Integrative Precision Oncology using Non-Parametric Mixture Modeling

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

Cancer is the leading cause of death by disease for children in the United States. The standard of care therapies are toxic and some childhood cancer types do not respond well to these therapies. Molecularly targeted therapies inhibit specific proteins that are necessary for cancer progression and may be able to overcome the limitations of current standard of care therapies. Current big data analyses treat specific diseases as being homogeneous, but significant heterogeneity exists within specific cancer types. This has led to the development of molecular subtypes. To facilitate the development of therapies specifically designed for each molecular subtype, we have developed a mixture model that deconvolutes subtypes within large, public cancer gene expression profiles. Often, it is difficult to acquire a normal sample for pediatric gene expression analysis or the cell of origin for some pediatric cancers is unknown. Our approach does not require a normal sample because it identifies which gene expression programs are differentially activated or repressed across cancer subtypes. Once the mixture model is learned, it is straightforward to classify new patients into molecular subtypes. We hope this work will accelerate personalized medicine approaches for pediatric and adult cancers.

Introduction

Childhood cancer patients need therapies that cure disease while also safeguarding development and future health. Approximately, 16,000 children are diagnosed with cancer each year in the United States. Despite significant improvements in childhood cancer therapies, one in eight children will die of cancer. Some forms of childhood cancer respond better to standard of care therapies than others. There are forms of pediatric brain tumors that have survival rates around ~10 %.

The standard of care therapies are also harmful to the long-term health of childhood cancer survivors. For instance, children respond well to high-dose chemotherapy, but chemotherapeutic agents are toxic and damage healthy tissue. Life-long side effects develop in ~60 % of the childhood cancer survivors. Childhood cancer survivors are more likely to develop other forms of cancer, heart and lung problems, stunted growth, and learning disabilities[1–3]. There are ~380,000 childhood cancer survivors in the United States and 60% of them are facing life-long disabilities as a result of their cancer therapy.

A more personalized approach may overcome the shortcomings of current standard of care therapies. Molecularly targeted therapies identify rare alterations within a patient's cancer that can be specifically inhibited to prevent cancer progression. Targeted therapies are biologically active at a lower dose than many standard of care therapies which makes them less toxic. While targeted therapies have induced tumor remissions, cancer cells are prone to become resistant to targeted therapies and the cancer returns. Research into the molecular mechanisms of drug resistance as well as development of more pediatric targeted inhibitors may yield novel therapeutic directions that yield better outcomes for patients with fewer harmful side effects[4].

A more targeted approach identifies specific molecular alterations that make cancer cells susceptible to targeted therapies. An example of a successful targeted therapy is imatinib (Gleevec) for BCR-ABL driven leukemia. BCR-ABL is a fusion protein that couples the oncogenic ABL1 gene with a constitutively expressed BCR gene. This increases the concentration of the oncogenic ABL1 gene to drive cancer progression. Imatinib can correct for this alteration by binding to the ABL active site and preventing ABL's biological function. BCR-ABL positive cancer cells depend on the ABL protein to proliferate, so inhibition of ABL's function halts cancer progression. The BCR-ABL fusion occurs in a fraction of leukemia patients, but application of imatinib to BCR-ABL positive leukemias has been proven to improve treatment outcomes[5,6].

Raw RNA sequencing data is in FASTQ format. FASTQ format is a simple text format that lists each sequence with the sequencer's confidence score for calling each base in the sequence. After preprocessing and quality control, the next step in gene expression analysis is to map the sequencing data to a reference genome or transcriptome using sequence similarity. The human genome is well-annotated, and the annotation is used to assign sequencing data to specific genes. There are many algorithms for mapping sequencing data to reference genomes, but one of the most widely adopted algorithms is called STAR[7].

After alignment, gene quantification algorithms count the number of reads that mapped to each gene or transcript. To improve transcript-level quantifications, some algorithms like the RSEM algorithm try to maximize the likelihood of observing the data and estimate an expected count for each gene[8].

