Title: Genomic features of NF1-associated peripheral nerve sheath tumors: a cohort analysis from the Johns Hopkins NF1 biospecimen repository.

Authors: Jineta Banerjee [1*], Yang Lyu [2*], Stavriani C. Makri [3*], Alexandra J. Scott [1], Lindy Zhang [3,4], Ana Calizo [5], Kai Pollard [6], Kuangying Yang [2], John M. Gross [7], Jiawan Wang [3], Carlos G. Romo [3,8], Robert Allaway [1], Jaishri O. Blakeley [3,8], Angela C. Hirbe [2], Christine A. Pratilas [3,4]

*These authors contributed equally.

Affiliations:

Sage Bionetworks, Seattle, WA, USA (1);

Division of Oncology, Department of Internal Medicine, Siteman Cancer Center, Washington University St. Louis (2);

Departments of Oncology (3), Pediatrics (4), Pathology (7), and Neurology (8), Johns Hopkins University School of Medicine;

Cellular and Molecular Medicine Program, Johns Hopkins University (5);

Personal Genome Diagnostics (6).

Abstract

Background: Neurofibromatosis type 1 (NF1) is a prevalent inherited neurocutaneous condition that predisposes to the development of peripheral nerve sheath tumors (PNST) including cutaneous neurofibromas (CNF), plexiform neurofibromas (PNF), atypical neurofibromatous neoplasms with unknown biological potential (ANNUBP), and malignant peripheral nerve sheath tumors (MPNST). Historically, therapeutic progress for PNF and MPNST has been limited in part due to restricted availability of primary tissues from patients with NF1. The successful advancement of therapeutic development for NF1-associated PNST necessitates ongoing efforts in the systematic acquisition and analysis of human tumor specimens and their corresponding model systems.

Methods: Patients with clinically or genetically confirmed NF1 having a clinically indicated surgical resection or biopsy of any NF1-associated tumor were invited to participate in an institutional review board (IRB) approved study for the collection and sharing of tissues and specimens. Tumors were assessed by the study pathologist, and banked in the laboratory as flash frozen tissues, paraffin embedded blocks or slides, DNA and RNA, or single cell suspensions. Efforts were made to create cell lines and patient derived xenografts (PDX) from primary human tissues. Clinical data for participating patients were fully annotated in a database that corresponds to banked tissue specimens. Applications for access to biospecimens, genomic data, and disease models, as well as de-identified clinical and molecular data are reviewed and approved with IRB oversight, to allow internal and external sharing to promote research collaboration.

Results: Since the inception of the JH NF1 biospecimen repository in 2016, 357 unique samples have been banked (from 183 unique patients) and include PNF (n=89), ANNUBP

(n=6), MPNST (n=62), CNF (n=103), and diffuse neurofibroma (diffuse NF, n=44). Xenografts have successfully been generated from seven MPNST samples and cell lines have been generated from three PNF and seven MPNST. RNA sequencing (RNAseq) and whole genome sequencing (WES) data were generated from 73 and 114 primary human tumor samples, respectively. These pre-processed data, standardized for immediate computational analysis, are accessible through the NF Data Portal, allowing immediate interrogation. Our analysis herein highlights key genetic variants and alterations in gene expression patterns, linked to pathways implicated in the pathology of the NF1-associated tumor types represented in the dataset. This work also combines new sample data with previously released samples, offering a comprehensive view of the entire cohort sequenced to date. Somatic variants in genes including *NF1*, *SUZ12* and *LRP1* and *LRP2* were identified in MPNST. Enrichment of RAS-RTK signaling pathways was identified through analysis of variants in both PNF and MPNST, however, MPNST demonstrated unique enrichment in pathways associated with extracellular matrix organization and cell cycle regulation.

Conclusion: Analysis of primary human tissue samples is critical for identification of therapeutically relevant molecular alterations. As a dedicated effort to systematically bank tumor samples from people with NF1 undergoing surgery, in collaboration with molecular geneticists and computational biologists who seek to advance understanding of NF1 biology, the Johns Hopkins NF1 biospecimen repository offers access to samples and genomic data to the NF1 research community to promote advancement of NF1-related therapies.

Materials and Methods:

Tumor sample collection and sequencing

Patient enrollment

All research was conducted according to a Johns Hopkins Hospital (JHH) institutional review board (IRB)-approved protocol, #IRB00096544.

Patients with clinically or genetically confirmed NF1 having a clinically indicated surgical resection or biopsy of any NF1-associated tumor (including, but not limited to, CNF, PNF, ANNUBP, MPNST, other NF1-associated neoplasm) were invited to participate. Written informed consent was obtained from all patients. The JHH IRB-approved consent form includes a statement about the voluntary nature of the research, a description of corresponding clinical data that will be collected for future analysis, and a statement that samples may be shared beyond JHH. Blood was collected from most patients on the day of surgery.

Medical records were retrospectively reviewed for demographic information and information related to the NF1 clinical condition (genetic diagnosis, family history, age of diagnosis), characteristics (phenotypic findings, symptoms), and tumor burden (number and size of NF1-associated tumors). These clinical data are stored in a password protected and de-identified database at JHH.

Tumor preservation and quality control

Surgical specimens were couriered to surgical pathology immediately after resection. The study neuropathologist performed immediate inspection of the tumor to ensure that

adequate tumor existed for clinical diagnostic needs. Upon approval, tumor pieces were sampled for banking and transported to the research laboratory in harvest medium (RPMI with 20% fetal bovine serum, supplemented with 1% penicillin-streptomycin and glutamine).

Specimens were sized into 5-10mm aliquots under sterile conditions in a biosafety cabinet. Individual aliquots were placed into 10% neutral buffered formalin, cell freezing media (Sigma #C6295), RNAlater (Qiagen #76104), and/ or placed into an empty vial and snap frozen on dry ice. Formalin-fixed paraffin-embedded tissues were stained with hematoxylin and eosin (H&E). H&E slides from each specimen were reviewed by the study neuropathologist to confirm an accurate diagnosis of each preserved specimen.

Tumor sample preparation for DNA sequencing

Genomic DNA from blood and tumor was extracted using the QIAGEN DNeasy Blood & Tissue kit. DNA library was constructed using KAPA HyperPrep Kits for NGS DNA Library Prep (Roche #7962363001). Exomes were captured by IDT xGen Exome Hyb Panel v1.0 (IDT #1056115) and the libraries were sequenced by NovaSeq6000 S4 with ~100X coverage for normal samples and tumor samples.

Tumor sample preparation for RNA sequencing

RNA was isolated from tumor by using TRIZOL. Samples containing at least 100ng total RNA with RNA Integrity Number (RIN) > 6.5 were sequenced. For RNA extraction, the first batch of samples were prepared with MGI/GTAC RiboErase method. Depletion was

performed with KAPA RiboErase HMR (Roche #07962274001). Library preparation was performed with a WUSTL in-house preparation protocol. For the second batch, samples were prepared with Illumina TruSeq® Stranded Total RNA Library Prep Gold kit (#20020599) which includes RiboZero Gold depletion. Library preparation was performed according to the manufacturer's protocol. Fragments from both batches were sequenced on an Illumina NovaSeq-6000 using paired-end reads extending 150 bases.

Whole exome sequencing data processing for somatic variant calling and analysis

Raw fastq data files were quality checked using FastQC v0.11.9 and a report was generated using MultiQC v1.8. Fastq files were aligned to GRCh38 using BWA 0.7.17-r1188 (1). Duplicates were marked using GATK MarkDuplicates, and bases recalibrated using GATK BaseRecalibrator and GATK ApplyBQSR (GATK v4.3.0.0) (2-4). Somatic single nucleotide variants (SNVs) were then called using Strelka2 software (Strelka v2.9.10) (5) and Mutect2 (GATK v4.4.0.0) (6). Strelka 2 was shown to have high precision and recall for SNVs compared to others as tested in the precision FDA challenge (5, 7). The variants were annotated using Variant Effect Predictor (VEP v99.2) and converted to mutation annotation format (MAF) files using vcf2maf (vcf2maf v1.6.21). All of these steps were completed on Nextflow Tower running the standardized nf-core pipeline sarek v3.1.2 (8).

