Title

Bayesian Framework for Detecting Gene Expression Outliers in Individual Samples.



Journal

JCO Clinical Cancer Informatics, 4(4)

ISSN

2473-4276

Authors

Vivian, John

Eizenga, Jordan M

Beale, Holly C

et al.

Publication Date

2020-11-01

DOI

10.1200/cci.19.00095

Peer reviewed

original reports

# Bayesian Framework for Detecting Gene Expression Outliers in Individual Samples

John Vivian, PhD[1]; Jordan M. Eizenga, MS[1]; Holly C. Beale, PhD[2]; Olena M. Vaske, PhD[2]; and Benedict Paten, PhD[1]

abstract

**PURPOSE** Many antineoplastics are designed to target upregulated genes, but quantifying upregulation in a single patient sample requires an appropriate set of samples for comparison. In cancer, the most natural comparison set is unaffected samples from the matching tissue, but there are often too few available unaffected samples to overcome high intersample variance. Moreover, some cancer samples have misidentified tissues of origin or even composite-tissue phenotypes. Even if an appropriate comparison set can be identified, most differential expression tools are not designed to accommodate comparisons to a single patient sample.

**METHODS** We propose a Bayesian statistical framework for gene expression outlier detection in single samples. Our method uses all available data to produce a consensus background distribution for each gene of interest without requiring the researcher to manually select a comparison set. The consensus distribution can then be used to quantify over- and underexpression.

**RESULTS** We demonstrate this method on both simulated and real gene expression data. We show that it can robustly quantify overexpression, even when the set of comparison samples lacks ideally matched tissue samples. Furthermore, our results show that the method can identify appropriate comparison sets from samples of mixed lineage and rediscover numerous known gene-cancer expression patterns.

**CONCLUSION** This exploratory method is suitable for identifying expression outliers from comparative RNA sequencing (RNA-seq) analysis for individual samples, and Treehouse, a pediatric precision medicine group that leverages RNA-seq to identify potential therapeutic leads for patients, plans to explore this method for processing its pediatric cohort.

**JCO Clin Cancer Inform 4:160-170. © 2020 by American Society of Clinical Oncology**

## INTRODUCTION

RNA sequencing (RNA-seq) has been used in the cancer field for a number of purposes: To examine differences between tumor and normal tissue; to classify cancers for diagnostics; and—with the advent of single-cell RNA-seq—to characterize tumor heterogeneity.[1-6] Precision medicine researchers have also begun exploring RNA-seq's potential to aid in target selection and drug repositioning by identifying clinically actionable aberrations in tumor samples.[7-10] Clinical studies have demonstrated actionable findings for up to 50% of patients through RNA-seq analysis, particularly for pediatric patients who often do not possess actionable coding DNA mutations.[11-15] This has led to efforts like Treehouse, a precision medicine initiative for pediatric cancer, that evaluates the utility of RNA-seq analysis to inform clinical interpretation. Treehouse has created a large compendium of open access cancer gene expression data, which is incorporated into its analysis.[16-18]

Protocols for such precision medicine initiatives involve the identification of upregulated druggable gene targets as therapeutic leads. Differential expression is commonly used to identify up- and downregulation of genes between two groups of samples. However, most differential expression tools operate best under experimental conditions where both groups consist of several technical replicates or if lacking that, biologic replicates.[19-22] Thus, most existing tools are poorly suited to the clinical setting, where one group consists of only a single biologic replicate from one patient (N of 1), and the other comparison group is a library of diverse potential comparison samples. In particular, none of the existing methods have any way of suggesting what an appropriate subset of the sample library should be used for comparison. This limitation is especially acute in cancer, where uncertainty about the cell of origin, histologic complexity, and metastasis can make it difficult to identify the appropriate reference tissues for a sample.[23] While some work exists to address statistical uncertainty of working

ASSOCIATED CONTENT

**Data Supplement**

Author affiliations and support information (if applicable) appear at the end of this article.

**Accepted on January 14, 2020 and published at ascopubs.org/journal/cci on February 25, 2020: DOI https://doi.org/10.1200/CCI.19.00095**

## CONTEXT

**Key Objective**

How can we identify targetable genes in individual patients with cancer?

**Knowledge Generated**

We discuss a novel Bayesian method that compares an individual RNA sequencing (RNA-seq) cancer sample to a large background of normal data. The model dynamically selects a background data set on the basis of similarity to the individual sample, which is then used to identify expression outliers among a set of genes of interest using posterior predictive *P* values.

**Relevance**

This method can be applied to tumor RNA-seq samples of individual patients with cancer to generate a ranked list of potential therapeutic targets.

with N-of-1 samples,[24] we focus on solving the second problem, which is the principled selection of an appropriate comparison set.

Existing N-of-1 protocols compare targeted genes in an N-of-1 sample to an outlier cutoff generated from a large compendium of either cancer samples or unaffected tissue to determine whether a gene is upregulated.[16,25-27] While this outlier cutoff method is fast, there are some notable drawbacks. Application of a cutoff binarizes data, which makes it difficult to meaningfully rank outliers or to be aware of samples just short of meeting the cutoff. This cutoff method is also intended for Gaussian distributions, which are empirically common for gene expression within a tissue group but not typical when considering the distribution of expression across tissues.[25]

Ultimately, the most difficult problem is justifying the choice of what samples constitute the comparison set that generates the cutoff because different comparison sets will identify different genes as outliers. Many comparison data sets are small (almost half of The Cancer Genome Atlas's [TCGA's] normal tissues have ≤ 10 samples), so they lack the statistical power to characterize the variability of the expression landscape in the normal tissue on their own.[28] This power can be increased by also including samples from different tissues, but including tissues with larger sample sizes can drown out the information from the matched tissue. In addition, it is unclear which other tissues should be included in the pooled comparison set.

These concerns led us to propose a new approach for identifying outliers for N-of-1 samples. In contrast to previous methods, our method adaptively constructs a meaningful comparison set and avoids selection bias by automatically weighting the background sets to generate a consensus distribution of expression. It then uses the consensus distribution to quantify overexpression for genes of interest.

## METHODS

The core of our method is a Bayesian statistical model for the N-of-1 sample's gene expression. The model implicitly assumes that the sample's gene expression can be approximated by a convex mixture of the gene expression of the background data sets. The coefficients of this mixture are shared across genes, much like a linear model in which each data point is the vector of expressions for a gene across the background data sets. In addition, we model expression for each gene from each background data set as a random variable itself. This allows us to incorporate statistical uncertainty from certain background sets' small sample size directly in the model without drowning out any background set's contribution through pooling (Fig 1).

