

DISCOVERING PATHWAY AND CELL TYPE SIGNATURES IN TRANSCRIPTOMIC  
COMPENDIA WITH MACHINE LEARNING

Gregory Philip Way

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

---

Casey S. Greene, Ph.D.  
Associate Professor of Pharmacology

Graduate Group Chairperson

---

Benjamin F. Voight, Ph.D.  
Associate Professor of Pharmacology

Dissertation Committee:

Chair: John M. Maris, M.D., Professor of Pediatrics  
Yoseph Barash, Ph.D., Associate Professor of Genetics  
Nancy R. Zhang, Ph.D., Professor of Statistics  
Josh Stuart, Ph.D., Professor of Biomedical Engineering

DISCOVERING PATHWAY AND CELL TYPE SIGNATURES IN TRANSCRIPTOMIC  
COMPENDIA WITH MACHINE LEARNING

COPYRIGHT

2019

Gregory Philip Way

This work is licensed under the Creative Commons Attribution-NonCommercial-  
ShareAlike 3.0 License

To view a copy of this license, visit

<https://creativecommons.org/licenses/by-nc-sa/3.0/us/>



*To Jessica, John, Lourdes, Ruth and Mercedes*

## ACKNOWLEDGMENTS

I would like to first thank Casey Greene for being a thoughtful and inspiring mentor and leader. I would also like to thank my thesis committee: John Maris, Yoseph Barash, Nancy Zhang, and Josh Stuart for providing invaluable guidance and insight as my projects were coming together. Thank you to the GCB administration of Li San Wang, Ben Voight, Hanna Chervitz, and Maureen Kirsch for their leadership, support, and for making my transfer to Penn from Dartmouth as smooth as possible. Thank you also to the 2014 Dartmouth QBS administrators Jason Moore, Anna Greene, and Krissy Giffin for helping me to navigate a PhD program. I would also like to thank my Penn GCB student cohort of Zerry Zhou, Di Zhang, Apexa Modi, Samantha Klasfeld, and Onur Yoruk. Thank you to my Dartmouth QBS cohort of Jeff Thompson, Chris Rees, Elle Nutter, Jen Franks, Sara Lundgren, and Mavra Nasir as well. Both groups provided incredible support and friendship. I would also like to extend a special thanks to Maya Bucan and Junhyong Kim for excellent training opportunities and insightful discussions.

The Greenelab was an amazing environment to grow as a scientist and to learn computational biology and data science. I would like to specifically thank Daniel Himmelstein and alumni Jaclyn Taroni, Jie Tan, and Brett Beaulieu-Jones for consistent and amazing feedback and for their great scientific intuition. Many others in the Greenelab including YoSon Park, Qiwen Hu, David Nicholson, Alex Lee, Michael Zietz, Amy Campbell, Tim Chang, Kathy Chen, Chris Williams, and James Rudd were instrumental in fostering a scientifically rigorous yet comfortable environment. I would also like to thank the software engineers in the Greenelab including Dongbo Hu, Matt Huyck, Rene Zelaya, Kurt Wheeler, and Deepa Prasad for their amazing support and expertise.

I am also extremely grateful to all of my collaborators that I have had the pleasure of working with and learning from over the past several years. I would like to specifically thank Jen Doherty, Yolanda Sanchez, Brock Christensen, Struan Grant, Elana Fertig, and Jo Lynne Rokita for their scientific insight and productive collaborations. I would also like to thank members of The Cancer Genome Atlas Research Network including Chen Wang and Nikki Schultz for their great organizational skills, inspiration, and expertise.

Lastly, I would like to thank my family – past, present, and future. Thank you to my parents, John and Lourdes, for raising me to have a strong work ethic and grit. Thank you to my brother, Stewart, and sister-in-law, Paulina, for all of your support and internet memes. Jessica, thank you for riding this roller coaster with me through each and every thrill, dip, and turn. Also thank you Myshkin for your stress-erasing greetings upon returning home from work every day. I would also like to thank my grandparents Ruth, Mercedes, and Felipe. Without your sacrifices, I would not be standing here today.

## ABSTRACT

### DISCOVERING PATHWAY AND CELL TYPE SIGNATURES IN TRANSCRIPTOMIC COMPENDIA WITH MACHINE LEARNING

Gregory Philip Way

Casey S. Greene

*Gene expression measurements capture downstream biological responses to molecular perturbations. This systems biology perspective can be investigated using both supervised and unsupervised machine learning approaches to rapidly derive insight, including cell type and pathway signatures, from transcriptomic compendia. Machine learning applied to transcriptomic compendia can aid in biological discovery, hypothesis generation, and precision medicine. We introduce these topics and discuss their impact in Chapter 1. In Chapters 2-4, we describe and extend a supervised learning approach to detect aberrant gene and pathway activity in cancer. We apply this approach to identify patient tumors, cell lines, and patient derived xenograft models with TP53 loss of function, Ras signaling activation, and NF1 loss. This approach facilitates the discovery of phenocopying variants and potential hidden responders to specific therapies. In Chapters 5-6, we focus on deriving transcriptomic signatures using unsupervised learning. We show that unsupervised learning can identify disease subtypes and can be used to develop gene expression signatures without the need to specify labels a priori. In Chapter 5, we assess the reproducibility of high grade serous ovarian cancer (HGSC) gene expression subtypes across populations and clustering algorithms. In Chapter 6, we train a variational autoencoder on patient tumors and use latent space arithmetic to identify gene signatures most distinguishing HGSC subtypes. Lastly, in Chapter 7, we develop an approach to rapidly interpret compressed features*

*engineered in unsupervised learning algorithms. We train a series of unsupervised models across a wide range of latent space dimensions and develop a network-based method for interpreting these compressed gene expression features. Using this approach, we observe that modifying the hidden layer dimensionality impacts the identification of specific geneset and cell-type activation patterns in cancer and normal tissue. Machine learning models scale to large genomic datasets and have provided state of the art results in a variety of biomedical domains. However, model interpretation is critical to build knowledge and to generate hypotheses.*

## TABLE OF CONTENTS

<b>ACKNOWLEDGMENTS</b> .....	<b>iv</b>
<b>ABSTRACT</b> .....	<b>vi</b>
<b>LIST OF TABLES</b> .....	<b>xii</b>
<b>LIST OF ILLUSTRATIONS</b> .....	<b>xiii</b>
<b>Chapter 1.</b> .....	<b>1</b>
<b>An Introduction to discovering pathway and cell type signatures in transcriptomic compendia with machine learning</b> .....	<b>1</b>
<b>1.1. Introduction</b> .....	<b>1</b>
<b>1.2. Supervised learning to isolate expression signatures</b> .....	<b>4</b>
1.2.1. <i>A brief overview of supervised machine learning methodology</i> .....	5
1.2.2. <i>Initial successes of supervised machine learning applied to transcriptome data</i> .....	5
1.2.3. <i>Supervised machine learning to derive cell type and pathway signatures</i> .....	6
<b>1.3. Unsupervised learning to discover hidden expression states</b> .....	<b>10</b>
1.3.1. <i>A brief overview of unsupervised machine learning algorithms</i> .....	11
1.3.2. <i>Unsupervised machine learning to uncover cell types</i> .....	12
1.3.3. <i>Unsupervised machine learning reveals underlying gene expression states</i> .....	14
<b>1.4. Interpreting machine learning models applied to transcriptomes</b> .....	<b>16</b>
1.4.1. <i>Supervised learning models reveal differences between sample statuses</i> .....	16
1.4.2. <i>Unsupervised learning models require interpretation of compressed features</i> .....	17
<b>1.5. Conclusion</b> .....	<b>21</b>
<b>1.6. Acknowledgements</b> .....	<b>21</b>
<b>Chapter 2.</b> .....	<b>22</b>
<b>A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma</b> .....	<b>22</b>
<b>2.1. Abstract</b> .....	<b>22</b>
2.1.1. <i>Background</i> .....	22
2.1.2. <i>Results</i> .....	23
2.1.3. <i>Conclusions</i> .....	23
<b>2.2. Background</b> .....	<b>24</b>
<b>2.3. Methods</b> .....	<b>25</b>
2.3.1. <i>The Cancer Genome Atlas data used for building the classifier</i> .....	25
2.3.2. <i>Hyperparameter optimization of the logistic regression classifier</i> .....	26
2.3.3. <i>Ensemble classifier construction and application to the validation set</i> .....	27
2.3.4. <i>Effect sizes and power analysis</i> .....	28
2.3.5. <i>Validation sample acquisition</i> .....	28
2.3.6. <i>Cell culture</i> .....	29
2.3.7. <i>RNA microarray</i> .....	30
2.3.8. <i>Validation sample processing</i> .....	30

2.3.9. Western blotting .....	31
2.3.10. Reproducibility of computational analyses .....	32
<b>2.4. Results .....</b>	<b>33</b>
2.4.1. Classifier performance .....	33
2.4.2. Identification and characterization of NF1 deficient glioblastoma tumor samples.....	33
2.4.3. Highly contributing genes .....	36
<b>2.5. Discussion .....</b>	<b>38</b>
<b>2.6. Conclusions .....</b>	<b>41</b>
<b>2.7. Acknowledgements .....</b>	<b>41</b>
<b>Chapter 3. ....</b>	<b>42</b>
<b>Machine learning detects pan-cancer Ras pathway activation in The Cancer Genome Atlas .....</b>	<b>42</b>
3.1. Summary .....	42
3.2. Introduction.....	43
3.3. Results .....	45
3.3.1. Machine learning models to predict pathway activity .....	45
3.3.2. Detecting Ras activation pan cancer .....	46
3.3.3. Ras classifier benchmarking analyses .....	50
3.3.4. Detecting Ras activation in cell lines .....	52
3.3.5. Other Ras pathway variants phenocopy Ras activation .....	55
3.4. Discussion .....	57
3.5. Methods .....	60
3.5.1. Contact for reagent and resource sharing .....	60
3.5.2. Training machine learning classifiers to detect aberrant gene events.....	61
3.5.3. Evaluating machine learning classifiers .....	63
3.5.4. Classifier benchmarking analyses.....	64
3.5.5. Differential expression analysis .....	65
3.5.6. Cell line validation.....	65
3.5.7. Ras pathway and oncogenicity curation.....	66
3.5.8. Quantification and statistical analyses.....	66
3.5.9. Data and software availability .....	67
3.6. Acknowledgements .....	67
<b>Chapter 4. ....</b>	<b>69</b>
<b>Machine learning derived expression signature predicts TP53 inactivation .....</b>	<b>69</b>
4.1. Introduction.....	69
4.2. Results .....	70
4.3. Methods .....	74
4.3.1. In-silico prediction of TP53 inactivation .....	74
4.4. Conclusions .....	76
<b>Chapter 5. ....</b>	<b>78</b>

<b>Comprehensive cross-population analysis of high-grade serous ovarian cancer supports no more than three subtypes .....</b>	<b>78</b>
<b>5.1. Abstract .....</b>	<b>78</b>
<b>5.2. Introduction.....</b>	<b>79</b>
<b>5.3. Methods .....</b>	<b>81</b>
5.3.1. <i>Data inclusion.....</i>	81
5.3.2. <i>Clustering.....</i>	82
5.3.3. <i>Identification of analogous clusters within and across studies .....</i>	83
5.3.4. <i>Clustering analysis of randomized data .....</i>	83
5.3.5. <i>Assessing the reproducibility of single population studies .....</i>	83
5.3.6. <i>Data availability.....</i>	84
<b>5.4. Results .....</b>	<b>84</b>
5.4.1. <i>Clustering.....</i>	84
5.4.2. <i>Correlation of cluster-specific expression patterns.....</i>	84
5.4.3. <i>Comparison with previously-identified HGSC clusters.....</i>	88
5.4.4. <i>Meta-research into previous HGSC subtyping studies .....</i>	90
<b>5.5. Discussion .....</b>	<b>91</b>
<b>5.6. Acknowledgements .....</b>	<b>93</b>
<b>Chapter 6. ....</b>	<b>95</b>
<b>Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders.....</b>	<b>95</b>
<b>6.1. Abstract .....</b>	<b>95</b>
<b>6.2. Introduction.....</b>	<b>96</b>
<b>6.3. Methods .....</b>	<b>98</b>
6.3.1. <i>Model summary .....</i>	98
6.3.2. <i>Model implementation.....</i>	98
6.3.3. <i>Parameter selection .....</i>	100
6.3.4. <i>Input data.....</i>	101
6.3.5. <i>Interpretation of gene weights .....</i>	101
6.3.6. <i>The latent space of ovarian cancer subtypes.....</i>	102
6.3.7. <i>Enabling exploration through visualization .....</i>	102
6.3.8. <i>Reproducibility.....</i>	103
<b>6.4. Results .....</b>	<b>103</b>
6.4.1. <i>Tumors were encoded in a lower dimensional space .....</i>	103
6.4.2. <i>Features represent biological signal .....</i>	104
6.4.3. <i>Interpolating the lower dimensional manifold of HGSC subtypes .....</i>	106
<b>6.5. Conclusions .....</b>	<b>109</b>
<b>6.6. Acknowledgements .....</b>	<b>110</b>
<b>Chapter 7. ....</b>	<b>111</b>
<b>Sequential compression across latent space dimensions enhances gene expression signatures.....</b>	<b>111</b>
<b>7.1. Abstract .....</b>	<b>111</b>
7.1.1. <i>Background .....</i>	111



7.1.2. Results .....	111
7.1.3. Conclusions .....	112
<b>7.2. Introduction.....</b>	<b>112</b>
<b>7.3. Results .....</b>	<b>114</b>
7.3.1. BioBombe implementation .....	114
7.3.2. Assessing compression algorithm reconstruction.....	114
7.3.3. Evaluating model stability and similarity within and across latent dimensions .....	114
7.3.4. Sequential compression can optimize gene expression signature discovery .....	121
7.3.5. Assessing gene set coverage of compressed features .....	122
7.3.6. Assessing sample type correlation differences across latent dimensions ....	125
7.3.7. Interpretation of GTEx blood with VAE compression features.....	129
7.3.8. Validating GTEx neutrophil and monocyte signatures in external datasets..	130
7.3.9. Using compressed features in supervised learning applications.....	132
<b>7.4. Discussion .....</b>	<b>133</b>
<b>7.5. Conclusions .....</b>	<b>139</b>
<b>7.6. Methods .....</b>	<b>140</b>
7.6.1. Transcriptomic compendia acquisition and processing .....	140
7.6.2. Training unsupervised neural networks.....	141
7.6.3. Optimizing training hyperparameters in neural network architectures.....	142
7.6.4. Training compression algorithms with sequential latent dimensions .....	142
7.6.5. Evaluating compression algorithm performance .....	143
7.6.7. Using BioBombe as a signature discovery tool.....	144
7.6.8. Gene network construction and processing .....	145
7.6.9. Rapid interpretation of compressed gene expression data .....	146
7.6.10. Calculating gene set coverage of sequentially compressed gene expression data .....	147
7.6.11. Downloading and processing publicly available expression data for neutrophil GTEx analysis.....	148
7.6.12. Downloading and processing publicly available expression data for monocyte GTEx analysis.....	149
7.6.13. Machine learning classification of cancer types and gene alterations in TCGA.....	150
<b>7.7. Acknowledgements .....</b>	<b>152</b>
<b>Chapter 8. ....</b>	<b>153</b>
<b>Conclusions .....</b>	<b>153</b>
<b>BIBLIOGRAPHY .....</b>	<b>156</b>

## LIST OF TABLES

Table 5.1: Characteristics of the populations included in the five HGSC data sets.....	82
Table 5.2: SAM moderated t score vector Pearson correlations between analogous clusters across populations.....	87
Table 5.3: Distributions of sample membership in the clusters identified in our study compared to the original cluster assignments in the TCGA, Tothill, and Konecny studies.....	89
Table 6.1: Summary of significantly overrepresented pathways separating HGSC subtypes identified by latent space arithmetic with VAE features.....	108

## LIST OF ILLUSTRATIONS

Figure 1.1: RNA sequencing (RNA-seq) provides a systems biology perspective.....	2
Figure 1.2: Supervised machine learning to derive cell type and pathway signature.....	7
Figure 1.3: Unsupervised machine learning is used to discover cell type proportion and pathway signatures.....	14
Figure 1.4: Interpretation of compressed gene expression features.....	19
Figure 2.1: Ensemble classifier errors over 100 iterations for TCGA GBM RNAseq.....	34
Figure 2.2: Performance of our classifier on an external validation set.....	35
Figure 2.3: Genes that contribute to the NF1 classifier performance.....	37
Figure 3.1: Supervised machine learning and data integration for TCGA PanCanAtlas.....	46
Figure 3.2: Ras pathway alteration percentages in TCGA PanCanAtlas.....	47
Figure 3.3: Evaluating machine learning classification of Ras activation.....	49
Figure 3.4: Benchmarking PanCanAtlas Ras classifiers.....	51
Figure 3.5: Cell line predictions of Ras activity by PanCanAtlas Ras classifier.....	54
Figure 3.6: Ras activation across Ras variants and alternative Ras pathway members.....	55
Figure 3.7: TCGA PanCanAtlas NF1 classification performance.....	56
Figure 3.8: Predicting BRAF status with the TCGA PanCanAtlas Ras classifier.....	58
Figure 4.1: Machine learning to predict TP53-inactivating mutations in cancer.....	71
Figure 4.2: Pan-Cancer TP53 classifier scores by cancer-type.....	72
Figure 4.3: TP53 exon-exon junctions for samples with c.375G>T mutations in TP53..	74
Figure 5.1: Sample by sample Pearson correlation matrices across HGSC Populations.....	85
Figure 5.2: NMF consensus matrices for HGSC datasets when $k = 2$ , $k = 3$ , and $k = 4$ .....	86
Figure 5.3: Pearson correlation heatmaps reveal consistency across HGSC datasets.....	86
Figure 5.4: Pearson correlation heatmaps of randomly shuffled HGSC datasets.....	87
Figure 5.5: Pearson correlations comparing $k$ means and NMF clustering HGSC subtypes.....	88
Figure 5.6: Comparing NMF consensus clustering in the Tothill dataset.....	91
Figure 6.1: A variational autoencoder (VAE) applied to gene expression data.....	99
Figure 6.2: Samples encoded by a variational autoencoder retain biological signals.....	104
Figure 6.3: Specific examples of Tybalt features capturing biological signals.....	106
Figure 6.4: Largest mean differences in HGSC subtype vector subtraction for each subtype.....	108
Figure 7.1: Representing our BioBombe implementation workflow.....	115
Figure 7.2: Overview of the BioBombe approach.....	116
Figure 7.3: Reconstruction cost across datasets, algorithms and dimensions.....	116
Figure 7.4: Assessing algorithm and dimension stability with singular vector canonical correlation analysis (SVCCA).....	118
Figure 7.5: Across algorithm stability as measured by singular vector canonical correlation analysis (SVCCA).....	119
Figure 7.6: Across latent dimension stability as measured by singular vector canonical correlation analysis (SVCCA).....	120
Figure 7.7: Using BioBombe as a signature discovery tool.....	122

Figure 7.8: Assessing gene set coverage of specific gene set collections.....	124
Figure 7.9: Absolute ranking of the top gene set BioBombe z scores across algorithms.....	125
Figure 7.10: Tracking the dimensions of highest BioBombe enrichment signal.....	126
Figure 7.11: Tracking sample correlation across latent dimensions.....	127
Figure 7.12: Pearson correlation between input and reconstructed samples in real and permuted data.....	128
Figure 7.13: Interpreting compressed features learned from GTEx using xCell gene sets.....	131
Figure 7.14: Using compressed features as features in supervised machine learning.....	134

## Chapter 1.

### **An Introduction to discovering pathway and cell type signatures in transcriptomic compendia with machine learning**

This chapter was adapted from: Way, Gregory, P. and Greene, Casey, S.

*“Discovering pathway and cell type signatures in transcriptomic compendia with machine learning.”* To appear in Annual Review of Biomedical Data Science, 2019. Preprint:

<https://peerj.com/preprints/27229/>

#### **1.1. Introduction**

The quantity of biological data and the pace of their generation have increased dramatically over the past several years (1). Biological data are also increasing in complexity, as multiple genomic modalities are being measured with improving resolution. One such modality measures the transcriptome—the complete RNA products of about 30,000 genes in a given organism, tissue, or cell. From the relatively low sample sizes and early days of microarray technology to the large data sets currently generated through RNA sequencing (RNA-seq) today, researchers have used transcriptome measurements to interrogate various biological hypotheses (2). RNA measurements can be used to investigate changes to specific expression patterns of single genes or pathways. RNA measurements can also be examined from a systems biology perspective, in which entire biological systems are studied rather than individual parts. From this perspective, the transcriptome represents downstream molecular consequences of perturbation or disease and captures alterations to gene regulatory networks and environmental stimuli (3). In this dissertation, we consider the systems biology perspective that transcriptome measurements provide (Figure 1.1).

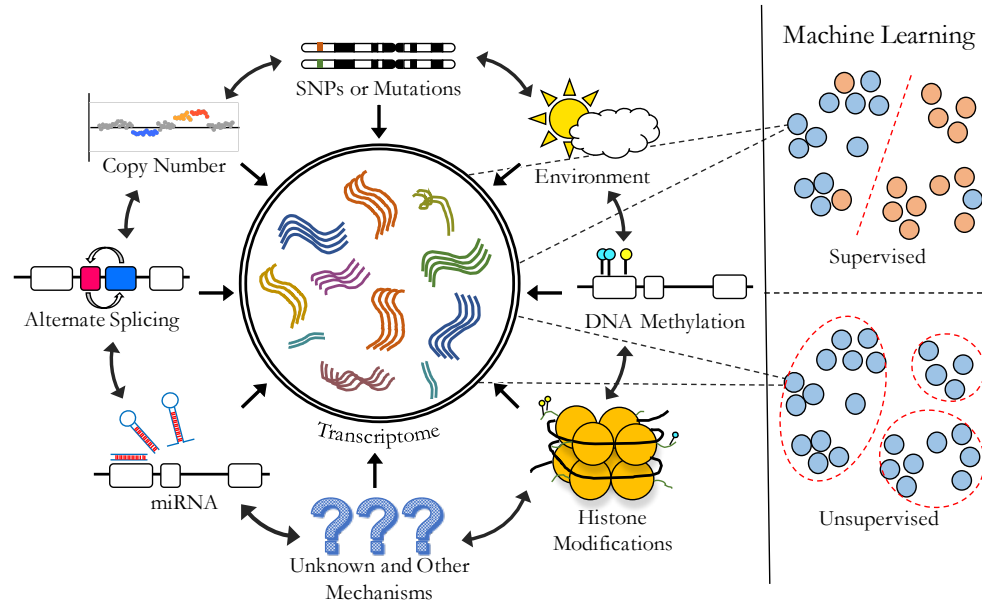


Figure 1.1: RNA sequencing (RNA-seq) provides a systems biology perspective

The downstream response to various molecular and environmental perturbations can be captured as signals in RNA-seq data. Supervised and unsupervised machine learning applied to RNA-seq interrogates this property to reveal expression signatures of cell type and pathway activity.

A significant challenge to transcriptome analyses is making sense of the high-dimensional data. After data processing, there are many mechanisms by which hypotheses can be tested and generated (4). One strategy uses machine learning, which is capable of rapidly deriving insights and providing accurate results. Machine learning is a branch of computer science used to derive solutions based on high-dimensional input data and a target goal. By optimizing the target goal, or objective function, the computer automatically learns a specific, and potentially insightful, solution. There are many different machine learning algorithms, each with different costs and benefits, including logistic regression, support vector machines (SVMs), random forests (RFs), neural networks (NNs), principal components analysis (PCA), non-negative matrix factorization (NMF), k-means clustering, and many more. Within each algorithm exists a

series of specific tunable knobs called hyperparameters. These knobs control how fast an algorithm learns, how many features are learned, how many times to cycle through data, and many other important considerations. Hyperparameter decisions can be configured through cross validation (CV) in a dataset specific fashion. CV optimizes performance by training on one portion of the data, evaluating performance on the remaining set, and alternating which portion of the data is removed from training. A common challenge in training these models is that the model performs well in training but fails to generalize to new data. To mitigate this problem, termed overfitting, researchers withhold a portion of the data from training and evaluate it later.

There are two basic classes of machine learning: supervised and unsupervised learning. Each class can be used with varying goals, but the fundamental purpose of each is the same: to test how well the model captures the underlying target biology and to determine if the biology is consistent when the model is applied to new data. While there are other classes of machine learning, such as semisupervised learning, reinforcement learning, distantly and weakly supervised learning, and others (5), we focus here on supervised and unsupervised learning. We apply supervised learning approaches in Chapters 2, 3, and 4, and discuss unsupervised learning projects in Chapters 5 and 6. Lastly, we discuss mechanisms to explore the dimensionality of latent spaces and interpret unsupervised compression models in Chapter 7.

Early efforts applying supervised machine learning to transcriptome data were largely successful. However, the approaches involved relatively simple supervised classification tasks such as cancer versus normal detection (6, 7), outcome prediction (8), or gene module detection (9, 10). Additionally, unsupervised tasks like cancer subtype discovery (11) and gene pattern identification (12) were also applied in early research. These pioneering studies included relatively few samples, and the target

biology resulted in large sources of variation. Larger data sets have allowed investigators to test more specific hypotheses and extract more subtle expression patterns. Many current machine learning algorithms applied to transcriptome data involve more subtle tasks, including the detection and characterization of pathway- and cell type–based signatures that exist in an underlying subspace of the observable data.

The extraction of pathway– and cell type–specific gene expression signatures can reveal the function and heterogeneity of transcriptome data, and these signatures are often the result of molecular perturbations that may be important to a disease or phenotype of interest (13–16). Machine learning methods can extract biological signals (17). In this introduction, we highlight specific machine learning techniques applied to transcriptomic compendia to reveal underlying patterns representing cell type and pathway signatures. We discuss supervised and unsupervised machine learning for tasks including cell type deconvolution, expression signature discovery for the prediction of pathway activity, and the use of dimensionality reduction, or compression, to uncover and explain hidden cellular states. We also discuss recent machine learning approaches to extract pathway activity in single-cell data and recent deep learning algorithm advancements. Lastly, we focus on specific challenges associated with interpreting machine learning models.

### ***1.2. Supervised learning to isolate expression signatures***

Supervised machine learning applied to transcriptome data is a powerful approach to test hypotheses about a given model system and to make predictions based on target biology. Leveraging the ability of the transcriptome to capture the differential mechanisms underlying biological states (see Figure 1.1), supervised machine learning can stratify samples and states that are based on specific cell type or pathway signatures. In the following subsection, we (a) broadly introduce supervised learning



methodology, (b) briefly discuss initial landmark studies applying supervised machine learning to transcriptome data, and (c) conclude with a review of current studies that train supervised models on large transcriptomic compendia to derive pathway and cell type signatures.

#### *1.2.1. A brief overview of supervised machine learning methodology*

The goal of supervised machine learning is to train a computer to determine the status of a known sample and to make accurate predictions on a new sample (18). Generally, the models receive as input an  $n \times p$  data matrix  $X$  and a vector  $y$  of length  $n$ . Here,  $n$  is the number of samples,  $p$  is the number of features, and  $y$  represents the predefined status, or target classes. In many supervised learning algorithms, the models reach a solution of weights  $w$  that are optimized against the classification or regression task, often through an iterative learning process, such as stochastic gradient descent. Additionally, various algorithms place different emphasis on the training process and restricting, or regularizing, the solution of weights. For example, one common algorithm is logistic regression, which can add penalty terms like Lasso or elastic net into the objective function, which will enforce sparse solutions (19, 20). SVMs maximize the distance between class labels in feature space, and RFs will determine over many iterations features used to split samples based on information content (21, 22). There have been many applications of supervised machine learning across a variety of domains. Here, we focus on supervised learning applied to deriving cell type and pathway signatures. In Chapters 2, 3, and 4, we apply supervised learning approaches to detect aberrant pathway activity in cancer.

#### *1.2.2. Initial successes of supervised machine learning applied to transcriptome data*

Various supervised learning algorithms have been applied to transcriptome data for nearly two decades (23). In this setting, the input matrix  $X$  is typically  $n$  samples by  $p$

gene expression features, and the vector  $y$  is defined by a target hypothesis or measured value. When it is important that only a few genes explain the target hypothesis, a researcher may prefer models that are constrained to provide sparse solutions, whereby only a small percentage of measurable genes contribute to performance. Sparsity may be helpful to define biomarker panels for downstream analyses. For example, a sparse classifier predicted metastases in breast cancer (24). This discovery led to the 70-gene Mammaprint panel, demonstrating that only 70 genes need to be measured to predict breast cancer severity. However, careful validation of prognostic signatures must be performed, as over 90% of gene signatures with 100 random genes were associated with breast cancer outcomes (25). Additional pioneering applications of supervised learning to gene expression data have identified top genes that differentiate acute lymphoblastic leukemia from acute myeloid leukemia (7), distinguished tumor from normal biopsies (6), predicted treatment response in lymphoma (8), and predicted the function of novel yeast open reading frames (9). These studies were performed on microarray data and were limited to small sample sizes. Therefore, the target goals of these approaches required that the two classes contain large differences in signal. While these studies did not directly interrogate hypotheses relating to cell type and pathway activity, the signals identified may have represented differential cell type or pathway expression. Current applications train machine learning models on data sets that are orders of magnitude larger, and can thus detect more subtle signatures hidden in the data.

### *1.2.3. Supervised machine learning to derive cell type and pathway signatures*

Applying supervised machine learning to large transcriptomic compendia allows researchers to test specific hypotheses about cell type and pathway signatures (Figure 1.2). For example, many cell type deconvolution methods perform supervised learning to

estimate cell type proportions in samples from bulk tissue expression. In a supervised setting, deconvolution uses regression and borrows information from sets of predefined marker genes or proportion estimates associated with specific cell types. One method, CIBERSORT, requires an input signature matrix of immune cell marker genes that, through support vector regression (SVR), deconvolves an input gene expression matrix

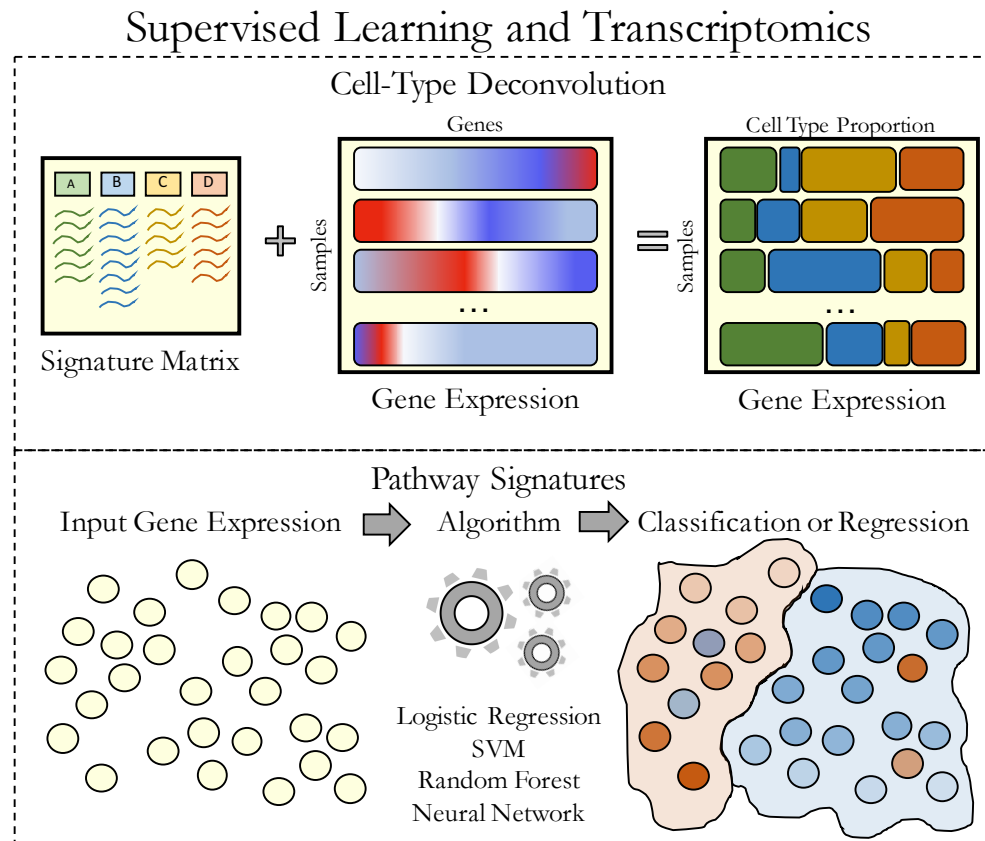


Figure 1.2: Supervised machine learning to derive cell type and pathway signatures

(top) Supervised cell type deconvolution methods require a signature matrix as input that has predefined marker genes or proportion estimates of cell types. Some form of linear regression incorporates this information to generate estimates of cell type proportion. (bottom) Supervised learning applied to large transcriptomic compendia with a targeted hypothesis can stratify samples based on pathway activity. The models can be used in classification or regression to provide binary labels or continuous activation estimates, respectively.

from bulk tissue (26). Similar approaches use linear regression based on other predefined cell type signature matrices to deconvolve immune cell types. This approach has been applied to bulk cancer and systemic lupus erythematosus gene expression data (27, 28). Other deconvolution algorithms implement least squares regression with input proportion matrices predefined in various ways. For example, the matrices can be defined by cell type–specific probes (29), by using purified reference samples (30), or from a pathologist’s estimation (31). In a related study with different goals, an in-silico dissection approach trained an SVM on bona fide cell type–specific genes to identify other genes in a guilt-by-association analysis (32). Other cell type deconvolution methods exist (reviewed in Reference 33), and many are based on unsupervised learning to reveal underlying patterns (discussed in Section 1.3).

Another use case for supervised learning stratifies samples based on pathway activity (Figure 1.2). A key step in this process is to assign accurate labels to samples that exhibit pathway misregulation. Assigning the correct status to a sample is costly, difficult, and often inaccurate. Therefore, this assignment is usually determined through orthogonal means (e.g., pathway mutation status in cancer). Despite this challenge, many studies have revealed interesting insights. For example, Guinney et al. (34) trained an elastic net classifier on colon cancer transcriptomes to detect *KRAS*-mutated tumors resistant to EGFR (epidermal growth factor receptor) inhibition therapy. The model generalized to unseen data sets, and misregulation was associated with survival and response to MEK inhibition. In other words, the model identified a subspace that separated wild-type *KRAS* samples from *KRAS*-mutant samples, which was validated in an external cell line data set. In Chapter 2, we discuss a similar approach applied to detecting NF1 loss of function in glioblastoma patients. We found that this model generalized to a series of patient-derived xenograft models (35). In this study, because

there was a relatively low number of positive examples, an ensemble logistic regression model was implemented. An ensemble machine learning model trains several classifiers on a single task and can help assess solution stability (36). In Chapter 3, we introduce a machine learning Ras classifier based on logistic regression with an elastic net penalty. We trained the model using data from The Cancer Genome Atlas (TCGA) PanCanAtlas project (37). The model predicted Ras activation across a variety of cancer types, including colon cancer, and generalized to alternative data sets and tissues. Additionally, sensitivity to MEK inhibition was strongly correlated with classifier scores in wild-type Ras cell lines. We also discuss a similar model applied to detecting TP53 inactivation in Chapter 4 (38). This model revealed an inactivating silent mutation in the splice donor of *TP53* exon 4, which was corroborated by orthogonal exon–exon splice junction evidence (39).

Other supervised learning algorithms and custom modifications have been applied to detecting pathway activity in transcriptomes. For example, custom SVM variants and boosting methods have been applied to identify mechanisms that increase malignancy in tumors (40). Including biological knowledge a priori in the classification task during training can also aid in feature selection and pathway activity stratification (41). Furthermore, one-class learning regression algorithms train models on gold standard gene expression of specific tissues or pathways, and can generalize to other data sets without knowledge of negative labels (42). This approach was recently applied to predict oncogenic potential, or stemness, in TCGA PanCanAtlas tumors (43). A similar approach, termed positive unlabeled learning, uses gold standard positively labeled genes alone to implicate other disease-associated genes (44). Supervised learning has also been applied to single-cell transcriptome data. For example, supervised learning has been applied to detect marker genes in neocortical cells (45). An NN-based

approach can also be used to predict cellular state and cell type (46). Generative adversarial networks, which train two competing NNs (47), have been trained to simulate single-cell gene expression profiles, which can identify rare cell populations (48, 49). In conclusion, supervised learning can determine specific cell type and pathway activity and can test hypotheses directly. However, sample labels are costly and often inaccurate. It is also important to assess the performance of these models in alternative data sets and to provide orthogonal biological evidence when making conclusions.

### ***1.3. Unsupervised learning to discover hidden expression states***

Unsupervised machine learning identifies underlying structures in data without the need for sample labels (50). The goals of unsupervised learning include clustering samples into similar groups and identifying hidden, or latent, variables present in lower-dimensional subspaces. Applied to gene expression data, unsupervised learning has been used to identify disease subtypes (11), deconvolve cell types (33), and extract underlying gene expression modules present in various percentages in lower-dimensional data representations (51). In the following subsection, we (a) broadly introduce unsupervised learning methodology, (b) discuss the extraction of cell types from expression data in an unsupervised manner, and (c) review a series of recent publications that train dimensionality reduction, or compression, models on large transcriptomic compendia to uncover hidden representations in data that reflect pathway activity. In Chapter 5, we apply unsupervised learning approaches to determine the concordance of high grade ovarian cancer subtypes (HGSC) across populations. In Chapter 6, we train a variational autoencoder (VAE) on gene expression data and perform latent space arithmetic to reveal underlying differences between these HGSC subtypes.

### 1.3.1. *A brief overview of unsupervised machine learning algorithms*

In many unsupervised algorithms, the models learn through minimizing reconstruction cost, in an  $n \times p$  input data matrix  $X$ , where  $n$  and  $p$  are defined as above. The algorithms reconstruct the input matrix after passing the data through one or more intermediate layers and projecting the matrix back onto input feature space. Most often, the intermediate layers have fewer dimensions than the number of input features and are considered bottleneck layers. Additionally, most algorithms use only a single-bottleneck layer. Dimensionality-reduction algorithms such as PCA, independent components analysis (ICA), NMF, and autoencoders are often evaluated by their ability to reconstruct input data. Researchers can add various constraints on the reconstruction loss to help increase feature sparsity or penalize the model to enforce specific feature learning. In each compression algorithm, there are two distinct and valuable matrices extracted that require interpretation. The matrices represent the learned components scores across samples, as well as the relative contribution of each expression feature to each component. In all cases, the researcher must select the bottleneck dimensionality or rely on heuristics.

The application of unsupervised machine learning to growing transcriptomic compendia has facilitated the rapid generation of biological hypotheses. Compression algorithms receive input gene expression from thousands of samples and apply a bottleneck layer to learn the most important sources of variation. These sources are learned in different ways. For example, PCA learns sources of variation that are orthogonal and that explain a decreasing amount of variation in the data. ICA solves a signal processing problem of disentangling sources of independent signals, which are not necessarily orthogonal. NMF, which has widely been used in the deconvolution literature, identifies so-called metagenes, or modules of genes with coordinated

expression patterns (52). NMF is also popular for cell type deconvolution because cell types exist in positive, linear proportions in bulk tissue. NN-based compression algorithms, such as autoencoders and their many variations, also compress data into lower dimensions (53, 54). These methods compress data with a nonlinear activation and can therefore learn subtle, nonlinear patterns in gene expression data given enough samples. Applied to transcriptomic compendia, compression algorithms have provided insights into underlying pathway activity.

Other instances of unsupervised learning algorithms involve clustering, including k-means clustering, Gaussian mixture models, hierarchical clustering, t-distributed stochastic neighbor embedding (t-SNE), and many more (55). These models use distance measures in various ways to group similar samples together for class stratification and class discovery. There are many examples of unsupervised learning applied to cluster gene expression data for subtype identification and gene module detection. For example, Hoadley et al. grouped tens of thousands of tumor samples from TCGA to highlight subtypes found independent of tissue of origin (56). We specifically discuss an application of k-means clustering and non-negative matrix factorization to various HGSC datasets in Chapter 5 (57). However, in this introductory section, we do not focus on clustering applications and instead focus on compression algorithms applied to uncover cell type and pathway signatures.

### *1.3.2. Unsupervised machine learning to uncover cell types*

Unsupervised learning can be used as a powerful approach to extract cell type signatures in transcriptomic compendia (Figure 1.3). Several unsupervised algorithms have been used for cell type deconvolution, including self-organizing maps, hierarchical clustering, and matrix decomposition methods like NMF and singular value decomposition (33, 52, 58). NMF is used to deconvolve gene expression data to identify



differentially expressed genes when no marker genes or reference data exist (59, 60). The NMF core algorithm can be guided to identify cell types by restricting the component matrix columns to sum to one (61). Additionally, a Markov chain Monte Carlo approach has been proposed to estimate cell type proportions in an unsupervised fashion (62). Nearest shrunken centroids, a technique that minimizes the number of genes required to describe subtypes (63), was also used to deconvolve tumors into malignant, nonmalignant, and stroma components (64). It is likely that other compression algorithms, in addition to NMF, also capture cell type associations in their compressed latent spaces. However, proper interpretation of learned gene expression components is required to determine if the observed signatures are representative of cell type expression.

One mechanism to obviate cell type deconvolution is to directly measure single-cell expression profiles. There has been a recent explosion of unsupervised learning algorithms, including NMF and autoencoders, applied to derive insights from single-cell transcriptome data (65–74). The application goals are usually batch correction, imputation, visualization, cell state identification, or identifying pathway activity underlying homogeneous cell type populations. These differential patterns of pathway activity can aid in cell state identification. For example, differential pathway activity in a homogenous population of B cells in lupus patients was predictive of patient outcome (75). Additionally, by applying methods to increase the distance between points in a homogeneous cell type population of *Schistosoma* parasites, Tarashansky et al. identified subsets of cells that do not express specific marker genes previously thought to be omnipresent (71). Therefore, unsupervised models can keep pace with expanding data and extract patterns at increasing resolution.

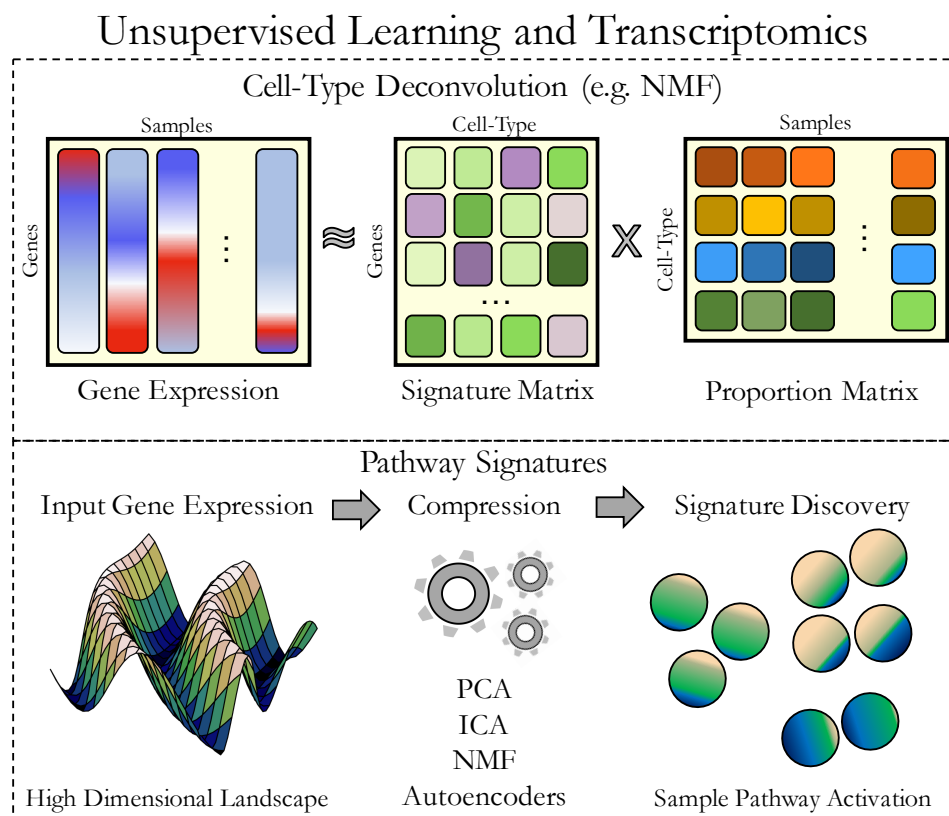


Figure 1.3: Unsupervised machine learning is used to discover cell type proportion and pathway signatures

(top) An illustrated example of unsupervised machine learning for cell type deconvolution. The input gene expression matrix is compressed into two component matrices, a signature matrix and a proportion matrix. The proportion matrix can be associated with cell type proportion in a given sample, and the signature matrix represents gene contributions to each signature. (bottom) Compression algorithms applied to a high-dimensional input matrix will automatically aggregate gene features into a lower-dimensional latent space. These latent spaces may represent pathway activities and other biological processes. Abbreviations: ICA, independent components analysis; NMF, non-negative matrix factorization; PCA, principal components analysis.

### 1.3.3. Unsupervised machine learning reveals underlying gene expression states

Compression algorithms applied to transcriptome data reveal pathway signatures hidden in latent spaces that represent a lower-dimensional data manifold (Figure 1.3). For example, PCA applied to a large compendium of nearly 80,000 transcriptomes showed a strong contribution of copy number alterations to disruptive gene signatures in

cancer (76). ICA has also been applied to transcriptome data to assign genes to gene modules and to identify pathway signatures and other hidden transcriptional programs (51, 77, 78); reviewed in Reference (79). In a direct comparison, ICA outperformed PCA in identifying gene modules significantly related to pathway activity in breast cancer samples (78). NMF is increasingly becoming the method of choice to derive pathway- and cell type-specific signatures from transcriptomic compendia (80–82). NMF does not constrain solutions to be orthogonal, and can therefore identify interconnected biological processes. A similar constrained latent variable approach provides interpretable pathway signatures and can identify pathway-specific activities while isolating technical artifacts (83). This method, called PLIER (pathway-level information extractor), has also been applied to large compendia to train a model that can provide insights into rare diseases through transfer learning (84). Other similar methods use Bayesian optimizations of matrix factorization to uncover patterns of biological processes hidden in transcriptome data (82, 85).

NMF identifies nonorthogonal linear patterns in data, which can be helpful in many tasks. Different techniques can use nonlinear activation functions to identify pathway activity from transcriptomic compendia. For example, denoising autoencoders (DAEs) trained on a large compendia of publicly available *Pseudomonas* transcriptomes were able to uncover biological pathways associated with the pathogen's response to media and oxygen exposure (86, 87). In this setting, the DAE was shallow, consisting of only one hidden latent space layer with a nonlinear activation function. DAE and stacked DAEs were also applied to yeast transcriptome data to reveal cell cycle expression signatures (88). DAEs compress input data through noise corruption and then reconstruct the original input through a nonlinear bottleneck layer (89). The corruption process provides regularization, permitting increased generalizability. More recent

applications have converted the autoencoder architecture into a generative model. A generative model learns a specific latent code that can be sampled from to simulate new data. VAEs are generative models (90, 91) and have gained popularity in transcriptome applications for a variety of purposes, including improving visualization and extracting hidden patterns underlying data (92). In Chapter 6, we discuss a VAE trained on TCGA PanCanAtlas expression data. This model revealed biological patterns associated with patient sex and various patterns of cell type and pathway activity, including immune cell infiltration (93). VAEs have also identified patterns of response to drug treatment in a panel of cell lines (94). However, it remains to be determined what other features are being compressed from transcriptomic compendia and what other signals representing known and potentially novel biology are being aggregated. In conclusion, unsupervised machine learning applied to transcriptomic compendia can reveal underlying patterns of cell type and pathway variation.

#### ***1.4. Interpreting machine learning models applied to transcriptomes***

Machine learning models enable the accurate detection of cellular states and robust predictions of pathway activity. In addition, interpreting supervised and unsupervised models can reveal important biology. Model interpretation is crucial to the success of any machine learning algorithm applied to transcriptome data. In Chapter 7, we discuss a novel approach to interpret compressed gene expression features using network projection.

