Hydra: a mixture modeling framework for subtyping pediatric cancer cohorts using multimodal gene expression signatures

Jacob Pfeil[1,2*], Lauren M. Sanders[1,2,3], Ioannis Anastopoulos[1,2] A. Geoffrey Lyle[2,3] Alana S. Weinstein[1,2] Yuanqing Xue[1,2] Andrew Blair[1,2] Holly C. Beale[2,3] Alex Lee[4] Stanley G. Leung[4] Phuong T. Dinh[4] Avanthi Tayi Shah[4] Marcus R. Breese[4] W. Patrick Devine[5] Isabel Bjork[2] Sofie R. Salama[1,2,6] E. Alejandro Sweet-Cordero[4] David Haussler[1,2,6] Olena Morozova Vaske[2,3]

1 Department of Biomolecular Engineering, University of California, Santa Cruz, California, United States of America
2 Genomics Institute, University of California, Santa Cruz, California, United States of America
3 Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, California, United States of America
4 Department of Pediatrics, Division of Hematology and Oncology, University of California, San Francisco, California, United States of America
5 Department of Anatomic Pathology, University of California, San Francisco, California, United States of America
6 Howard Hughes Medical Institute, University of California, Santa Cruz, California, United States of America

* Corresponding author: jpfeil@ucsc.edu

Abstract

Precision oncology has primarily relied on coding mutation status as a readout to define potential therapeutic benefit. Incorporation of transcriptome analysis into precision oncology workflows has proven to be challenging, as relative rather than absolute gene expression level needs to be considered, requiring differential expression analysis across samples. However, cell-of-origin and tumor microenvironment effects limit the effectiveness of these approaches. To address these challenges, we developed an unsupervised clustering approach for discovering differential pathway expression within cancer cohorts using gene expression measurements. Hydra is an unsupervised gene clustering approach that models differential expression as a multi-modal distribution. Multivariate clustering of multimodally expressed genes reveals differential pathway expression and tumor subtype signatures. We demonstrate that the hydra approach is more sensitive than widely-used gene set enrichment approaches for detecting multimodal expression signatures. We applied the hydra pipeline to high-risk neuroblastoma and osteosarcoma samples and discovered expression signatures associated with changes in the tumor microenvironment. These expression signatures were consistent with pathology review of the H&E slide images. Furthermore, we identified an association between ATRX deletions and elevated immune marker expression in high-risk neuroblastoma samples. Hydra is available as a Docker container for easy deployment (`https://hub.docker.com/r/jpfeil/hydra`). The source code is available on GitHub (`www.github.com/jpfeil/hydra`).

Author summary

Our work in pediatric precision oncology found a large number of multimodally expressed genes, but the methods being used in the translational research community often assume a single mode. To fit the data to a single mode leads to several problems, including overestimating uncertainty and potentially leading to spurious results. To enhance clinical utility, bioinformatic tools need to reflect the data being analyzed as accurately as possible. Mixture models that rely on computationally intensive MCMC sampling or do not incorporate prior biological knowledge in appropriate ways. There is a need for a computational approach that can scale to whole-transcriptome datasets, discover tumor subtypes and assign these subtypes to a given patient, and can incorporate available pathway gene set databases. We have developed an approach that uses variational methods at the gene-level to quickly and accurately identify multimodally expressed genes, which we use to define subtypes with respect to user-defined gene sets.

Introduction

Large cancer sequencing projects, including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET), have facilitated the development of cancer gene expression compendia [1–5] but these compendia often lack corresponding normal tissue expression data. Without the the normal comparator, Hoadley et al. (2018) found that cell-of-origin signals drive integrative clustering of TCGA data, with the exception of tumors that have high immune and stromal expression clustering irrespective of the cell-of-origin. Strong cell-of-origin and tumor microenvironment (TME) signals may complicate the interpretation of gene expression results for precision oncology applications, so careful modeling of the data is necessary to infer accurate conclusions from the data.

The TME includes tumor cells, stromal fibroblasts, infiltrating immune cells, and vasculature [6]. Similarities in TME composition across tumor samples have led to the identification of TME states (i.e. inflamed, immune-excluded, immune-desert). These states are dynamic and may change as the tumor progresses, but may shed light on the immunogenicity of tumor cells and correlate with response to cancer immunotherapies [7]. The TME cellular composition can be inferred from gene expression data since host cell RNA is sequenced along with the cancer cell RNA. Not accounting for the immune signal when performing gene expression analysis may confound the interpretation of clustering results [8–10].

Tumor progression and response to therapies is associated with features of the TME, so targeting the TME therapeutically may improve the treatment of some cancers. Immunotherapies that activate the host immune system to eradicate tumors have improved the treatment of several cancer types [11, 12]. Pediatric cancers are thought to be less immunogenic because pediatric cancers have lower mutation burdens in general. Gene overexpression is another source of tumor-specific antigens which may play an important role in pediatric tumor immunogenicity. In addition to infiltrating immune cells, cancer-associated fibroblasts assist in extracellular matrix remodeling and activation of growth factor signaling that facilitate tumor growth, metastasis, and resistance to therapies. Methods are needed to both infer and characterize gene expression subtypes that correlate with tumor microenvironment states to accelerate the development of novel therapies for pediatric cancers. Tumor/normal differential expression analysis in which a cohort of tumor tissues is compared to normal counterparts is an effective approach for identifying gene expression biomarkers [13–15], but it is often not possible to conduct this analysis in a clinical setting. Sufficient

biological and technical replicates are limited by tumor tissue availability, and healthy neighboring tissue often cannot be isolated. In addition, for many pediatric cancers, the cell-of-origin and thus the appropriate reference normal tissue is not known. Besides differential expression analysis, single-sample pathway analysis can be used to identify upregulation of biological gene sets in tumor subtypes. The most widely used pathway analysis approach is gene set enrichment analysis (GSEA). GSEA identifies coordinated expression of pathway genes using gene ranks and a Kolmogorov-Smirnov-like test statistic [?, 16]. GSEA is usually performed on differentially expressed genes, but single-sample versions exist for identifying cancer subtypes. GSEA uses curated pathway gene sets like those in the Molecular Signatures Database (MSigDB) [17].

