Integrative analysis identifies candidate tumor microenvironment and intracellular signaling pathways that define tumor heterogeneity in NF1

Authors

Jineta Banerjee[1]*, Robert J Allaway[1]*, Jaclyn N Taroni[2], Aaron Baker[1,4], Xiaochun Zhang[5], Chang In Moon[5], Jaishri O Blakeley[6], Justin Guinney[1], Angela Hirbe[5], Casey S Greene[2,3], Sara JC Gosline[1]^

*co-first authors
^corresponding author

Affiliations

[1]Computational Oncology, Sage Bionetworks, Seattle, WA, 98121
[2]Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, PA, 19102
[3]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104
[4]University of Wisconsin-Madison, Madison, WI, 53715
[5]Division of Oncology, Washington University Medical School, St. Louis, MO 63110
[6]Neurology, Neurosurgery and Oncology, Johns Hopkins University, Baltimore, MD 21287

Abstract

Neurofibromatosis type 1 is a monogenic syndrome that gives rise to numerous symptoms including cognitive impairment, skeletal abnormalities, and growth of benign nerve sheath tumors. Nearly all NF1 patients develop cutaneous neurofibromas (cNFs), which occur on the skin surface, while 40-60% of patients develop plexiform neurofibromas (pNFs) which are deeply embedded in the peripheral nerves. Patients with pNFs have a ~10% lifetime chance of these tumors becoming malignant peripheral nerve sheath tumors (MPNSTs). These tumors have a severe prognosis and few treatment options other than surgery. Given the lack of therapeutic options available to patients with these tumors, identification of druggable pathways or other key molecular features could aid ongoing therapeutic discovery studies. In this work, we used statistical and machine learning methods to analyze 77 genomically characterized NF1 tumors with the goal of characterizing key signaling pathways that distinguish these tumors and identifying candidates for drug development. We identified subsets of latent gene expression variables that may be important in the identification and etiology of cNFs, pNFs, other

neurofibromas, and MPNSTs. Furthermore, we characterized the association between these latent variables and genetic variants, immune deconvolution predictions, and protein activity predictions.

Keywords

neurofibromatosis type 1, nerve sheath tumor, cancer, latent variables, machine learning, supervised learning, transfer learning, random forest, metaVIPER, tumor deconvolution

Introduction

Neurofibromatosis type 1 is a neuropathic syndrome that occurs in approximately 1 in 3000 patients worldwide and gives rise to cognitive impairment, skeletal abnormalities and various nerve tumors including gliomas and neurofibromas and is caused by a mutation or deletion in one NF1 allele [1–3]. Nerve sheath tumors affect more than 90% of NF1 patients, mostly in the form of cutaneous neurofibromas (cNFs). These tumors grow at the skin surface and can range in number from 10s to 100s of tumors in a given patient [4]. Neurofibromas that occur deeper in the body, including subcutaneous neurofibromas or plexiform neurofibromas (pNFs), occur in 40-60% of NF1 patients and can cause pain and disfigurement among other symptoms [1,4]. Patients with pNFs have a 10% lifetime risk of these tumors developing into malignant peripheral nerve sheath tumors (MPNSTs) which have a 5-year-survival rate of 40-50% [5,6].

The rise of high-throughput genomic and transcriptomic sequencing has enabled many advances in understanding the molecular etiology of NF1 tumor types [7–11]. Genomic studies of NF1-derived tumors, particularly MPNSTs, have identified key features of tumor growth that could point to potential therapeutic avenues. For example, genomic approaches were recently used to identify the loss of function of polycomb repressor 2 complex components *EED* or *SUZ12*, alongside *CDKN2A* and *NF1* mutations as crucial co-mediators of MPNST transcriptional dysregulation, pathogenesis, and sensitivity to BRD4 inhibitors [9,11]. Others using genomic approaches to explore nerve sheath tumor biology identified *MET* and *HGF* amplification in MPNSTs; furthermore, models of *MET* amplified MPNSTs were subsequently sensitive to the MET inhibitor capmatinib [7]. Transcriptomics-focused approaches have also identified molecular features like MEK signaling, type 1 interferon signaling, and Aurora kinase A as putative therapeutic targets in NF1 tumors [12–14]. Taken together, these and other studies suggest that an integrative approach that combines multiple transcriptomic and genomic datasets might be well poised to identify new therapeutic avenues in MPNSTs and other NF1 related nerve sheath tumors.

Previous genomic profiling studies have demonstrated that many NF1 nerve sheath tumors (MPNST being the exception) are genetically quiet [3,15–17] and lack specific signatures that are predictive of drug response. An approach to compensate for the lack of genetic hotspots in NF1 tumors is to focus on combinations of transcriptomic signatures that may be unique to specific tumor types. In other tumor types, transcriptomic landscapes across cancer datasets

[18,19] have shown that combining RNA-seq data from similar diseases can identify expression profiles that correlate with prognosis [20], predict drug response [21,22] or identify key tumor biology [23].

However, comprehensive analysis of genomic data in NF1 tumors is limited [3]. To enable a larger landscape analysis of NF1 nerve sheath tumors, samples from different studies were collated into a single resource as part of the NF Open Science Initiative, a collaboration between NF-related funding agencies. This resource has been made publicly available through the NF Data Portal which houses high throughput data for neurofibromatosis 1, neurofibromatosis 2, and schwannomatosis [24].

In this work, we reprocessed and analyzed RNA-seq data from 77 NF1 nerve sheath tumor samples to understand the biological differences that give rise to distinct tumor types in NF1 patients. Given the low sample size compared to the large feature space of RNA-seq data, we applied a transfer learning-inspired approach to meaningfully reduce the feature space with minimal decrease in information content. Transfer learning techniques can leverage large well curated datasets such as recount2 [25] to identify latent variables (LVs): groups of genes that show common transcriptomic patterns that are relevant to a specific subset of samples [26,27]. We transferred a machine learning model [27] trained on recount2 to assess LV expression in the NF1 nerve sheath tumor dataset. While many of these LVs map to known signatures (i.e. KEGG, Gene Ontology), others are uncharacterized and may allow the detection of novel and meaningful transcriptomic patterns in NF data. We used supervised machine learning with random forests [28] to isolate combinations of such latent variables to identify specific molecular signatures that may describe the underlying biology unique to each tumor type. Finally, we integrated this information with sample-matched variant data, immune cell signatures [29–31], and protein activity predictions [32] to provide additional biological context to the most important latent variables. This approach reveals biological patterns that underlie different NF1 nerve sheath tumor types and candidate genes and cellular signatures associated with NF1 tumor heterogeneity.

Results

We collected mRNA sequencing, exome sequencing, and whole genome sequencing data from previously published or publicly available resources as depicted in Table 1. We applied a combination of methods to identify biological mechanisms of interest in the samples that had patient-derived transcriptomic data. Briefly, we employed a transfer learning-inspired approach to group transcripts into latent variables (LVs), and then selected the LVs that best separate out tumors by tumor type using an ensemble of random forest models. We evaluated the selected LVs for patterns of immune cell gene expression, protein activity, and gene variants.

Tumor Type	Individuals	Samples	# with genomic variant data
Cutaneous Neurofibroma	11	33	23
MPNST	13	13	1
Neurofibroma	12	12	11
Plexiform Neurofibroma	19	19	5

Table 1 - Summary of individuals and samples for the NF1 nerve sheath tumors used in this study. All samples have gene expression data and a subset have genomic data derived from whole-exome sequencing or whole genome sequencing.

Pan-NF transcriptomic analysis identifies most variable latent variables in NF1

To measure the diversity of the nerve sheath tumors at the transcriptomic level, we re-processed RNA-seq data from three published datasets [33–35] and one unpublished dataset. Despite having four types of nerve sheath tumors (cNFs, pNFs, NFs and MPNSTs) across four datasets, we observed confounding batch effects (Figure 1A) as some tumor types (eg. cNF and NF) were derived from separate studies. These batch effects, together with a lack of transcriptomic data from normal tissue from NF1 patients, precluded us from carrying out any meaningful differential gene expression analysis and motivated us to seek alternate approaches to pathway identification.

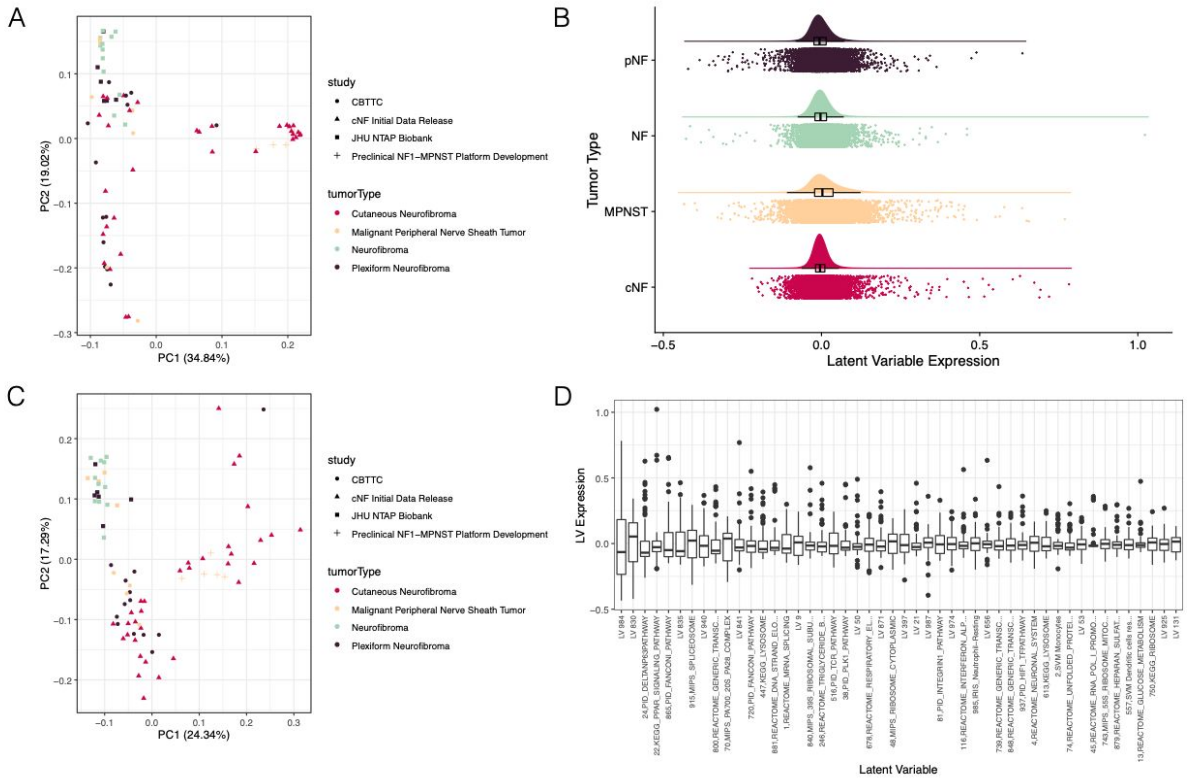


Figure 1 - Transfer learning reduces dimensionality and adds additional context to gene expression datasets. (a) Principal components analysis (PCA) of gene expression data indicates that counts-level data may be batch confounded. (b) The relative distributions of latent variable expression across the four tumor types using a density plot indicates that the majority of LVs have an expression value near 0 and that the four tumor types have similar latent variable expression distributions. (c) PCA of LVs indicates that batch effects, while reduced, may still exist in the LV data (d) A look at the 5% most variable LVs across the cohort of gene expression data indicates that the latent variables represent a wide swath of biological 1processes, as well as some LVs that have no clear association to a defined biological pathway.