##### ***1.4.1. Supervised learning models reveal differences between sample statuses***

Supervised learning models assign weights, or importance scores, to each gene expression feature given a classification or regression task. For example, an RF model will determine important gene expression features to split classes. Many methods have been developed to rank RF feature importance, including an integration of Gene

Ontology (GO) terms to predict gene expression changes. This technique has been applied to determine important genes in the aging process and response to chemical compounds in *Caenorhabditis elegans* (95, 96). Likewise, regression models and SVMs identify a subspace that represents specific activation patterns in the input feature space. The magnitude of these features can be interpreted as the most important genes for the classification task. Several methods penalize scores using recursive feature elimination and use hinge loss penalties to reduce the number of explanatory genes (97–99). A logistic regression model predicting Ras pathway activation identified similar genes as a differential expression analysis comparing Ras wild-type to mutant tumors (37). However, caution must be exercised when interpreting gene importance scores, since the algorithms can rely heavily on initializations, and different solutions are likely to implicate different genes (100). Models may select correlated genes and ignore causal genes, which is detrimental to downstream interpretation. NN models are also particularly difficult to interpret. The often black box models learn many layers of features with increasing complexity, and it is important not to over interpret what the models are learning. For instance, a sparse stacked autoencoder trained on yeast transcriptomes revealed transcription factor machinery in intermediate layers, but hidden layers are especially difficult to interpret (101).

#### 1.4.2. *Unsupervised learning models require interpretation of compressed features*

Compression algorithms applied to transcriptome data output features with different combinations of gene weights, or importance scores, that can be interpreted to represent biological processes. There are many mechanisms by which ranked gene lists can be interpreted, including overrepresentation pathway analysis and gene set enrichment analysis (102). However, the interpretation of compressed features in gene expression space has many open-ended questions. When trained on the same data set, the

distribution of feature importance scores across different algorithms has different skews and kurtosis values (Figure 1.4A). Therefore, it is not clear that interpreting compressed features is equivalent across algorithms. Furthermore, with the exception of the positive values learned by NMF, all other algorithms learn positive and negative signatures. It is not apparent if these values represent one general feature, two independent features, or something else. It is also not clear if the compressed features are learning single sources of variation, entangled sources of variation, or noise associated with technical artifacts. Thus far, researchers have attempted to interpret compressed features from a variety of algorithms in several ways (Figure 1.4B). For example, one can set a cutoff on gene importance scores based on two or three standard deviations above or below the mean (87, 103). Another strategy consists of sequentially removing top weighted genes from positive and negative tails and performing Lilliefors test of normality until the compressed feature resembles a normal distribution (77, 104). The removed genes represent a ranked gene list of the feature-specific genes. Another strategy is to use counterfactual analysis to observe which genes are strongly associated with covariates and to weight their importance to the biological source (105). In Chapter 7, we introduce a network projection approach that considers the full distribution of compressed gene expression features. We build gene set networks from publically available gene set compendia and determine enrichment of gene sets compared to permuted networks.

Another important question concerns how many compressed features exist. In other words, how many sources of variation to be compressed are there that contain important biology in a population? Researchers using a gene expression compendium of over 5,000 human tissues determined that only the first three principle components of a PCA contained biologically relevant information (106). However, a follow-up study using the same data extracted additional biologically meaningful features and reported that the low

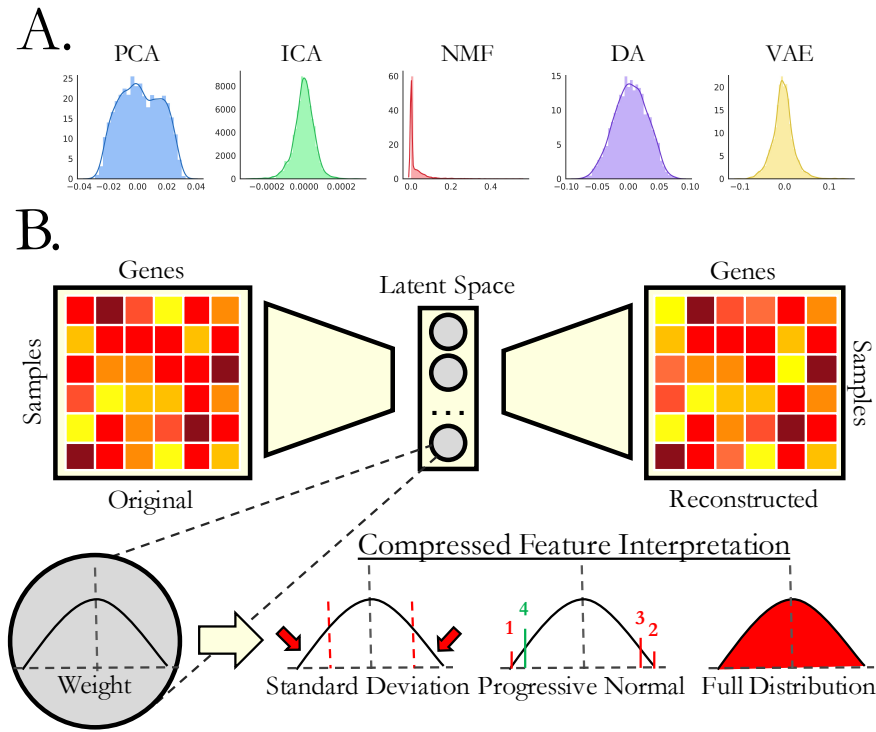


Figure 1.4: Interpretation of compressed gene expression features

(A) An example of a single random encoded feature of five different compression algorithms reveals the heterogeneity of the feature importance distribution. The input data are from The Cancer Genome Atlas PanCanAtlas gene expression data from 33 different tissue types spanning over 10,000 patients. (B) Defining genes that contribute to compressed features. These genes can be extracted in different ways. After the feature-associated genes are defined, there are various options for interpreting these compressed features, including various pathway- and network-based options. Abbreviations: DAE, denoising autoencoder; ICA, independent components analysis; NMF, non-negative matrix factorization; PCA, principal components analysis; VAE, variational autoencoder.

number of relevant compressed features was a sampling bias effect (107). Furthermore, an application of ICA to over 9,000 microarray samples revealed 423 components significantly associated with GO terms (51). A more recent analysis applying ICA to over 97,000 microarray samples revealed a total of 139 reproducible transcriptome modules (108). An issue common to many compression algorithms is the requirement to set an internal dimensionality. Mao et al. included extra capacity in the bottleneck layer to pool

technical artifacts in regions without prior biological knowledge constraints (83). In fact, it has been posited that gene expression consists of a series of compressed composite measurements (109). Nevertheless, it is clear that compression algorithms extract sources of variation in the underlying biology that are dependent on the strength of the signal, the number of samples that contain the biology, the assumptions of the model (e.g., linear versus nonlinear), and the predefined internal dimensionality. In Chapter 7, we investigate the dimensionality of gene expression data by serially compressing input matrices with an increasing bottleneck layer. More specifically, compress the data into 2 dimensions, 3 dimensions, 4 dimensions and so on up to 200. We project gene set networks onto the compressed features to quickly determine enriched gene sets captured in these features and determine how the bottleneck layer contributes to their identification.

Lastly, the stability of unsupervised learning solutions is of utmost importance. Because many unsupervised models are trained through an iterative process, the solutions identified will be different depending on internal conditions. Therefore, it is important to recognize stable patterns identified across various initializations. To this end, a method called stability NMF evaluates solutions from multiple starting points and determines stable basis vectors, or principle patterns, if they are consistently identified and correlated (110). Ensemble models have been used to aggregate solutions into a single model (87). Other methods have also been proposed to assess the stability of solutions, including adding dropout to NN models at test time (111). Nevertheless, interpreting machine learning models, investigating model stability, and associating compressed features with real biology are of paramount importance.



### **1.5. Conclusion**

Machine learning applied to transcriptomic compendia reveals interesting substructures in high-dimensional data that often represent cell type and pathway signatures. Both supervised and unsupervised machine learning models have been successfully applied to derive expression signatures with a variety of goals. As transcriptomic compendia continue to grow in size and resolution, so will the need for rapid insight generation and decision making abilities. In many models, there are no restrictions on which signals the machine learning models use to learn, so they can include artifacts and batch effects. Therefore, models must be applied to independent data sets to confirm the learned target biology. In addition to testing alternative data, orthogonal evidence supporting the discovered biology can help determine which signals are accurately interpreted and repeatable, and then additional molecular experiments can confirm the model's ability to identify biological signals. We are also in an age where computational experiments should be made reproducible (112). Therefore, software to reproduce machine learning models should be provided with publications to enable other researchers to quickly build upon work. Transcriptomic compendia contain vast amounts of signal and value, and machine learning is one technology that can tap into this resource.

### **1.6. Acknowledgements**

The authors would like to acknowledge funding from The Gordon and Betty Moore Foundation under GBMF 4552 and the National Institutes of Health's National Human Genome Research Institute under R01 HG010067 and the National Institutes of Health under T32 HG000046.

## Chapter 2.

### **A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma**

This chapter was originally published as: Way, Gregory, P.\*, Allaway, Robert, J.\*, Bouley, Stephanie, J., Fadul, Camilo, E., Sanchez, Yolanda, and Greene, Casey, S. “A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma.” BMC Genomics 18:127 (2017). doi: 10.1186/s12864-017-3519-

7. \*indicates co-first authors

Conceptualization: G.P.W., R.J.A., C.E.F., Y.S., C.S.G.; Methodology: G.P.W., R.J.A., C.S.G.; Software: G.P.W.; Investigation: G.P.W., R.J.A.; Writing – Original Draft: G.P.W., R.J.A., S.J.B., Y.S., C.S.G.; Writing – Review and Editing: G.P.W., R.J.A., S.J.B., C.E.F., Y.S., C.S.G.; Resources: C.E.F., Y.S., C.S.G.; Visualization: G.P.W., R.J.A.

#### **Contributions:**

In the paper Way, Allaway et al. 2017, I was a co-first author. Specifically, I trained and evaluated the machine learning approach to detect NF1 loss of function. The other co-first author, Allaway, performed the molecular validation experiments. Allaway also wrote methods sections 2.3.5, 2.3.6, 2.3.7, and 2.3.9 and produced the western blot in Figure 2.2. He also wrote and interpreted the gene and pathway analysis in section 2.4.3 and 2.5. I drafted all other sections and compiled all other figures. All authors provided comments on various revision versions and helped to design the study.

#### **2.1. Abstract**

##### *2.1.1. Background*

We have identified molecules that exhibit synthetic lethality in cells with loss of the neurofibromin 1 (*NF1*) tumor suppressor gene. However, recognizing tumors that have inactivation of the *NF1* tumor suppressor function is challenging because the loss may occur via mechanisms that do not involve mutation of the genomic locus. Degradation of the NF1 protein, independent of *NF1* mutation status, phenocopies inactivating mutations to drive tumors in human glioma cell lines. NF1 inactivation may alter the

transcriptional landscape of a tumor and allow a machine learning classifier to detect which tumors will benefit from synthetic lethal molecules.

### 2.1.2. *Results*

We developed a strategy to predict tumors with low NF1 activity and hence tumors that may respond to treatments that target cells lacking NF1. Using RNAseq data from The Cancer Genome Atlas (TCGA), we trained an ensemble of 500 logistic regression classifiers that integrates mutation status with whole transcriptomes to predict NF1 inactivation in glioblastoma (GBM). On TCGA data, the classifier detected *NF1* mutated tumors (test set area under the receiver operating characteristic curve (AUROC) mean = 0.77, 95% quantile = 0.53 – 0.95) over 50 random initializations. On RNA-Seq data transformed into the space of gene expression microarrays, this method produced a classifier with similar performance (test set AUROC mean = 0.77, 95% quantile = 0.53 – 0.96). We applied our ensemble classifier trained on the transformed TCGA data to a microarray validation set of 12 samples with matched RNA and NF1 protein-level measurements. The classifier's NF1 score was associated with NF1 protein concentration in these samples.

### 2.1.3. *Conclusions*

We demonstrate that TCGA can be used to train accurate predictors of NF1 inactivation in GBM. The ensemble classifier performed well for samples with very high or very low NF1 protein concentrations but had mixed performance in samples with intermediate NF1 concentrations. Nevertheless, high-performing and validated predictors have the potential to be paired with targeted therapies and personalized medicine.

## 2.2. Background

Genomic tools allow investigators to devise therapies targeting specific molecular abnormalities in tumors. One such alteration is the loss of neurofibromin 1 (NF1), an important tumor suppressor that regulates the activity of *RAS* GTPases (113, 114). Heterozygous mutation or deletion of *NF1* causes neurofibromatosis type 1 (NF), one of the most frequently inherited genetic disorders (115). NF patients often develop plexiform neurofibromas (PNs), benign nerve tumors for which the only therapy is surgery. However, resection is often impossible due to the tumor's intimate association with peripheral and cranial nerves (116). PNs can transform to malignant peripheral nerve sheath tumors (MPNSTs), which are chemo- and radiation-resistant sarcomas with a dismal 20% 5-year survival (117). In addition, patients with NF are susceptible to a broad spectrum of other tumors including low-grade/pilocytic astrocytomas, pheochromocytomas, optic nerve gliomas, and juvenile myelomonocytic leukemias (118). Many aggressive non-NF associated (sporadic) tumors have recently been shown to harbor *NF1* mutations, including glioblastoma (GBM), neuroblastoma, melanoma, thyroid, ovarian, breast, and lung cancers (119). Therefore, somatic and inherited loss of NF1 function is emerging as a driver of tumors from different organ sites.

Several groups including our own have been working to develop therapeutic approaches to target tumors with loss of NF1. Previously, our lab developed a high throughput approach using yeast and mammalian screening platforms to identify tool compounds and drug targets for cancer cells in which NF1 loss drives tumor formation. Our pipeline identified small molecules that selectively kill or stop the growth of MPNST cells carrying a mutation in *NF1* or yeast lacking the *NF1* homolog *IRA2* (120). We also developed an assay in yeast to identify the targets of our lead tool compounds and found that one of these compounds (UC-1) shares a mechanism (phosphorylation of RNA Pol

II CTD Ser2/5) with experimental drugs in clinical trials (120). UC-1 impacts CTD phosphorylation, which is regulated by the CTD kinase Ctk1, the yeast homolog of human Cdk9. We showed that deletion of *CTK1* was synthetic lethal with loss of the yeast *NF1* homolog *IRA2*. Furthermore, we have found that inhibitors of this process (dinaciclib, SNS-032) can inhibit other types of RAS-dysregulated tumor cells (121).

However, relying on genetic data alone to identify tumors that may be susceptible to therapies targeting NF1 loss may leave a proportion of potentially actionable tumors unrecognized. NF1 tumor suppressor activity can be lost via mutation of the genomic locus, proteasome-mediated degradation, inhibition by miRNA, *de novo* insertion of an ALU element, and C→U editing of the *NF1* mRNA (122–126). This complexity presents challenges when trying to identify tumors that will benefit from molecules that exert synthetic lethality with dysregulation of NF1/RAS pathways.

The Cancer Genome Atlas (TCGA) has released a large volume of data on several cancer tissues measured on a variety of genomic platforms. In the present study, we leverage TCGA GBM RNAseq expression data with matched mutation calls to construct a classifier capable of identifying an NF1 inactivation signature. This strategy sidesteps the problem of functional characterization of mutations by evaluating a regulator's downstream gene expression activity. We applied this signature to predict NF1 inactivation in a cohort of biobanked GBMs. In general, this approach can be translatable to any gene producing measurable downstream transcriptome-wide effects.

### **2.3. Methods**

#### *2.3.1. The Cancer Genome Atlas data used for building the classifier*

We downloaded RNAseq and mutation data from TCGA Pan Cancer project from the UCSC Xena data portal (127) and subset each dataset to only the GBMs (128). The data consists of 607 GBMs; of which 291 have mutation calls, 172 have RNAseq

measurements, and 149 have both RNAseq and mutation calls. Of these 149 samples, 15 have inactivating *NF1* mutations (10.1%) and were used as gold standard positives in building the classifier. Additionally, to reduce dimensionality while avoiding unexpressed and invariant genes, we subset to the top 8,000 most variably expressed genes by median absolute deviation. We z-scored all gene expression measurements. This resulted in the final input matrix with dimension 149 samples by 8,000 genes. For use in platform independent predictions, we used Training Distribution Matching (TDM) to transform the TCGA RNAseq data to match a microarray expression distribution (129).

Since we are also aware of the *NF1* mutation status for each of the samples, we form a supervised learning task – predicting when a sample has loss of NF1 activity. Our  $X$  matrix is formed by the RNAseq measurements for all 149 samples measured by 8,000 genes, which are the features in the model. Our  $y$  vector consists of  $\{0, 1\}$  elements where a 1 corresponds to a sample with an inactivating *NF1* mutation and a 0 is an *NF1* wildtype sample. The machine learning task is to find the feature weights, or gene coefficients, that best minimize our objective function. Along with these feature weights corresponding to the genes' importance in the learning task, we also output a probability estimate for each sample that they have loss of NF1 activity.

### 2.3.2. Hyperparameter optimization of the logistic regression classifier

Using the GBM RNAseq data, we trained logistic regression classifiers with an elastic net penalty using stochastic gradient descent to detect tumors with NF1 inactivation. We chose a penalized regression model because it is simple to train and has easily interpretable outputs including importance scores for each gene (feature weights) associated with the downstream consequences of NF1 loss of function and a probability for each sample that NF1 is lost. An elastic net logistic regression model has also been successfully implemented in similar studies (34, 130, 41).

We identified high-performing alpha and L1 mixing parameters using 5-fold cross validation ensuring balanced membership of *NF1* mutations in each fold. Briefly, alpha controls how weight penalty and the L1 mixing parameter tunes the amount of test set regularization by controlling the sparsity of the features. An L1 mixing parameter value of zero corresponds to the L2 penalty and a value of one corresponds to the L1 penalty, with L1 bringing a sparser solution. We used python 3.5.1 and Sci-kit Learn for machine learning implementations (131).

### 2.3.3. *Ensemble classifier construction and application to the validation set*

After selecting optimal hyperparameters, we constructed 500 classifiers that would compose our ensemble model. Specifically, across 100 different random initializations, we subset the full TCGA GBM data into 5 folds and trained a single classifier for each training fold.

We borrowed terminology from the epidemiology field to describe data partitioning. We trained our models on a “training” partition and assessed model performance on a “test” partition, which refers to the held out cross-validation fold. The independent “validation set” refers to the GBM dataset generated in a different lab (see Figure 2.1A).

Because of the small number of gold standard positive training examples, we were concerned about the stability of our model solutions. Therefore, we constructed an ensemble classifier from the 500 models. Specifically, we assigned each classifier a weight using the specific randomization’s “test set” cross-validation AUROC. Lastly, for the final *NF1* inactivation prediction, we used the mean of the weighted predictions across all iterations as the *NF1* inactivation prediction. We applied this ensemble classifier to the validation set in which *NF1* protein levels were directly measured.

#### 2.3.4. *Effect sizes and power analysis*

We calculated the decision function of each ensemble classifier applied to all samples in the training and testing 5-fold cross validation folds to calculate Cohen's D effect size between predicted *NF1* wildtype and NF1 inactive samples (132). The Cohen's D metric quantifies the difference between *NF1* wildtype and NF1 inactive samples according to the mean classifier score and directly demonstrates how different the ensemble model predicts the two groups to be.

Moreover, we were also concerned that our relatively small validation set would not provide us with enough power to observe a detectable effect in the ensemble model's final prediction. We performed a one-tailed Welch two-sample t-test comparing the NF1 protein concentration of our validation samples that were predicted to be either NF1 wildtype or NF1 deficient. Using the given sample size, Cohen's D effect size, and a significance threshold of  $\alpha = 0.05$ , we calculated the power of the prediction scores on the validation set. The power analysis was two-sample, one-tailed and incorporated unequal sample sizes in each group.

#### 2.3.5. *Validation sample acquisition*

Thirteen flash-frozen, de-identified GBM samples were obtained from the Maine Medical Center Biobank. Samples were received on dry ice and stored at  $-80^{\circ}\text{C}$  until isolation of DNA/RNA/protein. To isolate DNA, tumor fragments of approximately 20 mg in mass were harvested on an aluminum block pre-chilled on dry ice. Samples were then immediately transferred to a mortar and pestle containing a small volume of liquid nitrogen. The fragments were pulverized in the mortar and pestle, and the liquid nitrogen was allowed to evaporate. Next, samples were immediately processed with a DNA/RNA/Protein Purification Plus kit (Norgen Biotek) following the standard operating protocol for animal tissue. DNA concentration and quality were assessed using an ND-



1000 (Nanodrop), a Qubit Fluorometer (Thermo Scientific), and a Fragment Analyzer (Advanced Analytical Technologies). To isolate RNA, -80 °C tumor fragments were placed in 5-10 volumes of RNeasy Lysis Buffer (Qiagen) and placed at -20 °C until RNA extraction with a mirVana miRNA isolation kit, without phenol, following the standard operating protocol (Thermo Scientific). Samples were homogenized using a manual homogenizer in the presence of mirVana lysis buffer. RNA concentration and quality were determined using a Qubit Fluorometer (Thermo Scientific) and a Fragment Analyzer (Advanced Analytical Technologies). To isolate protein, small tumor fragments were pulverized and lysed in approximately 3 volumes of ice-cold radioimmunoprecipitation assay (RIPA) buffer (150 mM sodium chloride, 1% v/v nonidet P40, 0.5% w/v sodium deoxycholate, 0.05% w/v sodium dodecyl sulfate, 50 mM Tris pH 8.0) containing 1 mM sodium orthovanadate, 1 mM sodium fluoride, 1 mM phenylmethylsulfonyl fluoride, and 1X protease inhibitor cocktail (0.1 µg/mL leupeptin, 100 µM benzamidine HCl, 1 µM aprotinin, 0.1 µg/mL soybean trypsin inhibitor, 0.1 µg/mL pepstatin, 0.1 µg/mL antipain). Samples were passed through a 25 5/8 g needle and subsequently sonicated on ice to promote efficient lysis and DNA shearing. After a 30-minute incubation on ice, lysates were cleared by centrifuging at 16100 x g for 20 minutes. HEK293T, U87-MG, and U87-MG cells treated for two hours with 1 micromolar bortezomib (Selleckchem) and 10 micromolar MG132 (Selleckchem) were also prepared in RIPA buffer. Protein samples were stored at -80 °C until analysis.

#### 2.3.6. Cell culture

U87-MG and HEK293T cells were purchased from ATCC. Cell lines were regularly passaged and were cultured in Dulbecco's Modified Eagle Medium (Corning) with 10% v/v fetal bovine serum (Gibco) at 37 °C in 5% CO<sub>2</sub>.

Recent data regarding the U87MG cell line published by Allen *et al* suggest that the U87MG cell line distributed by ATCC is not from the same tumor as the cell line that was originally isolated in Uppsala. Transcriptome analysis comparing ATCC U87MG cell line to known tumor transcriptomes indicate that the ATCC U87MG cell line is a central nervous system tumor and is likely a glioblastoma cell line (133).

In the present study, we employ this cell line as a control representing an NF1-deficient tumor cell line. Previous studies have shown that the U87MG cell line has elevated proteasome-mediated degradation of NF1 and that this cell line required the loss of NF1 protein to promote tumorigenesis in xenograft tumor models (122). Given that the ATCC U87MG cell line is a well-characterized and broadly-used model of NF1 deficient tumor cells (122, 134–136), we propose that the use of the ATCC U87MG cell line is an appropriate control for Figure 2.2.

#### *2.3.7. RNA microarray*

After RNA isolation and QC, samples were labeled for the GeneChip Human Transcriptome Array 2.0 (HTA 2.0, Affymetrix). Labeling was performed with Affymetrix Proprietary DNA Label (biotin-linked) using a WT Plus Kit (Affymetrix) provided with the HTA 2.0, following the standard operating protocol for HTA 2.0, including PolyA controls. Hybridization, washing, and staining were performed with the WT Plus Kit, following the standard operating protocol for HTA 2.0. Washing and staining were performed using a GeneChip Fluidics 450. Scanning was performed with a GeneChip Scanner 3000. These data were deposited in the Gene Expression Omnibus under accession GSE85033.

#### *2.3.8. Validation sample processing*

We applied a quality control pipeline (137) to all CEL files generated by the HTA 2.0. All validation samples passed processing quality control, which included an inspection of spatial artifacts, MA plots, probe distributions, and sample comparison boxplots. We

summarized transcript intensities using robust multi-array analysis (RMA) (138). We determined batch normalization was unnecessary after a guided principal components analysis (gPCA) using sample processing date and array plate ID as potential batch effect confounders (139). Lastly, we collapsed HTA2.0 transcripts into gene level measurements using the ``collapseRows()`` function with the “maxmean” method from the R package WGCNA (140). We used the pd.hta.2.0 platform design file (version 3.12.1) and the Bioconductor package “hta20sttranscriptcluster.db” (version 8.3.1) to map manufacturer transcript IDs to genes. We performed all preprocessing steps using R version 3.2.3.

#### 2.3.9. *Western blotting*

Prior to sodium dodecyl sulfate polyacrylamide gel electrophoresis, protein sample concentration was determined using a Pierce BCA Protein Assay Kit (Thermo Scientific). Protein samples were prepared with 1X Laemmli sample buffer (50 mM Tris pH 6.8, 0.02% w/v bromophenol blue, 2% w/v SDS, 10% v/v glycerol, 1% v/v beta-mercaptoethanol, 12.5 mM EDTA) and 50 µg of tumor protein. Volumes were normalized with RIPA buffer including the protease/phosphatase inhibitors described above. SDS-PAGE was performed using a 4-15% Mini-PROTEAN TGX gel (Bio-Rad) for 1 hour at 120V. The samples were then transferred to a nitrocellulose membrane for 2 hours and 45 minutes at 400 mA in cold transfer buffer (384 mM glycine, 50 mM Tris, 20% methanol, 0.005% w/v sodium dodecyl sulfate). Following this, the blots were then blocked in 5% w/v BSA or 5% w/v nonfat dry milk in Tris-buffered saline (137 mM NaCl, 2.7 mM KCl, 19 mM Tris, 0.05% v/v Tween 20, pH 7.4) for 25 minutes. Immunoblotting was performed with the following antibodies and conditions (vendor, species, diluent, dilution, incubation time, incubation temperature): anti-NF1 D7R7D (Cell Signaling, rabbit, 2% BSA, 1:1000, overnight, 4°C), anti-tubulin B-1-2-5 (Santa Cruz, mouse, 2%

milk, 1:10000, 1 hour, RT), anti-EGFR D38B1 (Cell Signaling, rabbit, 2% milk, 1:1000-1:2000, 1h, RT), p-ERK ½ (p44/42 MAPK) #9101 (Cell Signaling, rabbit, 2% BSA, 1:2000, overnight, 4°C), SUZ12 D39F6 #3737 (Cell Signaling, rabbit, 2% milk, 1:1000, overnight, 4°C). Anti-NF1 D7R7D was a kind gift from Cell Signaling Technologies, Inc.

The binding of the primary antibodies was detected by incubation with secondary antibodies goat anti-rabbit HRP 1:20000 or goat anti-mouse HRP 1:10000 (Jackson ImmunoResearch Laboratories Inc.) at room temperature in 2% milk in TBST and detection of HRP activity using Pierce ECL Western Blotting substrate (Thermo Scientific), or in the case of *NF1*, SuperSignal West Femto Maximum Sensitivity Substrate (Thermo Scientific). The chemiluminescent signal was captured with MED-B medical x-ray film (Med X Ray Company Inc.). Between primary antibodies, the membrane was stripped twice for 10 minutes at room temperature using a mild stripping buffer containing 1.5% w/v glycine, 0.1% w/v SDS, 1% v/v Tween 20 at pH 2.2 (Abcam). One sample was eliminated due to low yield, and apparent degradation as determined by western blotting (all proteins examined were undetectable with the exception of tubulin, not shown). Densitometry was performed using Li-COR Image Studio Lite 5.0. Briefly, intensity measurements for *NF1* and tubulin were taken using equally-sized regions for all bands. The background was subtracted using the local median intensity from the left and right borders (size=2) of each measurement region. *NF1* values were divided by tubulin intensity to adjust for protein loading. All measurement ratios were then normalized by dividing values by the “U87+PI” measurement for each blot, respectively.

#### 2.3.10. Reproducibility of computational analyses

We provide software with a permissive open source license to reproduce all computational analyses (141). Ensuring a stable compute environment, we performed all

analyses in a Docker image (142). This image and source code can be used to freely confirm, modify, and build upon this work.

## **2.4. Results**

### *2.4.1. Classifier performance*

Using 5-fold cross validation across a parameter sweep, we identified optimal hyperparameters at  $\alpha = 0.15$  and L1 mixing = 0.1. To assess model performance, we performed 100 random initializations of five-fold cross-validation (Figure 2.1A). These models had mean test area under the receiver operating characteristic curve (AUROC) of 0.77 (95% Quantiles: 0.53 – 0.95) and a mean train AUROC of 0.997 (95% Quantile: 0.98 – 1.00). We repeated this procedure after TDM transformation and achieved comparable results with  $\alpha = 0.15$  and L1 mixing = 0.1 (mean test AUROC = 0.77, 95% Quantiles: 0.51 – 0.96; mean train AUROC = 0.998, 95% Quantiles: 0.99 – 1.00) (Figure 2.1). Because the validation set was measured by microarray, we used the classifier trained on TDM transformed data to construct our ensemble classifier. We also determined the Cohen's D effect size estimate for all training and testing partitions across all 5-fold cross validation iterations of the TDM transformed model. The classifier consistently and robustly separated *NF1* wildtype and *NF1* inactivated GBM samples with high effect sizes (Training: mean Cohen's D = 3.07, 95% CI = 2.24 – 4.16; Testing: mean Cohen's D = 1.27, 95% CI = 0.19 – 2.67).

### *2.4.2. Identification and characterization of NF1 deficient glioblastoma tumor samples*

We characterized *NF1* protein concentrations as well as other molecules involved in RAS signaling in the 12 GBM samples (Figure 2.2A). Two samples (CB2, 3HQ) had no apparent *NF1* protein. Eight other samples had similar or less *NF1* signal than the U87-MG *NF1*-low control (H5M, LNA, YXL, VVN, R7K, TRM, UNY, W31). Two samples

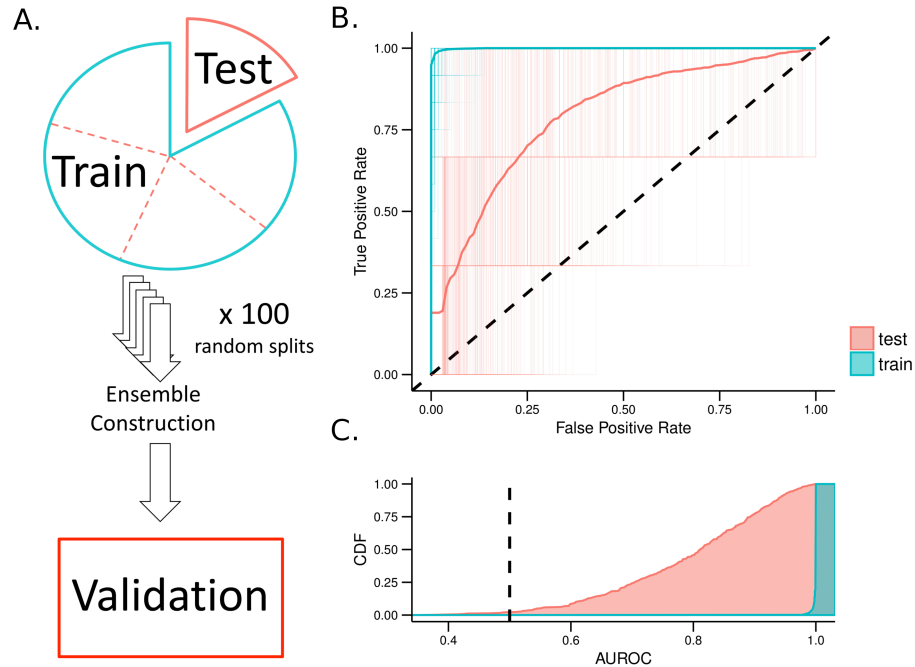


Figure 2.1: Ensemble classifier errors over 100 iterations for TCGA GBM RNAseq

**(A)** Schematic describing the terms used for training, testing, and validating our model. We applied 5-fold cross validation to the full dataset which consists of training and testing splits in each fold. The model is then applied as an ensemble classifier on a set of in-house samples (validation set) **(B)** Receiver operating characteristic (ROC) curves for all 500 classifiers that make up the ensemble model applied to both training and testing set. Also shown is the aggregate performance of the ensemble classifier. **(C)** The cumulative density of area under the ROC curve (AUROC) for training and testing partitions.

(PBH, RIW) had equal or greater NF1 than the positive control, U87-MG + proteasome inhibitors (preventing NF1 degradation). We also observed variable EGFR content in these samples, with non-existent to low levels (3HQ, YXL, R7K), or medium to large EGFR signal (CB2, H5M, PBH, LNA, YXL, VVN, RIW, TRM, UNY, W31).

All GBM samples had high concentrations of phospho-ERK1/2 signal relative to cell line controls. Samples with increased phospho-ERK1/2 may have greater Ras pathway activation. This can be attributed to multiple factors, including increased EGFR expression and/or NF1 inactivation.

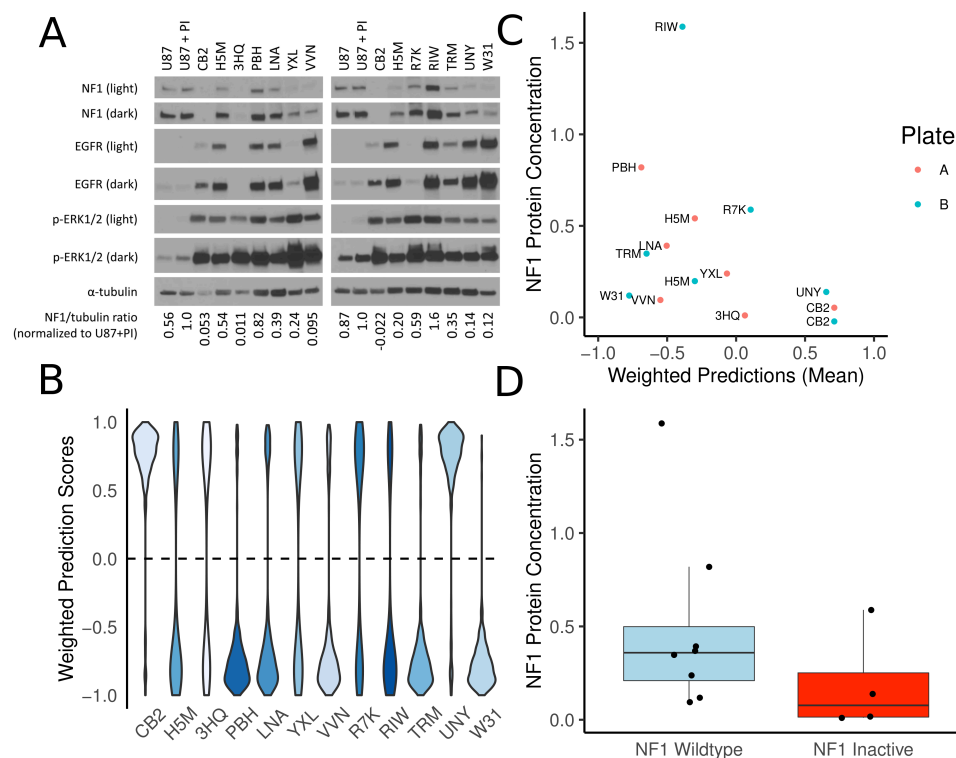


Figure 2.2: Performance of our classifier on an external validation set

**(A)** Two distinct western blots for each of our twelve samples. The controls are U87-MG, an *NF1* WT glioblastoma cell line that exhibits proteasomal degradation of the NF1 protein. U87+PI are U87-MG cells treated with the proteasome inhibitors (PI) MG-132 and bortezomib to block proteasome-mediated degradation of NF1. We used the NF1/tubulin ratio normalized to U87+PI as our NF1 protein level estimate. **(B)** Prediction scores for each of the 500 classifiers weighted by cross validation test set AUROC where a negative number indicates *NF1* wildtype and a positive number indicates *NF1* inactivation. Increasing color intensity indicates higher observed NF1 protein concentrations. **(C)** We quantify protein against U87+PI and provide the mean of the weighted predictions. **(D)** Based on weighted predictions, we show the abundance of NF1 protein compared to U87+PI.

Our ensemble classifier predicted four samples to have NF1 inactivation (CB2, UNY, R7K, and 3HQ) and eight samples to be NF1 wildtype (W31, TRM, PBH, VVN, LNA, RIW, H5M, and YXL) (Figure 2.2B). Because two samples, (CB2 and H5M) were measured on both western blots (Figure 2.2C), we used the mean of their NF1 protein level across both experiments.

We performed a one-tailed t-test to determine if NF1 protein concentrations were significantly higher in NF1 wildtype versus NF1 deficient samples based on our classifier predictions (Figure 2.2D). We did not observe a significant difference across groups ( $t = -1.38$ ,  $p = 0.098$ , effect size = 0.699). Additionally, while the effect size was fairly large, a power analysis indicated that we required 22 samples per group to achieve a power = 0.8. With a lack of glioblastoma samples with quantified NF1 protein available, the trend of less protein present in NF1 inactivated samples nevertheless remains promising. One of the samples predicted to be NF1 inactive contains detectable NF1 protein (R7K), suggesting that this sample may have NF1 inactivation not detectable by assaying protein, have a different alteration that phenocopies NF1 loss, or is incorrectly predicted by the classifier. Conversely, there are three samples predicted to be NF1 wildtype that have low or undetectable protein (YXL, VVN, W31), which either indicates unknown elements that confound the detection of some NF1 dysregulated tumors or a classification error.

#### 2.4.3. *Highly contributing genes*

We observed several genes that consistently contributed to the ensemble classifier performance (Figure 2.3). Since we applied several classifiers to the validation set as an ensemble, we took the sum of all classifier's gene weights across all 500 iterations to define these consistently contributing genes. While the data indicate that these genes have an impact on classifier performance, the data do not indicate whether changes in the expression of these genes are a direct consequence in changes in NF1 signaling. Expression of genes such as *TXNIP*, *ARRDC4*, *ISPD*, *C10orf107*, and *DUSP18* appear to be predictive of intact NF1 signaling. Among the list of genes that appear to be expressed in tumors with loss of NF1 function are *QPRT*, *ATF5*, *HUS1B*, *PEG10*, *HMGA2*, *RSL1D1*, and *NRG1*.



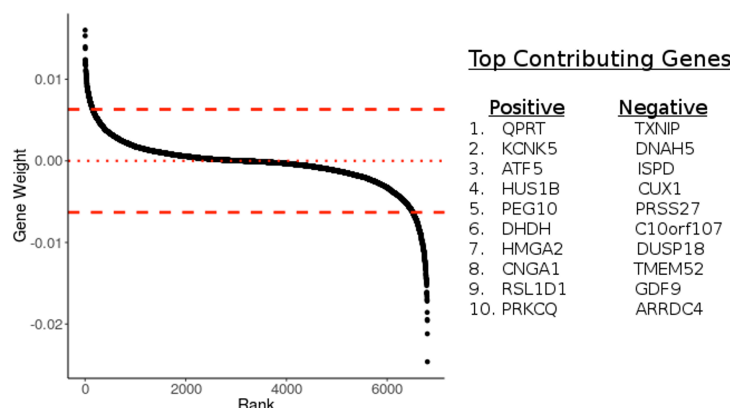


Figure 2.3: Genes that contribute to the NF1 classifier performance

Genes are shown ranked by their weighted contribution to the ensemble classifier. Weights are scaled to unit norm. The top 10 positive and top 10 negative contributing high weight genes are given on the right.

We also performed over-representation analysis of the most influential genes in the classifier to identify gene ontology (GO) sets and pathways that may be predictive of NF1 status (143–146). For high-weight genes predictive of intact NF1 signaling, we observed GO sets involved in plasma membrane-localized proteins (GO:0005886, GO:0071944, GO:0016324) and homeostasis (GO:0048871, GO:0001659, GO:0048873, GO:0031224), among others. Annotated pathways associated with genes from this dataset include hematopoietic stem cell differentiation, thyroid cancer, voltage-gated potassium channels, and RHO GTPase functional pathways.

For high-weight genes predictive of NF1 loss of function, we observed GO sets related to cellular adhesion (GO:0007155, GO:0098742), negative regulation of signaling (GO:0009968, GO:0023507, GO:0010648), and nervous system development (GO:0051962, GO:0007416, GO:0050808), among others. These genes were also enriched for elements of the phototransduction cascade and thyroxine production pathways.

## 2.5. Discussion

A machine learning classifier, based on gene expression data, can capture signal associated with the inactivation of a tumor suppressor. Our classifier is able to detect subtle downstream changes in gene expression as a result of the tumor responding to NF1 loss of function. This finding supports using mRNA as a summary measurement capable of capturing system-wide responses to molecular events beyond transcription factor alterations. Machine learning has been applied to gene expression in a variety of studies with various goals (23, 86, 147–149). In a similar study, Guinney *et al.* trained a classifier to model RAS activity in colorectal cancer and demonstrated its clinical utility by predicting response to MEK inhibitors and anti-EGFR based treatments (34). With a wealth of signal embedded in gene expression and a rapidly growing library of datasets, the performance of machine learning models is likely to rapidly improve. An increase in performance leads to more reliable clinical applications that would potentially predict the effectiveness of pathway-specific targeted therapies.

While our classifier was able to predict NF1 inactivation status to an extent, its performance is far from being clinically actionable. A major difficulty in developing a reliable classifier in this case is contamination in gold standard positives and negatives. While we aim to detect NF1 inactivation events, our gold standard positives can only include samples with known *NF1* mutation status. Conversely, we expect that negative samples (about 90% of the data) are also contaminated with NF1 inactivated samples due to protein loss and other mechanisms. We cannot determine scenarios where NF1 is inactivated beyond mutation at scale in the TCGA data. Another challenge with the construction of classifiers from such data is overfitting. Even after hyperparameter optimization we observed substantial overfitting (Figure 2.2), which has also been observed in competitions (see, for example, Supplementary Figure S2 of Noren *et al.*

2016 (150) in which the best performing algorithms also overfit). Finally with a small number of positive examples the model performance is unstable, which demonstrates high variability in gold standard samples used to train the model (151). We employed ensemble classification to mitigate this issue as averaging over heterogeneous models would result in a relatively stable classifier (see Figure 2.2B). In summary, our results are promising but these challenges are substantial and significant work remains to reach a robust classifier with clinical utility.

The performance of the classifier appears to be impacted by many cancer related genes. For example, genes such as *TXNIP* and *ARRDC4*, which are both indicative of lactic acidosis, correlate with better clinical outcomes, and contribute to predicting tumors with intact NF1 signaling (152). We also observed transcripts that are more highly expressed in brain tissue than either other normal tissue (*ISPD*, *C10orf107*), or more highly expressed in normal brain tissue than glioma (*EPHA5*) (153–155). *DUSP18* contributes to the prediction of *NF1* wildtype status and is a negative regulator of ERK phosphorylation, possibly by regulating *SHP2* phosphorylation (156). It is unclear whether the expression of these genes is a direct result of *NF1* expression, the result of signaling downstream of *NF1*, or a consequence of other phenomena (such as expression of *SPRED1*, an *NF1* binding partner that is essential for NF1 signaling). Future studies could elucidate the potential connections between *NF1* and the genes identified as important for the performance of this classifier.

Over-representation analysis of these data highlighted changes in potassium channel expression. It was previously demonstrated that *NF1* wild-type Schwann cells have altered K<sup>+</sup> channel activity as compared to *NF1*<sup>-/-</sup> Schwann cells suggesting that this may be one factor by which *NF1* mutant and wild-type cells can be distinguished (157).

Regarding prediction of NF1 inactivated tumors, we observed several genes that have been linked to cancer such as *QPRT*, which is highly expressed in malignant pheochromocytomas as compared to benign; *RSL1D1* (CSIG), which stabilizes *c-myc* in hepatocellular carcinoma; *PPEF*, which is highly expressed in astrocytic gliomas as compared to normal brain tissue (158–160); and *PEG10*, a poor prognostic marker and regulator of proliferation, migration, and invasion in several tumor types (161–163). We also observed *ATF5*, a gene for which expression in malignant glioma is correlated with poor survival (164). Knockdown of *ATF5* in GBM cells causes cell death *in vitro* and *in vivo* (165). Analysis of genes that contribute to the prediction of NF1 inactivation yielded several GO terms related to neural development. It is well established that loss of NF1 can result in abnormal neural development and/or tumorigenesis (126, 166, 167). We also observed genes associated with the mesodermal commitment pathway, components of which are linked to the epithelial to mesenchymal transition in human cancer cells (168–170). Analysis of this pathway may be informative in identifying tumors with NF1 loss because mesenchymal GBMs are enriched for tumors with NF1 loss (171).

Our ensemble classifier was able to robustly detect the samples with the highest and lowest NF1 protein concentrations, but it struggled with samples of intermediate NF1 concentrations. This could be a result of an enrichment of mechanisms causing NF1 inactivation beyond protein abundance, an overrepresentation of mesenchymal tumors in NF1 inactivated samples contaminating dataset splits (171), poor classifier generalizability, or incomplete data transformation between RNAseq and microarray data. Because training and testing performance were similar between transformed and non-transformed data, we don't anticipate performance to be impacted much by platform differences or classifier generalizability. Nevertheless, we demonstrated the ability of

system-wide gene expression measurements to capture downstream consequences of a complex biological mechanism that would otherwise require several different types of data acquisition to capture.

## **2.6. Conclusions**

A machine learning classifier for transcriptomic data was able to detect signal associated with the inactivation of *NF1*, a tumor suppressor gene. The gene is an important regulator of the oncogene *RAS* and is inactivated frequently in GBM and in other tumors. The measurement of NF1 inactivity cannot be comprehensively captured by any single genomic characterization such as targeted sequencing or fluorescence in situ hybridization. This difficulty arises from diverse and complex biological mechanisms that inactivate the tumor suppressor in a variety of ways. However, we demonstrated that measuring system-wide RNA can capture subtle downstream changes that occur in response to NF1 inactivation. Improving classification performance is required before transitioning such a model into clinical use, but our method could be used to characterize cell lines or patient-derived xenograft (PDX) models with inactive NF1. Eventually, with more data and improved classification, we expect machine-learning models constructed on system-wide transcriptomics will translate into clinically relevant predictions that will guide targeted therapy.

## **2.7. Acknowledgements**

This work was supported by the MMC BioBank, a core facility of the Maine Medical Center Research Institute, and the Dartmouth Genomics Shared Resource, a core facility of the Norris Cotton Cancer Center.

## Chapter 3.

### Machine learning detects pan-cancer Ras pathway activation in The Cancer Genome Atlas

This chapter was originally published as: Way, Gregory, P., Sanchez-Vega, Francisco, La, Konnor, Armenia, Joshua, Chatila, Walid, K., Luna, Augustin, Sander, Chris, Cherniack, Andrew, D., Mina, Marco, Ciriello, Giovanni, Schultz, Nikolas, The Cancer Genome Atlas Network, Sanchez, Yolanda, and Greene, Casey, S. “*Machine learning detects pan-cancer Ras pathway activation in The Cancer Genome Atlas.*” Cell Reports 23 (2018) 172-180. doi: 10.1016/j.celrep.2018.03.046.

Conceptualization: G.P.W., Y.S., and C.S.G.; Methodology: G.P.W. and C.S.G.; Software: G.P.W.; Investigation: G.P.W. and C.S.G.; Curation and Resources: F.S.V., K.L., J.A., W.K.C., N.S., A.L., C.S., A.D.C., M.M., and G.C.; Writing – Original Draft: G.P.W. and C.S.G.; Writing – Review and Editing: F.S.V., K.L., J.A., W.K.C., N.S., A.L., C.S., A.D.C., M.M., G.C., and Y.S.

#### Contributions:

In the paper Way et al. 2018, I was the first author. Specifically, I trained and evaluated the machine learning approach to detect Ras activity in the PanCanAtlas. I wrote the full manuscript and created all figures. The other co-authors contributed as specified above.