### Model Specification

Suppose we have $n$ background data sets for expression, which we will call $X_1,\ldots,X_n$. Within each background set

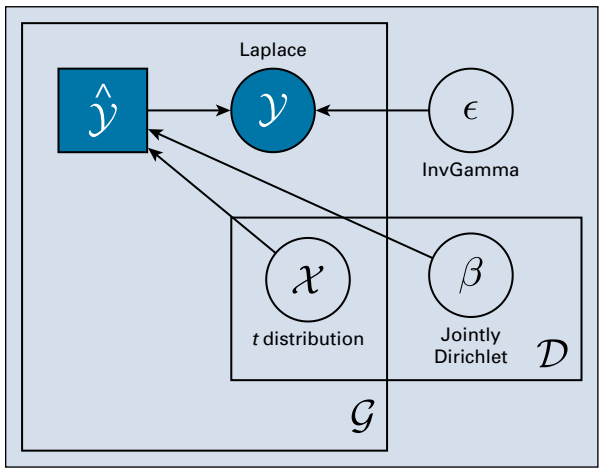


**FIG 1.** Bayesian plate notation of the model, where G denotes gene and D denotes data set. $x$ represents gene expression for one background data set and is multiplied by $\beta^T$ to produce the convex combination $\hat{y}$. We specify jointly Dirichlet because there is not one Dirichlet distribution per background data set. Instead, each background data set is one component of the Dirichlet $\alpha$ vector. We add Laplacian error $\varepsilon$ to the expected expression $\hat{y}$ when modeling the observed expression of a new sample $y$. The scale of the Laplacian error follows an inverse gamma (InvGamma) distribution.

$d = 1, 2, \ldots, n$, we model the expression of gene $g$ as a normal-distributed random variable:

$$x_{d,g} \sim \mathcal{N}(\mu_{d,g}, \sigma_{d,g}).$$

The parameters of each of these normal distributions are distributed according to a shallow but proper normal-inverse gamma prior:

$$\mu_{d,g}, \sigma^2_{d,g} \sim \mathcal{N}\text{-}\mathcal{IG}(0, 1, 1/2, 1).$$

Next, we assume that there is another unobserved random variable $\hat{x}_{d,g}$ from the same distribution $\mathcal{N}(\mu_{d,g}, \sigma_{d,g})$. Conceptually, this corresponds to the expression value that the background distribution would influence the N-of-1 sample to take for that gene. We model the expected expression $\hat{y}_g$ from a new sample as a convex combination of the unobserved expression values across the data sets:

$$\hat{y}_g = \beta_1 \hat{x}_{1,g} + \cdots + \beta_n \hat{x}_{n,g}$$

$$\beta \sim \text{Dirichlet}(1, \cdots, 1).$$

Note that $\beta$ is shared across all genes. A Dirichlet distribution was chosen for $\beta$ to enforce the convexity constraint. Finally, we model the observed expression of the new sample $y_g$, which adds Laplacian error to the expected expression $\hat{y}_g$:

$$y_g = \hat{y}_g + \varepsilon_g$$

$$\varepsilon_g \sim \text{Laplace}(0, \tau)$$

$$\tau \sim \mathcal{IG}(1, 1).$$

The distribution of the model error $\varepsilon$ is shared between genes and incorporates uncertainty into the posterior to account for variance generated by a poor match of the N of 1 to any particular background group, weak model fitting, and biologic or technical noise. We use a Laplace distribution instead of the more conventional normal distribution because we are interested in identifying expression outliers. The Laplace distribution is heavier tailed, so it will fit to outliers less aggressively and thereby preserve their outlier status.

This model fits the data well for most cases. However, it behaves poorly on genes that have large variances in the background data set. The reason is that the $\varepsilon$ parameter shares a scale parameter ($\tau$) across all genes. This causes differences in expression that are modest relative to the large background variance to appear to be highly significant. Thus, the model fits aggressively to the N-of-1 expression on these genes rather than to preserving the background distribution (Data Supplement). To address this limitation, we normalize the background data sets for variance but not for location. This normalization step must be incorporated into the model specification because it is not known a priori which background data sets the model will learn to be important, and different background data sets have different variances. This leads to the following equation:

$$\sum_d \beta_d \left( \frac{y_g - \mu_d}{\sigma_d} + \mu_d \right) = \sum_d \beta_d \left( \frac{\hat{x}_d - \mu_d}{\sigma_d} + \mu_d \right) + \varepsilon_g,$$

which simplifies to

$$y_g = \frac{\sum_d \frac{\beta_d}{\sigma_d} \hat{x}_d + \varepsilon_g}{\sum_d \frac{\beta_d}{\sigma_d}}.$$

The model can be explored using Markov chain Monte Carlo (MCMC) to obtain samples for $y_g$ that approximate its posterior distribution. If we have an observed expression value for a gene of interest (from the N-of-1 cancer sample), we can compare it to the sampled values. The proportion of sampled values for $y_g$ that are greater (or lesser) than the observed value is an estimate of the posterior predictive $P$ value for this expression value. The posterior predictive $P$ value can be seen as a measure of how much of an outlier the expression is given the expectations of the comparison set.

The model is implemented in PyMC3, and each N-of-1 sample is trained using the No-U-turn MCMC sampler.[29,30] Because of the computational burden of sampling from the model, we use a couple of computational tricks to reduce runtime to a tractable level. First, we integrate out the $\mu_{d,g}$ and $\sigma_{d,g}$ parameters so that each $\hat{x}_{d,g}$ is distributed according to a posterior predictive $t$ test. Given our choice of a Dirichlet distribution for $\beta$, most of the background data sets are assigned 0 weight, which means that it is inefficient to include all $n$ background data sets for every training run. Instead, background data sets are heuristically ranked for similarity to the N-of-1 sample by a combination of analysis of variance and pairwise distance and then iteratively added until the posterior predictive $P$ values converge to Pearson correlation $> 0.99$ (Data Supplement). The model is available as a Python package for convenience, a Docker container for reproducibility, and a Toil workflow for scalability. The software also provides comprehensive output to aid users in interpreting model results (Data Supplement).