Variant calling from samples with paired normal in a separate batch

The samples contained in syn52659564 and **Supplementary Table 1** were special cases where the tumor samples and the paired normal samples were sequenced in two different batches with different sequencing library kits and thus had two different browser extensible data (BED) files for target capture. For the above samples, first the BED file from JH_batch1 was lifted over from hg19 to hg38 coordinates using the UCSC liftOver tool (9) and the liftover chain file named 'hg19ToHg38.over.chain.gz' and then sorted using the sort function from bedtools suite. Then a common region BED file was generated using the intersect function in bedtools suite with at least 50% overlap in intervals between the BED file from JH_batch1 and WU_batch1. The common region BED file (syn52594249) was then used to call somatic variants from these tumor-normal pairs using the sarek v3.2.2 pipeline and Strelka2 and Mutect2 variant callers.

Consensus variant calling using SomaticSeq

To avoid detection of false positive variant calls in the samples, we identified and reported the variant calls that had consensus between Strelka2 and Mutect2 callers. To identify these consensus calls we used a previously benchmarked method called SomaticSeq (10). Once consensus calls for SNVs and insertions and deletions (indels) were identified, the variants in respective variant call formats (VCFs), were annotated using vcf2maf in the sarek v3.2.2 pipeline to generate a merged MAF file that was used for further analysis. The consensus variants were filtered for high confidence calls (PASS filtered) with population allele frequency below 0.05 to identify robust variant calls excluding common variants.

RNAseq data processing and analysis

Raw fastq files were processed using nf-core/rnaseq (v3.11.2) (11) and quantified using salmon: v1.10.1 (12). ComBat from sva R package (v3.42.0) was used for batch correction (13). DESeq2 (v1.34.0) was applied to call the differentially expressed genes (|fold change| > 1.2, adjust $P < 0.05$), and ggplot2 R package (v3.3.6) was applied to draw volcano plots. Principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) (14) were applied for the visualization of samples. Gene set enrichment analysis (GSEA) was conducted by using R-package fgsea (v1.20.0) [13]. ggplot2 (v3.3.6) was used for the gene volcano plot and pathway enrichment plot [14]. Pheatmap (v1.0.12) was applied to plot gene expression heatmap. Singscore (v1.14.0) was applied to calculate single sample gene signature score (15).

Evaluation of relatedness in genomic and transcriptomic data

Somalier (v0.2.17) (16) was used to evaluate the relatedness between WES and RNAseq samples. The *extract* function was used to identify informative sites from the WES cram files and RNAseq bam files, and then the *relate* function was used to calculate statistics on the number of shared genotypes and the coefficient of relatedness between all samples. One WES sample, JH-2-009-2578C-A, did not contain data at a sufficient number of polymorphic sites to calculate relatedness to other samples and was excluded from further analysis. The pheatmap R package (v1.0.01) was used to generate heatmaps of the relatedness comparing each primary human WES sample exactly once

to each primary human RNAseq sample and the ggplot2 package was used to generate scatter plots. Cell line and PDX samples were excluded from the heatmap.

Results

Efforts to generate a genomically and clinically annotated NF1 biospecimen repository, for the purpose of sharing resources, to promote collaborative research, and make specimens and genomic data available to the NF1 research community, were previously reported (17). As of December 2023, the Johns Hopkins NF1 biospecimen repository includes 357 banked tumor tissues from 183 unique patients with NF1. A full specimen inventory is available online (https://bit.ly/nf1biobank) and investigators can request specimens by providing a statement of scientific rationale and intended use. More than 65 requests for sharing of available specimens, models, and/ or genomic data have been fulfilled to date.

Patient and sample information

Clinical and demographic information on the patients whose tumors have been genomically characterized, including sex and age by range, are included in **Table 1**. Specific ages at time of tumor collection are not provided to protect patient identities given the rarity of some NF1-associated conditions. Sample and file information can be found in Synapse.org datasets (WES: syn53132831.1, RNAseq: syn53133024.1; for more details, refer to the Data Availability selection) and accessed on the NF Data Portal for further analysis.

198

199 Relatedness of WES and RNAseq data from individual patients was confirmed as a
200 measure of quality control.

201 As a quality and validation step, we used Somalier (16) to evaluate relatedness between
202 WES and RNAseq samples, comparing each WES sample exactly once to each RNAseq
203 sample. We found that all sample comparisons that originated from different individuals
204 had a relatedness coefficient <0.9. Based on this and the clinical annotations of the
205 samples, we set a threshold of relatedness coefficient ≥0.9 to be the confirmation that a
206 pair of WES and RNAseq samples were indeed from the same individual (**Figure 1A**).

207 With this threshold in place, all but three corresponding samples from the same patient
208 were found to meet the definition of relatedness (**Figure 1A,B**). One RNAseq sample
209 from a recurrent tumor, JH-2-079-CB92G-FG226, had a relatedness coefficient of 0.86 to both
210 the normal WES sample from the same individual and to a WES sample from a different tumor in
211 the same individual. However, it had high relatedness (0.997) to the WES sample generated from
212 the same recurrent tumor. Two additional RNAseq samples that originated from one
213 individual had lower than expected relatedness to other samples from that individual. On
214 examination of the clinical annotations, we found that JH-2-002-GAF53-2E57B was an
215 RNAseq sample of a patient-derived cell line and JH-2-002-GAF53-3BAGC was an
216 RNAseq sample of a xenograft developed from the tumor sample. This distinction
217 suggested that the threshold of relatedness coefficient determined by our analysis was
218 sensitive enough to detect differences between primary samples and the cell lines and
219 xenografts derived from them.

Unbiased exploration of somatic variants identifies NF1 among the top genes with variants using WES

WES data from the human tumor samples were haplotype matched to their respective normal samples (**Supplementary Table 2**). To explore the variant landscape of the human tumor samples from the biobank, we consulted six highly cited reports (17-22) to generate a list of 60 priority genes that are deemed to be of key biological importance in NF1-MPNST. Given the differences in detecting limits of variant caller algorithms, we used multiple variant callers to identify single nucleotide polymorphisms (SNPs) and short indels for these genes of interest in the tumor samples. Variants detected using Strelka2 (**Figure 2A,D,E)** exceeded those detected using Mutect2 (**Figure 2B,F,G**). The median number of variants found in the samples using Strelka2 and Mutect2, respectively, are shown in **Figure 2D** (median =129) and **2F** (median=11) and highlight that the majority of the variants detected were missense variants. The top ten genes with variants detected using Strelka2 and Mutect2 callers are shown as **Figures 2E,G**. To find variants that were identified regardless of the variant caller used, we identified consensus variant calls with SomaticSeq – (**Figure 2C,H,I)**. Samples in which no high-confidence variants were detected using at least one variant caller were excluded from further analysis, and those that had consensus calls for variants using SomaticSeq are visualized in **Figure 2C**. A median of 6 variants including missense, nonsense, splice-site, nonstop, and translation start-site variants were found per sample when only consensus calls were considered (**Figure 2H)**. In the consensus calls, *NF1, SUZ12, LRP1, and LRP2* genes were found within the top 10 genes with variants with moderate or high impact on gene function among the samples (**Figure 2I**).

Given that PNF and MPNST tumor types represented the greatest number of samples in our cohort, and PNF may undergo malignant transformation to MPNST, we focused on exploring the consensus genomic variants identified in these two tumor types with attention to their similarities, and differences that may relate to malignant transformation.

Somatic variants in known oncogenic pathways identified in PNF and MPNST samples

We identified somatic SNVs in the *NF1* gene in 31% of PNF samples and 26% of MPNST samples examined. The genetic loci of *NF1* variants identified in our samples and the predicted effect on protein expression are shown schematically alongside a protein structure representation of the *NF1* gene (**Figure 3A**, PNF shown at top, MPNST shown at bottom). **Figure 3B** represents a diagrammatic rendition of known functional domains in the NF1 protein for illustration, modified from (23).

Additionally, we compared the presence of variants in known oncogenic pathways from The Cancer Genome Atlas (TCGA) (24) to the consensus genetic variants found in our cohort. Specifically, we tested the enrichment of genes where genetic variants were found in PNF samples and in MPNST samples separately to understand which pathways may be affected by the presence of variants in each of these tumor types. SNVs in our cohort of PNF samples identified enrichment of variants in genes related to the RAS-RTK pathway (**Figure 3C**). Fourteen out of 36 PNF samples had variants in at least one gene involved in RAS-RTK signaling pathways. In MPNST samples, however, five out of 19 MPNST samples had variants in at least one gene related to the RAS-RTK pathway, and

in several samples, genetic variants related to the Hippo, TP53, MYC, and PI3K pathways were enriched (**Figure 3D).**

MPNSTs demonstrate enrichment in pathways associated with extracellular matrix organization and cell cycle regulation

We employed PCA and UMAP to generate a clustering visualization, based on the expression of all genes in all tumor samples (**Figure 4A,B**). The majority of CNF tumors fell into a distinct cluster, appearing less related to MPNST, indicating divergent gene expression profiles between the two tumor types. Some PNF and MPNST samples however, clustered nearby each other, possibly indicating a range of biological behavior in PNF, which are known to have a propensity for malignant progression in some patients, or possibly MPNST with less aggressive biology or potential for metastasis. Interrogation of the clinical outcomes for specific patients from whom these PNF were obtained would take years of follow up and is therefore outside the scope of this analysis. For other tumor types, strong discernible patterns were not evident, due to the constraints imposed by limited sample numbers.