#### 3.1. Summary

Precision oncology uses genomic evidence to match patient with treatment, but often fails to identify all patients who may respond. The transcriptome of these “hidden responders” may reveal responsive molecular states. We describe and evaluate a machine learning approach to classify aberrant pathway activity in tumors, which may aid in hidden responder identification. The algorithm integrates RNA-seq, copy number, and mutations from 33 different cancer-types across The Cancer Genome Atlas (TCGA) PanCanAtlas project to predict aberrant molecular states in tumors. Applied to the Ras pathway, the method detects Ras activation across cancer-types and identifies

phenocopying variants. The model, trained on human tumors, can predict response to MEK inhibitors in wild-type Ras cell-lines. We also present data that suggest multiple hits in the Ras pathway confer increased Ras activity. The transcriptome is underused in precision oncology and, combined with machine learning, can aid in the identification of hidden responders.

### **3.2. Introduction**

Precision oncology matches cancer patients to specific therapies based on genomic evidence, but has benefited only a relatively low proportion of cancer patients to date (172, 173). While clinically promising, precision oncology lacks complete and accurate matching strategies and fails to identify many patients that could be matched using alternative approaches (174). Cataloging transcriptome measurements across thousands of tumors enables a systems biology perspective into the downstream consequences of molecular perturbation. Detecting these perturbations using transcriptomic states can improve precision oncology efforts toward more accurate and complete pairing of patients to effective treatments (175).

In the largest uniformly processed cancer dataset to date, TCGA PanCanAtlas has released multi-platform genomic measurements across thousands of tumors from 33 different cancer-types (176). With this scale of data, researchers can build and evaluate statistical models that stratify tumors based on aberrant gene and pathway function. Previously, strategies have been explored using expression signatures to stratify patients (177). Some strategies have used data from individual cancer-types. For example, gene expression signatures in colon adenocarcinoma (COAD) and glioblastoma (GBM) stratified tumors with aberrant *KRAS* and *NF1* function, respectively (34, 35). Furthermore, data integration approaches incorporating pathway connectivity, including PARADIGM, are used to characterize pathway activity and infer gain or loss of

function events (41, 178, 179). An unsupervised approach decomposing gene expression states in cell lines to map pathway activity has been proposed (180). Here, we introduce an elastic net penalized logistic regression classifier to learn signatures of gene or pathway alterations from gene expression assays of tumor biopsies across cancer-types. Our method is applied across cancer-types to learn an independent, pan-cancer signature of pathway aberration. Our method can be used to identify phenocopying variants and requires only gene expression data for inference on new data. We apply our method to detect Ras pathway activation pan-cancer.

The Ras pathway is frequently altered in many different cancer-types (181). When the pathway is activated, often by gain of function *KRAS*, *NRAS*, or *HRAS* mutations or through *NF1* loss of function events, cells increase their translational output and unchecked cellular proliferation occurs (182, 183). Certain cancer-types, such as pancreatic adenocarcinoma (PAAD), skin cutaneous melanoma (SKCM), thyroid carcinoma (THCA), lung adenocarcinoma (LUAD), and colon adenocarcinoma (COAD) are known to be largely driven by mutations in Ras pathway genes (184–187). Additionally, mutations in the Ras pathway have been observed to be early events driving tumorigenesis and have also been associated with poor survival and treatment resistance (188–191). Because the Ras pathway is ubiquitously misregulated, developing specific therapeutic targets is one of the National Cancer Institute’s key initiatives. However, Ras is also notoriously difficult to therapeutically target and accurate detection of its malfunction is paramount (192).

The most direct method of assessing Ras activation is by targeted sequencing of Ras. However, these methods would fail to detect unknown variants in other genes that phenocopy Ras activating mutations. Detecting such tumors may enable more patients to be targeted therapeutically. In the following study, we describe our machine learning



approach that integrates bulk RNA-seq, copy number, and mutation data from the PanCanAtlas. We apply the method to Ras genes and demonstrate that our method can detect Ras activation pan-cancer. The classifier also identifies *NF1* phenocopying events in TCGA and prioritizes Ras wild-type cell lines that respond to MEK inhibitors. Manually curated oncogenic variants in Ras pathway genes were assigned higher classification scores than variants with unknown significance. Our method can be applied to other cancer-associated genes and pathways as well. For example, the DNA Damage Repair PanCanAtlas analysis working group (AWG) applied this approach to detecting *TP53* inactivation (38).

### **3.3. Results**

#### *3.3.1. Machine learning models to predict pathway activity*

We developed a machine learning approach to detect aberrant pathway activity in tumors. The method integrated RNA-seq, copy number, and mutation data. The models were trained using tumors from TCGA PanCanAtlas with a complete set of these measurements; which included 9,075 tumors across 33 different cancer-types. The method is based on a logistic regression classifier framework regularized with an elastic net penalty. We used RNA-seq as a measurement describing the expression state of a tumor, and trained the classifier to detect downstream gene expression patterns consistent with aberrant pathway activity (Figure 3.1A). The algorithm learned a combination of gene importance scores, or weights ( $w$ ), that together learn to best separate aberrant from wild-type expression patterns. As input during training, tumors with any non-silent somatic variants in target genes were included in the positive set (Figure 3.1B). We also included copy number gains for oncogenes, and deep copy number loss for tumor suppressor genes (Figure 3.1B). For complete details about the model and training approach, refer to the methods section (section 3.4). In principle, this

approach could be applied to predict other gene or pathway events. Here, we applied the method to classifying Ras activity.

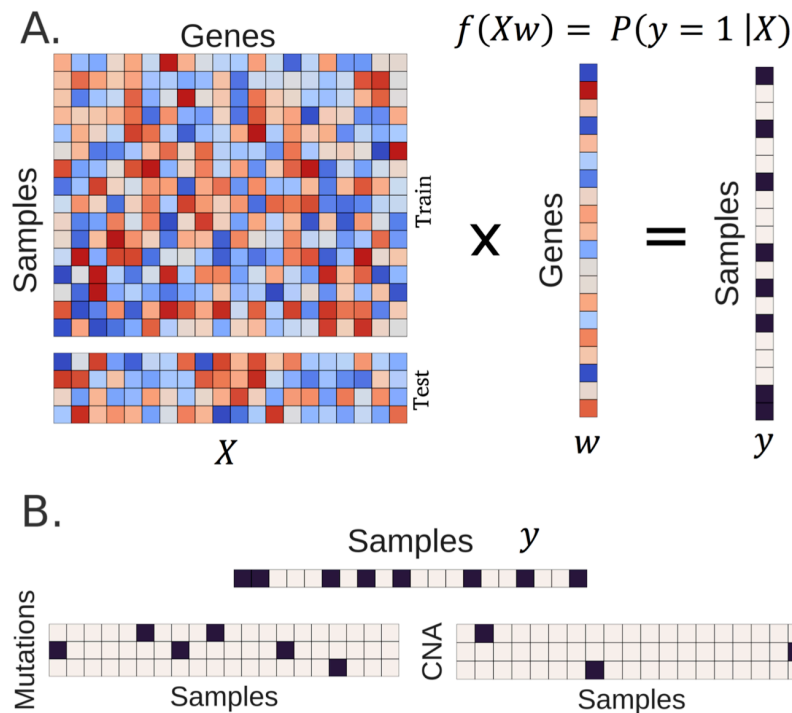


Figure 3.1: Supervised machine learning and data integration for TCGA PanCanAtlas

**(A)** RNAseq data ( $X$ ) is multiplied by a vector of gene weights ( $w$ ) where the optimization task is to find the optimal  $w$  to correctly classify the pathway status matrix ( $y$ ). We train the model with the train partition and evaluate performance on a held-out test set. **(B)** The status matrix  $y$  is constructed by integrating mutations and copy number alterations (CNA). We consider activating or loss of function mutations and high copy number gain and deep copy number loss for oncogenes and tumor suppressor genes, respectively. Black squares indicate aberrant events. For the Ras classifier, we used non-silent somatic mutations and high copy gains in the oncogenes *KRAS*, *NRAS*, and *HRAS*.

### 3.3.2. Detecting Ras activation pan cancer

We trained a classifier to detect aberrant Ras activity in tumors using knowledge of *KRAS*, *HRAS*, and *NRAS* mutations and copy number gains (see Figure 3.1). These 3 core Ras genes differed greatly in variant prevalence across cancer-types. In the PanCanAtlas, *KRAS* mutations were widespread in PAAD (72%), COAD (45%), rectum adenocarcinoma (READ, 42%), and LUAD (31%), while *NRAS* mutations were common



held out 10% of the samples ( $n = 476$ ) to create a test set. The test set was selected to have the same proportion of cancer-types and Ras statuses as the training set. The training set consisted of the remaining 90% ( $n = 4,283$ ), which included 3,374 Ras wild-type and 909 tumors with non-silent somatic Ras variants. Within the training set we performed five-fold cross validation (CV). We report training (“training”), cross-validation (“CV”), and held-out test set (“testing”) performance using these cancer-types. We also evaluated the final classifier on cancer-types that were initially filtered from training.

Overall, the classifier showed high performance, with an area under the receiver operating characteristic (AUROC) curve above 84% and an area under the precision recall (AUPR) curve above 63% in the CV and testing sets (Figure 3.3B). For the samples initially filtered from training, we also observed reasonable performance, with an AUROC = 75.2% and an AUPR = 24.7%. Therefore, the classifier detected Ras activation signal in tissues it was not exposed to during training. Applying the final classifier to all 9,075 samples, we observed an 86.7% AUROC and a 61.2% AUPR.

The Ras classifier consisted of automatically learned gene weights, or importance scores. Training with an elastic net penalty resulted in a sparse classifier, with only 185 genes contributing to classification. Genes and covariates with weights above zero can be interpreted as being up-regulated in tumors with activated Ras while negative weight genes are characteristic of tumors with wild-type Ras (Figure 3.3C). However, caution must be exercised in interpreting these coefficients as our elastic net regularization approach induces sparsity, which means that the solution represents a subset of genes associated with, and therefore useful for identifying, Ras activation. A differential expression analysis of Ras aberrant to wild-type tumors would reveal these downstream genes.

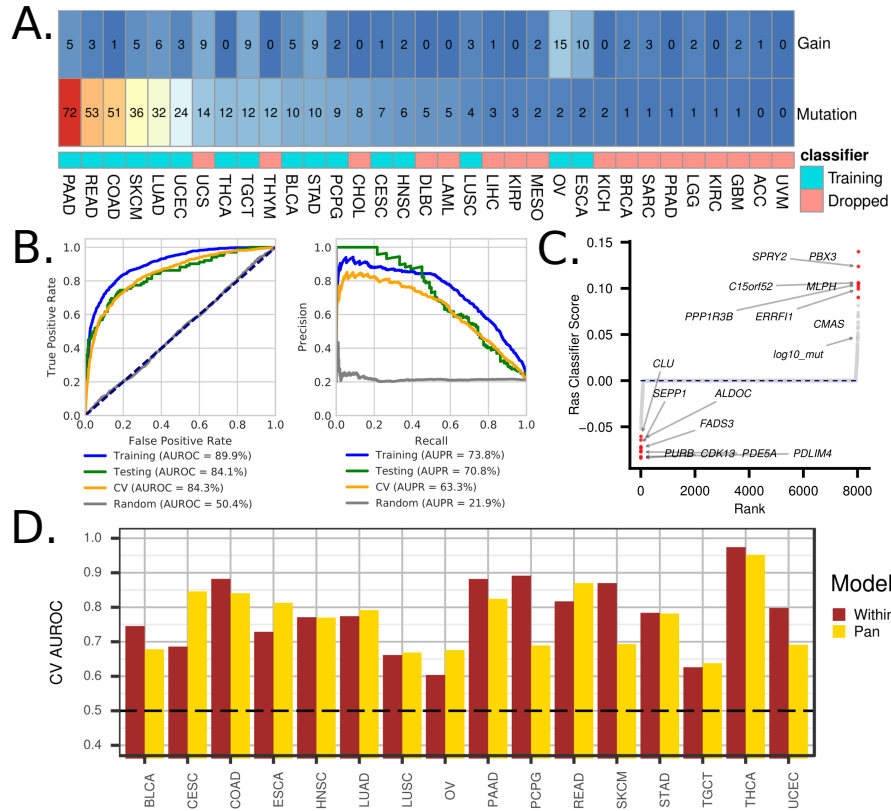


Figure 3.3: Evaluating machine learning classification of Ras activation

**(A)** Cancer-type specific percentages of Ras aberration by copy number gain and deleterious mutation in *KRAS*, *HRAS*, or *NRAS*. The colored squares indicate if the cancer-type was included in model training. **(B)** Predicting Ras pathway activation metrics. The grey lines represent classifier predictions on a randomly shuffled gene expression matrix. *Left*: Receiver operating characteristic (ROC) curve and Area under the ROC (AUROC) curve given for training, testing, and cross-validation (CV) sets. The dotted navy line represents a hypothetical random classifier. *Right*: Precision Recall (PR) Curve and corresponding area under the PR (AUPR) curve for each evaluation set. **(C)** Sparse classifier coefficients indicate which genes impact classifier performance. Log10\_mut represents tumor-specific non-silent mutation rate. **(D)** Cancer-type specific performance for the pan-cancer model compared to separate models trained on each cancer-type independently.

Nevertheless, many of the classifier implicated genes are known modulators of the Ras/MAPK pathway. For instance, high expression of *ERFF1* contributed to predicting tumors with activated Ras. *ERFF1* is a tumor suppressor of various receptors in the Ras pathway (194). The top positive gene, *PBX3*, is a transcription factor previously implicated in certain astrocytomas (195). The second top positive gene, *SPRY2*, inhibits

*FGFR* signaling and interacts with *ERBB1*. The negatively associated genes are indicative of expression profiles of wild-type Ras tumors. For example, *CDK13* was the most predictive gene and is involved in regulating transcription; which potentially indicates an alternative mechanism driving transcriptional disruption in wild-type Ras tumors. We also compared pan-cancer classification with classifiers trained independently within each cancer-type. Both the cancer-type specific and pan-cancer classifiers had variable performance across cancer-types, with the pan-cancer model outperforming the models optimized within cancer-types approximately half of the time (Figure 3.3D).

### 3.3.3. *Ras classifier benchmarking analyses*

We performed several analyses to evaluate the robustness of the Ras classifier. A null model trained on a randomly shuffled gene expression matrix performed with about 50% AUROC and 20% AUPR in holdout test and CV sets, which indicates strong performance of the model over this baseline (Figure 3.4A-B). We also assessed performance of the classifier for detecting Ras mutations and Ras copy number gains separately. Performance was similar with the mutations-only model performing better than the combined model and the copy number-only model performing worst (Figure 3.4C). Our model was robust to dropping *KRAS*, *NRAS*, and *HRAS* and 11 other Rasopathy genes from the gene expression matrix (Figure 3.4D). Lastly, performance was not impacted by covariate information (Figure 3.4E).

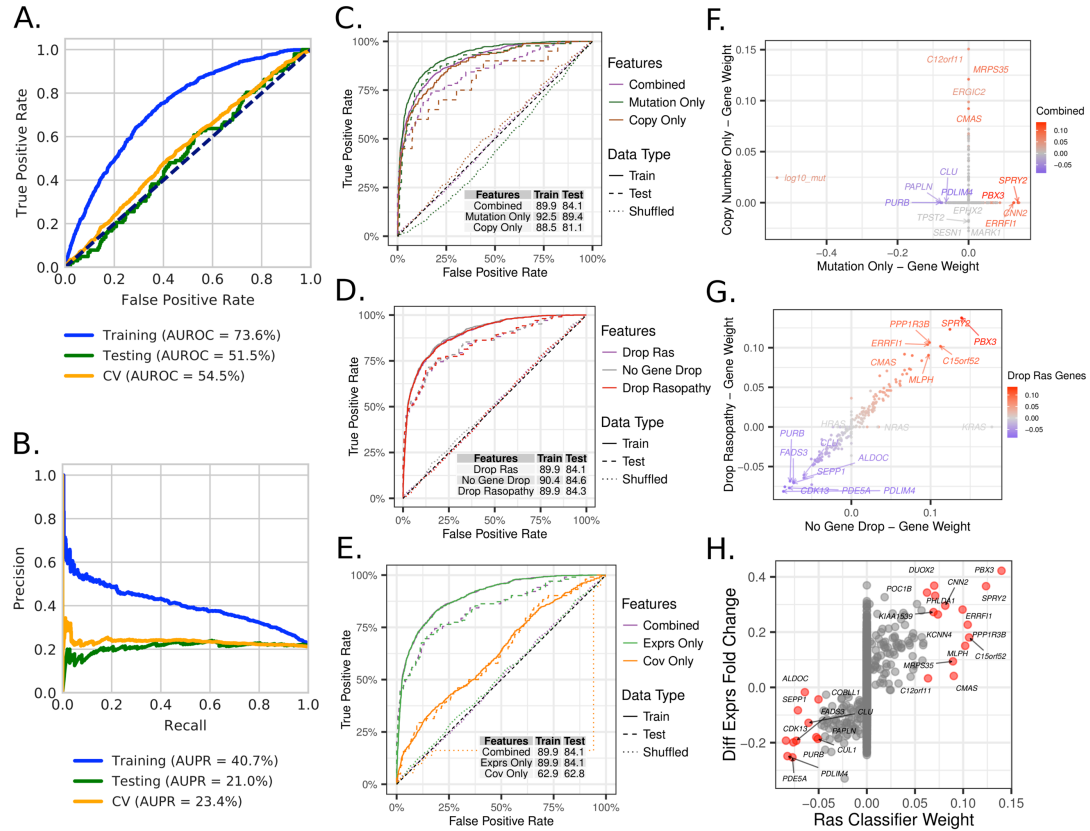


Figure 3.4: Benchmarking PanCanAtlas Ras classifiers

**(A)** Receiver operating characteristic (ROC) curve and **(B)** Precision recall (PR) curve for a null model trained on a randomly shuffled RNAseq matrix. Also provided are the area under the ROC (AUROC) and area under the PR (AUPR) curves for training, testing, and cross validation sets. **(C)** ROC curve for three models predicting: 1) Ras mutations only; 2) Ras copy number gains only; 3) Combined data (model in Figure 2). The AUROC is provided for both training and testing sets. **(D)** ROC/AUROC across train and test sets for dropping different genes from the RNAseq matrix. The Drop Ras model is the model provided in Figure 2. **(E)** ROC/AUROC across train and test sets for using expression data or covariates only. The combined model is the model provided in Figure 2. In all ROC curves, the dashed navy line represents a hypothetical random guess classifier. Gene coefficients for the models presented in **(F)** panel C and in **(G)** panel D. The points are colored by the model presented in Figure 3.3. **(H)** Differential fold change for tumors with active Ras against tumors with wild-type Ras compared against the Ras classifier gene coefficients. Red points correspond to labelled genes.

We also explored gene coefficient relationships across models. The high weight positive genes in the copy-only model included *C12orf11* (*ASUN*), *MRPS35*, *ERGIC2*, and *CMAS*; all of which are located on chromosome 12p near *KRAS*, which may indicate

artifacts of common copy gain events and be a result of low sample size in the positive copy-only set (Figure 3.4F). Gene coefficients were similar across models when dropping different Ras pathway genes (Figure 3.4G). Lastly, we compared our machine learning approach to a differential expression analysis of Ras mutant vs. wild-type tumors controlled by cancer-type. The differential expression scores aligned closely with the learned Ras classifier coefficients, but identified many more genes than the sparse classifier (Figure 3.4H). In summary, the Ras classifier differed depending on data-type inclusion, but was robust to input genes in the expression matrix, did not rely on covariate data, and included similar but fewer genes than a differential expression analysis.

#### 3.3.4. Detecting Ras activation in cell lines

We sought to determine whether or not predictions from the Ras classifier trained with TCGA tumors generalized to cell lines. We applied the classifier to two cell line datasets. First, we applied the classifier to 10 small-airway epithelial cell RNAseq profiles (GSE94937) (180). The set consisted of 4 wild-type profiles and 6 *KRAS* G12V expressing mutant profiles. Our classifier correctly classified 9 out of 10 profiles and ranked all mutant profiles higher than all wild-type profiles ( $p = 1.16e-2$ ) (Figure 3.5A). Though the PanCanAtlas data does not include gene edited tumors that would allow us to directly evaluate Ras oncogenicity, the cell lines from this independent test set are induced to stably express a *bona fide* oncogenic *KRAS* variant.

Next, we applied our Ras classifier to RNAseq profiles from 737 different cell lines from the Cancer Cell Line Encyclopedia (CCLE) with matched expression and mutation data (196) (Figure 3.5B). The Ras classifier assigned significantly higher scores to Ras mutated (*KRAS*, *HRAS*, or *NRAS*) cell lines than Ras wild-type cell lines ( $p = 6.35e-36$ ). Of the 393 cell lines predicted to be wild-type, 357 were labelled wildtype (negative



predictive value = 90.8%). However, only 153 of 344 cell lines predicted to be Ras mutated were labeled Ras mutant (precision = 44.5%). In total, 510 of 737 (69.2%) cell lines were predicted correctly. In this case, the low precision could indicate either that the classifier failed to generalize or that the classifier successfully identified phenocopying events, which were negatives from the point of view of evaluations but also what we aimed to capture.

We sought to differentiate between these two possibilities by using independent information that was not provided to the classifier. First, we examined mutation status for *BRAF*, a well characterized oncogene downstream of Ras genes (197). *BRAF* mutations that phenocopy Ras would be counted as negatives, and, if they were highly ranked, would reduce the observed precision. Indeed, the classifier assigned *BRAF* mutant cell lines with significantly higher scores compared to *BRAF* wild-type cell lines ( $p = 1.16 \times 10^{-11}$ ) (Figure 3.5B). Of all 191 false positives, 56 had *BRAF* mutations (29.3%). The remaining false positives either indicated tumors incorrectly assigned, or tumors that harbored other phenocopying variants. Next, we tested CCLE pharmacological response data to determine if Ras classifier scores were predictive of sensitivity to *MEK* inhibitors. We observed a strong correlation of the Ras classifier scores with sensitivity to two *MEK* inhibitors Selumetinib (AZD6244) and PD-0325901 (Figure 3.5C-D). The correlation was primarily driven by cell lines wild-type for Ras genes, implicating several drug sensitive cell lines that may have otherwise been missed by direct sequencing of Ras genes. Taken together, the evaluation of additional mutations and the drug response data for Ras wild-type cell lines strongly suggested that the low precision in this case was related to the identification of phenocopying events.

Lastly, the classifier scored 34 cell lines harboring Ras mutations as Ras wild-type. We observed that 22 of these 34 false negatives harbored variants annotated in the

COSMIC database (64%) (198). Conversely, 144 of 152 true positives harbored COSMIC variants (95%), which is significantly higher than the proportion in false negatives ( $\chi^2 = 26.1$ ,  $p = 3.2\text{e-}7$ ). Therefore, our classifier detected signal at variant level resolution.

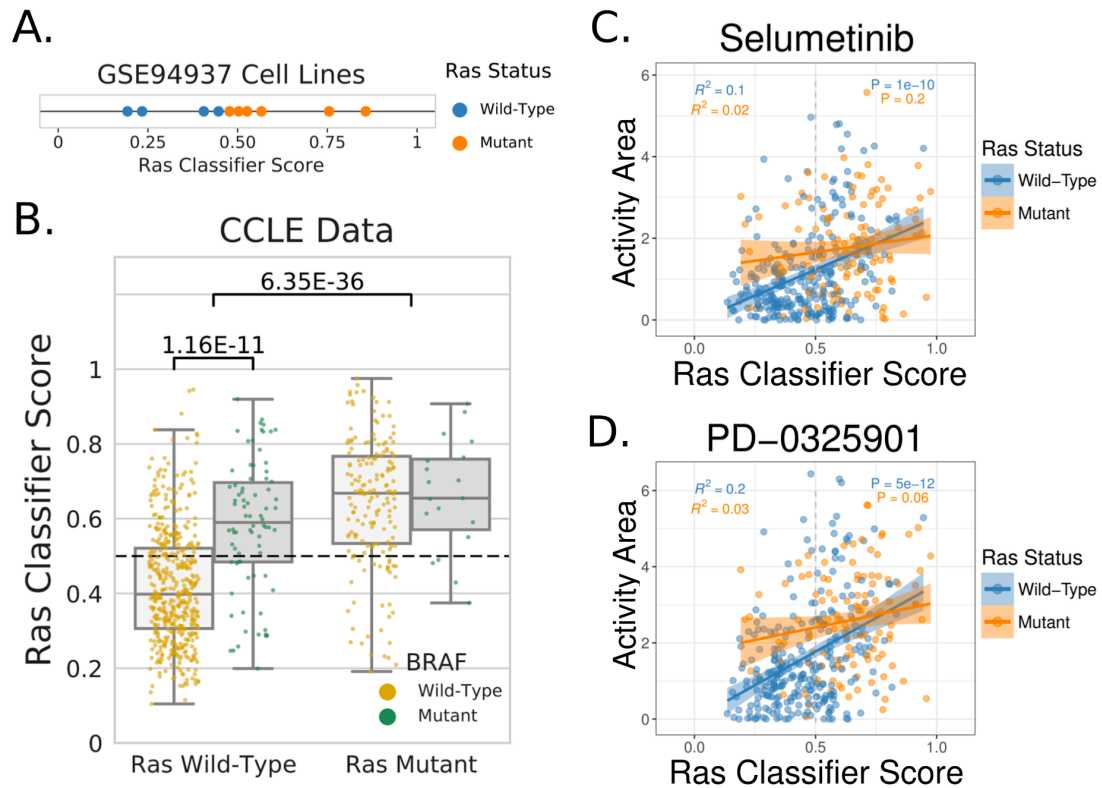


Figure 3.5: Cell line predictions of Ras activity by PanCanAtlas Ras classifier

**(A)** Ras classifier trained on PanCanAtlas tumors applied to a dataset of small airway epithelial cells. The mutant cells included a stably expressed *KRAS* G12V mutation. **(B)** Ras classifier trained on PanCanAtlas tumors applied to 737 cell lines from CCLE. Cell lines with *KRAS*, *HRAS*, or *NRAS* mutations are shown in the right boxes and wild-type tumors are shown in the left boxes. Scores for cell lines with *BRAF* mutations (green) and wild-type *BRAF* (gold) are also shown. Drug activity area for **(C)** Selumetinib (AZD6244) and **(D)** PD-0325901 compared against Ras classifier scores for 388 CCLE cell lines with both gene expression and pharmacologic profiling data. Cell lines with mutant (orange) or wild-type (blue) *KRAS*, *HRAS*, and *NRAS* is shown.

### 3.3.5. Other Ras pathway variants phenocopy Ras activation

The Ras classifier was able to detect *NF1* loss events particularly well in central nervous system tumors (GBM, low grade glioma (LGG), and pheochromocytoma & paraganglioma (PCPG)). Performance was comparable to *NF1* classifiers built using cancer-type specific and pan-cancer models (Figure 3.6A). These tumors were not included in training the Ras classifier. Detection of *NF1* inactivating events was also improved in COAD, OV, and uterine corpus endometrial carcinoma (UCEC) as compared to *NF1* specific classifiers (Figure 3.6A).

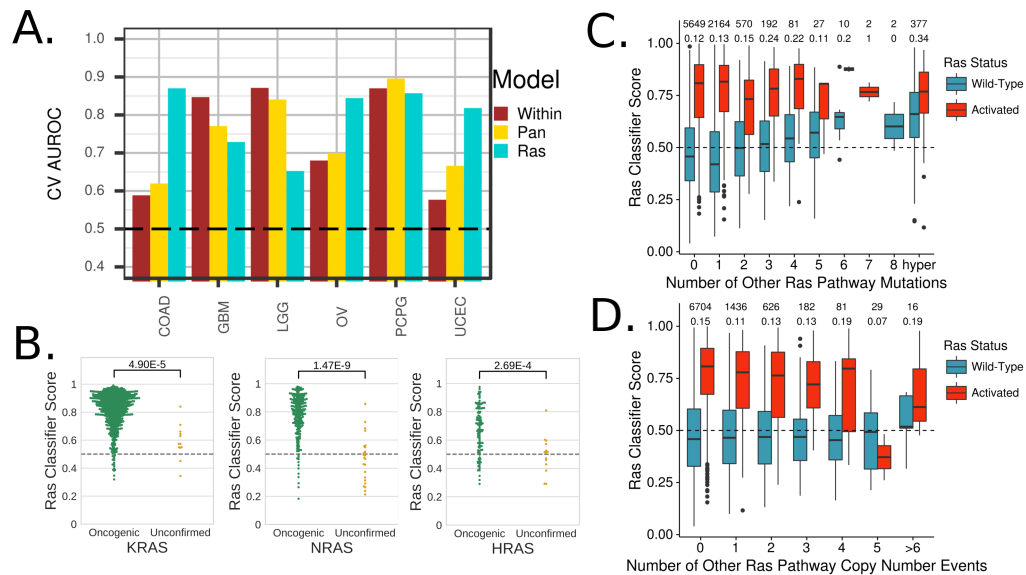


Figure 3.6: Ras activation across Ras variants and alternative Ras pathway members

**(A)** Cross validation area under the receiver operating characteristic curve for predicting *NF1* inactivation. Within and pan-cancer models are classifiers trained to detect *NF1* inactivation. The Ras model is the classifier trained in Figure 3.3. The *NF1* model is the classifier trained in Figure 3.7 **(B)** Ras classifier scores for samples with oncogenic or unconfirmed variants in *KRAS*, *HRAS*, and *NRAS*. Variant oncogenicity designations are based on curation (see methods). Ras classifier scores stratified by Ras activity (*KRAS*, *NRAS*, *HRAS*) status and number of **(C)** aberrant mutations or **(D)** copy number alterations in other Ras pathway members. The two rows of numbers above each graph indicate number of samples in each group (top) and percentage of samples assigned to active Ras (bottom).

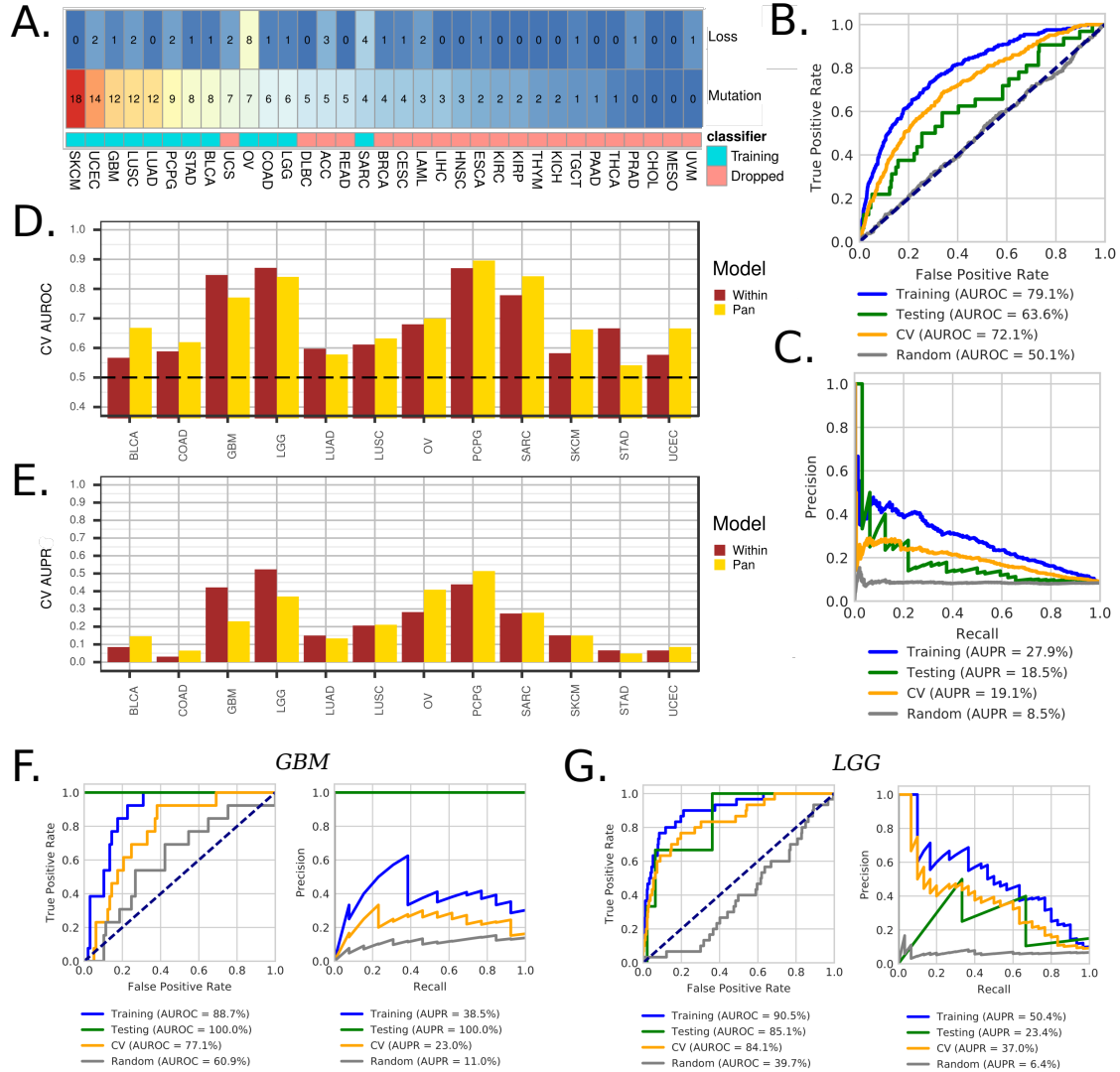


Figure 3.7: TCGA PanCanAtlas *NF1* classification performance

**(A)** Cancer-type specific percentages of *NF1* inactivation by copy number loss and deleterious mutation. The colored squares indicate if the cancer type was included in model training. **(B)** Receiver operating characteristic (ROC) curve and Area under the ROC curve (AUROC) given for training, testing, and cross-validation (CV) sets. **(C)** Precision Recall (PR) curve and corresponding area under the PR (AUPR) curve for each evaluation set. Cancer-type specific CV **(D)** AUROC and **(E)** AUPR for the *NF1* pan-cancer model compared to separate models trained on each cancer type independently. ROC and PR curves for predicting *NF1* inactivation in **(F)** GBM and **(G)** LGG using the pan-cancer model. The grey lines represent predictions made on a shuffled gene expression matrix.

The Ras classifier's performance predicting *NF1* loss of function was comparable to distinct pan-cancer models trained specifically to detect *NF1* loss of function events (Figure 3.7).

We applied the Ras classifier to curated variants in 38 core Ras pathway genes, which consisted of 34 oncogenes and 4 tumor suppressor genes (199, 200). We observed an enrichment of high scores in tumors with oncogenic variants in *KRAS*, *NRAS*, and *HRAS* (Figure 3.6B). Scores for oncogenic *BRAF* variants were also enriched (Figure 3.8A). However, we noted that *BRAF* V600E mutations in THCA were overwhelmingly predicted to be Ras wild-type (Figure 3.8B). We trained a classifier for which we removed both of the *BRAF* dominated cancer-types (THCA and SKCM) (Figure 3.8C). In this model, we observed that THCA *BRAF* V600E mutations were predicted to have Ras activation, which aligns with previous understanding of *BRAF* function and our cell line analysis (Figure 3.8D).

Lastly, in samples wildtype for *KRAS*, *NRAS*, and *HRAS* (blue bars), we observed that Ras classifier scores increased after subsequent mutations in other pathway genes (Figure 3.6C). In samples with a *KRAS*, *NRAS*, or *HRAS* mutation (red bars), classifier scores did not increase after additional mutations to other genes in the pathway (Figure 3.6C). However, more copy number events in other Ras pathway genes led to lower Ras classifier scores in Ras mutated samples (Figure 3.6D). These results potentially suggest that multiple hits in Ras pathway genes outside of Ras genes themselves may confer an increased Ras activation phenotype.

### **3.4. Discussion**

We described a machine learning method to detect malfunctioning genes and pathways in cancer and applied our method to detecting Ras activation. The method has variable performance across cancer-types, but is generally sensitive and specific overall,

is generalizable to cell line data, largely aligns with curated variant oncogenicity, and identifies phenocopying events leading to activated Ras. The approach can be applied generally to other genes and pathways.

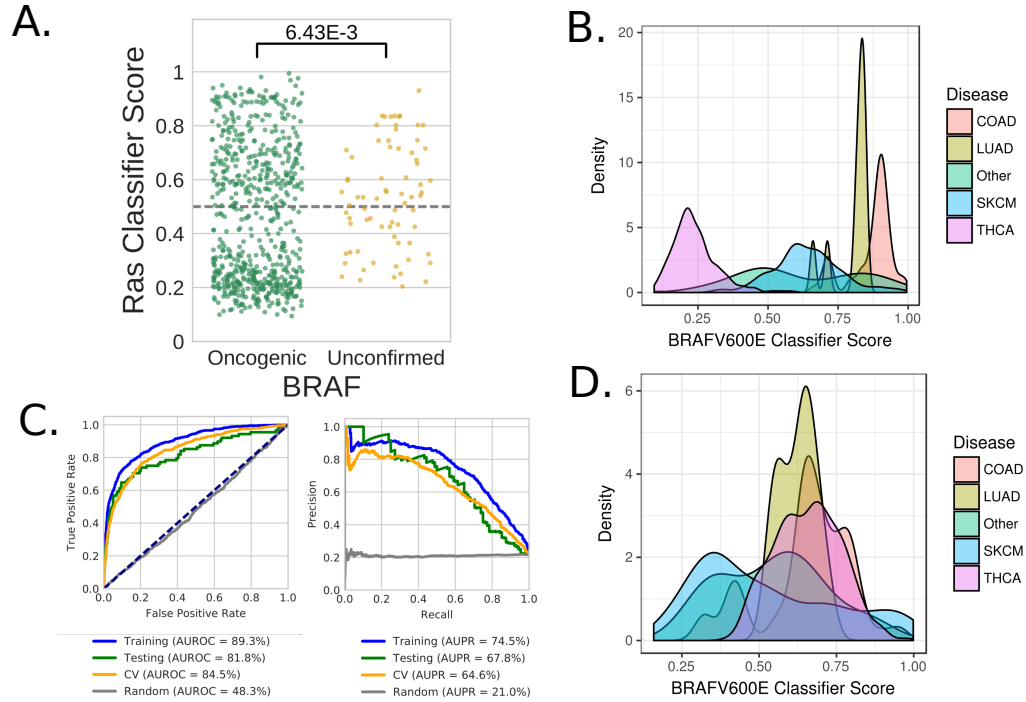


Figure 3.8: Predicting *BRAF* status with the TCGA PanCanAtlas Ras classifier

**(A)** Predictions for tumors with oncogenic or unconfirmed variants in *BRAF* given by the Ras classifier evaluated in Figure 2. **(B)** Ras classifier scores assigned to samples with *BRAF* V600E mutations stratified by cancer type. A score above 0.5 indicates a prediction of activated Ras. **(C)** Ras classifier evaluation after removing THCA and SKCM from training. ROC and PR curves for the Ras classifier without THCA and SKCM does not indicate reduced performance. The grey lines represent predictions made on a shuffled gene expression matrix. **(D)** Ras classifier without THCA and SKCM classify *BRAF* V600E as Ras wildtype in THCA, but not in SKCM.

The cell line evaluation included accurately detecting isogenic lines transfected to express activating *KRAS* mutations and identifying CCLE cell lines with known Ras and *BRAF* mutations. We also demonstrated that CCLE Ras classifier scores were correlated with the drug activity of two MEK inhibitors (Selumetinib and PD-0325901). In

clinical trials, Selumetinib did not increase overall survival in *KRAS* mutant advanced non-small cell lung cancer (NSCLC) patients (201, 202). PD-0325901 also failed to meet efficacy endpoints in *KRAS* mutant NSCLC patients (203). Selumetinib and PD-0325901 have also been tested across many different cancer-types including ovarian, thyroid, skin, hepatocellular, breast, and colon cancers (202, 204–207). Selumetinib has shown promising results in treating children with *NF1* mutant plexiform neurofibromas (208) while PD-0325901 has shown efficacy in treating *NF1* mutant neurofibromas in mice and human-derived malignant peripheral nerve sheath xenografts (190). Furthermore, the classifier automatically learns similar gene coefficients of an 18 gene panel previously curated using a targeted differential expression analysis to predict Selumetinib sensitivity (210). Overall, our results suggest a useful biomarker application to potentially reveal hidden responders that may have otherwise been missed by sequencing.

Our approach to detecting Ras activation is supervised and, as any supervised approach, is penalized by inaccurate labels. We encountered this limitation when detecting *BRAF* mutations in THCA. *BRAF* mutations are known to activate *ERK*, and should not be classified as wild-type Ras (211). Our results suggest that in situations with predicted confounding mutations, it may be best to withhold a cancer-type entirely during training. Withholding such data, as opposed to re-building a new classifier post-hoc that uses *BRAF* V600E mutations as positive examples, may help to prevent a process of classifier-creep in which the classifier is continually expanded to improve metrics. Additionally, it is unclear how to best adjust for hypermutated phenotypes as these tumors are more likely have Ras mutations by chance. Unsupervised or semi-supervised methods to automatically retrieve gene expression signatures may overcome labeling issues and may sidestep some of the difficulties modeling hypermutated tumors by first separating sources of variation.

While mutual exclusivity analyses across pathways drives hypotheses and reveals etiological insights (212, 213), our findings suggest that when multiple mutations occur in Ras pathway genes, tumors exhibit a transcriptional profile associated with increased Ras activity. This is the opposite observation for copy number events as more events outside of *KRAS*, *NRAS*, and *HRAS* appear to confer lower scores, which may either indicate some sort of dosage response counteracting the effects of hyperactivation or alternative events that dampen accurate Ras classification. Furthermore, tumors harboring specific Ras pathway isoforms curated by the PanCanAtlas Pathways AWG are generally predicted to have higher scores than unconfirmed variants.

In conclusion, we presented a machine learning method to predict Ras activity in individual bulk tumors using transcriptomes. Our approach may side-step requirements to profile multiple genomic measurements to detect Ras activation and identify more patients with activated Ras. Our approach can be used as an additional method to improve precision oncology (175). Sub-clonal mutations may also prevent accurate Ras classification by gene sequencing. Training classifiers with single cell RNA-seq data may enable detection rare events and can help to characterize intratumor heterogeneity. As data increase in scale and algorithms are better constructed to model disease heterogeneity, the ability to research downstream responses of pathway misregulation and identify multi-model therapies targeting various vulnerabilities of individual tumors will improve.

### **3.5. Methods**

#### **3.5.1. Contact for reagent and resource sharing**

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Casey S. Greene (csgreene@upenn.edu). The Cancer Genome Atlas will provide instructions on how to access publicly available data.



### 3.5.2. Training machine learning classifiers to detect aberrant gene events

We integrated Illumina RNAseq, multi-center mutation calls (MC3), and GISTIC2.0 copy number threshold calls from The Cancer Genome Atlas (TCGA) PanCanAtlas project to classify aberrant pathway function (214). We downloaded TCGA datasets from the Genome Data Commons (GDC). In total, there were 9,075 tumors that were measured on all three platforms that passed quality control filtering. We subset the gene expression matrix to the 8,000 most variably expressed genes by median absolute deviation (MAD), as genes that do not vary are unlikely to be useful for classification and to reduce training time. We dropped the target genes of interest (e.g. *KRAS*, *NRAS*, *HRAS* or *NF1*) when training the models to prevent the model from potentially relying too heavily on dosage-specific effects of these genes instead of the downstream response to their activation. We also removed the samples with the highest mutation burden to remove potential false positives. We defined these samples based on five standard deviations above the log10 total non-silent somatic mutation count per sample. Because we were interested in a balanced training set based on aberrant gene events, we further filtered samples to include only cancer-types with greater than 15 target gene events and a proportion of negatives to positives no less than 5%.

Using this data, we trained a supervised elastic net penalized logistic regression classifier with stochastic gradient descent (20). Our model is trained on RNAseq gene expression ( $X$ ) to predict gene status ( $y$ ). To control for tumors with a hypermutator phenotype and potential tissue-specific expression patterns, we included cancer-type dummy variables and per sample log10 mutation count in the model as covariates. We defined gold standard gene status using loss of function mutation and deep copy number losses for tumor suppressor genes and gain of function mutations and large copy number gains for oncogenes. For simplicity and to reduce the requirement for

extensive manual curation, we considered any non-silent mutation including insertion-deletions in the gene body or mutations in splice site regions of target genes. For the specific focus of the paper, we integrated gain of function mutation and copy number gains for the oncogenes (*KRAS*, *NRAS*, and *HRAS*), and loss of function and deep copy number losses for the tumor suppressors (*NF1*). For example, if a tumor had a deleterious mutation or copy number amplification in one of these genes, we considered the Ras status equal to one.

The objective of the classifier is to determine the probability a given sample ( $i$ ) has a Ras event given the sample's RNAseq measurements ( $X_i$ ). In order to achieve the objective, the classifier learns a vector of coefficients or gene-specific weights ( $w$ ) that optimize the following penalized logistic function.

$$P(y_i = 1 | X_i) = f(X_i w) = \frac{1}{1 + e^{-w X_i}}$$

$$\text{negative loglikelihood} = L = - \sum_{i=1}^n y_i \log P(y_i = 1 | X_i) + (1 - y_i) \log P(y_i = 0 | X_i)$$

$$w = \text{argmin}(L + \alpha \sum ||w||_l)$$

Where  $\alpha$  and  $l$  are regularization and elastic net mixing hyperparameters that are only active during training, respectively. Using a training set consisting of 90% of the full data set, equally balanced for different proportions of included cancer-types and Ras status, we performed cross validation over the hyperparameter grid:  $l = \{0.15, 0.155, 0.16, 0.2, 0.25, 0.3, 0.4\}$  and  $\alpha = \{0.1, 0.13, 0.15, 0.18, 0.2, 0.25, 0.3\}$ . We used balanced 5-fold cross validation based on the highest cross-validation area under the receiver operating characteristic (AUROC).

We trained the Ras classifier using optimal hyperparameters ( $l = 0.15$  and  $\alpha = 0.1$ ) and assessed performance on training, testing (held out 10% of data) and across 5-fold

cross-validation intervals. In 5-fold cross-validation, the data are partitioned into five even sets (balanced by Ras status and cancer-type). Four of the folds, called training intervals, are used to construct the model. The model is then evaluated on the fifth fold, which is called the evaluation fold. The reported training performance comes from the folds used for training, while the cross-validation performance uses the evaluation fold. Therefore, performance on cross-validation intervals are the predictions reported on the training set samples when they were included in the internal cross-validation evaluation fold.

### 3.5.3. *Evaluating machine learning classifiers*

We evaluated the pan-cancer classifiers in various ways. For every evaluation, we reported the AUROC and area under the precision-recall (AUPR) curve. We also compared gene specific classifiers built using pan-cancer data to classifiers trained independently using only data from individual cancer-types. In these cases, each cancer-type specific model was optimized individually. We compared how the pan-cancer model performed on individual cancer-types compared to individual cancer-type optimizations. Additionally, we cataloged the performance of the Ras classifier to predict *NF1* inactivation in various cancer-types. *NF1* is a tumor suppressor of Ras and we postulated that it would have similar downstream consequences that could be captured by the Ras classifier. Therefore, we performed the same procedure of filtering datasets and training pan and within cancer-type classifiers for *NF1*. We compared these *NF1* evaluations against the Ras classification. Lastly, we evaluated the Ras classifier on predicting aberrant mutations of other genes and variants in the Ras pathway and in two different cell line datasets.

#### 3.5.4. Classifier benchmarking analyses

We determined the robustness of the classifier by evaluating performance under various input features and prediction tasks. We evaluated potential inflation of performance metrics by training a null model on a randomly shuffled input gene expression matrix. We did not shuffle the covariate information or the  $y$  matrix. Performance on the random shuffling of genes, while maintaining the same ratio of Ras mutations, provides insight into how the model would be expected to perform in a scenario lacking Ras activation signal. We also performed the same shuffling and classifier testing procedure as internal negative controls in every pan-cancer model and report ROC/PR curves and AUROC/AUPRs in each figure.

