Sci.* **38**, 15–24 (2017).

188. Zhang, X., Sharma, P. K., Peter Goedegebuure, S. & Gillanders, W. E. Personalized cancer vaccines: Targeting the cancer mutanome. *Vaccine* **35**, 1094–1100 (2017).

189. Nakatsura, T. Era of cancer immunotherapy has come. *Nihon Rinsho Meneki Gakkai Kaishi* **39**, 164–171 (2016).

190. Overwijk, W. W. *et al.* Mining the mutanome: developing highly personalized Immunotherapies based on mutational analysis of tumors. *J. Immunother. Cancer* **1**, 11 (2013).

191. Zitvogel, L. & Kroemer, G. Targeting PD-1/PD-L1 interactions for cancer immunotherapy. *Oncoimmunology* **1**, 1223–1225 (2012).

192. Poschke, I., Flossdorf, M. & Offringa, R. Next-generation TCR sequencing - a tool to understand T-cell infiltration in human cancers. *J. Pathol.* **240**, 384–386 (2016).

193. Murray, E. *et al.* Quantitative proteomic profiling of pleomorphic human

sarcoma identifies CLIC1 as a dominant pro-oncogenic receptor expressed in diverse sarcoma types. *J. Proteome Res.* **13**, 2543–2559 (2014).

194. Rajasagi, M. *et al.* Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* **124**, 453–462 (2014).

195. Vlenterie, M. *et al.* Next generation sequencing in synovial sarcoma reveals novel gene mutations. *Oncotarget* **6**, 34680–34690 (2015).

196. Fang, D. *et al.* The histone H3.3K36M mutation reprograms the epigenome of chondroblastomas. *Science* **352**, 1344–1348 (2016).

197. Lu, C. *et al.* Histone H3K36 mutations promote sarcomagenesis through altered histone methylation landscape. *Science* **352**, 844–849 (2016).

198. Liu, X. *et al.* Next-Generation Sequencing of Pulmonary Sarcomatoid Carcinoma Reveals High Frequency of Actionable MET Gene Mutations. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **34**, 794–802 (2016).

199. Crago, A. M. *et al.* Near universal detection of alterations in CTNNB1 and Wnt pathway regulators in desmoid-type fibromatosis by whole-exome sequencing and genomic analysis. *Genes. Chromosomes Cancer* **54**, 606–615 (2015).

200. Behjati, S. *et al.* Recurrent PTPRB and PLCG1 mutations in angiosarcoma. *Nat. Genet.* **46**, 376–379 (2014).

201. JBS, J. A boneless arm. *Boston Med Surg J* **18**, 368–9 (1838).

202. Gorham, L. W. & Stout, A. P. Massive osteolysis (acute spontaneous absorption of bone, phantom bone, disappearing bone); its relation to hemangiomatosis. *J. Bone Joint Surg. Am.* **37-A**, 985–1004 (1955).

203. Dellinger, M. T., Garg, N. & Olsen, B. R. Viewpoints on vessels and vanishing bones in Gorham-Stout disease. *Bone* **63**, 47–52 (2014).

204. Franco-Barrera, M. J. *et al.* Gorham-Stout Disease: a Clinical Case Report and Immunological Mechanisms in Bone Erosion. *Clin. Rev. Allergy Immunol.* **52**, 125–132 (2017).

205. de Keyser, C. E., Saltzherr, M. S., Bos, E. M. & Zillikens, M. C. A Large Skull Defect Due to Gorham-Stout Disease: Case Report and Literature Review on Pathogenesis, Diagnosis, and Treatment. *Front. Endocrinol.* **11**, (2020).

206. Möller, G. *et al.* The Gorham-Stout syndrome (Gorham's massive osteolysis). A report of six cases with histopathological findings. *J. Bone Joint Surg. Br.* **81**, 501–506 (1999).

207. Tong, A. C. K., Leung, T. M. & Cheung, P. T. Management of massive osteolysis of the mandible: a case report. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* **109**, 238–241 (2010).

208. Hu, P., Yuan, X., Hu, X., Shen, F. & Wang, J. Gorham-Stout syndrome in mainland China: a case series of 67 patients and review of the literature. *J. Zhejiang Univ. Sci. B* **14**, 729–735 (2013).

209. Zheng, M.-W. *et al.* Gorham-Stout syndrome presenting in a 5-year-old girl with a successful bisphosphonate therapeutic effect. *Exp. Ther. Med.* **4**, 449–451 (2012).