To reduce the dimensionality of the RNA-seq data and increase the biological interpretability, we applied MultiPLIER, a transfer learning approach trained on the recount2 dataset . This analysis quantified the expression of 962 latent variables (LVs) in all 77 samples using their gene expression data (Figure 1B, Supplemental Figure 1, Supplemental Table 1). These latent variables were derived using a model trained on recount2, an independent dataset comprising thousands of gene expression experiments. As long as technical confounders in our dataset are independent of the technical confounders in the training dataset, we anticipate that we can find some latent variables that are independent of batch effects. We observed that expression of the latent variables was uniformly distributed across the four tumor types (Figure 1B). Principal components analysis suggested that this method reduced the size of the previously observed

batch effects (Figure 1A,C) as the within-cluster distances are significantly reduced (Supplemental Figure 2, p=2.1e-182 Wilcoxon test).

To assess which LVs characterize the differences between individual tumors we identified the most variable LVs across the entire cohort of samples. To find these, we examined the LVs with the 5% largest standard deviation across all LVs (Figure 1D). Many of the resulting latent variables were found to be associated with known biological mechanisms including metabolic, immune, and transcription related signatures. However, this analysis did not explore the association between tumor types and latent variables. Therefore, we performed additional analyses to (1) identify the gene expression features that best define each tumor type and (2) elucidate the biological mechanisms underlying the most important latent variables.

Ensemble of random forests identify latent variables that robustly describe individual tumor types

In the absence of differential expression and latent variables that readily associate with specific tumor types, we sought to identify combinations of latent variables that correlated with specific tumor types. We used random forest models due to their ability to characterize complex decision spaces better than basic clustering approaches [28]. Random forests achieve this by building decision trees that account for both the values of the features and conditional logic at each 'branch' of the tree. While random forest models are generally used to classify samples, in this study we leverage the capacity of random forest models to identify features in the data that provide meaningful information. Therefore, in this study our focus was to identify important features rather than building the most accurate classifier.

To determine which features would be most informative in distinguishing the different tumor types, we generated a random forest using NF specific latent variable scores (Supplemental Table 1) as described in Methods and depicted in Supplemental Figure 3. Given the limited number of samples in our dataset, particularly for some tumor types, we anticipated high variability in the performance of our supervised learning model. To obtain a distribution of model performance, we generated 500 training sets (using stratified sampling from the full dataset without replacement), each training their own random forest. We used the distribution of model accuracy measurements and feature importance scores from these forests to estimate overall model performance and feature importance (Figure 2A). We then compared the class-specific accuracy measurements (median accuracy scores: cNF=0.97, MPNST=0.67, NF=0.67, pNF=0.57) (Figure 2A) to those from forests built with only the top 40 LVs from each class (a total of 98 LVs) from the first forest ensemble. Using the top 40 set of LVs, we were able to improve model performance for each tumor type on an independent test set (median accuracy scores: cNF=1.00, MPNST=0.86, NF=0.86, pNF=0.75) (Figure 2B). For each tumor type, reducing the random forest to these 98 LVs significantly improved he median accuracy scores for the models (Mood's median test p-values: cNF < 2.2e-16 , MPNST= 3.789e-11 , NF= 7.297e-09 , pNF< 2.2e-16). This suggested that the 98 selected LVs with high importance

scores for each tumor-type were sufficient for characterizing the specific tumor types (Figure 2C).

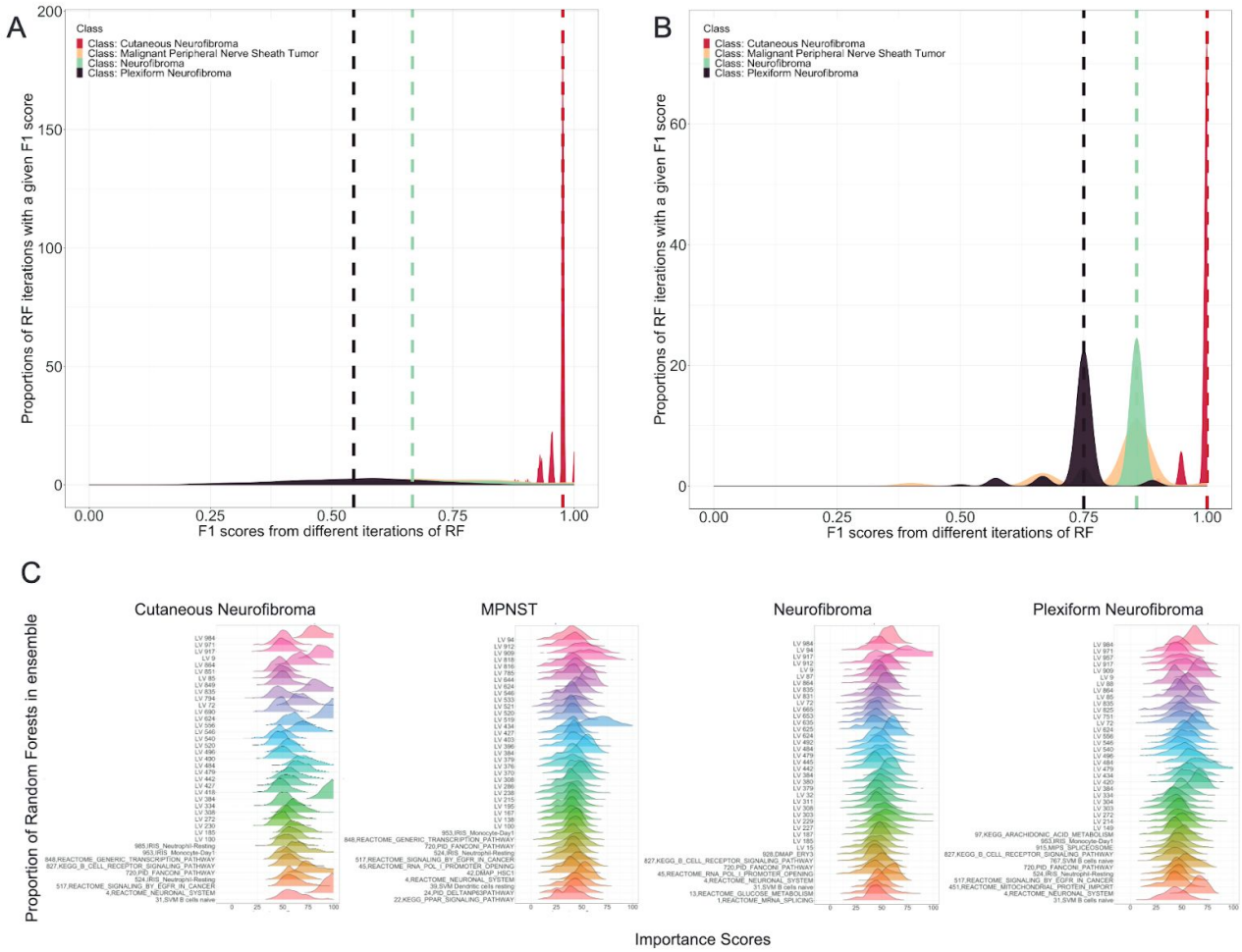


Figure 2: An ensemble of random forests selects the most important latent variables for classifying different tumor types in NF1. (A) Density plot showing the distribution of F1 scores of 500 iterations of independent random forest models using all latent variables. (B) Density plot showing the distribution of F1 scores of 500 iterations of independent random forest models trained using only the top 40 features with high importance scores for each class obtained from models included in (A). (C) Ridgeplots of top 40 latent variables for each tumor type and their importance scores for each class that were selected for later analyses.

Selected latent variables represent distinct biology of nerve sheath tumor types

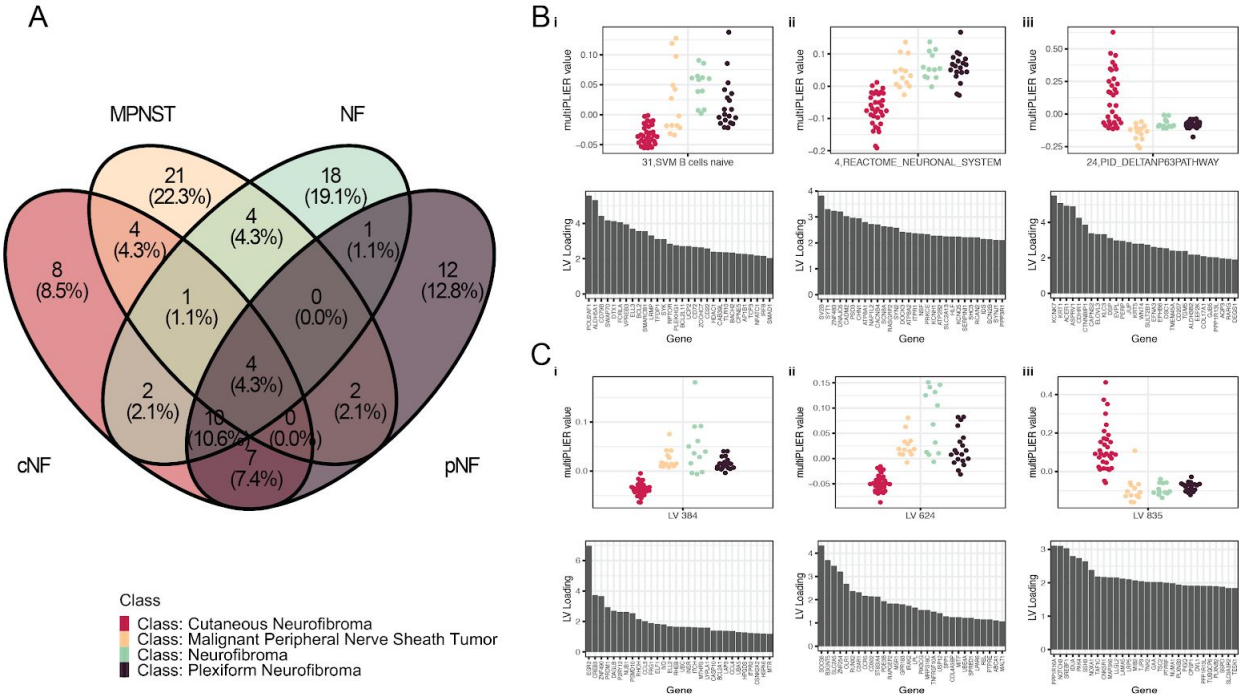


Figure 3: Selected latent variables represent gene combinations unique to each tumor type. (A) Venn diagram showing the distribution of the top 40 LVs from each tumor type. (B i-iii) Genes constituting the latent variables associated with known cell signaling pathways. (C i-iii) Genes constituting the uncharacterized latent variables.

To further probe the biology underpinning the distinction between NF1 tumor types, we focused on the 98 latent variables selected by the ensemble of random forests, depicted in Figure 3 and listed in Supplemental Table 2. We evaluated the tumor-wise LV expression as well as the contributing genes (loadings) for each LV. For example, some latent variables that were predictive of all four tumor types (Figure 3A) showed differences in expression between various tumor types (dot plots in Figures 3Biii and 3Ciii), while others are less distinct (Figures 3Bi and 3i). By investigating the loadings of each LV we tried to map them to known biological pathways. For each LV, the loadings of the constituent genes denote the contribution of that gene to the particular LV. These gene lists associated with the selected LVs can be used to find testable candidates for each tumor type for downstream analysis. For example, one of the four latent variables predictive of all four tumor types was enriched in genes associated with neuronal signaling (Figure 3Bii), suggesting that the presence of neuronal tissue varies across tumor types. Other functionally enriched latent variables, such as those depicted in Figure 4Bi and Figure 3Biii, implicate other biological pathways in tumor growth. However, many of the selected latent variables still remain uncharacterized (Figure 3 Ci-iii, top 3 shown here).