To assess value added in combining mutation and copy number data in the prediction task (altering the  $y$  matrix), we trained pan-cancer classifiers with the same procedure described above to predict Ras mutations and Ras copy number gains separately. The combined model presented here is the same model trained in Figure 3.3. To test the effect of dropping *KRAS*, *HRAS*, and *NRAS* from the model (altering the  $X$  matrix), we trained models with the previously described procedure with the input gene expression matrix without dropping Ras genes. We also tested a classifier after dropping 14 genes from the Expanded RASopathy Panel (215). The genes included *BRAF*, *CBL*, *HRAS*, *KRAS*, *MAP2K1*, *MAP2K2*, *NF1*, *NRAS*, *PTPN11*, *RAF1*, *SHOC2*, *SOS1*, *SPRED1*, and *RIT1*. For the two previous comparisons, we compared the learned gene expression coefficients to the classifier trained in Figure 3.3. For the dropping genes analysis, we added back all dropped genes as zero weights. We also compared the performance of gene expression-only and covariate-only models (altering the  $X$  matrix) to the combined model presented in Figure 3.3. The  $y$  matrix remained the same, but each model was trained on only a subset of the combined  $X$  matrix. The differentially

expressed genes visualized in Figure 3.4H were obtained from the differential expression analysis described below.

#### 3.5.5. *Differential expression analysis*

We performed a differential expression analysis using the limma Bioconductor package (216). We adjusted the model by cancer-type by including cancer-type indicator variables in the limma design matrix. We considered all 9,074 samples and 20,500 genes in this analysis. We zero-one normalized the input matrix by gene prior to fitting with limma.

#### 3.5.6. *Cell line validation*

We applied the Ras classifier to two independent cell line datasets. The first dataset was generated by (180) and was deposited in the Gene Expression Omnibus (217) with the identifier GSE94937. We used the preprocessed form of the data from (180). We also used data from 737 cell lines from the CCLE that had matching RNAseq and mutation data (196). Of these 737, 708 also had variant level annotations. In order to apply the classifier to both cell-line datasets, we z-score normalized gene expression values and subset the data to classifier genes, independently. 177 out of 185 (96%) of the features were in common to classifier genes in both datasets, so we proceeded to make predictions with this subset. In order to apply the predictions, we used the following transformation:

$$s = f(X_i w) = \frac{1}{1 + e^{-wX}}$$

Where  $s$  is the classifier prediction,  $w$  is the gene weights, and  $X$  is the corresponding subset cell line gene expression matrix.

We used the CCLE pharmacologic profiling data, which measured the activity of 24 drugs across 504 CCLE cell lines (CCLE\_NP24.2009\_profiling\_2012.02.20.csv). Data were accessed from <https://portals.broadinstitute.org/ccle/data> (196).

#### *3.5.7. Ras pathway and oncogenicity curation*

We used the PanCanAtlas Pathways Working Group definition of 38 core Ras pathway genes (200). We obtained oncogenicity assignments for mutations in these genes using OncoKB (199) and additional manual curation by the PanCanAtlas Pathways AWG. The manual curation included referencing MutSig (218), hotspot analyses (219), and GISTIC Peaks (214).

#### *3.5.8. Quantification and statistical analyses*

We performed all machine learning model training, testing, and evaluations using scikit learn (version 0.18.1) with python 3.5.2 (131). We processed data using a combination of pandas (version 0.20.3) and dplyr (version 0.7.1) and visualized results using a combination of seaborn (version 0.7.1), ggplot2 (version 2.2.1), and PathwayMapper (220). R packages were run on R version 3.4.0. Please refer to the Key Resources Table and the available GitHub repository (<https://github.com/greenelab/pancancer>) for full software version details (221). We evaluated all classifiers using AUROC and AUPR. The AUROC is a metric describing the overall trade-off between true positive and false positive rates, while the AUPR measures precision against recall for a given classifier. An AUROC of 0.5 constitutes random guessing. We describe specific filtering steps for each analysis in various places in the methods section. We describe overall sample and gene filtering in the section 3.5.2. We discuss additional gene filtering for evaluating all alternative genes in section 3.5.3. We set random seeds in all computational analyses in order to preserve reproducibility. We performed independent t-tests with unequal variances when

comparing classifier scores for curated variants versus variants of unknown significance per Ras pathway gene. We performed the same test comparing CCLE cell line Ras classifier scores for Ras wildtype versus Ras (*KRAS*, *HRAS*, or *NRAS*) mutant samples and for Ras wildtype, *BRAF* wildtype versus Ras wildtype, *BRAF* mutant. Using the up to 388 cell lines with both gene expression and pharmacology data measured, we fit linear regression models comparing drug activity vs. Ras classifier scores for all 24 drugs to Ras wild-type and Ras mutant cell lines individually. Using a Bonferroni adjusted p value ( $0.05 / (24 * 2) = 0.001$ ), we implicated two high correlated drugs (AZD6244 (Selumetinib) and PD-0325901). Selumetinib was tested on 387 cell lines while PD-0325901 was tested on 388 cell lines. We also used a chi square test for proportions of Ras mutations annotated as COSMIC variants in true positives compared to false negatives with a null hypothesis that both sets of samples have the same proportion of COSMIC variants.

#### 3.5.9. *Data and software availability*

All analytical results can be reproduced using the code available at <https://github.com/greenelab/pancancer> (221). Here, we provide instructions to replicate the computing environment, download versioned data, and all scripts to reproduce the entire analysis pipeline. The pipeline is modular and amendable to generate classifiers and predictions for any combination of genes, pathways, and TCGA PanCanAtlas cancer-types.

### 3.6. **Acknowledgements**

We thank Daniel Himmelstein, Jie Tan, and Amy Campbell for helpful code review. This work was funded in part by grants from the Gordon and Betty Moore Foundation (GBMF 4552) to C.S.G. and the National Institutes of Health (R01 NS095411) to Y.S. G.P.W. was supported in part by a training grant from the National Institutes of Health

(T32 HG000046). This work was also supported by the US National Cancer Institute funding of TCGA (U54 HG003273, U54 HG003067, U54 HG003079, U24 CA143799, U24 CA143835, U24 CA143840, U24 CA143843, U24 CA143845, U24 CA143848, U24 CA143858, U24 CA143866, U24 CA143867, U24 CA143882, U24 CA143883, U24 CA144025, P30 CA016672)



## Chapter 4.

### Machine learning derived expression signature predicts TP53 inactivation

Portions of this chapter were originally published as: Knijnenburg, Theo, A., Wang, Linghua, Zimmerman, Michael, T., Chambwe, Nyasha, Gao, Galen, F., Cherniack, Andrew, D., Fan, Huihui, Shen, Hui, Way, Gregory, P., Green, Casey, S., Liu, Yuexin, Akbani, Rehan, Feng, Bin, Donehower, Lawrence A., Miller, Chase, Shen, Yang, Karimi, Mostafa, Chen, Haoran, Kim, Pora, Jia, Peilin, Shinbrot, Eve, Zhang, Shaojun, Liu, Jianfang, Hu, Hai, Bailey, Matthew, H., Yau, Christina, Wolf, Dinse, Zhao, Zhongming, Weinstein, John, N., Li, Lei, Ding, Li, Mills, Gordon B., Laird, Peter, W., Wheeler, David A., Shmulevich, Ilya, The Cancer Genome Atlas Network, Monnat, Raymond, J. Jr., Xiao, Yonghong, Wang, Chen. “*Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas.*” *Cell Reports* 23 (2018) 239-254. doi: 10.1016/j.celrep.2018.03.076.

#### Contributions:

The paper cited above (Knijnenberg et al. 2018) was a large consortium paper. In the paper, I trained and evaluated a machine learning classifier to detect TP53 inactivation. This was the same approach we used in the Ras pathway paper discussed in Chapter 3. My contributions included drafting and editing the TP53 classifier section and generating Figure 5 and Supplementary Figure 6 in the original publication. This chapter only includes the aforementioned section and an abridged introduction.

#### 4.1. Introduction

*TP53* is the most frequently mutated gene in cancer. The gene is intricately involved in many cellular processes, including response to DNA damage (222). The Cancer Genome Atlas (TCGA) has profiled many different datatypes and biological processes across 33 different cancer-types totaling over 10,000 tumors (176). One effort profiled deficiencies in the DNA repair and response pathway, which included the *TP53* mutation landscape (38). Machine learning can be used to detect when tumors express specific

gene expression signatures (37). In the following chapter, we train a logistic regression classifier to detect samples with TP53 loss of function. We demonstrate that certain copy number events phenocopy TP53 inactivation, and we implicate a silent mutation in a *TP53* splice donor site that appears to ablate TP53 function in a dominant negative fashion.

## **4.2. Results**

The loss of TP53 function across many cancer types has significant functional consequences as measured by genomic instability in association with a higher somatic copy number alterations (SCNA) burden and increased HRD scores. Cancer-associated *TP53* mutations may promote these consequences through simple loss of function, as well as by altering transcription or through dominant-negative, gain-of-function mechanisms (222–226). A subset of these consequences can also be phenocopied by other genomic alterations. In order to better predict the consequences of TP53 inactivation and identify potential phenocopies of TP53 loss, we constructed a TP53 classifier that predicts inactivation status from RNA sequencing expression data, adjusted for cancer type and mutation burden, then used this to analyze cancer types with comparable numbers of TP53 alterations. The resulting classifier was highly sensitive and specific (Figure 4.1A), and when trained using PanCanAtlas data, it outperformed individual cancer models in 14 out of 19 cases (Figures 4.2).

Individual weights of the TP53 classifier identified 10 top negative-weighted genes, of which 9 are confirmed TP53 target genes (222) (Figure 4.1B). The remaining gene, *MPDU1*, may have been identified by virtue of being located ~80 kb downstream of *TP53* and thus sensitive to TP53 copy loss. Of note, our classifier was able to predict TP53 deficiency independent of cancer type with a high AUROC (area under the

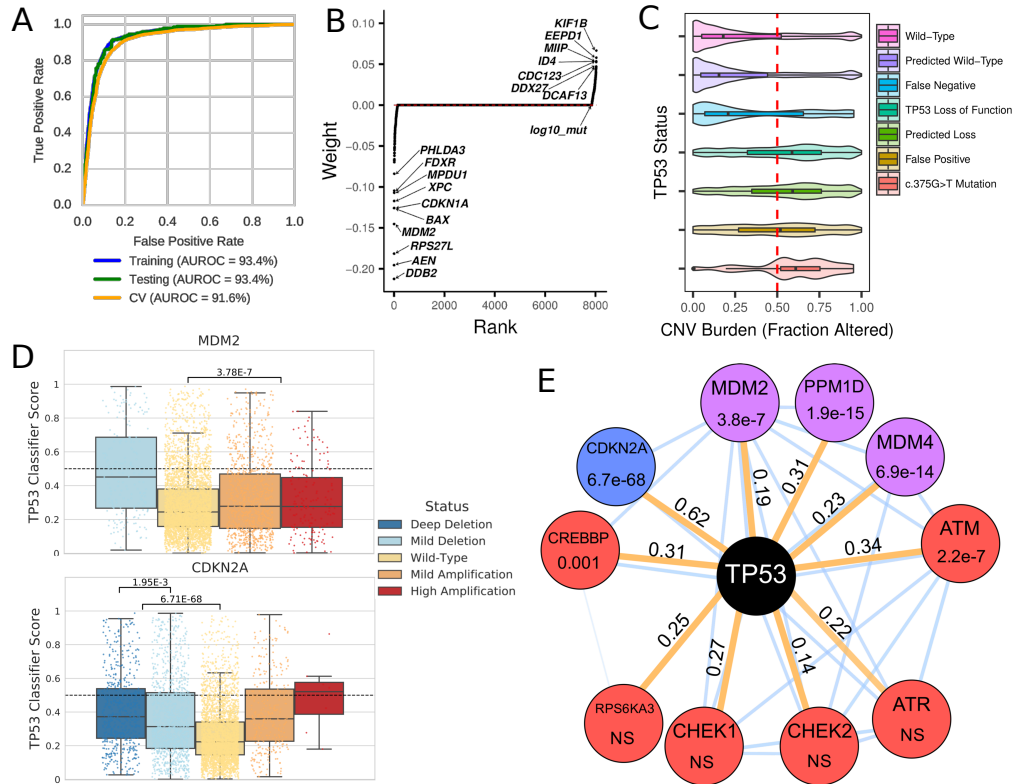


Figure 4.1: Machine learning to predict TP53-inactivating mutations in cancer

**(A)** Robust classifier performance by receiver operating characteristic (ROC) and area under the ROC curve (AUROC). Training data, cross validation assessment, and held out test set (10%) for 19 cancer types were used. **(B)** Model-derived gene weighting. Classifier weights indicate individual gene influence on classification accuracy. Negative weights indicate increased gene expression in TP53 wild-type samples. **(C)** SCNA burden is correlated with known/predicted TP53 status. Plots show SCNA/CNV burden as fraction altered for known or predicted TP53 status. The SCNA profile for TP53 mutation c.375G>T in TP53 exon 4 appears similar to other TP53 loss events. **(D)** SCNA in TP53-interacting genes MDM2 and CDKN2A phenocopies TP53 loss. Results shown are for PanCanAtlas TP53 wild-type samples. **(E)** TP53 network gene alterations phenocopy TP53 deficiency. Mutations were manually curated and selected a priori. All mutation tests including only TP53 wild-type/non-hypermutated cancers are indicated by orange edges. Node color indicates event class (red, mutation; blue, copy-number loss; and purple, copy-number amplification); edge values indicate Cohen's d effect size. Thin blue edges indicate predicted interactions from the STRING database. NS is "not significant" with  $p > 0.005$ .

receiver operating characteristic curve; 0.94), and in samples initially removed from training. These included cancer types with few TP53 events (THCA and UVM), as well as those dominated by TP53 events (OV and UCS) (Figures 4.2C – 4.2F).

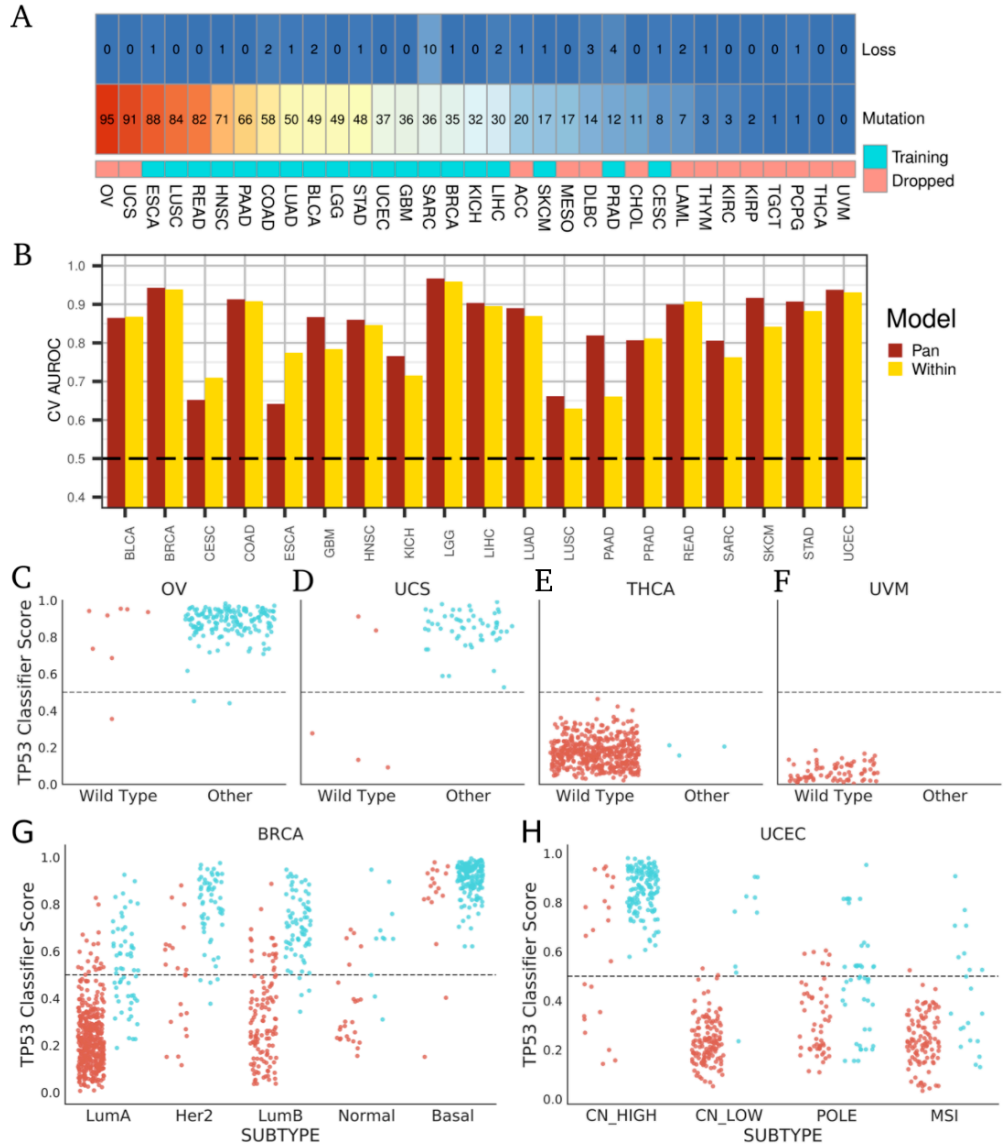


Figure 4.2: Pan-Cancer TP53 classifier scores by cancer-type

**(A)** Cancer types display a broad distribution of TP53 deep copy number loss and deleterious mutation events. Loss represents  $GISTIC \leq -2$ . The bottom bar indicates cancer types included during model training (teal fill). **(B)** Cancer-specific models may identify tissue-level effects. CV AUROC for a pan-cancer model (Pan) compared to models optimized individually for each cancer-type (Within). **(C-F)** Pan-cancer TP53 classifier applied to cancer-types with imbalanced class sizes and not used in training. “Other” tumors have either TP53 mutation or deep copy loss. **(G-H)** Pan-cancer TP53 classifier applied to BRCA and UCEC stratified by subtype. Samples in red indicate wild-type TP53 and samples in blue indicate TP53 loss of function (mutation or deep copy loss).

The classifier was also able to distinguish *TP53* mutant from wild-type BRCA and UCEC, with nearly all basal-subtype BRCA cancers predicted to be *TP53* deficient (Figures 4.2G and 4.2H). We used analogous approach has been used to predict RAS pathway activation in PanCanAtlas cancers (37).

The classifier enabled the identification of phenocopying mutations both in *TP53* and in other functionally related genes. Consistent with previous pan-cancer analyses (227), we observed that predicted *TP53* loss-of-function samples, including cancers with synonymous *TP53* c.375G>T mutation, had an increased SCNA burden when compared with wild-type samples (Figure 4.1C). This synonymous mutation may act by altering a splice donor to produce alternatively spliced transcripts that compromise *TP53* function (228, 229). Samples with c.375G>T or c.375G>A mutations were also enriched for a 200-base pair truncation in exon 4 when compared with wild-type *TP53* samples (Figure 4.3; OR (odds ratio) = 61.9,  $p < 2.2e-16$ ). This mutation/truncation pairing was previously observed in a pancreatic cancer cell line and as a SNP (rs55863639) likely pathogenic for Li-Fraumeni syndrome (230).

Significantly increased classifier scores were also noted for cancers with *MDM2* copy-number amplification and *CDK2NA* copy-number deletion in an analysis including only non-hypermuted cancers without deleterious *TP53* mutation (Figure 4.1D). We had observed a copy-number dosage effect for *CDK2NA* copy-number deletions, where loss of the *CDKN2A*-encoded P14ARF protein can phenocopy *TP53* alterations (231). Among eight other tested genes, *MDM4* and *PPM1D* copy-number amplification and *ATM* and *CREBBP* gene mutations were associated with increased *TP53* classifier scores, while *ATR*, *CHEK1/2*, or *RP6SKA3* mutations were not (Figure 4.1E). These results suggest the general utility of this approach, even in circumstances where a diversity of molecular events and potential downstream consequences might occur.

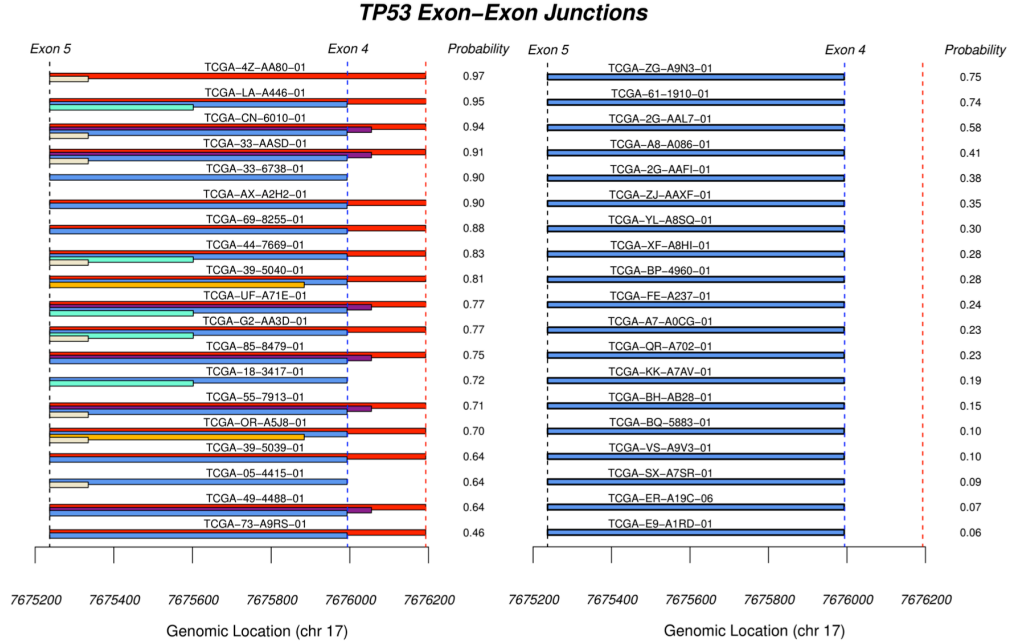


Figure 4.3: *TP53* exon-exon junctions for samples with c.375G>T mutations in *TP53*

Shown are *TP53* exon-exon junctions on chromosome 17 between canonical exons 4 and 5. All samples are annotated as wild-type *TP53*. The horizontal bars indicate different exon-exon junctions in this region. The same sample can have multiple observed junctions. The blue bar represents the canonical exon-exon junction between exon 5 (black dotted line) and exon 4 (blue dotted line). The red dotted line indicates a junction event occurring exactly 200 base pairs upstream of the exon 4 splice donor that corresponds to the observation in Suwa et al., 1994 (232). Also listed are each sample's expression classifier probability of *TP53* loss. Left – The 19 samples with the c.375G>T mutation resulting in ablation of the traditional splice donor site on exon 4 (blue dotted line). c.375G>T is the only *TP53* mutation in these samples. We observe many more alternative splice forms in these samples, including an enrichment of a previously reported splicing event at the red site. Right – 19 randomly selected wild-type samples without the c.375G>T mutation show only canonical *TP53* transcript splicing at this site.

### 4.3. Methods

#### 4.3.1. In-silico prediction of *TP53* inactivation

We trained a classifier to use RNA-seq expression data to predict *TP53* status.

Specifically, we trained a logistic regression classifier with an elastic net penalty using the sci-kit learn implementation of stochastic gradient descent (131). The labels ( $y$ ) for the supervised task included samples with MC3 annotated deleterious *TP53* mutations (samples with silent mutations were considered *TP53* wild-type) and samples with *TP53*

deep copy number loss as predicted by the GISTIC2.0 algorithm (214, 233). We included cancer-types in the model that had greater than 15 samples in each class and between 5% and 95% of samples in both classes and removed all others (see Figure 4.2A). The features ( $X$ ) consisted of the 8000 most variably expressed genes by median absolute deviation (MAD). We dropped expression of *TP53* itself from the features to prevent the model from relying on the target gene. MAD genes were z-scored and concatenated with binarized dummy variables for all cancer types and mutation burden (total log<sub>10</sub> mutation count) to adjust for potential confounding factors. To reduce the effect of mutation burden confounding, we also removed the outlier samples with the most extreme hypermutation phenotypes (greater than 5 standard deviations above the mean log<sub>10</sub> mutation count). The goal of the classification scheme was to determine the weights ( $w$ ) that minimize the objective function described in section 3.5.2.

We selected optimal hyperparameters by balanced 5-fold cross validation with the goal of inducing a sparse solution. We also used a balanced 10% held out set to test the performance of the classifier on data never used for training or hyperparameter optimization. We fit the final model on the remaining 90% of the data and report performance using receiver operating characteristic (ROC) curves and area under the ROC curve (AUROC) metrics.

We manually selected an a priori set of genes known to interact with TP53 for our phenocopying experiment (Lawrence Donehower, personal communication). We tested *MDM2*, *MDM4*, and *PPM1D* amplifications, *CDKN2A* deletions, and *ATM*, *ATR*, *CHEK1*, *CHEK2*, *CREBBP*, and *RPS6KA3* mutations. For the copy number tests, we included both deep and shallow alterations in the altered set compared to tumors with wild-type profiles only. We removed tumors with deleterious *TP53* mutations or deep copy number loss ( $n = 4,037$ ). From the remaining 5,629 tumors, we removed 219 hypermutated

tumors leaving an analytic set of 5,410 tumors. We performed independent t-tests and calculated Cohen's D effect sizes comparing the assigned TP53 classifier scores for wild-type against altered tumors. We considered variants significant if they were less than a Bonferroni adjusted p value ( $p > 0.005$ ). We visualized the results in a network diagram presented in Figure 4.1E. The underlying interaction network was downloaded from the STRING database (version 10.5). The thickness of edges in the STRING network display interaction confidence and were generated by experimental data. Note that there are no direct interaction edges between *RPS6KA3* and *TP53* and *PPM1D* and *TP53*. We provide materials under an open source license to reproduce and expand upon this analysis at <https://github.com/greenelab/pancancer>.

#### **4.4. Conclusions**

*TP53* is the most mutated gene in cancer and is central to many essential tumor suppressing processes, including the coordination of the response to DNA damage (224). However, it remains unclear how *TP53* mutations alter the molecular state of tumors, and *TP53* mutations have many unknown downstream consequences. Therefore, we sought to develop an algorithm that can detect when a tumor has aberrant *TP53*. Using TCGA Pan Cancer data, we trained an elastic net penalized logistic regression classifier to detect when *TP53* is misregulated. Performance on the training, cross validation, and held out test set was highly sensitive and specific. We also observed that the Pan Cancer model outperformed the within-disease type models in 14 out of 19 cases. In 3 of the 5 cases where the within-disease type model was better, we observed only marginal gains. However, in 2 of the cases, esophageal (ESCA) and cervical (CESC) cancer types, performance increased substantially. This result could suggest that TP53 aberrations play a tissue specific role in both diseases. For example,



results for CESC could indicate a signature of inactivation specific to human papillomavirus infection (234).

The elastic net penalty induced sparsity in the features, selecting only 319 genes. Many of these genes are well-known regulators and targets of *TP53* including *MDM2* and *CDKN1A*, well known apoptosis associated genes including *AEN* and *BAX*, genes associated with homologous recombination (*EEPD1*), and cell cycle related genes such as *CDC123*. The genes associated with negative weights represent genes that are commonly upregulated in *TP53* wild-type tumors while genes with positive weights are genes that are upregulated in *TP53* aberrant tumors.

We observed that one annotated silent mutation, c.375G>T, occurred 27 times (19 times it was the only *TP53* mutation in a non-hypermuted sample) across all Pan Cancer samples and was consistently predicted to have TP53 loss of function. We also observed a c.375G>A silent mutation 6 times, which was also predicted deleterious. These mutations occur in the last nucleotide on the 3' end of TP53 exon 4. This location is a SNP (rs55863639) that has been previously associated with Li Fraumeni syndrome (235, 236). While the mutation does not alter the threonine amino acid, c.375G>T was observed to impact splicing in a pancreatic cell line (232). Therefore, we were able to identify a TP53 loss of function phenocopying variant that was previously annotated as a silent variant. This approach can be extended to other genes and pathways to identify variants and patients who may harbor specific variants that would have been missed by alternative means.

## Chapter 5.

### **Comprehensive cross-population analysis of high-grade serous ovarian cancer supports no more than three subtypes**

This chapter was originally published as: Way, Gregory, P., Rudd, James, Wang, Chen, Hamidi, Habib, Fridley, Brooke, L., Konecny, Gottfried, E., Goode, Ellen, L., Greene, Casey, S., Doherty, Jennifer, A. “*Comprehensive cross-population analysis of high-grade serous ovarian cancer supports no more than three subtypes.*” G3: Genes, Genomes, Genetics 6 (2016) 4097 – 4103. doi: 10.1534/g3.116.033514

Conceptualization: G.P.W., J.R., C.S.G., J.A.D.; Methodology: G.P.W., C.S.G., J.A.D.; Software: G.P.W.; Investigation: G.P.W.; Writing – Original Draft: G.P.W., J.R., C.S.G., J.A.D.; Writing – Review and Editing: G.P.W., J.R., C.W., H.H., B.L.F., G.E.K., E.L.G., C.S.G., J.A.D.; Resources: C.W., H.H., B.L.F., G.E.K., E.L.G.; Visualization: G.P.W.

#### **Contributions:**

For the paper, I performed all analyses and wrote the full manuscript. The additional authors contributed as stated above.

#### **5.1. Abstract**

Four gene expression subtypes of high-grade serous ovarian cancer (HGSC) have been previously described. In these early studies, a fraction of samples that did not fit well into the four subtype classifications were excluded. Therefore, we sought to systematically determine the concordance of transcriptomic HGSC subtypes across populations without removing any samples. We created a bioinformatics pipeline to independently cluster the five largest mRNA expression datasets using k-means and nonnegative matrix factorization (NMF). We summarized differential expression patterns to compare clusters across studies. While previous studies reported four subtypes, our cross-population comparison does not support four. Because these results contrast with previous reports, we attempted to reproduce analyses performed in those studies. Our

results suggest that early results favoring four subtypes may have been driven by the inclusion of serous borderline tumors. In summary, our analysis suggests that either two or three, but not four, gene expression subtypes are most consistent across datasets.

## **5.2. Introduction**

Invasive ovarian cancer is a heterogeneous disease typically diagnosed at a late stage, with high mortality (237). The most aggressive and common histologic type is HGSC (238), which is characterized by extensive copy number variation and *TP53* mutation (239). Given the genomic complexity of these tumors, mRNA expression can be thought of as a summary measurement of these genomic and epigenetic alterations, to the extent that the alterations influence gene expression in either the cancer or stroma.

Four gene expression subtypes with varying components of mesenchymal, proliferative, immunoreactive, and differentiated gene expression signatures have been reported in all studies of HGSC to date (239–243). Two of these studies also observed survival differences across subtypes (240, 243). Tothill et al. (2008) first identified four HGSC subtypes (as well as two other subtypes that largely included low-grade serous and serous borderline tumors) in an Australian population using k-means clustering (240). Later, The Cancer Genome Atlas (TCGA) used NMF and also reported four subtypes that were labeled as: “mesenchymal,” “differentiated,” “proliferative,” and “immunoreactive” (239). The TCGA group also applied NMF clustering to the Tothill data and observed similar subtypes (239). Konecny et al. (2014) applied NMF to cluster an independent set of HGSC samples and reported four subtypes, which they labeled as C1–C4 (243). These subtypes were similar to those in the TCGA, but a subtype classifier trained on these subtypes better differentiated survival in their own data, data from TCGA, and Bonome et al. (2008).

Despite the extensive research in the area, work to date has several limitations. In both the TCGA and Tothill studies, 8–15% of samples were excluded from analyses. A reanalysis of the TCGA data showed that over 80% of the samples could be assigned to more than one subtype (244). In more recent TCGA analyses by the Broad Institute Genome Data Analysis Center (GDAC) Firehose initiative, with the largest number of HGSC cases evaluated to date ( $n = 569$ ), three subtypes fit the data better than four (245, 246). This uncertainty in HGSC subtyping led us to determine if four homogeneous subtypes exist across study populations.

Our goal is to rigorously assess the number of HGSC subtypes. We reanalyze data from the five largest independent studies to date (and add an analysis of our own collection of samples) using a standardized bioinformatics pipeline. We apply k-means clustering as well as NMF to each population and do not remove “hard-to-classify” samples, as was done in previous studies (239, 240). We perform independent analyses within each dataset and compare subtyping results across studies. We summarize each subtype’s expression patterns using moderated t-score vectors and comprehensively characterize correlations between subtypes across populations. This method contrasts with earlier reanalyses that pooled HGSC datasets together to identify subtypes (242). We sidestep gene expression platform or dataset biases, which could affect clustering if under or overcorrected, by comparing dataset- and subtype-specific summary statistics instead of pooling raw gene expression data.

Our cross-population comparative analysis does not support the conclusion that four HGSC subtypes exist; rather, the data more strongly support an interpretation that there are either two or three subtypes. We show that the support for four subtypes observed in TCGA’s reanalysis of the Tothill data (239) is lost when serous borderline tumors, which have very different genomic profiles and survival compared to HGSC (241, 247), are

excluded before clustering. Our work also highlights the impact that a single study can have on the trajectory of subtyping research and suggests the importance of periodic histopathologic review and rigorous reanalysis of existing data for cross-study commonalities.

### **5.3. Methods**

#### *5.3.1. Data inclusion*

We used data from the R package, *curatedOvarianData* (248), and our own dataset (“Mayo”). A subset of these data has been published previously (GSE53963) (243), but the present dataset (GSE74357) contains 343 more samples. Briefly, these criteria selected HGSC samples from studies including at least 130 cases assayed on standard microarrays. We included only HGSC and high-grade endometrioid samples, which are molecularly similar to HGSC (249) as identified by study-specific pathological review. Data from the new Mayo HGSC samples, as well as other samples with mixed histologies and grades, for a total of 528 additional ovarian tumor samples, were deposited in NCBI’s Gene Expression Omnibus (GEO) (217); these data can be accessed with the accession number GSE74357 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74357>). All study participants provided written informed consent, and this work was approved by the Mayo Clinic and Dartmouth College Institutional Review Boards.

After applying the unified inclusion criteria, our final analytic datasets included: TCGA (n = 499) (239, 245); Mayo (n = 379; GSE74357) (243); Yoshihara (n = 256; GSE32062.GPL6480) (250); Tothill (n = 242; GSE9891) (240); and Bonome (n = 185; GSE26712) (241) (Table 5.1). We restricted analyses to the 10,930 genes measured successfully in all five populations.

	<i>TCGA</i>	<i>Mayo</i>	<i>Yoshihara</i>	<i>Tothill</i>	<i>Bonome</i>
<i>GEO</i>		GSE74357	GSE32062	GSE9891	GSE26712
<i>Platform</i>	HGU1133	Agilent 4x44K	Agilent 4x44K	HGU1133	HGU1133
<i>Population</i>	USA	USA	Japan	Australia	USA
<i>Original n =</i>	578	528	260	285	195
<i>Analytic n =</i>	499	379	256	242	185
<i>Age [Mean (SD)]</i>	60.0 (11.6)	62.9 (11.3)	NA	60.3 (10.3)	61.5 (11.9)
<i>Stage:</i>					
<i>I</i>	10 (2%)	7 (3%)	0 (0%)	11 (5%)	0 (0%)
<i>II</i>	17 (4%)	11 (3%)	0 (0%)	8 (4%)	0 (0%)
<i>III</i>	351 (80%)	275 (73%)	202 (79%)	178 (83%)	146 (80%)
<i>IV</i>	63 (14%)	86 (23%)	54 (21%)	17 (8%)	36 (20%)
<i>Grade:</i>					
<i>2</i>	55 (12%)	3 (1%)	130 (51%)	80 (37%)	NA
<i>3</i>	386 (88%)	376 (99%)	126 (49%)	134 (63%)	NA
<i>Debulking</i>					
<i>Optimal</i>	325 (74%)	287 (76%)	101 (39%)	132 (62%)	89 (49%)
<i>Suboptimal</i>	116 (26%)	87 (23%)	155 (61%)	82 (38%)	93 (51%)

Table 5.1: Characteristics of the populations included in the five HGSC data sets

### 5.3.2. Clustering

We performed independent clustering within each dataset to avoid potential biases from different platforms or studies. We identified the 1500 genes with the highest variance from each dataset and used the union of these genes ( $n = 3698$ ) for clustering. We performed clustering within each dataset using each potential  $k$  from 2 to 4 clusters. We performed  $k$ -means clustering in each population using the R package “cluster” (version 2.0.1) (251) with 20 initializations. We repeated these analyses using NMF in the R package “NMF” (version 0.20.5) (52) with 100 different random initializations for each  $k$ . As done in prior studies, we calculated cophenetic correlation coefficients to select appropriate  $k$  for each dataset after NMF clustering with 10 consensus runs. The cophenetic correlation identifies appropriate solutions and tends to decrease with increasing  $k$  unless a more accurate solution is observed at a larger  $k$ .

#### *5.3.3. Identification of analogous clusters within and across studies*

We performed significance analysis of microarray (SAM) (252, 253) analysis on all clusters from each study using all 10,930 genes. This resulted in a cluster-specific moderated  $t$  statistic for each of the input genes (254). To summarize the expression patterns of all 10,930 genes for a specific cluster in a specific population, we combined gene-wise moderated  $t$  statistics into a vector of length 10,930. We repeated the SAM analysis using only the MAD subset genes and the results were similar. The TCGA subtype labels have become widely used in the field. To generate comparable labels across  $k$  and across studies, we mapped our TCGA subtype assignments back to the original TCGA labels to define reference clusters at  $k = 4$  (that is, mesenchymal-like, proliferative-like, etc.). Clusters in other populations that were most strongly correlated with the TCGA clusters were assigned the same label.

#### *5.3.4. Clustering analysis of randomized data*

Any clustering procedure is expected to induce strong correlational structure across clusters within a dataset, even if there is no true underlying structure. However, if there is no true underlying structure, clusters across datasets are not expected to be correlated. To assess this, we used the same datasets but shuffled each gene's expression vector to disrupt the correlative structure. We performed within- and cross study analyses of cluster identification using this set of data that were parallel to those performed using the nonrandomized data.

#### *5.3.5. Assessing the reproducibility of single population studies*

We compared our sample assignments at  $k = 2-4$  to the four subtypes reported in the Tothill, TCGA, and Konecny publications (239, 240, 243). Because the labels that were assigned in TCGA's reanalysis of the Tothill data were not available, we performed

NMF consensus clustering of Tothill's data without removing low malignant potential (LMP) samples in order to generate labels for comparison.

#### *5.3.6. Data availability*

We provide software under a permissive open source license to download the required data and reproduce our analyses (255). Analyses were run in a Docker container, allowing the computing environment to be recreated (142). Our Docker image can be pulled from: [https://hub.docker.com/r/gregway/hgsc\\_subtypes/](https://hub.docker.com/r/gregway/hgsc_subtypes/). This allows interested users to freely download the software, reproduce the analyses, and then build on this work. All data used in this analysis is publicly available including data we generated (accessible under GEO accession GSE74357).

### **5.4. Results**

#### *5.4.1. Clustering*

To visually inspect the consistency and distinctness of clusters, we compared sample-by-sample correlation heatmaps. For  $k = 2-4$  within each study, we observed high sample-by-sample correlations within clusters and relatively low sample-by-sample correlations across clusters (Figure 5.1). Clustering results using NMF were similar to  $k$  means results (Figure 5.2).

#### *5.4.2. Correlation of cluster-specific expression patterns*

Across datasets, we observed strong positive correlations of moderated  $t$  score vectors between analogous clusters in TCGA, Tothill, Mayo, and Yoshihara (Figure 5.3 and Table 5.2). However, clustering of the Bonome data did not correlate strongly with clusters identified in the other datasets (Table 5.2). We believe that we were unable to assign parallel subtypes in Bonome because of either RNA contamination or inappropriate grading assignments. However, more work is required in order to identify exactly why we were unable to classify.



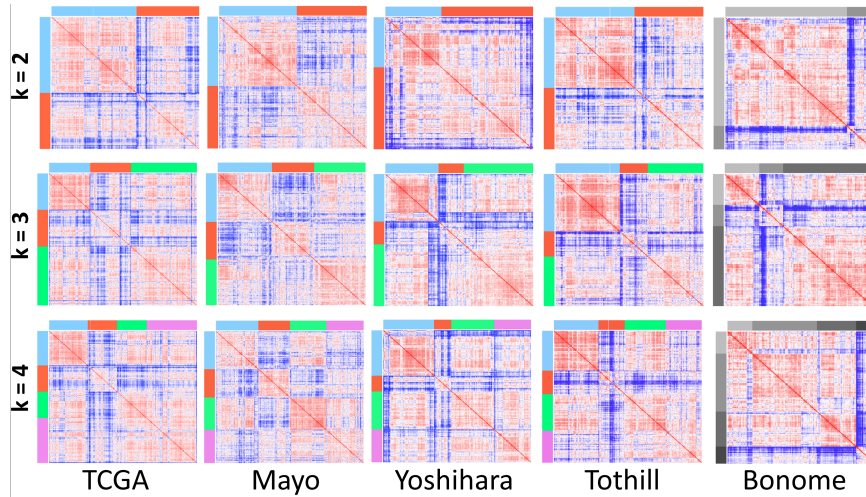


Figure 5.1: Sample by sample Pearson correlation matrices across HGSC populations

Top panel:  $k = 2$ . Middle panel:  $k = 3$ . Bottom panel:  $k = 4$ . The color bars are coded as blue for cluster 1, red for cluster 2, green for cluster 3, and purple for cluster 4. In the matrices, red represents high correlation, blue low correlation, and white intermediate correlation. The scales are slightly different in each population because of different correlational structures. The clusters in the Bonome study are depicted in grey scale because in cross-population analyses to identify analogous clusters, those from Bonome did not correlate with those observed in the four other studies.

In contrast to our analyses, which independently cluster data from each study, Konecny et al. (2014) assigned subtypes to the Bonome data by applying a Predictive Analysis of Microarray (PAM) (63) to their own subtypes to define reduced, subtype-specific predictive gene lists. They then assigned Bonome samples based on the highest Spearman correlation against subtype centroids (243). To assess our analytical approach, we performed an analysis using randomized data. This showed that within-population correlation structure was induced by clustering, but structure between populations was not (Figure 5.4). The off-diagonals in this figure are close to, but not exactly, zero. Permutation induces more independent features than in real gene expression data and therefore may produce much lower correlations if structure is present in real data. Comparing Figure 5.2 with Figure 5.4, we observed much higher

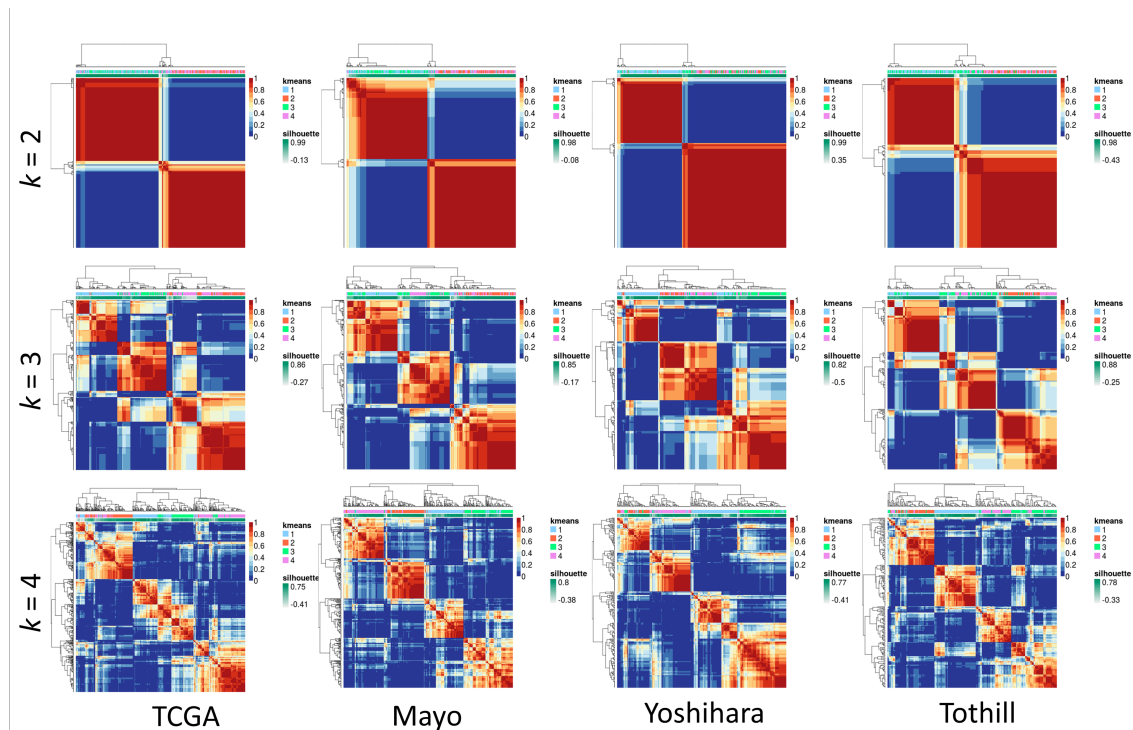


Figure 5.2: NMF consensus matrices for HGSC datasets when  $k = 2$ ,  $k = 3$ , and  $k = 4$

The first track represents cluster membership for  $k$  means clusters and the second track represents silhouette widths. Note that NMF clusters are not ordered in the same way as the  $k$  means clusters.

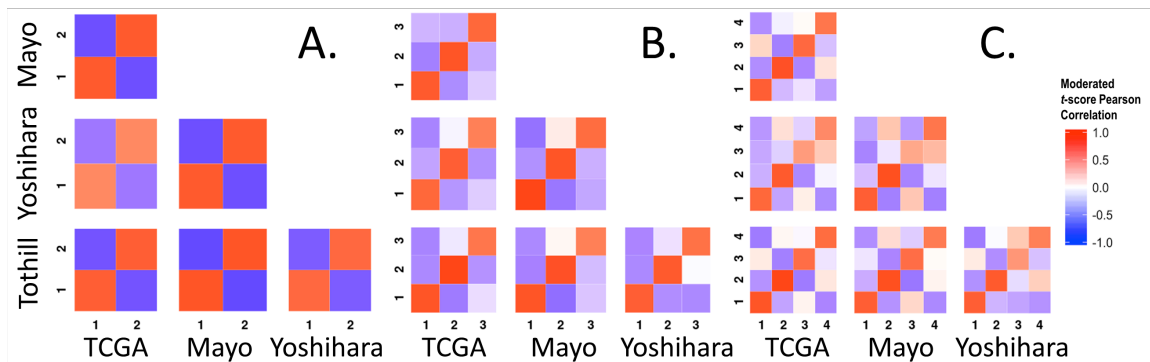


Figure 5.3: Pearson correlation heatmaps reveal consistency across HGSC datasets

**(A)** Correlations across datasets for  $k$  means  $k = 2$ . **(B)** Correlations across datasets for  $k$  means  $k = 3$ . **(C)** Correlations across datasets for  $k$  means  $k = 4$ . TCGA, The Cancer Genome Atlas.

correlation across datasets (Figure 5.2), which was lost after randomization (Figure 5.4). For example, for  $k = 2$ , the TCGA and Mayo cluster correlations for analogous clusters was high (top left panel in Figure 5.3). Conversely, the same relationship in randomized data (second row, first column panel in Figure 5.4) showed correlations near zero. This indicates that the high correlations observed across datasets in Figure 5.3 are induced by similar underlying structure in the data.