210. Nikolaou, V. S., Chytas, D., Korres, D. & Efstathopoulos, N. Vanishing bone disease (Gorham-Stout syndrome): A review of a rare entity. *World J. Orthop.* **5**, 694–698 (2014).

211. Ellati, R., Attili, A., Haddad, H., Al-Hussaini, M. & Shehadeh, A. Novel approach of treating Gorham-Stout disease in the humerus--Case report and review of literature. *Eur. Rev. Med. Pharmacol. Sci.* **20**, 426–432 (2016).

212. Koto, K., Inui, K., Itoi, M. & Itoh, K. Gorham-Stout disease in the rib and thoracic spine with spinal injury treated with radiotherapy, zoledronic acid, vitamin D, and propranolol: A case report and literature review. *Mol. Clin. Oncol.* **11**, 551–556 (2019).

213. Li, M. *et al.* Successful Management of Gorham–Stout Disease in Scapula and Ribs: A Case Report and Literature Review. *Orthop. Surg.* **10**, 276–280 (2018).

214. Schneider, K. N. *et al.* Gorham–Stout disease: good results of bisphosphonate treatment in 6 of 7 patients. *Acta Orthop.* **91**, 209–214 (2020).

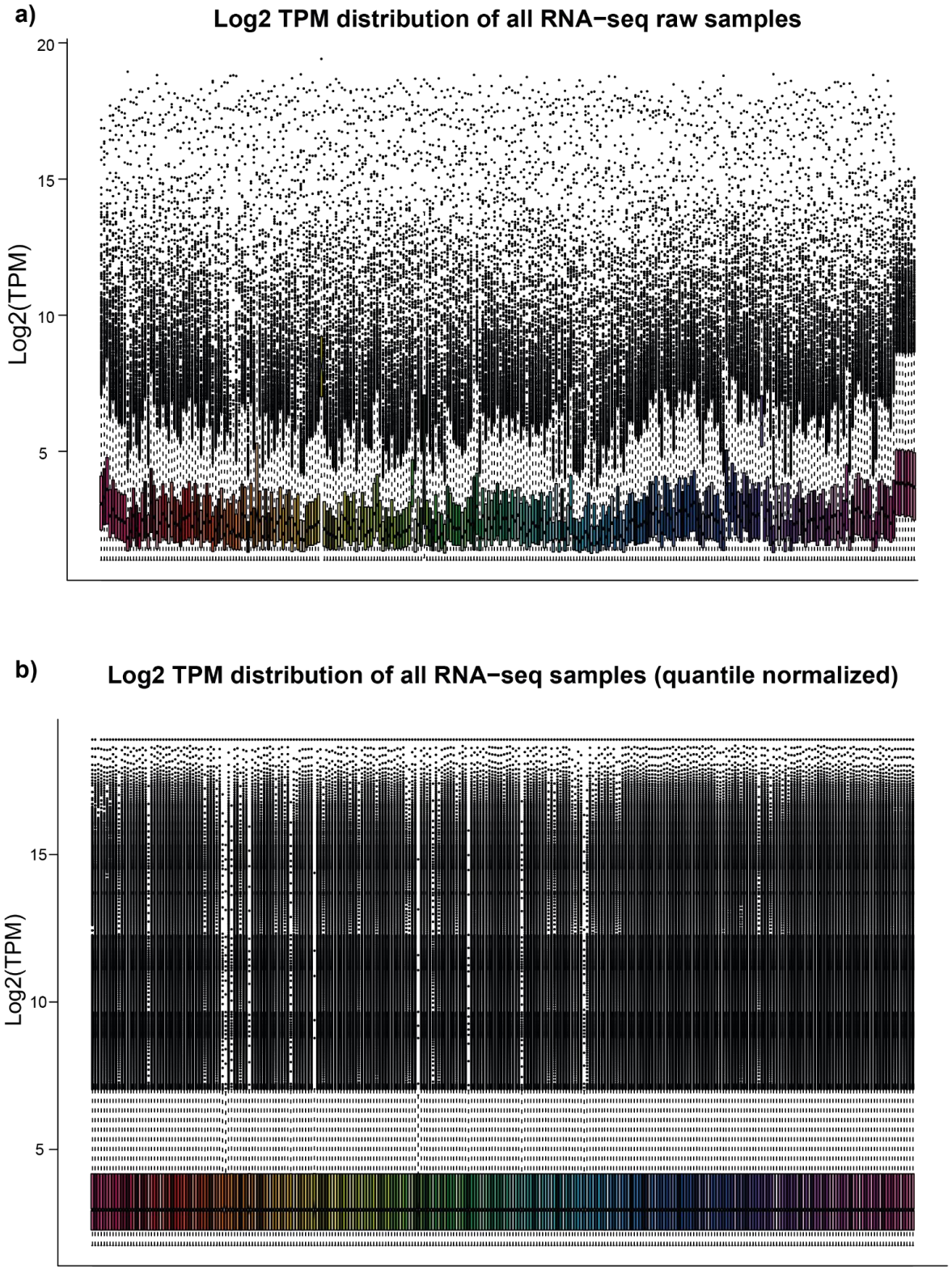
215. Liang, Y. *et al.* Gorham-Stout disease successfully treated with sirolimus (rapamycin): a case report and review of the literature. *BMC Musculoskelet. Disord.* **21**, 577 (2020).