Cancer gene expression subtypes are traditionally identified using unsupervised clustering analysis such as consensus clustering analysis [?, 18, 19]. These methods are generally underpowered because the number of genes greatly exceeds the number of samples. Dimensionality reduction approaches such as Principal Component Analysis (PCA) have been found to underestimate the dimensionality of gene expression data. Lenz at al. (2016) found two cases in which PCA fails to identify a biological signal: when the size of the cluster is small and when the effect size is small. Lenz et al. (2016) suggest investigating multimodally expressed genes to improve identification of subtypes. Cancer subtypes naturally lead to multimodal expression patterns where each subtype expresses genes with distinct levels and correlation patterns. Furthermore, enriching for multimodally expressed genes before clustering has been shown to improve correlations with clinical features of interest [20].

Gaussian mixture models are a powerful class of unsupervised clustering algorithms for detecting multimodally expressed genes [21]. A Gaussian mixture model is appropriate when the expression data can be modeled by two or more Gaussian distributions [?]. One limitation of Gaussian mixture models in this context is that the number of clusters in the data is often not known beforehand, so a parameter search is used to identify the best-performing model. However, this is a computationally expensive approach. This problem can be overcome by placing a Dirichlet process prior on the number of expression clusters. The number of clusters is then inferred while fitting the mixture model using Markov chain Monte Carlo sampling [?]. This approach has not been widely used in clinical cancer research because these algorithms are also computationally intensive, but recent advances in approximate sampling methods make this approach scalable for precision oncology applications.

$$G \sim DP(M, G_0) \tag{1}$$

$$y_1, y_2, ..., y_n | G \sim G \tag{2}$$

Here, we describe a Dirichlet process mixture model pipeline called hydra for identifying clinically relevant expression subtypes and classifying N-of-1 tumor samples. The hydra pipeline uses the nonparametric Bayesian Python library bnpy [22] and incorporates state-of-the-art cluster profiling software and biological pathway gene sets. We provide an overview of the hydra pipeline, assess performance for detecting differential pathway expression, and apply the approach to better understand expression patterns in high-risk neuroblastoma and osteosarcoma. We propose a novel framework for N-of-1 tumor gene expression analysis and show that this framework can identify distinct immune and stromal expression signatures that differentiate pediatric cancer samples.

Materials and methods

Synthetic Data Generation and Validation

We tested hydra's ability to detect differential pathway expression using synthetic cancer data. We compared hydra to two widely used gene set enrichment tools: single-sample gene set enrichment analysis (ssGSEA) and gene set variation analysis (GSVA) [23–25]. Both methods are implemented in the GSVA R package [24]. We modeled cancer gene expression as a multivariate Gaussian distribution, and then estimated the mean vector and covariance matrix using the TARGET MYCN wild-type neuroblastoma cohort (n=70). This approach allowed us to model pediatric cancer gene expression data while also controlling for subtype-related expression variation. We downloaded the RSEM-quantified TPM normalized gene expression measurements from the UCSC Xena Browser [3]. To reduce heteroscedasticity and the effect of outlier expression levels, we then transformed the expression data to log2(TPM + 1) [26].

We defined an expression subtype as a subset of samples with a distinct expression signature. We modelled expression subtypes across the top 20 most highly expressed MSigDB Hallmark gene sets with at least 100 genes. We tested a range of synthetic data parameters related to the number of differentially expressed genes within a gene set and the effect size for these genes. We randomly generated the covariance matrix for the cancer subtype expression data. We tested the effect of having 10% and 25% of genes within a gene set being differentially expressed (%DEG). In addition to these parameters, we tested a range of effect sizes: 0.25 (least different), 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, and 3.0 (most different). We randomly sampled 225 expression profiles from the compendium multivariate Gaussian distribution and 75 expression profiles from the subtype multivariate Gaussian distribution. This process was repeated twice for each gene set to create synthetic training and test data, which were then analyzed with the hydra pipeline using the supervised gene set clustering mode. The mean expression filter removed any genes with a mean expression of fewer than 1.0 log2(TPM + 1) to avoid lowly-expressed genes that may have unstable expression measurements. The prior on the hydra covariance matrix was the identity scaled by 2.0, and the prior on the number of clusters was 2. We analyzed the enrichment score and posterior probability thresholds using receiver operator curves.

Fig 1. Bold the figure title. Figure caption text here, please use this space for the figure panel descriptions instead of using subfigure commands. A: Lorem ipsum dolor sit amet. B: Consectetur adipiscing elit.