	<i>Cluster 1</i>	<i>Cluster 2</i>	<i>Cluster 3</i>	<i>Cluster 4</i>
$k = 2$	0.62 – 0.81	0.62 – 0.81	NA	NA
$k = 3$	0.77 – 0.85	0.80 – 0.90	0.65 – 0.77	NA
$k = ^a$	0.77 – 0.85	0.83 – 0.89	0.51 – 0.76	0.61 – 0.75
<i>Bonome</i> $k = 2$	-0.08 – 0.24	-0.08 – 0.24	NA	NA
<i>Bonome</i> $k = 3$	0.45 – 0.46	-0.02 – 0.12	0.22 – 0.42	NA
<i>Bonome</i> $k = 4$	0.50 – 0.57	-0.04 – 0.04	0.13 – 0.29	0.26 – 0.43

Table 5.2: SAM moderated  $t$  score vector Pearson correlations between analogous clusters across populations

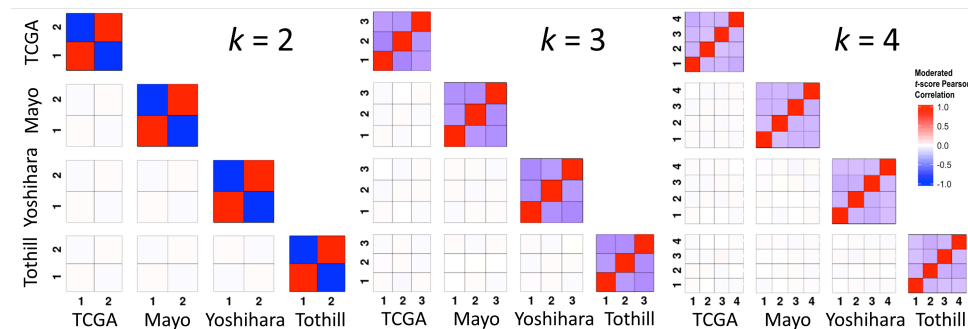


Figure 5.4: Pearson correlation heatmaps of randomly shuffled HGSC datasets

The within dataset correlations are artificially induced because the clustering algorithm will find clusters even without true underlying structure. However, the across dataset clusters are not correlated in the randomized data indicating that the results we observe in Figure 1 are not artifacts of the clustering algorithm.

Across studies, positive correlations between analogous clusters and negative correlations between nonanalogous clusters were stronger for clusters identified when  $k = 2$  and  $k = 3$  than when  $k = 4$  (Figure 5.3), with comparable statistical precision. These cross-population comparisons suggested that two and three subtypes fit HGSC gene expression data more consistently than the four widely accepted subtypes.

Within each population, clusters identified by NMF were similar to those identified using k-means clustering (Figure 5.5), suggesting that these results were independent of clustering algorithm. With NMF, both positive and negative correlations were stronger for  $k = 2$  and  $k = 3$  than for  $k = 4$ . Across  $k = 3$  and  $k = 4$ , correlations were strongest for clusters 1 and 2.

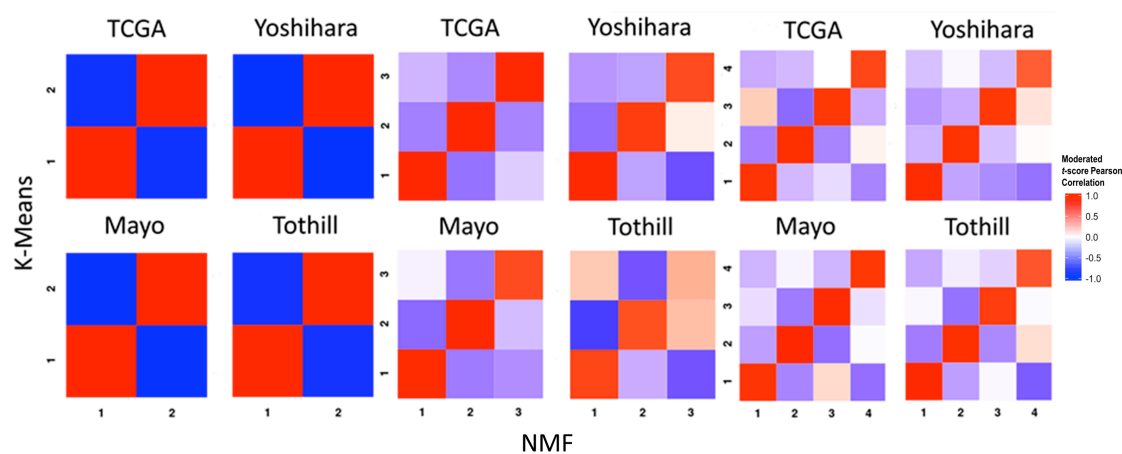


Figure 5.5: Pearson correlations comparing k means and NMF clustering HGSC subtypes

Within dataset results are shown for both methods when setting each algorithm to find 2, 3, and 4 clusters. NMF, nonnegative matrix factorization; TCGA, The Cancer Genome Atlas.

#### 5.4.3. Comparison with previously-identified HGSC clusters

Our clustering results for the Tothill, TCGA, and Mayo datasets were highly concordant with the clustering described in the original publications (239, 240, 243) as

evidenced by the high degree of consistent overlap in sample assignments to the previously-defined clusters (Table 5.3). Our cross-study cluster 1 was mostly mapped to the “Mesenchymal” label from TCGA, “C1” from Tothill, and “C4” from Mayo. This cluster was the most stable in our analysis within all datasets, across  $k = 2, 3$ , and 4, and across clustering algorithms. Cross-study cluster 2, which was also observed consistently, was most similar to the “Proliferative” label from TCGA, “C5” from Tothill, and “C3” from Mayo. Cross-study cluster 3 for  $k = 3$  was associated with both the “Immunoreactive” and “Differentiated” TCGA labels, “C2” and “C4” in Tothill, and “C1” and “C2” in Mayo.

		$k = 2$			$k = 3$				$k = 4$			
		1	2		1	2	3		1	2	3	4
TCGA	Mes	98	1		98	1	0		97	1	0	1
	Pro	7	127		2	111	21		4	85	5	40
	Imm	93	2		20	0	75		12	0	80	3
	Dif	68	60		11	11	106		12	0	3	113
	NC	21	22		6	16	21		5	13	12	13
Tothill	C1	78	0		77	1	0		74	1	3	0
	C2	39	5		22	0	22		0	0	42	2
	C3	1	5		0	0	6		0	0	0	6
	C4	0	44		0	3	41		0	1	1	42
	C5	0	35		0	35	0		0	34	1	0
	C6	0	2		0	2	0		0	2	0	0
Konecny	NC	11	22		6	5	22		0	5	14	14
	C1	36	6		16	0	26		7	0	29	6
	C2	21	39		13	16	31		12	9	6	33
	C3	2	41		2	36	5		3	31	0	9
	C4	26	0		26	0	0		25	0	1	0
	NA	114	94		82	56	70		62	41	57	48

Table 5.3: Distributions of sample membership in the clusters identified in our study compared to the original cluster assignments in the TCGA, Tothill, and Konecny studies

For analyses where  $k = 4$ , the third cluster was associated with “Immunoreactive”, “C2,” and “C1,” while the fourth cluster was associated with “Differentiated,” “C4,” and “C2” for TCGA, Tothill, and Mayo, respectively.

#### 5.4.4. *Meta-research into previous HGSC subtyping studies*

Each of the publications that only considered high-grade samples (239, 243) found clustering coefficients consistent with  $k = 2$ ,  $k = 3$ , and  $k = 4$ . Nevertheless, each publication concludes the existence of four subtypes, while our cross-population analysis suggested that two or three clusters fit HGSC data better than four clusters. The only results in previous studies that contradicted this work were from TCGA’s reanalysis of the Tothill data. According to Figure S6.2 in the TCGA paper, the reanalysis included serous borderline tumors (i.e., tumors with low malignant potential) ( $n = 18$ ). The inclusion of these tumors in the TCGA HGSC reanalyses was done even though, in the original Tothill paper, the serous borderline tumors had a unique gene expression pattern and clustered entirely in a group labeled “C3.”

To assess the extent to which serous borderline tumors inclusion drove the TCGA reanalysis results, we reproduced TCGA’s reanalysis of the Tothill dataset, including the serous borderline tumors ( $n = 18$ ); we indeed observed that the cophenetic correlation is higher for  $k = 4$  than  $k = 3$  (Figure 5.6A). However, when we appropriately removed these serous borderline tumors, we observed an increase in the  $k = 3$  cophenetic correlation (Figure 5.6B). The results that support four subtypes were generated during clustering of HGSC and serous borderline tumors combined. Subtyping analyses of HGSC alone reveal less than four subtypes.

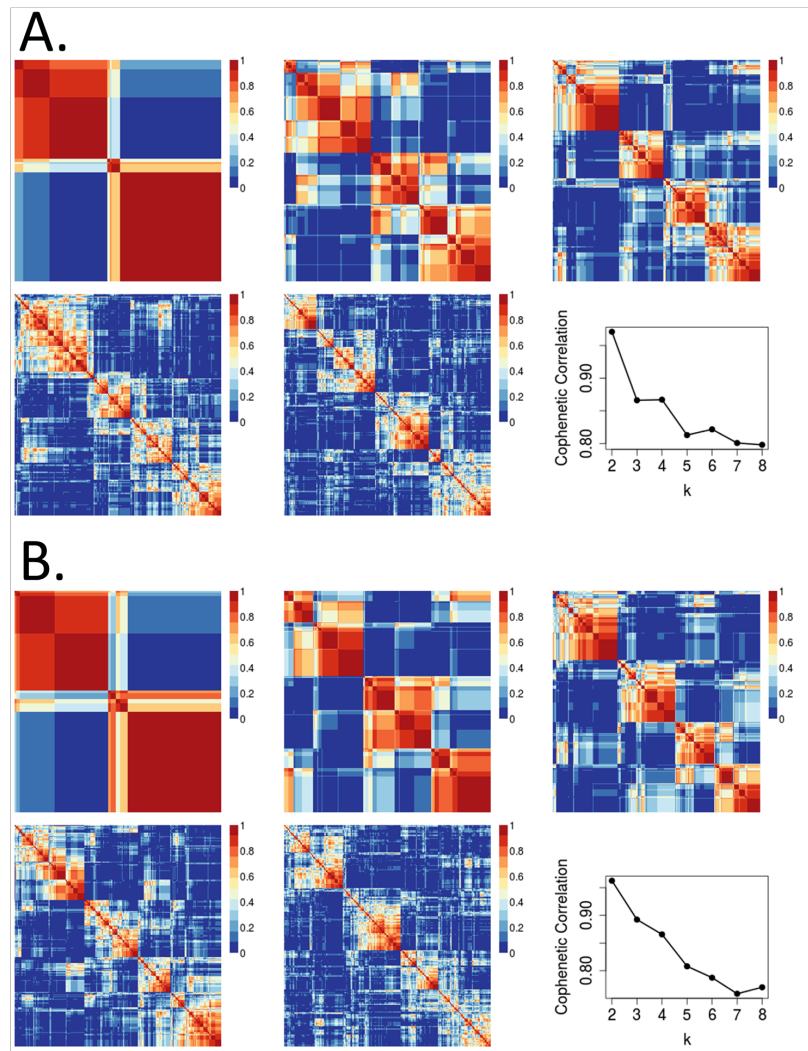


Figure 5.6: Comparing NMF consensus clustering in the Tothill dataset

**(A)** Tothill dataset (n = 260) with borderline samples (n = 18) not removed prior to clustering. **(B)** Tothill dataset with borderline samples removed (n = 242).

### 5.5. Discussion

Although prior studies have reported the existence of four molecular subtypes of HGSC ovarian cancer (239, 240, 243, 245), our analysis suggests the existence of only two or three subtypes. This conclusion is based on our observation that concordance of analogous subtypes across study populations was stronger for two or three clusters as opposed to four. Previous studies used either k-means or NMF clustering, and because

our results contradicted prior work, we performed analyses using both of these methods. Results for each population were similar for the k means and NMF clustering algorithms, suggesting that the clustering algorithm did not drive the observed differences.

In the previous literature, the only report that suggested four subtypes represented the data better than three was TCGA's reanalysis of the Tothill data (Figure S6.2 in their publication); the cophenetic coefficient dropped dramatically at  $k = 3$  before recovering at  $k = 4$  (239). Notably, TCGA's figure legend for this supplemental result indicates that they did not remove serous borderline tumors from the Tothill data. Our analysis of the Tothill data differed from TCGA's in that we excluded serous borderline tumors, and instead supports the existence of two or three subtypes. To evaluate the influence of these serous borderline tumors in the Tothill data, we repeated our analyses including serous borderline tumors, and observed a drop in the cophenetic coefficient for  $k = 3$  relative to  $k = 4$  (Figure 5.6). This suggests that the four subtypes observed in TCGA's analysis of the Tothill data may be due, in part, to the inclusion of serous borderline tumors.

There are several limitations to note in the HGSC data we analyzed. Given the intratumor heterogeneity that is likely to exist (256), our approach would be strengthened by having data on multiple areas of the tumors. Additionally, since histology and grade classification have changed over time (257, 258), it is unclear whether the populations we studied used comparable guidelines to determine histology and grade. We attempted to exclude all low-grade serous and low-grade endometrioid samples because they often have very different gene expression patterns and more favorable survival compared to their higher-grade counterparts (238). It is unclear why the Bonome clusters did not correspond to the clusters observed in other populations. Lack of consistency could result from unreported biological differences.



In summary, our study demonstrates that two clusters of HGSC, “mesenchymal-like” and “proliferative-like,” are clearly and consistently identified within and between populations. This suggests that there are two reproducible HGSC subtypes that are either etiologically distinct, or acquire phenotypically determinant alterations through their development. Our study also suggests that the previously described “immunoreactive-like” and “differentiated-like” subtypes appear to be more variable across populations, and tend to be collapsed into a single category when three subtypes are specified. These may represent, for example, steps along an immunoreactive continuum or could represent the basis of a third, but more variable, subtype. Understanding the underlying biology of the robust, well-defined “mesenchymal-like” and “proliferative-like” subtypes universally observed across populations could lead to targeted treatments that might influence survival. More work needs to be done to determine whether the heterogeneous samples that do not fall into one of these clear groups can be classified into homogeneous subtypes using other characteristics such as methylation markers or a combination of genomic measures. Our analysis reveals the importance of critically reassessing molecular subtypes across multiple large study populations using parallel analyses and consistent inclusion criteria. New systematic approaches hold promise for the implementation of such analyses (259, 260). Our results underscore the importance of ovarian cancer histopathology, contradict the four HGSC subtype hypothesis, and suggest that there may be fewer HGSC molecular subtypes with variable immunoreactivity and stromal infiltration.

## **5.6. Acknowledgements**

We thank Sebastian Armasu and Hsiao-Wang Chen for help with statistical analyses and data processing and Emily Kate Shea for helpful discussions. This work was supported by the National Cancer Institute at the National Institutes of Health (R01

CA168758 to J.A.D., F31 CA186625 to J.R., and R01 CA122443 to E.L.G.); the Mayo Clinic Ovarian Cancer Specialized Program of Research Excellence grant (P50 CA136393 to E.L.G.); the Mayo Clinic Comprehensive Cancer Center-Gene Analysis Shared Resource (P30 CA15083); the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative (grant number GBMF 4552 to C.S.G.); the American Cancer Society (grant number IRG 8200327 to C.S.G.); and by Norris Cotton Cancer Center Developmental Funds.

## Chapter 6.

### Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders

This chapter was originally published as: Way, Gregory, P., Greene, Casey, S. “*Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders*”. Pacific Symposium on Biocomputing 23 (2018) 80 – 91. doi:10.1142/9789813235533\_0008

Conceptualization: G.P.W., C.S.G.; Methodology: G.P.W., C.S.G; Software: G.P.W.; Investigation: G.P.W.; Writing – Original Draft: G.P.W.; Writing – Review and Editing: G.P.W., C.S.G.; Visualization: G.P.W.

#### 6.1. Abstract

The Cancer Genome Atlas (TCGA) has profiled over 10,000 tumors across 33 different cancer-types for many genomic features, including gene expression levels. Gene expression measurements capture substantial information about the state of each tumor. Certain classes of deep neural network models are capable of learning a meaningful latent space. Such a latent space could be used to explore and generate hypothetical gene expression profiles under various types of molecular and genetic perturbation. For example, one might wish to use such a model to predict a tumor's response to specific therapies or to characterize complex gene expression activations existing in differential proportions in different tumors. Variational autoencoders (VAEs) are a deep neural network approach capable of generating meaningful latent spaces for image and text data. In this work, we sought to determine the extent to which a VAE can be trained to model cancer gene expression, and whether or not such a VAE would capture biologically-relevant features. In the following report, we introduce a VAE trained on TCGA pan-cancer RNA-seq data, identify specific patterns in the VAE encoded

features, and discuss potential merits of the approach. We name our method “Tybalt 🐱” after an instigative, cat-like character who sets a cascading chain of events in motion in Shakespeare’s “Romeo and Juliet”. From a systems biology perspective, Tybalt could one day aid in cancer stratification or predict specific activated expression patterns that would result from genetic changes or treatment effects.

## **6.2. Introduction**

Deep learning has improved the state of the art in many domains, including image, speech, and text processing, but it has yet to make significant enough strides in biomedicine for it to be considered transformative (261). Nevertheless, several studies have revealed promising results. For instance, Esteva et al. used convolutional neural networks (CNNs) to diagnose melanoma from skin images and Zhou and Troyanskaya trained deep models to predict the impact of non-coding variants (262, 263). However, several domain specific limitations remain. In contrast to image or text data, validating and visualizing learning in biological datasets is particularly challenging. There is also a lack of ground truth labels in biomedical domains, which often limits the efficacy of supervised models. New unsupervised deep learning approaches such as generative adversarial nets (GANs) and variational autoencoders (VAEs) harness the modeling power of deep learning without the need for accurate labels (47, 90, 91). Unlike traditional CNNs, which model data by minimizing inaccurate class predictions, autoencoder models, including VAEs, learn through data reconstruction. Reconstructing gene expression input data using autoencoder frameworks has been previously shown to reveal novel biological patterns (86, 87, 101).

VAEs and GANs are generative models, which means they learn to approximate a data generating distribution. Through approximation and compression, the models have

been shown to capture an underlying data manifold – a constrained, lower dimensional space where data is distributed – and disentangle sources of variation from different classes of data (264, 265). For instance, a recent group trained adversarial autoencoders on chemical compound structures and their growth inhibiting effects in cancer cell lines to learn manifold spaces of effective small molecule drugs (266, 267). Additionally, Rampasek et al. trained a VAE to learn a gene expression manifold of reactions of cancer cell lines to drug treatment perturbation (94). The theoretical basis for modeling cancer using lower dimensional manifolds is established, as it has been previously hypothesized that cancer exists in “basins of attraction” defined by specific pathway aberrations that drive cells toward cancer states (3). These states could be revealed by data driven manifold learning approaches.

The Cancer Genome Atlas (TCGA) has captured several genomic measurements for over 10,000 different tumors across 33 cancer-types (176). TCGA has released this data publicly, enabling many secondary analyses, including the training of deep models that predict survival (268). One data type amenable to modeling manifold spaces is RNA-seq gene expression because it can be used as a proxy to describe tumor states and the downstream consequences of specific molecular aberration. Biology is complex, consisting of multiple nonlinear and often redundant connections among genes, and when a specific pathway aberration occurs, the downstream response to the perturbation is captured in the transcriptome. In the following report, we extend the autoencoder framework by training and evaluating a VAE on TCGA RNA-seq data. We aim to demonstrate the validity and specific latent space benefits of a VAE trained on gene expression data. We do not aim to comprehensively profile all learned pan-cancer VAE features nor survey clinical implications. We also do not compare our approach to alternate dimensionality reduction algorithms, but instead present our model as an

additional tool in the toolkit for extracting knowledge from gene expression. We shall name this model “Tybalt 🐱”.

### **6.3. Methods**

#### *6.3.1. Model summary*

VAEs are data driven, unsupervised models that can learn meaningful latent spaces in many contexts. In this work, we aim to build a VAE that compresses gene expression features and reveals a biologically relevant latent space. The VAE is based on an autoencoding framework, which can discover nonlinear explanatory features through data compression and nonlinear activation functions. A traditional autoencoder consists of an encoding phase and a decoding phase where input data is projected into lower dimensions and then reconstructed (89). An autoencoder is deterministic, and is trained by minimizing reconstruction error. In contrast, VAEs are stochastic and learn the distribution of explanatory features over samples. VAEs achieve these properties by learning two distinct latent representations: a mean and standard deviation vector encoding. The model adds a Kullback-Leibler (KL) divergence term to the reconstruction loss, which also regularizes weights through constraining the latent vectors to match a Gaussian distribution. In a VAE, these two representations are learned concurrently through the use of a reparameterization trick that permits a back propagated gradient (90). Importantly, new data can be projected onto an existing VAE feature space enabling new data to be assessed.

#### *6.3.2. Model implementation*

VAEs have been shown to generate “blurry” data compared with other generative models, including GANs, but VAEs are also generally more stable to train (269). We trained our VAE model, Tybalt, with the following architecture: 5,000 input genes

encoded to 100 features and reconstructed back to the original 5,000 (Figure 6.1A). The 5,000 input genes were selected based on highest variability by median absolute deviation (MAD) in the TCGA pan-cancer dataset.

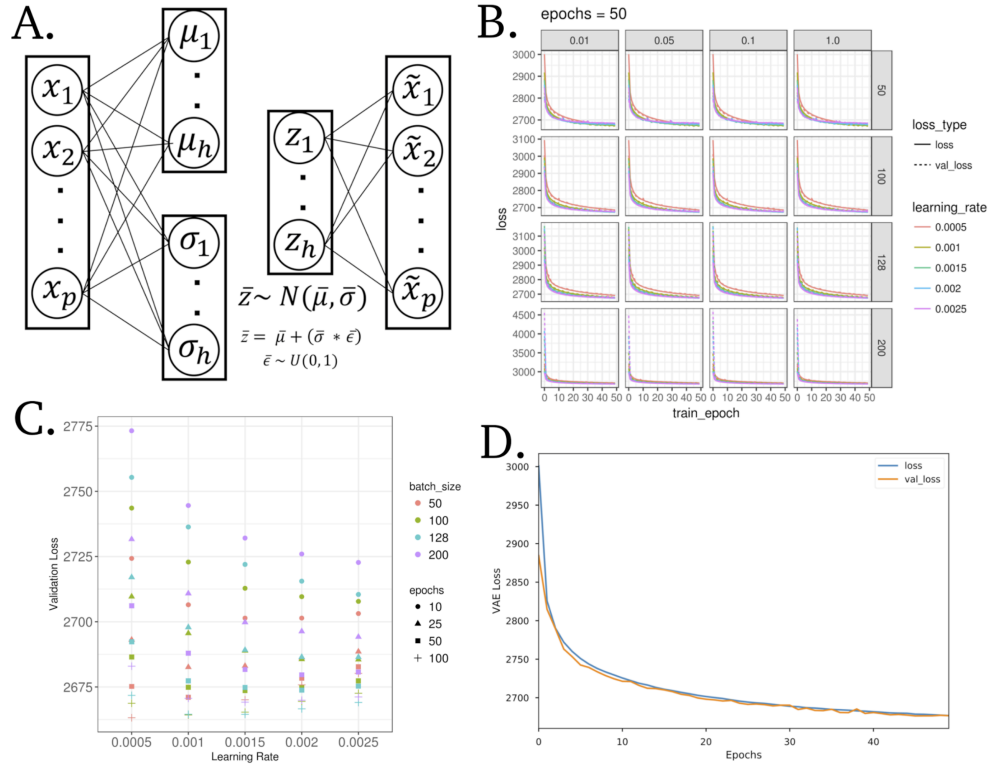


Figure 6.1: A variational autoencoder (VAE) applied to gene expression data

**(A)** Model wire diagram of Tybalt encoding a gene expression vector ( $p = 5,000$ ) into mean and standard deviation vectors ( $h = 100$ ). A reparameterization trick (90, 91) enables learning  $z$ , which is then reconstructed back to input. **(B)** Training and validation VAE loss across training epochs (full pass through all training data). Shown across vertical and horizontal facets are values of kappa and batch size, respectively. **(C)** Final validation loss for all parameters with kappa = 1. **(D)** VAE loss for training and testing sets through optimized model training.

We initially trained Tybalt without batch normalization (270), but observed that when we included batch normalization in the encoding step, we trained faster and with heterogeneous feature activation. Batch normalization in machine learning is distinct from normalizing gene expression batches together in data processing. In machine learning, batch normalization adds additional feature regularization by scaling activations

to zero mean and unit variance, which has been observed to speed up training and reduce batch to batch variability thus increasing generalizability. We trained Tybalt with an Adam optimizer (271), included rectified linear units (272) and batch normalization in the encoding stage, and sigmoid activation in the decoding stage. We built Tybalt in Keras (version 2.0.6) (273) with a TensorFlow backend (version 1.0.1) (274). For more specific VAE illustrations and walkthroughs refer to an extended tutorial (275) and these intuitive blog posts (276, 277).

### 6.3.3. *Parameter selection*

We performed a parameter sweep over batch size (50, 100, 128, 200), epochs (10, 25, 50, 100), learning rates (0.005, 0.001, 0.0015, 0.002, 0.0025) and warmups (kappa) (0.01, 0.05, 0.1, and 1). Kappa controls how much the KL divergence loss contributes to learning, which effectively transitions a deterministic autoencoder to a VAE (278, 279). For instance, a kappa = 0.1 would add 0.1 to a weight on the KL loss after each epoch. After 10 epochs, the KL loss will have equal weight as the reconstruction loss. We did not observe kappa to influence model training (Figure 6.1) so we kept kappa = 1 for downstream analyses. We evaluated train and test set loss at each epoch. The test set was a random 10% partition of the full data. In general, training was relatively stable for many parameter combinations, but was consistently worse for larger batches, particularly with low learning rates. Ultimately, the best parameter combination based on validation loss was batch size 50, learning rate 0.0005, and 100 epochs (Figure 6.1C). Because training stabilized after about 50 epochs, we terminated training early. Training and testing loss across all 50 epochs is shown in Figure 6.1D. We performed the parameter sweep on a cluster of 8 NVIDIA GeForce GTX 1080 Ti GPUs on the PMACS cluster at The University of Pennsylvania.



#### 6.3.4. *Input data*

The input data consisted of level 3 TCGA RNA-seq gene expression data for 9,732 tumors and 727 tumor adjacent normal samples (10,459 total samples) measured by the 5,000 most variably expressed genes. The full dataset together is referred to as the pan-cancer data. The level 3 RNA-seq data consists of a preprocessed and batch-corrected gene abundance by sample matrix measured by  $\log_2(\text{FPKM} + 1)$  transformed RSEM values. The most variably expressed genes were defined by median absolute deviation (MAD). In total, there were 33 different cancer-types (including glioblastoma, ovarian, breast, lung, bladder cancer, etc.) profiled, each with varying number of tumors. We accessed RNA-seq data from the UCSC Xena data browser on March 8th, 2016 and archived the data in Zenodo (280). To facilitate training, we min-maxed scaled RNA-seq data to the range of 0 to 1. We used corresponding clinical data accessed from the Snaptron web server (39).

#### 6.3.5. *Interpretation of gene weights*

Much like the weights of a deterministic autoencoder, Tybalt's decoder weights captured the contribution of specific genes to each learned feature (86, 281, 101). For most features, the distribution of gene weights was similar: Many genes had weights near zero and few genes had high weights at each tail. In order to characterize patterns explained by selected encoded features of interest, we performed overrepresentation pathway analyses (ORA) separately for both positive and negative high weight genes; defined by greater than 2.5 standard deviations above or below the mean, respectively. We used WebGestalt (282), with a background of the 5,000 assayed genes, to perform the analysis over gene ontology (GO) biological process terms (145). P values are presented after an Benjamini-Hochberg FDR adjustment.

### 6.3.6. *The latent space of ovarian cancer subtypes*

Image processing studies have shown the remarkable ability of generative models to mathematically manipulate learned latent dimensions (283, 284). For example, subtracting the image latent representation of a neutral man from a smiling man and adding it to a neutral woman, resulted in a vector associated with a smiling woman. We were interested in the extent to which Tybalt learned a manifold representation that could be manipulated mathematically to identify state transitions across high grade serous ovarian cancer (HGSC) subtypes. The TCGA naming convention of these subtypes is mesenchymal, proliferative, immunoreactive, and differentiated (239). To characterize the largest differences between the mesenchymal/immunoreactive and proliferative/differentiated HGSC subtypes, we performed a series of mean HGSC subtype vector subtractions in Tybalt latent space:

$$\bar{\theta}_k = \frac{\sum_{i=1}^n z_{i,1}(i_k = k)}{n_k}, \dots, \frac{\sum_{i=1}^n z_{i,100}(i_k = k)}{n_k}$$
$$\bar{\theta}_{immunoreactive} - \bar{\theta}_{mesenchymal} = \bar{\theta}_{immuno-mes}$$
$$\bar{\theta}_{differentiated} - \bar{\theta}_{proliferative} = \bar{\theta}_{diff-prolif}$$

Where  $i_k = k$  is an indicator function if sample  $i$  has membership with subtype  $k$  and  $z$  is the encoded layer. We used tumor subtype assignments provided for TCGA samples in Verhaak et al. 2013 (244). If Tybalt learned a biological manifold, this subtraction would result in the identification of biologically relevant features stratifying tumors of specific subtypes with a continuum of expression states.

### 6.3.7. *Enabling exploration through visualization*

We provide a Shiny app to interactively visualize activation patterns of encoded Tybalt features with covariate information at [https://gregway.shinyapps.io/pancan\\_plotter/](https://gregway.shinyapps.io/pancan_plotter/).

#### 6.3.8. *Reproducibility*

We provide all scripts to reproduce and to build upon this analysis under an open source license at <https://github.com/greenelab/tybalt> (285).

### 6.4. **Results**

Tybalt compressed tumors into a lower dimensional space, acting as a nonlinear dimensionality reduction algorithm. Tybalt learned which genes contributed to each feature, potentially capturing aberrant pathway activation and treatment vulnerabilities. Tybalt was unsupervised; therefore, it could learn both known and unknown biological patterns. In order to determine if the features captured biological signals, we characterized both sample- and gene-specific activation patterns.

#### 6.4.1. *Tumors were encoded in a lower dimensional space*

The tumors were encoded from original gene expression vectors of 5,000 MAD genes into a lower dimensional vector of length 100. To determine if the sample encodings faithfully recapitulated large, tissue specific signals in the data, we visualized sample-specific Tybalt encoded features (z vector for each sample) by t-distributed stochastic neighbor embedding (t-SNE) (55). We observed similar patterns for Tybalt encodings (Figure 6.2A) as compared to 0-1 normalized RNA-seq data (Figure 6.2B). Tybalt geometrically preserved well known relationships, including similarities between glioblastoma (GBM) and low grade glioma (LGG). Importantly, the recapitulation of tissue-specific signal was captured by non-redundant, highly heterogeneous features (6.2C). Based on the hierarchical clustering dendrogram, the features appeared to be capturing distinct signals. For instance, tumor versus normal and patient sex are large signals present in cancer gene expression, but they were distributed uniformly in the clustering solution indicating non-redundant feature activations.

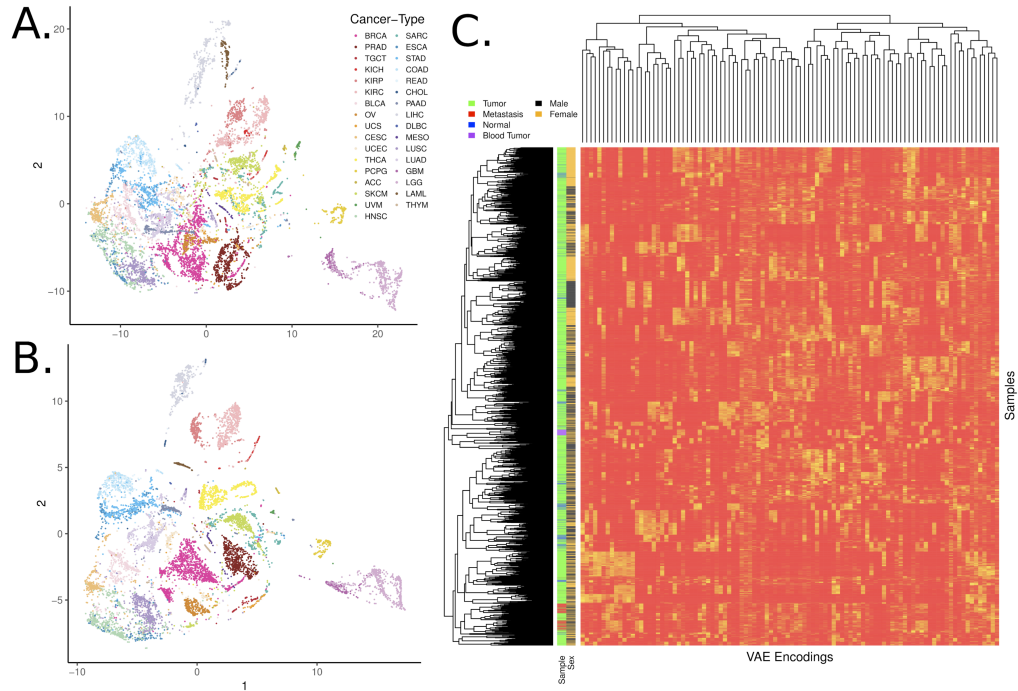


Figure 6.2: Samples encoded by a variational autoencoder retain biological signals

**(A)** t-distributed stochastic neighbor embedding (t-SNE) of TCGA pan-cancer tumors with Tybalt encoded features. **(B)** t-SNE of 0-1 normalized gene expression features. Tybalt retains similar signals as compared to uncompressed gene expression data. **(C)** Full Tybalt encoding features by TCGA pan-cancer sample heatmap. Given on the y axis are the patient's sex and type of sample.

#### 6.4.2. Features represent biological signal

Our goal was to train and evaluate Tybalt on its ability to learn biological signals in the data and not to perform a comprehensive survey of learned features. Therefore, we investigated whether or not Tybalt could distinguish patient sex and patterns of metastatic activation. We determined that the model extracted patient sex robustly (Figure 6.3A). Feature encoding 82 nearly perfectly separated samples by sex. Furthermore, we identified a set of nodes that together identified skin cutaneous melanoma (SKCM) tumors of both primary and metastatic origin (Figure 6.3B). The weights used to decode the hidden layer (z vector) back into a high-fidelity

reconstruction of the input can capture important and consistent biological patterns embedded in the gene expression data (86, 101, 281). For instance, there were only 17 genes needed to identify patient sex (Figure 6.3C). These genes were mostly located on sex chromosomes. The two positive weight genes were X inactivation genes *XIST* and *TSIX*, while the negative weight genes were mostly Y chromosome genes such as *EIF1AY*, *UTY*, and *KDM5D*. This result served as a positive control that the unsupervised model was able to construct a feature that described a clearly biological source of variance in the data.

There were several genes contributing to the two encoded features that separated the SKCM tumors (Figure 6.3D). Several genes existed in the high weight tails of each distribution for feature encodings 53 and 66. We performed an ORA on the high weight genes. In general, several pathways were identified as overrepresented in the set as compared to random. The samples had intermediate to high levels of feature encoding 53, which did not correspond to any known GO term, potentially indicating an unknown but important biological process. The samples also had intermediate to high levels of encoding 66 which implicated GO terms related to cholesterol, ethanol, and lipid metabolism including “regulation of intestinal cholesterol absorption” ( $adj. p = 3.0e^{-2}$ ), “ethanol oxidation” ( $adj. p = 4.0e^{-02}$ ), and “lipid catabolic process” ( $adj. p = 4.0e^{-02}$ ). SKCM samples had consistently high activation of both encoded features, which separated them from other tumors. Nevertheless, more research is required to determine how VAE features could be best interpreted in this context.

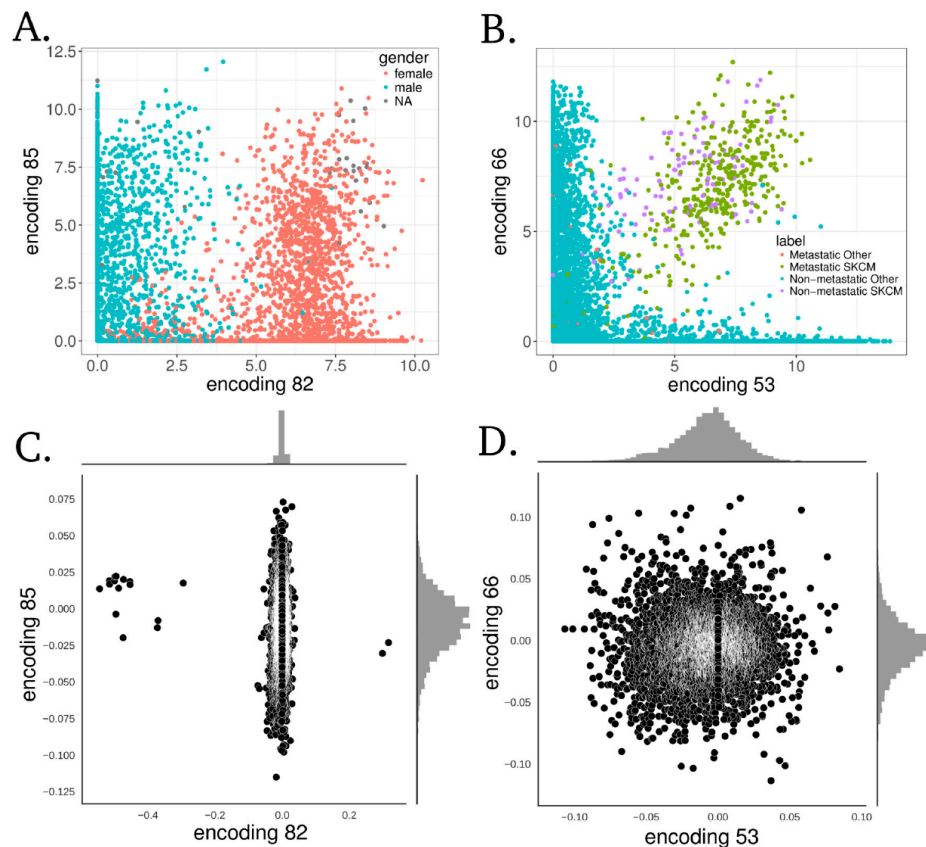


Figure 6.3: Specific examples of Tybalt features capturing biological signals

**(A)** Encoding 82 stratified patient sex. **(B)** Together, encodings 53 and 66 separated melanoma tumors. Distributions of gene coefficients contributing to each plot above for **(C)** patient sex and **(D)** melanoma. The gene coefficients consist of the Tybalt learned weights for each feature encoding.

#### 6.4.3. Interpolating the lower dimensional manifold of HGSC subtypes

We performed an experiment to test whether or not Tybalt learned manifold differences of distinct HGSC subtypes. Previously, several groups identified four HGSC subtypes using gene expression (239, 240, 243). However, the four HGSC subtypes were not consistently defined across populations; the data suggested the presence of three subtypes or fewer (57). The study observed that the immunoreactive/mesenchymal and differentiated/proliferative tumors consistently collapsed together when setting clustering algorithms to find 2 subtypes (57). This observation may suggest the presence

of distinct gene expression programs existing on an activation spectrum driving differences in these subtypes. Therefore, we hypothesized that Tybalt would learn the manifold of gene expression spectra existing in differential proportions across these subtypes.

The largest feature encoding difference between the mean HGSC mesenchymal and the mean immunoreactive subtype ( $\bar{\theta}_{\text{immuno-mes}}$ ) was encoding 87 (Figure 6.4A). Encoding 77 and encoding 56 (Figure 6.4B) also distinguished the mesenchymal and immunoreactive subtypes. The largest feature encoding differences between the mean proliferative and the mean differentiated subtype ( $\bar{\theta}_{\text{diff-prolit}}$ ) were contributed by encoding 79 (Figure 6.4C) and encoding 38 (Figure 6.4D). Interestingly, encoding 38 had high mean activation in both the immunoreactive and differentiated subtypes.

The mesenchymal subtype had the highest encoding 87 activation. Encoding 87 was associated with the expression of genes involved in collagen and extracellular matrix processes (Table 6.1), which has been previously observed to be an important marker of the mesenchymal subtype (239, 240). Encoding 56 was associated with immune system responses (Table 6.1), and the immunoreactive subtype displayed the highest activation. Encoding 79 is mostly expressed in the proliferative subtype and has low activation in differentiated tumors. The high weight negative genes of encoding 79 were associated with glucuronidation processes (Table 6.1). The negative genes of encoding 38, which also distinguished differentiated from proliferative tumors but in the opposite direction, were also associated with glucuronidation. Previously, glucuronidation processes were observed to be associated with response to chemotherapy and survival in colon cancer patients (286, 287). Our results indicate that differential activation of glucuronidation is a strong signal distinguishing HGSC subtypes.

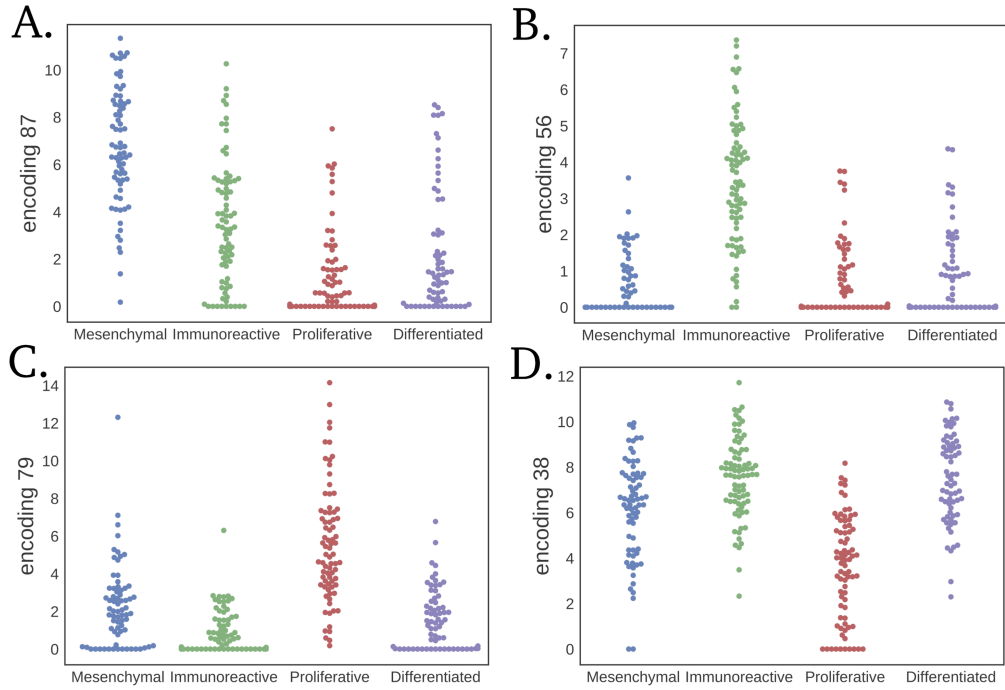


Figure 6.4: Largest mean differences in HGSC subtype vector subtraction for each subtype

Subtracting the mesenchymal subtype by the immunoreactive result in distribution differences in **(A)** feature encoding 87 and **(B)** encoding 56. Subtracting the proliferative subtype by the differentiated subtype results in differences between **(C)** feature encoding 79 and **(D)** encoding 38.

<i>Encoding</i>	<i>Tail</i>	<i>Subtype</i>	<i>Enriched Pathway</i>	<i>Adj.P Value</i>
87	+	Mesenchymal	Collagen Catabolic Process	1.8e-9
87	+	Mesenchymal	Extracellular Matrix Organization	4.2e-6
87	-	Immunoreactive	Urate Metabolic Process	1.5e-2
56	+	Immunoreactive	Immune Response	1.3-12
56	+	Immunoreactive	Defense Response	2.9e-12
56	-	Mesenchymal	No Sig. Pathways	
79	+	Proliferative	Chemical Synaptic Transmission	9.1e-3
79	-	Differentiated	Xenobiotic Glucuronidation	2.1e-9
38	+	Differentiated	No Sig. Pathways	
38	-	Proliferative	Xenobiotic Glucuronidation	7.2e-6

Table 6.1: Summary of significantly overrepresented pathways separating HGSC subtypes identified by latent space arithmetic with VAE features



This observation may also help to explain increased survival in HGSC patients with differentiated tumors (243). Lastly, encoding 77 also separated immunoreactive from mesenchymal tumors and did not display any significant terms, which may indicate novel biology explaining undiscovered subtype differences.

### **6.5. Conclusions**

Tybolt is a promising model but still requires careful validation and more comprehensive evaluation. We observed that the encoded features recapitulated tissue specific patterns. We determined that the learned features were generally non-redundant and could disentangle large sources of variation in the data, including patient sex and SKCM. It is also likely that the features learn tissue specific patterns distinguishing other cancer-types (our shiny app enables full exploration of VAE features by cancer-type). While we identified specific features separating HGSC subtypes, there are likely several other features that describe other important biological differences across cancer-types including differentiation state and activation states of specific pathways. Interpretation of the decoding layer weights helped to identify the contribution of different genes and pathways promoting disparate biological patterns. However, interpretation by pathway analysis must be performed with caution as these analyses rely on incomplete pathway databases and may contain many false positive results.

VAEs provide similar benefits as autoencoders, but they also have the ability to learn a manifold with meaningful relationships between samples. This manifold could represent differing pathway activations, transitions between cancer states, or indicate particular tumors vulnerable to specific drugs. We performed initial testing to determine if we could traverse the underlying manifold by subtracting out cancer-type specific mean activations. While we identified several promising functional relationships existing in a spectrum of activation patterns, rigorous experimental testing would be required to draw

strong conclusions about the biological implications. The specific subtype associations must be confirmed in independent datasets and the processes must be confirmed experimentally. It must also be assessed if Tybalt features learned from TCGA pan-cancer are generalizable to other, potentially more heterogeneous datasets. Further testing is required to confirm that Tybalt catalogued an interpretable manifold capable of interpolation between cancer states. In the future, we will develop higher capacity models and increased evaluation/interpretation efforts to catalog Tybalt encoded RNA-seq expression patterns present in specific cancer-types. This effort may lead to widespread stratification of expression patterns and enable accurate detection of patients who may benefit from specific targeted therapies.

#### **6.6. Acknowledgements**

This work was supported by NIH grant T32 HG000046 (GPW) and GBMF 4552 from the Gordon and Betty Moore Foundation (CSG). We would like to thank Brett K. Beaulieu-Jones for helpful discussions and Jaclyn N. Taroni and David Nicholson for code review. We would also like to thank four anonymous reviewers for their insightful comments.

## Chapter 7.

### **Sequential compression across latent space dimensions enhances gene expression signatures**

This chapter will be submitted for publication as: Way, Gregory, P., Zietz, Michael, Himmelstein, Daniel, S., and Greene, Casey, S. “*Sequential compression across latent space dimensions enhances gene expression signatures*”.

Conceptualization: G.P.W., C.S.G.; Methodology: G.P.W., C.S.G.; Software: G.P.W., M.Z., D.S.H.; Investigation: G.P.W., C.S.G.; Writing – Original Draft: G.P.W.; Visualization: G.P.W.

#### **7.1. Abstract**

##### *7.1.1. Background*

Unsupervised machine learning algorithms applied to gene expression data extract latent, or hidden, signals representing technical and biological sources of variation. However, these algorithms require a user to select a biologically-appropriate dimensionality.

##### *7.1.2. Results*

We compressed gene expression data from three large transcriptomic datasets consisting of adult normal tissue, adult cancer tissue, and pediatric cancer tissue. We compressed these data using principal components analysis (PCA), independent components analysis (ICA), non-negative matrix factorization (NMF), denoising autoencoders (DAE), and variational autoencoders (VAE). Rather than selecting a single latent dimensionality, we sequentially compressed input data into many dimensions ranging from 2 to 200. Each algorithm has various tradeoffs. We observed high model stability and model similarity between PCA, ICA, and NMF algorithms across latent dimensions. We identified more unique biological signatures in ensembles of DAE and VAE models. Using all compressed features across algorithms, ensembles, and latent

dimensions captured the highest proportion of biological features. We used compressed features across algorithms and dimensions to identify gene expression signatures representing sample sex, neuroblastoma MYCN amplification, and blood cell types, which generalized to external datasets. In supervised machine learning tasks, compressed features can be used to predict cancer type and gene alteration status. In this setting, the best performing supervised models used features from different dimensionalities and compression algorithms indicating that there was no single best dimensionality or compression algorithm.

#### *7.1.3. Conclusions*

Ensembles of features from different unsupervised algorithms discovers biological signatures in large transcriptomic datasets. In order to optimize biological signature discovery, rather than compressing input data into a single pre-selected dimensionality, it is best to perform compression on input data over many different latent dimensionalities.

