Tools for large-scale genomic analysis and gene expression outlier modeling for precision therapeutics

John Vivian

6-6-19

NIH Big Data to Knowledge (BD2K)

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Outline

01

Background

02

Large-scale Compute

03

Outlier Detection

01

Background

Genomics is the leader in “big data”

Source	Size	Data Type
TCGA	2.5 petabytes	Genomic, epigenomic, transcriptomic, proteomic
SRA	11 petabytes	Sequence data
MinION	10-30 Gb per flow cell	DNA sequence data
Human Cell Atlas	> 10 petabytes	Primarily sequence data

Table of Bytes

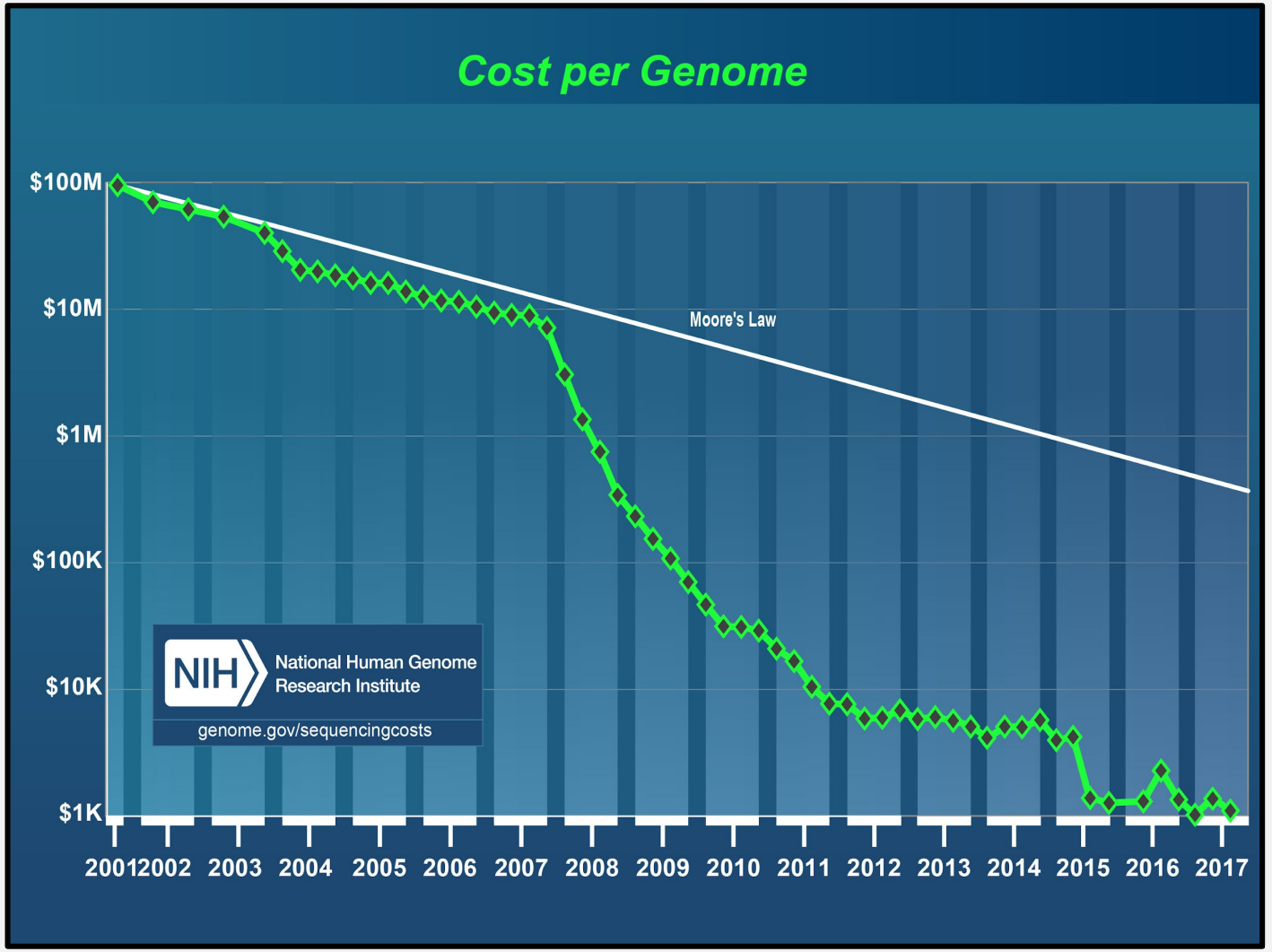
Bytes	Unit
10^6	megabyte
10^9	gigabyte
10^{12}	terabyte
10^{15}	petabyte
10^{18}	exabyte
10^{21}	zettabyte

Projected annual storage and compute needs of genomics by 2025

Acquisition	Storage	Variant Calling	Genome Alignments
1 zetta-bases / year	2-40 EB / year	~2 trillion CPU hours	~10,000 trillion CPU hours

Stephens et al. (2015), Oxford Nanopore Technologies, NIH

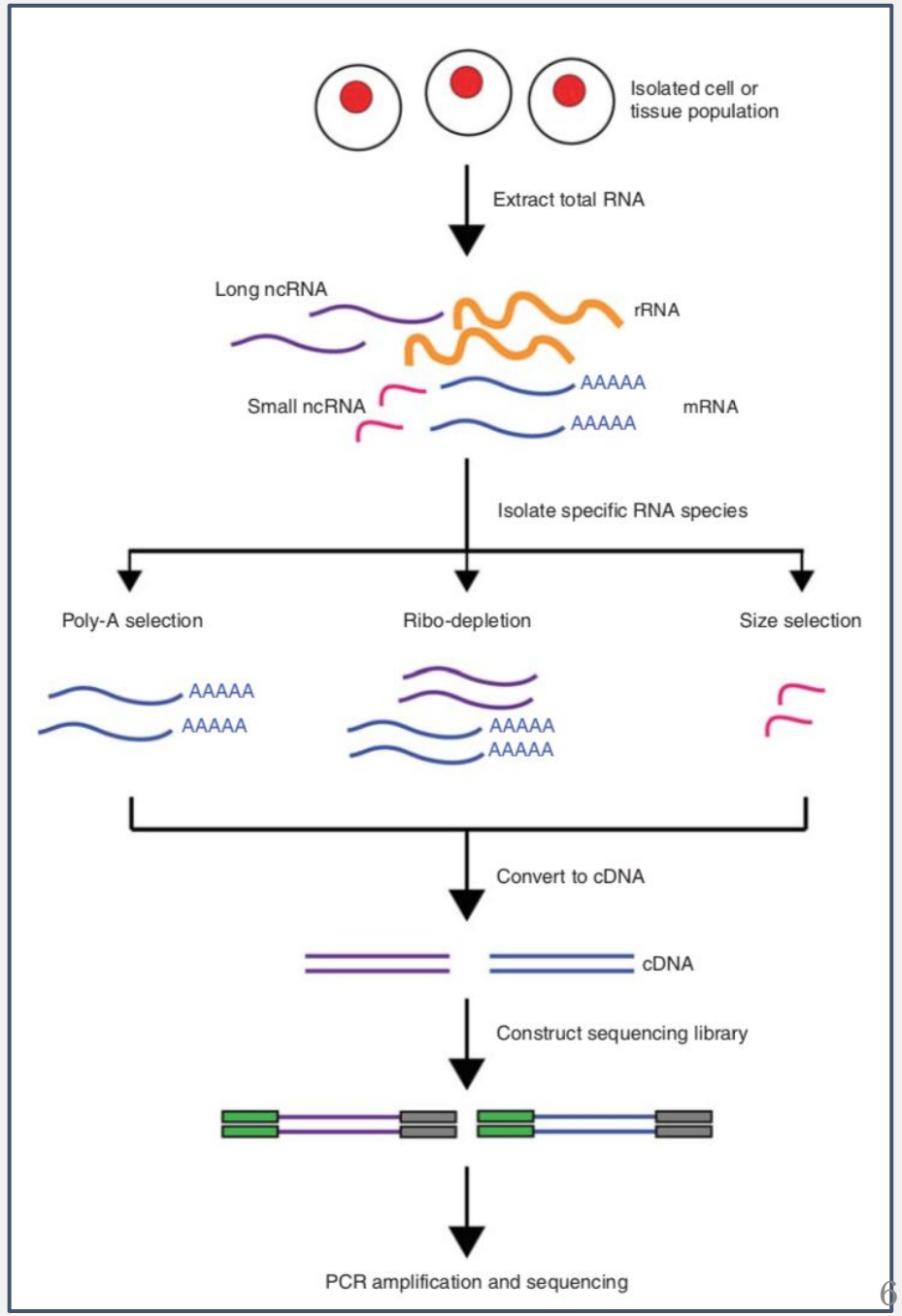
Explosion in genomic data is driven by sequencing costs