LV scores may be attributed to specific gene variants for specific tumor-types

To characterize the 73 LVs with no associated pathway information, we focused on the 40 samples from our dataset (Table 1) that had matched gene variant data (WGS or exome-Seq) to assess if there were any genes that, when mutated, caused a significant change in LV expression. The results of this calculation are in Supplemental Table 2.

We identified 22 latent variables that were significantly (Benjamini-Hotchberg adjusted $p<0.01$) associated with single gene variants (Supplemental Table 3). These latent variables, along with the genes whose variants are associated with altered expression, are depicted in Figure 4A. This approach failed to identify any genes that were mutated across multiple tumor types. In the list of genes in which variants were associated with changes in latent variable expression, we identified 9 genes with variants in cNFs and 2 genes with variants in one neurofibroma sample. An example of how these variants are associated with the expression of a latent variable are depicted in Figure 4B. Mutations in the 9 cNF-variant genes are associated with lower expression of LV 851. Since most of these variants occur in cNF samples but not in the other tumor types, it's not surprising that this latent variable is down-regulated across all cNFs (Figure 4C). Figure 4D shows that *MAP1B,* which has been reported to play a role in glioblastoma [36] but has not been studied in NF1-linked tumors, is a major contributor towards this LV.

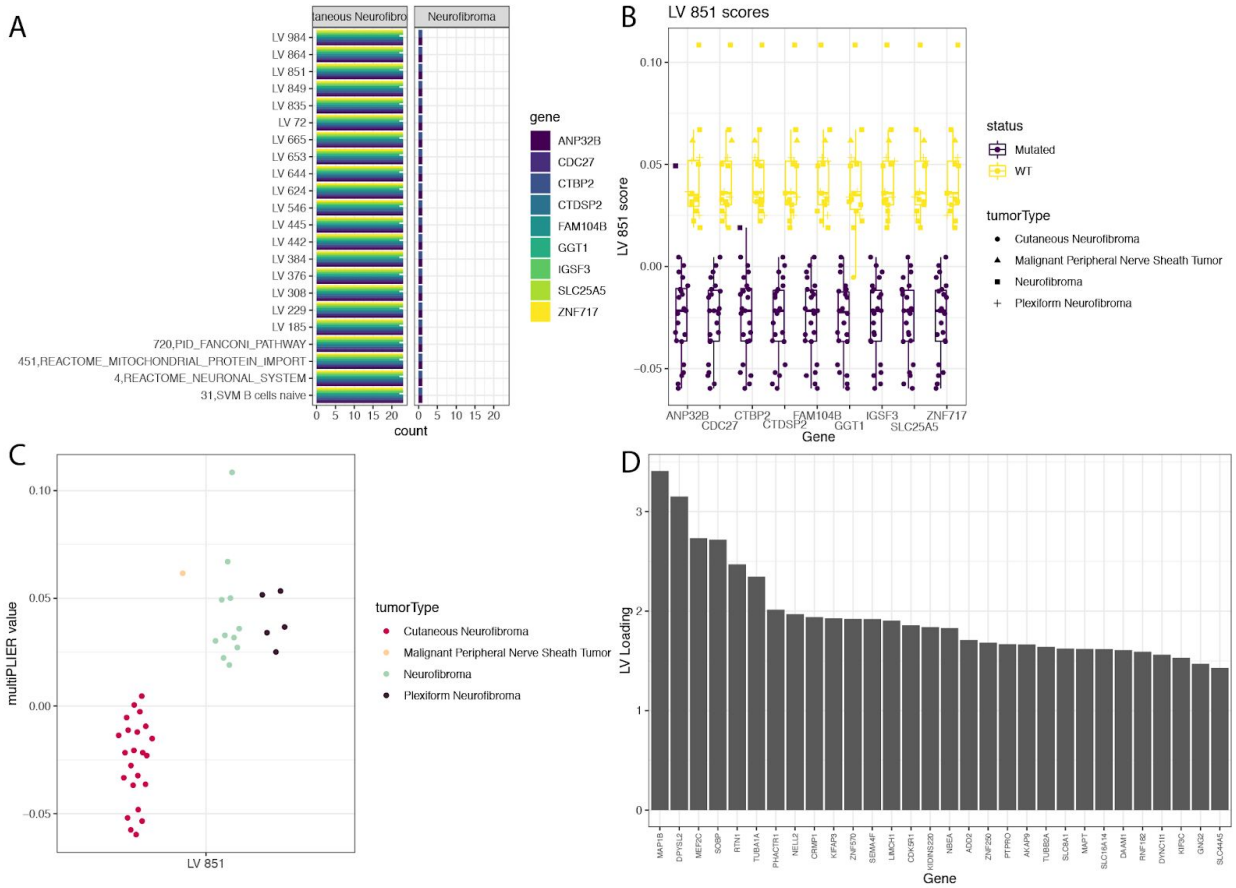


Figure 4: Some genes significantly distinguish expression of latent variables. (A) Latent variables (y-axis) whose values are significantly altered by mutations in specific genes. (B) Expression of LV 851 across tumor samples. (C) Expression of 851 across all samples (D) loadings of LV 851

Selected latent variables represent specific immune cell types in the tumor microenvironment

Given the limited representation of different NF1 tumor types in our genomic variant dataset, we used alternate gene expression metrics to assess the biological underpinnings of the 73 uncharacterized latent variables. We performed tumor immune cell deconvolution analysis [29–31] to identify potential immune infiltration signatures present in the individual tumors using CIBERSORT and MCP-counter (Supplemental Table 4). Specifically, CIBERSORT deconvolution indicated the presence of resting mast cells and M2 macrophages in all tumor types (Figure 5A). The analysis further suggested that all of the tumors have a population of resting CD4+ memory T cells. The results from MCP-counter, depicted in Figure 5B, show a complementary view of the tumor cell types due to the slightly different categorization of cell types. Specifically we see the presence of fibroblasts across all tumors that are not captured in CIBERSORT.

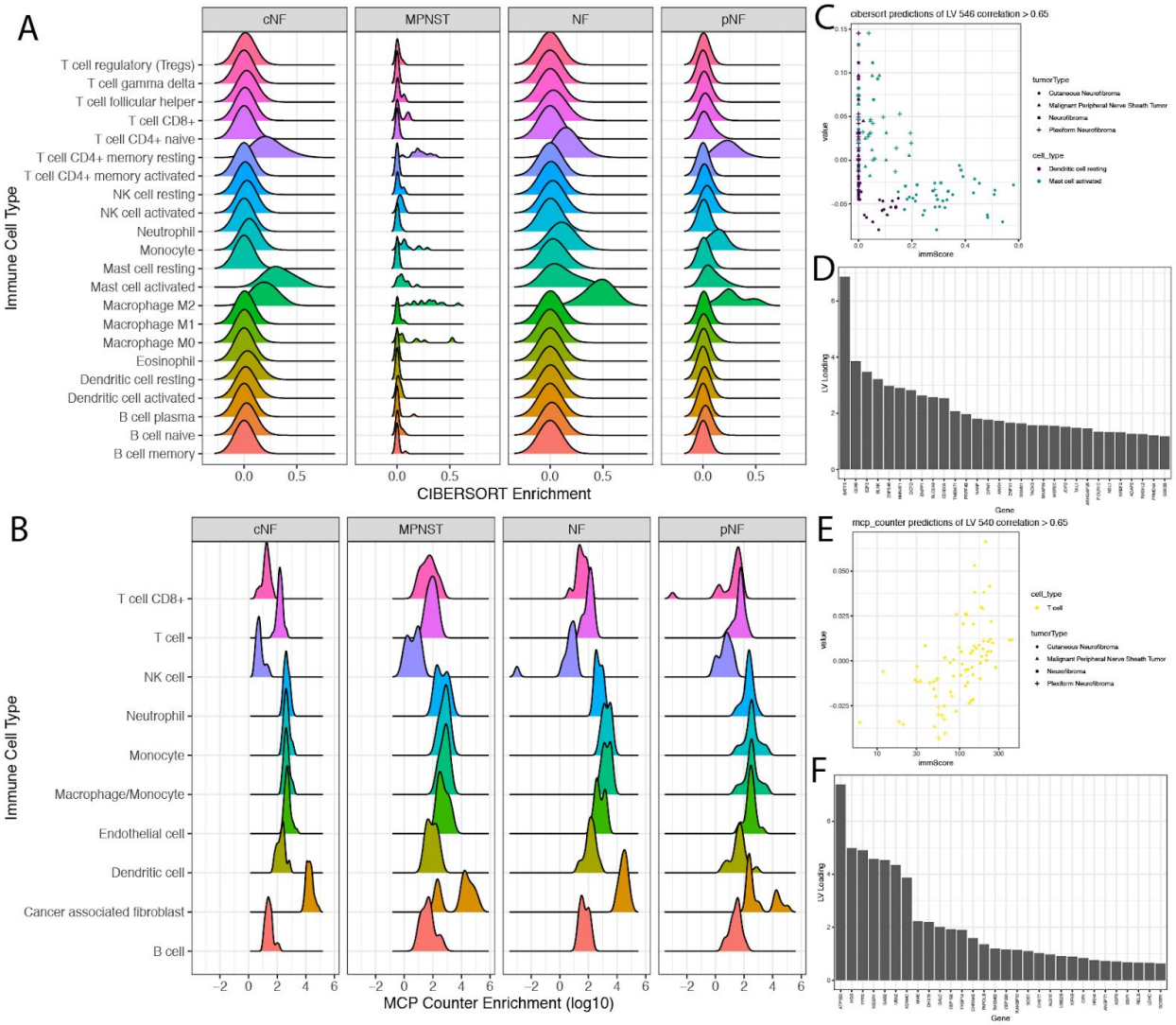


Figure 5: Various immune cell signatures correlate to specific LVs that differentiate tumor types in NF1. (A) CIBERSORT deconvolution of bulk nerve sheath tumor expression data predicted the presence of resting mast cells and M2 macrophages and resting CD4+ memory T cells in all of the tested tumor types. (B) MCP-counter deconvolution of bulk nerve sheath tumor expression data predicted the presence of fibroblast across all tumor types, and diversity in T cell population across tumor types. (C) Correlation of CIBERSORT immune score (x-axis) with expression of latent variable 546 highlights the increased presence of Mast cells and resting dendritic cells in cNFs (circles). (D) Loadings of LV 546. (E) Correlation of MCP-counter score of Tell infiltration (x-axis) with LV 540 (F) Loadings of LV 540.

We then used the immune scores from CIBERSORT and MCP-counter to probe some of the latent variables selected by the random forest model to better understand their role in NF1 tumor biology. We first searched for tumor deconvolution scores that were correlated with latent variable expression across all tumors. Figure 5C and E show specific LVs where the immune scores of the samples highly correlated with a predicted immune cell type. LV 546, for example, (Figure 5C) has a high correlation of activated mast cells in cNF samples. Figure 5E shows that LV 540 has immune scores strongly correlated with T cells in all NF1 tumor-types. Complete results of immune scores across all tumors can be found in Supplemental Table 4. Together, these results suggest that the selected LVs capture signatures from the tumor microenvironment in NF1 samples.