0.1 TARGET Gene Expression Analysis

The TARGET neuroblastoma and osteosarcoma RNA-Seq gene expression data were downloaded from the UCSC Xena Browser (Goldman et al. 2018). The TARGET clinical data were downloaded from the TARGET Data Matrix (https://ocg.cancer.gov/programs/target/data-matrix). We trained the hydra pipeline on 70 International Neuroblastoma Staging System (INSS) stage 4, MYCN non-amplified neuroblastoma and 74 osteosarcoma expression profiles. The resulting models were tested using an independent dataset of pediatric RNA-Seq data shared through the UCSC Treehouse compendium. The Treehouse compendium uses the same gene expression pipeline as the TARGET samples from the UCSC Xena Browser [1].

The neuroblastoma unsupervised enrichment analysis included all genes with a mean expression greater than 1.0 log2(TPM + 1), a minor expression component probability of at least 20%, and a minimum effect size of 1.0. We used a smaller expression

component probability for osteosarcoma to show the ability of the hydra approach to ₁₂₄
discover smaller but clear subtype expression signatures. The osteosarcoma ₁₂₅
unsupervised enrichment analysis included all genes with mean expression greater than ₁₂₆
1.0 log2(TPM + 1), a minor expression component probability of at least 10%, and a ₁₂₇
minimum effect size of 1.0. ClusterProfiler identified statistically significant enrichment ₁₂₈
of GO Biological Process (GOBP) terms (FDR ¡ 0.01) (Yu et al. 2012). The ₁₂₉
multivariate mixture model gamma dispersion parameter was set to 5.0; the prior on ₁₃₀
the covariance matrix was set to the identity scaled by 2.0. The prior parameter for the ₁₃₁
number of clusters was 5 clusters. We correlated hydra expression clusters with the ₁₃₂
results of the tumor microenvironment profiling tools xCell [8] and ESTIMATE [?]. We ₁₃₃
also applied the consensus clustering method M3C [19] to the TARGET neuroblastoma ₁₃₄
and osteosarcoma data using the 5000 genes with the largest median absolute deviation ₁₃₅
(MAD). The number of clusters was inferred to be the smallest statistically significant ₁₃₆
value. ₁₃₇

0.2   Statistical Analysis ₁₃₈

A Kruskal-Wallis test was used to identify statistically significant differences across two ₁₃₉
or more groups, and a Mann-Whitney U test was used for pairwise tests using a ₁₄₀
Holm-Sidak correction for multiple hypothesis testing (Pedregosa et al. 2011; Jones et ₁₄₁
al. 2001). We used the scipy stats implementation of the Kruskal-Wallis test and the ₁₄₂
scikit-learn post hoc processing implementation of pairwise Mann-Whitney U tests. ₁₄₃
Spearman rank and Pearson correlation values were calculated using the scipy library ₁₄₄
(Jones et al. 2001). Correlations between clinical features and clusters were identified ₁₄₅
using the Fisher Exact Test implemented in the R stats package (R Core Team 2013). ₁₄₆
Survival analysis was done using the survminer package (Kassambara et al. 2017). ₁₄₇

0.3   H&E Slide Preparation and Pathologist Review ₁₄₈

Pediatric tumor samples were flash frozen, embedded in OCT, and 5um cryosections ₁₄₉
were collected. Slides were hematoxylin and eosin (H&E) stained and imaged on a Leica ₁₅₀
DMi8, equipped with a HC PL APO 40x/0.85 NA objective and DFC7000T camera. ₁₅₁
H&E slides were reviewed by a licensed pathologist for evidence of inflammation and ₁₅₂
graded as having either minimal (¡ 10%) or moderate inflammation (20-30%). ₁₅₃

0.4   Hydra Method ₁₅₄

We developed a Bayesian non-parametric clustering pipeline for identifying biological ₁₅₅
and technical variation in large cancer gene expression datasets without the need for a ₁₅₆
reference normal dataset (Fig 2). Bayesian non-parametric models have been proposed ₁₅₇
for analyzing gene expression, but to our knowledge, this is the first reproducible and ₁₅₈
widely deployable implementation of a non-parametric mixture model pipeline designed ₁₅₉
for precision oncology gene expression analysis. The hydra pipeline is an open source ₁₆₀
software tool hosted on GitHub ((www.github.com/jpfeil/hydra)). A Docker container ₁₆₁
is available for deployment across environments ((www.dockerhub.com/jpfeil/hydra)). ₁₆₂

Fig 2. Overview of the Hydra Approach. The hydra pipeline uses Bayesian
non-parametric statistics to identify multimodally expressed genes. The pipeline can be
run in two modes. The first mode clusters multimodally expressed genes for
user-defined gene sets using a multivariate mixture model. The second mode searches
for the enrichment of biological gene sets, including Gene Ontology (GO) terms, before
performing multivariate mixture model analysis.

