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.

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 discover smaller but clear subtype expression signatures. The osteosarcoma unsupervised enrichment analysis included all genes with mean expression greater than 1.0 log2(TPM + 1), a minor expression component probability of at least 10%, and a minimum effect size of 1.0. ClusterProfiler identified statistically significant enrichment

of GO Biological Process (GOBP) terms (FDR ¡ 0.01) (Yu et al. 2012). The multivariate mixture model gamma dispersion parameter was set to 5.0; the prior on the covariance matrix was set to the identity scaled by 2.0. The prior parameter for the number of clusters was 5 clusters. We correlated hydra expression clusters with the results of the tumor microenvironment profiling tools xCell [8] and ESTIMATE [**?**]. We also applied the consensus clustering method M3C [19] to the TARGET neuroblastoma and osteosarcoma data using the 5000 genes with the largest median absolute deviation (MAD). The number of clusters was inferred to be the smallest statistically significant value.

0.2 Statistical Analysis

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

0.3 H&E Slide Preparation and Pathologist Review

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

0.4 Hydra Method

We developed a Bayesian non-parametric clustering pipeline for identifying biological and technical variation in large cancer gene expression datasets without the need for a reference normal dataset (Fig 1). Bayesian non-parametric models have been proposed for analyzing gene expression, but to our knowledge, this is the first reproducible and widely deployable implementation of a non-parametric mixture model pipeline designed for precision oncology gene expression analysis. The hydra pipeline is an open source software tool hosted on GitHub ((www.github.com/jpfeil/hydra)). A Docker container is available for deployment across environments ((www.dockerhub.com/jpfeil/hydra)).

Fig 1. 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 mixture models (DP-GMM) to cancer gene expression data. While most clustering algorithms require setting the number of clusters before fitting the data, the DP-GMM learns the number of clusters while clustering the samples. The Dirichlet process is a distribution for modeling distributions. This feature makes the Dirichlet process an

effective prior for the number of clusters in a mixture model. The DP-GMM learns the cluster mean vectors and covariance matrices and assigns each sample to a cluster [22]. This process automates the identification of multimodally expressed genes and multivariate expression signatures. Coordinated expression of multimodally expressed genes, which is encoded in the correlation structure across genes, defines subtype expression signatures. The DP-GMM is also effective when clusters overlap, which is a beneficial property in gene expression data analysis. The parameters learned from fitting cancer gene expression data can be used to classify new samples for precision medicine applications.

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

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

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

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 States

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 cluster was about 6% in our analysis.

We next applied the cluster profiling gene set enrichment analysis (see Hydra Method section) to each cluster using all genes from the pre-filtered expression matrix. Cluster 1 was enriched for adaptive immune response gene sets, cluster 2 was enriched for proliferative signaling gene sets, and cluster 3 was enriched for cancer-associated fibroblast gene sets (Figure 3B). Two of the three clusters were enriched for tumor microenvironment-associated expression. To further validate this signal, we correlated the hydra clusters with enrichment scores from the tumor microenvironment profiling tools xCell [8] and ESTIMATE [33]. Cluster 1 had higher average xCell enrichment

scores associated with adaptive immune cell types including B-cells, CD4+ naive T-cells, and CD8+ naive T-cells (Kruskal-Wallis: p ¡ 0.001; Supplementary Tables 3-5). Cluster 2 was characterized by the absence of immune and stromal expression and higher tumor purity. The average ESTIMATE tumor purity for each cluster was 88%, 96% and 82%, respectively. Cluster 3 was enriched for fibroblast-associated expression by xCell analysis (Kruskal-Wallis: p ¡ 0.001). Clusters 1 and 3 had higher ESTIMATE immune-associated expression levels than cluster 2 (average ImmuneScore per cluster: 58, -612, 56), but cluster 3 had the highest stromal expression signature (average StromalScore per cluster: -1027, -1310, -135). We found no difference in patient survival outcomes across clusters (log-rank test, p ¿ 0.05). We investigated associations with clinical covariates, including mutation burden, age, and tumor content as assessed by a clinical pathologist, but found no statistically significant differences (Kruskal-Wallis: p ¿ 0.05; Supplementary Figure 2). We then investigated associations between the hydra clusters and neuroblastoma-associated molecular aberrations and clinical features (Supplementary Table 6). ATRX gene deletions were enriched in cluster 1 (Fisher's Exact Test: p ¡ 0.05). MKI low tumors were enriched in cluster 2 and 3 (Fisher's Exact Test: p ¡ 0.01). Chromosome 17 wild type tumors were enriched in clusters 2 and 3 (Fisher's Exact Test: p ¡ 0.01).