To further characterize the MPNST and PNF cohorts, we evaluated the expression of genes previously implicated in MPNST biology [29]. We found that *ERBB3, S100B, CNP, PMP22, PDGFB,* and *TGFBR2* were significantly downregulated in MPNST compared to PNF, while *CCNB2, TWIST1, COL6A3, PTK7, GAS1,* and *TGFB2* were significantly upregulated in MPNST (**Figure 4C,D**) (|fold change| > 1.2 with adjusted $P < 0.05$), consistent with previous reports [29, 30].

To elucidate the underlying pathways implicated in MPNST biology, we employed gene set enrichment analysis (GSEA) on all genes differentially expressed between MPNST and PNF in our cohort (**Figure 5**). Our analysis revealed alterations in six pathways within Hallmark gene sets in MPNST versus PNF (adjusted $P < 0.01$) (**Figure 5A and Supplementary Table 3**). Notably, the three most statistically significant pathways involved upregulation of genes related to epithelial to mesenchymal transition, G2M checkpoint, and myogenesis (25). Volcano plots depicting the involved genes are shown in **Figure 5C,D,E** and heatmaps showing corresponding gene expression are shown in **Figure S1 A,B,C.** Differentially expressed genes in these pathways are listed in **Supplementary Tables 5, 6, 7** (adjusted $P < 0.01$ & $|FC|>1.2$). Additionally, we evaluated pathways within Reactome gene sets (26) and identified 23 pathways that were significantly different in MPNST vs. PNF (adjusted $P < 0.01$) (**Figure 5B and Supplementary Table 4**). The three most statistically significant pathways were associated with upregulation of genes involved in extracellular matrix (ECM) organization, ECM degradation, and collagen degradation. Genes involved in these pathways are shown in the volcano plots in **Figure 5F,G,H** and corresponding heatmaps of gene expression are shown in **Figure S1 D,E,F**. Differentially expressed genes in these pathways are listed in **Supplementary Tables 8, 9, 10** (adjusted $P < 0.01$ and $|FC| > 1.2$). Taken together, these findings suggest that MPNST exhibit a stronger mesenchymal phenotype compared to PNF and that MPNST display increased ECM and activated collagen degradation, potentially fostering tumor metastasis (27-29).

Further, as expected, we observed significant activation of several pathways previously reported in the literature, which have clear links to MPNST biology including "H3K27me3-associated", "metastasis-associated", and "mTOR-associated" pathways in MPNST (**Supplementary Figure 2**). The single sample gene signature score for each of these gene sets was significantly higher in MPNST versus PNF (**Supplementary Figure 3,** $p<0.01$, Wilcox test).

Discussion

The Johns Hopkins NF1 biospecimen repository continues to preserve human tissues from patient volunteers with NF1 undergoing surgical resection of any tumor or tissue for the purposes of: 1) analysis of current and emerging diagnostic criteria for PNST; 2) histologic and genomic characterization of these samples; and 3) creating a publicly accessible, clinically and genomically annotated biospecimen repository that enables therapeutic discovery research for NF1 across internal and external collaborators. The goal of this project is to collect and store blood and tumor samples with accompanying clinical data from 15-20 patients per year, including those with CNF, PNF, ANNUBP, MPNST, and other NF1-associated neoplasms. We expect ongoing accrual at this rate, given the volume of patients with NF1 seen at the Johns Hopkins Comprehensive NF Center annually and patient interest. At present, we have banked 357 tissue samples from a total of 183 unique patients with NF1 since the inception of the biospecimen repository in 2016. These tissues, as well as cell lines and PDX models derived from human tissue collection, and accompanying genomic data, represent publicly available resources which are shared routinely with investigators upon request and free of charge,

to promote ongoing collaborative science focused on therapeutic discovery in the NF1 research community. A total of 187 samples (plus 71 corresponding normal samples) from banked tissues have been sequenced, including whole exome on a total of 114 tumors (33 MPNST, 4 ANNUBP, 57 PNF, 9 diffuse NF, and 11 CNF) and RNA sequencing on a total of 73 tumors (18 MPNST, 1 ANNUBP, 38 PNF, 9 diffuse NF, 6 CNF, nodular NF 1). These genomic data are now publicly available to NF1 researchers to support an array of important lines of scientific inquiry.

With this publication, we are releasing curated datasets of pre-processed WES and RNAseq data that are ready for general re-use. These datasets are processed using standardized publicly vetted processing workflows to reduce the efforts of data re-use. Multiple variant callers and well known RNAseq quantification methods have been used to generate datasets designed for maximum reusability. Additionally, these datasets are well annotated with clinical metadata to enable appropriate re-analysis of the data as needed.

Here, we present an analysis of 133 primary tumor samples sequenced using WES (82 total: MPNST 24; ANNUBP 1; PNF 45; diffuse NF 1; CNF 10, nodular NF 1) and RNAseq (51 total: MPNST 14; ; PNF 29; diffuse NF 1; CNF 6, nodular NF 1). All existing RNAseq and WES raw data for these tumors, as well as 4 MPNST cell lines and 2 PDX, have been deposited in the NF Data Portal. The analysis presented herein is a preliminary demonstration of whole exome and RNA sequencing analysis, meant to provide the scientific community with an overview of available samples and their potential utility. Several aspects of NF1 associated tumors were confirmed in this work. First, as expected the *NF1* gene was the most commonly mutated gene in MPNST and PNF. We note that

355 variants in the *NF1* gene were detected in 35% of the samples rather than the expected
356 100% of samples, given their origin from people with the NF1 clinical condition. The
357 incomplete detection of *NF1* variants is likely due to a combination of factors: 1) detection
358 sensitivity of WES is limited by lack of coverage on intronic regions; 2) WES analysis is
359 not sensitive for microdeletion or copy number changes; 3) somatic variant analysis will
360 be unable to call a variant in tumor samples with loss of heterozygosity of the *NF1* gene
361 since both tumor and normal samples will have the same genetic variants; and 4) variants
362 will not be called with high confidence in tumor samples with low tumor cell purity. These
363 limitations suggest that targeted high depth clinical sequencing may be more sensitive in
364 capturing *NF1* variants in these tumors (30-33). The inability to detect *NF1* variants in all
365 samples despite a clinical diagnosis highlights the current limitations with WES
366 technology and the many ways in which the *NF1* gene can be altered resulting in the
367 variable clinical presentation of NF1. Second, we demonstrated that the main pathway in
368 which oncogenic variants are detected in PNF involves RTK-RAS signaling. In MPNST
369 samples, the bulk of genetic variants affected not only RAS-RTK signaling, but also the
370 Hippo and TP53 pathways. This finding agrees with prior work indicating that progression
371 to MPNST involves genetic changes beyond those affecting RAS biology. Third, through
372 our RNAseq analysis, we demonstrated an increase in activation of pathways involving
373 extracellular matrix organization and degradation as well as cell cycle checkpoints in
374 MPNST compared to PNF. These biologic insights are in accordance with the aggressive
375 nature of MPNST. While it would be validating to find that these expression signatures
376 are even further enhanced in recurrent or metastatic MPNST, compared to those that

377 remained localized or did not recur, the numbers of samples included herein limit the
378 ability to achieve meaningful significance in this analysis.

379 In summary, this study showcases the release of WES and RNA sequencing data from
380 human NF1 associated peripheral nerve sheath tumors, along with pre-processed data
381 in analysis-ready format to facilitate reuse without additional efforts. Our exploration of
382 the data confirms findings that are concordant with available literature regarding the
383 genomics of NF1 associated tumors and validate this biospecimen repository as a rich
384 resource for the scientific community.