## RESULTS

### TCGA and Genotype-Tissue Expression Consortium Validation

We evaluated whether the model would identify comparison sets on the basis of tissue type. To do so, we ran the model on 977 TCGA tumor samples spanning 10 different tissues. We selected these 10 tissues on the basis of two criteria: They had corresponding tissues in the Genotype-Tissue Expression Consortium (GTEx) data set, and they had sufficiently many tumor samples to get a good estimate of the average behavior of our model. We used the entire set of normal tissue samples from GTEx and TCGA

 

as two different background data sets to compare with these randomly selected TCGA samples[31] (Fig 2; Data Supplement).

For every group of samples within a tissue type, the matched tissue in GTEx or TCGA-normal was afforded a majority of the model weight, with only two groups of samples receiving < 60% of all total weight: bladder and stomach. Dimensionality reduction reveals that bladder and stomach samples tend to cluster near other tissue groups that the model assigns weight to (Data Supplement). This behavior is especially impressive considering that the comparison is between GTEx and TCGA: two independent projects. It is likely that there are some hidden batch effects between these data sets (which would be challenging to correct for), but true biologic signal tends to overwhelm any batch effects that exist.

### Negative Control

As a negative control experiment, we ran 100 samples across 10 GTEx tissues using three different backgrounds: TCGA-tumor, TCGA-normal, and GTEx. Our expectation was that there would be relatively few outliers when either normal (noncancer) data set is used as the background comparison set relative to TCGA-tumor. Figure 3A shows that when either GTEx or TCGA-normal were used as the background data set, the gene *P* values shrink toward the middle, and outliers are rarely identified. The model tends to assign almost all weight to the N of 1's matched tissue in GTEx or TCGA-normal, and the N of 1 does not deviate significantly from other samples in that tissue group, with few exceptions.

### Testing Model Robustness by Removing Matched Tissues

In most cases, our method is robust to situations in which there is no obvious matched normal tissue (Fig 3B). To demonstrate this, we used our method with a comparison data set in which we artificially removed the tissue matched to the sample and then compared the results with the restricted data set to the results we obtain with the full data set. The model will often go from assigning almost all of the weight to the matched normal tissue to distributing it among several other phenotypically similar tissues. However, in

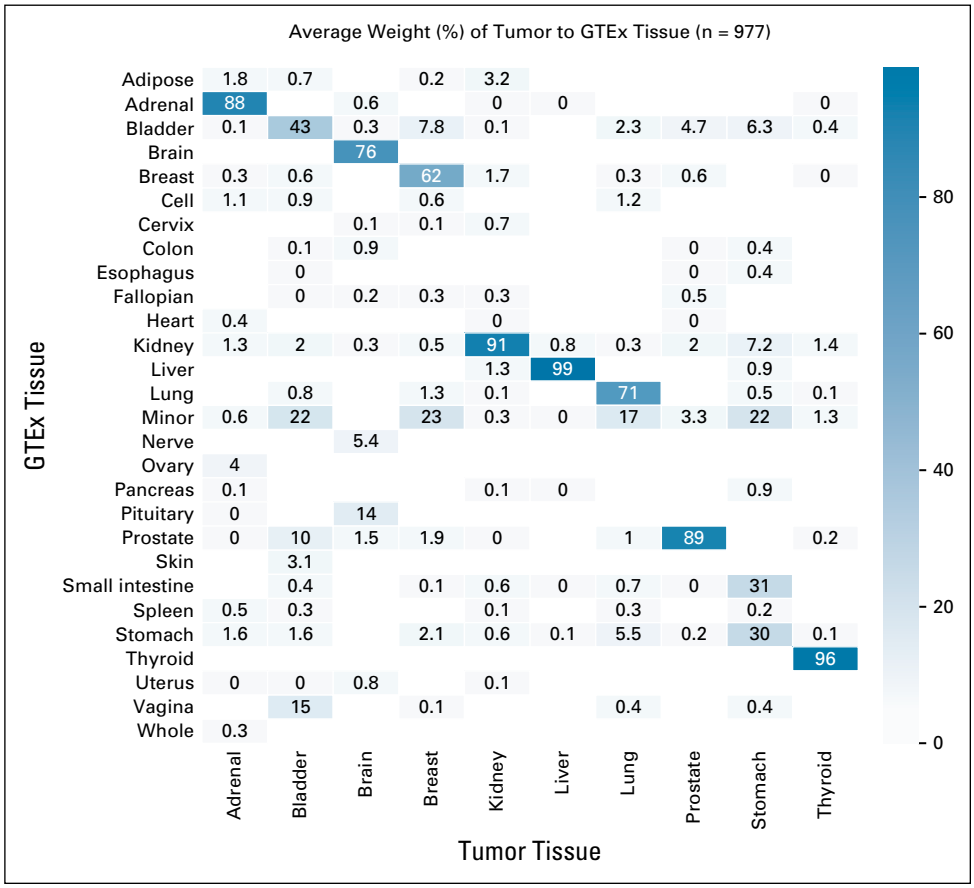


**FIG 2.** Heat map of the average model weight assigned to Genotype-Tissue Expression Consortium (GTEx) tissues across tumor subtypes in The Cancer Genome Atlas (TCGA). The model assigns a majority of weight to the matched tissue in GTEx for every tumor subtype. Only two tumor tissues—bladder and stomach—received < 60% of the average model weight. GTEx has only nine bladder samples, and principal component analysis shows that those bladder samples cluster on top of minor salivary gland and vaginal samples, which helps to explain its lower average weight relative to other tumor types (Data Supplement).