RNA Sequencing

RNA-seq — reveals the presence and quantity of RNA in a biological sample at a given moment in time

Different library preparations select for different types of RNA, but once converted into cDNA, can leverage high-throughput DNA sequencing technology

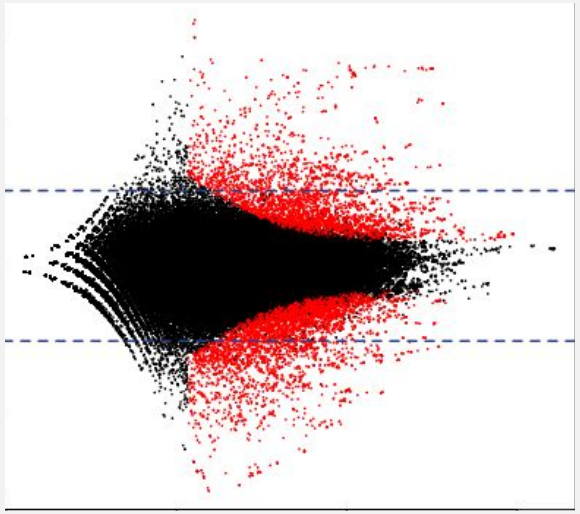


Kukurba and Montgomery (2015)

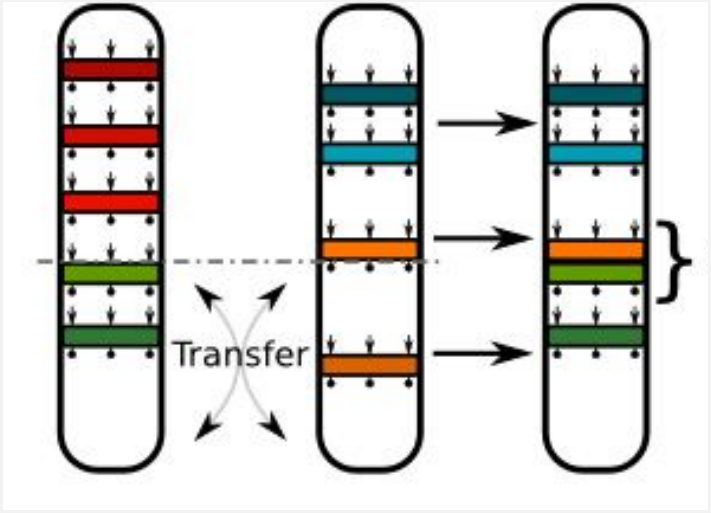
How does the cancer field leverage RNA-seq?

RNA-seq - reveals the presence and quantity of RNA in a biological sample at a given moment

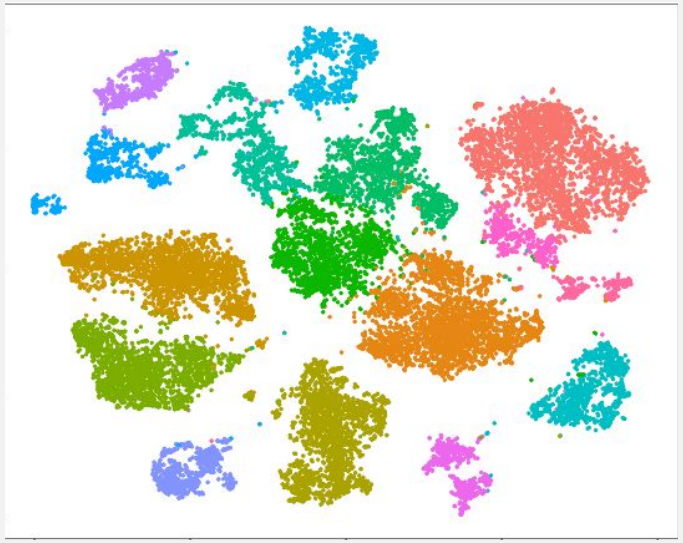
DIfferential Expression



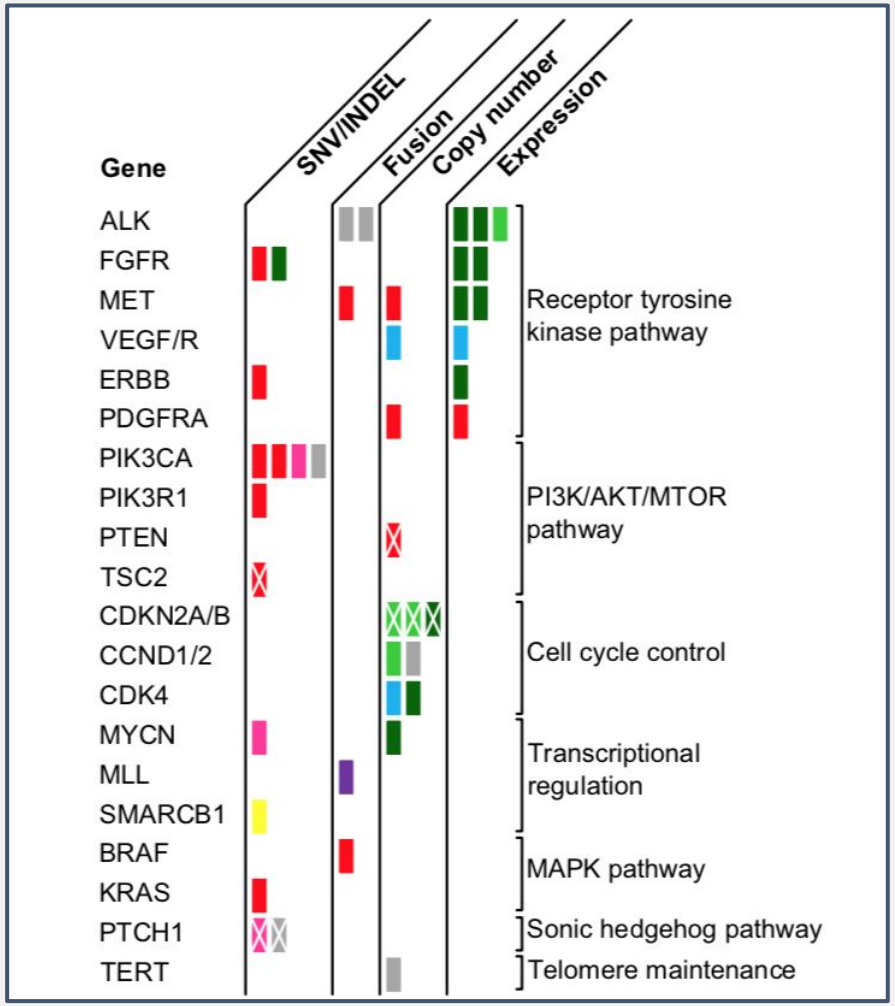
Fusion Genes



Tumor Heterogeneity



RNA-seq in precision medicine



Worst et al. (2016)

