Non-Parametric Mixture Modeling for Pediatric Precision Oncology

Jacob Pfeil[1,*], Geoff Lyle[1], Lauren Sanders[1], Katrina Learned[1], Ellen Kephart[1], Ann Durbin[1], Holly Beale[1], Olena Morozova[1], Sofie Salama[1], and David Haussler[1]

[1]University of California, Santa Cruz, Biomolecular Engineering, Santa Cruz, 95064, United States
[*]jpfeil@ucsc.edu

ABSTRACT

Precision oncology is changing the way medicine is practiced by incorporating high-throughput genomic analyses and data analytics. The field has focused on genomic variants, but gene expression data is becoming an additional tool for clinicians to improve patient outcomes. Here, we discuss a novel computational approach for identifying gene expression subtypes that may be helpful for future drug development and risk stratification. We apply this approach to synthetic data as well as patient data from the Cancer Genome Atlas Project.

Introduction

Cancer gene expression analysis has traditionally been used to compare the expression between two groups. The heterogeneity of tumor populations warrants an analysis that can identify differences across multiple subtypes with overlapping features. Differentiating expression at this level has been a challenge, so single-sample approaches have become more popular. However, the single-sample approach does not identify structure within a disease population that would improve the analysis of individual patients.

Many researchers have proposed a mixture modeling approach for differential expression analysis, but a mixture modeling approach has not been adopted by the gene expression analysis community (TODO: ADD REFERENCES TO GENE EXPRESSION MIXTURE MODEL PAPERS). One reason for this is the lack of tools to facilitate this type of analysis. These models are difficult to implement and pre-implemented tools are not well designed for cancer gene expression analysis. There has also been an insufficient number of tissue-specific gene expression profiles to differentiate molecular subtypes. However, the work of TCGA, TARGET, and public repositories not make more sophisticated gene expression analyses possible.

Most genes' expression distribution is approximately normally distributed once normalized using a log transformation[1].

Commonly used gene expression analyses are used to compare two populations or single samples, but methods are needed to find subtypes within a disease population. Supervised learning approaches like linear regression use known categorical information to learn differences between groups, but these methods are not able to detect subtle differences across a disease population without knowing the differences beforehand. Unsupervised methods are needed to subtype diseases, but one of the limitations in an unsupervised approach is that it is often necessary to specify the number of expected groups beforehand, which limits the ability to discover novel subtypes.

The goal of the hydra algorithm is to enrich for genes of interest by removing genes that are unimodally expressed across a disease cohort. Biomarker development requires gene expression distributions that are detectable in noisy gene expression data. Thus, potential biomarkers must be multimodally expressed and we can use this assumption to identify biomarkers for subtyping tumors and identifying novel drug targets. Multimodally expressed genes may be useful for subtyping tumors, but often there is a variable of interest that covaries with an expression distribution. Once the feature space has been reduced to a smaller size, then the analysis become powered to detected coordinated expression by identifying multivariate clusters. This allows the researchers to investigate the correlation of several genes associated with a variable of interest which will lead to robust detection of expression subtypes and robust gene expression signatures.

Hydra is built upon the Bayesian non-parametric python toolbox bnpy[2]. The benefit of a non-parametric approach is that the number of clusters does not need to specified because the algorithm is able to infer the number of clusters from the data. The bnpy library provides a flexible set of tools for performing Dirichlet process mixture modeling. Here, we use the Dirichlet process mixture modeling functions to identify multimodally expressed genes in the univariate setting and a bivariate setting if a covariate variable is provided. The bnpy implementation is fast and can scale to tens of thousands of genes.