### **7.2. Introduction**

Dimensionality reduction algorithms compress input data into feature representations that capture major sources of variation. Applied to gene expression data, compression algorithms identify latent biological and technical processes. These processes reveal important information about the samples and can help to generate hypotheses that are difficult or impossible to observe in the original genomic space. For example, applying PCA to a large cancer transcriptomic compendium determined the influence of copy number alterations in gene expression measurements (76). ICA applied to transcriptome data aggregated gene modules to identify core pathways and hidden transcriptional programs (51, 79). NMF is often applied to estimate cell type proportion in bulk gene expression data (33, 288). DAEs have revealed latent signals representing oxygen

exposure and transcription factor targets in gene expression data (87, 101). VAEs have identified biologically relevant latent features discriminating cancer subtypes and drug response (93, 94). Nevertheless, a major challenge to all compression applications is the fundamental requirement that a researcher must first determine the number of latent dimensions ( $k$ ) to compress the input data into.

We hypothesize that different latent space dimensionalities and algorithms best capture various biological signatures. Therefore, in the following paper, we train and evaluate compression models across a wide range of latent space dimensionalities, from  $k = 2$  to  $k = 200$ , using PCA, ICA, NMF, DAE, and VAE models. We use RNAseq gene expression data from three different datasets: The Cancer Genome Atlas (TCGA) PanCanAtlas (176), the Genome Tissue Expression Consortium Project (GTEx) (289), and the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) Project (290). We integrate gene set networks using Molecular Signatures Database (MSigDB) and xCell data to interpret the biological signals activated in compressed latent features (102, 291, 292).

Across algorithms and latent dimensionalities, we report training and testing performance, including reconstruction cost, model stability, and gene set coverage. We demonstrate various tradeoffs between models, and we determine that compressing gene expression data using various latent dimensions and algorithms enhances biological signature discovery. We name our sequential compression approach BioBombe after the large mechanical device developed by Alan Turing and other cryptologists in World War II to decode encrypted messages sent by Enigma machines. BioBombe sequentially compresses gene expression input with increasing dimensions to optimize biological signature discovery and decipher biological signals embedded within compressed gene expression features.

### **7.3. Results**

#### *7.3.1. BioBombe implementation*

We compressed RNAseq data from TCGA, GTEx, and TARGET using PCA, ICA, NMF, DAE, and VAE across 28 different latent dimensions ( $k$ ) ranging from  $k = 2$  to  $k = 200$ . We used real and permuted data and initialized each model five times per latent dimension resulting in a total of 4,200 different compression models (Figure 7.1). We evaluated hyperparameters for DAE and VAE models across dimensions and trained models using optimized parameter settings. See Figure 7.2 for an outline of our approach. We provide full results for all compression models for both real (293–295) and permuted data (296–298) as publicly available resources.

#### *7.3.2. Assessing compression algorithm reconstruction*

Reconstruction cost, a measurement of the difference between the input and output matrices, is often used to describe the ability of compression models to capture fundamental processes in latent space features that recapitulate the original input data. We tracked the reconstruction cost for the training and testing data partitions for all datasets, algorithms, latent dimensions, and random initializations. As expected, we observed lower reconstruction costs in models trained with real data and with higher latent dimensions (Figure 7.3). Because PCA and ICA are rotations of one another, we used these scores as a positive control. All compression algorithms had similar reconstruction costs, with the highest variability at low latent dimensions (Figure 7.3).

#### *7.3.3. Evaluating model stability and similarity within and across latent dimensions*

We applied singular vector canonical correlation analysis (SVCCA) to algorithm weight matrices to assess model stability within algorithm initializations, and to determine model similarity between algorithms (299). Briefly, SVCCA calculates the average similarity between two compression algorithm weight matrices and identifies the

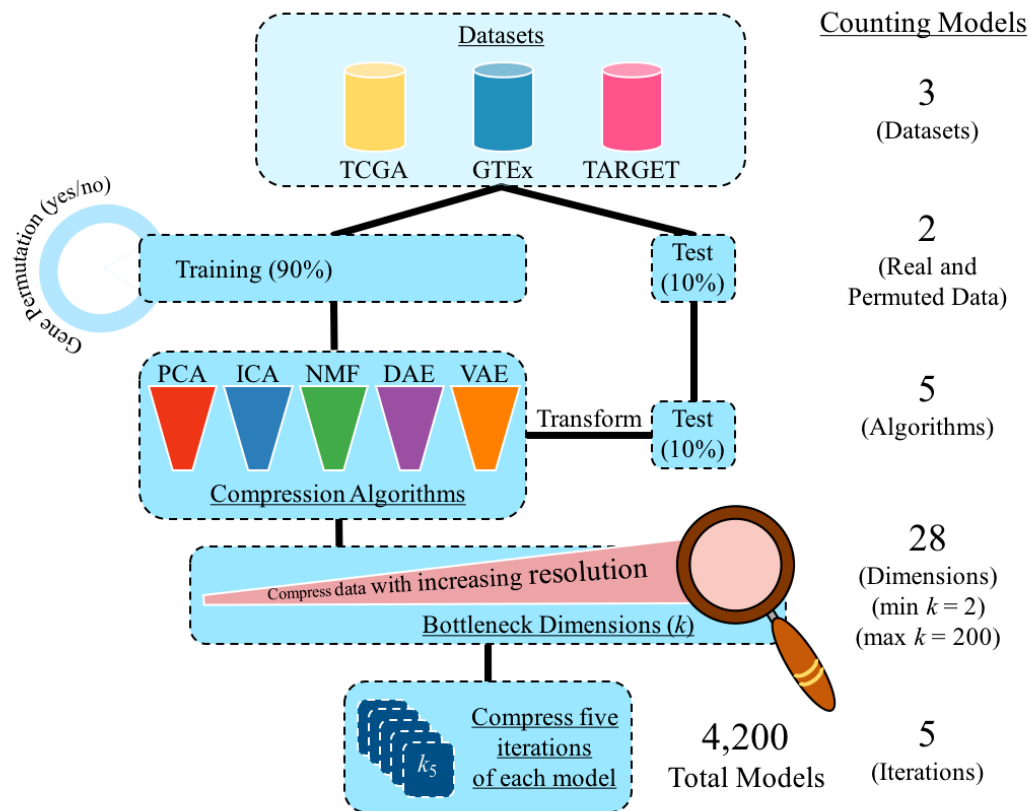


Figure 7.1: Representing our BioBombe implementation workflow

We independently apply our approach to three transcriptome compendia including The Cancer Genome Atlas PanCanAtlas Project (TCGA), Genome-Tissue Expression Project (GTEx), and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative. For each dataset, we split 90% of the data into a training data partition and 10% of the data into a testing data partition. The data is split to match the proportion of cancer-types or tissue-types in each partition. We also randomly permute the gene expression values by gene for all samples in the training set. We proceed with the downstream approach for both real and permuted data in parallel. We apply five compression algorithms including principle components analysis (PCA), independent components analysis (ICA), non-negative matrix factorization (NMF), denoising autoencoders (DAE), and variational autoencoders (VAE). We compress the testing data partition using the trained weights learned from the training set. We sequentially compress the input data into various bottleneck dimensions ( $k$ ) from 2 dimensions to 200 dimensions. We use  $k = 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150$ , and 200 for a total of 28 different dimensions. For each model, we train five independent times using five different random seed initializations. Combined, this yields a total of 4,200 different compression matrices that can be interpreted for biological processes.

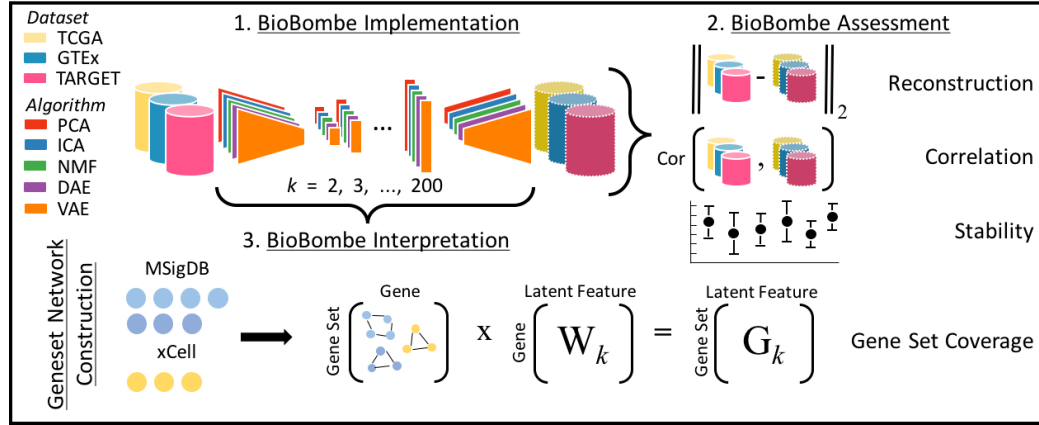


Figure 7.2: Overview of the BioBombe approach

We implemented BioBombe on three datasets using five different algorithms. We sequentially compressed input data into various bottleneck dimensions. We calculated various metrics that describe different benefits and trade-offs of the algorithms. Lastly, we implemented a network projection approach to interpret the compressed latent features. We used MSigDB collections and xCell gene sets in our network.

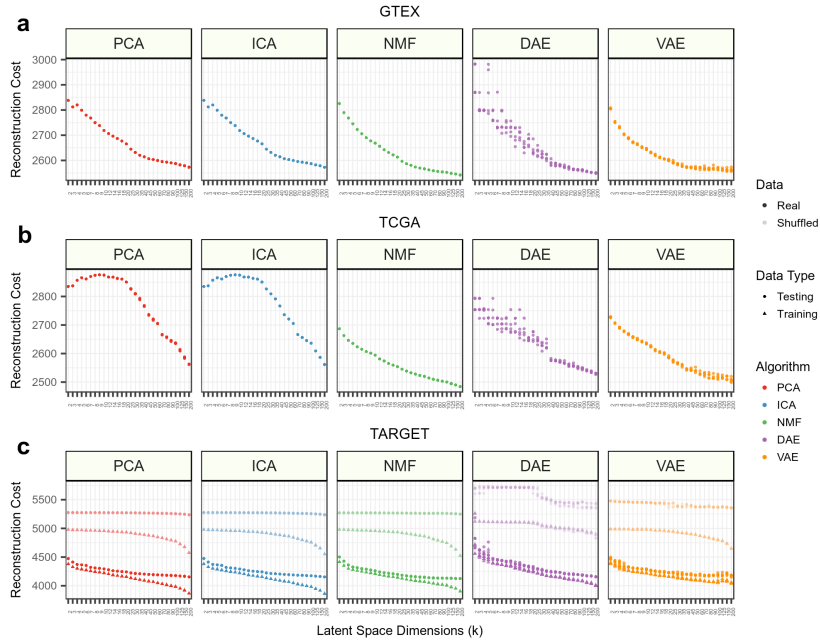


Figure 7.3: Reconstruction cost across datasets, algorithms and dimensions

Reconstruction performance for **(A)** Genome-Tissue Expression Project (GTEx) **(B)** The Cancer Genome Atlas PanCanAtlas Project (TCGA) and **(C)** Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative data. Only real testing data is shown for GTEx and TCGA to highlight specific performance differences that would be unable to visualize with the other data present.



highest correlating feature across all latent space features. Training with TCGA data, we observed highly stable models within algorithms and within all latent dimensions for PCA, ICA, NMF (along the matrix diagonal in Figure 7.4A). VAE models were also largely stable, with some decay in higher latent dimensions. However, DAE models were highly unstable, particularly at low latent dimensions (Figure 7.4A). PCA and ICA were highly similar, and because the two algorithms are rotations of one another, we used this as a positive control for SVCCA estimates. NMF was also highly similar to PCA and ICA, particularly at low latent dimensions (Figure 7.4A). The AE models were less similar to other algorithms. VAE models were more similar to PCA, ICA, and NMF than DAE models, particularly at low latent dimensions, and the instability patterns within DAE models also lead to large differences across algorithms (Figure 7.4A). We observed similar patterns in GTEx and TARGET data (Figure 7.5)

We also used SVCCA to compare the similarity of weight matrices across latent dimensions. Both PCA and ICA found highly similar solutions across all dimensions (Figure 7.4B). This is not surprising since the solutions are deterministic and are arranged with decreasing amounts of variance. NMF also identified highly similar solutions in low dimensions, but solutions were less similar in higher dimensions. DAE solutions were the least similar, with intermediate dimensions showing the lowest mean similarity. VAE models displayed relatively high model similarity, but there were regions of modest model stability in intermediate and high dimensions (Figure 7.4B). We observed similar patterns in GTEx and TARGET data, despite TARGET containing only about 700 samples (Figure 7.6).

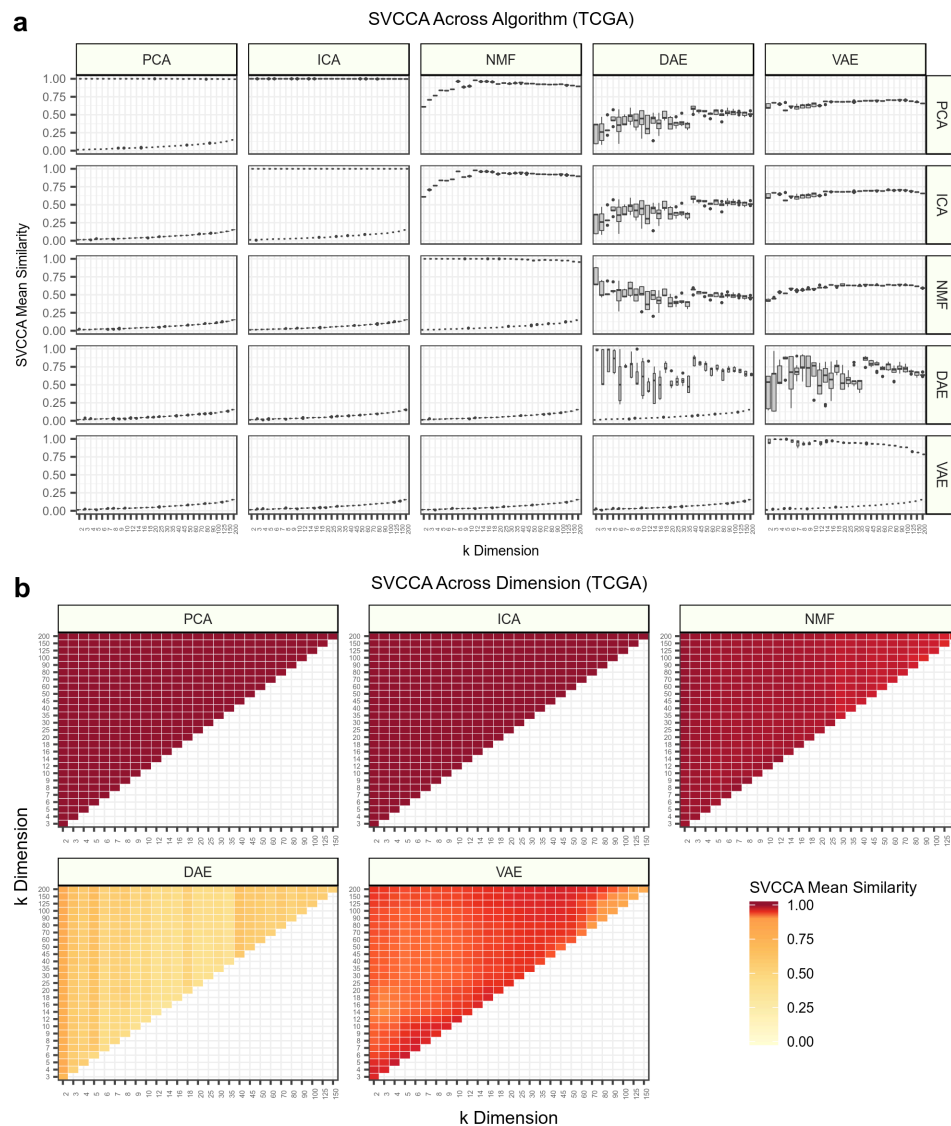


Figure 7.4: Assessing algorithm and dimension stability with singular vector canonical correlation analysis (SVCCA)

**(A)** SVCCA applied to the weight matrices learned by each compression algorithm in gene expression data from The Cancer Genome Atlas (TCGA). The mean of all canonical correlations comparing independent iterations is shown. The distribution of mean similarity represents a comparison of all pairwise iterations within and across algorithms. The upper triangle represents SVCCA applied to real gene expression data, while the lower triangle represents permuted expression data. Both real and permuted data are plotted along the diagonal. **(B)** Mean correlations of all iterations within algorithms but across  $k$  dimensions. SVCCA will identify  $\min(i, j)$  canonical vectors for bottleneck dimensions  $k_i$  and  $k_j$ . The mean of all pairwise correlations is shown for all combinations of  $k$  dimensions.

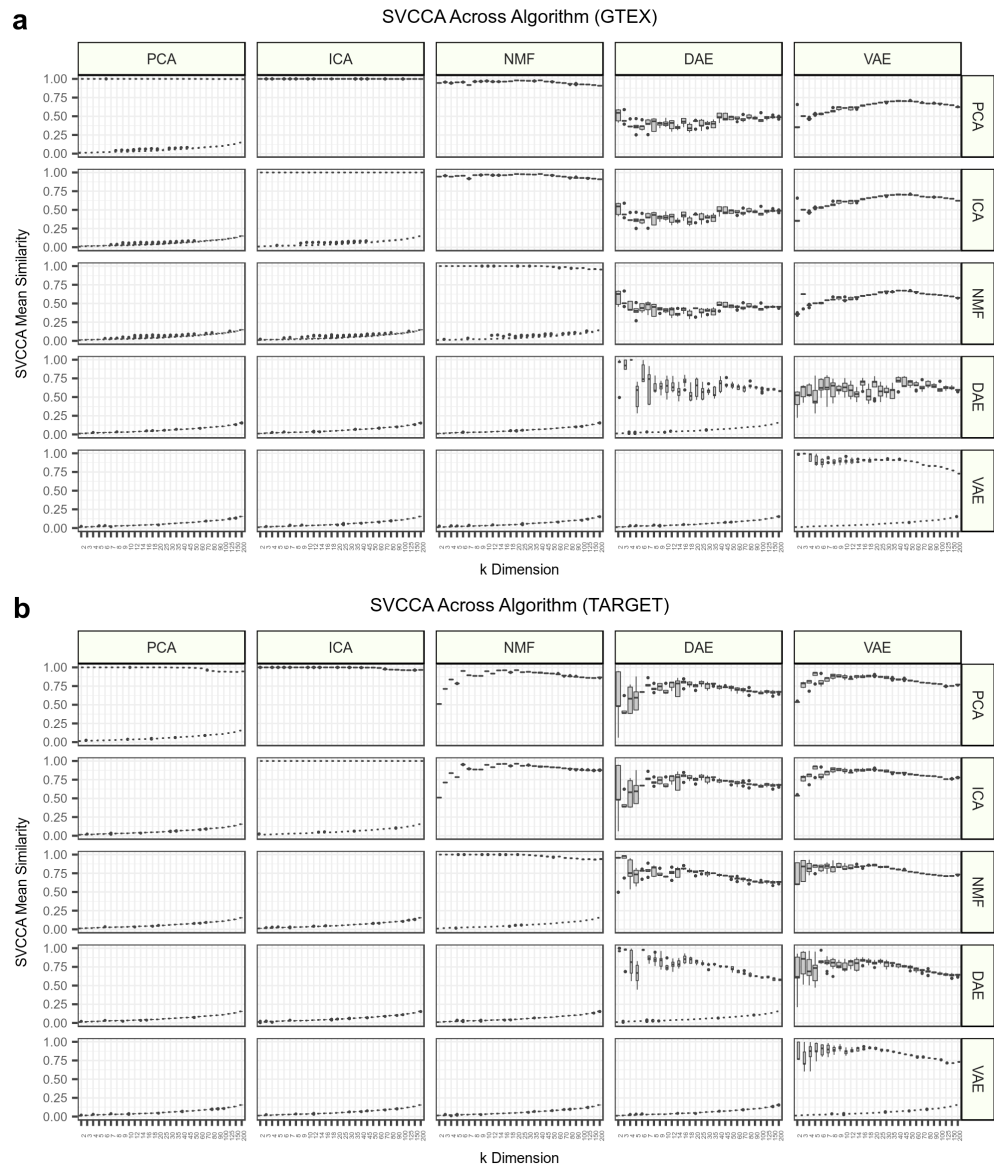


Figure 7.5: Across algorithm stability as measured by singular vector canonical correlation analysis (SVCCA)

Stability is measured for the weight matrices in **(a)** Genome-Tissue Expression Project (GTEx) and **(b)** Therapeutically Applicable Research to Generate Effective Treatments (TARGET). The boxplots represent all pairwise estimates of SVCCA mean similarity for all initializations (across seeds) for real data (upper triangle) and permuted data (lower triangle).

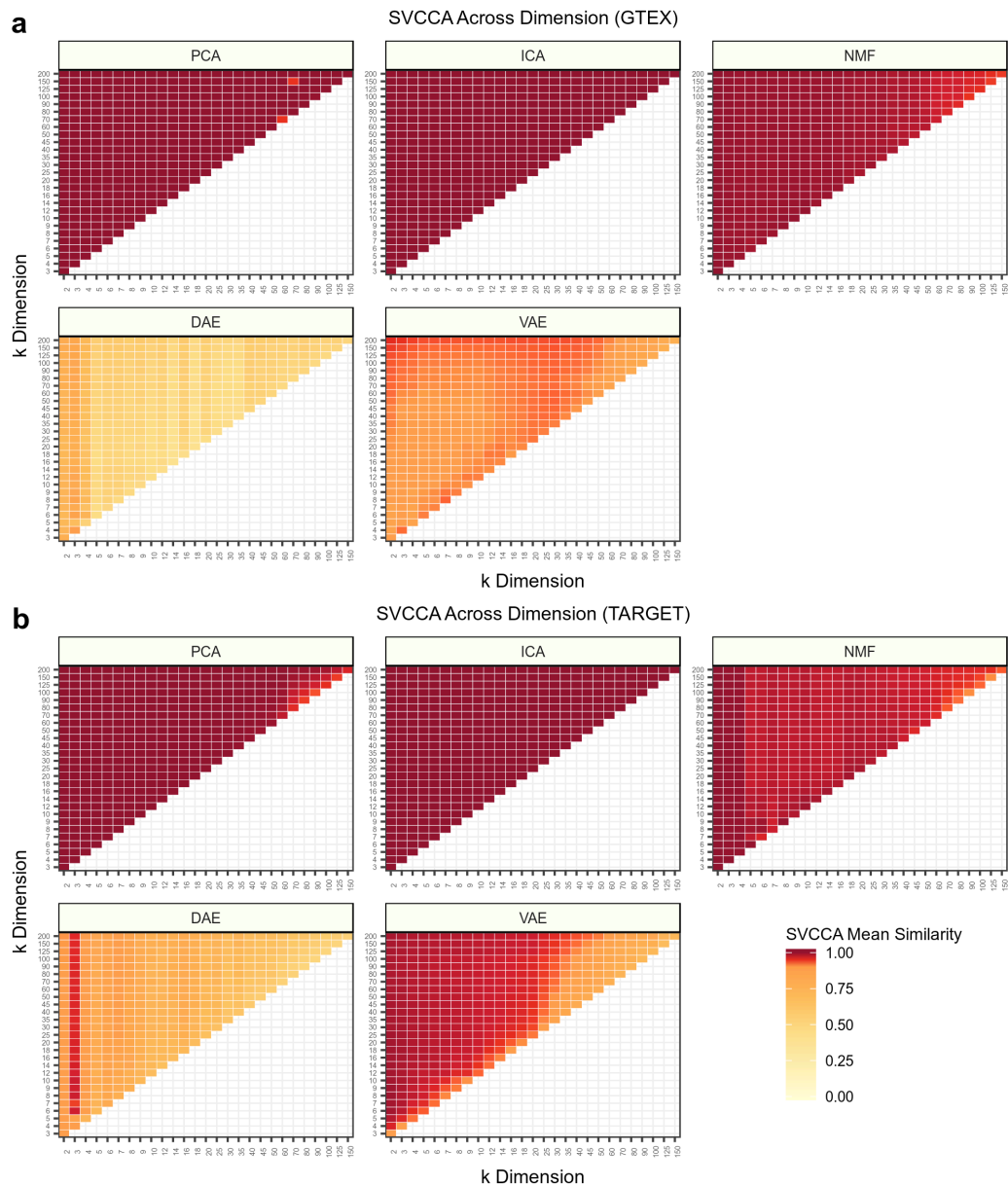


Figure 7.6: Across latent dimension stability as measured by singular vector canonical correlation analysis (SVCCA)

Stability is measured for the weight matrices in **(A)** Genome-Tissue Expression Project (GTEx) and **(B)** Therapeutically Applicable Research to Generate Effective Treatments (TARGET). SVCCA can be measured in two weight matrices of different dimensions. The mean similarity represents the mean of all pairwise estimates across all algorithm initializations. There is some numerical instability observed in the PCA assessment of both plots.

#### 7.3.4. Sequential compression can optimize gene expression signature discovery

We tested the ability of the BioBombe sequential compression approach to isolate biological signatures. First, we sought to identify a latent space feature that identified sample sex, which has been previously observed to be captured in latent spaces (69, 93, 300). We performed a two-tailed t-test comparing male and female samples in GTEx across all initializations, algorithms, and latent dimensions. This signal was optimally identified in higher latent dimensions, particularly in VAE and NMF models (Figure 7.7A). The top feature separating GTEx males and females was VAE feature 108 in  $k = 200$  ( $t = 49.0$ ,  $p = 2.7 \times 10^{-285}$ ) (Figure 7.7B). We performed the same approach using sequentially compressed features in TCGA data. Whereas the largest models appeared to capture sex optimally in GTEx data, intermediate latent dimensions best captured sex in TCGA data (Figure 7.7C). Additionally, the top latent dimension identified was not consistent across algorithms. The top feature distinguishing TCGA males and females was VAE feature 16 in the  $k = 20$  model ( $t = -13.9$ ,  $p = 1.8 \times 10^{-40}$ ) (Figure 7.7D).

We also tested the ability of the sequential compression approach to distinguish MYCN amplification in neuroblastoma (NBL) tumors. MYCN amplification is a biomarker associated with poor prognosis in NBL patients (301). Using latent features derived from the full TARGET data, we performed a two-tailed t-test comparing MYCN amplified vs. MYCN not amplified NBL tumors. Each algorithm discovered optimal signal at various latent dimensions, but the best aligned feature was identified in VAE models at  $k = 200$  (Figure 7.7E). Although there were some potentially mischaracterized samples, feature 111 in VAE  $k = 200$  robustly separated MYCN amplification status in NBL tumors ( $t = 17.5$ ,  $p = 3.0 \times 10^{-37}$ ) (Figure 7.7F). This feature also robustly separated MYCN amplification status in NBL cell lines (302) that were previously unseen by the compression model (Figure 7.7G).

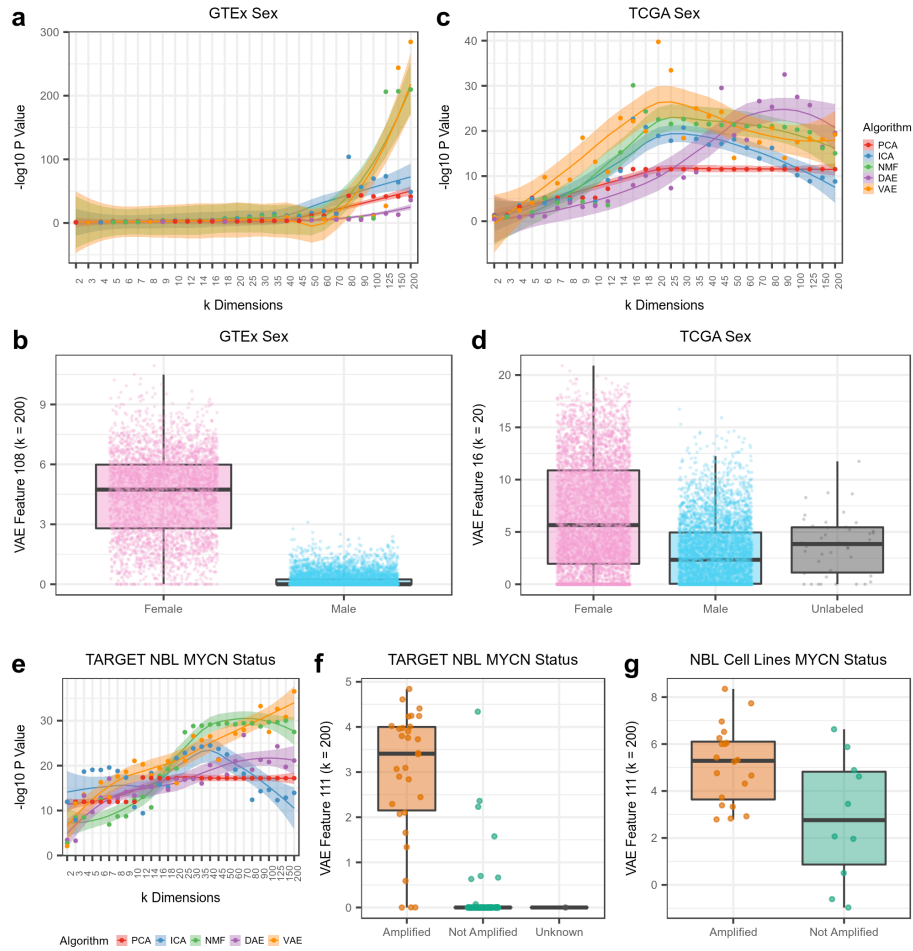


Figure 7.7: Using BioBombe as a signature discovery tool

Detecting GTEx sample sex across **(A)** various latent dimensions and algorithms, and **(B)** the latent feature with the highest enrichment. Detecting TCGA patient sex across **(C)** various latent dimensions, and **(D)** the latent feature with the highest enrichment. Detecting TARGET MYCN amplification in neuroblastoma (NBL) tumors **(E)** across various latent dimensions, and **(F)** the latent feature with the highest enrichment. **(G)** Applying the MYCN signature to an external dataset of NBL cell lines implicates MYCN amplified cell lines.

#### 7.3.5. Assessing gene set coverage of compressed features

We sought to identify biological patterns present in compressed latent features learned across all latent dimensions, algorithms, and initializations. Using various collections as curated by the molecular signatures database (MSigDB) and xCell

genesets, we generated gene set networks. We projected these networks onto compressed gene expression features to assess the proportion of gene sets covered by the various compression features (see methods for more details). Specifically, we tracked coverage of three MSigDB gene set collections representing transcription factor (TF) targets, cancer modules, and Reactome pathways across latent dimensions in TCGA data (Figure 7.8). In all cases, we observed higher gene set coverage in models with larger latent dimensionalities. Using individual models, we observed higher coverage in the linear methods. In particular, ICA seemed to outperform all other algorithms (Figure 7.8A). However, while the linear methods showed the highest coverage, the features identified had relatively low enrichment scores compared to other algorithms (Figure 7.9).

Aggregating all five random initializations into an ensemble model, we observed substantial AE coverage increases (Figure 7.8B). VAE models had high coverage for all gene sets in intermediate dimensions, while DAE improved in higher dimensions. However, at the highest dimensions, ICA demonstrated the highest coverage. NMF consistently had the highest enrichment scores, but the lowest coverage (Figure 7.8B). When considering all models combined (forming an ensemble of algorithm ensembles) within latent dimensions, we observed substantially increased coverage of all gene sets. However, most of the unique gene sets were contributed by the AE models (Figure 7.8B). Lastly, when we aggregated all features across all algorithms and all dimensions together into a single large ensemble model, we observed the highest gene set coverage (Figure 7.8C). While models compressed with larger latent space dimensions had higher gene set coverage, many individual gene sets were captured with the highest enrichment in models with low and intermediate dimensions (Figure 7.10). These results

indicated that optimal biological signature discovery occurs using various compression algorithms with various latent space dimensions.

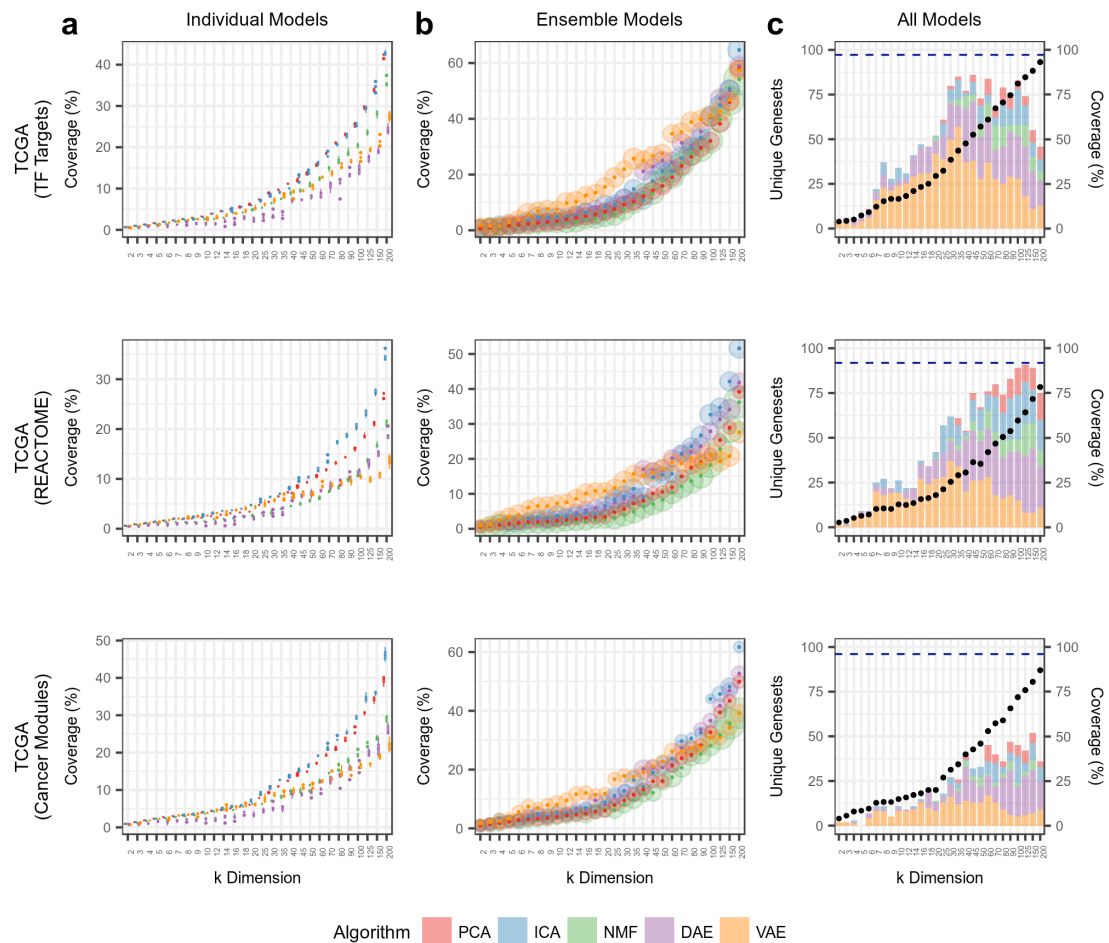


Figure 7.8: Assessing gene set coverage of specific gene set collections

Tracking results in TCGA data for three gene set collections representing transcription factor (TF) targets (C3TFT), Reactome pathways (C2CPREACTOME), and cancer modules (C4CM). **(A)** Tracking coverage in individual models, which represents the distribution of scores across five algorithm iterations. **(B)** Tracking coverage in ensemble models, which represents coverage after combining all five iterations into a single model. The size of the point represents relative enrichment strength. **(C)** Tracking coverage in all models combined within  $k$  dimensions. The number of algorithm-specific unique gene sets identified is shown as bar charts. Coverage for all models combined across all  $k$  dimensions is shown as a dotted navy blue line.



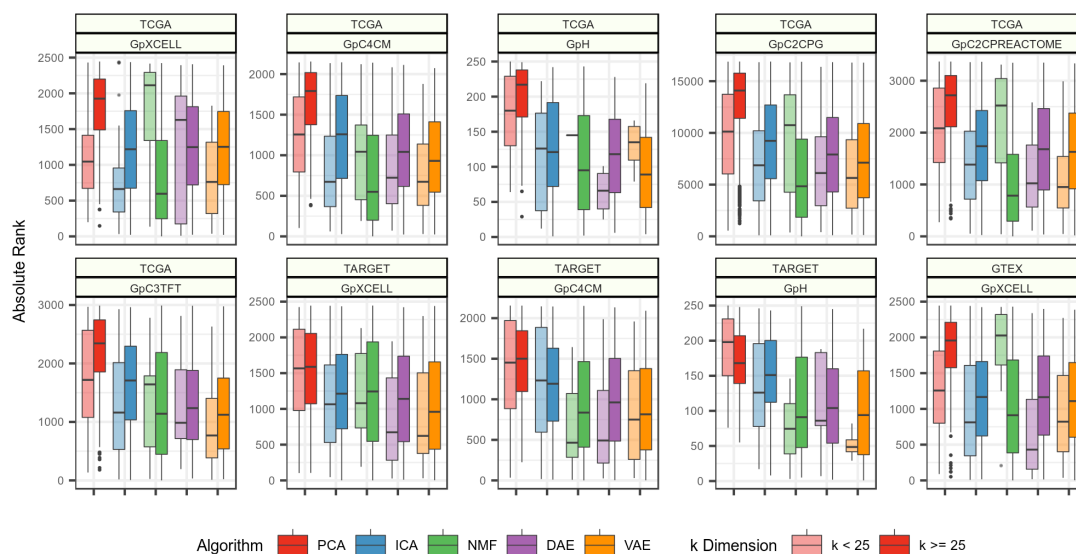


Figure 7.9: Absolute ranking of the top gene set BioBombe z scores across algorithms

We ranked all BioBombe z scores of top scoring gene sets within specific collections across algorithms. All gene sets within a specific collection are visualized within each algorithm box plot and whether or not they were identified as a top feature in model dimensions less than  $k = 25$ .

### 7.3.6. Assessing sample type correlation differences across latent dimensions

We measured the Pearson correlation between all samples' gene expression input and reconstructed output. In TCGA data, we observed increased mean correlation and decreased variance as the latent dimensions increased (Figure 7.11A). We also observed similar patterns in GTEx and TARGET data (Figure 7.12). Across all datasets, in randomly permuted data, we observed correlations near zero (Figure 7.12). The correlation with real data was not consistent across all algorithms as PCA, ICA, and NMF generally outperformed the AE models. We also tracked correlation differences to determine the latent dimensions at which specific sample types could be detected. Most cancer types, including breast invasive carcinoma (BRCA) and colon adenocarcinoma (COAD), displayed relatively gradual increases in sample correlation as the latent dimensions increased (Figure 7.11B). However, in other cancer types, such as low

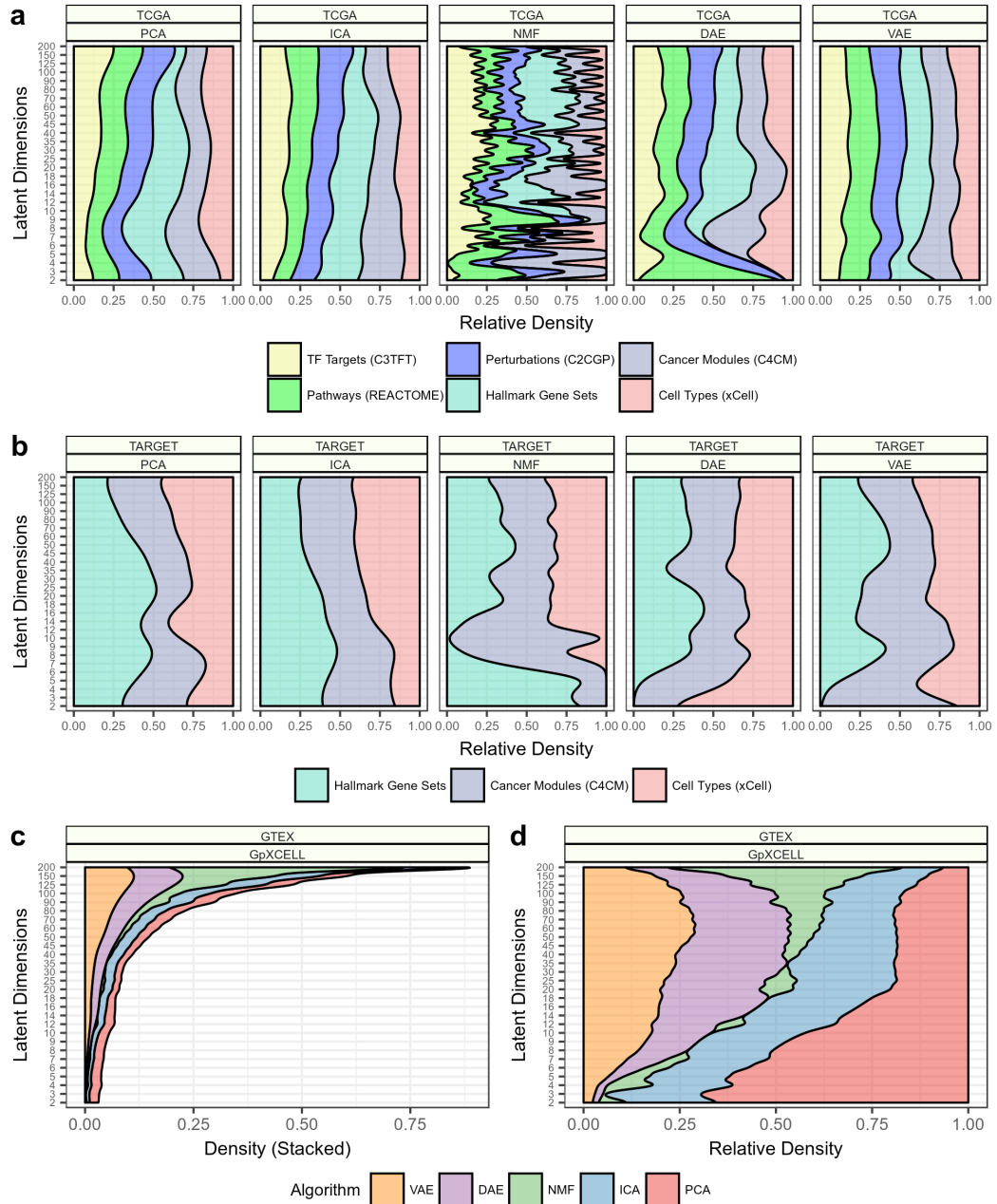


Figure 7.10: Tracking the dimensions of highest BioBombe enrichment signal

The latent space dimension at which a gene set was identified with the highest enrichment across  $k$  dimensions is shown. Observing the relative density of top features identified for several gene set collections across algorithms in (A) TCGA (B) TARGET data. Comparing (C) total counts and (D) relative density of xCell gene sets enrichment across  $k$  dimension in GTEx data.

grade glioma (LGG), pheochromocytoma and paraganglioma (PCPG), and acute myeloid leukemia (LAML), we observed large correlation gains with a single increase in latent dimension (Figure 7.11C). We also observed similar performance spikes in GTEx data for several tissues including liver, pancreas, and blood (Figure 7.11D). This sudden and rapid increase in correlation in specific tissues occurred at different latent dimensions for different algorithms, but was consistent across algorithm initializations. In some cases, certain structure present in data was captured by increasing model capacity by a single  $k$ , but the specific  $k$  at which this happened varied across methods.

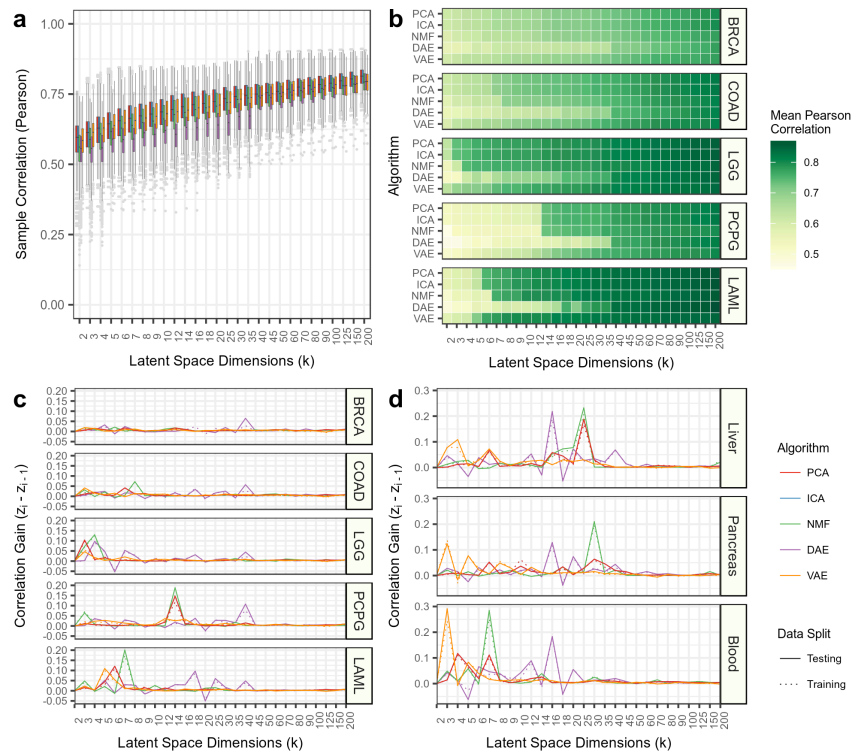


Figure 7.11: Tracking sample correlation across latent dimensions

**(A)** Sample Pearson correlation for all data in the testing data partition for The Cancer Genome Atlas (TCGA). The different algorithms follow the legend provided in panel d. **(B)** Mean Pearson correlation for select cancer types in the testing data partition. Pearson correlation gain between sequential latent dimensions for **(c)** select cancer types in TCGA and **(d)** select tissue-types in GTEx.