385

386

387 Author Contribution:

388 Authorship was determined using ICMJE recommendations.

389 Conceptualization: JB, ACH, JOB, CAP

390 Data curation: JB, YL, SCM, AC, KP, LZ

391 Formal Analysis: JB, YL, AJS

392 Funding acquisition: JOB, RJA, JB, CAP

393 Investigation: all authors

394 Methodology: JB, YL, SCM, AC, KP, LZ, AJS

395 Project administration: ACH, JOB, CAP

396 Resources: JB, YL,

397 Software: JB, YL, AJS

398 Supervision: ACH, JOB, RJA, CAP

399 Validation: not applicable

400 Visualization: JB, YL, AJS

401 Writing: original draft: JB, YL, SCM, AJS, ACH, CAP

402 Writing: review & editing: all authors

403

404 Code and Data Availability:

All the figures in this article are generated through reproducible R code which is available on Github (https://github.com/nf-osi/biobank-release-2). All raw data files were processed for analyses using nf-core pipelines available through https://nf-co.re/ and specific versions used in this study have been noted in the methods section. All the raw data files are available on Synapse.org (WES: dataset format: syn53132831.1, RNAseq: dataset format: syn53133024.1). All processed data visualized in this article are available on Synapse.org (WES: Somatic Variants Mutect2 – syn53149144.2, Somatic Variants Strelka2 – syn53149128.1, RNAseq quantification files: syn53140231.1, RNAseq counts files: syn53141534.1).

418 **Table 1.** List of samples with RNAseq and/ or WES, including age (in in years by range),
419 sex, and type of NF1-associated PNST, included in the current analysis of samples and
420 for which sequencing data are available (inclusive of samples released with Pollard et al,
421 *Sci Data*, 2020 (17)). WES and RNA sequencing are indicated by an "x" and "y" symbol,
422 respectively.

Patient ID	Age at collection	Sex	Normal (blood)	CNF	Nodular NF	Diffuse NF	PNF	ANNUBP	MPNST
JH-2-001		F	x				x, y		
JH-2-002		M	x*				x*, y		x*, y*
JH-2-003		M					y		
JH-2-004		F	x				x, y		
JH-2-006		F	x				x, y		
JH-2-007		M		y					
JH-2-009		M	x				x		x*, y*
JH-2-010		M	x				x, y		
JH-2-012		M	x				x, y		
JH-2-013		F	x				x, y		
JH-2-014		M	x				x, y		
JH-2-015		M	x				x*		x, y
JH-2-016		F	x				x, y		x, y
JH-2-017		M	x				x, y		
JH-2-019		F	x				x, y		
JH-2-021		M	x		x, y				
JH-2-023		M	x				x		x*, y*
JH-2-026		F	x				x, y		
JH-2-029		M	x	x, y					
JH-2-031		M	x						x*, y*
JH-2-032		F	x$				x, y		
JH-2-044		F	x$				x, y		
JH-2-045		F	x$	x, y					x
JH-2-054		M	x				x*		
JH-2-055		F	x				x		x*, y
JH-2-057		F	x				x		
JH-2-060		M	x				x*		
JH-2-061		F	x				x, y		
JH-2-068		M	x				x*, y		
JH-2-072		M	x						x

JH-2-074		M	x				x		
JH-2-075		M	x	x			x, y		
JH-2-076		F	x	x			x, y		
JH-2-077		F	x				x		
JH-2-079		F	x						x*, y*
JH-2-082		M	x				x, y		x, y
JH-2-085		M	x				x, y		
JH-2-086		M	x				x, y		
JH-2-089		M	x				x, y		
JH-2-090		M	x			x, y		x	
JH-2-091		M	x				x, y		
JH-2-092		F	x	x*,y*					
JH-2-093		M	x	x*, y					
JH-2-094		F	x				x, y		x
JH-2-095		M	x				x		
JH-2-096		M	x						x
JH-2-099		F	x				x		
JH-2-100		F	x				x, y		
JH-2-102		F	x	x					
JH-2-103		M	x						x
JH-2-104		M	x				x, y		
JH-2-107		M					y		
JH-2-111		M	x				x, y		
JH-2-113		F	x						x

Key:	Age (years)
	0-10
	11-20
	21-30
	31-40
	41-50
	>50

x WES

y RNAseq

* more than one same sample was sequenced \

$ normal tissue was sequenced (not blood)

Figure legends

Figure 1. Relatedness between WES and RNAseq samples. A) A scatter plot of the relatedness coefficient versus the number of sites with the same genotype between two samples, as calculated by Somalier. Comparisons between two samples annotated as being from the same individual are shown in red and those from different individuals are colored blue. The shape of the point indicates the tissue of origin of the samples compared. The vertical line indicates the cutoff value (≥ 0.9) used to determine whether two samples are from the same individual based on the relatedness coefficient. B) A heatmap of the relatedness coefficient of pairwise comparisons between WES and RNAseq data from primary human samples as shown in (A), excluding cell lines or xenografts derived from these samples. RNAseq samples are shown on the x-axis and WES samples are on the y-axis. The relatedness coefficient is displayed according to the color scale bar shown in the figure.

Figure 2. Summary of somatic variants detected in PNST (10 CNF, 1 nodular NF, 1 diffuse NF, 36 PNF, 1 ANNUBP, and 19 MPNST). A) Oncoplot of variants in selected genes of interest (see methods) in the cohort of biobank patients detected using Strelka2. B) Oncoplot of variants in selected genes of interest in the cohort of biobank patients detected using Mutect2. C) Oncoplot of the consensus of variants in selected genes of interest in the cohort of biobank patients detected using SomaticSeq. The status of *NF1* functional inactivation in the patients was determined using either clinical genetic testing or clinical diagnosis and is provided using colored bars in the bottom panel of each oncoplot. (D, F, H) Barplot showing the number of variants per sample, by SNV class

(missense, nonsense, splice site, nonstop or translation start site). (E, G, I) Bar plot of the top 10 genes with variants of moderate or high impact in the cohort.

Figure 3. Somatic variants in plexiform neurofibromas and malignant peripheral nerve sheath tumors. A) Lollipop plot showing different positions and protein consequences of variants in the *NF1* gene as detected in this cohort using WES. B) Schematic representation of neurofibromin protein with various functional domains. Domains: CBD caveolin-binding domains, CSRD cysteine and serine rich domain, CTD C-terminal domain, GRD RAS-GAP-related domain, HLR HEAT-like repeats, LRD leucine-rich domain, PH pleckstrin homology, SBR syndecan-binding region, TBD tubulin-binding domain (adapted from Mo et al, 2022, with permission) (23). C) Plots showing top oncogenic pathways that are affected by genomic variants in PNF samples. D) Plots showing top oncogenic pathways that are affected by genomic variants in MPNST samples.

Figure 4. A) Principal component analysis (PCA) plot illustrating the distribution of all tumors based on RNAseq gene counts. B) Uniform manifold approximation and projection (UMAP) plot depicting the tumor distribution using RNAseq gene counts. C) Volcano plot highlighting differentially expressed genes associated with MPNST compared to PNF ($|FC| > 1.2$ and $P < 0.05$). D) Heatmap displaying the expression patterns of differentially expressed genes associated with MPNST in comparison to PNF.

Figure 5. Pathway enrichment analysis in MPNST versus PNF (RNAseq). A) Dot plot showing significant pathways in Hallmark gene sets in MPNST versus PNF (adjusted $P < 0.01$). B) Dot plot showing significant pathways in Reactome gene sets in MPNST versus PNF (adjusted $P < 0.01$ & $|FC| > 1.2$). C,D,E) Volcano plots for top three significant Hallmark pathways (epithelial-mesenchymal transition, G2M checkpoint, and myogenesis) highlighting differentially expressed genes in MPNST vs. PNF (adjusted $P < 0.01$). F,G,H) Volcano plots for top three significant Reactome pathways (extracellular matrix organization, collagen degradation, and extracellular matrix degradation) highlighting differentially expressed genes in MPNST vs. PNF (adjusted $P < 0.05$ & $|FC|>1.2$).

Supplementary Materials

Supplementary figures

Supplementary Figure 1. Heatmap of gene expression in significant pathways in MPNST vs. PNF tumors. Heatmap illustrating the expression of differentially expressed genes in significant Hallmark pathways (A,B,C – epithelial-mesenchymal transition, G2M check-point, and myogenesis) and Reactome pathways (D,E,F – extra cellular matrix organization, collagen degradation, and extracellular matrix degradation) in MPNST vs. PNF tumors.