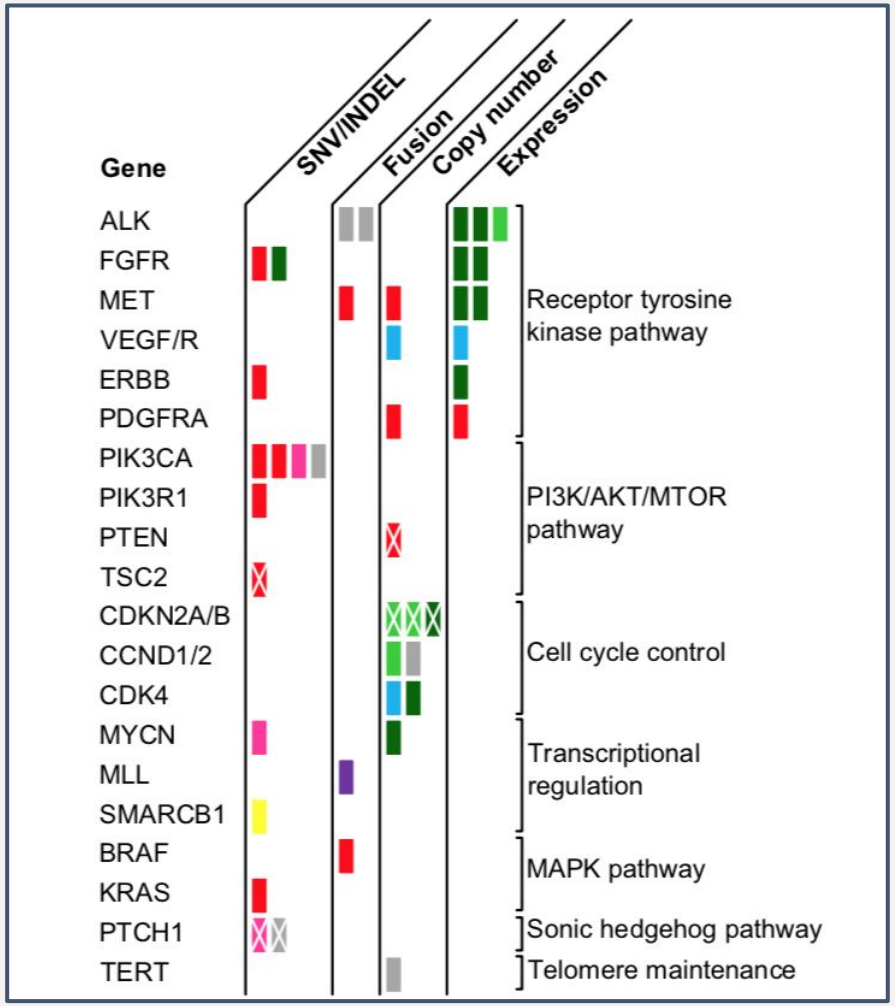
"RNA-seq discoveries alone accounted for almost **20% of the actionable findings in our study**, which would have been missed otherwise."

Mody et al. (2016)

"RNA-seq was clinically impactful in 37/65 patients (57%) providing diagnostic and/or prognostic information for 17 patients (26%) and **identified therapeutic targets in 15 patients (23%)**"

Oberg et al. (2016)

RNA-seq in precision medicine



Worst et al. (2016)

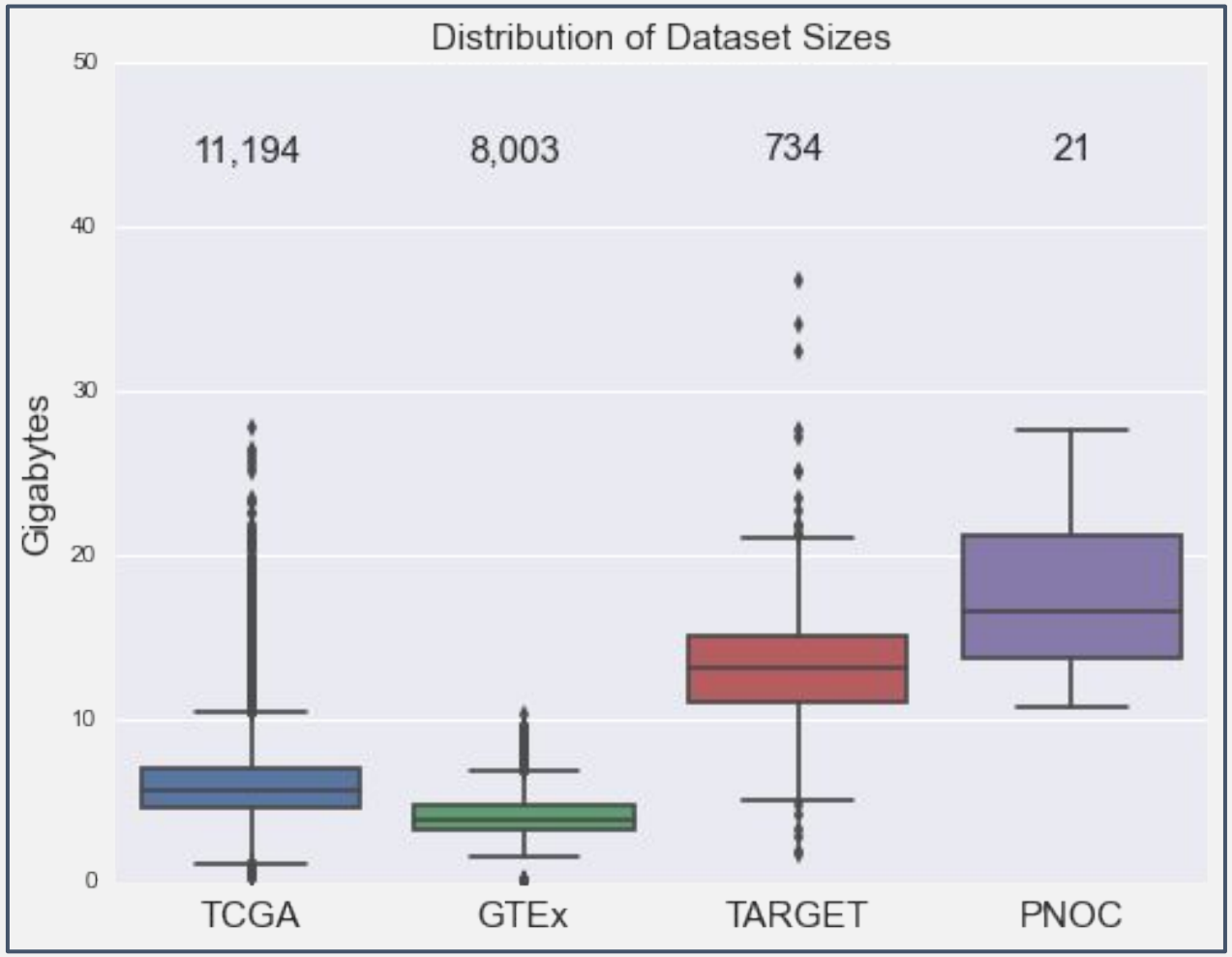
"RNA-seq discoveries alone accounted for almost **20% of the actionable findings in our study**, which would have been missed otherwise."