Selected latent variables capture protein regulatory networks in NF1 tumors

Of the 98 latent variables selected by the random forest, 55 could not be characterized via enrichment of the gene loadings (Figure 3), mutational patterns (Figure 4) or immune subtypes (Figure 5). To characterize them, we applied the metaVIPER algorithm [32] to identify specific protein activity measurements that correlated with latent variable expression. This algorithm leverages previously published regulatory network information [37] to infer protein activity in each sample to assign numerical scores of activity across 6168 proteins for each of the 77 samples (Supplemental Table 5). We then measured the correlation of these scores and latent variable scores (Supplemental Table 6, Figure 6A) to identify functional aspects of the latent variables.

As seen in previous work [38], we observed that multiple latent variables exhibit similar correlation patterns with active proteins (Figure 6A). We clustered the correlation scores to see if we could group the latent variables with similar protein activity predictions (Figure 6B). In clustering the latent variable-protein activity scores, we observed 5 distinct clusters of latent variables with similar predicted protein activities (Figure 6B). Aggregation of each cluster of latent variables (mean expression within each cluster) demonstrated that these functional clusters were differentially expressed in the different NF1 tumor types (Figure 6C).

We then assessed the druggability of each of these 5 clusters by taking the average correlation of each protein within the cluster and performing gene set enrichment analysis against a database of small molecules with known biological targets [39]. This enabled the identification of drugs and drug-like compounds that are significantly enriched for targets in each cluster (Supplemental Table 7). For example, cluster 2, which is expressed in pNF, NF, and MPNST more than it is expressed in cNF, has correlated VIPER proteins that are enriched for both clinically approved drugs (dovitinib) and druglike small molecules (Figure 6D). Furthermore, we found that cluster 3 was enriched for compounds that affect cell cycle progression (e.g. dinaciclib, abemaciclib), while clusters 1 and 5 were enriched for molecules that inhibit histone deacetylases like CUDC-101 and analogs (Supplemental Figure 4).

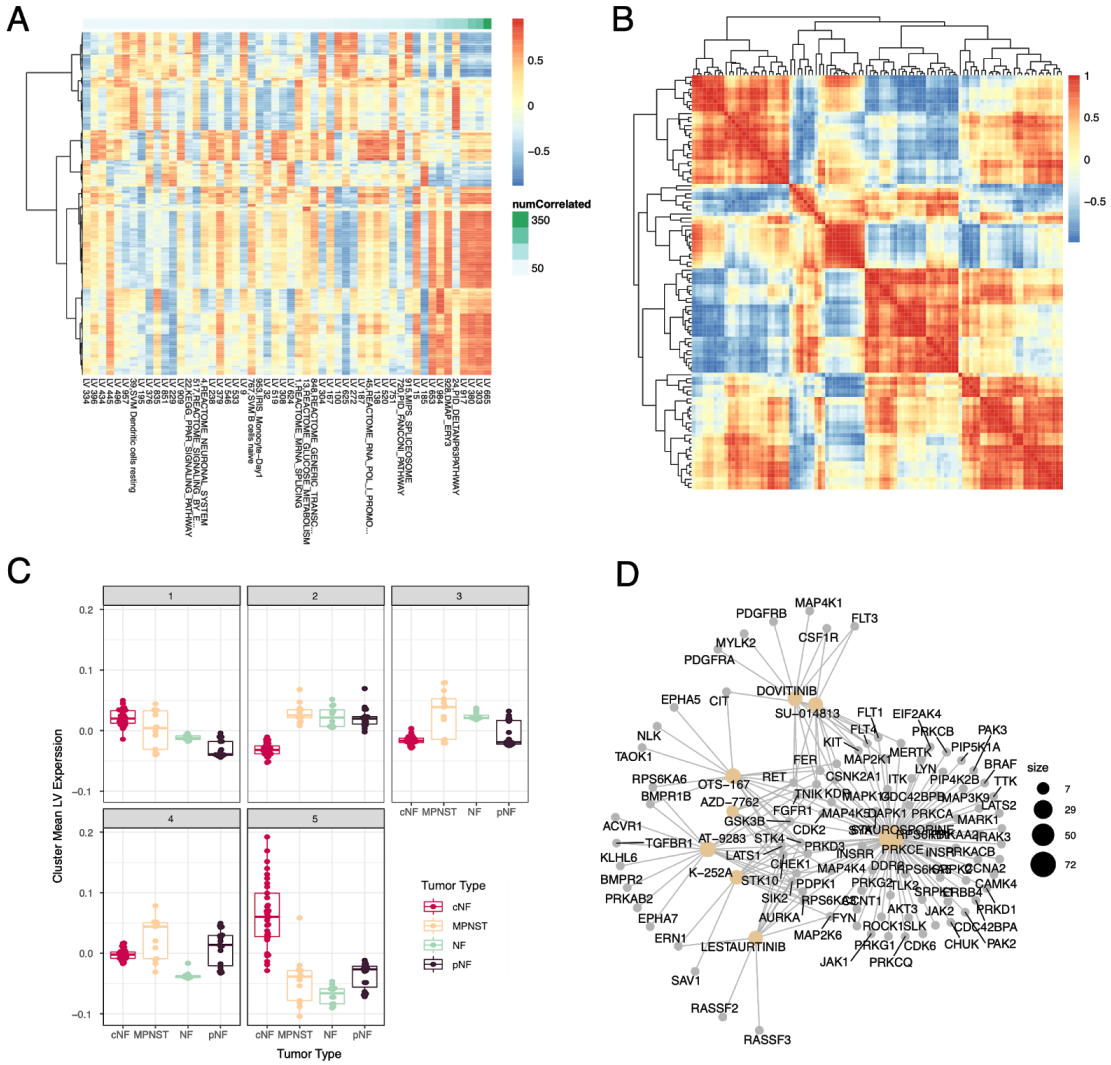


Figure 6: Integration of protein activity information with LVs can identify candidate drug targets for different NF1 tumor types. (A) A heatmap of correlation scores of known proteins with regulatory networks (or *regulons*) that are represented in the characterized and uncharacterized LVs selected above. The green bar across the top depicts how many protein activity scores have a Spearman correlation greater than 0.65. (B) Clustering of the LV-correlated VIPER proteins highlights 5 clusters of latent variables with similar VIPER protein predictions, suggesting that these 5 clusters may have functional overlap. (C) Mean LV expression within the clusters highlights differential expression within the clusters across tumor types. (D) Drug set enrichment analysis of the average VIPER protein correlation of cluster 2 identifies some drugs and preclinical molecules that are enriched with targets in this cluster.

Discussion

NF1 is the most common of all neurofibromatosis syndromes and is caused by the loss of function of *NF1* gene (a known tumor suppressor) due to mutation or deletion. However, NF1 patients show a great deal of phenotypic heterogeneity [1,37,38]. Identification of candidate cellular signaling pathways that differentiate between various tumor types is key for understanding the biology underlying such phenotypic diversity as well as predicting progression of tumor types towards malignancy. In this study we have integrated various *in silico* resources and analytical techniques to identify candidate genes or pathways unique to different tumor types in an attempt to generate testable hypotheses. By capturing complex gene expression patterns using latent variables (LVs), we identified combinations of LVs that were important to classify tumor types. An ensemble of random forests was then used to select 98 latent variables that were important and sufficient for identifying and classifying the various tumor types with reasonably high accuracy (Figure 2). The selected LVs were then subjected to downstream analyses using different data modalities to gain insight into the composition and relevance of the LVs in the context of NF1 (Figure 3-6). The selected latent variables that correlated with known pathways using tumor deconvolution methods confirmed the presence of previously described tumor microenvironment components. Investigation into previously uncharacterised LVs in the context of NF1 suggested the presence of (a) candidate genes for targeted experiments, (b) previously known as well as unknown tumor microenvironment components, (c) candidate tissue-specific protein regulatory networks for future drug screening experiments.

To interpret the results of these analyses we first evaluated the gene loadings of the latent variables that were associated with one or more tumor types. We found that two of the top LVs with lower expression in cNF but higher expression in other tumor types, LV 384 and LV 624 (Figure 3C, Figure 6), had ties to known Schwann cells and NF1 tumor biology. Specifically, *EGR2,* one of the major components of LV 384, has been implicated in diseases associated with the myelin sheath like Charcot-Marie Tooth Disease (another disease of the Schwann cell) and is thought to play a role in pathways associated with myelination [38,39]. Indeed, clinical case studies of patients with concurrent NF1-induced tumors and Charcot-Marie Tooth disease have been reported in the literature suggesting a possible overlap in underlying biology [40–42]. While the role of *EGR2* in Schwann cell differentiation is still under active investigation [43], *EGR2*-driven pathways have been found to be significantly downregulated in Lats1/2 deficient Schwann cell based MPNST models [44]. Furthermore, EGR2 expression in immune cells has been shown to be important for activation of M1/M2 macrophages [45] as well as regulation of CD4+ T cells [46]. Our downstream analysis, using all the selected LVs and tumor deconvolution techniques, identifies the enrichment of M2 macrophage and CD4+ T cell markers as tumor microenvironment components in our samples (Figure 5A). Similarly, *RUNX2* (a major component of LV 624), has been shown to drive neurofibromagenesis by repression of *Pmp22* (a Schwann cell component) [52].

Alternatively, we also evaluated LVs that uniquely associated with specific tumor types. For example LV 24 was found to be an important feature for classifying MPNST tumor samples but not the other benign tumor samples (Figure 3Biii). LV 24 significantly associated with the ΔNp63 pathway (FDR < 0.05), a pathway with implications in determining malignancy and poor survival in various subtypes of pancreatic and squamous cell carcinomas [47]. ΔNp63 signaling in the central nervous system is believed to play a role in neural precursor cell survival and neural stem cell dynamics [48,49], but its role in formation or maintenance of malignant NF1 tumors like MPNSTs is relatively unexplored. Overall, the evidence surrounding EGR2, RUNX2 and ΔNp63 pathway suggest that LV 384, LV 624, and LV 24 may be promising candidates for experimental followup.

Beyond looking at individual genes that comprise the latent variables, we employed orthogonal algorithms that measured tumor immune activity and regulatory protein activity to identify specific signatures represented by the latent variables. Tumor deconvolution methods (CIBERSORT, MCP-counter) confirmed previously described observations like the presence of mast cells in NF1 model systems and patient tumors [50–56]. They also suggested the presence of T cells in human tumor samples. This has previously been described in mouse models of NF1 tumors [57]. Furthermore, in humans, systemic T-cell burden has been found to correlate with NF1 nerve sheath tumor progression; T-cells have also been observed in NF1 gliomas [10,58].

Through the metaVIPER protein regulatory activity predictions, we were able to identify putative therapeutic candidates for further evaluation. Specifically, we found two clusters of latent variables expressed in NF, pNF, and MPNST that had regulatory proteins enriched in targets of dovitinib and drug-like small molecules that inhibit receptor tyrosine kinases, as well as cyclin-dependent kinase (CDK) inhibitors such as abemaciclib or dinaciclib. These findings are consistent with previous studies that identify CDK inhibitors as useful in models of *NF1*-deficient or RAS-dysregulated tumors [59–62]. Since latent variables that are correlated with HDACs were found to be expressed most highly in cNFs (Cluster 5, Figure 6C, Supplemental Figure 4C) and MPNSTs (cluster 1, Figure 6C, Supplemental Figure 4A), compounds like CUDC-101 and analogs could be potential candidates for treating cNF and MPNST. Indeed, HDAC inhibitors were previously found to be efficacious in *in vitro* and *in vivo* models of MPNSTs [59,61]. Thus our results further suggest that this therapeutic approach might be feasible in both MPNSTs and cNFs. Drugs found in other clusters such as dovitinib and lestaurinib in cluster 2, which is expressed most in MPNSTs, pNFs, and NFs (Figure 6C, 6D), may also merit further study.