Absolute gene expression is difficult to analyze, so a common analysis method is to compare absolute gene expression of two groups of data and identify differences in expression. Differential expression analysis for cancer studies typically estimate gene expression in two groups of samples, typically a healthy control and disease group, and identifies differences in gene expression. Differential expression analysis can be used to find cancer genes by comparing tumor expression to matched healthy tissue expression. When a tumor is biopsied or resected, the surgeon often takes a sample of healthy tissue for comparison. For many cancer types, it is not feasible to take a matched normal sample. In our experience, pediatric gene expression data rarely has matched normal data, so other methods are needed to identify differentially expressed genes.

Differential expression analysis also requires defining two conditions. For cancer, the two conditions are usually cohort of paired healthy tissue, or normal samples, and the second condition is a cohort of disease samples. In addition to having limited cancer tissue, in our experience, it is more difficult to obtain paired normal pediatric tissue. Therefore, there is not a control group to compare pediatric cancer expression to. This is one reason to assemble the Treehouse compendium of adult and pediatric cancer because we can use other pediatric cancer samples to identify patterns in expression for pediatric tissue.

Alternative methods to differential expression analysis include GFOLD and Cancer Outlier Profile Analysis (COPA). GFOLD is the state-of-the-art method for ranking genes based on fold-change. GFOLD prioritizes genes that have high fold change relative to controls and a large number of read counts. GFOLD performs better than differential expression algorithms when working with a single biological replicate[9]. The Treehouse algorithm is similar to GFOLD in that a gene expression outlier needs to be expressed at a much higher level than the median and be in the top 5% of all expressed genes.

Many differential expression tools are based on a t-test for comparing two means. One challenge with this approach is that some samples in a cohort may have differential gene expression that is not consistent with the overall population. For a particular disease, patient A may have MYC over-expression and normal levels of CDK4, but patient B may have CDK4 over-expression and normal levels of MYC. The COPA method was designed to find subtle patterns of differential expression compared to a normal cohort. The COPA method assumes that the healthy cohort will not have pathogenic expression, but samples within the experimental disease cohort will show mutually exclusive expression for pairs of genes[10,11]. This approach fails for Treehouse analysis because our control cohort includes cancer samples that will likely have over-expression of oncogenic genes.

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 from scratch and pre-implemented tools are not well designed for cancer gene expression analyses. 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[12].

Results

Overview of Method

The goal of the hydra algorithm is to enrich for genes of interest by removing genes that are unimodally expressed across a disease cohort and thus not differentially expressed. The assumption is that a unimodally distributed gene has limited information because the signal is diluted by the variability in sequencing technology. However, multi-modally expressed genes have a signal that clearly breaks through the noise of sequencing technology and thus should be prioritized in a genome-wide gene expression analysis. This method is unbiased and can be used to identify novel biomarkers for disease. We can additionally filter genes that covary with a variable of interest such as response to a drug or survival in order to identify subtype specific expression that is related to the research question. 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.

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.

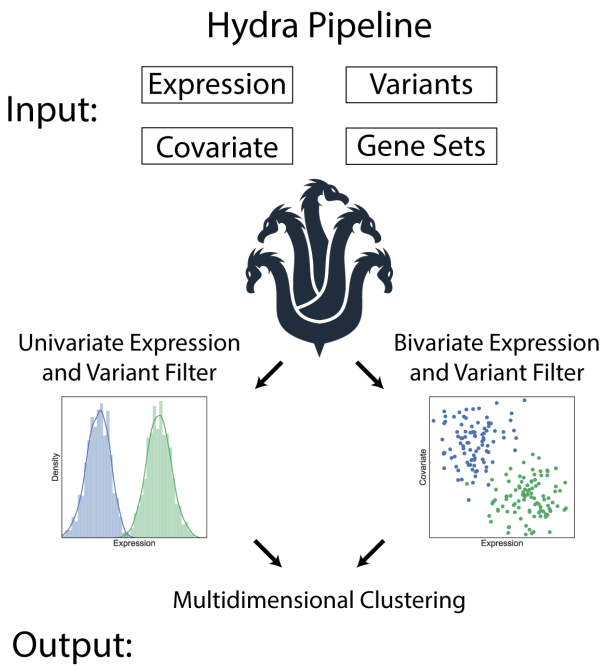


Figure 1. Hydra Method Overview

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[13] and gene set variation analysis[14]. 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[15]. 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[16]. We then randomly sampled GTEx skeletal muscle 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[17]. 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