Supplementary Figure 2. Volcano plot and heatmap of gene expression in three pathways with established biological activity in MPNST vs. PNF tumors ("H3K27me3-associated", "metastasis-associated", and "mTOR-associated"). Genes with fold change > 2 (adjusted $P < 0.05$) are labeled in the volcano plots (A,B,C) and gene expression patterns of differentially expressed genes in MPNST vs. PNF (adjusted $P < 0.05$ & |FC| >1.2|) are shown in heatmap (D,E,F).

Supplementary Figure 3. Boxplot of the single sample gene signature score (singscore) in three significant pathways in MPNST vs. PNF tumors as shown in Fig S2. These three pathways as shown in aggregate score, are upregulated in MPNST compared to PNF (A,B,C, Wilcox test, $P < 0.05$).

Supplementary tables

Supplementary table 1. Tumor samples with paired-normal samples sequenced in two different batches.

Supplementary table 2. HLA typing results of WES data.

Supplementary table 3. Hallmark pathways that meet significance thresholds in MPNST vs. PNF (adjusted $P < 0.05$).

Supplementary table 4. Reactome pathways that meet significance thresholds in MPNST vs. PNF (adjusted $P <0.05$).

Supplementary table 5. Differentially expressed genes in "HALLMARK_EPITHELIAL_ MESENCHYMAL_TRANSITION" in MPNST vs. PNF (adjusted $P < 0.05$, $|FC| > 1.2$).

Supplementary table 6. Differentially expressed genes in "HALLMARK_G2M_CHECKPOINT" in MPNST vs. PNF (adjusted $P < 0.05$ & $|FC| > 1.2$).

Supplementary table 7. Differentially expressed genes in "HALLMARK_MYOGENESIS" in MPNST vs. PNF (adjusted $P < 0.05$ & $|FC| > 1.2$).

Supplementary table 8. Differentially expressed genes n "REACTOME_EXTRACELLULAR_ MATRIX_ORGANIZATION" in MPNST vs. PNF (adjusted $P < 0.05$ & $|FC| > 1.2$).

Supplementary table 9. Differentially expressed genes in "REACTOME_COLLAGEN_ DEGRADATION" in MPNST vs. PNF (adjusted $P < 0.05$ & $|FC| > 1.2$).

Supplementary table 10. Differentially expressed genes in "REACTOME_DEGRADATION_OF_ THE_EXTRACELLULAR_MATRIX" in MPNST vs. PNF (adjusted $P < 0.05$ & $|FC| > 1.2$).

Acknowledgements

Funding: This publication was supported by funding from the Neurofibromatosis Therapeutic Acceleration Program (NTAP) (https://www.n-tap.org/) at the Johns Hopkins University School of Medicine, to C.A.P. and to A.C.H (Grant ID: 212027, 152009). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of The Johns Hopkins University School of Medicine. Bioinformatics analyses were supported by funding from NTAP to R.J.A. and J.B (Grant ID: 229101). The funders were involved in reviewing the manuscript and providing feedback.

References

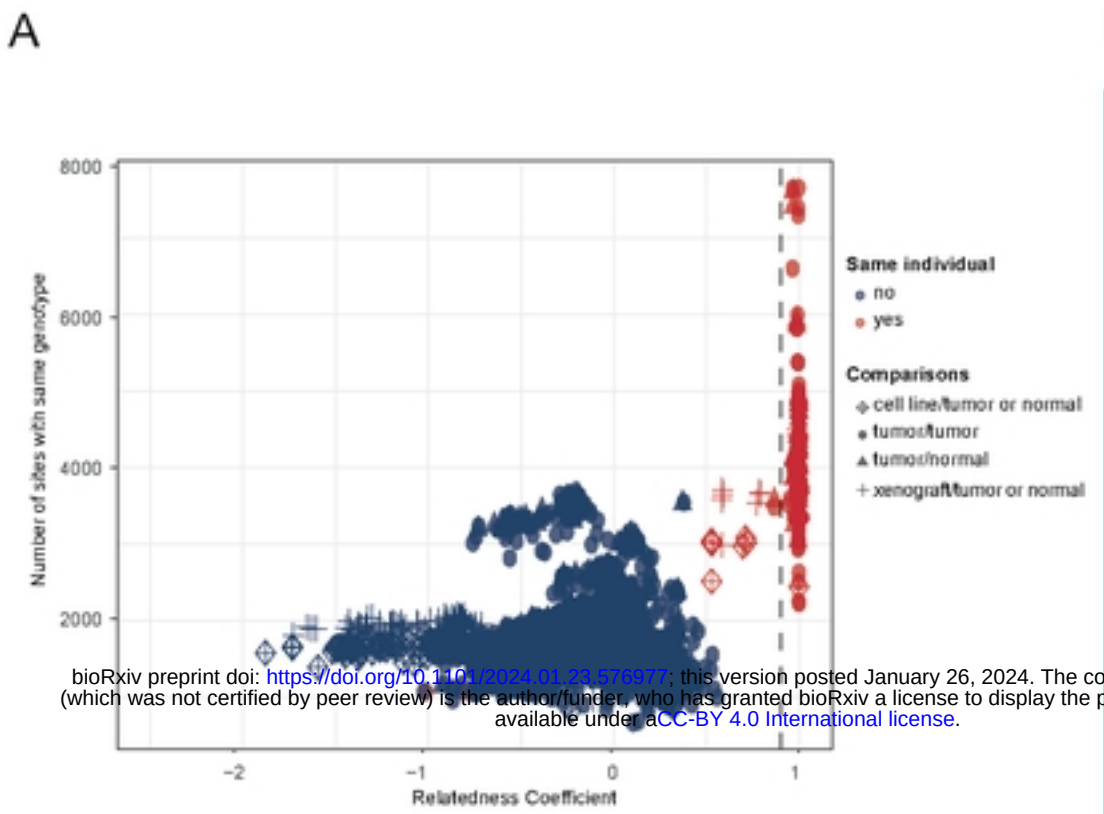
1. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM 2013 [
2. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
3. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8.
4. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11 0 1- 0 33.
5. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Kallberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15(8):591-4.
6. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213-9.
7. Kandoth C, Gao JJ, Qwang M, Mattioni M, Struck A, Boursin Y, et al. mskcc/vcf2maf: vcf2maf v1.6.16 2018 [Available from: doi:10.5281/zenodo.593251.
8. Garcia M, Juhos S, Larsson M, Olason PI, Martin M, Eisfeldt J, et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. F1000Res. 2020;9:63.
9. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14(2):144-61.
10. Fang LT, Afshar PT, Chhibber A, Mohiyuddin M, Fan Y, Mu JC, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biol. 2015;16(1):197.
11. Patel HE, P., Peltzer, A. Botvinnik, O., Manning, J., Sturm, G., Garcia, M. Moreno, D. Vemuri, P., nf-core bot, Pantano, L., Binzer-Panchal, M., Zepper, M., Syme, R., Kelly, G., Hanssen, F., rfenouil, Espinosa-Carrasco, J., marchoeppner, Miller, E., Zhou, P., Guinchard, S., Gabarnet, G., Mertes, C., Straub, D., DiTommaso, P. nf-core/rnaseq: nf-core/rnaseq v3.13.2 - Cobalt Colt 2023 [updated November 21, 2023. Version 3.13.2:[
12. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-9.
13. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3.
14. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction: arXiv; 2018 [
15. Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. BMC Bioinformatics. 2018;19(1):404.