216. Nozawa, A. *et al.* A somatic activating KRAS variant identified in an affected lesion of a patient with Gorham–Stout disease. *J. Hum. Genet.* **65**, 995–1001 (2020).

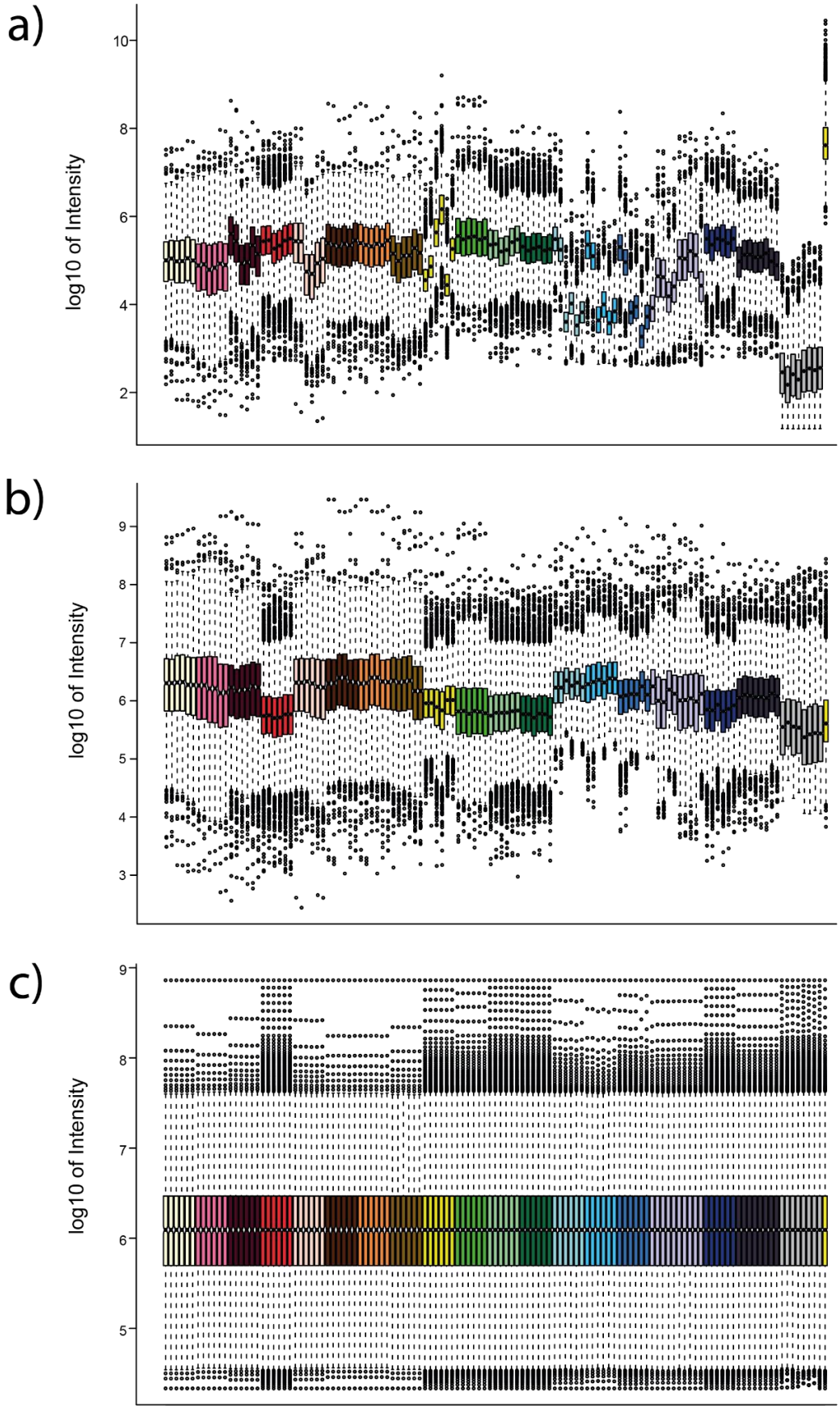
217. Homayun-Sepehr, N. *et al. KRAS*-driven model of Gorham-Stout disease effectively treated with trametinib. *JCI Insight* **6**, (2021).