The hydra pipeline uses the bnpy Python library to fit Dirichlet process Gaussian 163
mixture models (DP-GMM) to cancer gene expression data. While most clustering 164
algorithms require setting the number of clusters before fitting the data, the DP-GMM 165
learns the number of clusters while clustering the samples. The Dirichlet process is a 166
distribution for modeling distributions. This feature makes the Dirichlet process an 167
effective prior for the number of clusters in a mixture model. The DP-GMM learns the 168
cluster mean vectors and covariance matrices and assigns each sample to a cluster [22]. 169
This process automates the identification of multimodally expressed genes and 170
multivariate expression signatures. Coordinated expression of multimodally expressed 171
genes, which is encoded in the correlation structure across genes, defines subtype 172
expression signatures. The DP-GMM is also effective when clusters overlap, which is a 173
beneficial property in gene expression data analysis. The parameters learned from 174
fitting cancer gene expression data can be used to classify new samples for precision 175
medicine applications. 176

The hydra pipeline starts with gene expression preprocessing by mean centering and 177
subsetting to genes of interest. The pipeline reads gene set annotations and 178
standardizes gene symbols using the NCBI gene alias annotations [27]. Next, hydra 179
performs multimodal feature selection across all genes of interest. Multimodal feature 180
selection has been shown to improve clustering performance, and the resulting 181
clustering correlates better with clinical features [20]. The multimodal filter removes 182
unimodal expression distributions, selecting for multimodally expressed genes that 183
harbor statistically significant differences in expression in multiple groups of samples 184
within the cohort. The multimodally expressed genes are then used in downstream 185
multivariate clustering. 186

The hydra pipeline can be run in two modes. Mode 1 uses multimodally expressed 187
genes to identify differential pathway expression within user-specified gene sets. Mode 2 188
looks for the enrichment of biological gene sets before multivariate clustering using the 189
clusterProfiler tools [28]. We use the gene set analysis (mode 1) for identifying known 190
gene expression signatures and the gene set enrichment analysis (mode 2) for 191
discovering unknown sources of variation. The pipeline is equipped with commonly used 192
gene sets, including the Molecular Signatures Database (MSigDB) [17], the Gene 193
Ontology Terms [29,30], and the EnrichmentMap gene sets [31]. The gene set database 194
is configurable to the user's research goals and additional gene sets can be added at 195
runtime. 196

The pipeline includes routines for cluster profiling and N-of-1 tumor analysis. 197
Cluster profiling analysis includes GSEA to identify pathway expression that 198
characterizes each cluster. GSEA uses all available genes since these methods require 199
non-differentially expressed genes to assess the significance of an enrichment score. A 200
t-statistic is calculated for each gene comparing gene expression values of samples 201
within and outside of a cluster. Cluster profiling GSEA uses the ranked gene-level 202
t-statistics to determine gene set enrichment. The N-of-1 tumor analysis routine 203
classifies a new gene expression profile into one of the inferred clusters, calculates a 204
gene-level z-score for that sample relative to the normalized expression distribution, and 205
performs GSEA. This procedure can identify subtle gene expression signatures that may 206
not be detectable using the entire expression cohort. 207

Results

0.5 Performance Assessment using Synthetic Gene Expression Data

To assess how well hydra detects differentially expressed pathways as compared to common GSEA approaches, we applied these methods to synthetically-generated cancer gene expression data. We generated synthetic cancer gene expression data based on the TARGET high-risk neuroblastoma cohort and the Hallmark MSigDB gene sets (see Synthetic Data Generation and Validation section). We tested a range of effect sizes and percent differentially expressed genes (%DEG) within the MSigDB gene sets. We generated receiver operator curves (ROC) and calculated the Area Under the receiver operator Curve (AUC) for each analysis. Overall, the hydra pipeline outperformed the single-sample GSEA approaches with a mean AUC of 0.97 (95% CI: 0.96 - 0.98). ssGSEA had a mean AUC of 0.73 (95% CI: 0.72 - 0.75) and GSVA had a mean AUC of 0.69 (95% CI: 0.67 - 0.70) (Figure 2A).

We further investigated the performance of these methods by plotting the AUC against the effect size at %DEG of 10 and 25% (Figure 2B). The hydra method performed better across all effect sizes, achieving near perfect performance above an effect size of 1.0 and 0.75 at %DEG of 10% and 25%, respectively. ssGSEA and GSVA performed similarly at low effect sizes, but ssGSEA performed better at %DEG of 10% and an effect size greater than 1.0. The performance difference between ssGSEA and GSVA was less pronounced at a higher %DEG of 25%. The ssGSEA and GSVA methods began to perform similarly to hydra at an effect size of 3.0 and %DEG of 25%. We next examined how the mean expression for a gene set, as a measure of the baseline expression, correlated with performance. We correlated the mean expression of the gene set with AUC and found that both the ssGSEA and GSVA AUC scores were negatively correlated with the mean expression (Pearson correlation: -0.21 and -0.18). The hydra method was positively correlated (Pearson correlation: 0.16) with the mean expression at low effect sizes (Figure 2C).

0.6 Hydra Analysis of High-Risk Neuroblastoma Identifies Distinct Tumor Microenvironment State

After completing the performance assessment with synthetic gene expression data, we applied the hydra unsupervised enrichment analysis to the TARGET high-risk neuroblastoma cohort. High-risk neuroblastoma is an aggressive disease and is resistant to intensive therapy. Further subtyping of high-risk neuroblastoma may identify novel therapeutic targets and improve risk stratification. We hypothesized that unsupervised clustering of Gene Ontology terms would identify expression subtypes of high-risk neuroblastoma tumors. TumorMap analysis showed that the MYCN-non-amplified (MYCN-NA) neuroblastoma samples clustered separately from MYCN-amplified (MYCN-A) and stage 4S neuroblastomas (Supplementary Figure 1) samples. We focused on the MYCN-NA neuroblastoma tumor samples because this is the largest set of samples (N=70), and variation within MYCN-NA tumors is not well understood [32].