16. Pedersen BS, Bhetariya PJ, Brown J, Kravitz SN, Marth G, Jensen RL, et al. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. Genome Med. 2020;12(1):62.
17. Pollard K, Banerjee J, Doan X, Wang J, Guo X, Allaway R, et al. A clinically and genomically annotated nerve sheath tumor biospecimen repository. Sci Data. 2020;7(1):184.
18. Zhang M, Wang Y, Jones S, Sausen M, McMahon K, Sharma R, et al. Somatic mutations of SUZ12 in malignant peripheral nerve sheath tumors. Nat Genet. 2014;46(11):1170-2.
19. Lee W, Teckie S, Wiesner T, Ran L, Prieto Granada CN, Lin M, et al. PRC2 is recurrently inactivated through EED or SUZ12 loss in malignant peripheral nerve sheath tumors. Nat Genet. 2014;46(11):1227-32.
20. De Raedt T, Beert E, Pasmant E, Luscan A, Brems H, Ortonne N, et al. PRC2 loss amplifies Ras-driven transcription and confers sensitivity to BRD4-based therapies. Nature. 2014;514(7521):247-51.
21. Sohier P, Luscan A, Lloyd A, Ashelford K, Laurendeau I, Briand-Suleau A, et al. Confirmation of mutation landscape of NF1-associated malignant peripheral nerve sheath tumors. Genes Chromosomes Cancer. 2017;56(5):421-6.
22. Cortes-Ciriano I, Steele CD, Piculell K, Al-Ibraheemi A, Eulo V, Bui MM, et al. Genomic Patterns of Malignant Peripheral Nerve Sheath Tumor (MPNST) Evolution Correlate with Clinical Outcome and Are Detectable in Cell-Free DNA. Cancer Discov. 2023;13(3):654-71.
23. Mo J, Moye SL, McKay RM, Le LQ. Neurofibromin and suppression of tumorigenesis: beyond the GAP. Oncogene. 2022;41(9):1235-51.
24. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018;173(2):321-37 e10.
25. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25.
26. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50(D1):D687-D92.
27. Brockman QR, Scherer A, McGivney GR, Gutierrez WR, Voigt AP, Isaacson AL, et al. PRC2 loss drives MPNST metastasis and matrix remodeling. JCI Insight. 2022;7(20).
28. Thomson CS, Pundavela J, Perrino MR, Coover RA, Choi K, Chaney KE, et al. WNT5A inhibition alters the malignant peripheral nerve sheath tumor microenvironment and enhances tumor growth. Oncogene. 2021;40(24):4229-41.
29. Errico A, Stocco A, Riccardi VM, Gambalunga A, Bassetto F, Grigatti M, et al. Neurofibromin Deficiency and Extracellular Matrix Cooperate to Increase Transforming Potential through FAK-Dependent Signaling. Cancers (Basel). 2021;13(10).
30. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12(1):91.

633 31. Viray H, Li K, Long TA, Vasalos P, Bridge JA, Jennings LJ, et al. A prospective,
634 multi-institutional diagnostic trial to determine pathologist accuracy in estimation of
635 percentage of malignant cells. Arch Pathol Lab Med. 2013;137(11):1545-9.
636 32. Smits AJ, Kummer JA, de Bruin PC, Bol M, van den Tweel JG, Seldenrijk KA, et
637 al. The estimation of tumor cell percentage for molecular testing by pathologists is not
638 accurate. Mod Pathol. 2014;27(2):168-74.
639 33. Dudley J, Tseng LH, Rooper L, Harris M, Haley L, Chen G, et al. Challenges
640 posed to pathologists in the detection of KRAS mutations in colorectal cancers. Arch
641 Pathol Lab Med. 2015;139(2):211-8.