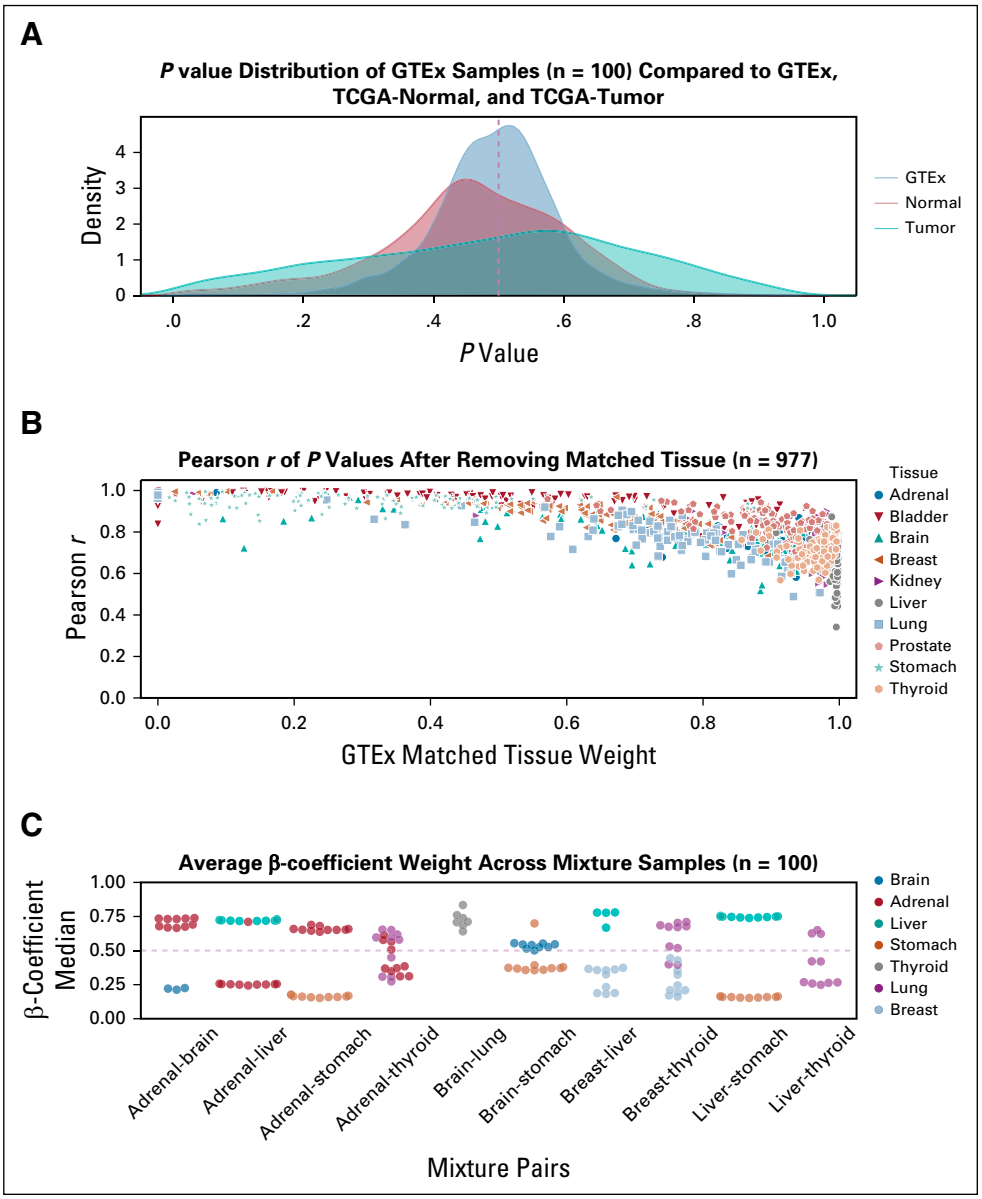


**FIG 3.** Different aspects of model robustness. (A) Robustness to false-positives. Negative control experiment where 100 Genotype-Tissue Expression Consortium (GTEx) samples were run using GTEx, The Cancer Genome Atlas (TCGA)-normal, and TCGA-tumor as different backgrounds. Because posterior predictive *P* values measure how different the observed result is from model expectations, we assume that using normal tissues as backgrounds for GTEx N-of-1 samples will result in a peak at approximately .5 and very few outliers compared with when TCGA-tumor is used as a background. (B) Robustness to imperfect comparison sets. The effect that removing a matched tissue has on the gene *P* values the model generates as measured by Pearson correlation. The *x*-axis is the weight assigned to the matched tissue by the model. Gene *P* values are relatively consistent, even when a matched tissue is removed, particularly if the model can redistribute that weight to tissues of similar phenotypes. (C) Robustness to mixed lineage samples. Average model weight of mixture samples generated from random pairings of GTEx tissues. Samples were generated by averaging gene expression between randomly sampled subsets of each tissue group. In most cases, the two tissues used to generate the mixture are assigned the majority of the model weight, which is the expected result. Three sets of mixture samples—adrenal-brain, brain-lung, and liver-thyroid—do not get the same result. Principal component analysis of those mixture samples, in the context of similar tissues, shows that the generated mixture samples happen to cluster closer to other tissues than one or both of the tissues used to generate the samples (Data Supplement). In these circumstances, we expect the model to assign weight to those tissues that are more similar to the mixture samples.

 

most tissues, the model largely compensates for the missing data in the final results: The *P* values remain highly correlated to those produced with the full data set (Data Supplement). That said, the *P* values do move slightly away from the tails, indicating lower power to detect outliers.

### Mixture Simulation

We used a simulation to validate the method's ability to identify comparison sets in tumors of nonspecific lineage. Simulated N-of-1 samples were created by randomly selecting tissue pairs from GTEx then averaging gene expression between random samples from those tissue pairs. Principal component analysis of the mixture samples shows a tight cluster in between the two clusters for the contributing tissues (Data Supplement). Mixture samples were run through the model, and the weights from the two contributing tissues were collected (Fig 3C). Ideally, 50% of the model weight should be assigned to each of the contributing tissues used to generate those mixture samples, which is true for a majority of the tissue pairs. We would not want the model to split weight evenly between the two contributing tissues if the generated mixtures happen to be more similar to other tissue types in the background data set. For mixture samples that did not match to the tissues used to generate them, dimensionality reduction clearly shows that other tissues happen to cluster closer to the mixtures than one or both of the contributing tissues (Data Supplement).