We applied the hydra unsupervised enrichment analysis to the MYCN-NA cohort. The multimodal expression filter identified 428 genes with a minor component probability greater than 20% (Supplementary Table 2). Gene Ontology analysis found enrichment for the following GO terms (FDR q ¡ 0.01): adaptive immune response (24 genes), mesenchyme development (12 genes), steroid hormone secretion (4 genes), and response to corticosterone (4 genes). Multivariate Dirichlet process mixture model analysis of the 44 enriched GO term genes identified three clusters of neuroblastoma samples (Figure 3A). The posterior probability for belonging to each cluster was 42%,

34%, and 17%, respectively. The posterior probability for a sample belonging to a new ²⁵⁷
cluster was about 6% in our analysis. ²⁵⁸

We next applied the cluster profiling gene set enrichment analysis (see Hydra ²⁵⁹
Method section) to each cluster using all genes from the pre-filtered expression matrix. ²⁶⁰
Cluster 1 was enriched for adaptive immune response gene sets, cluster 2 was enriched ²⁶¹
for proliferative signaling gene sets, and cluster 3 was enriched for cancer-associated ²⁶²
fibroblast gene sets (Figure 3B). Two of the three clusters were enriched for tumor ²⁶³
microenvironment-associated expression. To further validate this signal, we correlated ²⁶⁴
the hydra clusters with enrichment scores from the tumor microenvironment profiling ²⁶⁵
tools xCell [8] and ESTIMATE [33]. Cluster 1 had higher average xCell enrichment ²⁶⁶
scores associated with adaptive immune cell types including B-cells, CD4+ naive T-cells, ²⁶⁷
and CD8+ naive T-cells (Kruskal-Wallis: $p < 0.001$; Supplementary Tables 3-5). Cluster ²⁶⁸
2 was characterized by the absence of immune and stromal expression and higher tumor ²⁶⁹
purity. The average ESTIMATE tumor purity for each cluster was 88%, 96% and 82%, ²⁷⁰
respectively. Cluster 3 was enriched for fibroblast-associated expression by xCell ²⁷¹
analysis (Kruskal-Wallis: $p < 0.001$). Clusters 1 and 3 had higher ESTIMATE ²⁷²
immune-associated expression levels than cluster 2 (average ImmuneScore per cluster: ²⁷³
58, -612, 56), but cluster 3 had the highest stromal expression signature (average ²⁷⁴
StromalScore per cluster: -1027, -1310, -135). We found no difference in patient survival ²⁷⁵
outcomes across clusters (log-rank test, $p > 0.05$). We investigated associations with ²⁷⁶
clinical covariates, including mutation burden, age, and tumor content as assessed by a ²⁷⁷
clinical pathologist, but found no statistically significant differences (Kruskal-Wallis: $p >$ ²⁷⁸
0.05; Supplementary Figure 2). We then investigated associations between the hydra ²⁷⁹
clusters and neuroblastoma-associated molecular aberrations and clinical features ²⁸⁰
(Supplementary Table 6). ATRX gene deletions were enriched in cluster 1 (Fisher's ²⁸¹
Exact Test: $p < 0.05$). MKI low tumors were enriched in cluster 2 and 3 (Fisher's Exact ²⁸²
Test: $p < 0.01$). Chromosome 17 wild type tumors were enriched in clusters 2 and 3 ²⁸³
(Fisher's Exact Test: $p < 0.01$). ²⁸⁴

Consensus clustering is a widely used approach for identifying tumor subtypes using ²⁸⁵
gene expression data. We applied the M3C consensus clustering method, which is a ²⁸⁶
more sophisticated version of consensus clustering that uses a null distribution to assess ²⁸⁷
the statistical significance of the clustering (John et al. 2018; Wilkerson and Hayes ²⁸⁸
2010). M3C clustering of the MYCN-NA expression data using the 5000 genes with the ²⁸⁹
largest median absolute deviation (MAD) resulted in the identification of two ²⁹⁰
statistically significant clusters. We found that the M3C clusters correlated with the ²⁹¹
hydra clusters with lower ESTIMATE TumorPurity, but were not able to separate the ²⁹²
adaptive immune cell and fibroblast infiltrated clusters. We also applied kmeans ²⁹³
clustering using the gap statistic approach for estimating the number of clusters, but ²⁹⁴
this approach did not identify any clusters (Tibshirani et al. 2001; Maechler et al. ²⁹⁵
2019).These results suggest that the hydra approach is more sensitive at detecting ²⁹⁶
distinct tumor microenvironment states. ²⁹⁷

0.7   N-of-1 Tumor Analysis for Pediatric Neuroblastoma ²⁹⁸

We investigated the predictive performance of the hydra approach for identifying ²⁹⁹
clinically relevant signals in the N-of-1 tumor analysis setting. We obtained tumor gene ³⁰⁰
expression data from five stage 4, MYCN-NA neuroblastoma samples (Figure 4). The ³⁰¹
age at diagnosis ranged from 2 to 6 years. Four out of five samples had a deletion in the ³⁰²
ATRX gene. Samples 1D and 1R are diagnosis and resection samples from the same ³⁰³
patient (Figure 4). Three of the ATRX-deleted samples clustered with the high immune ³⁰⁴
expression cluster (cluster 1) and one clustered in the low immune, high proliferative ³⁰⁵
signaling cluster (cluster 2). Hydra analysis assigned sample 1D to cluster 1 and sample ³⁰⁶
1R to cluster 2. The resection sample 1R was extracted from lymph node tissue, which ³⁰⁷

has a significantly different immune background than the training data. Another 308
possible explanation for this change is that the tumor microenvironment is dynamic and 309
the tumor may evade immune recognition as the disease progresses. We performed 310
GSEA comparing samples 1D and 1R to investigate potential mechanisms leading to 311
immune evasion in sample 1R. GSEA analysis found downregulation of the MHC Class 312
I Antigen Processing & Presentation GO term in sample 1R (adjusted p-value ¡ 0.002). 313
Loss of antigen processing functions is a common mechanism of immune evasion across 314
cancer types (Reeves and James 2017). 315