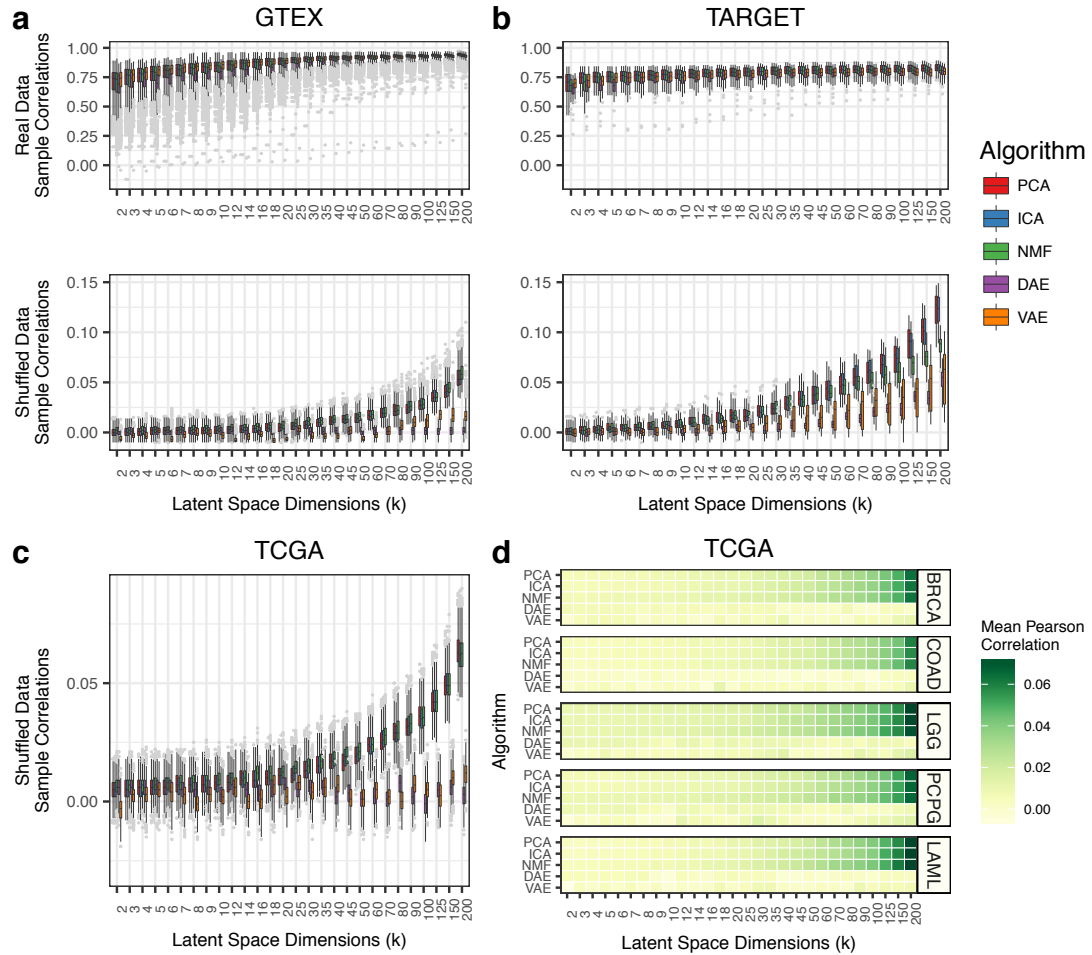


Figure 7.12: Pearson correlation between input and reconstructed samples in real and permuted data

Pearson sample correlation for real (top) and permuted (bottom) data in **(A)** Genome-Tissue Expression Project (GTEx) and **(B)** Therapeutically Applicable Research to Generate Effective Treatments (TARGET). **(C)** Pearson correlations in permuted data from The Cancer Genome Atlas PanCanAtlas Project (TCGA). **(D)** Pearson correlations between input and reconstructed output in permuted data to mirror select cancer-types in Figure 3. The data are permuted before input to the compression algorithms. Results across all specific cancer-types and tissue-types for GTEx, TARGET, and TCGA are provided in: <https://github.com/greenelab/BioBombe/blob/master/4.analyze-components/>

### 7.3.7. Interpretation of GTEx blood with VAE compression features

We examined the sharp increase in GTEx blood tissue correlation observed in VAE models between latent space dimensions 2 and 3 (See Figure 7.11D). We assigned enrichment scores using an xCell gene set network across all compressed features in both VAE models. xCell gene sets represent computationally derived cell type signatures and we sought to identify specific signatures detected by each compressed gene expression feature (291). The top features identified for the VAE  $k = 2$  model included skeletal muscle, keratinocyte, and neuronal gene sets (Figure 7.13A). Skeletal muscle was the likely most significant gene set identified because it is the most represented tissue type in GTEx. Similar gene sets were enriched in the  $k = 3$  model, but we also observed new enrichment for a specific neutrophil gene set (“Neutrophils\_HPCA\_2”) (Figure 7.13A). Neutrophils represent 50% of all cell types in blood, which may explain the increased correlation in blood tissue observed in VAE  $k = 3$  models.

We also calculated the mean absolute value z scores for xCell gene sets in all compression features for VAE models with  $k = 2$  and  $k = 3$  dimensions (Figure 7.13B). Again, we observed skeletal muscle, keratinocytes, and neuronal gene sets to be enriched in both models. Importantly, we also observed a cluster of monocyte gene sets with modest enrichment in  $k = 3$ , but low enrichment in  $k = 2$  (Figure 7.13B). Monocytes are also important cell types found in blood tissue, and it is possible these signatures also contributed to the increased correlation in VAE  $k = 3$  models.

We scanned all other algorithms and latent dimensions to identify other compression features with high enrichment scores in the “Neutrophils\_HPCA\_2” gene set (Figure 7.13C) and “Monocytes\_FANTOM\_2” gene set (Figure 7.13D). We observed the same sharp increase in neutrophil signature enrichment between VAE  $k = 2$  and  $k = 3$  (Figure 7.13C). We also observed stronger enrichment of the “Neutrophil\_HPCA\_2” gene set in

AE models compared to PCA, ICA, and NMF, especially at lower latent dimensions. We observed the highest score for the “Neutrophil\_HPCA\_2” gene set at  $k = 14$  in VAE models (Figure 7.13C). Conversely, PCA, ICA, and NMF identified the “Monocytes\_FANTOM\_2” signature with higher enrichment than the AE models (Figure 7.13D). We also observed a large spike at  $k = 7$  for both PCA and NMF models, but the highest enrichment for “Monocytes\_FANTOM\_2” occurred at  $k = 200$  in NMF models.

#### *7.3.8. Validating GTEx neutrophil and monocyte signatures in external datasets*

We downloaded a processed gene expression dataset (GSE103706) that applied two treatments to induce neutrophil differentiation in two leukemia cell lines (303). We hypothesized that transforming the dataset by the learned “Neutrophil\_HPCA\_2” signature would reveal differential scores in the treated cell lines. We observed large differences in sample activations of treated vs untreated cell lines in the top Neutrophil signature (VAE  $k = 14$ ) (Figure 7.13E). We also tested the “Monocytes\_FANTOM\_2” signature on a different publicly available dataset (GSE24759) measuring gene expression of isolated cell types undergoing hematopoiesis (304). We observed increased scores for isolated monocyte cell population (MONO2) and relatively low scores for several other cell types for top VAE features (Figure 7.13F). Applying all top compressed feature signatures to each dataset, we observed various dimensions and algorithms that optimally isolated differences between each group (Figure 7.13G). These separation patterns were associated with network projection scores in NMF models, but were not consistent in other algorithms (Figure 7.13H). Taken together, we determined that features capturing Neutrophil and Monocyte activity patterns improved signal detection in GTEx blood tissues, signatures are optimally learned at various latent dimensions across algorithms, and that the signatures generalized to datasets that were not encountered during training.

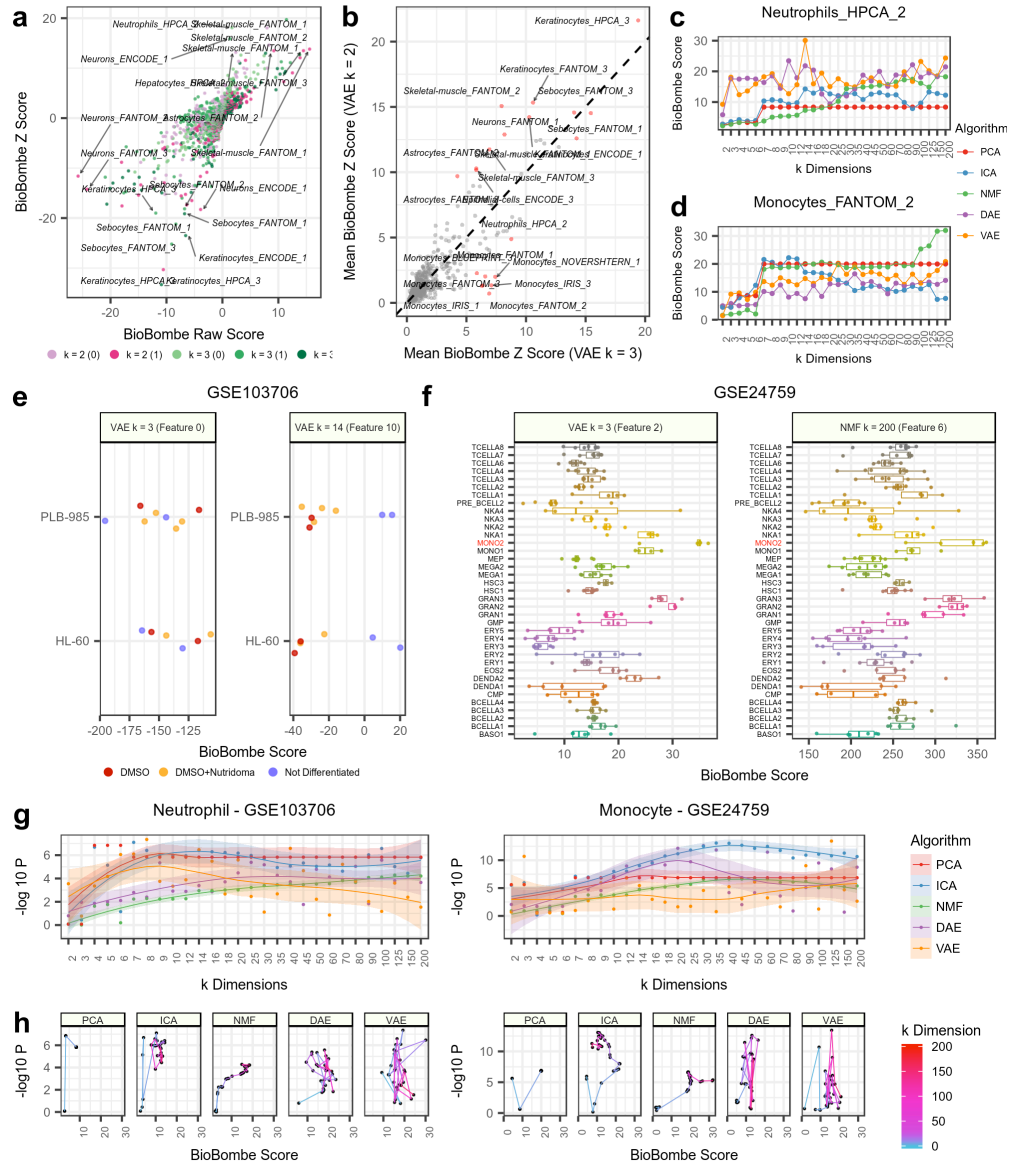


Figure 7.13: Interpreting compressed features learned from GTEx using xCell gene sets

**(A)** Comparing BioBombe scores of all compressed latent features for variational autoencoder (VAE) models when bottleneck dimensions are set to  $k = 2$  and  $k = 3$ . **(B)** Comparing mean BioBombe Z scores of aggregated latent features across two VAE models with  $k$  dimensions 2 and 3. Tracking the BioBombe Z scores of **(C)** “Neutrophils\_HPCA\_2” and **(D)** “Monocytes\_FANTOM\_2” gene sets across dimensions and algorithms. Only the top scoring feature per algorithm and dimension is shown. **(E)** Projecting the VAE feature  $k = 3$  feature and the highest scoring feature (VAE  $k = 14$ ) that best captures a neutrophil signature to an external dataset measuring neutrophil differentiation treatments (GSE103706). **(F)** Projecting the VAE  $k = 3$  feature that best captures monocytes and the feature of the top scoring model (NMF  $k = 200$ ) to an

external dataset of isolated hematopoietic cells (GSE24759). **(G)** Tracking neutrophil and monocyte signatures across all top dimensions. **(H)** Observing how the BioBombe z scores correlated with t test estimates of top dimension correlations.

#### *7.3.9. Using compressed features in supervised learning applications*

We used the latent features generated from the compression algorithms as input features into supervised machine learning tasks. We first trained logistic regression models using the compressed features within each algorithm iteration to predict each of the 33 different cancer types in TCGA. All cancer types could be predicted with high precision and recall using compressed features. We observed multiple performance spikes at varying dimensions for different cancer types and algorithms, and typically in small latent dimensions (Figure 7.14A). We also input the unsupervised compression features into the supervised classification framework to predict samples with alterations in the top 50 most mutated genes in TCGA. We focused on the prediction performance of four cancer genes and one negative control; *TP53*, *PTEN*, *PIK3CA*, *KRAS*, and *TTN* (Figure 7.14B). *TTN* is a particularly large gene and is associated with high passenger mutation burden and should provide no predictive signal (305). As expected, we did not observe any signal in *TTN* across latent dimensions (Figure 7.14B). Again, we observed performance increases at varying model capacities across algorithms. However, predictive signal for mutations occurred at higher latent dimensions compared to cancer types (Figure 7.14C, D). Compared to features trained within algorithm and within iteration, an ensemble of five VAE models and an ensemble of five models representing one iteration of each algorithm, identified cancer type and mutation status in earlier dimensions compared to single model iterations (Figure 7.14C, D).

We also tracked the logistic regression coefficients assigned to each compression feature. Many models were sparse, meaning they included a high percentage of coefficients with zero weights (Figure 7.14E). DAE models consistently displayed sparse



models. The VAE ensemble and model ensemble also induced high sparsity (Figure 7.14E). Lastly, we trained logistic regression classifiers using all 30,850 compressed features generated across iterations, algorithms, and latent dimensions. These logistic regression models were sparse and high performing; comparable to logistic regression models trained using raw features (Figure 7.14E, F, G). Of all 30,850 compressed features in this model, only 317 were assigned non-zero weights (1.03%). We applied the network projection approach with Hallmark gene sets to interpret the biological signatures of the top supervised model coefficients. The top positive feature was derived from a VAE trained with  $k = 200$ . The top hallmarks of this feature included “HALLMARK\_ESTROGEN\_RESPONSE\_EARLY”, “HALLMARK\_ESTROGEN\_RESPONSE\_LATE”, and “HALLMARK\_P53\_PATHWAY”. The top negative feature was derived from a VAE trained with  $k = 150$  and was associated with hallmark genesets including “HALLMARK\_BILE\_ACID\_METABOLISM”, “HALLMARK\_EPITHELIAL\_MESENCHYMAL\_TRANSITION”, and “HALLMARK\_FATTY\_ACID\_METABOLISM”. Overall, the features selected by the logistic regression classifier were distributed across algorithms and latent dimensions suggesting that combining signatures across dimensionalities and algorithms provided the best representation of the signal (Figure 7.14H).

#### **7.4. Discussion**

Unsupervised learning algorithms applied to gene expression data extract biological and technical signals present in input samples. When applying these algorithms, researchers must determine how many latent dimensions to compress their input data into. A study that applies compression algorithms to gene expression data can have a variety of goals. If the goal is visualization, compression algorithms can be used to

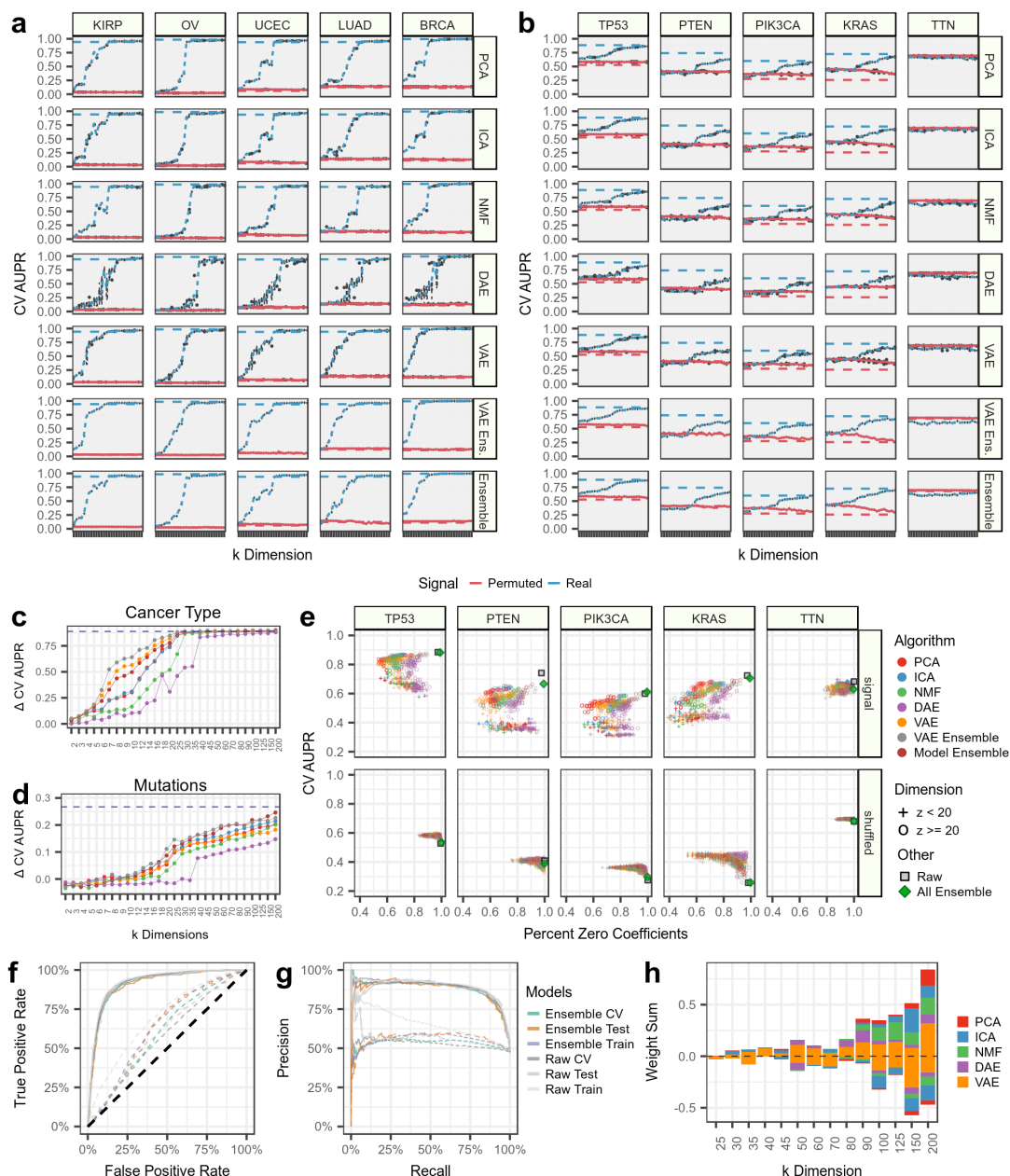


Figure 7.14: Using compressed features as features in supervised machine learning

Predicting **(A)** cancer-type status and **(B)** gene mutation status for select cancer-types and important cancer genes using five compression algorithms and two ensemble models. The area under the precision recall (AUPR) curve for cross validation (CV) data partitions is shown. The blue lines represent predictions made with permuted data input into each compression algorithm. The dotted lines represent AUPR on untransformed RNAseq data. The dotted gray line represents a hypothetical random guess. *TTN* is used as a negative control. Tracking the average change in AUPR between real and permuted data across latent dimensions and compression models in predicting **(C)**

cancer types and **(D)** mutation status. The average includes the five cancer types and mutations tracked in panels a and b. **(E)** Tracking the sparsity and performance of supervised models using BioBombe compressed features in real and permuted data. **(F)** Receiver operating characteristic (ROC), **(G)** PR curves, **(H)** and the average absolute value weight per algorithm for the all-compression-feature ensemble model predicting *TP53* alterations.

stratify samples revealing the largest sources of variation (55, 306–310). For

visualization tasks, selecting a small number of latent dimensions is best, and there is no need for sequential compression. However, if the analysis goals include learning biological signatures that are differentially active in input samples, then there may not be a single optimal latent dimension or optimal algorithm. While it is likely that compressing data into a single latent dimension will capture many biological signals, the “correct” dimension is not always clear, and several biological signatures may be better revealed in alternative latent dimensions.

In the current paradigm, a researcher will use one or many mechanisms to decide upon an optimal latent dimension. Measurements such as Akaike information criterion (AIC), Bayesian information criterion (BIC), stability, and cross validation (CV) can be applied to a series of latent dimensions and a heuristic, like the elbow method, can enable model selection (311, 312). Other algorithms, like Dirichlet processes, can naturally arrive at an appropriate dimension through several algorithm iterations (313). In unsupervised neural networks, hidden layer dimensions are tunable hyperparameters that a user must define based on input data complexity and performance expectations. In recent genomic applications, researchers have used a variety of methods to estimate the latent dimensions. For example, through a combination of outlier detection and PCA, the method Thresher is used to identify optimal number of clusters (314). Stability of compression modules is also considered when determining the optimal number of dimensions (110). Applied to nearly 100,000 publicly available gene expression profiles,

ICA revealed a total of 139 reproducible modules (108). Researchers analyzing a transcriptome compendium of over 5,000 samples determined that only the first three PCA components represented biological signatures (106). However, biological signature discovery was impacted by sample types proportion (107). Instead, we argue that it is best to maximize signature discovery using a sequential compression approach that compresses input gene expression data into many different latent space dimensions.

In an application to predict sample sex and MYCN amplification, we demonstrated that BioBombe maximized biological signature discovery. In each case, various dimensions and different algorithms identified automatically learned biological features at varying association strengths. We also demonstrated that the highest coverage of various gene set collections was achieved by using a combination of models across dimensions and algorithms. We showed that subtle differences in compression model dimensionality impacted identification of tissue specific signatures, including neutrophil and monocyte signatures in blood. Compressed features can also be used to predict cancer type and gene mutations in TCGA gene expression data, and a sparse classifier implicated features across latent dimensions and algorithms. Although performance was higher in models trained using raw gene expression features, compression feature models used less features to generate predictions. These isolated features also offer important clues into the biological processes activated in samples with the specific alteration, and the supervised learning approach can associate these features with specific sample types. The analysis also revealed insight into the impact of latent dimensionality on capturing different biology. For instance, cancer types were predicted with high accuracy, and models arrived at good solutions at low latent dimensions. Conversely, gene mutations were predicted at higher latent dimensions. In genomic applications of supervised learning, the labels of the samples are often inaccurate.

Mutations may be missed by the specific caller, the gene may be activated by alternative means, or there is incomplete knowledge on the pathway being studied (315). Therefore, unsupervised approaches that aggregate validated signatures may also be useful in overcoming sample label limitations.

An additional benefit of compressing gene expression data is to identify novel genes involved in specific biological functions. The compressed features aggregate input signals, and can be used as evidence linking genes together with similar functions. A major benefit of unsupervised algorithms is they do not require external datasets or other resources. Therefore, they can subvert biases present in incomplete gene set collections or other potentially noisy resources. It is possible that many genes we aggregated are part of processes that have been previously undiscovered. For example, in features with high enrichment among specific gene sets, we observed many other unassigned genes with similar weights. Therefore, it is possible that these genes participate in similar biological functions. Additionally, analyzing and extracting knowledge from rapidly expanding publicly available resources will require automated approaches. These approaches can learn signal across different datasets, which can then be applied to smaller datasets that lack power to identify robust biological signatures (84). Extracted from large transcriptomic compendia, compression features can help researchers to interpret and stratify samples in their own datasets. While we did not assess these questions directly, we provide all compression models as publicly available resources for others to test and validate various hypotheses.

Nevertheless, there are many limitations to our approach and analysis. First, our approach takes a long time to run. We are training many different algorithms across many different latent dimensions and iterations, which requires a lot of compute time. However, because we are training many models independently, this task can be

parallelized. Additionally, we did not evaluate dimensions above  $k = 200$ . It is likely that many more signatures can be learned, and possibly with even higher association strengths in higher dimensions. Additionally, we did not explore adding hidden layers in AE models. Many models trained on gene expression data have benefited from using multiple hidden layers in neural network architectures (68, 101). Additional methods, like DeepLift, can be used to reveal gene importance values in internal representations of deep networks (46, 316).

An additional challenge is interpreting the biological content of the compressed gene expression features. Overrepresentation analysis (ORA) and gene set enrichment analysis (GSEA) are commonly applied but have significant limitations (102, 282). ORA requires a user to select a cutoff, typically based on standard deviation, to build representative gene sets from each feature. ORA tests also do not consider the weights, or gene importance scores, in each compression feature. Conversely, GSEA operates on ranked features, but often requires many permutations to establish significance. Furthermore, ORA requires each tail of the compressed feature distribution to be interpreted separately in algorithms that also learn negative weights. The weight distribution is dependent on the specific compression algorithm, and the same cutoff may not be appropriate for all algorithms and all compressed features. Instead, we implemented a network based approach to interpret compressed latent gene expression features (317, 318). The network projection approach is applied to the full and continuous distribution of gene weights, operates independently of the algorithm feature distribution, does not require arbitrary thresholds, and obviates the need to consider both tails of the distribution separately. Nevertheless, additional downstream experimental validation is required to determine if the constructed feature actually represents the biology it has been assigned. We also do not have a mechanism to detect compressed

features that represent technical artifacts. While we showed that compressed signatures representing MYCN amplification, neutrophils, and monocytes generalized to external datasets, more research is required and additional validation should be performed.

The algorithms we used had various tradeoffs. The linear models consistently displayed lower reconstruction costs and higher correlations between input and output samples compared with AE models. The AE models were also not as stable as the linear methods. DAE models were particularly unstable in low latent dimension. However, this likely benefited the AE models in their ability to capture biological signatures in ensemble models. In the NMF models we observed a particularly higher gene set enrichment in high latent dimensions. If training an NMF model on gene expression data, it is best to fit models with many latent dimensions to maximize biological signature discovery. Furthermore, ICA captured the most biological signatures when applying individual models, especially at high latent dimensions. ICA outperformed all other algorithms across datasets and gene set collections. However, when detecting biological signatures using ensemble models, the AE models often outperformed other algorithms, particularly in intermediate latent dimensions. Nevertheless, when combining all models together across latent dimensionalities, we identified nearly 100% of gene sets in many collections. Additionally, the highest performing supervised algorithms used features derived from various algorithms across latent dimensionalities. Therefore, combining features across our BioBombe sequential compression approach optimized biological signature discovery.

### **7.5. Conclusions**

To enhance biological signature discovery, it is best to compress gene expression data using several algorithms and many different latent space dimensionalities. These signatures represent important biological signals including various cell types,

phenotypes, and biomarkers. We present BioBombe as an approach to sequentially compress gene expression data to enhance biological signature discovery. BioBombe can be considered an ensemble of ensemble models that can be used to engineer many different gene expression signatures. We showed, through several experiments tracking gene set coverage and supervised learning performance, that optimal gene expression signatures are learned using a variety of latent space dimensionalities and different compression algorithms. As unsupervised machine learning continues to be applied to derive insight from biomedical datasets, researchers should shift focus away from optimizing a single model based on certain mathematical heuristics, and instead towards learning good, reproducible biological representations that generalize to alternative datasets regardless of compression algorithm and latent dimensionality.

## **7.6. Methods**

### *7.6.1. Transcriptomic compendia acquisition and processing*

We downloaded all transcriptomic compendia from publicly available resources. We downloaded the batch-corrected TCGA PanCanAtlas RNAseq data from the National Cancer Institute Genomic Data Commons (<https://gdc.cancer.gov/about-data/publications/pancanatlas>). These data consisted of 11,069 samples with 20,531 measured genes quantified with RSEM and normalized with log transformation. We converted Hugo Symbol gene identifiers into Entrez gene identifiers and discarded non-protein coding genes and genes that failed to map. We also removed tumors that were measured from multiple sites. This resulted in a final TCGA PanCanAtlas gene expression matrix with 11,060 samples and 16,148 genes, which included 33 different cancer-types.

We downloaded the TPM normalized GTEx RNAseq data (version 7) from the GTEx data portal (<https://gtexportal.org/home/datasets>). There were 11,688 samples and



56,202 genes in this dataset. After selecting only protein-coding genes and converting Hugo Symbols to Entrez gene identifiers, we considered 18,356 genes. There are 53 different detailed tissue-types described in GTEx.

Lastly, we retrieved the TARGET RNAseq gene expression data from the UCSC Xena data portal (127). The TARGET data was processed through the FPKM UCSC Toil RNA-seq pipeline and was normalized with RSEM and log transformed (319). The original matrix consists of 734 samples and 60,498 Ensembl gene identifiers. We converted the Ensembl gene identifiers to Entrez gene names and retained only protein-coding genes. This procedure resulted in a total of 18,753 genes measured in TARGET. There are 7 cancer-types profiled in TARGET. All specific downloading and processing steps can be viewed and reproduced at <https://github.com/greenelab/BioBombe/tree/master/0.expression-download>.

#### *7.6.2. Training unsupervised neural networks*

Autoencoders (AE) are unsupervised neural networks that learn through minimizing the reconstruction of input data after passing the data through one or several intermediate layers (320). Typically, these layers are of a lower dimension than the input, so the algorithms must learn the most important sources of variation in the data. Denoising autoencoders (DAE) add noise to input layers during training to regularize solutions and improve generalizability (89). Variational autoencoders (VAE) add regularization through an additional penalty term imposed on the objective function (90, 91). In a VAE, the latent space dimensions ( $k$ ) are penalized with a Kullback-Leibler (KL) divergence penalty restricting the distribution of samples in the latent space to Gaussian distributions. We independently optimized each AE model across a grid of hyperparameter combinations including 6 representative bottleneck dimensions.

### 7.6.3. *Optimizing training hyperparameters in neural network architectures*

We applied BioBombe using five compression algorithms. Two of the five models, variational autoencoders (VAE) and denoising autoencoders (DAE), are based on autoencoder (AE) frameworks and include several hyperparameters that must be tuned to optimize signal reconstruction. Our primary concern in training the AE models was to ameliorate potential performance biases as the bottleneck dimension increased if hyperparameters were kept static. In other words, we sought to isolate performance differences to the effects of changing  $k$  dimensions. Therefore, we performed a grid search around several hyperparameters for both AE models including 6 representative  $k$  dimensions.

Training autoencoders, and neural networks in general, requires architectural and hyperparameter decisions to optimize learning important signals in input data. We searched through a grid of various combinations of learning rates, epochs, batch sizes, sparsity, and noise parameters for DAE models, and learning rates, epochs, batch sizes, and kappa values for VAE models. We included 6 representative  $k$  dimensions in this grid ( $k = 5, 25, 50, 75, 100$ , and  $125$ ). We selected the hyperparameter combinations for the top performing models and used these in training downstream models.

### 7.6.4. *Training compression algorithms with sequential latent dimensions*

Independently for each dataset (TCGA, GTEx, and TARGET), we performed the following procedure to train the compression algorithms. First, we randomly split data into 90% training and 10% testing partitions. We balanced each partition by cancer type or tissue type, which meant that each split contained relatively equal representation of tissues. Before input into the compression algorithm, we transformed the gene expression values by gene to a range between 0 and 1 independently for the testing and training partitions. We used the training set to train each compression algorithm. We

used the Sci-Kit Learn implementations of PCA, ICA, and NMF, and the Tybalt implementations of VAE and DAE (see Chapter 6 for more details) (93, 131).

After learning optimized compression models with the training data, we transformed the testing data using these models. We assessed performance metrics using both training and testing data to reduce bias. In addition to training with real data, we also trained all models with randomly permuted data. To permute the training data, we randomly shuffled the gene expression values for all genes independently. We also transformed testing partition data with models trained using randomly permuted data. Training with permuted data removes the correlational structure in the data and can help set performance metric baselines.

One of our goals was to assess differences in performance and biological signal detection across a range of latent dimensions ( $k$ ). To this end, we trained all algorithms with various  $k$  dimensionalities including  $k = 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150$ , and 200 for a total of 28 different dimensions. All of these models were trained independently. Lastly, for each  $k$  dimension we trained five different models initialized with five different random seeds. In total, considering the three datasets, five algorithms, randomly permuted training data, all 28  $k$  dimensions, and five initializations, we trained 4,200 different compression models. Therefore, in total, we generated 185,100 different compression features.

#### *7.6.5. Evaluating compression algorithm performance*

We evaluated all compression algorithms on three main tasks: Reconstruction, sample correlation, and weight matrix stability. First, we evaluated how well the input data is reconstructed after passing through the bottleneck layer. Because the input data was transformed to a distribution between 0 and 1, we used binary cross entropy to measure the difference between algorithm input and output as a measure of

reconstruction cost. The lower the reconstruction cost, the higher fidelity reconstruction, and therefore the higher proportion of signals captured in the latent space features. We also assessed the Pearson correlation of all samples comparing input to reconstructed output. This value is similar to reconstruction and can be quickly tracked at an individual sample level. Lastly, we used singular vector canonical correlation analysis (SVCCA) to determine model stability within and model similarity between algorithms and across latent dimensions (299). The SVCCA method consisted of two distinct steps. First, singular value decomposition (SVD) was performed on two input weight matrices. The singular values that combined to reconstruct 98% of the signal in the data were retained. Next, the SVD transformed weight matrix was input into a canonical correlation analysis (CCA). CCA aligned different features in the weight matrix based on maximal correlation after learning a series of linear transformations. Taken together, SVCCA outputs a single metric comparing two input weight matrices that represents stability across model initializations and average similarity of two different models. Because we used the weight matrices, the similarity describes biological signature discovery. We use the distribution of SVCCA similarity measures across all pairwise algorithm initializations and latent dimensionalities to indicate model stability (299).

#### *7.6.7. Using BioBombe as a signature discovery tool*

We tested the ability of BioBombe sequentially compressed features to distinguish sample sex in GTEx and TCGA data, and MYCN amplification in TARGET NBL data. We performed a two-tailed independent t-test assuming equal variance comparing male and female samples, and NBL samples with and without MYCN amplification. We applied the t-test to all compression features identified across algorithms, initializations, and dimensions. Shown in the figures are the top scoring feature per latent space dimension and algorithm.

We applied each optimal signature learned in GTEx, TCGA, and TARGET to alternative datasets. We applied the GTEx sex feature to TCGA data and vice versa. We applied the TARGET MYCN amplification signature to a series of publicly available NBL cell lines (302). The data were processed using STAR, and we accessed the processed FPKM matrix from figshare (321). We transformed the datasets with the identified signatures using the following operation:

$$S_{g'}^T * D_{g' \times n} = D_{s \times n}'$$

Where  $D$  represents the respective RNAseq data to transform,  $S$  represents the specific signature,  $g'$  represents the overlapping genes measured in both datasets,  $n$  represents samples, and  $D'_s$  represents the signature scores in the transformed dataset.

#### 7.6.8. Gene network construction and processing

We constructed networks using gene set collections compiled by version 6.2 of the Molecular Signatures Database (MSigDB) and cell types derived from xCell (102, 291, 292). These gene sets represent a series of genes that are involved in specific biological processes and functions. We integrated all openly licensed MSigDB collections which included hallmark gene sets (H), positional gene sets (C1), curated gene sets (C2), motif gene sets (C3), computational gene sets (C4), Gene Ontology (GO) terms (C5), oncogenic gene sets (C6) and immunologic gene sets (C7). We omitted KEGG, BioCarta, and AAAS/STKE gene sets because of copyright restrictions. The C2 gene set database was split into chemical and genetic perturbations (C2.CPG) and Reactome (C2.CP.Reactome). The C3 gene set was split into microRNA targets (C3.MIR) and transcription factor targets (C3.TFT). The C4 gene set was split into cancer gene neighborhoods (C4.CGN) and cancer modules (C4.CM). Lastly, the C5 gene set was split into GO Biological Processes (C5.BP), GO Cellular Components (C5.CC), and GO

molecular functions (C5.MF). xCell represents a gene set compendia of 489 computationally derived gene signatures from 64 different human cell types. In BioBombe network projection, only a single collection is projected at a time.

To build the gene set network, we used heterogeneous network (hetnet) software (322). Briefly, hetnets are networks that include multiple node types and edge relationships. We used only a single edge relationship in this application, which indicated if a gene participated in a given gene set. We used hetnets to build a single network containing all MSigDB collections and xCell gene sets listed above. The network consisted of 17,451 unique gene sets and 2,159,021 edges representing gene set membership among 20,703 unique gene nodes. In addition to generating a single hetnet using curated gene sets, we also generated 10 permuted hetnets. The hetnets are permuted using the XSwap algorithm, which preserves node degree, or the amount of gene set relationships per gene (324). Therefore, the permuted networks are not restricted by biases induced by uneven gene participation. We compared the real hetnet score and against the distribution of permuted network scores to interpret the biological signatures in each compression feature.

#### 7.6.9. *Rapid interpretation of compressed gene expression data*

Our goal was to quickly interpret the automatically generated compressed latent features learned by each unsupervised algorithm. To this end, we constructed gene set adjacency matrices with specific MSigDB or xCell gene set collections using hetnet software. We then performed the following matrix multiplication against a given compressed weight matrix to obtain a raw score for all gene sets for each latent feature.

$$H_{c \times n} * W_{n \times k} = G_{c \times k}$$

Where  $H$  represents the gene set adjacency matrix,  $c$  is the specific gene set collection, and  $n$  represents genes.  $W$  represents the specific compression algorithm weight matrix, which includes  $n$  genes and  $k$  latent space features. The output of this matrix multiplication,  $G$ , is represented by  $c$  gene sets and  $k$  latent dimensions. Through a single matrix multiplication, the matrix  $G$  tracks raw BioBombe scores.

Because certain hub genes are more likely to be implicated in gene sets and longer gene sets will receive higher raw scores, we compared  $G$  to the distribution of permuted scores against all 10 permuted hetnets.

$$H_p^{1-10} * W_{n \times k} = G_p$$

$$G_{z-score} = \frac{G_{c \times k} - \overline{G_p}}{\sigma(G_p)}$$

Where  $H_p^{1-10}$  represents the adjacency matrices for all 10 permuted hetnets and  $G_p$  represents the distribution of scores for the same  $k$  features for all permutations. We calculated the z score for all gene sets by latent features ( $G_{z-score}$ ). This score represents the BioBombe Score. Other network based gene set methods consider geneset influence based on network connectivity of gene set genes (317, 318). Instead, we used the latent feature weights derived from unsupervised compression algorithms as input, and the compiled gene set networks to assign biological function.

#### 7.6.10. *Calculating gene set coverage of sequentially compressed gene expression data*

We were interested in determining the proportion of gene sets within gene set collections that were captured by the features derived from various compression algorithms. We considered a gene set “captured” by a compression feature if it had the

highest positive or highest negative BioBombe z score compared to all other gene sets in that collection. We converted BioBombe z scores into p values using the `pnorm()` R function using a two-tailed test. We removed gene sets from consideration if their p values were not lower than a Bonferroni adjusted value determined by the total number of  $k$  dimensions in the model. We calculated coverage ( $C$ ) by considering all unique top gene sets ( $U$ ) identified by all features in the compression model ( $w$ ) and dividing by the total number of gene sets in the collection ( $T_c$ ).

$$C = \frac{U_w}{T_c}$$

We calculated the coverage metric for all models independently ( $C_i$ ), for ensembles, or individual algorithms across all five iterations ( $C_e$ ), and for all models across  $k$  dimensions ( $C_k$ ). We also calculated the total coverage of all BioBombe features combined in a single model ( $C_a$ ). A larger coverage value indicated a model that captured a larger proportion of the signatures present in the given gene set collection.

#### *7.6.11. Downloading and processing publicly available expression data for neutrophil GTEx analysis*

We used an external dataset to validate the neutrophil feature that we identified to contribute to detecting blood signatures in GTEx. To assess the performance of this neutrophil signature, we downloaded data from the Gene Expression Omnibus (GEO) with accession number GSE103706 (303). RNA was captured in this dataset using Illumina NextSeq 500. The dataset measured the gene expression of several replicates of two neutrophil-like cell lines, HL-60 and PLB-985, which were originally derived from acute myeloid leukemia (AML) patients. The PLB-985 cell line was previously identified as a subclone of HL-60, so we expect similar signature activity between the two lines



(324). Gene expression of the two cell lines was measured with and without neutrophil differentiation treatments. In this dataset, DMSO treatment was used to induce neutrophil differentiation. Gene expression was also collected in Nutridoma supplemented media, which has also been used to induce neutrophil differentiation. We tested the hypothesis that our neutrophil signature would distinguish the samples with and without neutrophil differentiation treatment. We transformed external datasets with the following operation:

$$W_{k \times g'}^T * D_{g' \times n} = D'_{k \times n}$$

Where  $D$  represents the processed RNAseq data from GSE103706. Of 8,000 genes measured in  $W$ , 7,664 were also measured in  $D$  (95.8%). These 7,664 genes are represented by  $g'$ . All of the “Neutrophils\_HPCA\_2” signature genes were measured in  $W$ .  $D'$  represents the GSE103706 data transformed along the specific compression feature. Each sample in  $D'$  is then considered transformed by the specific signature captured in  $k$ .

#### *7.6.12. Downloading and processing publicly available expression data for monocyte GTEx analysis*

We used an additional external dataset to validate the identified monocyte signature. We accessed processed data for the publicly available GEO dataset with accession number GSE24759 (304). The dataset was measured by Affymetrix HG-U133A (early access array) and consisted of 211 samples representing 38 distinct and purified populations of cells, including monocytes, undergoing various stages of hematopoiesis. The samples were purified from 4 to 7 independent donors each. Many xCell gene sets were computationally derived from this dataset as well (291). Not all genes in the weight matrices were measured in the GSE24759 dataset. For this application, 4,645 genes

(58.06%) corresponded with the genes used in the compression algorithms. Additionally, 168 out of 178 genes (94.38%) in the “Monocyte\_FANTOM\_2” gene set were measured. We investigated the “Monocytes\_FANTOM\_2” signature because of its high enrichment in VAE  $k = 3$  and low enrichment in VAE  $k = 2$ .

#### 7.6.13. *Machine learning classification of cancer types and gene alterations in TCGA*

We trained supervised machine learning models to predict cancer type from RNAseq features in TCGA PanCanAtlas RNAseq data. We implemented a logistic regression classifier with an elastic net penalty. More details about the specific implementation are described in Chapter 3 and in Way et al. 2018 (37). Here, we predicted all 33 cancer types using all 11,060 samples. These predictions were independent per cancer type, which meant that we trained models with the same input gene expression data, but used 33 different status matrices.

We also trained models to predict gene alteration status in the top 50 most mutated genes in the PanCanAtlas. We defined the status in this task using all non-silent mutations identified with a consensus mutation caller (233). We also considered large copy number amplifications for oncogenes and deep copy number deletions for tumor suppressor genes as previously defined (325). We also used the threshold GISTIC2.0 calls for large copy amplifications (score = 2) and deep copy deletions (score = -2) in defining the status matrix (214). For each gene alteration prediction, we removed samples with a hypermutator phenotype, defined by having log10 mutation counts greater than five standard deviations above the mean. For the mutation prediction task, we also did not include certain cancer types in training. We omitted cancer types if they had less than 5% or more than 95% representation of samples with the given gene alteration. The positive and negative sets must have also included at least 15 samples.

We filtered out cancer types in this manner to avoid the classifiers from artificially detecting differences induced by unbalanced training sets.

We trained models with raw RNAseq data subset by the top 8,000 most variably expressed genes by median absolute deviation. The training data used was the same training set used for the sequential compression procedure. We also trained models using all compression matrices for each  $k$  dimension, and using real and permuted data. We combined compressed features together to form three different types of ensemble model. The first type grouped all five iterations of VAE models per latent dimension to make predictions. The second type grouped features of five different algorithms (PCA, ICA, NMF, DAE, VAE) of a single iteration together to make predictions. The third ensemble aggregated all features learned by all algorithms, all initializations, and across all latent dimensions, which included a total of 30,850 features. In total, considering the 33 cancer types, 50 mutations, 28  $k$  dimensions, ensemble models, raw RNAseq features, real and permuted data, and 5 initializations per compression, we trained and evaluated 32,868 different supervised models.

We optimized each model independently using 5-fold cross validation (CV). We searched over a grid of elastic net mixing and alpha hyperparameters. The elastic net mixing parameter represents the tradeoff between  $l_1$  and  $l_2$  penalties (where mixing = 0 represents an  $l_2$  penalty) and controls the sparsity of solutions (20). Alpha is a penalty tuning the impact of regularization, with higher values inducing higher penalties on gene coefficients. We searched over a grid for both hyperparameters (alpha = 0.1, 0.13, 0.15, 0.2, 0.25, 0.3 and mixing = 0.15, 0.16, 0.2, 0.25, 0.3, 0.4) and selected the combination with the highest CV AUROC. For each model, we tested performance using the original held out testing set that was also used to assess compression model performance.

### **7.7. Acknowledgements**

This work was supported by NIH grant T32 HG000046 (GPW), GBMF 4552 from the Gordon and Betty Moore Foundation (CSG), and the National Institutes of Health's National Human Genome Research Institute under R01 HG010067 (CSG). We would like to thank Jaclyn Taroni, Yoson Park, and Alexandra Lee for insightful discussions and code review.

## Chapter 8.

### Conclusions

Signal embedded in transcriptome data can be used to inform disease subtypes, cell type activation patterns, pathway misregulation, response to molecular and environmental perturbation, and many other important biological signatures. By viewing the transcriptome from a systems biological perspective, researchers can develop many different biomedical applications and biological hypotheses. As transcriptome data continues to be generated at a rapid pace, analysis methods to identify patterns and generate hypotheses are becoming increasingly important. Machine learning is one class of tools that can be helpful with these problems.

There are two major classes of machine learning: supervised and unsupervised learning. Both tools are useful in biological applications. Supervised learning can be used to target specific hypotheses. For example, supervised learning can be used to prioritize genes or target compounds, to inform treatment of patients who are likely to respond to specific therapies, and many other important applications. In the first aim of my dissertation (Chapters 2 – 4), we used transcriptome data to distinguish tumors with *NF1*, *TP53*, and Ras pathway aberration. We showed that supervised machine learning can be used to detect wild type Ras cell lines sensitive to MEK inhibitors. Measuring gene mutation status in these cell lines alone would have missed many potential responders to this therapy.

Labels in biomedical and biological applications are not often reliable. Class assignment is noisy, labels are often expensive to acquire, and biological signatures are seldom discrete. Therefore, data driven methods, such as unsupervised machine learning, can be helpful to generate hypotheses and identify patterns of activity in transcriptome data. In the second aim of my dissertation (Chapters 5 and 6), we applied

unsupervised learning to gene expression data. We showed that unsupervised learning can help detect high grade serous ovarian cancer subtypes across different populations. Other, more recently developed methods, often coming from different fields, can be repurposed and used effectively in biological domains. For instance, we trained a variational autoencoder, a model developed primarily for image processing applications, on gene expression data and, leveraging the compressed latent space, isolated continuous gene expression signatures that were differentially active in HGSC subtypes.

While unsupervised learning methods alleviate certain biases present in labelled data, there remain many obstacles to successful applications. Unsupervised models compress input data into a lower dimensional representation that aggregates various biological and technical signatures that describe input data. Major challenges include determining the number of latent features to compress and interpreting the biological signal embedded in compressed features. Therefore, in the third aim of my dissertation (Chapter 7), we developed an approach to sequentially compress gene expression data using several compression algorithms and many different bottleneck dimensions. We constructed gene set and cell type networks to rapidly interpret the biological signatures captured in the compressed features. We observed that different compression algorithms and various latent dimensions capture biological signatures at variable association strengths. Rather than training models to optimize traditional performance metrics, biomedical researchers should shift focus to models that identify useful biological representations. However, experimental validation is essential to confirm that the signatures identified are sound, reproducible, and valuable.

Gene expression data captures important signatures activated in biological data. However, other data types capture additional, and potentially orthogonal, signals as well. For example, measuring DNA methylation, protein, DNA sequence, biological images,

and other data modalities can provide additional insight. There are many data types and views that can be leveraged across biomedical domains. Applying a holistic perspective, and embracing an integrative approach, will aid in the next generation of target discovery, drug development, and healthcare decisions. We are entering an exciting time with many unknowns and a deluge of data. Computational scientists are developing new analysis tools while molecular biologists are developing new data collection methods to measure new and exciting biological data types. As the open science movement continues to grow, tools and data will continue to be shared, and biomedical progress and discoveries can help advance patient care. Machine learning and transcriptomics are currently underused in biomedicine, and they can play an important role in advancing human health. Biomedical integration of different data types and interdisciplinary collaboration will improve the pace of progress.

## BIBLIOGRAPHY

1. Altman RB, Levitt M. 2018. What is Biomedical Data Science and Do We Need an Annual Review of It? *Annu. Rev. Biomed. Data Sci.* 1(1):i–iii
2. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. 2017. Transcriptomics technologies. *PLOS Comput. Biol.* 13(5):e1005457
3. Huang S, Ernberg I, Kauffman S. 2009. Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. *Semin. Cell Dev. Biol.* 20(7):869–76
4. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:
5. Alpaydin E. 2016. *Introduction to Machine Learning: Selected Papers of Lionel W. McKenzie*. Cumberland: MIT Press, The
6. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. 2000. Tissue classification with gene expression profiles. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 7(3–4):559–83
7. Golub TR. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science.* 286(5439):531–37
8. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, et al. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* 8(1):68–74
9. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.* 97(1):262–67



10. Li J, Wong L. 2002. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*. 18(5):725–34
11. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. 2000. Molecular portraits of human breast tumours. *Nature*. 406(6797):747–52
12. Liebermeister W. 2002. Linear modes of gene expression determined by independent component analysis. *Bioinforma. Oxf. Engl.* 18(1):51–60
13. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* 17(5):257–71
14. Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. 2017. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. *Int. J. Mol. Sci.* 18(8):
15. Chibon F. 2013. Cancer gene expression signatures - the rise and fall? *Eur. J. Cancer Oxf. Engl.* 1990. 49(8):2000–2009
16. Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10(1):57–63
17. Leung MKK, DeLong A, Alipanahi B, Frey BJ. 2016. Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. *Proc. IEEE*. 104(1):176–97
18. Kotsiantis S. 2007. *Supervised Machine Learning: A Review of Classification Techniques*, Vol. Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. The Netherlands: IOS Press Amsterdam. 3-24 pp.