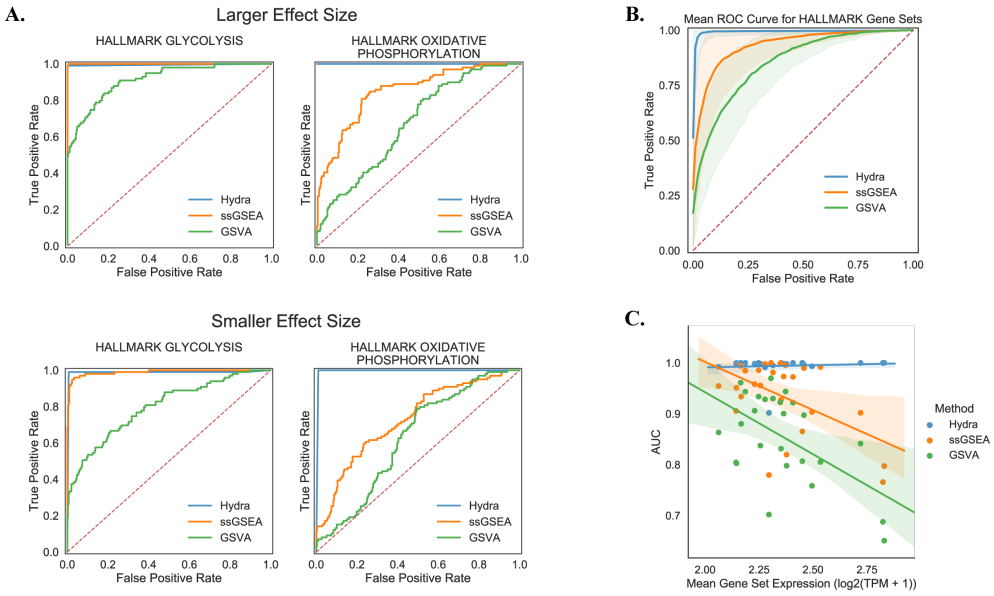


Figure 2. ROC Plot Curves for Assessing the Performance of Pathway Enrichment Tools.

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[18–20], 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 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[21,22].

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.