We obtained H&E slide images for these samples; the images were reviewed by a 316
licensed pathologist and scored for evidence of inflammation. The N-of-1 predictive 317
function of the hydra pipeline was used to classify samples into the subtypes discovered 318
by the TARGET neuroblastoma analysis. Most of the samples (4 out of 5) clustered in 319
cluster 1, which is characterized by higher immune marker expression. The hydra 320
analysis agreed with the pathologist review in 4 out of 5 samples (Figure 4). Sample 4 321
was the only discordant sample. Sample 4 is also from lymph node tissue, which may 322
have higher immune expression because of the tissue type. Notably, concordant 323
ESTIMATE values were present in 3 out of 5 samples scored by the pathologist: 324
samples 1D, 1R, and 2. 325

To further investigate expression patterns within the hydra-identified tumor 326
microenvironment subtypes, we performed GSEA by z-score normalizing each tumor's 327
gene expression data to its tumor microenvironment cluster. This approach revealed 328
additional signal not present at the cohort level (Supplementary Figure 4). For example, 329
enrichment of immune expression signatures within cluster 2 predicted differences in 330
overall survival such that patients with higher immune expression had a better survival 331
rate. Similarly, an elevated cell cycle signal within cluster 3 predicted worse survival 332
compared to other cluster 3 samples with lower cell cycle expression. This approach 333
provides a more appropriate background distribution for determining the significance of 334
gene expression patterns and survival statistics. 335

0.8   Hydra Analysis Discovers Complex Tissue Signatures 336

While the MYCN-NA neuroblastoma analysis focused on immune and fibroblast 337
expression signatures, the hydra enrichment pipeline is unsupervised and is not 338
restricted to immune or fibroblast signatures. For example, we applied the hydra 339
enrichment analysis to the TARGET osteosarcoma cohort (N=74) and discovered 340
enrichment of the GO striated muscle contraction term (FDR ¡ 0.01) (Supplementary 341
Figure 5). Multivariate clustering for the GO striated muscle contraction gene set 342
identified two clusters. xCell analysis of the osteosarcoma cohort found significant 343
enrichment of skeletal muscle expression in the second cluster (Mann-Whitney U test, p 344
¡ 0.001). Surprisingly, the M3C clustering approach was not able to detect the strong 345
muscle signature using the 5000 genes with the largest MAD (p ¿ 0.05). We identified a 346
similar expression signal in an independent cohort of osteosarcoma tumor samples and 347
subsequently confirmed with a licensed pathologist that the tumor sample did in fact 348
contain muscle tissue. Explaining these sources of variation is necessary to derive 349
clinically relevant conclusions from gene expression data. In addition to the muscle 350
signature, we were able to identify innate immune and stromal subtypes in the 351
TARGET osteosarcoma cohort that correlated with differences in overall survival rates 352
(data not shown). Therefore, the hydra approach reveals important gene expression 353
signatures reflecting cell content within the tumor sample and is widely applicable for 354
revealing important gene expression signatures in complex tumor samples. 355

Discussion   ₃₅₆

The hydra pipeline uses model-based clustering to identify recurrent expression patterns   ₃₅₇
within cancer gene expression cohorts. We leveraged recent improvements in   ₃₅₈
model-based clustering algorithms to identify differentially expressed genes without a   ₃₅₉
matched normal distribution. We modeled differential expression as a multimodal   ₃₆₀
Gaussian distribution using nonparametric Bayesian statistics. We then enriched for   ₃₆₁
biologically-annotated Gene Ontology terms and performed multivariate clustering to   ₃₆₂
reveal cancer subtyping expression signatures. The hydra framework can be used for   ₃₆₃
identifying expression subtypes and classifying N-of-1 tumors. The hydra pipeline   ₃₆₄
outperformed standard gene set enrichment tools for identifying overexpression of the   ₃₆₅
MSigDB Hallmark cancer gene sets in synthetic cancer gene expression data and   ₃₆₆
identified tumor microenvironment gene expression signatures in the TARGET pediatric   ₃₆₇
neuroblastoma and osteosarcoma datasets that were not detected by consensus   ₃₆₈
clustering analysis.   ₃₆₉

Multivariate gene expression analysis is typically underpowered because the number   ₃₇₀
of genes greatly exceeds the number of samples. To address this limitation, we propose   ₃₇₁
selecting for multimodally expressed genes before performing multivariate analysis. The   ₃₇₂
hydra multimodal filter reduces the number of genes and enriches for genes that   ₃₇₃
participate in known biological processes, including those curated in the Gene Ontology   ₃₇₄
database. As Yi Li et al. (2005) found in their study on unsupervised clustering of gene   ₃₇₅
expression data, we also found that reducing expression data to multimodally expressed   ₃₇₆
genes improves clustering of clinical subtypes in pediatric cancers. For example,   ₃₇₇
multimodally expressed genes separate neuroblastoma subtypes by TumorMap analysis   ₃₇₈
better than the standard approach of using all expressed genes. Furthermore, we   ₃₇₉
showed that the hydra approach is more sensitive at resolving tumor microenvironment   ₃₈₀
subtypes than the M3C consensus clustering approach.   ₃₈₁