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

0.7 N-of-1 Tumor Analysis for Pediatric Neuroblastoma

We investigated the predictive performance of the hydra approach for identifying clinically relevant signals in the N-of-1 tumor analysis setting. We obtained tumor gene expression data from five stage 4, MYCN-NA neuroblastoma samples (Figure 4). The age at diagnosis ranged from 2 to 6 years. Four out of five samples had a deletion in the ATRX gene. Samples 1D and 1R are diagnosis and resection samples from the same patient (Figure 4). Three of the ATRX-deleted samples clustered with the high immune expression cluster (cluster 1) and one clustered in the low immune, high proliferative signaling cluster (cluster 2). Hydra analysis assigned sample 1D to cluster 1 and sample 1R to cluster 2. The resection sample 1R was extracted from lymph node tissue, which has a significantly different immune background than the training data. Another possible explanation for this change is that the tumor microenvironment is dynamic and the tumor may evade immune recognition as the disease progresses. We performed GSEA comparing samples 1D and 1R to investigate potential mechanisms leading to immune evasion in sample 1R. GSEA analysis found downregulation of the MHC Class I Antigen Processing & Presentation GO term in sample 1R (adjusted p-value ¡ 0.002). Loss of antigen processing functions is a common mechanism of immune evasion across cancer types [34].

We obtained H&E slide images for these samples; the images were reviewed by a licensed pathologist and scored for evidence of inflammation. The N-of-1 predictive

function of the hydra pipeline was used to classify samples into the subtypes discovered by the TARGET neuroblastoma analysis. Most of the samples (4 out of 5) clustered in cluster 1, which is characterized by higher immune marker expression. The hydra analysis agreed with the pathologist review in 4 out of 5 samples (Figure 4). Sample 4 was the only discordant sample. Sample 4 is also from lymph node tissue, which may have higher immune expression because of the tissue type. Notably, concordant ESTIMATE values were present in 3 out of 5 samples scored by the pathologist: samples 1D, 1R, and 2.

Fig 2. Hydra method correlates with histopathology review of tumor H&E slides. The tumor microenvironments of stage 4, MYCN-NA neuroblastoma patient samples were analyzed using gene expression and H&E slide image data. Inflammation levels in the same tumor samples were assessed from H&E slide images at 20x magnification (Moderate inflammation: 20-30% lymphocyte content; Minimal inflammation: ¡ 10% lymphocyte content). ATRX mutation status, hydra cluster assignment, and ESTIMATE ImmuneScore value are also indicated. Concordant and discordant predictions are marked with a positive (+) and negative (-) sign, respectively.