While the present *in silico* study brings forth various candidate genes and pathways for further follow up, it also presents a few limitations that could be mitigated with future studies, particularly for tumor types with limited samples such as MPNSTs and neurofibromas. Most notably, analyzing genetic variants across tumor types failed to identify relevant variant signatures (Figure 4). This highlights the challenges in variant analyses using samples with limited class representation and motivates our focus on transcriptional signatures. Additional

genomic and transcriptomic data from the same biobanks or additional tumor datasets will improve our ability to identify recurrent genetic markers of tumor type. Furthermore, additional data from tumor-adjacent normal tissue would greatly add value to additional analyses based on differential gene expression. Such differential expression analyses were not possible within the scope of this work since these data are not currently available. Additionally, future studies comparing genomic signatures identified here to other publicly-available tumor expression and variant data (e.g. the Cancer Genome Atlas or the International Cancer Genome Consortium [19]) may identify genomic similarities between peripheral nerve sheath tumors and other tumors.

In conclusion, this work proposes a short list of testable hypotheses involving specific biological signatures for NF1 deficient nerve sheath tumors. Verification of these mechanisms in *in vitro* and *in vivo* models of nerve sheath tumors as well as human NF1 nerve sheath tumor tissue needs active and extensive experimental work. While we have analyzed tumor datasets from four different studies, the addition of other neurofibromatosis-driven tumor datasets will greatly aid identification of commonalities or critical differences to inform therapeutic decisions across the family of neurofibromatoses. This study, together with future work, will guide the repositioning of clinically approved drugs in the context of NF1.

Materials and Methods

Implementation and data and code availability

All analysis was performed using the R programming language. A comprehensive list of packages used and their versions are available in an *renv* lockfile in the GitHub repository. Key packages used include tidyverse [65], PLIER [26], synapser [66], tximport [67], immundeconv [31], and viper [32]. All data analyzed in this article is stored on the NF Data Portal [24] (http://nfdataportal.org) with analyses stored at http://synapse.org/nf1landscape. To recapitulate the analysis from these data, all relevant code can be found at https://github.com/Sage-Bionetworks/NF_LandscapePaper_2019.

Sequencing data collection and processing

Gene expression data were collected from four independent studies and processed via a workflow at https://github.com/Sage-Bionetworks/rare-disease-workflows/tree/master/rna-seq-workflow to be stored on the NF Data Portal (Table 2). Specifically raw fastq files were downloaded from Synapse and transcripts were quantified using the Salmon pseudo-alignment tool [63] with Gencode V29 transcriptome. Links to specific datasets and the access teams required to download them can be found using Synapse ID *syn21221980*.

Dataset Name	Project Name	Table Name	Access Team
WashU Biobank	Preclinical NF1-MPNST Platform Development	WashU Biobank RNA-seq data	WUSTL MPNST PDX Data Access
JHU Biobank [35]	A Nerve Sheath Tumor Bank from Patients with NF1	Biobank RNASeq Data	JHU Biobank Data Access
cNF Patient Data [33]	Cutaneous Neurofibroma Data Resource	cNF RNASeq Counts	CTF cNF Resource Data Access Group
CBTTC Data [34]	Children's Brain Tumor Tissue Consortium	CBTTC RNASeq Counts	CBTTC Data Access Group

Table 2: Description of the gene expression datasets used in the present article

Genomic variant data was collected from exome-Seq [35] or whole-genome sequencing [33]. VCFs were processed via vcf2maf (https://github.com/mskcc/vcf2maf) according to the workflow located at https://github.com/Sage-Bionetworks/rare-disease-workflows/tree/master/gene-variant-workflow and then uploaded to the NF Data Portal. A list of datasets and the access teams required to download them can be found in Table 3 or syn21266269.

Dataset Name	Assay	Table Name	Access Team	Project
JHU Biobank Exome-Seq Data	exomeSeq	Biobank ExomeSeq Data	JHU Biobank Data Access	A Nerve Sheath Tumor Bank from Patients with NF1
cNF WGS Data	wholeGenomeSeq	cNF WGS Harmonized Data	CTF cNF Resource Data Access Group	Cutaneous Neurofibroma Data Resource

Table 3: Description of the genomic variant datasets used in the present article

Latent variable calculation and selection

We analyzed transcriptomic data from NF in the context of latent variables from MultiPLIER, a machine learning resource designed to aid in rare disease analyses [27]. Raw transcriptomic data from NF were retrieved from the NF Data Portal, reprocessed, and stored in Synapse as described above. Salmon output files (quant.sf) files were imported into an R session and converted to HUGO gene names using the tximport [64] and org.Hs.eg.db [65] packages.

MultiPLIER reuses models that were trained on large public compendia. We retrieved a model [66] that was previously trained on the recount2 RNA-seq dataset [25,67] and then used it to assess the expression of latent variables in the pan-NF dataset. We retrieved code for this analysis from the public repository for MultiPLIER [27] (https://github.com/greenelab/multi-plier). To project the NF data into the MultiPLIER model, we used the GetNewDataB function. This analytical approach is described in more detail in a machine learning training module (https://github.com/AlexsLemonade/training-modules) produced by the Alex's Lemonade Stand Foundation's Childhood Cancer Data Lab under an open source license.

In addition, to ensure orthogonality in the final set of latent variables [68], we calculated the pairwise Pearson correlation of all latent variables by comparing the gene loadings (Supplemental Figure 1) and identified non-self correlations greater than 0.5. We eliminated one of each of these highly-intercorrelated (Pearson correlation >0.5) latent variables. This process resulted in a final set of 962 latent variables for further analysis. Code for our implementation of this material is available on GitHub at https://github.com/Sage-Bionetworks/NF_LandscapePaper_2019.

We used the R function `prcomp` to compute the principal components for Figures 1A and 1C both using genes (Figure 1A) and latent variables (Figure 1C).

Generation of ensemble of Random Forests for feature selection

To select gene expression patterns of interest we used ensembles of random forests to sufficiently resample our modestly sized dataset. We compared random forest models built using gene expression data as well as latent variables.

Algorithm implementation: The main algorithm was implemented using the *caret* and *randomforest* packages in R [69,70]. Supplemental Figure 3 outlines the steps involved in the generation of the ensemble of random forests. Briefly, the full dataset was first split into 80% *model* set and 20% *independent test* set. The function *createDataPartition* was used to create balanced splits of the data according to the tumor type. We tuned two parameters to the random forest algorithm, *mtrys*, and *ntrees*, using an iterative approach, evaluating *mtrys* values of 1 to

100 and *ntrees* values of 250, 500, 1000, and 2000. We selected the optimal values using 5-fold cross-validation, using both gene expression data and latent variables as input features.The gene expression-based forest exhibited best accuracy with the parameters *mtry* = 29 and *ntrees* = 500, and the latent variable forest model had parameters of , *mtry*=51, *ntrees*=1000). We then split the *model* training set further to generate 500 samples of *training* data (75%) and hold-out *test* data (25%) (balanced splits randomly sampled without replacement). Each of these *training* and *test* datasets were used to train separate random forests to obtain a distribution of F1 scores and feature importance scores (n=500). Given our noisy dataset with limited sample size, the distribution of feature and F1 scores enabled estimation of confidence intervals for the features importance as well as model performance.

Feature Selection: To select the top features for a specific class (i.e. tumor type), we calculated the median importance score of each feature from the distribution of importance scores generated through 500 iterations of random forests. The top 40 features were selected according to the mean decrease of the Gini index. The union of top features from all classes was then used as a restricted feature set to train another 500 iterations of random forests as described above. However, each new forest trained with restricted feature-set was tested using the *independent test set* to examine the performance of the model on a completely unseen dataset. The median F1 scores of the new ensemble of forests were compared to the previous ensemble of random forests. Improvement of median F1 scores for each class in the final ensemble of random forests compared to the earlier one suggested that the selected features from each class were sufficiently informative for classification. This subset of features were then selected for downstream analyses.

Immune subtype prediction

To understand the relative immune infiltration across the nerve sheath tumors studied we used two tumor deconvolution tools: CIBERSORT and MCP-counter, as implemented through the *immunedeconv* R package [29–31]. Analysis is located at https://github.com/Sage-Bionetworks/NF_LandscapePaper_2019 and results were uploaded to a Synapse table (syn21177277) that includes the tumor-specific immune cell scores for both algorithms as well as associated tumor metadata.

MetaVIPER

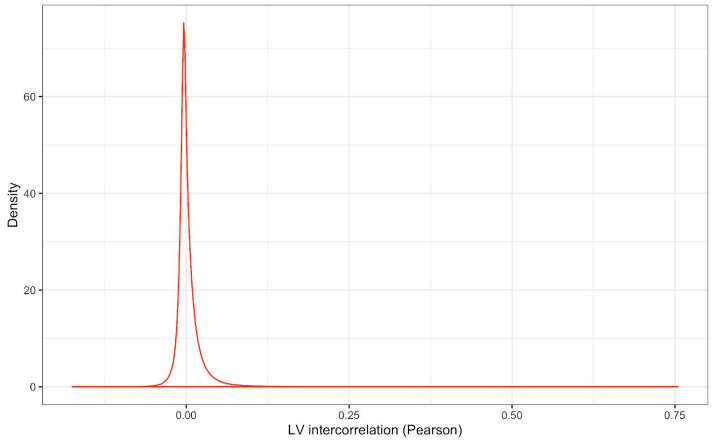
We applied the metaVIPER algorithm [32] to infer protein regulator activity based on the tumor gene expression profiles. This algorithm builds transcriptional regulatory networks across the cancer genome atlas (TCGA) [37,69] and uses these to build consensus predictions for a sample of other origin. The resulting analysis is uploaded to Synapse and stored at syn21259610 along with tumor-specific metadata.

VIPER correlation clustering and drug enrichment analysis

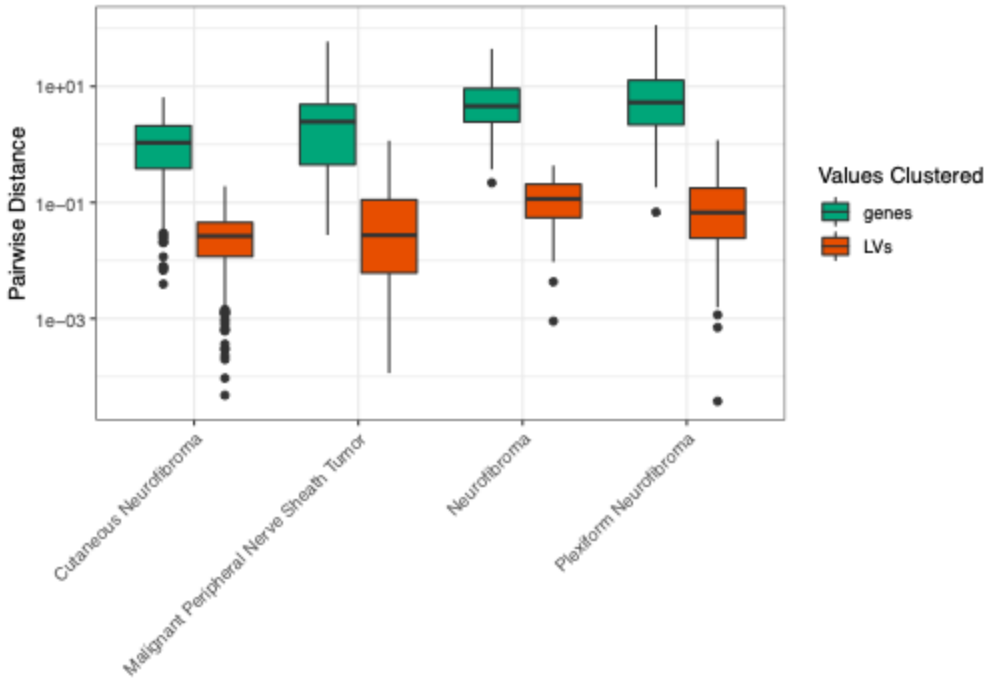
A heatmap and subclusters of latent variables that had similar VIPER protein predictions were generated using the pheatmap R package [71] (Figure 6B). We observed 5 clear clusters of latent variables; these clusters were defined using the R cutree function to isolate the 5 clusters and their contents. We then calculated the mean correlation of each VIPER protein within each cluster to generate a consensus protein activity prediction for each latent variable cluster. Then, we used gene-set enrichment analysis (via the clusterProfiler R package [72]) to assess whether drug targets were enriched in the 5 consensus protein lists. Drugwise target lists were obtained from the Drug Target Explorer database [73]. Significant enrichment was defined as any positively enriched drug (i.e. a VIPER protein positively correlated with the latent variable cluster) with a Benjamini-Hotchberg corrected p-value < 0.05. Results were plotted using ggplot and enrichplot [74,75]. To plot the LV cluster expression by tumor type, we calculated the mean expression of all latent variables for each cluster and for each tumor sample.