Mody et al. (2016)

"RNA-seq was clinically impactful in 37/65 patients (57%) providing diagnostic and/or prognostic information for 17 patients (26%) and **identified therapeutic targets in 15 patients (23%)**"

Oberg et al. (2016)

Size of RNA-seq datasets



TCGA — The Cancer Genome Atlas

GTEx — The Genotype Tissue Expression Consortium

TARGET — Therapeutically Applicable Research to Generate Effect Treatment

PNOC — Pacific Pediatric Neuro-Oncology Consortium

Vivian et al. (2017)

Motivating question

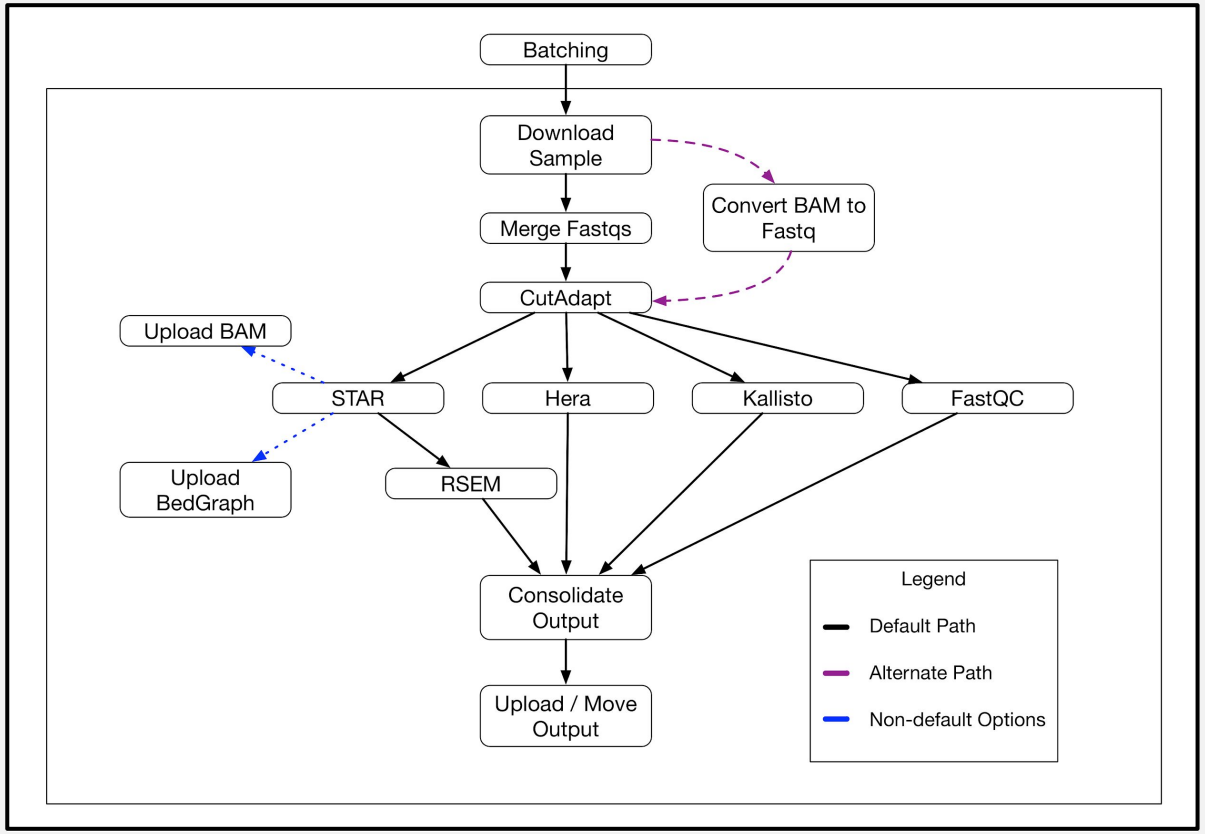
How can massive amounts of genomic data be consistently, efficiently, and reproducibly processed at scale?

02

Large-scale Compute

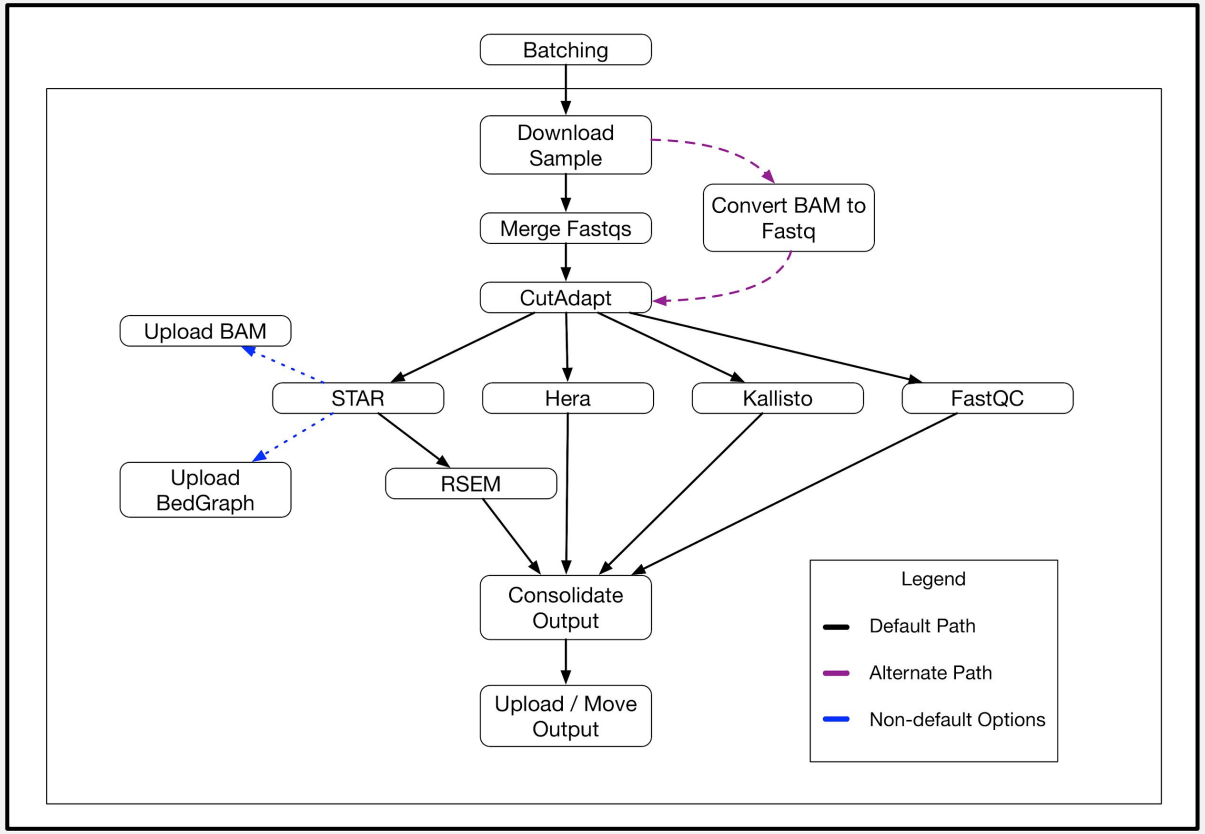
What is a workflow?

- **Workflow —** Comprises a set of tasks, or *jobs*, that are orchestrated by specification of a set of dependencies that map the inputs and outputs between jobs
- **Node —** Represents a *job*, an atomic unit of work
- **Edge —** Represents dependency between jobs