### Upregulated Gene Outlier Counts Across Tumor Subtypes

Gene amplification and overexpression are common hallmarks of cancer cells, resulting from extra copies of a locus (amplicon) as well as from other genetic and epigenetic changes. In many cases, these changes occur in genes that are specific to their tissue of origin.[32-34] Many of these commonly mutated genes can be targeted by existing drugs.[35] Eighty-five such druggable genes, mostly receptor tyrosine kinases, were curated and provided to us by Treehouse. We calculated *P* values for these genes using our method across the 977 TCGA samples (Fig 4). Genes with *P* values below a cutoff ($< 0.05$) that also appeared in more than a third of the tumor samples within a subtype were all identified as known biomarkers in the literature (Data Supplement). These include *AURKA* in both bladder and breast cancer[36-38]; *AURKB*, *CDK4*, *EGFR*, and *PDGFRA* in brain cancers and gliomas[39-43]; *MET* in thyroid carcinomas and gastric cancers[44-47]; and *ROS1* in lung cancers.[48]

### Exploring Results for a Single Sample

To illustrate how our method is used in practice, we demonstrate the model on a single sample rather than on summary statistics over many samples. Figure 5 compares our method on a random tumor sample from TCGA to Treehouse's standard practice approach of pooling normal samples and applying a cutoff on the basis of the interquartile range on a selection of 85 cancer genes that could be targeted by an available therapy.[49,50] The random sample is labeled thyroid carcinoma in TCGA. More than 8,000 samples from the GTEx data set were used as the comparative normal data set, categorized by tissue type. The model automatically weights each tissue group and assigns a majority of the weight to thyroid tissue in GTEx. Where the pan-normal cutoff method returns a binary classification for each of the selected genes, our method returns a posterior predictive *P* value generated from a distribution informed by the background data sets that are most similar to the N-of-1 sample. Where there is disagreement in outlier classification between the two methods (*PIK3CB* and *CCND2*), the posterior distribution can be examined in the context of the highest weighted background data sets to clearly understand how the *P* value was generated. For example, our method does not identify *PIK3CB* as an outlier (given a *P* value cutoff of .05) because the method down-weights nonthyroid tissues, which have lower average expression for this gene than normal thyroid tissue.

## DISCUSSION

Our method avoids selection bias introduced by having to choose a single comparison background data set. It also provides continuous *P* values for genes that can be ranked and avoids missing borderline cases that would be ignored by existing cutoff methods. Moreover, in addition to under- and overexpression, the model quantifies the similarity of the analyst's sample to background comparison sets. Researchers can use this feature as a diagnostic: Diffuse weight distributions suggest that the model did not identify a strong set of matches among the background data sets. The model has also been demonstrated to be robust to false-positives, incomplete comparison datasets, and analysis of samples of mixed lineage.

These benefits do come with a trade-off in computation. The calculation of outliers through other methods can be very fast, whereas the runtime of this method increases quadratically with the number of genes and data sets. Moreover, this method uses MCMC, which is computationally demanding even after we used tricks to reduce runtime. The method is only appropriate for analyzing small targeted gene sets—fewer than approximately 200 ideally—because of the way model complexity scales. After approximately 200 genes, it is better to parallelize multiple runs for a single sample, which is facilitated by a Toil-based version of the workflow that makes scaling trivial and allows the method to run hundreds of samples in parallel on both standard local and cloud-based clusters.[51] The software provides intermediate output at every step in the workflow so that users can validate model convergence, assess the model's similarity metrics for background data sets to the N of 1, and examine every model parameter to reproduce how *P* values were calculated for every gene.

The model makes certain mathematical assumptions that we know to be unrealistic. First, it implicitly assumes

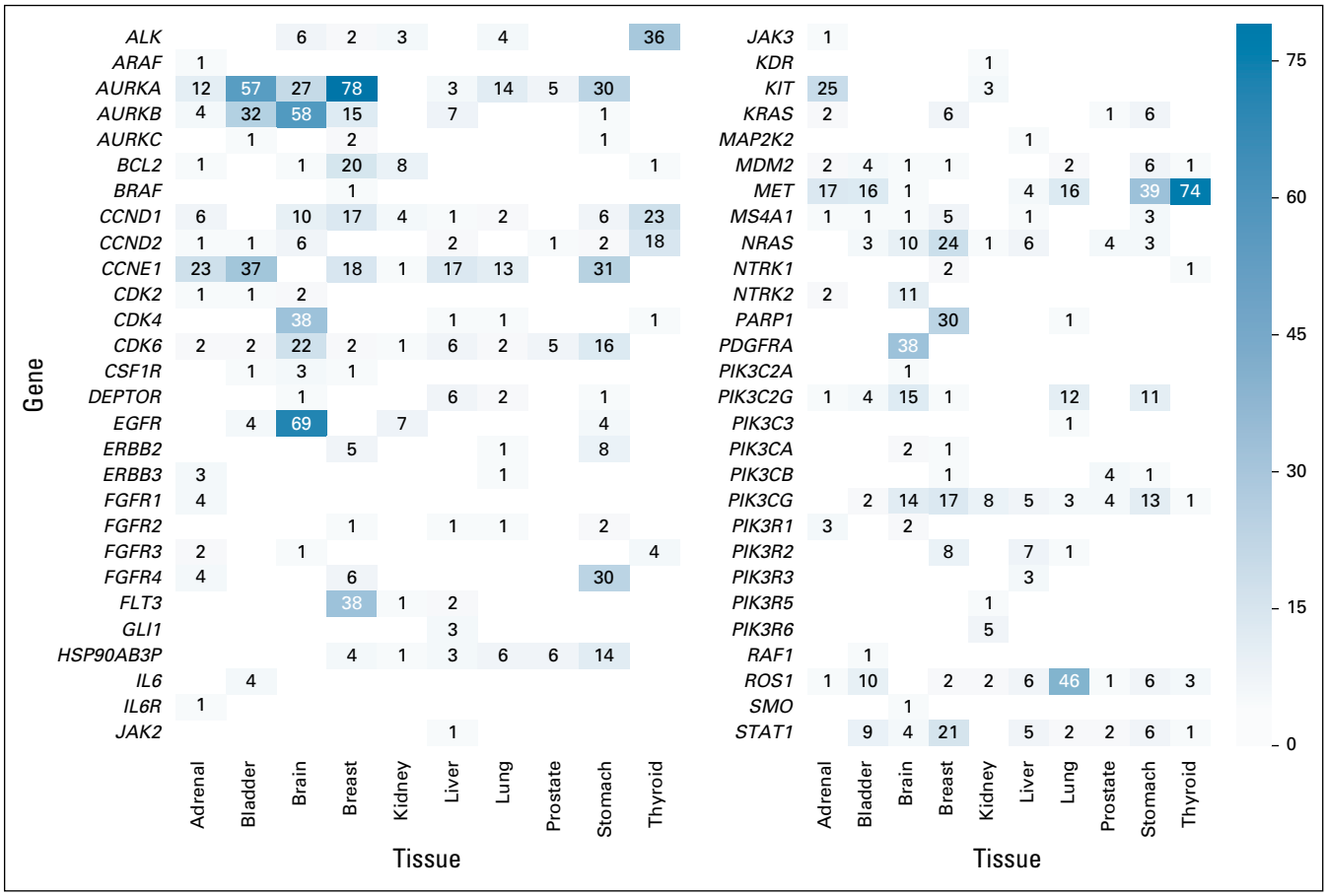


**FIG 4.** Outlier counts given a *P* value cutoff of .05. Eighty-five druggable genes curated by Treehouse were used as the target gene set for this analysis. All genes with counts for more than half of the samples within a tumor subtype were identified as known tumor biomarkers within that subtype in the literature (Data Supplement).

independence among the N-of-1 sample's genes. It also assumes that the sample's gene expression can be approximated by a convex mixture of the gene expression of the background data sets, which aids interpretability at the expense of descriptive accuracy. These limitations could be addressed by extending the model.