The input for the hydra approach is a genes by samples matrix. The Dirichlet process mixture model has the property that as the number of samples increases, the number of clusters identified also increases. The first pass of the hydra pipeline is to

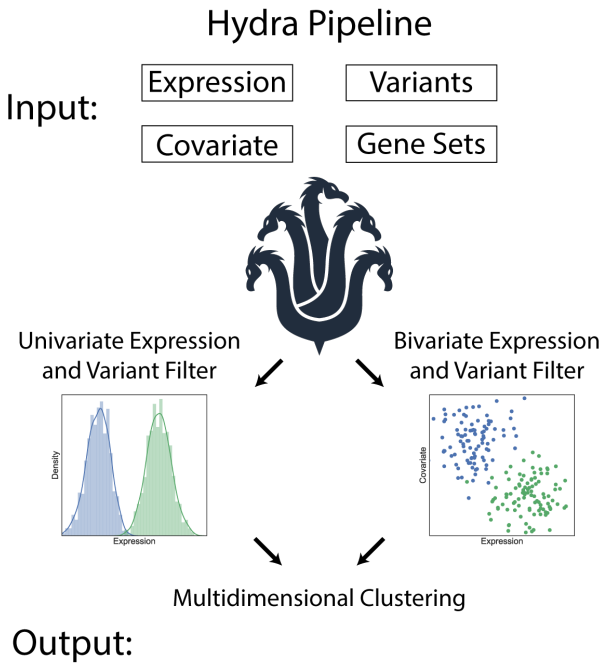


Figure 1. Hydra Method Overview

identify all of the genes that are multimodally expressed and thus a potential biomarker. If a covariate is provided, then the hydra method does a bivariate clustering to identify genes that are differentially expressed and statistically different in the covariate dimension using a Kruskal-Wallis test[3]. The goal of this analysis is to identify the genes that have a strong signal and may subtype tumors. The multimodally expressed genes are then assembled into a multivariate mixture modeling analyis to find coordinated expression of genes that subtype tumors. We have found that these expression networks show related biological function and can be used to identify pathway-level expression for subtyping tumors.

Results

Here we apply this method to synthetic data and compare this approach to other common analysis methods for subtyping tumors samples. One of the benefits of this approach is that this provides a framework for understanding a population of patients and can identify subtypes that may benefit from future drug development. Investigation into subtype specific expression may also identify opportunities to repurpose available therapies for other diseases. This is particularly useful for pediatric cancer gene expression analysis where drug development has lagged and there is a need for novel therapies.

Synthetic Data Analysis

Synthetic data was generated to assess the performance of the hydra mixture modeling approach. We have also included a few comparable gene expression analysis tools to validate our approach. Our first validation is to show the sensitivity and specificity of the hydra method compared to the common endpoint of bioinformatic analysis which is gene set enrichment analysis (GSEA). As mentioned in the introduction, gene set enrichment analysis is a challenge for precision oncology, especially in the pediatric field, since access to matched normal tissue is lacking. This limits the number of methods for inferring differentially expressed genes to ranking approaches such as single sample gene set enrichment analysis[4] and gene set variation analysis[5]. These methods are commonly used for N-of-1 analysis and have performed surprisingly well given limitations in their design.

We compared the predictive performance for identifying enriched genes using healthy tissue samples collected as part of the Genotype-Tissue Expression (GTEx) project[6]. The Molecular Signatures Database (MSigDB) provides a large set of curated gene lists for identifying biological functions for precision medicine applications. We used the Hallmark MSigDB gene sets to select genes with related biological functions and correlated expression[7]. We then randomly sampled GTEx skeletal muscle