Supplementary Materials

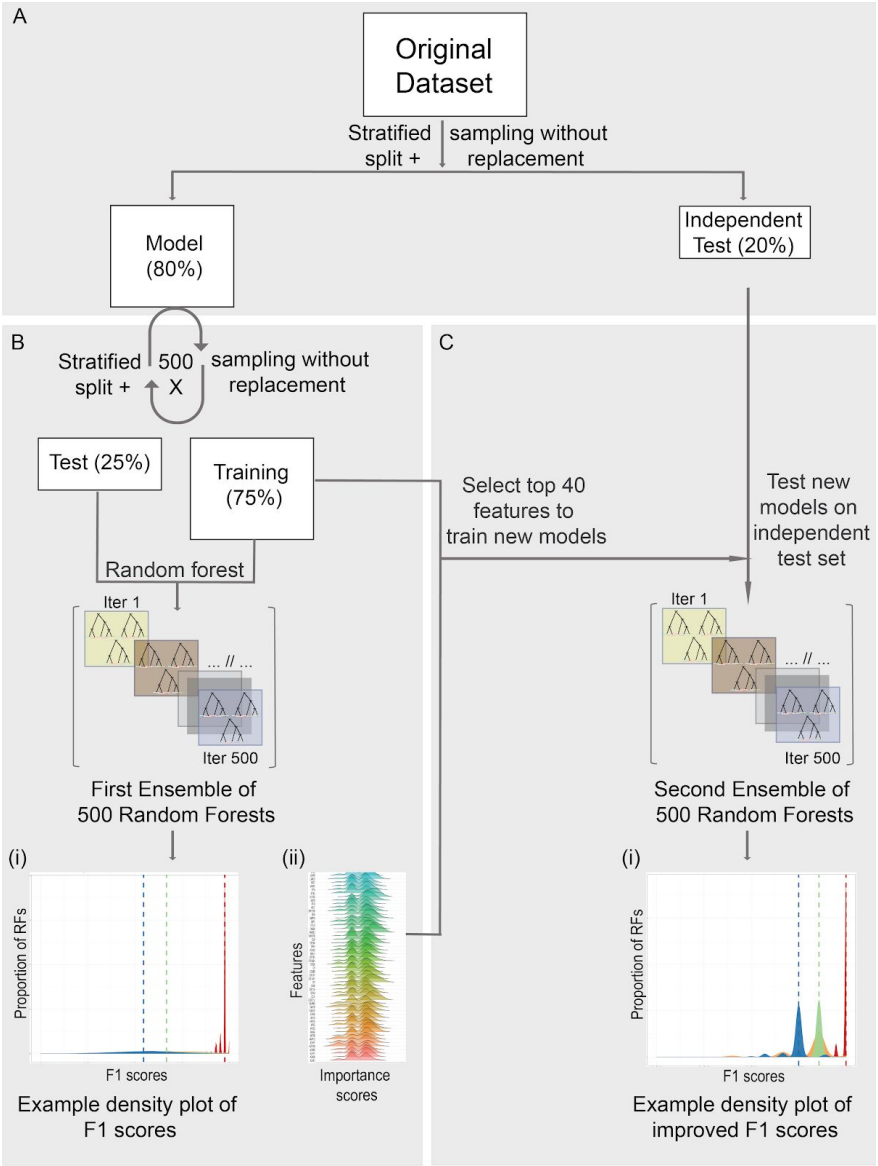
Supplemental Table 1: Pan-NF1 MultiPLIER results across RNA-seq data for 77 nerve sheath tumors.
Supplemental Table 2: Significance of latent variable status being correlated with gene mutation status.
Supplemental Table 3: Summary of latent variables selected by random forest model and their correlated genes and signatures.
Supplemental Table 4: Tumor deconvolution scores across NF1 patient samples
Supplemental Table 5: metaVIPER protein scores across all NF1 patient samples
Supplemental Table 6: Spearman correlation scores of RF-selected latent variables with metaVIPER scores.
Supplemental Table 7: Drug set enrichment analysis of latent variables.



Supplemental Figure 1: (a) A density plot indicating the pairwise Pearson correlation of LVs (not including self-self pairs). A threshold of 0.5 was selected to eliminate non-orthogonal latent variables.

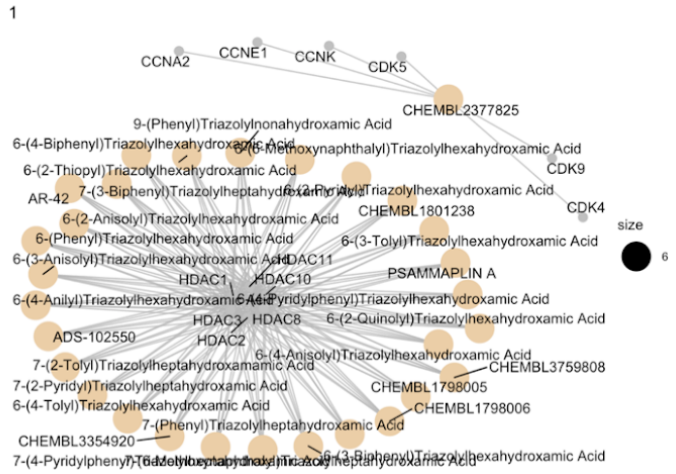


Supplemental Figure 2: A boxplot representing the decrease in pairwise distances between all tumor samples of the same tumor type from Figure 1A (green, based on gene clustering) to 1C (orange, based on LV clustering).

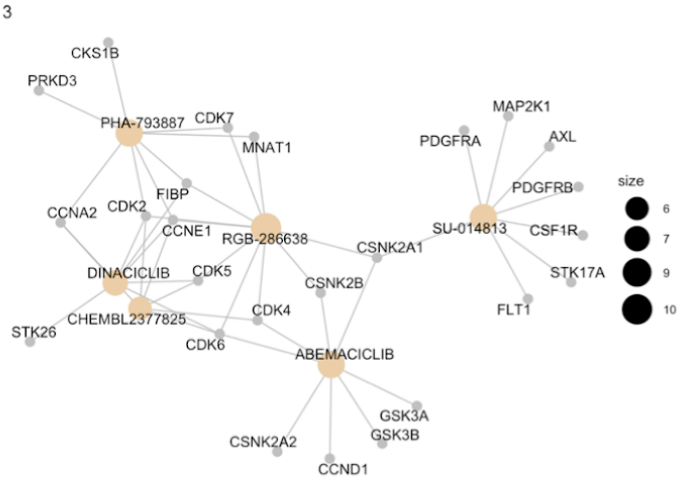

A
Original Dataset
Stratified split +
sampling without replacement
Model (80%)
Independent Test (20%)
B
Stratified split +
500 X
sampling without replacement
Test (25%)
Training (75%)
Random forest
Iter 1
... // ...
Iter 500
First Ensemble of 500 Random Forests
(i)
Proportion of RFs
F1 scores
Example density plot of F1 scores
(ii)
Features
Importance scores
C
Select top 40 features to train new models
Test new models on independent test set
Iter 1
... // ...
Iter 500
Second Ensemble of 500 Random Forests
(i)
Proportion of RFs
F1 scores
Example density plot of improved F1 scores

Supplemental Figure 3: Schematic showing the generation of the ensemble of random forests. (A) Stratified split of original dataset into "model" set and "independent test" set. (B) Generation of first ensemble of random forests using training set and test set generated from the "model" set for all 962 latent variables. (i) example density plot showing distribution of model accuracy (F1) scores for each class. (ii) example plot showing distribution of feature importance scores from the 500 random forests. (C) Generation of the second ensemble of random forests using restricted feature set of only 98 latent variables. The newly generated models were tested on the "independent test" set. (i) example plot showing the distribution of model accuracy scores from the second ensemble of random forests.

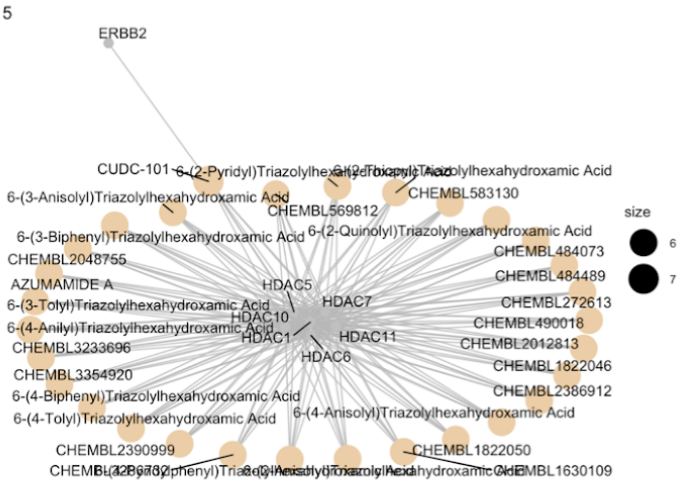
A


1
CCNA2
CCNE1
CCNK
CDK5
CHEMBL2377825
9-(Phenyl)Triazolylnonahydroxamic Acid
6-(4-Biphenyl)Triazolylhexahydroxamic Acid
6-(6-Methoxynaphthalyl)Triazolylhexahydroxamic Acid
CDK9
6-(2-Thiopyl)Triazolylhexahydroxamic Acid
AR-42
CHEMBL1801238
CDK4
6-(2-Anisolyl)Triazolylhexahydroxamic Acid
6-(Phenyl)Triazolylhexahydroxamic Acid
6-(3-Tolyl)Triazolylhexahydroxamic Acid
6-(3-Anisolyl)Triazolylhexahydroxamic Acid
HDAC11
HDAC1
HDAC10
PSAMMAPLIN A
6-(4-Anilyl)Triazolylhexahydroxamic Acid
HDAC3
HDAC8
6-(2-Quinolyl)Triazolylhexahydroxamic Acid
ADS-102550
HDAC2
6-(4-Anisolyl)Triazolylhexahydroxamic Acid
7-(2-Tolyl)Triazolylheptahydroxamamic Acid
CHEMBL3759808
7-(2-Pyridyl)Triazolylheptahydroxamic Acid
CHEMBL1798005
6-(4-Tolyl)Triazolylhexahydroxamic Acid
CHEMBL1798006
7-(Phenyl)Triazolylheptahydroxamic Acid
CHEMBL3354920
size
6