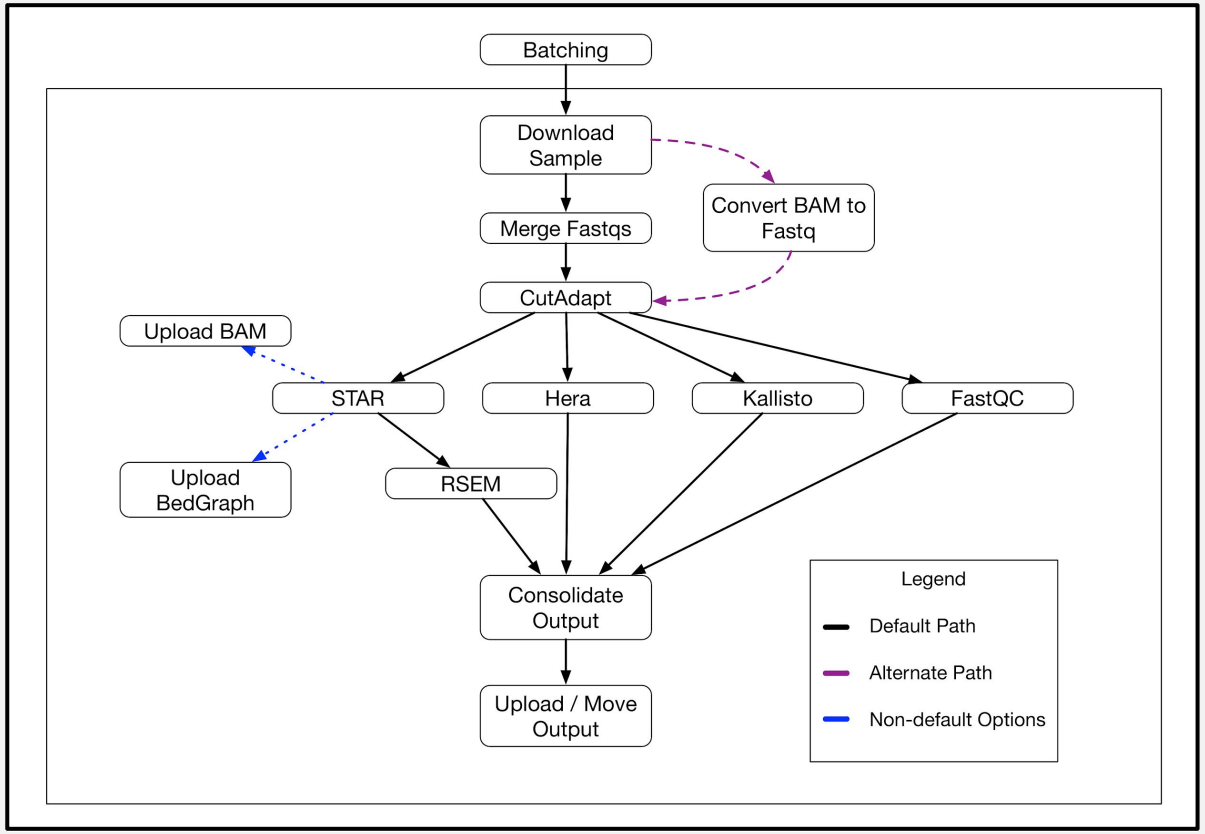
What is a workflow?

- **Workflow —** Comprises a set of tasks, or *jobs*, that are orchestrated by specification of a set of dependencies that map the inputs and outputs between jobs
- **Node —** Represents a *job*, an atomic unit of work
- **Edge —** Represents dependency between jobs



What is a workflow?

- **Workflow —** Comprises a set of tasks, or *jobs*, that are orchestrated by specification of a set of dependencies that map the inputs and outputs between jobs
- **Node —** Represents a *job*, an atomic unit of work
- **Edge —** Represents dependency between jobs



Toil: scalable workflow execution engine

- Efficient
  - Fast and scalable
- Python
  - Easy implementation
- Robust
  - Resume from failure
- Cloud Support
  - Auto-scaling based on resource demands
- Static and Dynamic Workflows
  - Elegant and powerful

Toil: scalable workflow execution engine

- Efficient
  - Fast and scalable
- Python
  - Easy implementation
- Robust
  - Resume from failure
- Cloud Support
  - Auto-scaling based on resource demands
- Static and Dynamic Workflows
  - Elegant and powerful

Toil: scalable workflow execution engine

- Efficient
  - Fast and scalable
- Python
  - Easy implementation
- Robust
  - Resume from failure
- Cloud Support
  - Auto-scaling based on resource demands
- Static and Dynamic Workflows
  - Elegant and powerful

Toil: scalable workflow execution engine

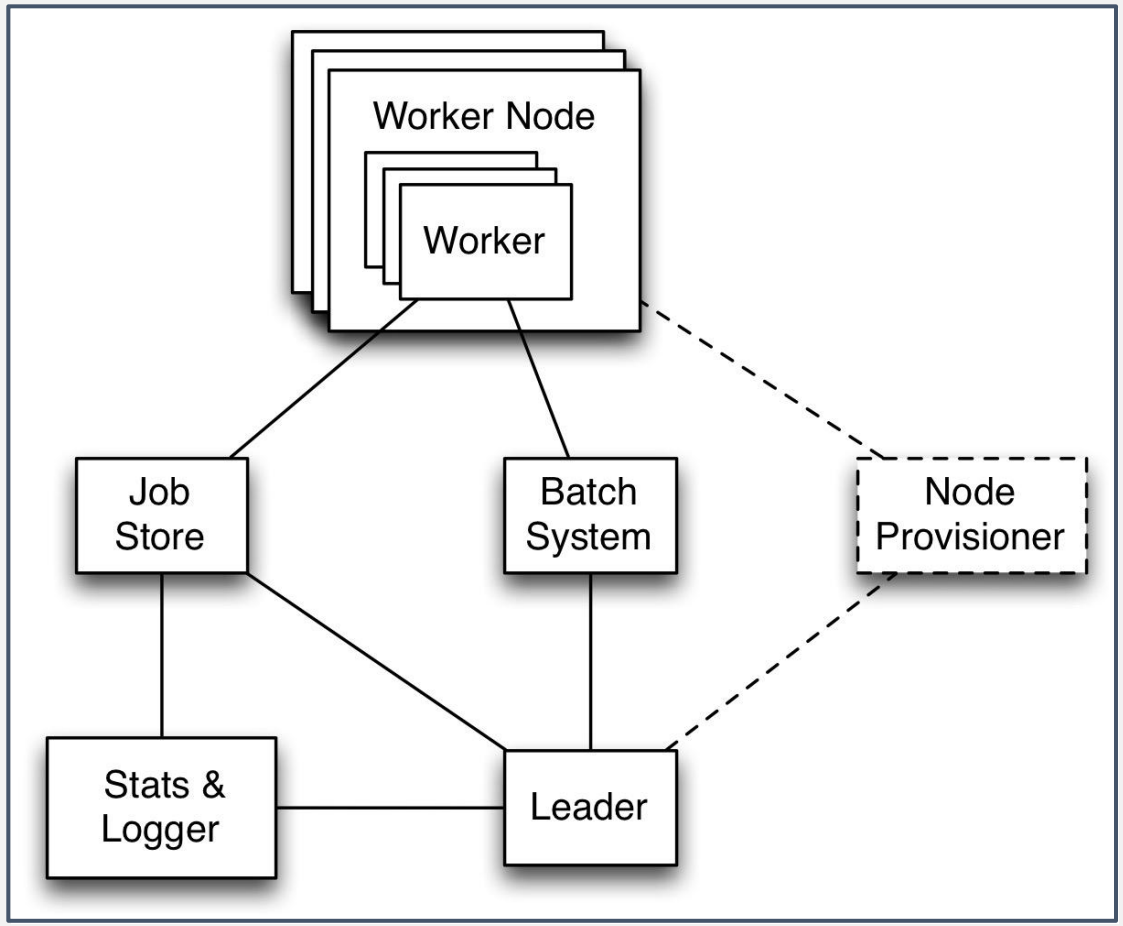
- Efficient
  - Fast and scalable
- Python
  - Easy implementation
- Robust
  - Resume from failure
- Cloud Support
  - Auto-scaling based on resource demands
- Static and Dynamic Workflows
  - Elegant and powerful

Toil: scalable workflow execution engine

- Efficient
  - Fast and scalable
- Python
  - Easy implementation
- Robust
  - Resume from failure
- Cloud Support
  - Auto-scaling based on resource demands
- Static and Dynamic Workflows
  - Elegant and powerful