642

Figure 1



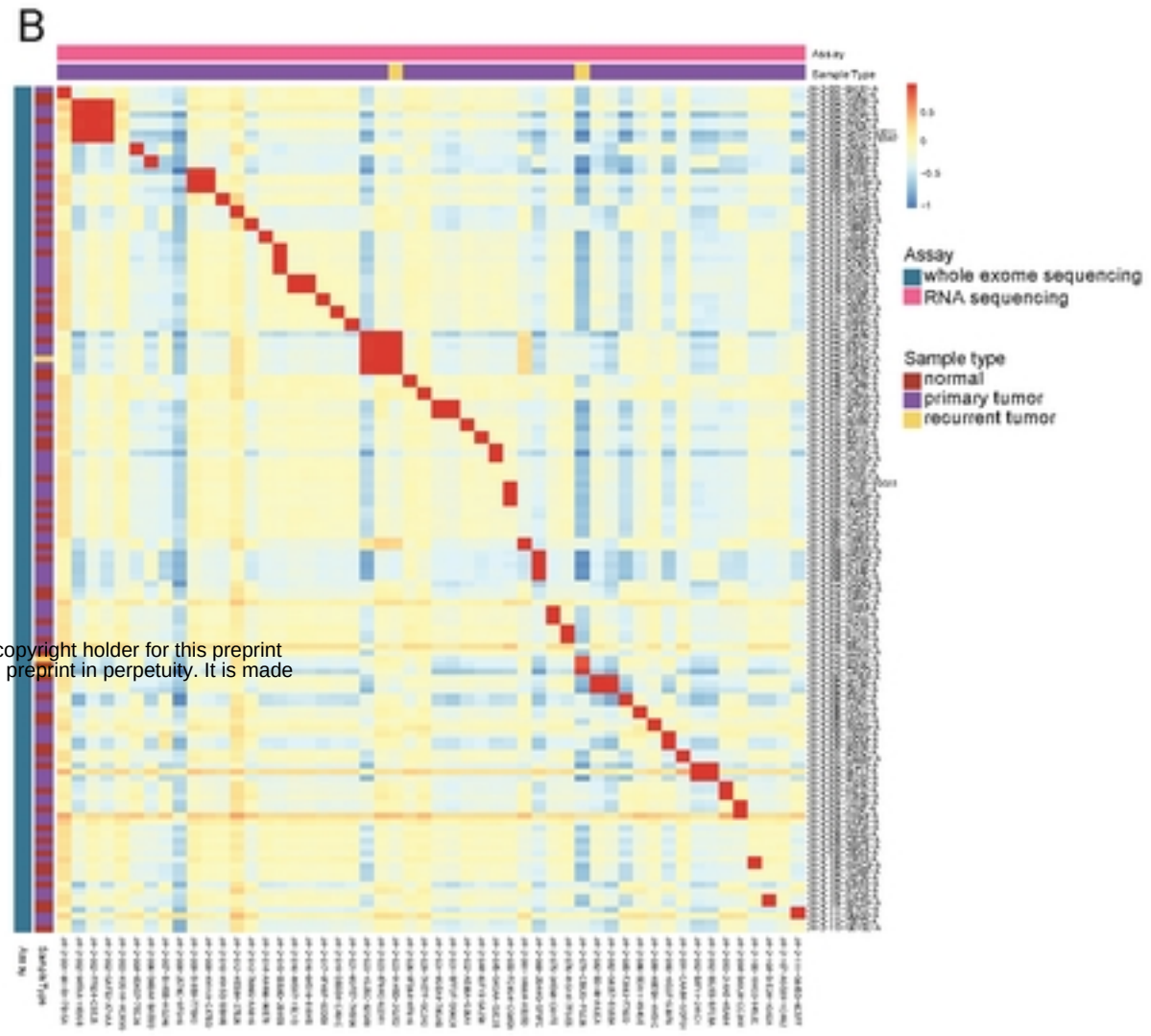


Figure 2

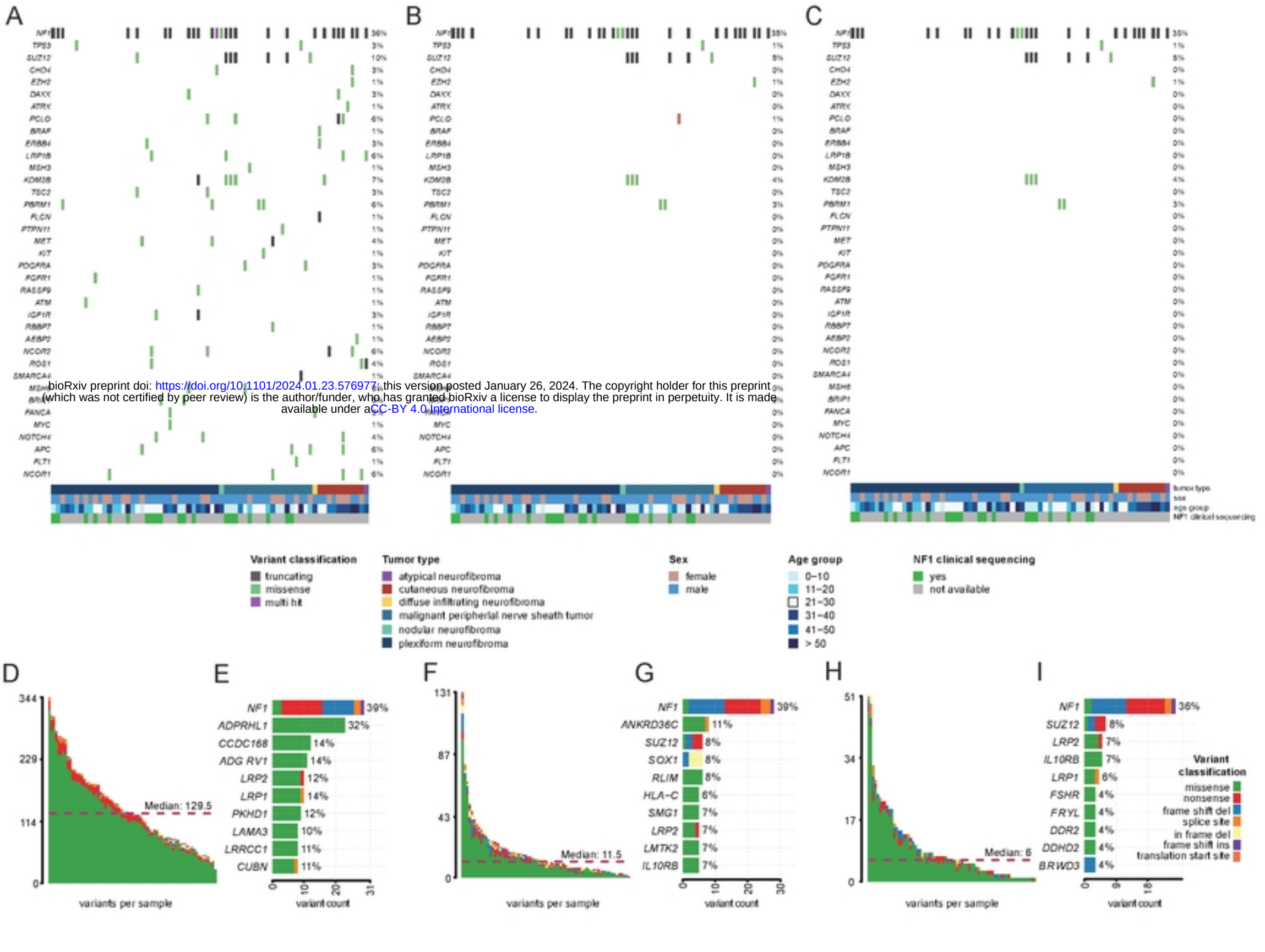
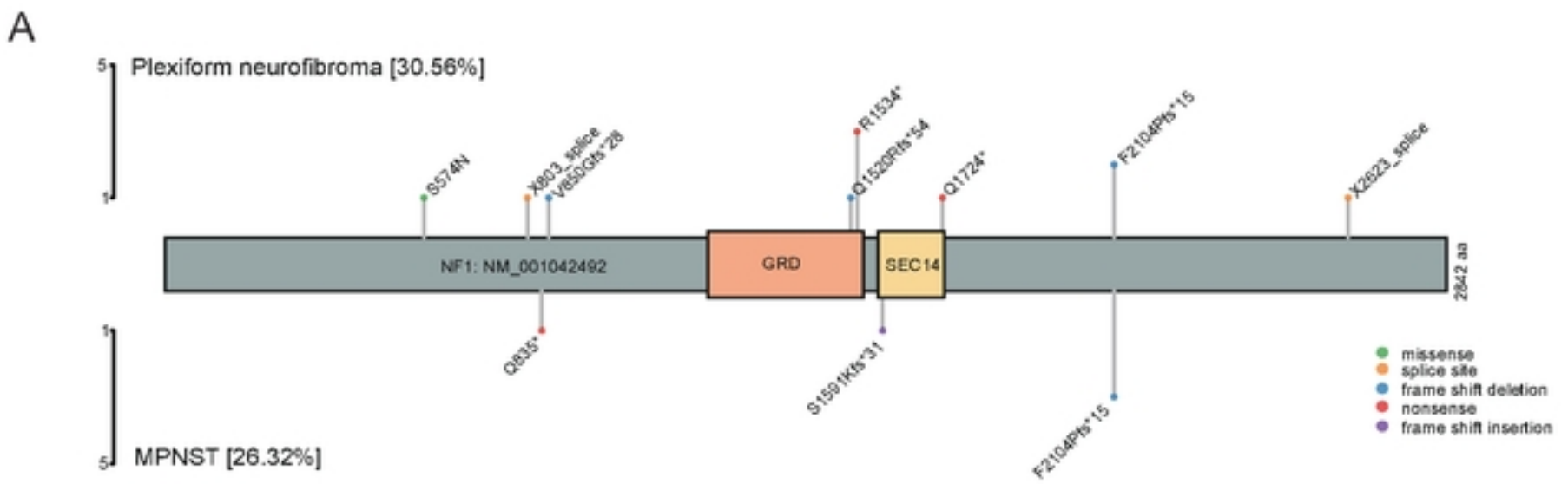
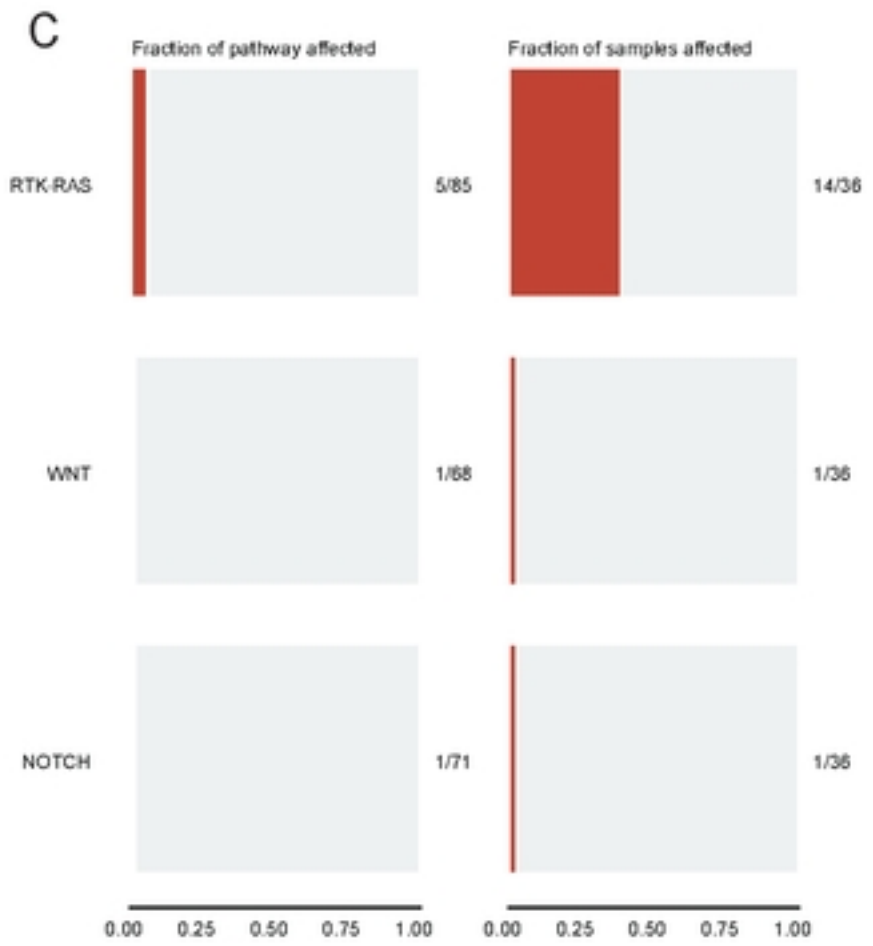