To incorporate correlation between genes, random variables for each $x_{d,g}$ could be replaced with multivariable distributions that are shared among all genes belonging to that group of co-expressed genes, where groups could be formed through a clustering process or according to existing annotations. Prior knowledge could also be used to introduce nonindependence between genes into the model. For instance, the independent error terms could be replaced by errors that are structured according to the Laplacian matrix of a gene interaction network.[52]

The method is also limited by the availability of ethnically diverse background data sets. For instance, 85% of GTEx samples are of European ancestry. If other subpopulations harbor genetic variants that affect expression, the baseline computed from a primarily European sample will be miscalibrated. While the magnitude of such differences in expression are small compared with differences between tissues (studies in lymphoblastic cell lines attributed only 2%-3% of variance in expression to ancestry[53]), this miscalibration has the potential to bias results in underrepresented populations. However, we note that the model can easily accommodate more granular subdivisions of the background data set by ancestry as such data become available.

Our model could be theoretically extended to single-cell RNA-seq analysis in addition to bulk. However, without any modifications to the existing model, this would require training the model for each cell, which would be computationally expensive. A faster alternative would be to cluster the cells and sample a small number of representative cells from each cluster to run through the model to get summary information about each of the single-cell clusters. However, the high levels of technical noise, biologic variability between cells, and dropouts may require too many training genes to obtain robust estimates of the model parameters.[54] Finally, the distribution of the random variable $x_{d,g}$ would need to be replaced with a more appropriate distribution to

 

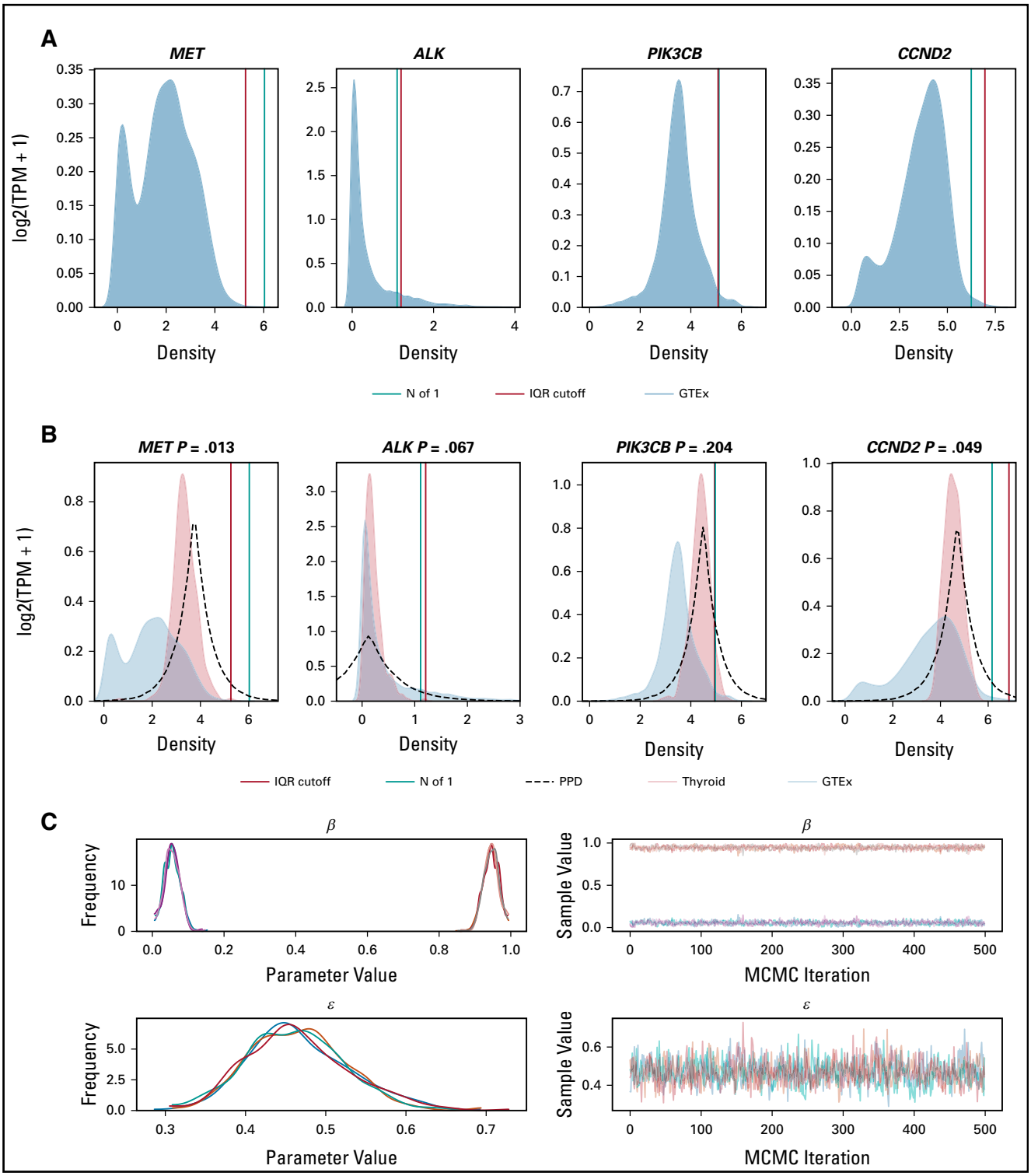


**FIG 5.** Comparison of results between a pan-normal cutoff approach and our model for a single sample. (A) Gene expression distributions for four genes using the Genotype-Tissue Expression Consortium (GTEx) data set overlaid with the gene expression value from our N-of-1 tumor in green and an outlier cutoff in red generated from Tukey's method (quartile 3 + interquartile range [IQR] × 1.5). (B) Same plots as in A along with the gene expression distribution for thyroid in GTEx (orange) and the posterior predictive distribution (dashed curve). The posterior distribution closely matches the GTEx thyroid distribution because the model assigned almost all of the weight to that tissue. Posterior predictive *P* values for each gene were added to the plot titles, which are based on the relative proportion of samples in the posterior distribution greater than the value from the N-of-1 sample. (C) Trace plots for $\beta$ and $\varepsilon$ from the Markov chain Monte Carlo (MCMC) sampling step. $\beta$ is a Dirichlet distribution and sums to 1 and, in this case, is assigned essentially all model weight to thyroid in GTEx. PPD, posterior predictive distribution; TPM, transcripts per million.