Significant progress has been made in subtyping neuroblastomas and adapting   ₃₈₂
therapy for aggressive subtypes, but unexplained heterogeneity remains. Not accounting   ₃₈₃
for this heterogeneity decreases the power of standard methods to detect important   ₃₈₄
expression patterns. Identifying biomarkers using genome-wide technology may lead to   ₃₈₅
improved risk stratification and the discovery of novel drug targets [32]. Hydra analysis   ₃₈₆
of the TARGET MYCN-NA neuroblastoma cohort (N=70) found differential expression   ₃₈₇
of tumor microenvironment markers, including markers of the adaptive immune   ₃₈₈
response. Pediatric cancers are generally thought to be non-immunogenic because they   ₃₈₉
have lower mutation burden than adult cancers, but the immunogenicity of pediatric   ₃₉₀
cancer has not been sufficiently investigated [34]. Our analysis found significant   ₃₉₁
variation in immune marker expression and identified ATRX deletions as a potential   ₃₉₂
biomarker of immune infiltrated tumors. Further investigation into gene expression   ₃₉₃
signatures and molecular aberrations that predict response to immunotherapy in   ₃₉₄
pediatric cancers is needed. Hydra analysis may facilitate the development of novel   ₃₉₅
therapies by grouping patients with similar tumor microenvironment properties.   ₃₉₆

We found significant differences in immune and stromal expression that may inform   ₃₉₇
precision medicine applications. The tumor microenvironment has become an important   ₃₉₈
therapeutic consideration, but few methods account for the tumor microenvironment   ₃₉₉
directly. Tumor purity has been identified as a confounding factor in cancer gene   ₄₀₀
expression subtyping efforts [9]. For example, tumor purity and tumor   ₄₀₁
microenvironment expression have been shown to correlate with pancreatic cancer   ₄₀₂
subtypes [35]. Furthermore, Aran et al. (2018) found that tumor purity was correlated   ₄₀₃
with the mesenchymal glioblastoma subtype and recommended a differential expression   ₄₀₄
approach to computationally remove the tumor purity signal. However, standard   ₄₀₅
approaches for subtracting the tumor purity effect may not be the best approach   ₄₀₆
because several mechanisms may influence tumor purity, and each mechanism may   ₄₀₇

result in a different expression pattern. For instance, our analysis of MYCN-NA 408
neuroblastoma identified two gene expression signatures that correlated with lower 409
predicted tumor purity. Cluster 1 had an adaptive immune expression signature and 410
cluster 3 had a cancer-associated fibroblast signature. Therefore, the estimated tumor 411
purity signal should not be subtracted without first accounting for the different 412
mechanisms influencing tumor purity. 413

Conclusion 414

Precision oncology aims to differentiate tumors of the same diagnosis in order to match 415
patients with the best treatment. We have developed the hydra method to discover 416
subtle but recurrent expression patterns within a cancer type. Our approach may help 417
to uncover the biology underlying tumor progression and response to therapy. We have 418
shown that hydra is more sensitive than standard gene set enrichment approaches for 419
detecting differential pathway expression. Additionally, we applied the unsupervised 420
hydra analysis to pediatric neuroblastoma and osteosarcoma data and discovered 421
distinct tumor microenvironment states. The hydra pipeline is a sensitive unsupervised 422
clustering approach for N-of-1 tumor analysis and will facilitate pediatric precision 423
oncology by discovering expression subtype signatures. 424

Supporting information 425

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item 426
(optional). 427

S2 Fig. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. 428
Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. 429
Curabitur fringilla pulvinar lectus consectetur pellentesque. 430

S1 File. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. 431
Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. 432
Curabitur fringilla pulvinar lectus consectetur pellentesque. 433

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. 434
Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. 435
Curabitur fringilla pulvinar lectus consectetur pellentesque. 436

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices 437
gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec 438
euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. 439

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. 440
Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. 441
Curabitur fringilla pulvinar lectus consectetur pellentesque. 442

Acknowledgments 443

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada 444
fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi 445
malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae. 446

References

1. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil Enables Reproducible, Open Source, Big Biomedical Data Analyses. Nature Biotechnology. 2017;35(4):314–316. doi:10.1038/nbt.3772.

2. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, et al. The Genetic Landscape of High-Risk Neuroblastoma. Nature Genetics. 2013;45(3):279–284. doi:10.1038/ng.2529.

3. Goldman M, Craft B, Kamath A, Brooks A, Zhu J, Haussler D. The UCSC Xena Platform for Cancer Genomics Data Visualization and Interpretation. bioRxiv. 2018; p. 326470. doi:10.1101/326470.

4. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. Nature Genetics. 2013;45:1113–1120. doi:10.1038/ng.2764.

5. Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, et al. TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. Cancer Research. 2017;77(21):e111–e114. doi:10.1158/0008-5472.CAN-17-0580.

6. Joyce JA, Fearon DT. T Cell Exclusion, Immune Privilege, and the Tumor Microenvironment. Science (New York, NY). 2015;348(6230):74–80. doi:10.1126/science.aaa6204.