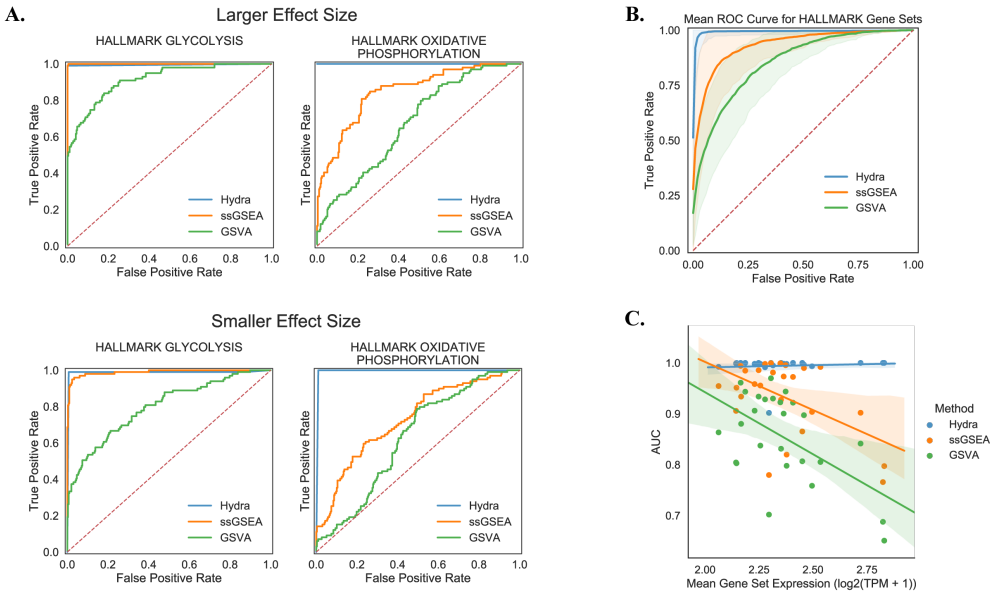


Figure 2. ROC Plot Curves for Assessing the Performance of Pathway Enrichment Tools.

samples, modified their expression values for a subset of the gene set genes. We did this process twice to generate synthetic training and test data. The same genes were used in the test and training data, but the values were sampled independently from a normal distribution at varying mean differences.

We applied this analysis to all of the Hallmark gene sets, but we are showing two illustrative examples here (TODO: add figure ref). The Hallmark Glycolysis gene set includes 199 genes involved in glycolysis and gluconeogenesis. We sample 25% of these genes to be differentially expressed and sampled a difference in expression value from a normal $\mathcal{N}(0.5, 0.5)$ or a $\mathcal{N}(1, 0.5)$. If this difference caused a negative expression value, then we set the expression value to be zero. The GTEx expression TPM values were generated using the Toil RNA-Seq pipeline[8]. The reciever operator curves (ROC) were generated for the hydra, ssGSEA, and GSVA methods. The hydra method performed the best with an AUC (area under the curve) of X, but the ssGSEA method was comparable with an AUC of blank. The GSVA method performed the worst with an AUC of BLANK.

One of the limitations of a non-parametric ranking approach to gene set enrichment is that it cannot account for the dynamic range in expression for a gene in a single-sample context. Therefore, we hypothesized that a GSEA method approach that does not account for this information will suffer if the background expression for a gene set is high. We see this in the HALLMARK Oxidative Phosphorylation analysis where the ssGSEA suffers from poor performance because the background expression is higher and thus is not well suited to identify subtle changes in expression.

TODO: Need to compare the small effect size and the large effect size

Cell Line Validation

Immune Subtype Validation

Whether pediatric cancer in general has a significant immune component is an active area or research. The success of checkpoint therapies in adult cancers has not been as strongly adopted in pediatric cancers because these therapies tend to do better in patients with a high mutation burden. Pediatric cancers generally have low mutation burdens. However, several groups have found elevated checkpoint marker expression in pediatric cancers[9–11], which suggests that immunosuppressive mechanisms are active in pediatric cancers and may need further characterization to identify therapeutic opportunities.