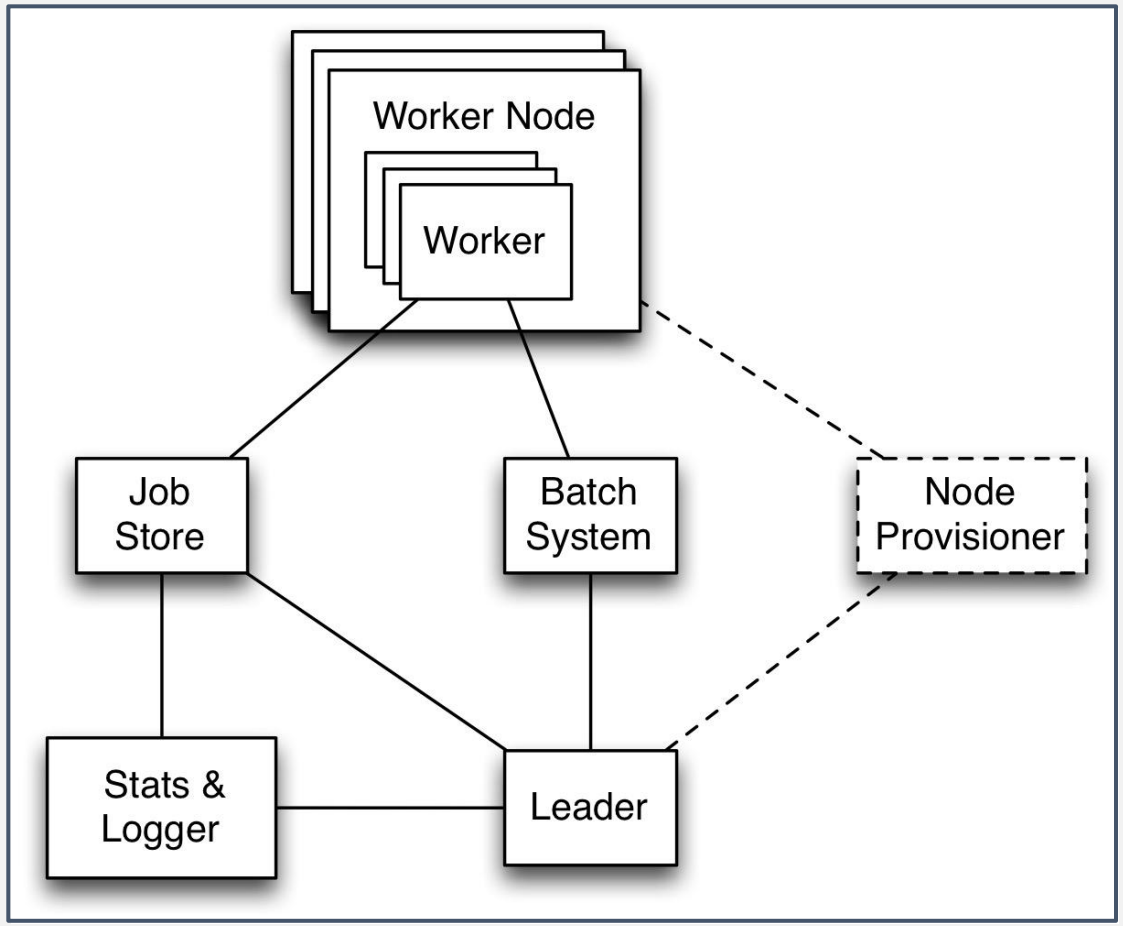
How does Toil work?

- **Leader —** Responsible for deciding which jobs should run by traversing the job graph
- **Worker** — Temporary process responsible for running one job at a time
- **Job Store** —During execution, intermediate files and job information are stored here atomically, which also permits resumption if failure occurs
- **Batch System** — Job execution orchestration



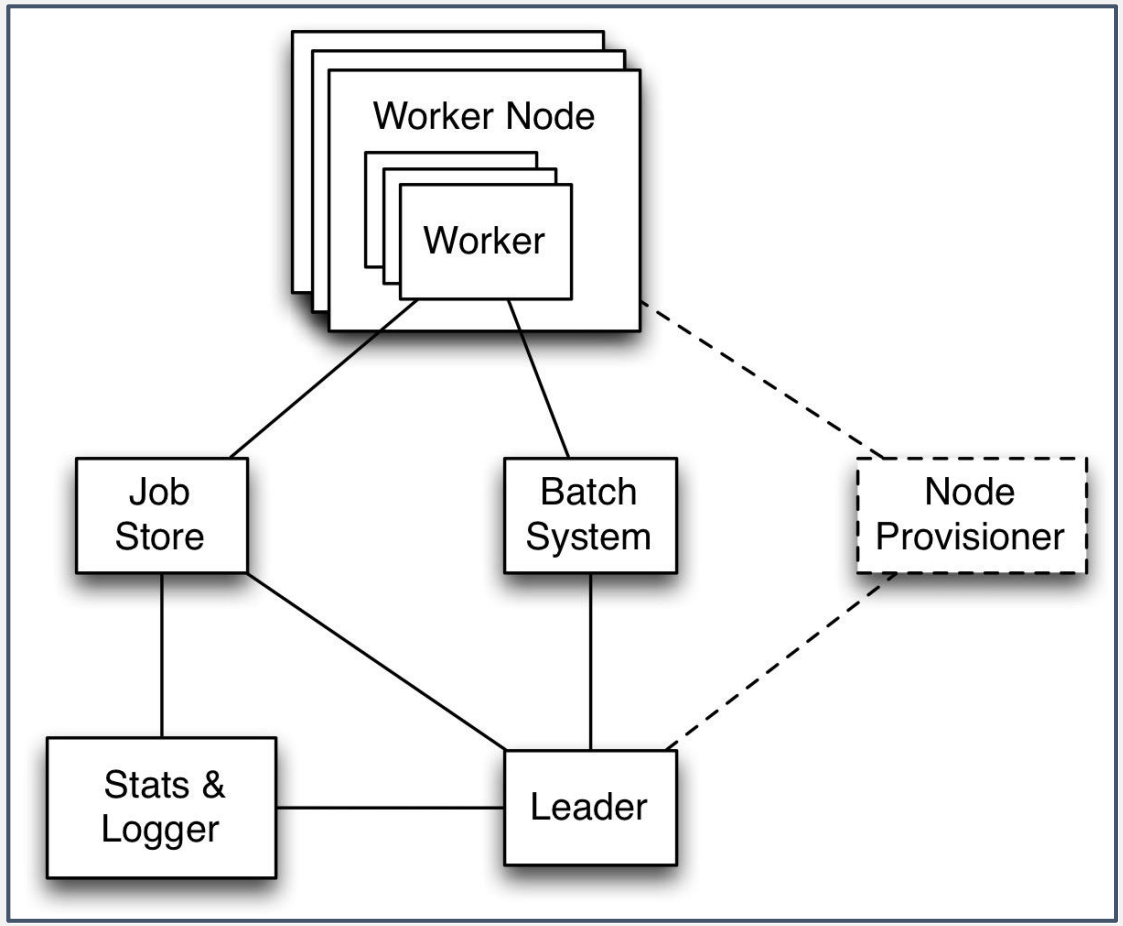
How does Toil work?

- **Leader —** Responsible for deciding which jobs should run by traversing the job graph
- **Worker** — Temporary process responsible for running one job at a time
- **Job Store** —During execution, intermediate files and job information are stored here atomically, which also permits resumption if failure occurs
- **Batch System** — Job execution orchestration