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

0.8 Hydra Analysis Discovers Complex Tissue Signatures

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

Discussion

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

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

Significant progress has been made in subtyping neuroblastomas and adapting therapy for aggressive subtypes, but unexplained heterogeneity remains. Not accounting for this heterogeneity decreases the power of standard methods to detect important expression patterns. Identifying biomarkers using genome-wide technology may lead to improved risk stratification and the discovery of novel drug targets [32]. Hydra analysis of the TARGET MYCN-NA neuroblastoma cohort (N=70) found differential expression of tumor microenvironment markers, including markers of the adaptive immune response. Pediatric cancers are generally thought to be non-immunogenic because they have lower mutation burden than adult cancers, but the immunogenicity of pediatric cancer has not been sufficiently investigated [35]. Our analysis found significant variation in immune marker expression and identified ATRX deletions as a potential biomarker of immune infiltrated tumors. Further investigation into gene expression signatures and molecular aberrations that predict response to immunotherapy in pediatric cancers is needed. Hydra analysis may facilitate the development of novel therapies by grouping patients with similar tumor microenvironment properties.

We found significant differences in immune and stromal expression that may inform precision medicine applications. The tumor microenvironment has become an important therapeutic consideration, but few methods account for the tumor microenvironment directly. Tumor purity has been identified as a confounding factor in cancer gene expression subtyping efforts [9]. For example, tumor purity and tumor microenvironment expression have been shown to correlate with pancreatic cancer subtypes [36]. Furthermore, Aran et al. (2018) found that tumor purity was correlated with the mesenchymal glioblastoma subtype and recommended a differential expression approach to computationally remove the tumor purity signal. However, standard approaches for subtracting the tumor purity effect may not be the best approach because several mechanisms may influence tumor purity, and each mechanism may

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

Conclusion

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

Supporting information

S1 Fig. Enriching for multimodally expressed genes improves clustering of established neuroblastoma subtypes. Standard TumorMap analysis (Newton et al. 2017) of the TARGET neuroblastoma dataset resulted in stage 4S samples clustering with stage 4 neuroblastoma samples (left). An alternative TumorMap based solely on 1,498 multimodally expressed genes separated the stage 4S samples into a distinct cluster (right).

S2 Fig. No statistically significant differences observed for age, mutation burden, or tumor purity features across MYCN-NA neuroblastoma hydra clusters. We hypothesized that older patients may have a better trained immune system and correlate with elevated immune infiltrate, but we did not find evidence to support this hypothesis. Mutation burden has been shown to correlate with immune infiltration, but we did not observe a correlation between elevated adaptive immune infiltrate and mutation burden in MYCN-NA neuroblastoma. Finally, we did not identify a significant difference in the pathology assessed tumor content, which is surprising because our gene expression analysis revealed a strong difference in estimated tumor purity. Investigation of these clinical features of MYCN-NA tumor tumor microenvironment revealed no statistically significant difference (Kruskal-Wallis test; $p > 0.05$).

S3 Fig. Gene set enrichment analysis (GSEA) of MYCN-NA neuroblastoma identifies survival differences within hydra cluster 2 and cluster 3. (A) Top 5 differentially enriched gene sets for each cluster comparing the entire MYCN-NA neuroblastoma cohort (cohort-level GSEA) and the corresponding hydra cluster (cluster-level GSEA). GSEA analysis of the hydra clusters found immune and cell cycle signals that were not identified in the cohort-level analysis. (B & C) GSEA separated cluster 2 into high and low immune subtypes and cluster 3 into high and low cell cycle subtypes. These signals correlated with differences in overall survival.

S4 Fig. Hydra analysis of TARGET osteosarcoma cohort reveals skeletal muscle signature. Hydra enrichment analysis on the TARGET osteosarcoma cohort revealed a subset of patients with high skeletal muscle expression. (A) Clustered heatmap shows the muscle signature genes identified by hydra unsupervised enrichment analysis. (B) xCell tumor microenvironment profiling identified significant differences in skeletal muscle expression compared to background (p ¡ 0.001). (C) H&E stained tumor slide for an independent osteosarcoma sample confirms presence of striated muscle tissue within the sequenced tumor sample.

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

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

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

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

Acknowledgments

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

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. Reeves E, James E. Antigen Processing and Immune Regulation in the Response to Tumours. Immunology. 2017;150(1):16–24. doi:10.1111/imm.12675.

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

36. 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.