Figure 3





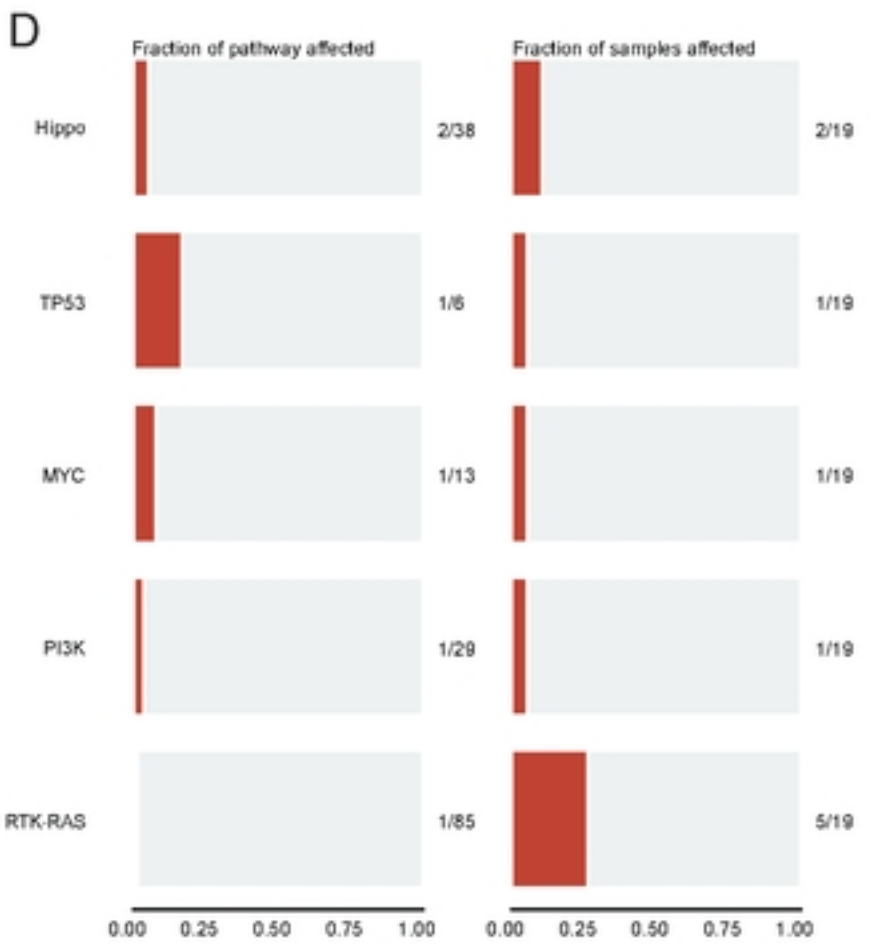
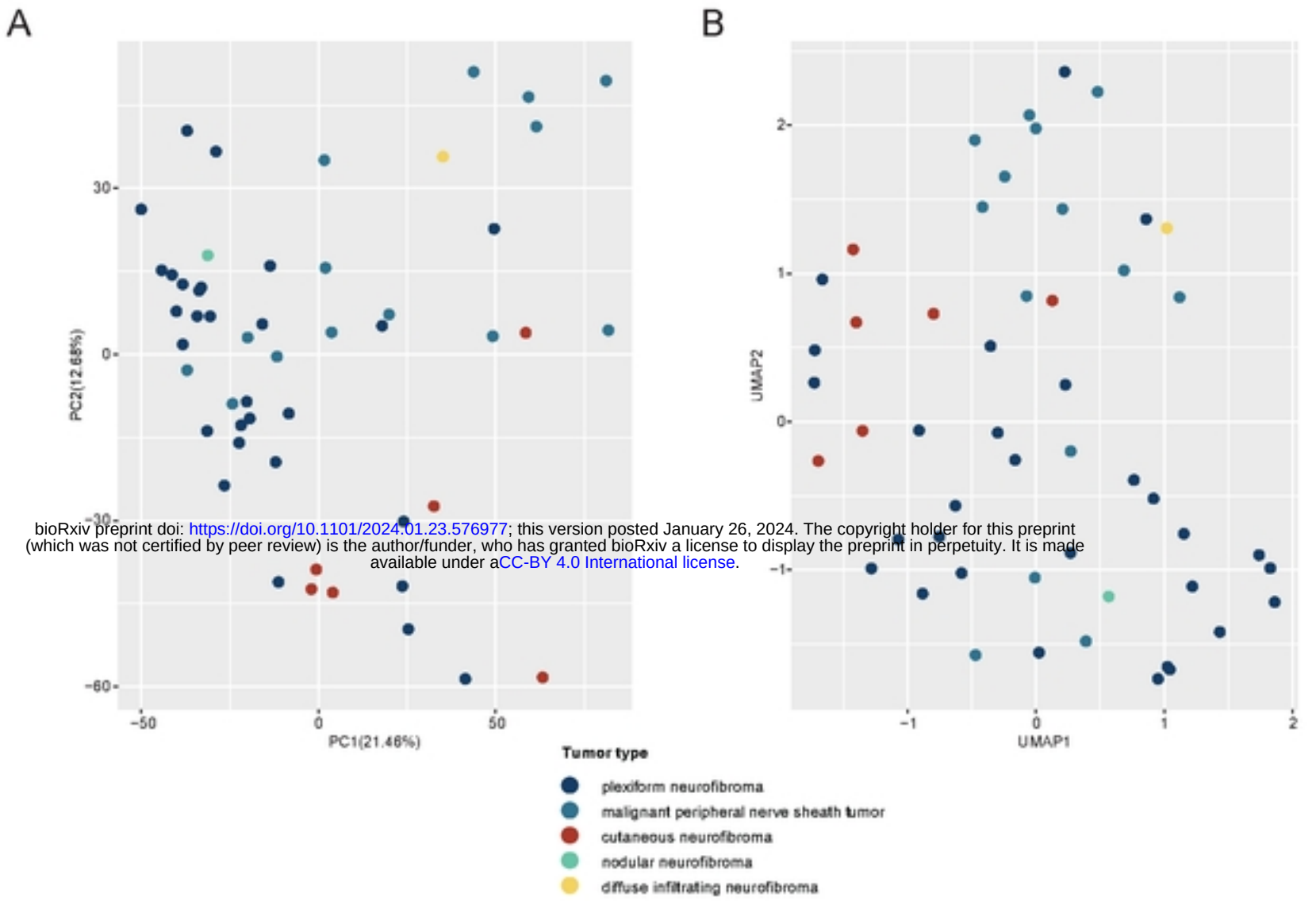
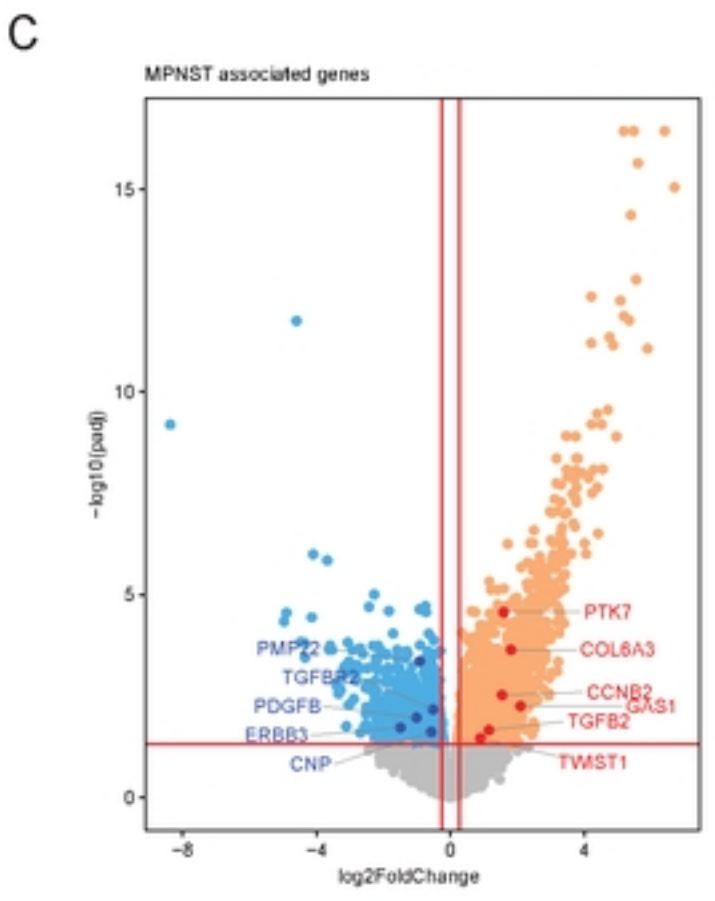


Figure 4





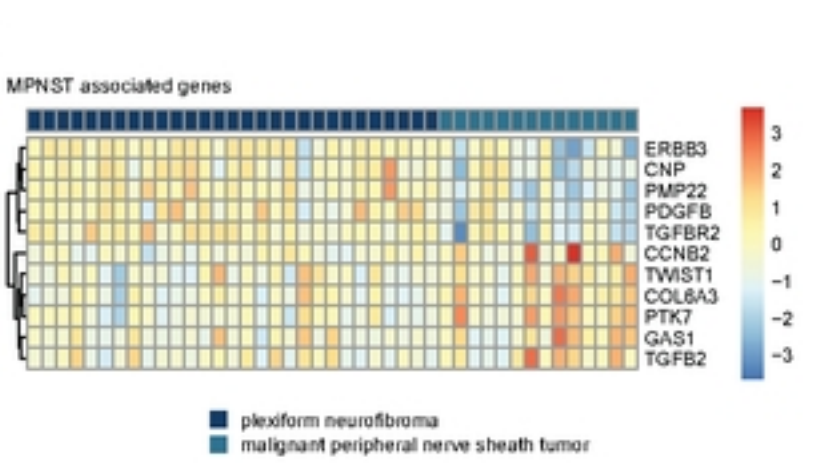


Figure 5