We show that the hydra approach is able to subtype patients using immune gene sets in a way that correlates with the computational staining of haemotoxylin and eosin (H&E) staining. We analyzed the TCGA skin cutaneous melanoma data set (TODO: N=X)

Neuroblastoma Transposable Element Analysis

One theory for way ATRX deleted samples have higher immune marker infiltrate is that dysregulation of ATRX function leads to increased expression of transposable elements which induces a DAMP response. Transposable elements have been shown to induce innate and adaptive immune responses, so the nature of the transposable elements may play a role in the type of immune

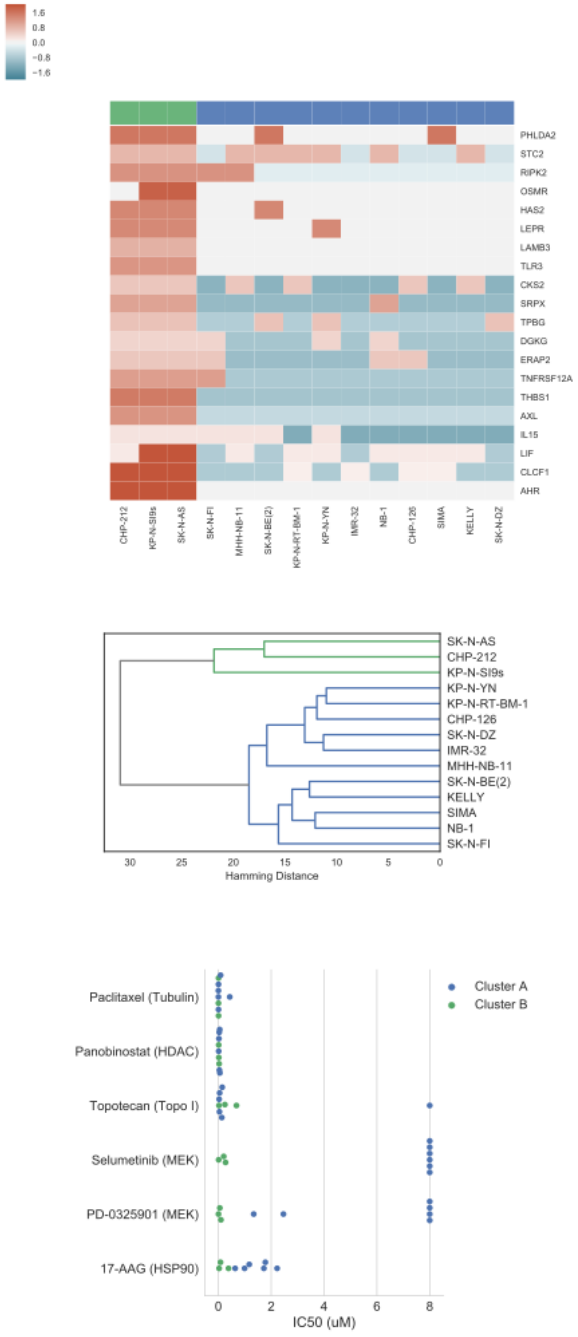


Figure 3. Cell Line Validation

response the occurs. The transposable element sequences may lead to the generation of antigens that can be used for treatment of neuroblastoma.

In order to investigate the transposable element expression in neuroblastoma, we applied two state-of-the-art methods. The first is the salmonTE algorithm and the second is the BLANK algorithm. We used two background cohorts to quantify transposable element expression. The first is tissue matched patients that were identified as having significantly lower expression of immune markers. The second is the UVM TCGA cohort which has been used in other studies as a negative control for immune active tumors[12,13].

Discussion

The Discussion should be succinct and must not contain subheadings.

Methods

Topical subheadings are allowed. Authors must ensure that their Methods section includes adequate experimental and characterization data necessary for others in the field to reproduce their work.

Synthetic Data Generation

The MSigDB Hallmark gene sets were used to simulate biological pathway expression that reflects the gene set properties of a typical gene expression analysis[7,14]. We also needed a large cohort of human tissue samples to infer differentially expressed genes. We used the GTEx skeletal muscle cohort[15]. To simulate subtype specific expression, we randomly sampled 25% of the patient population and then we sampled from a normal distribution centered at 0.5 with a standard deviation of 0.5. We then added this value to 25% of the genes in each Hallmark gene set to simulate coordinated expression of genes within a biological gene set.