8. Appendix

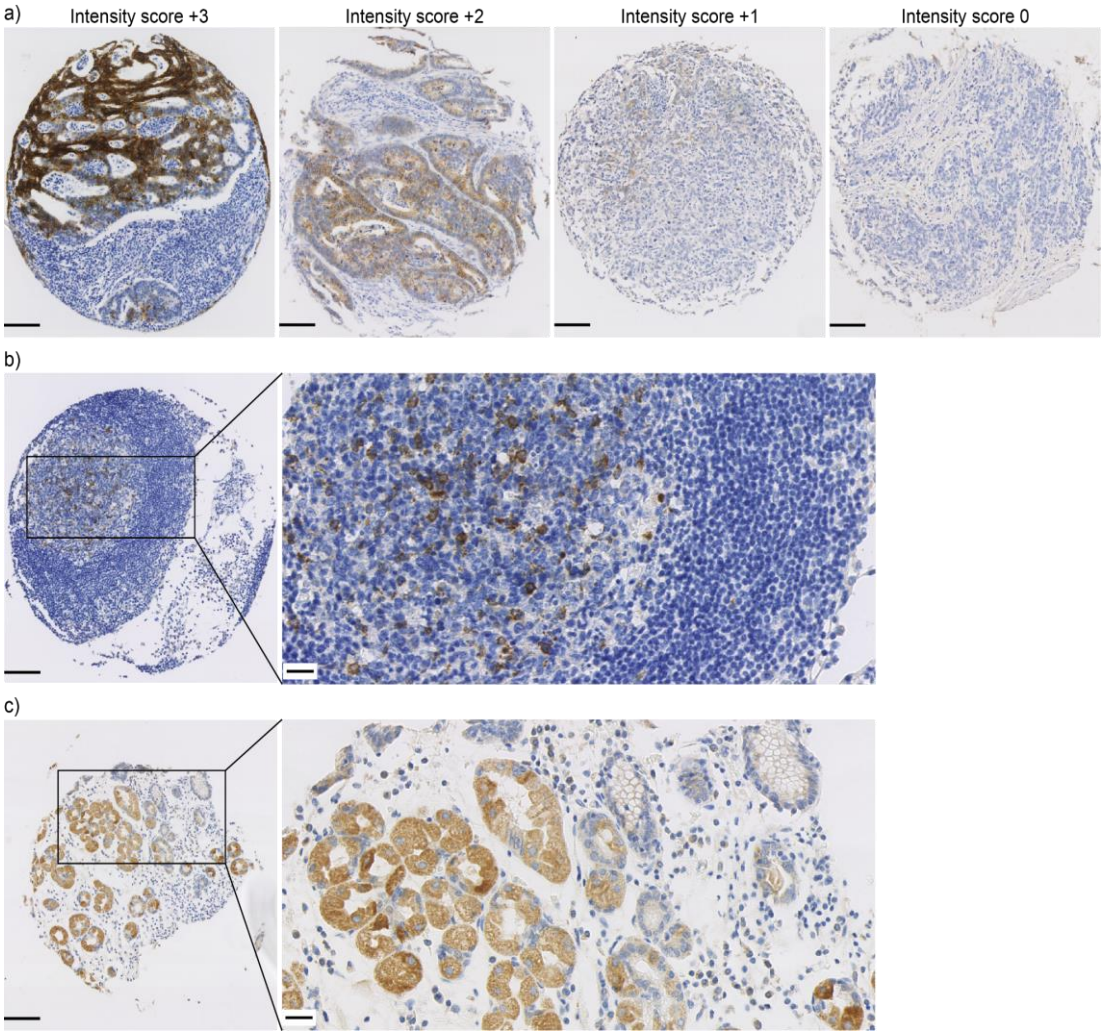
Supplementary Figure 1



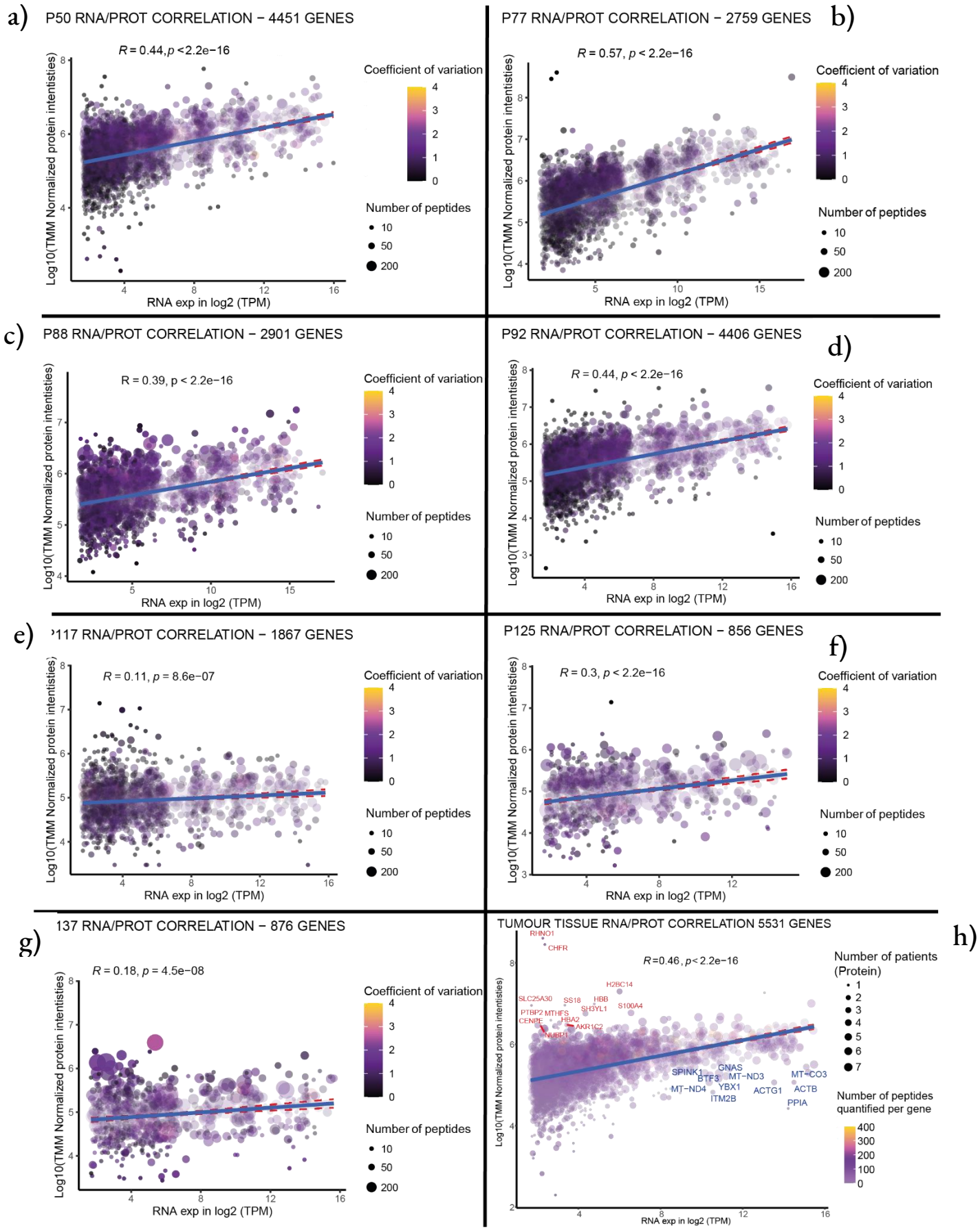
Supplementary Figure 2



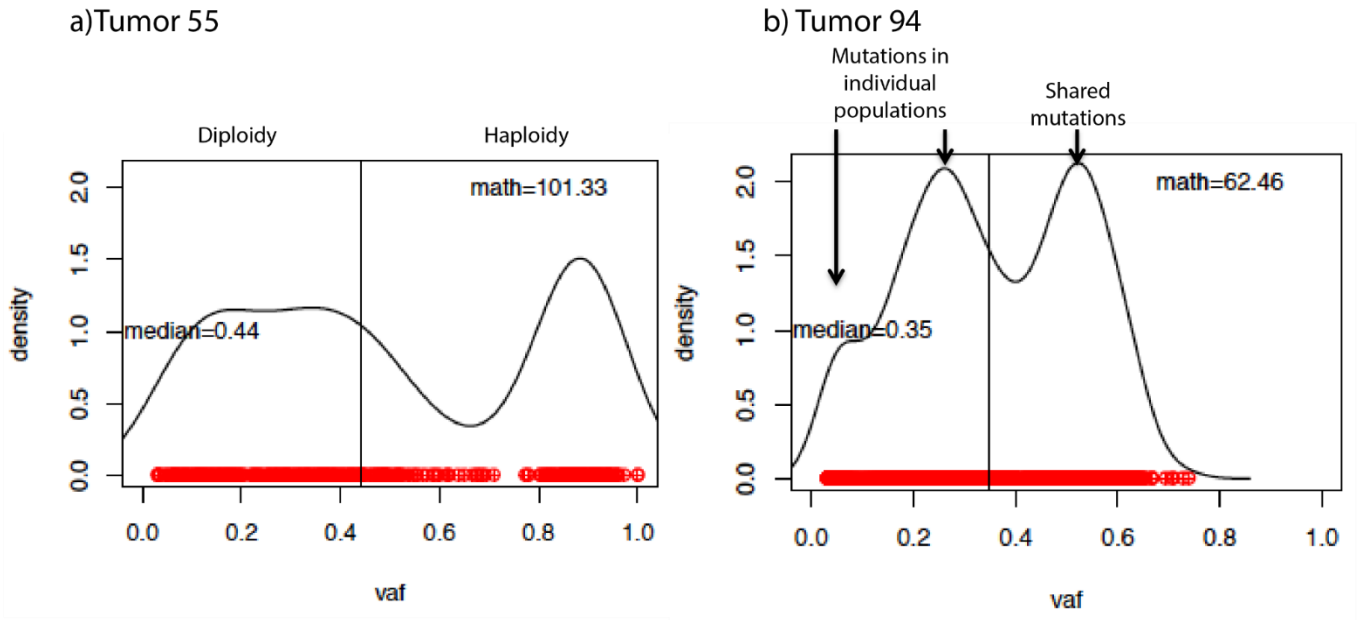
Supplementary Figure 3



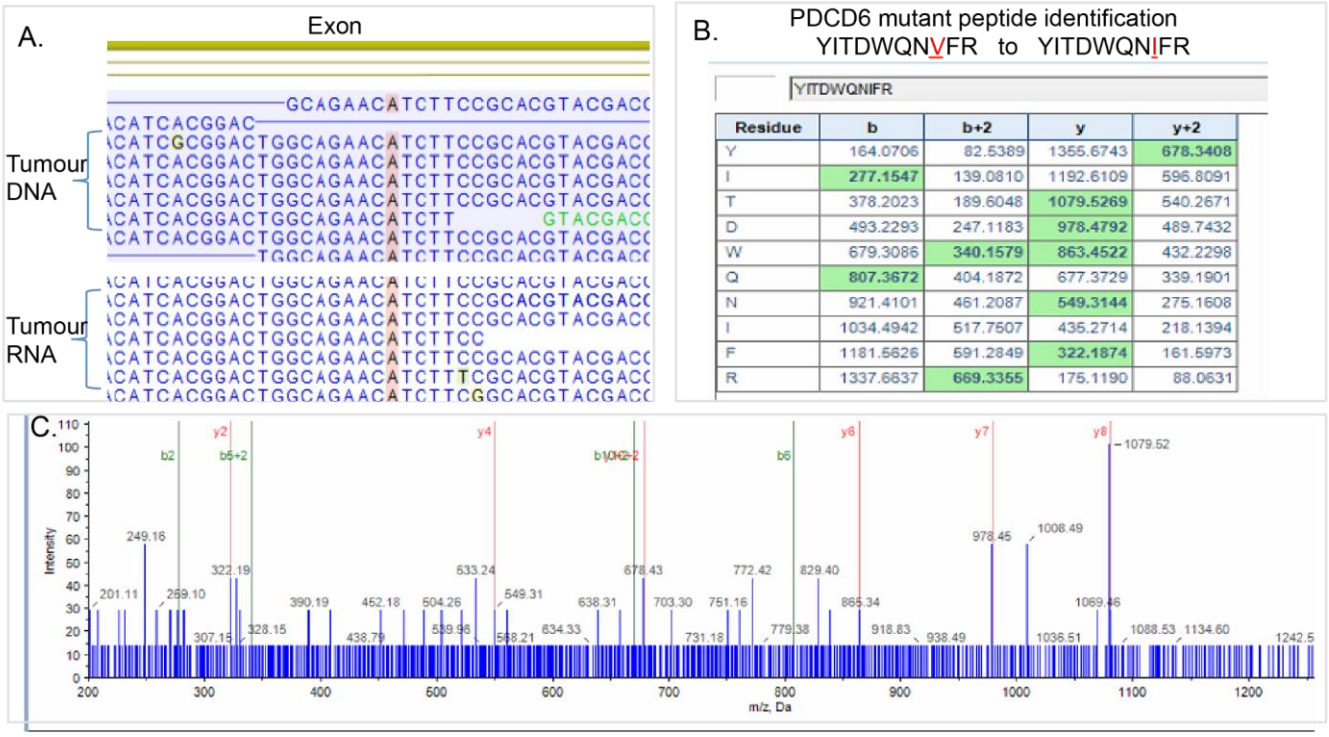
Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6



Residue	b	b+2	y	y+2
Y	164.0706	82.5389	1355.6743	678.3408
I	277.1547	139.0810	1192.6109	596.8091
T	378.2023	189.6048	1079.5269	540.2671
D	493.2293	247.1183	978.4792	489.7432