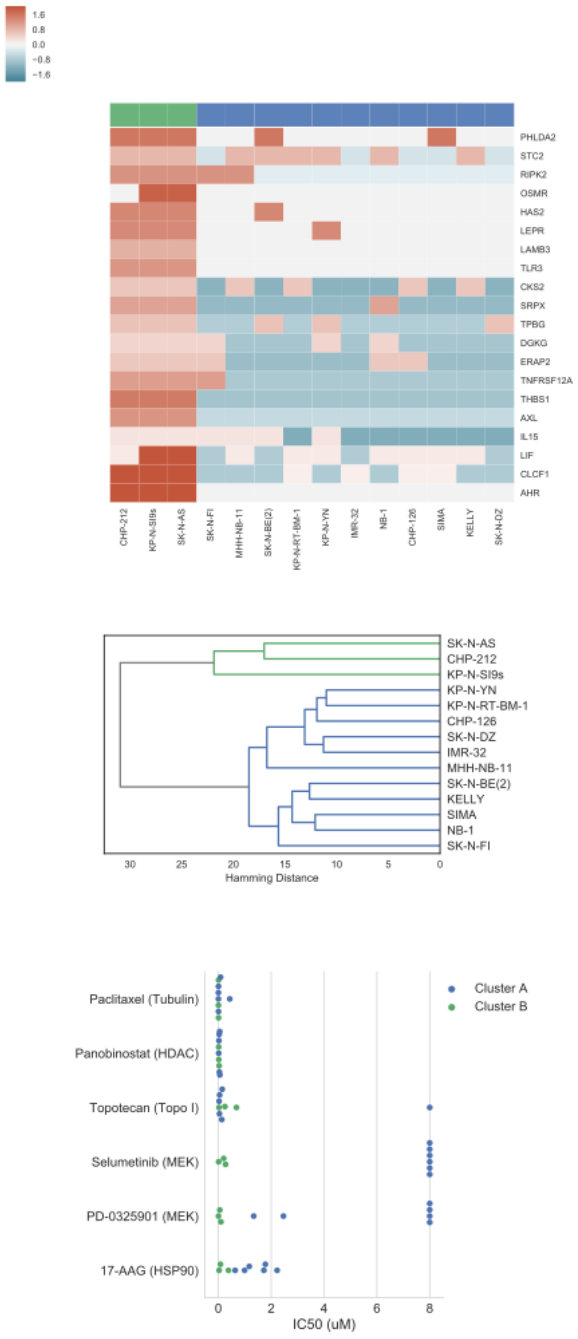


Figure 3. Cell Line Validation

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[16,23]. We also needed a large cohort of human tissue samples to infer differentially expressed genes. We used the GTEx skeletal muscle cohort[24]. 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. Berger, E. Late and long-term effects of treatment of childhood leukemia (2016). URL https://www.cancer.org/cancer/leukemia-in-children/after-treatment/long-term-effects.html. [Online; accessed 28-April-2017].
2. Kopp, L. M., Gupta, P., Pelayo-Katsanis, L., Wittman, B. & Katsanis, E. Late effects in adult survivors of pediatric cancer: a guide for the primary care physician. *The Am. journal medicine* **125**, 636–641 (2012).
3. The American Cancer Society. Cancers that develop in children. http://www.cancer.org/cancer/cancer-in-children/types-of-childhood-cancers.html (2016). Accessed: 2016-01-19.
4. Norris, R. E. & Adamson, P. C. Challenges and opportunities in childhood cancer drug development. *Nat. Rev. Cancer* **12**, 776 (2012).
5. Wong, S. & Witte, O. N. The bcr-abl story: bench to bedside and back. *Annu. Rev. Immunol.* **22**, 247–306 (2004).
6. Bernt, K. M. & Hunger, S. P. Current concepts in pediatric philadelphia chromosome-positive acute lymphoblastic leukemia. *Front. oncology* **4**, 54 (2014).
7. Dobin, A. *et al.* Star: ultrafast universal rna-seq aligner. *Bioinforma.* **29**, 15–21 (2013).
8. Li, B. & Dewey, C. N. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).
9. Feng, J. *et al.* Gfold: a generalized fold change for ranking differentially expressed genes from rna-seq data. *Bioinforma.* **28**, 2782–2788 (2012).
10. Macdonald, J. W. & Ghosh, D. Copa—cancer outlier profile analysis. *Bioinforma.* **22**, 2950–2951 (2006).
11. Wang, C. *et al.* mcopa: analysis of heterogeneous features in cancer expression data. *J. clinical bioinformatics* **2**, 22 (2012).
12. Zwiener, I., Frisch, B. & Binder, H. Transforming rna-seq data to improve the performance of prognostic gene signatures. *PloS one* **9**, e85150 (2014).
13. Barbie, D. A. *et al.* Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nat.* **462**, 108 (2009).
14. Hänzelmann, S., Castelo, R. & Guinney, J. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics* **14**, 7 (2013).
15. Lonsdale, J. *et al.* The genotype-tissue expression (gtex) project. *Nat. genetics* **45**, 580 (2013).
16. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell systems* **1**, 417–425 (2015).
17. Vivian, J. *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314–316 (2017).
18. 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).
19. 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).
20. 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).
21. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immun.* **39**, 782–795 (2013).
22. Thorsson, V. *et al.* The immune landscape of cancer. *Immun.* **48**, 812–830 (2018).
23. Liberzon, A. *et al.* Molecular signatures database (msigdb) 3.0. *Bioinforma.* **27**, 1739–1740 (2011).