B


3
CKS1B
PRKD3
PHA-793887
CDK7
MNAT1
MAP2K1
PDGFRA
AXL
PDGFRB
FIBP
CCNA2
CDK2
CCNE1
RGB-286638
SU-014813
CSF1R
CSNK2A1
DINACICLIB
CDK5
STK17A
CSNK2B
CHEMBL2377825
CDK4
FLT1
STK26
CDK6
ABEMACICLIB
GSK3A
GSK3B
CSNK2A2
CCND1
size
6
7
9
10

C


5
ERBB2
CUDC-101
6-(2-Pyridyl)Triazolylhexahydroxamic Acid
6-(3-Anisolyl)Triazolylhexahydroxamic Acid
CHEMBL583130
CHEMBL569812
6-(3-Biphenyl)Triazolylhexahydroxamic Acid
6-(2-Quinolyl)Triazolylhexahydroxamic Acid
CHEMBL2048755
CHEMBL484073
AZUMAMIDE A
HDAC5
CHEMBL484489
6-(3-Tolyl)Triazolylhexahydroxamic Acid
HDAC7
CHEMBL272613
HDAC10
6-(4-Anilyl)Triazolylhexahydroxamic Acid
HDAC1
HDAC11
CHEMBL490018
CHEMBL3233696
HDAC6
CHEMBL2012813
CHEMBL1822046
CHEMBL3354920
CHEMBL2386912
6-(4-Biphenyl)Triazolylhexahydroxamic Acid
6-(4-Tolyl)Triazolylhexahydroxamic Acid
6-(4-Anisolyl)Triazolylhexahydroxamic Acid
CHEMBL2390999
CHEMBL1822050
CHEMBL1630109
size
6
7

Supplemental Figure 4: Small molecule-target networks show enrichment of LV-dependent target classes. (A) Cluster 1 of LV-correlated VIPER proteins is enriched for HDAC inhibitors. (B) Cluster 3 of LV-correlated VIPER proteins is enriched for kinase inhibitors, particularly inhibitors of kinases responsible for cell cycle progression (CDKs). (C) Cluster 5 of LV-correlated VIPER proteins is enriched for HDAC inhibitors.

Acknowledgements

This work was supported with funds from the Neurofibromatosis Therapeutic Acceleration Program. We would also like to acknowledge our colleagues at Sage Bionetworks, Matthew Wall and Michael Mason, for many helpful conversations.

Author Contributions

RJA, JB, JNT, CSG, AB, JG, SJCG designed and performed the analyses. JB, RJA, CSG, AB, JG, and SJCG wrote and edited the manuscript. XZ, CIM, and AH performed and supervised sequencing of previously-unpublished tumor data. JOB, JG, SJCG supervised the study and guided the analyses.

Conflicts of Interest

The authors have no conflicts of interest to declare, with the exception of Dr. Hirbe, who is serving as a co-editor for this issue of *Genes*. Dr. Hirbe has abstained from any editorial duties relating to this manuscript.

References

1. Friedman, J. *Neurofibromatosis 1*; University of Washington, Seattle, 1993;
2. Evans, D.G.; Howard, E.; Giblin, C.; Clancy, T.; Spencer, H.; Huson, S.M.; Lalloo, F. Birth incidence and prevalence of tumor-prone syndromes: estimates from a UK family genetic register service. *Am. J. Med. Genet. A.* **2010**, *152A*, 327–332.
3. Allaway, R.J.; Gosline, S.J.C.; La Rosa, S.; Knight, P.; Bakker, A.; Guinney, J.; Le, L.Q. Cutaneous neurofibromas in the genomics era: current understanding and open questions. *Br. J. Cancer* **2018**, *118*, 1539–1548.
4. Plotkin, S.R.; Bredella, M.A.; Cai, W.; Kassarjian, A.; Harris, G.J.; Esparza, S.; Merker, V.L.; Munn, L.L.; Muzikansky, A.; Askenazi, M.; et al. Quantitative Assessment of Whole-Body Tumor Burden in Adult Patients with Neurofibromatosis. *PLoS ONE* **2012**, *7*.
5. Stucky, C.-C.H.; Johnson, K.N.; Gray, R.J.; Pockaj, B.A.; Ocal, I.T.; Rose, P.S.; Wasif, N. Malignant Peripheral Nerve Sheath Tumors (MPNST): The Mayo Clinic Experience. *Ann. Surg. Oncol.* **2012**, *19*, 878–885.

6. Yuan, Z.; Xu, L.; Zhao, Z.; Xu, S.; Zhang, X.; Liu, T.; Zhang, S.; Yu, S. Clinicopathological features and prognosis of malignant peripheral nerve sheath tumor: a retrospective study of 159 cases from 1999 to 2016. *Oncotarget* **2017**, *8*, 104785–104795.
7. Peacock, J.D.; Pridgeon, M.G.; Tovar, E.A.; Essenburg, C.J.; Bowman, M.; Madaj, Z.; Koeman, J.; Boguslawski, E.A.; Grit, J.; Dodd, R.D.; et al. Genomic Status of MET Potentiates Sensitivity to MET and MEK Inhibition in NF1-Related Malignant Peripheral Nerve Sheath Tumors. *Cancer Res.* **2018**, *78*, 3672–3687.
8. Brohl, A.S.; Kahen, E.; Yoder, S.J.; Teer, J.K.; Reed, D.R. The genomic landscape of malignant peripheral nerve sheath tumors: diverse drivers of Ras pathway activation. *Sci. Rep.* **2017**, *7*, 14992.
9. Lee, W.; Teckie, S.; Wiesner, T.; Ran, L.; Prieto Granada, C.N.; Lin, M.; Zhu, S.; Cao, Z.; Liang, Y.; Sboner, A.; et al. PRC2 is recurrently inactivated through EED or SUZ12 loss in malignant peripheral nerve sheath tumors. *Nat. Genet.* **2014**, *46*, 1227–1232.
10. D'Angelo, F.; Ceccarelli, M.; Tala; Garofano, L.; Zhang, J.; Frattini, V.; Caruso, F.P.; Lewis, G.; Alfaro, K.D.; Bauchet, L.; et al. The molecular landscape of glioma in patients with Neurofibromatosis 1. *Nat. Med.* **2019**, *25*, 176–187.
11. De Raedt, T.; Beert, E.; Pasmant, E.; Luscan, A.; Brems, H.; Ortonne, N.; Helin, K.; Hornick, J.L.; Mautner, V.; Kehrer-Sawatzki, H.; et al. PRC2 loss amplifies Ras-driven transcription and confers sensitivity to BRD4-based therapies. *Nature* **2014**, *514*, 247–51.
12. Patel, A.V.; Eaves, D.; Jessen, W.J.; Rizvi, T.A.; Ecsedy, J.A.; Qian, M.G.; Aronow, B.J.; Perentesis, J.P.; Serra, E.; Cripe, T.P.; et al. Ras-Driven Transcriptome Analysis Identifies Aurora Kinase A as a Potential Malignant Peripheral Nerve Sheath Tumor Therapeutic Target. *Clin. Cancer Res.* **2012**, *18*, 5020–5030.
13. Jessen, W.J.; Miller, S.J.; Jousma, E.; Wu, J.; Rizvi, T.A.; Brundage, M.E.; Eaves, D.; Widemann, B.; Kim, M.-O.; Dombi, E.; et al. MEK inhibition exhibits efficacy in human and mouse neurofibromatosis tumors. *J. Clin. Invest.* **2013**, *123*, 340–347.
14. Choi, K.; Komurov, K.; Fletcher, J.S.; Jousma, E.; Cancelas, J.A.; Wu, J.; Ratner, N. An inflammatory gene signature distinguishes neurofibroma Schwann cells and macrophages from cells in the normal peripheral nervous system. *Sci. Rep.* **2017**, *7*, 43315.
15. Pemov, A.; Li, H.; Patidar, R.; Hansen, N.F.; Sindiri, S.; Hartley, S.W.; Wei, J.S.; Elkahloun, A.; Chandrasekharappa, S.C.; Boland, J.F.; et al. The primacy of NF1 loss as the driver of tumorigenesis in neurofibromatosis type 1-associated plexiform neurofibromas. *Oncogene* **2017**, *36*, 3168–3177.
16. Pemov, A.; Hansen, N.F.; Sindiri, S.; Patidar, R.; Higham, C.S.; Dombi, E.; Miettinen, M.M.; Fetsch, P.; Brems, H.; Chandrasekharappa, S.; et al. Low mutation burden and frequent loss of CDKN2A/B and SMARCA2, but not PRC2, define pre-malignant neurofibromatosis type 1-associated atypical neurofibromas. *Neuro-Oncol.* **2019**.
17. Thomas, L.; Kluwe, L.; Chuzhanova, N.; Mautner, V.; Upadhyaya, M. Analysis of NF1 somatic mutations in cutaneous neurofibromas from patients with high tumor burden. *neurogenetics* **2010**, *11*, 391–400.
18. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **2013**, *45*, 1113–1120.
19. International network of cancer genome projects. *Nature* **2010**, *464*, 993–998.
20. Guinney, J.; Dienstmann, R.; Wang, X.; de Reyniès, A.; Schlicker, A.; Soneson, C.; Marisa, L.; Roepman, P.; Nyamundanda, G.; Angelino, P.; et al. The Consensus Molecular Subtypes of Colorectal Cancer. *Nat. Med.* **2015**, *21*, 1350–1356.
21. Ali, M.; Aittokallio, T. Machine learning and feature selection for drug response prediction in