19. Tibshirani R. 1994. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B*
20. Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67(2):301–20
21. Breiman L. 2001. *Random Forests*, Vol. 45. Kluwer Academic Publishers. 5-32 pp.
22. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. 1998. Support vector machines. *IEEE Intell. Syst. Their Appl.* 13(4):18–28
23. Pirooznia M, Yang JY, Yang MQ, Deng Y. 2008. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics.* 9(Suppl 1):S13
24. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 415(6871):530–36
25. Venet D, Dumont JE, Detours V. 2011. Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. *PLoS Comput. Biol.* 7(10):e1002240
26. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, et al. 2015. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods.* 12(5):453–57
27. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. 2009. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLoS ONE.* 4(7):e6098
28. Li B, Severson E, Pignon J-C, Zhao H, Li T, et al. 2016. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 17(1):

29. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, et al. 2010. Cell type-specific gene expression differences in complex tissues. *Nat. Methods.* 7(4):287–89
30. Wang M, Master SR, Chodosh LA. 2006. Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics.* 7:328
31. Wang Y, Xia X-Q, Jia Z, Sawyers A, Yao H, et al. 2010. In silico Estimates of Tissue Components in Surgical Samples Based on Expression Profiling Data. *Cancer Res.* 70(16):6448–55
32. Ju W, Greene CS, Eichinger F, Nair V, Hodgins JB, et al. 2013. Defining cell-type specificity at the transcriptional level in human disease. *Genome Res.* 23(11):1862–73
33. Shen-Orr SS, Gaujoux R. 2013. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.* 25(5):571–78
34. Guinney J, Ferte C, Dry J, McEwen R, Manceau G, et al. 2014. Modeling RAS Phenotype in Colorectal Cancer Uncovers Novel Molecular Traits of RAS Dependency and Improves Prediction of Response to Targeted Agents in Patients. *Clin. Cancer Res.* 20(1):265–72
35. Way GP, Allaway RJ, Bouley SJ, Fadul CE, Sanchez Y, Greene CS. 2017. A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma. *BMC Genomics.* 18:127
36. Yang P, Hwa Yang Y, B. Zhou B, Y. Zomaya A. 2010. A Review of Ensemble Methods in Bioinformatics. *Curr. Bioinforma.* 5(4):296–308

37. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, et al. 2018. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. *Cell Rep.* 23(1):172–180.e3
38. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, et al. 2018. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* 23(1):239–254.e6
39. Wilks C, Gaddipati P, Nellore A, Langmead B. 2018. Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics.* 34(1):114–16
40. Turki T, Wei Z. 2018. Boosting support vector machines for cancer discrimination tasks. *Comput. Biol. Med.*
41. Sokolov A, Carlin DE, Paull EO, Baertsch R, Stuart JM. 2016. Pathway-Based Genomics Prediction using Generalized Elastic Net. *PLOS Comput. Biol.* 12(3):e1004790
42. Sokolov A, Paull EO, Stuart JM. 2016. ONE-CLASS DETECTION OF CELL STATES IN TUMOR SUBTYPES. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 21:405–16
43. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, et al. 2018. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell.* 173(2):338–354.e15
44. Yang P, Li X-L, Mei J-P, Kwoh C-K, Ng S-K. 2012. Positive-unlabeled learning for disease gene identification. *Bioinformatics.* 28(20):2640–47
45. Hu Y, Hase T, Li HP, Prabhakar S, Kitano H, et al. 2016. A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. *BMC Genomics.* 17(S13):

46. Lin C, Jain S, Kim H, Bar-Joseph Z. 2017. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res.* 45(17):e156–e156
47. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative Adversarial Networks. *ArXiv14062661 Cs Stat*
48. Marouf M, Machart P, Magruder DSS, Bansal V, Kilian C, et al. 2018. Realistic in silico generation and augmentation of single cell RNA-seq data using Generative Adversarial Neural Networks. *bioRxiv*. <https://doi.org/10.1101/390153>:
49. Ghahramani A, Watt FM, Luscombe NM. 2018. Generative adversarial networks simulate gene expression and predict perturbations in single cells. *bioRxiv*. <https://doi.org/10.1101/262501>:
50. Maaten L van der, Postma E, Herik J van den. 2009. Dimensionality Reduction: A Comparative Review. *Tilburg Cent. Creat. Comput.*
51. Engreitz JM, Daigle BJ, Marshall JJ, Altman RB. 2010. Independent component analysis: Mining microarray data for fundamental human gene expression modules. *J. Biomed. Inform.* 43(6):932–44
52. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci.* 101(12):4164–69
53. Rumelhart DE, Hinton GE, Williams RJ. 1986. *Learning internal representations by error propagation*, Vol. 1. Cambridge, MA, USA: MIT Press. 318-362 pp.
54. Weng L. 2018. From Autoencoder to Beta-VAE. *Lil'Log*. <https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-html>:
55. Maaten L van der, Hinton G. 2008. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9(Nov):2579–2605

56. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, et al. 2014. Multi-platform analysis of 12 cancer types reveals molecular classification within and across tissues-of-origin. *Cell*. 158(4):929–44
57. Way GP, Rudd J, Wang C, Hamidi H, Fridley BL, et al. 2016. Comprehensive Cross-Population Analysis of High-Grade Serous Ovarian Cancer Supports No More Than Three Subtypes. *G3 Genes Genomes Genet*. g3.116.033514
58. Kohonen T. 1990. The self-organizing map. *Proc. IEEE*. 78(9):1464–80
59. Chikina M, Zaslavsky E, Sealfon SC. 2015. CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinforma. Oxf. Engl*. 31(10):1584–91
60. Repsilber D, Kern S, Telaar A, Walzl G, Black GF, et al. 2010. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconvolution approach. *BMC Bioinformatics*. 11(1):27
61. Gaujoux R, Seoighe C. 2012. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. *Infect. Genet. Evol*. 12(5):913–21
62. Ogundijo OE, Wang X. 2017. A sequential Monte Carlo approach to gene expression deconvolution. *PloS One*. 12(10):e0186167
63. Tibshirani R, Hastie T, Narasimhan B, Chu G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci*. 99(10):6567–72
64. Stuart RO, Wachsman W, Berry CC, Wang-Rodriguez J, Wasserman L, et al. 2004. In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proc. Natl. Acad. Sci*. 101(2):615–20

65. Amodio M, van Dijk D, Srinivasan K, Chen WS, Mohsen H, et al. 2019. Exploring Single-Cell Data with Deep Multitasking Neural Networks. *bioRxiv*.  
<http://biorxiv.org/lookup/doi/10.1101/237065>:
66. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. 2019. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* 10(1):
67. Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, et al. 2018. Identifying Gene Expression Programs of Cell-type Identity and Cellular Activity with Single-Cell RNA-Seq. *bioRxiv*. <https://doi.org/10.1101/310599>:
68. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nat. Methods.* 15(12):1053–58
69. Stein-O'Brien GL, Clark BS, Sherman T, Zibetti C, Hu Q, et al. 2018. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *bioRxiv*. <https://doi.org/10.1101/395004>:
70. Stumpf PS, MacArthur BD. 2019. Machine Learning of Stem Cell Identities From Single-Cell Expression Data via Regulatory Network Archetypes. *Front. Genet.* 10:
71. Tarashansky AJ, Xue Y, Quake SR, Wang B. 2018. Self-assembling Manifolds in Single-cell RNA Sequencing Data. *bioRxiv*. <https://doi.org/10.1101/364166>:
72. Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19(1):
73. Grønbech CH, Vording MF, Timshel PN, Sønderby CK, Pers TH, Winther O. 2019. scVAE: Variational auto-encoders for single-cell gene expression data: Supplementary materials. *bioRxiv*. <https://doi.org/10.1101/318295>:

74. Hu Q, Greene CS. 2018. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. *bioRxiv*. <https://doi.org/10.1101/385534>:
75. DeTomaso D, Jones M, Subramaniam M, Ashuach T, Ye CJ, Yosef N. 2018. Functional Interpretation of Single-Cell Similarity Maps. *bioRxiv*. <https://doi.org/10.1101/403055>:
76. Fehrmann RSN, Karjalainen JM, Krajewska M, Westra H-J, Maloney D, et al. 2015. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* 47(2):115–25
77. Frigyesi A, Veerla S, Lindgren D, Höglund M. 2006. Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinformatics*. 7:290
78. Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C. 2007. Elucidating the Altered Transcriptional Programs in Breast Cancer using Independent Component Analysis. *PLoS Comput. Biol.* 3(8):e161
79. Kong W, Vanderburg CR, Gunshin H, Rogers JT, Huang X. 2008. A review of independent component analysis application to microarray gene expression data. *BioTechniques*. 45(5):501–20
80. Li Y, Ngom A. 2013. The non-negative matrix factorization toolbox for biological data mining. *Source Code Biol. Med.* 8(1):10
81. Ochs MF, Fertig EJ. 2012. Matrix Factorization for Transcriptional Regulatory Network Inference. *IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol. Proc. IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol.* 2012:387–96



82. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, et al. 2018. Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet. TIG*. 34(10):790–805
83. Mao W, Harmann B, Sealfon SC, Zaslavsky E, Chikina M. 2017. Pathway-Level Information ExtractoR (PLIER) for gene expression data. *bioRxiv*. <https://doi.org/10.1101/116061>:
84. Taroni JN, Grayson PC, Hu Q, Eddy S, Kretzler M, et al. 2019. MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *bioRxiv*. <https://doi.org/10.1101/395947>:
85. Fertig EJ, Ding J, Favorov AV, Parmigiani G, Ochs MF. 2010. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics*. 26(21):2792–93
86. Tan J, Hammond JH, Hogan DA, Greene CS. 2016. ADAGE-Based Integration of Publicly Available *Pseudomonas aeruginosa* Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. *mSystems*. 1(1):e00025-15
87. Tan J, Doing G, Lewis KA, Price CE, Chen KM, et al. 2017. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. *Cell Syst*. 5(1):63–71.e6
88. Gupta A, Wang H, Ganapathiraju M. 2015. Learning structure in gene expression data using deep architectures, with an application to gene clustering. *2015 IEEE Int. Conf. Bioinforma. Biomed. BIBM*, pp. 1328–35
89. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. 2008. Extracting and composing robust features with denoising autoencoders. *Proc. 25th Int. Conf. Mach. Learn. - ICML 08*, pp. 1096–1103

90. Kingma DP, Welling M. 2013. Auto-Encoding Variational Bayes. *ArXiv13126114 Cs Stat*
91. Rezende DJ, Mohamed S, Wierstra D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *ArXiv14014082 Cs Stat*
92. Ding J, Condon A, Shah SP. 2018. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* 9(1):
93. Way GP, Greene CS. 2018. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 23:80–91
94. Rampasek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A. 2017. Dr.VAE: Drug Response Variational Autoencoder. *ArXiv170608203 Stat*
95. Fabris F, Doherty A, Palmer D, de Magalhães JP, Freitas AA. 2018. A new approach for interpreting Random Forest models and its application to the biology of ageing. *Bioinformatics.* 34(14):2449–56
96. Barardo DG, Newby D, Thornton D, Ghafourian T, de Magalhães JP, Freitas AA. 2017. Machine learning for predicting lifespan-extending chemical compounds. *Aging.* 9(7):1721–37
97. Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46:389–422
98. Zhang HH, Ahn J, Lin X, Park C. 2006. Gene selection using support vector machines with non-convex penalty. *Bioinforma. Oxf. Engl.* 22(1):88–95
99. Vanitha CDA, Devaraj D, Venkatesulu M. 2015. Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. *Procedia Comput. Sci.* 47:13–21

100. Okun O, Priisalu H. 2007. Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues. In *Pattern Recognition and Image Analysis*, ed J Martí, JM Benedí, AM Mendonça, J Serrat. 4478:483–90. Berlin, Heidelberg: Springer Berlin Heidelberg
101. Chen L, Cai C, Chen V, Lu X. 2016. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinformatics*. 17(1):S9
102. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102(43):15545–50
103. Lee S-I, Batzoglou S. 2003. Application of independent component analysis to microarrays. *Genome Biol.* 4(11):R76
104. Lilliefors HW. 1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *J. Am. Stat. Assoc.* 62(318):399
105. Wang H, van der Laan MJ. 2011. Dimension reduction with gene expression data using targeted variable importance measurement. *BMC Bioinformatics*. 12:312
106. Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, et al. 2010. A global map of human gene expression. *Nat. Biotechnol.* 28(4):322–24
107. Lenz M, Müller F-J, Zenke M, Schuppert A. 2016. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. *Sci. Rep.* 6(1):
108. Zhou W, Altman RB. 2018. Data-driven human transcriptomic modules determined by independent component analysis. *BMC Bioinformatics*. 19(1):

109. Cleary B, Cong L, Cheung A, Lander ES, Regev A. 2017. Efficient Generation of Transcriptomic Profiles by Random Composite Measurements. *Cell*. 171(6):1424–1436.e18
110. Wu S, Joseph A, Hammonds AS, Celniker SE, Yu B, Frise E. 2016. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proc. Natl. Acad. Sci.* 113(16):4290–95
111. Gal Y, Ghahramani Z. 2015. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *ArXiv150602142 Cs Stat*
112. Beaulieu-Jones BK, Greene CS. 2017. Reproducibility of computational workflows is automated using continuous analysis. *Nat. Biotechnol.* 35(4):342–46
113. Martin GA, Viskochil D, Bollag G, McCabe PC, Crosier WJ, et al. 1990. The GAP-related domain of the neurofibromatosis type 1 gene product interacts with ras p21. *Cell*. 63(4):843–49
114. Xu G, O'Connell P, Viskochil D, Cawthon R, Robertson M, et al. 1990. The neurofibromatosis type 1 gene encodes a protein related to GAP. *Cell*. 62(3):599–608
115. Boyd KP, Korf BR, Theos A. 2009. Neurofibromatosis type 1. *J. Am. Acad. Dermatol.* 61(1):1–14
116. Dogra B, Rana K. 2013. Facial plexiform neurofibromatosis: A surgical challenge. *Indian Dermatol. Online J.* 4(3):195
117. Evans DGR, Baser M, McGaughan J, Sharif S, Howard E, Moran A. 2002. Malignant peripheral nerve sheath tumours in neurofibromatosis 1. *J. Med. Genet.* 39(5):311–14
118. Rad E, Tee AR. 2016. Neurofibromatosis type 1: Fundamental insights into cell signalling and cancer. *Semin. Cell Dev. Biol.* 52:39–46

119. Ratner N, Miller SJ. 2015. A RASopathy gene commonly mutated in cancer: the neurofibromatosis type 1 tumour suppressor. *Nat. Rev. Cancer.* 15(5):290–301
120. Wood M, Rawe M, Johansson G, Pang S, Soderquist RS, et al. 2011. Discovery of a Small Molecule Targeting IRA2 Deletion in Budding Yeast and Neurofibromin Loss in Malignant Peripheral Nerve Sheath Tumor Cells. *Mol. Cancer Ther.* 10(9):1740–50
121. Allaway RJ, Fischer DA, de Abreu FB, Gardner TB, Gordon SR, et al. 2016. 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.* 7(13):17087–102
122. McGillicuddy LT, Fromm JA, Hollstein PE, Kubek S, Beroukhir R, et al. 2009. Proteasomal and Genetic Inactivation of the NF1 Tumor Suppressor in Gliomagenesis. *Cancer Cell.* 16(1):44–54
123. Subramanian S, Thayanithy V, West RB, Lee C-H, Beck AH, et al. 2010. Genome-wide transcriptome analyses reveal p53 inactivation mediated loss of miR-34a expression in malignant peripheral nerve sheath tumours. *J. Pathol.* 220(1):58–70
124. Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. 1991. A de novo Alu insertion results in neurofibromatosis type 1. *Nature.* 353(6347):864–66
125. Skuse GR, Cappione AJ, Sowden M, Metheny LJ, Smith HC. 1996. The Neurofibromatosis Type I Messenger RNA Undergoes Base-Modification RNA Editing. *Nucleic Acids Res.* 24(3):478–86
126. Cichowski K, Jacks T. 2001. NF1 tumor suppressor gene function: narrowing the GAP. *Cell.* 104(4):593–604

127. Goldman M, Craft B, Kamath A, Brooks AN, Zhu J, Haussler D. 2018. The UCSC Xena Platform for cancer genomics data visualization and interpretation. *bioRxiv*
128. Brennan CW, Verhaak RGW, McKenna A, Campos B, Nounshmehr H, et al. 2013. The Somatic Genomic Landscape of Glioblastoma. *Cell*. 155(2):462–77
129. Thompson JA, Tan J, Greene CS. 2016. Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ*. 4:e1621
130. Liang Y, Liu C, Luan X-Z, Leung K-S, Chan T-M, et al. 2013. Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. *BMC Bioinformatics*. 14(1):198
131. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12(Oct):2825–30
132. Cohen J. 1969. *Statistical power analysis for the behavioral sciences*. New York: Academic Press
133. Allen M, Bjerke M, Edlund H, Nelander S, Westermarck B. 2016. Origin of the U87MG glioma cell line: Good news and bad news. *Sci. Transl. Med.* 8(354):354re3-354re3
134. Dagniakatte GC, Gutmann DH. 2007. Neurofibromatosis-1 (Nf1) heterozygous brain microglia elaborate paracrine factors that promote Nf1-deficient astrocyte and glioma growth. *Hum. Mol. Genet.* 16(9):1098–1112
135. Hollstein PE, Cichowski K. 2013. Identifying the Ubiquitin Ligase Complex That Regulates the NF1 Tumor Suppressor and Ras. *Cancer Discov.* 3(8):880–93
136. Tan X, Wang S, Yang B, Zhu L, Yin B, et al. 2012. The CREB-miR-9 Negative Feedback Minicircuitry Coordinates the Migration and Proliferation of Glioma Cells. *PLoS ONE*. 7(11):e49570

137. Carvalho BS, Irizarry RA. 2010. A framework for oligonucleotide microarray preprocessing. *Bioinforma. Oxf. Engl.* 26(19):2363–67
138. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31(4):e15
139. Reese SE, Archer KJ, Therneau TM, Atkinson EJ, Vachon CM, et al. 2013. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics.* 29(22):2877–83
140. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, et al. 2011. Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics.* 12(1):322
141. Way G. 2016. Greenelab/Nf1\_Inactivation: Pipeline Ready.  
<https://zenodo.org/record/200713>
142. Boettiger C. 2015. An introduction to Docker for reproducible research. *ACM SIGOPS Oper. Syst. Rev.* 49(1):71–79
143. Kamburov A, Wierling C, Lehrach H, Herwig R. 2009. ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Res.* 37(Database issue):D623-628
144. Kamburov A, Stelzl U, Lehrach H, Herwig R. 2013. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* 41(D1):D793–800
145. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25(1):25–29
146. The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43(D1):D1049–56

147. Molla M, Waddell M, Page D, Shavlik J. 2004. Using Machine Learning to Design and Interpret Gene-Expression Microarrays. *AI Mag.* 25(1):23–44
148. Bastani M, Vos L, Asgarian N, Deschenes J, Graham K, et al. 2013. A Machine Learned Classifier That Uses Gene Expression Data to Accurately Predict Estrogen Receptor Status. *PLoS ONE.* 8(12):e82144
149. Chou W-C, Ma Q, Yang S, Cao S, Klingeman DM, et al. 2015. Analysis of strand-specific RNA-seq data using machine learning reveals the structures of transcription units in *Clostridium thermocellum*. *Nucleic Acids Res.* 43(10):e67–e67
150. Noren DP, Long BL, Norel R, Rhissorakrai K, Hess K, et al. 2016. A Crowdsourcing Approach to Developing and Assessing Prediction Algorithms for AML Prognosis. *PLOS Comput. Biol.* 12(6):e1004890
151. Yu B. 2013. Stability. *Bernoulli.* 19(4):1484–1500
152. Chen JL-Y, Merl D, Peterson CW, Wu J, Liu PY, et al. 2010. Lactic Acidosis Triggers Starvation Response with Paradoxical Induction of TXNIP through MondoA. *PLoS Genet.* 6(9):e1001093
153. Willer T, Lee H, Lommel M, Yoshida-Moriguchi T, de Bernabe DBV, et al. 2012. ISPD loss-of-function mutations disrupt dystroglycan O-mannosylation and cause Walker-Warburg syndrome. *Nat. Genet.* 44(5):575–80
154. Thierry-Mieg D, Thierry-Mieg J. 2006. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* 7 Suppl 1:S12.1-14
155. Almog N, Ma L, Raychowdhury R, Schwager C, Erber R, et al. 2009. Transcriptional Switch of Dormant Tumors to Fast-Growing Angiogenic Phenotype. *Cancer Res.* 69(3):836–44



156. Sacco F, Boldt K, Calderone A, Panni S, Paoluzi S, et al. 2014. Combining affinity proteomics and network context to identify new phosphatase substrates and adapters in growth pathways. *Front. Genet.* 5:
157. Xu Y, Chiamvimonvat N, Vázquez AE, Akunuru S, Ratner N, Yamoah EN. 2002. Gene-targeted deletion of neurofibromin enhances the expression of a transient outward K<sup>+</sup> current in Schwann cells: a protein kinase A-mediated mechanism. *J. Neurosci. Off. J. Soc. Neurosci.* 22(21):9194–9202
158. Thouënnon E, Elkahloun AG, Guillemot J, Gimenez-Roqueplo A-P, Bertherat J, et al. 2007. Identification of Potential Gene Markers and Insights into the Pathophysiology of Pheochromocytoma Malignancy. *J. Clin. Endocrinol. Metab.* 92(12):4865–72
159. Cheng Q, Yuan F, Lu F, Zhang B, Chen T, et al. 2015. CSIG promotes hepatocellular carcinoma proliferation by activating c-MYC expression. *Oncotarget.* 6(7):4733–44
160. Bageritz J, Puccio L, Piro RM, Hovestadt V, Phillips E, et al. 2014. Stem cell characteristics in glioblastoma are maintained by the ecto-nucleotidase E-NPP1. *Cell Death Differ.* 21(6):929–40
161. Deng X, Hu Y, Ding Q, Han R, Guo Q, et al. 2014. PEG10 plays a crucial role in human lung cancer proliferation, progression, prognosis and metastasis. *Oncol. Rep.* 32(5):2159–67
162. Li C-M, Margolin AA, Salas M, Memeo L, Mansukhani M, et al. 2006. PEG10 is a c-MYC target gene in cancer cells. *Cancer Res.* 66(2):665–72
163. Akamatsu S, Wyatt AW, Lin D, Lysakowski S, Zhang F, et al. 2015. The Placental Gene PEG10 Promotes Progression of Neuroendocrine Prostate Cancer. *Cell Rep.* 12(6):922–36

164. Sheng Z, Li L, Zhu LJ, Smith TW, Demers A, et al. 2010. A genome-wide RNA interference screen reveals an essential CREB3L2-ATF5-MCL1 survival pathway in malignant glioma with therapeutic implications. *Nat. Med.* 16(6):671–77
165. Greene LA, Lee HY, Angelastro JM. 2009. The transcription factor ATF5: role in neurodevelopment and neural tumors. *J. Neurochem.* 108(1):11–22
166. Zhu Y, Romero MI, Ghosh P, Ye Z, Charnay P, et al. 2001. Ablation of NF1 function in neurons induces abnormal development of cerebral cortex and reactive gliosis in the brain. *Genes Dev.* 15(7):859–76
167. Joseph NM, Mosher JT, Buchstaller J, Snider P, McKeever PE, et al. 2008. The Loss of Nf1 Transiently Promotes Self-Renewal but Not Tumorigenesis by Neural Crest Stem Cells. *Cancer Cell.* 13(2):129–40
168. Morishita A, Zaidi MR, Mitoro A, Sankarasharma D, Szabolcs M, et al. 2013. HMGA2 is a driver of tumor metastasis. *Cancer Res.* 73(14):4289–99
169. de Boeck M, Cui C, Mulder AA, Jost CR, Ikeno S, ten Dijke P. 2016. Smad6 determines BMP-regulated invasive behaviour of breast cancer cells in a zebrafish xenograft model. *Sci. Rep.* 6:24968
170. Salomonis N, Mshel016, Cirillo E, Hanspers K, Kutmon M. Mesodermal Commitment Pathway (Homo sapiens).  
<http://www.wikipathways.org/index.php/Pathway:WP2857>
171. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, et al. 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell.* 17(1):98–110
172. Prasad V. 2016. *Perspective: The precision-oncology illusion.* Nature.  
<https://www.nature.com/articles/537S63a>

173. Prasad V, Fojo T, Brada M. 2016. Precision oncology: origins, optimism, and potential. *Lancet Oncol.* 17(2):e81–86
174. Kumar-Sinha C, Chinnaiyan AM. 2018. Precision oncology in the age of integrative genomics. *Nat. Biotechnol.* 36(1):46
175. Cieřlik M, Chinnaiyan AM. 2018. Cancer transcriptome profiling at the juncture of clinical translation. *Nat. Rev. Genet.* 19(2):93
176. Weinstein JN, Collisson EA, Mills GB, Shaw KM, Ozenberger BA, et al. 2013. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45(10):1113–20
177. Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. 2006. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature.* 439(7074):353–57
178. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, et al. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics.* 26(12):i237–45
179. Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, et al. 2012. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics.* 28(18):i640–46
180. Kim JW, Abudayyeh OO, Yeerna H, Yeang C-H, Stewart M, et al. 2017. Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States. *Cell Syst.* 5(2):105–118.e9
181. Luca AD, Maiello MR, D'Alessio A, Pergameno M, Normanno N. 2012. The RAS/RAF/MEK/ERK and the PI3K/AKT signalling pathways: role in cancer pathogenesis and implications for therapeutic approaches. *Expert Opin. Ther. Targets.* 16(sup2):S17–27

182. McCormick F. 1989. ras GTPase activating protein: Signal transmitter and signal terminator. *Cell*. 56(1):5–8
183. Xu G, O'Connell P, Viskochil D, Cawthon R, Robertson M, et al. 1990. The neurofibromatosis type 1 gene encodes a protein related to GAP. *Cell*. 62(3):599–608
184. Goretzki PE, Lyons J, Stacy-Phipps S, Rosenau W, Demeure M, et al. 1992. Mutational activation of RAS and GSP oncogenes in differentiated thyroid cancer and their biological implications. *World J. Surg.* 16(4):576–81
185. Omholt K, Platz A, Kanter L, Ringborg U, Hansson J. 2003. NRAS and BRAF mutations arise early during melanoma pathogenesis and are preserved throughout tumor progression. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 9(17):6483–88
186. Pao W, Wang TY, Riely GJ, Miller VA, Pan Q, et al. 2005. KRAS Mutations and Primary Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib. *PLOS Med.* 2(1):e17
187. di Magliano MP, Logsdon CD. 2013. Roles for KRAS in Pancreatic Tumor Development and Progression. *Gastroenterology*. 144(6):1220–29
188. Garcia-Rostan G, Zhao H, Camp RL, Pollan M, Herrero A, et al. 2003. ras Mutations Are Associated With Aggressive Tumor Phenotypes and Poor Prognosis in Thyroid Cancer. *J. Clin. Oncol.* 21(17):3226–35
189. Vauthey J-N, Zimmitti G, Kopetz SE, Shindoh J, Chen SS, et al. 2013. RAS mutation status predicts survival and patterns of recurrence in patients undergoing hepatectomy for colorectal liver metastases. *Ann. Surg.* 258(4):

190. Dinu D, Dobre M, Panaitescu E, Bîrlă R, Iosif C, et al. 2014. Prognostic significance of KRAS gene mutations in colorectal cancer - preliminary study. *J. Med. Life.* 7(4):581–87
191. Hsu H-C, Thiam TK, Lu Y-J, Yeh CY, Tsai W-S, et al. 2016. Mutations of KRAS/NRAS/BRAF predict cetuximab resistance in metastatic colorectal cancer patients. *Oncotarget.* 7(16):22257–70
192. Stephen AG, Esposito D, Bagni RK, McCormick F. 2014. Dragging Ras Back in the Ring. *Cancer Cell.* 25(3):272–81
193. Davis J, Goadrich M. 2006. The relationship between Precision-Recall and ROC curves. *Proc. 23rd Int. Conf. Mach. Learn. - ICML 06*, pp. 233–40
194. Masoumi-Moghaddam S, Amini A, Morris DL. 2014. The developing story of Sprouty and cancer. *Cancer Metastasis Rev.* 33(2–3):695
195. Ho C-Y, Bar E, Giannini C, Marchionni L, Karajannis MA, et al. 2013. MicroRNA profiling in pediatric pilocytic astrocytoma reveals biologically relevant targets, including PBX3, NFIB, and METAP2. *Neuro-Oncol.* 15(1):69–82
196. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature.* 483(7391):603–7
197. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, et al. 2002. Mutations of the BRAF gene in human cancer. *Nature.* 417(6892):949–54
198. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, et al. 2017. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 45(D1):D777–83
199. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, et al. 2017. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.*, pp. 1–16

200. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, et al. 2018. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*. 173(2):321–337.e10
201. Jänne PA, Shaw AT, Pereira JR, Jeannin G, Vansteenkiste J, et al. 2013. Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. *Lancet Oncol*. 14(1):38–47
202. Jänne PA, van den Heuvel M, Barlesi F, Cobo M, Mazieres J, et al. 2016. Selumetinib in combination with docetaxel as second-line treatment for patients with KRAS-mutant advanced NSCLC: Results from the phase III SELECT-1 trial. *Ann. Oncol*. 27(suppl\_6):
203. Haura EB, Ricart AD, Larson TG, Stella PJ, Bazhenova L, et al. 2010. A Phase II Study of PD-0325901, an Oral MEK Inhibitor, in Previously Treated Patients with Advanced Non–Small Cell Lung Cancer. *Clin. Cancer Res*. 16(8):2450–57
204. Boasberg PD, Redfern CH, Daniels GA, Bodkin D, Garrett CR, Ricart AD. 2011. Pilot study of PD-0325901 in previously treated patients with advanced melanoma, breast cancer, and colon cancer. *Cancer Chemother. Pharmacol*. 68(2):547–52
205. Farley J, Brady WE, Vathipadiekal V, Lankes HA, Coleman R, et al. 2013. Selumetinib in women with recurrent low-grade serous carcinoma of the ovary or peritoneum: an open-label, single-arm, phase 2 study. *Lancet Oncol*. 2(14):134–40
206. Ho AL, Grewal RK, Leboeuf R, Sherman EJ, Pfister DG, et al. 2013. Selumetinib-Enhanced Radioiodine Uptake in Advanced Thyroid Cancer. *N. Engl. J. Med*. 368(7):623–32

207. O'Neil BH, Goff LW, Kauh JSW, Strosberg JR, Bekaii-Saab TS, et al. 2011. Phase II Study of the Mitogen-Activated Protein Kinase 1/2 Inhibitor Selumetinib in Patients With Advanced Hepatocellular Carcinoma. *J. Clin. Oncol.* 29(17):2350–56
208. Dombi E, Baldwin A, Marcus LJ, Fisher MJ, Weiss B, et al. 2016. Activity of Selumetinib in Neurofibromatosis Type 1–Related Plexiform Neurofibromas. *N. Engl. J. Med.* 375(26):2550–60
209. Jessen WJ, Miller SJ, Jousma E, Wu J, Rizvi TA, et al. 2013. MEK inhibition exhibits efficacy in human and mouse neurofibromatosis tumors. *J. Clin. Invest.* 123(1):340–47
210. Dry JR, Pavey S, Pratilas CA, Harbron C, Runswick S, et al. 2010. Transcriptional Pathway Signatures Predict MEK Addiction and Response to Selumetinib (AZD6244). *Cancer Res.* 70(6):2264–73
211. Oikonomou E, Koustas E, Goulielmaki M, Pintzas A. 2014. BRAF vs RAS oncogenes: are mutations of the same pathway equal? differential signalling and therapeutic implications. *Oncotarget.* 5(23):11752–77
212. Babur Ö, Gönen M, Aksoy BA, Schultz N, Ciriello G, et al. 2015. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol.* 16(1):
213. Mina M, Raynaud F, Tavernari D, Battistello E, Sungalee S, et al. 2017. Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies. *Cancer Cell.* 32(2):155–168.e6
214. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhi R, Getz G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12(4):R41

215. Genetic Testing Registry - NCBI. *Expanded RASopathy Panel (14 Genes)*.  
<https://www.ncbi.nlm.nih.gov/gtr/tests/GTR000521315/>
216. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43(7):e47–e47
217. Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30(1):207–10
218. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 499(7457):214–18
219. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, et al. 2016. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* 34(2):155–63
220. Bahceci I, Dogrusoz U, La KC, Babur O, Gao J, Schultz N. 2017. PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data. *Bioinformatics*
221. Way G, Greene C. 2018. Greenelab/Pancancer: Ras/Nf1 Classifier Submission - PancanAtlas Cell Reports. <https://doi.org/10.5281/zenodo.1186801>
222. Kasthuber ER, Lowe SW. 2017. Putting p53 in Context. *Cell.* 170(6):1062–78
223. Bouaoun L, Sonkin D, Ardin M, Hollstein M, Byrnes G, et al. 2016. TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Hum. Mutat.* 37(9):865–76



224. Olivier M, Hollstein M, Hainaut P. 2010. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harb. Perspect. Biol.* 2(1):a001008–a001008
225. Pfister NT, Prives C. 2017. Transcriptional Regulation by Wild-Type and Cancer-Related Mutant Forms of p53. *Cold Spring Harb. Perspect. Med.* 7(2):
226. Stracquadanio G, Wang X, Wallace MD, Grawenda AM, Zhang P, et al. 2016. The importance of p53 pathway genetics in inherited and somatic cancer genomes. *Nat. Rev. Cancer.* 16(4):251–65
227. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, et al. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* 45(10):1134–40
228. Wilks C, Gaddipati P, Nellore A, Langmead B. 2017. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *bioRxiv*, p. 97881
229. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, et al. 2017. Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* 35(4):319–21
230. Leroy B, Girard L, Hollestelle A, Minna JD, Gazdar AF, Soussi T. 2014. Analysis of TP53 Mutation Status in Human Cancer Cell Lines: A Reassessment. *Hum. Mutat.* 35(6):756–65
231. Sherr CJ. 2001. The INK4a/ARF network in tumour suppression. *Nat. Rev. Mol. Cell Biol.* 2(10):731–37
232. Suwa H, Yoshimura T, Yamaguchi N, Kanehira K, Manabe T, et al. 1994. K-ras and p53 alterations in genomic DNA and transcripts of human pancreatic adenocarcinoma cell lines. *Jpn. J. Cancer Res. Gann.* 85(10):1005–14

233. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandath C, et al. 2018. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* 6(3):271–281.e7
234. Nunobiki O, Ueda M, Toji E, Yamamoto M, Akashi K, et al. 2011. Genetic Polymorphism of Cancer Susceptibility Genes and HPV Infection in Cervical Carcinogenesis. *Pathol. Res. Int.* 2011:
235. Varley JM, Chapman P, McGown G, Thorncroft M, White GR, et al. 1998. Genetic and functional studies of a germline TP53 splicing mutation in a Li-Fraumeni-like family. *Oncogene.* 16(25):3291–98
236. Varley JM, Attwooll C, White G, McGown G, Thorncroft M, et al. 2001. Characterization of germline TP53 splicing mutations and their genetic and functional analysis. *Oncogene.* 20(21):2647–54
237. Kurman RJ, Shih I-M. 2010. The origin and pathogenesis of epithelial ovarian cancer: a proposed unifying theory. *Am. J. Surg. Pathol.* 34(3):433–43
238. Vang R, Shih I-M, Kurman RJ. 2009. Ovarian low-grade and high-grade serous carcinoma: pathogenesis, clinicopathologic and molecular biologic features, and diagnostic problems. *Adv. Anat. Pathol.* 16(5):267–82
239. The Cancer Genome Atlas Research Network. 2011. Integrated genomic analyses of ovarian carcinoma. *Nature.* 474(7353):609–15
240. Tothill RW, Tinker AV, George J, Brown R, Fox SB, et al. 2008. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 14(16):5198–5208
241. Bonome T, Lee J-Y, Park D-C, Radonovich M, Pise-Masison C, et al. 2005. Expression profiling of serous low malignant potential, low-grade, and high-grade tumors of the ovary. *Cancer Res.* 65(22):10602–12

242. Tan TZ, Miow QH, Huang RY-J, Wong MK, Ye J, et al. 2013. Functional genomics identifies five distinct molecular subtypes with clinical relevance and pathways for growth control in epithelial ovarian cancer. *EMBO Mol. Med.* 5(7):1051–66
243. Konecny GE, Wang C, Hamidi H, Winterhoff B, Kalli KR, et al. 2014. Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *J. Natl. Cancer Inst.* 106(10):
244. Verhaak RGW, Tamayo P, Yang J-Y, Hubbard D, Zhang H, et al. 2013. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.* 123(1):517–25
245. Broad Institute TCGA Genome Data Analysis Center. 2016. Analysis overview for ovarian serous cystadenocarcinoma (primary solid tumor cohort). *Broad Inst. MIT Harv.*
246. Broad Institute TCGA Genome Data Analysis Center. 2016. Clustering of mRNA expression: consensus NMF. *Broad Inst. MIT Harv.*
247. Ouellet V, Provencher DM, Maugard CM, Le Page C, Ren F, et al. 2005. Discrimination between serous low malignant potential and invasive epithelial ovarian tumors using molecular profiling. *Oncogene.* 24(29):4672–87
248. Ganzfried BF, Riester M, Haibe-Kains B, Risch T, Tyekucheva S, et al. 2013. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database.* 2013:
249. Köbel M, Kalloger SE, Lee S, Duggan MA, Kelemen LE, et al. 2013. Biomarker-based ovarian carcinoma typing: a histologic investigation in the ovarian tumor tissue analysis consortium. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* 22(10):1677–86

250. Yoshihara K, Tsunoda T, Shigemizu D, Fujiwara H, Hatae M, et al. 2012. High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 18(5):1374–85
251. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. 2014. *cluster: cluster analysis basics and extensions. R package version 1.15.3*
252. Tusher VG, Tibshirani R, Chu G, Tibshirani R, Tusher VG. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* 98(9):5116–21
253. Schwender H, Krause A, Ickstadt K. 2006. Identifying Interesting Genes with siggenes. *RNews.* 6:45–50
254. Schwender H. 2012. *siggenes: multiple testing using SAM and Efron's empirical Bayes approaches*
255. Way G, Rudd J, Greene C. 2015. Analytical code for “cross-population analysis of high-grade serous ovarian cancer reveals only two robust subtypes.”  
<https://zenodo.org/record/32906>
256. Blagden SP. 2015. Harnessing Pandemonium: The Clinical Implications of Tumor Heterogeneity in Ovarian Cancer. *Front. Oncol.* 5:149
257. Silverberg SG. 2000. Histopathologic grading of ovarian carcinoma: a review and proposal. *Int. J. Gynecol. Pathol. Off. J. Int. Soc. Gynecol. Pathol.* 19(1):7–15
258. Soslow RA. 2008. Histologic subtypes of ovarian carcinoma: an overview. *Int. J. Gynecol. Pathol. Off. J. Int. Soc. Gynecol. Pathol.* 27(2):161–74
259. Celik S, Logsdon BA, Battle S, Drescher CW, Rendi M, et al. 2016. Extracting a low-dimensional description of multiple gene expression datasets reveals a potential driver for tumor-associated stroma in ovarian cancer. *Genome Med.* 8:66

260. Planey CR, Gevaert O. 2016. CoINclDE: A framework for discovery of patient subtypes across multiple datasets. *Genome Med.* 8:27
261. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface.* 15(141):
262. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 542(7639):115–18
263. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods.* 12(10):931–34
264. Higgins I, Matthey L, Glorot X, Pal A, Uria B, et al. 2016. Early Visual Concept Learning with Unsupervised Deep Learning. *ArXiv160605579 Cs Q-Bio Stat*
265. Park E. Manifold Learning with Variational Auto-encoder for Medical Image Analysis. [http://www.cs.unc.edu/~eunbyung/papers/manifold\\_variational.pdf](http://www.cs.unc.edu/~eunbyung/papers/manifold_variational.pdf)
266. Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, et al. 2016. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget.* 8(7):10883–90
267. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. 2017. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharm.* 14(9):3098–3104
268. Chaudhary K, Poirion OB, Lu L, Garmire LX. 2018. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 24(6):1248–59
269. Lamb A, Dumoulin V, Courville A. 2016. Discriminative Regularization for Generative Models. *ArXiv160203220 Cs Stat*

270. Ioffe S, Szegedy C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv150203167 Cs*
271. Kingma DP, Ba J. 2014. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs*
272. Nair V, Hinton GE. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, pp. 807–814
273. Chollet F, others. 2015. *Keras*. GitHub
274. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, et al. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *ArXiv160304467 Cs*
275. Doersch C. 2016. Tutorial on Variational Autoencoders. *ArXiv160605908 Cs Stat*
276. Franz K. 2016. Variational Autoencoders Explained. .  
<http://kvfrans.com/variational-autoencoders-explained/>:
277. Saghir H. 2017. An intuitive understanding of variational autoencoders without any formula. . [https://hsaghir.github.io/data\\_science/denoising-vs-variational-autoencoder/](https://hsaghir.github.io/data_science/denoising-vs-variational-autoencoder/):
278. Raiko T, Valpola H, Harva M, Karhunen J. 2007. Building Blocks for Variational Bayesian Learning of Latent Variable Models. *J Mach Learn Res*. 8:155–201
279. Sønderby CK, Raiko T, Maaløe L, Sønderby SK, Winther O. 2016. Ladder Variational Autoencoders. *ArXiv160202282 Cs Stat*
280. Way G. 2016. Data Used For Training Glioblastoma Nf1 Classifier. *zenodo*.  
10.5281/zenodo.56735:
281. Tan J, Ung M, Cheng C, Greene CS. 2015. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with

- denoising autoencoders. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, pp. 132–43
282. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. 2017. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.* 45(W1):W130–37
  283. Dosovitskiy A, Springenberg JT, Brox T. 2015. Learning to generate chairs with convolutional neural networks. *IEEE*.  
<http://ieeexplore.ieee.org/document/7298761/>:1538–46
  284. Radford A, Metz L, Chintala S. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ArXiv151106434 Cs*
  285. Way G, Greene C. 2017. greenelab/tybalt: Initial Development Release. *Zenodo*.  
<https://zenodo.org/record/840199/>:
  286. Cummings J, Ethell BT, Jardine L, Boyd G, Macpherson JS, et al. 2003. Glucuronidation as a Mechanism of Intrinsic Drug Resistance in Human Colon Cancer: Reversal of Resistance by Food Additives. *Cancer Res.* 63(23):8443–50
  287. Cecchin E, Innocenti F, D'Andrea M, Corona G, De Mattia E, et al. 2009. Predictive Role of the UGT1A1, UGT1A7, and UGT1A9 Genetic Variants and Their Haplotypes on the Outcome of Metastatic Colorectal Cancer Patients Treated With Fluorouracil, Leucovorin, and Irinotecan. *J. Clin. Oncol.* 27(15):2457–65
  288. Gaujoux R, Seoighe C. 2013. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinforma. Oxf. Engl.* 29(17):2211–12
  289. GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45(6):580–85

290. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LAA, et al. 2009. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N. Engl. J. Med.* 360(5):470–80
291. Aran D, Hu Z, Butte AJ. 2017. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 18(1):
292. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* 1(6):417–25
293. Way G. 2018. TCGA BioBombe Results. *Zenodo*.  
<https://zenodo.org/record/2110752>:December 9, 2018
294. Way G. 2018. TARGET BioBombe Results. *Zenodo*.  
<https://zenodo.org/record/2222463>:December 12, 2018
295. Way G. 2018. GTEX BioBombe Results. *Zenodo*.  
<https://zenodo.org/record/2300616>:December 15, 2018
296. Way G. 2018. TCGA BioBombe Results - Randomly Permuted Data. *Zenodo*.  
<https://zenodo.org/record/2221216>:December 12, 2018
297. Way G. 2018. TARGET BioBombe Results - Randomly Permuted Data. *Zenodo*.  
<https://zenodo.org/record/2222469>:December 12, 2018
298. Way G. 2018. GTEX BioBombe Results - Randomly Permuted Data. *Zenodo*.  
<https://zenodo.org/record/2386816>:December 18, 2018
299. Raghu M, Gilmer J, Yosinski J, Sohl-Dickstein J. 2017. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. *Neural Inf. Process. Syst. NeurIPS*
300. Clark B, Stein-O’Brien G, Shiao F, Cannon G, Davis E, et al. 2018. Comprehensive analysis of retinal development at single cell resolution identifies



NFI factors as essential for mitotic exit and specification of late-born cells. *bioRxiv*.  
<https://doi.org/10.1101/378950>:

301. Huang M, Weiss WA. 2013. Neuroblastoma and MYCN. *Cold Spring Harb. Perspect. Med.* 3(10):a014415–a014415
302. Harenza JL, Diamond MA, Adams RN, Song MM, Davidson HL, et al. 2017. Transcriptomic profiling of 39 commonly-used neuroblastoma cell lines. *Sci. Data.* 4:170033
303. Rincón E, Rocha-Gregg BL, Collins SR. 2018. A map of gene expression in neutrophil-like cell lines. *BMC Genomics.* 19(1):573
304. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, et al. 2011. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell.* 144(2):296–309
305. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature.* 446(7132):153–58
306. Shi J, Luo Z. 2010. Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. *Comput. Biol. Med.* 40(8):723–32
307. Bartenhagen C, Klein H-U, Ruckert C, Jiang X, Dugas M. 2010. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics.* 11(1):
308. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, et al. 2018. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37(1):38–44
309. Kobak D, Berens P. 2018. The art of using t-SNE for single-cell transcriptomics. *bioRxiv*. <http://biorxiv.org/lookup/doi/10.1101/453449>:

310. McInnes L, Healy J, Melville J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426*
311. Ben-Hur A, Elisseeff A, Guyon I. 2002. A stability based method for discovering structure in clustered data. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, pp. 6–17
312. Wang J. 2010. Consistent selection of the number of clusters via crossvalidation. *Biometrika*. 97(4):893–904
313. Wang L, Wang X. 2013. Hierarchical Dirichlet process model for gene expression clustering. *EURASIP J. Bioinforma. Syst. Biol.* 2013(1):5
314. Wang M, Abrams ZB, Kornblau SM, Coombes KR. 2018. Thresher: determining the number of clusters while removing outliers. *BMC Bioinformatics*. 19(1):9
315. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, et al. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. 173(2):371–385.e18
316. Shrikumar A, Greenside P, Kundaje A. 2017. Learning Important Features Through Propagating Activation Differences. *ArXiv170402685 Cs*
317. Fang Z, Tian W, Ji H. 2012. A network-based gene-weighting approach for pathway analysis. *Cell Res*. 22(3):565–80
318. Dong X, Hao Y, Wang X, Tian W. 2016. LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. *Sci. Rep*. 6(1):
319. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, et al. 2017. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol*. 35(4):314–16

- 320. Baldi P, Hornik K. 1989. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Netw.* 2(1):53–58
- 321. Harenza JL. 2019. Transcriptomic profiling of 39 commonly-used neuroblastoma cell lines. . <https://figshare.com/articles/STAR-reads/7613975/3>:
- 322. Himmelstein DS, Lizée A, Hessler C, Brueggeman L, Chen SL, et al. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*. 6:
- 323. Hanhijärvi S, Garriga GC, Puolamäki K. 2009. Randomization Techniques for Graphs. *Proc. 2009 SIAM Int. Conf. Data Min.*, pp. 780–91
- 324. Drexler HG, Dirks WG, Matsuo Y, MacLeod R a. F. 2003. False leukemia-lymphoma cell lines: an update on over 500 cell lines. *Leukemia*. 17(2):416–26
- 325. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. 2013. Cancer Genome Landscapes. *Science*. 339(6127):1546–58