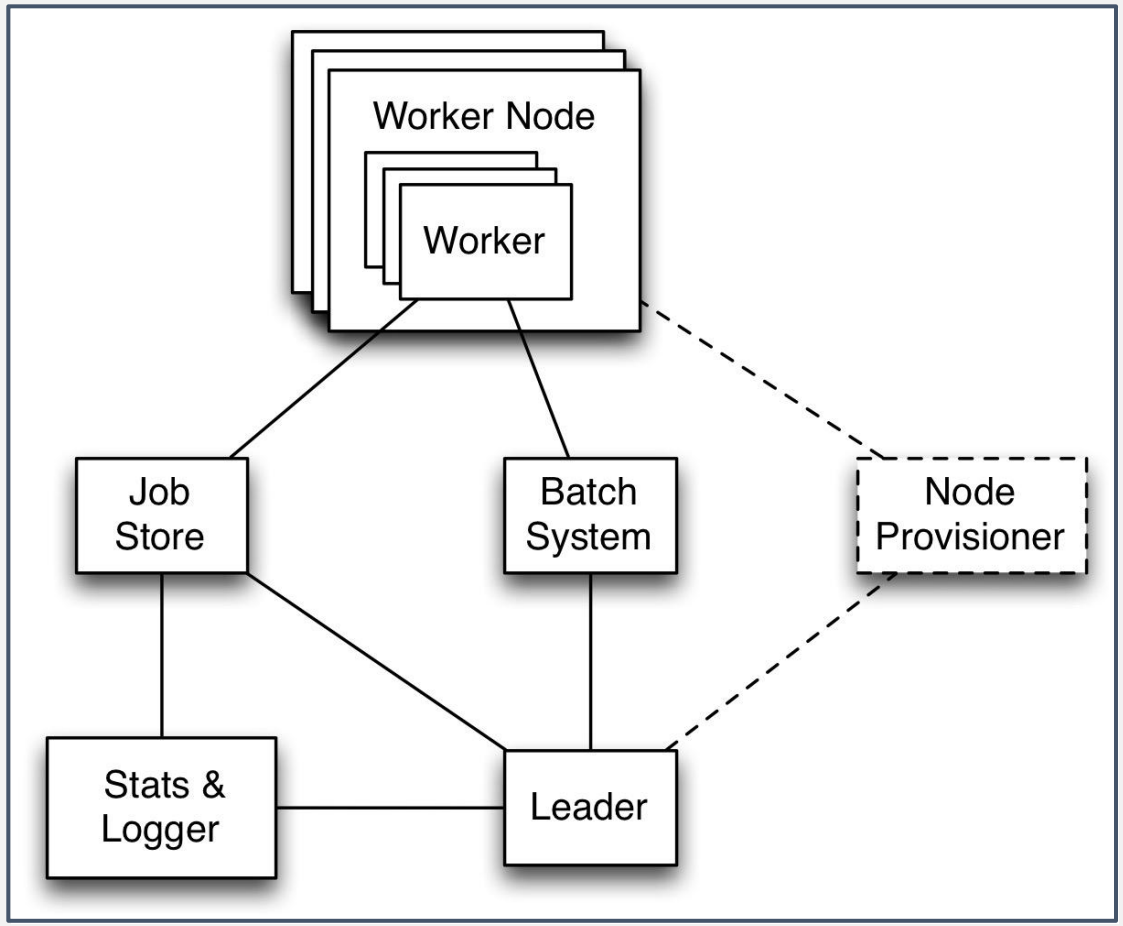
How does Toil work?

- **Leader —** Responsible for deciding which jobs should run by traversing the job graph
- **Worker** — Temporary process responsible for running one job at a time
- **Job Store —**During execution, intermediate files and job information are stored here atomically, which also permits resumption if failure occurs
- **Batch System** — Job execution orchestration



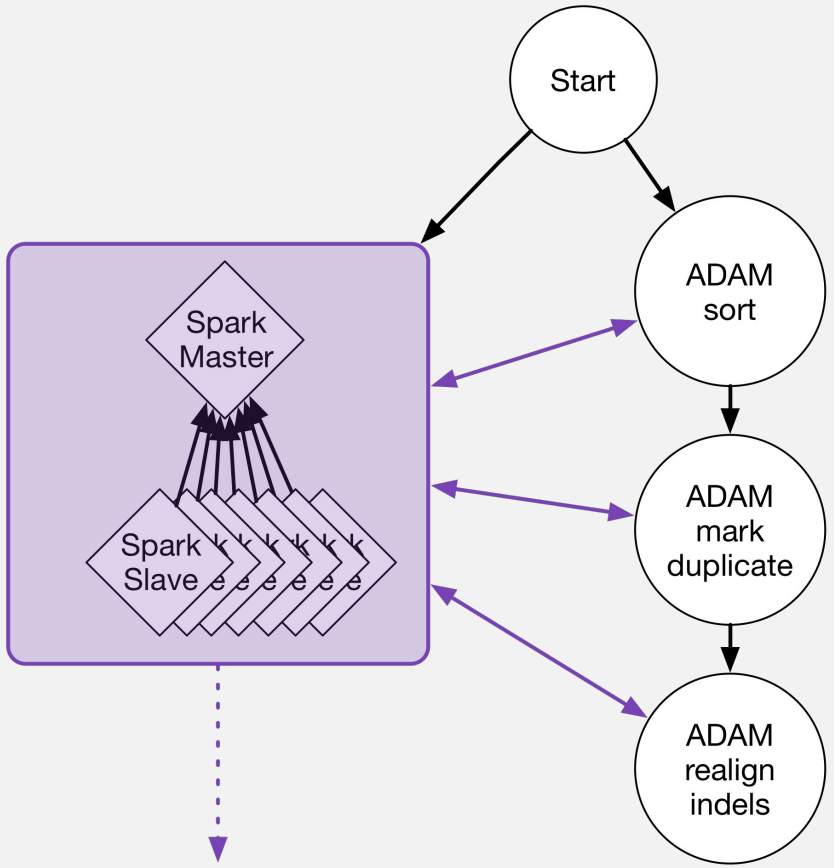
How does Toil work?

- **Leader —** Responsible for deciding which jobs should run by traversing the job graph
- **Worker** — Temporary process responsible for running one job at a time
- **Job Store —**During execution, intermediate files and job information are stored here atomically, which also permits resumption if failure occurs
- **Batch System —** Job execution orchestration



Toil has many nifty features and optimizations

- **Services** — Support for long-running services such as an external database or Spark cluster
- **Caching** — Caches job results so child jobs on the same node can reuse file objects which eliminates wasteful I/O
- **Autoscaling** — Automatic worker provisioning and scaling based on resource demands
- **3rd party language support** — Support for Common Workflow Language (CWL) and Workflow Description Language (WDL)
- Job chaining, preemptable node support, and more...

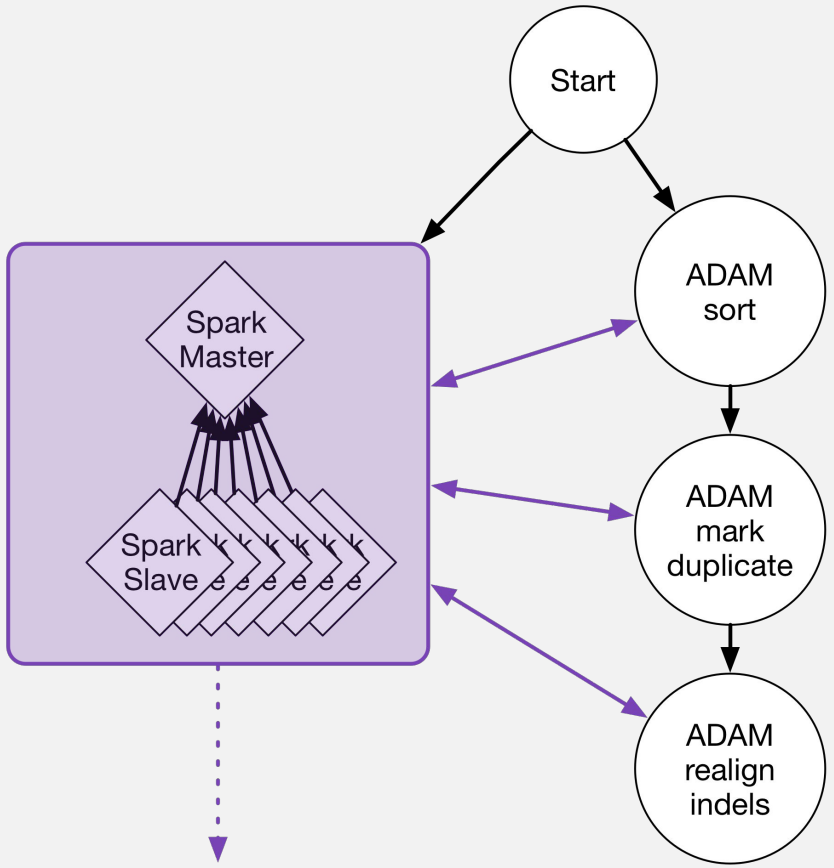


Spark cluster as a **service**

Vivian et al. (2017)

Toil has many nifty features and optimizations

- **Services** — Support for long-running services such as an external database or Spark cluster
- **Caching** — Caches job results so child jobs on the same node can reuse file objects which eliminates wasteful I/O
- **Autoscaling** — Automatic worker provisioning and scaling based on resource demands
- **3rd party language support** — Support for Common Workflow Language (CWL) and Workflow Description Language (WDL)
- Job chaining, preemptable node support, and more...