7. Chen DS, Mellman I. Elements of Cancer Immunity and the Cancer–Immune Set Point. Nature. 2017;541(7637):321–330. doi:10.1038/nature21349.

8. Aran D, Hu Z, Butte AJ. xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. Genome Biology. 2017;18(1):220. doi:10.1186/s13059-017-1349-1.

9. Rhee JK, Jung YC, Kim KR, Yoo J, Kim J, Lee YJ, et al. Impact of Tumor Purity on Immune Gene Expression and Clustering Analyses across Multiple Cancer Types. Cancer Immunology Research. 2018;6(1):87–97. doi:10.1158/2326-6066.CIR-17-0201.

10. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173(2):291–304.e6. doi:10.1016/j.cell.2018.03.022.

11. Mellman I, Coukos G, Dranoff G. Cancer Immunotherapy Comes of Age. Nature. 2011;480(7378):480–489. doi:10.1038/nature10673.

12. Page DB, Postow MA, Callahan MK, Allison JP, Wolchok JD. Immune Modulation in Cancer with Antibodies. Annual Review of Medicine. 2014;65:185–202. doi:10.1146/annurev-med-092012-112807.

13. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-Based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor. Nature Protocols. 2013;8(9):1765–1786. doi:10.1038/nprot.2013.099.

14. Anders S, Huber W. Differential Expression Analysis for Sequence Count Data. Genome Biology. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106.

15. Soneson C, Delorenzi M. A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data. BMC Bioinformatics. 2013;14(1):91. doi:10.1186/1471-2105-14-91.

16. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes. Nature Genetics. 2003;34(3):267–273. doi:10.1038/ng1180.

17. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular Signatures Database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–1740. doi:10.1093/bioinformatics/btr260.

18. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, et al. Clustering Algorithms: Their Application to Gene Expression Data. Bioinformatics and Biology Insights. 2016;10:237–253. doi:10.4137/BBI.S38316.

19. John CR, Watson D, Russ D, Goldmann K, Ehrenstein M, Lewis M, et al. M3C: A Monte Carlo Reference-Based Consensus Clustering Algorithm. bioRxiv. 2018; p. 377002.

20. Yi Li, Wing-Kin Sung, Miller LD. Multimodality as a Criterion for Feature Selection in Unsupervised Analysis of Gene Expression Data. In: Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05); 2005. p. 276–280.

21. Ghosh D. Mixture Models for Assessing Differential Expression in Complex Tissues Using Microarray Data. Bioinformatics (Oxford, England). 2004;20(11):1663–1669. doi:10.1093/bioinformatics/bth139.

22. Hughes MC, Sudderth EB. Bnpy : Reliable and Scalable Variational Inference for Bayesian Nonparametric Models; p. 4.

23. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA Interference Reveals That Oncogenic KRAS-Driven Cancers Require TBK1. Nature. 2009;462(7269):108–112. doi:10.1038/nature08460.

24. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. BMC Bioinformatics. 2013;14(1):7. doi:10.1186/1471-2105-14-7.

25. Tarca AL, Bhatti G, Romero R. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. PLOS ONE. 2013;8(11):e79217. doi:10.1371/journal.pone.0079217.

26. Zwiener I, Frisch B, Binder H. Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures. PLOS ONE. 2014;9(1):e85150. doi:10.1371/journal.pone.0085150.

27. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic acids research. 2012;41(D1):D36–D42.

28. Yu G, Wang LG, Han Y, He QY. clusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters. Omics: A Journal of Integrative Biology. 2012;16(5):284–287. doi:10.1089/omi.2011.0118.

29. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for the Unification of Biology. Nature genetics. 2000;25(1):25.

30. Consortium GO. The Gene Ontology Resource: 20 Years and Still GOing Strong. Nucleic acids research. 2018;47(D1):D330–D338.

31. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. PloS one. 2010;5(11):e13984.

32. Morgenstern DA, Bagatell R, Cohn SL, Hogarty MD, Maris JM, Moreno L, et al. The Challenge of Defining "Ultra-High-Risk" Neuroblastoma. Pediatric blood & cancer. 2019;66(4):e27556.

33. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. Nature Communications. 2013;4:2612. doi:10.1038/ncomms3612.

34. Majzner RG, Heitzeneder S, Mackall CL. Harnessing the Immunotherapy Revolution for the Treatment of Childhood Cancers. Cancer Cell. 2017;31(4):476–485.

35. Raphael BJ, Hruban RH, Aguirre AJ, Moffitt RA, Yeh JJ, Stewart C, et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer cell. 2017;32(2):185–203.

References

1. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil Enables Reproducible, Open Source, Big Biomedical Data Analyses. Nature Biotechnology. 2017;35(4):314–316. doi:10.1038/nbt.3772.

2. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, et al. The Genetic Landscape of High-Risk Neuroblastoma. Nature Genetics. 2013;45(3):279–284. doi:10.1038/ng.2529.

3. Goldman M, Craft B, Kamath A, Brooks A, Zhu J, Haussler D. The UCSC Xena Platform for Cancer Genomics Data Visualization and Interpretation. bioRxiv. 2018; p. 326470. doi:10.1101/326470.

4. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. Nature Genetics. 2013;45:1113–1120. doi:10.1038/ng.2764.

5. Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, et al. TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. Cancer Research. 2017;77(21):e111–e114. doi:10.1158/0008-5472.CAN-17-0580.