24. Consortium, T. G. The genotype-tissue expression (gtex) project. *Nat. genetics* **45**, 580 (2013).

25. Yildirim, E. *et al.* Xist rna is a potent suppressor of hematologic cancer in mice. *Cell* **152**, 727–742 (2013).

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[25]. 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.

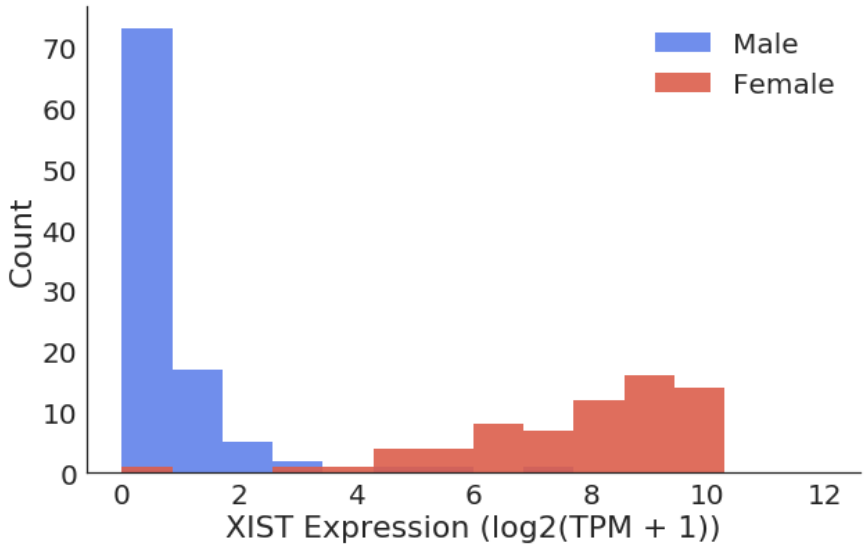
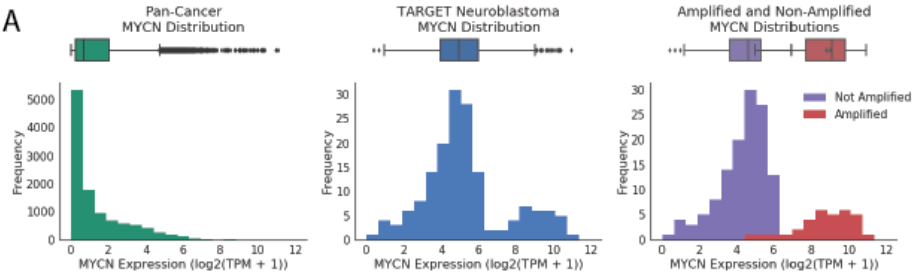
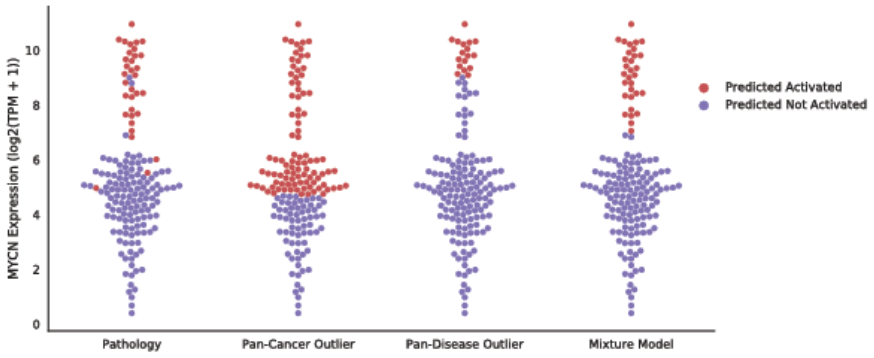


Figure S1. Example of gender-specific expression that can be modeled in a hierarchical model. The XIST gene is involved in X chromosome silencing, so XIST is not expressed for males. XIST has been linked to cancer, but the Treehouse model overestimates the variance in XIST expression because females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

A



B



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