precision oncology applications. *Biophys. Rev.* **2019**, *11*, 31–39.
22. Costello, J.C.; Heiser, L.M.; Georgii, E.; Gönen, M.; Menden, M.P.; Wang, N.J.; Bansal, M.; Ammad-ud-din, M.; Hintsanen, P.; Khan, S.A.; et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **2014**, *32*, 1202–1212.
23. Way, G.P.; Sanchez-Vega, F.; La, K.; Armenia, J.; Chatila, W.K.; Luna, A.; Sander, C.; Cherniack, A.D.; Mina, M.; Ciriello, G.; et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. *Cell Rep.* **2018**, *23*, 172-180.e3.
24. Allaway, R.J.; La Rosa, S.; Verma, S.; Mangravite, L.; Guinney, J.; Blakeley, J.; Bakker, A.; Gosline, S.J.C. Engaging a community to enable disease-centric data sharing with the NF Data Portal. *Sci. Data* **2019**, *6*.
25. Collado-Torres, L.; Nellore, A.; Kammers, K.; Ellis, S.E.; Taub, M.A.; Hansen, K.D.; Jaffe, A.E.; Langmead, B.; Leek, J.T. Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **2017**, *35*, 319–321.
26. Mao, W.; Zaslavsky, E.; Hartmann, B.M.; Sealfon, S.C.; Chikina, M. Pathway-level information extractor (PLIER) for gene expression data. *Nat. Methods* **2019**, *16*, 607–610.
27. Taroni, J.N.; Grayson, P.C.; Hu, Q.; Eddy, S.; Kretzler, M.; Merkel, P.A.; Greene, C.S. MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease. *Cell Syst.* **2019**, *8*, 380-394.e4.
28. Breiman, L. Random Forests. *Mach Learn* **2001**, *45*, 5–32.
29. Newman, A.M.; Liu, C.L.; Green, M.R.; Gentles, A.J.; Feng, W.; Xu, Y.; Hoang, C.D.; Diehn, M.; Alizadeh, A.A. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **2015**, *12*, 453–457.
30. Becht, E.; Giraldo, N.A.; Lacroix, L.; Buttard, B.; Elarouci, N.; Petitprez, F.; Selves, J.; Laurent-Puig, P.; Sautès-Fridman, C.; Fridman, W.H.; et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **2016**, *17*, 218.
31. Sturm, G.; Finotello, F.; Petitprez, F.; Zhang, J.D.; Baumbach, J.; Fridman, W.H.; List, M.; Aneichyk, T. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **2019**, *35*, i436–i445.
32. Ding, H.; Douglass, E.F.; Sonabend, A.M.; Mela, A.; Bose, S.; Gonzalez, C.; Canoll, P.D.; Sims, P.A.; Alvarez, M.J.; Califano, A. Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat. Commun.* **2018**, *9*, 1471.
33. Gosline, S.J.C.; Weinberg, H.; Knight, P.; Yu, T.; Guo, X.; Prasad, N.; Jones, A.; Shrestha, S.; Boone, B.; Levy, S.E.; et al. A high-throughput molecular data resource for cutaneous neurofibromas. *Sci. Data* **2017**, *4*, 170045.
34. Ijaz, H.; Koptyra, M.; Gaonkar, K.S.; Rokita, J.L.; Baubet, V.P.; Tauhid, L.; Zhu, Y.; Brown, M.; Lopez, G.; Zhang, B.; et al. Pediatric high-grade glioma resources from the Children's Brain Tumor Tissue Consortium. *Neuro-Oncol.*
35. Pollard, K.; Banerjee, J.; Doan, X.; Wang, J.; Guo, X.; Allaway, R.; Langmead, S.; Slobogean, B.; Meyer, C.F.; Loeb, D.M.; et al. A clinically and genomically annotated nerve sheath tumor biospecimen repository. *bioRxiv* **2019**, 2019.12.19.871897.
36. Laks, D.R.; Oses-Prieto, J.A.; Alvarado, A.G.; Nakashima, J.; Chand, S.; Azzam, D.B.; Gholkar, A.A.; Sperry, J.; Ludwig, K.; Condro, M.C.; et al. A molecular cascade modulates MAP1B and confers resistance to mTOR inhibition in human glioblastoma. *Neuro-Oncol.* **2018**, *20*, 764–775.
37. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Dalla Favera, R.; Califano, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **2006**, *7 Suppl 1*, S7.

38. Sevilla, T.; Sivera, R.; Martínez-Rubio, D.; Lupo, V.; Chumillas, M.J.; Calpena, E.; Dopazo, J.; Vílchez, J.J.; Palau, F.; Espinós, C. The EGR2 gene is involved in axonal Charcot−Marie−Tooth disease. *Eur. J. Neurol.* **2015**, *22*, 1548–1555.
39. Warner, L.E.; Mancias, P.; Butler, I.J.; McDonald, C.M.; Keppen, L.; Koob, K.G.; Lupski, J.R. Mutations in the early growth response 2 (EGR2) gene are associated with hereditary myelinopathies. *Nat. Genet.* **1998**, *18*, 382–384.
40. Roos, K.L.; Pascuzzi, R.M.; Dunn, D.W. Neurofibromatosis, Charcot-Marie-Tooth disease, or both? *Neurofibromatosis* **1989**, *2*, 238–243.
41. Lancaster, E.; Elman, L.B.; Scherer, S.S. A patient with Neurofibromatosis type 1 and Charcot-Marie-Tooth Disease type 1B. *Muscle Nerve* **2010**, *41*, 555–558.
42. Lupski, J.R.; Pentao, L.; Williams, L.L.; Patel, P.I. Stable inheritance of the CMT1A DNA duplication in two patients with CMT1 and NF1. *Am. J. Med. Genet.* **1993**, *45*, 92–96.
43. Lin, H.-P.; Oksuz, I.; Svaren, J.; Awatramani, R. Egr2-dependent microRNA-138 is dispensable for peripheral nerve myelination. *Sci. Rep.* **2018**, *8*.
44. Wu, L.M.N.; Deng, Y.; Wang, J.; Zhao, C.; Wang, J.; Rao, R.; Xu, L.; Zhou, W.; Choi, K.; Rizvi, T.A.; et al. Programming of Schwann Cells by Lats1/2-TAZ/YAP Signaling Drives Malignant Peripheral Nerve Sheath Tumorigenesis. *Cancer Cell* **2018**, *33*, 292-308.e7.
45. Veremeyko, T.; Yung, A.W.Y.; Anthony, D.C.; Strekalova, T.; Ponomarev, E.D. Early Growth Response Gene-2 Is Essential for M1 and M2 Macrophage Activation and Plasticity by Modulation of the Transcription Factor CEBPβ. *Front. Immunol.* **2018**, *9*, 2515.
46. Okamura, T.; Yamamoto, K.; Fujio, K. Early Growth Response Gene 2-Expressing CD4+LAG3+ Regulatory T Cells: The Therapeutic Potential for Treating Autoimmune Diseases. *Front. Immunol.* **2018**, *9*.
47. Hamdan, F.H.; Johnsen, S.A. DeltaNp63-dependent super enhancers define molecular identity in pancreatic cancer by an interconnected transcription factor network. *Proc. Natl. Acad. Sci.* **2018**, *115*, E12343–E12352.
48. Cancino, G.I.; Yiu, A.P.; Fatt, M.P.; Dugani, C.B.; Flores, E.R.; Frankland, P.W.; Josselyn, S.A.; Miller, F.D.; Kaplan, D.R. p63 Regulates Adult Neural Precursor and Newly Born Neuron Survival to Control Hippocampal-Dependent Behavior. *J. Neurosci.* **2013**, *33*, 12569–12585.
49. Packard, A.; Schnittke, N.; Romano, R.-A.; Sinha, S.; Schwob, J.E. ΔNp63 Regulates Stem Cell Dynamics in the Mammalian Olfactory Epithelium. *J. Neurosci.* **2011**, *31*, 8748–8759.
50. Carr, N.J.; Warren, A.Y. Mast cell numbers in melanocytic naevi and cutaneous neurofibromas. *J. Clin. Pathol.* **1993**, *46*, 86–87.
51. Zhu, Y.; Ghosh, P.; Charnay, P.; Burns, D.K.; Parada, L.F. Neurofibromas in NF1: Schwann cell origin and role of tumor environment. *Science* **2002**, *296*, 920–922.
52. Isaacson, P. Mast cells in benign nerve sheath tumours. *J. Pathol.* **1976**, *119*, 193–196.
53. Tucker, T.; Riccardi, V.M.; Sutcliffe, M.; Vielkind, J.; Wechsler, J.; Wolkenstein, P.; Friedman, J.M. Different patterns of mast cells distinguish diffuse from encapsulated neurofibromas in patients with neurofibromatosis 1. *J. Histochem. Cytochem. Off. J. Histochem. Soc.* **2011**, *59*, 584–590.
54. Greggio, H. Les cellules granuleuses (Mastzellen) dans les tissus normaux et dans certaines maladies chirurgicales. *Arch. Médecine Expérimentale Anat. Pathol.* **1911**, *23*, 323–375.
55. Chen, Z.; Mo, J.; Brosseau, J.-P.; Shipman, T.; Wang, Y.; Liao, C.-P.; Cooper, J.M.; Allaway, R.J.; Gosline, S.J.C.; Guinney, J.; et al. Spatiotemporal Loss of NF1 in Schwann Cell Lineage Leads to Different Types of Cutaneous Neurofibroma Susceptible to Modification by the Hippo Pathway. *Cancer Discov.* **2019**, *9*, 114–129.

56. Nürnberger, M.; Moll, I. Semiquantitative aspects of mast cells in normal skin and in neurofibromas of neurofibromatosis types 1 and 5. *Dermatol. Basel Switz.* **1994**, *188*, 296–299.
57. Brosseau, J.-P.; Liao, C.-P.; Wang, Y.; Ramani, V.; Vandergriff, T.; Lee, M.; Patel, A.; Ariizumi, K.; Le, L.Q. NF1 heterozygosity fosters de novo tumorigenesis but impairs malignant transformation. *Nat. Commun.* **2018**, *9*, 1–11.
58. Farschtschi, S.; Park, S.-J.; Sawitzki, B.; Oh, S.-J.; Kluwe, L.; Mautner, V.F.; Kurtz, A. Effector T cell subclasses associate with tumor burden in neurofibromatosis type 1 patients. *Cancer Immunol. Immunother.* **2016**, *65*, 1113–1121.
59. Kahen, E.J.; Brohl, A.; Yu, D.; Welch, D.; Cubitt, C.L.; Lee, J.K.; Chen, Y.; Yoder, S.J.; Teer, J.K.; Zhang, Y.O.; et al. Neurofibromin level directs RAS pathway signaling and mediates sensitivity to targeted agents in malignant peripheral nerve sheath tumors. *Oncotarget* **2018**, *9*, 22571–22585.
60. Allaway, R.J.; Fischer, D.A.; de Abreu, F.B.; Gardner, T.B.; Gordon, S.R.; Barth, R.J.; Colacchio, T.A.; Wood, M.; Kacsoh, B.Z.; Bouley, S.J.; et al. Genomic characterization of patient-derived xenograft models established from fine needle aspirate biopsies of a primary pancreatic ductal adenocarcinoma and from patient-matched metastatic sites. *Oncotarget* **2016**, *7*.
61. Wood, M.; Rawe, M.; Johansson, G.; Pang, S.; Soderquist, R.S.; Patel, A.V.; Nelson, S.; Seibel, W.; Ratner, N.; Sanchez, Y. Discovery of a small molecule targeting IRA2 deletion in budding yeast and neurofibromin loss in malignant peripheral nerve sheath tumor cells. *Mol. Cancer Ther.* **2011**, *10*, 1740–50.
62. Danilov, A.V.; Hu, S.; Orr, B.; Godek, K.; Mustachio, L.M.; Sekula, D.; Liu, X.; Kawakami, M.; Johnson, F.M.; Compton, D.A.; et al. Dinaciclib Induces Anaphase Catastrophe in Lung Cancer Cells via Inhibition of Cyclin-Dependent Kinases 1 and 2. *Mol. Cancer Ther.* **2016**, *15*, 2758–2766.
63. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat. Methods* **2017**, *14*, 417–419.
64. Soneson, C.; Love, M.I.; Robinson, M.D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **2016**, *4*, 1521.
65. Carlson, M. *org.Hs.eg.db: Genome wide annotation for Human*; 2019;
66. MultiPLIER Fileset 2019.
67. Collado-Torres, L.; Nellore, A.; Jaffe, A.E.; Taub, M.A.; Kammers, K.; Ellis, S.E.; Hansen, K.D.; Langmead, B.; Leek, J.T. *recount: Explore and download data from the recount project*; Bioconductor version: Release (3.10), 2019;
68. Tan, J.; Huyck, M.; Hu, D.; Zelaya, R.A.; Hogan, D.A.; Greene, C.S. ADAGE signature analysis: differential expression analysis with data-defined gene sets. *BMC Bioinformatics* **2017**, *18*, 512.
69. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 5.
70. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26.
71. Kolde, R. *pheatmap: Pretty Heatmaps*; 2019;
72. Yu, G.; Wang, L.-G.; Han, Y.; He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **2012**, *16*, 284–287.
73. Allaway, R.J.; La Rosa, S.; Guinney, J.; Gosline, S.J.C. Probing the chemical-biological relationship space with the Drug Target Explorer. *J. Cheminformatics* **2018**, *10*, 41.
74. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.; François, R.; Grolemund, G.;

Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the Tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686.
75. Yu, G. *enrichplot: Visualization of Functional Enrichment Result*; 2019;