model single-cell expression, such as a β-Poisson mixture model.[55]

In conclusion, as clinicians have begun to demonstrate that RNA-seq analysis can produce actionable findings for patients with cancer, it is necessary to have informed and principled analytic tools for an individual patient. The method we have proposed detects gene expression outliers among a panel of target genes. It also provides additional information for researchers to explore and validate the results through examination of the model's parameters. For portability, scalability, and reproducibility, we have made this open source tool available as a Python package, Docker container, and Toil workflow available at https://github.com/jvivian/gene-outlier-detection.

### Availability of Data and Material

Raw expression data used in these experiments is available at University of California, Santa Cruz (UCSC) Xena: https://toil.xenahubs.net

All data used to produce every figure and experiment is publicly available at UCSC: http://courtyard.gi.ucsc.edu/~jvivian/outlier-paper

Our model's code is open source and available on GitHub: https://github.com/jvivian/gene-outlier-detection


## AFFILIATIONS

[1]Computational Genomics Laboratory, University of California, Santa Cruz, Santa Cruz, CA
[2]Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, Santa Cruz, CA


Preprint version available on bioRxiv.


## CORRESPONDING AUTHOR

Benedict Paten, PhD, Computational Genomics Laboratory, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA 95064; e-mail: bpaten@ucsc.edu.



## SUPPORT

Supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health under award number R01HG009737; a subagreement from the European Molecular Biology Laboratory with funds provided by NHGRI agreement number 2U41HG007234; the National Heart, Lung, and Blood Institute under award number U01HL137183; the W.M. Keck Foundation under award number DT06172015; the NHGRI under award number U54HG007990; St Baldrick's Foundation Treehouse Childhood Cancer Project (427053); and the California Precision Medicine Initiative: California Kids Cancer Comparison (OPR014109). The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of National Institutes of Health, NHGRI, or the European Molecular Biology Laboratory.


## AUTHOR CONTRIBUTIONS

**Conception and design:** John Vivian, Jordan M. Eizenga, Holly C. Beale, Benedict Paten
**Financial support:** Benedict Paten
**Collection and assembly of data:** John Vivian, Benedict Paten
**Data analysis and interpretation:** John Vivian, Jordan M. Eizenga, Olena M. Vaske, Benedict Paten
**Manuscript writing:** All authors
**Final approval of manuscript:** All authors
**Accountable for all aspects of the work:** All authors

## AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

**Jordan M. Eizenga**
**Employment:** Cambridge Epigenetix

**Olena M. Vaske**
**Employment:** NantOmics (I)
**Stock and Other Ownership Interests:** NantHealth (I)

**Benedict Paten**
**Honoraria:** Genentech
**Consulting or Advisory Role:** Calico, Cambridge Epigenetix

No other potential conflicts of interest were reported.


## ACKNOWLEDGMENT

We thank the members of the Computational Genomics Laboratory at the UCSC Genomics Institute for numerous conversations that helped to shape this work as well as the Treehouse organization for its collaborative efforts.


 