References

1. Zwiener, I., Frisch, B. & Binder, H. Transforming rna-seq data to improve the performance of prognostic gene signatures. *PloS one* **9**, e85150 (2014).
2. Hughes, M. C. & Sudderth, E. B. Bnpy: Reliable and scalable variational inference for bayesian nonparametric models. In *NIPS Probabilistic Programimming Workshop* (2014).
3. McKight, P. E. & Najab, J. Kruskal-wallis test. *The corsini encyclopedia psychology* 1–1 (2010).
4. Barbie, D. A. *et al.* Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nat.* **462**, 108 (2009).
5. Hänzelmann, S., Castelo, R. & Guinney, J. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics* **14**, 7 (2013).
6. Lonsdale, J. *et al.* The genotype-tissue expression (gtex) project. *Nat. genetics* **45**, 580 (2013).
7. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell systems* **1**, 417–425 (2015).
8. Vivian, J. *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314–316 (2017).
9. Majzner, R. G. *et al.* Assessment of programmed death-ligand 1 expression and tumor-associated immune cells in pediatric cancer tissues. *Cancer* **123**, 3807–3815 (2017).
10. Saletta, F. *et al.* Programmed death-ligand 1 expression in a large cohort of pediatric patients with solid tumor and association with clinicopathologic features in neuroblastoma. *JCO Precis. Oncol.* **1**, 1–12 (2017).
11. Nowicki, T. S., Anderson, J. L. & Federman, N. Prospective immunotherapies in childhood sarcomas: Pd1/pdl1 blockade in combination with tumor vaccines. *Pediatr. research* **79**, 371 (2016).
12. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immun.* **39**, 782–795 (2013).
13. Thorsson, V. *et al.* The immune landscape of cancer. *Immun.* **48**, 812–830 (2018).
14. Liberzon, A. *et al.* Molecular signatures database (msigdb) 3.0. *Bioinforma.* **27**, 1739–1740 (2011).
15. Consortium, T. G. The genotype-tissue expression (gtex) project. *Nat. genetics* **45**, 580 (2013).
16. Yildirim, E. *et al.* Xist rna is a potent suppressor of hematologic cancer in mice. *Cell* **152**, 727–742 (2013).

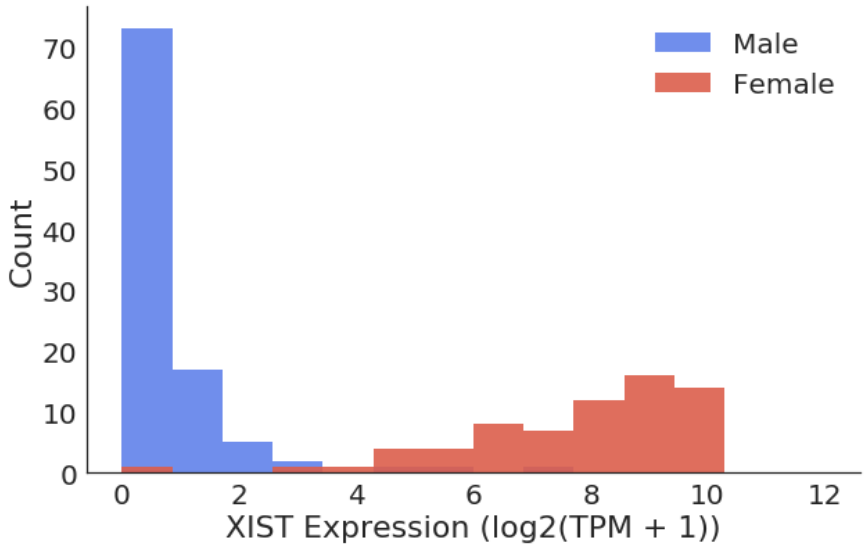


Figure S1. Example of gender-specific expression that can be modeled in a hierarchical model. The XIST gene is involved in X chromosome silencing, so XIST is not expressed for males. XIST has been linked to cancer, but the Treehouse model overestimates the variance in XIST expression because females and males are modeled together. The proposed hierarchical model learns the differences between male and female XIST expression for improved model fit.