W	679.3086	340.1579	863.4522	432.2298
Q	807.3672	404.1872	677.3729	339.1901
N	921.4101	461.2087	549.3144	275.1608
I	1034.4942	517.7507	435.2714	218.1394
F	1181.5626	591.2849	322.1874	161.5973
R	1337.6637	669.3355	175.1190	88.0631

Supplementary Table 1

Neoantigen sets defined using DNAseq with SWATH-MS (1-8) and including IDA (9-12)

Protein	t-value	p-value	fold change	Mutation	peptide	(nM)	HLA
1.CADM1	1.045371	0.40558	13.8	R379C	IILG**C**YFAR	406	A*03:01
					ILG**C**YFARH	4097	A*03:01
					CLLIILG**C**Y	212	A*29:02
					LLIILG**C**YF	611	A*29:02
2.IDH3G	1.156134	0.36707	4.752609	R242H	AA**H**YPQITF	277	B*35:01
					VAA**H**YPQIT	10490	B*35:01
3.SSCPDH	0.781501	0.51633	1.766	Y237C	WPIS**C**CREL	29	B*07:02
					WPIS**C**CREL	53	B*35:01
					RWPIS**C**CRE	40623	B*35:01
4.PLEC	-0.84274	0.48809	0.85	A2107V	A**V**RQRQLAA	125	B*07:02
					VRQRQLAAE	28704	B*07:02
5.PDCD6	2.832139	0.10534	4.400572	V98I	YITDWQN**I**F	357	A*29:02
					ITDWQN**I**FR	20864	A*29:02
					YITDWQN**I**F	71	B*35:01
					ITDWQN**I**FR	27462	B*35:01
					YITDWQN**I**F	150	C*07:02
					ITDWQN**I**FR	45570	C*07:02
6.LAMB1	5.861941	0.02789	6.602734	T1100M	NO BINDERS		
7.POFUT2	6.007655	0.02661	2.833805	R213Q	NO BINDERS		
8.PIP4K2B	0.749014	0.53196	2.3664157	R134W	NO BINDERS		
9.PABPC1				T319I	SPFG**I**ITSA	887	B*07:02
					PFG**I**ITSAK	29088	B*07:02
10.EEF1A2				A65T	NO BINDERS		
11.TP53				R248Q	NO BINDERS		
12.GOLIM4				H132Y	NO BINDERS		