## REFERENCES


1. Golub TR, Slonim DK, Tamayo P, et al: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286:531-537, 1999
2. Ramaswamy S, Tamayo P, Rifkin R, et al: Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 98:15149-15154, 2001
3. Han L, Yuan Y, Zheng S, et al: The pan-cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. Nat Commun 5:3963, 2014
4. Best MG, Sol N, Kooi I, et al: RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. Cancer Cell 28:666-676, 2015
5. Patel AP, Tirosh I, Trombetta JJ, et al: Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344:1396-1401, 2014
6. Morozova O, Hirst M, Marra MA: Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10:135-151, 2009
7. Laskin J, Jones S, Aparicio S, et al: Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers. Cold Spring Harb Mol Case Stud 1:a000570, 2015
8. Brown AS, Kong SW, Kohane IS, et al: ksRepo: A generalized platform for computational drug repositioning. BMC Bioinformatics 17:78, 2016
9. Chen B, Butte AJ: Leveraging big data to transform target selection and drug discovery. Clin Pharmacol Ther 99:285-297, 2016
10. Raynal NJ-M, Da Costa EM, Lee JT, et al: Repositioning FDA-approved drugs in combination with epigenetic drugs to reprogram colon cancer epigenome. Mol Cancer Ther 16:397-407, 2017
11. Rodon J, Soria JC, Berger R, et al: Challenges in initiating and conducting personalized cancer therapy trials: Perspectives from WINTHER, a Worldwide Innovative Network (WIN) Consortium trial. Ann Oncol 26:1791-1798, 2015
12. Mody RJ, Wu YM, Lonigro RJ, et al: Integrative clinical sequencing in the management of refractory or relapsed cancer in youth. JAMA 314:913-925, 2015
13. Chang W, Brohl AS, Patidar R, et al: multidimensional clinomics for precision therapy of children and adolescent young adults with relapsed and refractory cancer: A report from the Center for Cancer Research. Clin Cancer Res 22:3810-3820, 2016
14. Worst BC, van Tilburg CM, Balasubramanian GP, et al: Next-generation personalised medicine for high-risk paediatric cancer patients—the INFORM pilot study. Eur J Cancer 65:91-101, 2016
15. Oberg JA, Glade Bender JL, Sulis ML, et al: Implementation of next generation sequencing into pediatric hematology-oncology practice: Moving beyond actionable alterations. Genome Med 8:133, 2016
16. Morozova O, Newton, Y, Cline M, et al: Treehouse Childhood Cancer Project: A resource for sharing and multiple cohort analysis of pediatric cancer genomics data. Cancer Res 75, 2015 (suppl; abstr LB-212)
17. Beale H, Lam DL, Vivian J, et al: Identifying confidently measured genes in single pediatric cancer patient samples using RNA sequencing. Cancer Res 77, 2017 (suppl; abstr 2466)
18. Morozova O, Newton Y, Sha AT, et al: A pan-cancer analysis framework for incorporating gene expression information into clinical interpretation of pediatric cancer genomic data. Cancer Res 77, 2017 (suppl; abstr 4890)
19. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550, 2014
20. Li J, Tibshirani R: Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res 22:519-536, 2013
21. Zhou X, Lindsay H, Robinson MD: Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res 42:e91, 2014
22. Soneson C, Delorenzi M: A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14:91, 2013
23. Marusyk A, Polyak K: Tumor heterogeneity: Causes and consequences. Biochim Biophys Acta 1805:105-117, 2010
24. Feng J, Meyer CA, Wang Q, et al: GFOLD: A generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics 28:2782-2788, 2012
25. Beniger JR, Tukey JW: exploratory data analysis. Contemp Sociol 7:64, 1978
26. Kothari V, Wei I, Shankar S, et al: Outlier kinase expression by RNA sequencing as targets for precision therapy. Cancer Discov 3:280-293, 2013
27. Jones SJ, Laskin J, Li YY, et al: Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. Genome Biol 11:R82, 2010
28. Weinstein JN, Collisson EA, Mills GB, et al: The Cancer Genome Atlas pan-cancer analysis project. Nat Genet 45:1113-1120, 2013
29. Hoffman M, Gelman A: The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 15:1593-1623, 2014
30. Salvatier J, Wiecki TV, Fonnesbeck C: Probabilistic programming in Python using PyMC3. PeerJ Comput Sci 2:e55, 2016
31. GTEx Consortium: Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348:648-660, 2015
32. Axelsen JB, Lotem J, Sachs L, et al: Genes overexpressed in different human solid cancers exhibit different tissue-specific expression profiles. Proc Natl Acad Sci U S A 104:13122-13127, 2007 [Erratum: Proc Natl Acad Sci U S A 104:15168, 2007]
33. Lotem J, Netanely D, Domany E, et al: Human cancers overexpress genes that are specific to a variety of normal human tissues. Proc Natl Acad Sci U S A 102:18556-18561, 2005
34. Hogarty MD, Brodeur GM: Gene Amplification in human cancers: Biological and clinical significance, in Valle D, Antonarakis S, Ballabio A, et al: The Online Metabolic & Molecular Bases of Inherited Disease. New York, NY, McGraw-Hill, 2019
35. Chen Y, McGee J, Chen X, et al: Identification of druggable cancer driver genes amplified across TCGA datasets. PLoS One 9:e98293, 2014 [Erratum: PLoS One 9:e107646, 2014]
36. Shirodkar SP, Lokeshwar VB: Potential new urinary markers in the early detection of bladder cancer. Curr Opin Urol 19:488-493, 2009
37. Park H-S, Park WS, Bondaruk J, et al: Quantitation of aurora kinase A gene copy number in urine sediments and bladder cancer detection. J Natl Cancer Inst 100:1401-1411, 2008
38. Staff S, Isola J, Jumppanen M, et al: Aurora-A gene is frequently amplified in basal-like breast cancer. Oncol Rep 23:307-312, 2010
39. Bie L, Zhao G, Wang Y-P, et al: Kinesin family member 2C (KIF2C/MCAK) is a novel marker for prognosis in human gliomas. Clin Neurol Neurosurg 114:356-360, 2012
40. Jacques TS, Swales A, Brzozowski MJ, et al: Combinations of genetic mutations in the adult neural stem cell compartment determine brain tumour phenotypes. EMBO J 29:222-235, 2010

41. Broniscer A, Baker SJ, West AN, et al: Clinical and molecular characteristics of malignant transformation of low-grade glioma in children. J Clin Oncol 25:682-689, 2007
42. Zohrabian VM, Nandu H, Gulati N, et al: Gene expression profiling of metastatic brain cancer. Oncol Rep 18:321-328, 2007 https://doi.org/10.3892/or.18.2.321
43. Puputti M, Tynninen O, Sihto H, et al: Amplification of KIT, PDGFRA, VEGFR2, and EGFR in gliomas. Mol Cancer Res 4:927-934, 2006
44. Lee J, Seo JW, Jun HJ, et al: Impact of MET amplification on gastric cancer: Possible roles as a novel prognostic marker and a potential therapeutic target. Oncol Rep 25:1517-1524, 2011
45. Di Renzo MF, Olivero M, Ferro S, et al: Overexpression of the c-MET/HGF receptor gene in human thyroid carcinomas. Oncogene 7:2549-2553, 1992
46. Janjigian YY, Tang LH, Coit DG, et al: MET expression and amplification in patients with localized gastric cancer. Cancer Epidemiol Biomarkers Prev 20:1021-1027, 2011
47. Catenacci DVT, Henderson L, Xiao SY, et al: Durable complete response of metastatic gastric cancer with anti-Met therapy followed by resistance at recurrence. Cancer Discov 1:573-579, 2011
48. Chin LP, Soo RA, Soong R, et al: Targeting ROS1 with anaplastic lymphoma kinase inhibitors: A promising therapeutic strategy for a newly defined molecular subset of non-small-cell lung cancer. J Thorac Oncol 7:1625-1630, 2012
49. Vaske OM, Bjork I, Salama SR, et al: Comparative tumor RNA sequencing analysis for difficult-to-treat pediatric and young adult patients with cancer. JAMA Netw Open 2:e1913968, 2019
50. Newton Y, Rassekh SR, Deyell RJ, et al: Comparative RNA-sequencing analysis benefits a pediatric patient with relapsed cancer. JCO Precis Oncol
51. Vivian J, Rao AA, Nothaft FA, et al: Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol 35:314-316, 2017
52. Sokolov A, Carlin DE, Paull EO, et al: Pathway-based genomics prediction using generalized elastic net. PLOS Comput Biol 12:e1004790, 2016
53. Kelly DE, Hansen MEB, Tishkoff SA: Global variation in gene expression and the value of diverse sampling. Curr Opin Syst Biol 1:102-108, 2017
54. Kharchenko PV, Silberstein L, Scadden DT: Bayesian approach to single-cell differential expression analysis. Nat Methods 11:740-742, 2014
55. Vu TN, Wills QF, Kalari KR, et al: Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics 32:2128-2135, 2016