Acknowledgements (not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

J.P. conceived the analysis, conducted the analysis, and analyzed the data. O.M., H.B., S.S., and D.H. reviewed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing financial interests** (mandatory statement).

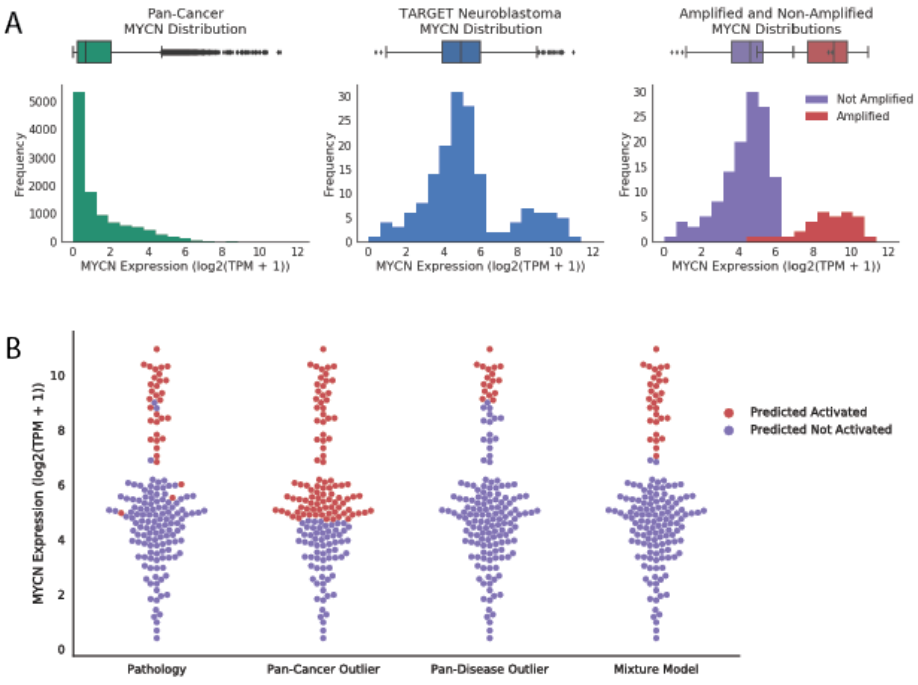
The corresponding author is responsible for submitting a competing financial interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Supplementary Information

Cancer Gene Expression Distributions are Multi-Modal

Modes Correspond to Sex-Specific Expression

Genes are expressed at different levels for different tissues. In addition to tissue specific expression, there are also biological features that influence gene expression across individuals. For example, age and gender are correlated with expression of some genes. A varying effects model where the mean and the effect of biological features change depending on the tissue can be used to make better predictions of gene expression. For example, a hierarchical model can identify sex-linked expression, but the current pan-cancer and pan-disease analyses are not able to detect sex-linked expression. An example of sex-linked expression that has been associated with cancer is the XIST gene[16]. XIST controls X-chromosome silencing in females and is not usually expressed in males (Figure **??**). This is a clear example where assuming male and female gene expression comes from the same distribution leads to an exaggerated estimation of the outlier threshold. It is therefore difficult to identify potential cases where under-expression of XIST in females may contribute to their cancer. While the incidence of cancer is equal across boys and girls, boys tend to respond worse to therapy. An investigation into sex-linked gene expression may yield insights into the differences in response to cancer therapies for boys and girls.



C

	Pan-Cancer	Pan-Disease	Mixture Model
Accuracy	62%	90%	96%
Precision	35%	100%	94%
Recall	100%	52%	88%
F1	52%	68%	91%

Figure S2. MYCN Validation