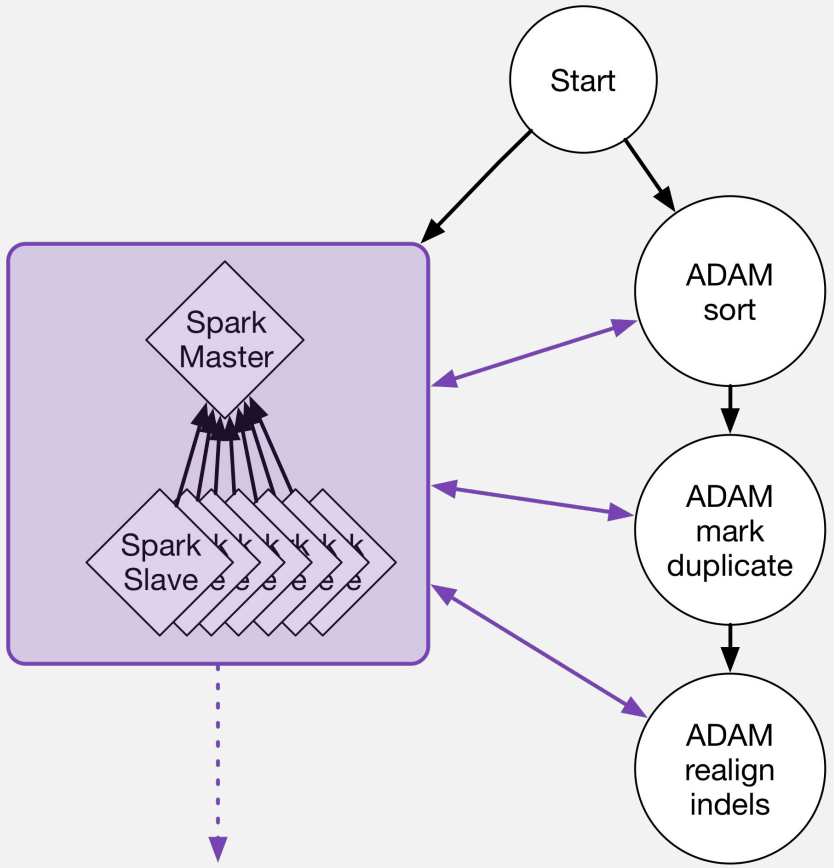


Spark cluster as a **service**

Vivian et al. (2017)

Toil has many nifty features and optimizations

- **Services** — Support for long-running services such as an external database or Spark cluster
- **Caching** — Caches job results so child jobs on the same node can reuse file objects which eliminates wasteful I/O
- **Autoscaling** — Automatic worker provisioning and scaling based on resource demands
- **3rd party language support** — Support for Common Workflow Language (CWL) and Workflow Description Language (WDL)
- Job chaining, preemptable node support, and more...

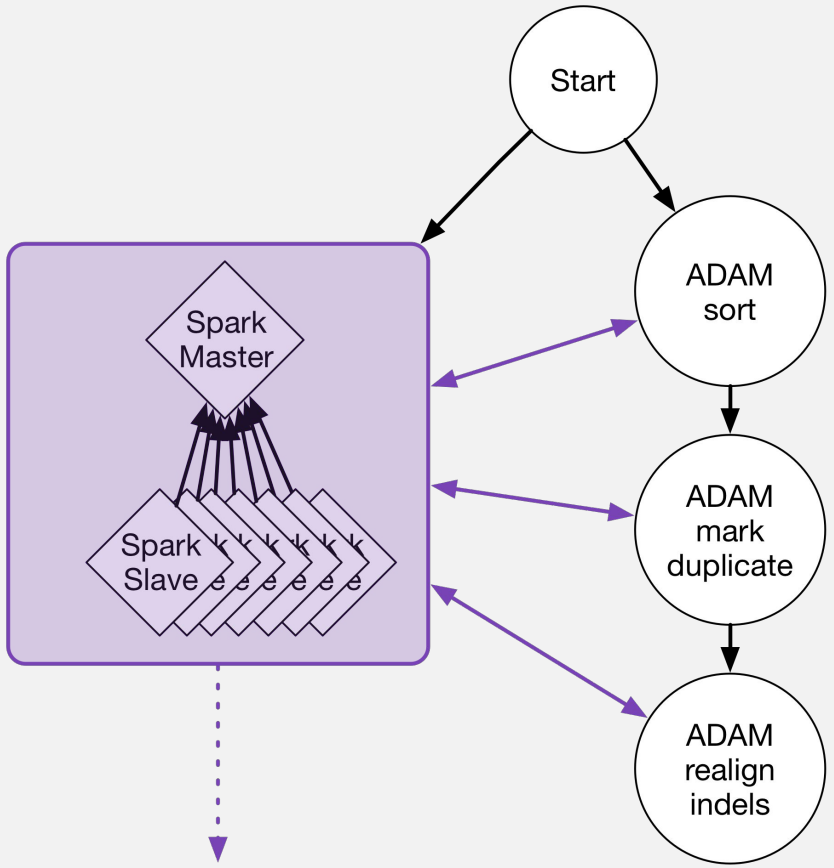


Spark cluster as a **service**

Vivian et al. (2017)

Toil has many nifty features and optimizations

- **Services** — Support for long-running services such as an external database or Spark cluster
- **Caching** — Caches job results so child jobs on the same node can reuse file objects which eliminates wasteful I/O
- **Autoscaling** — Automatic worker provisioning and scaling based on resource demands
- **3rd party language support** — Support for Common Workflow Language (CWL) and Workflow Description Language (WDL)
- Job chaining, preemptable node support, and more...

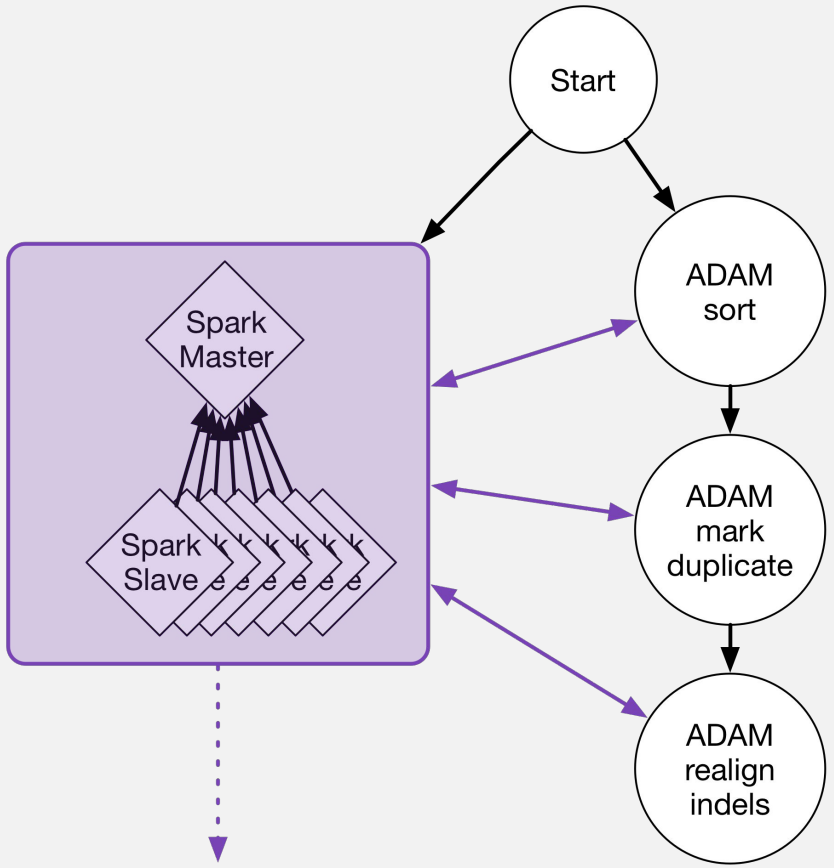


Spark cluster as a **service**

Vivian et al. (2017)

Toil has many nifty features and optimizations

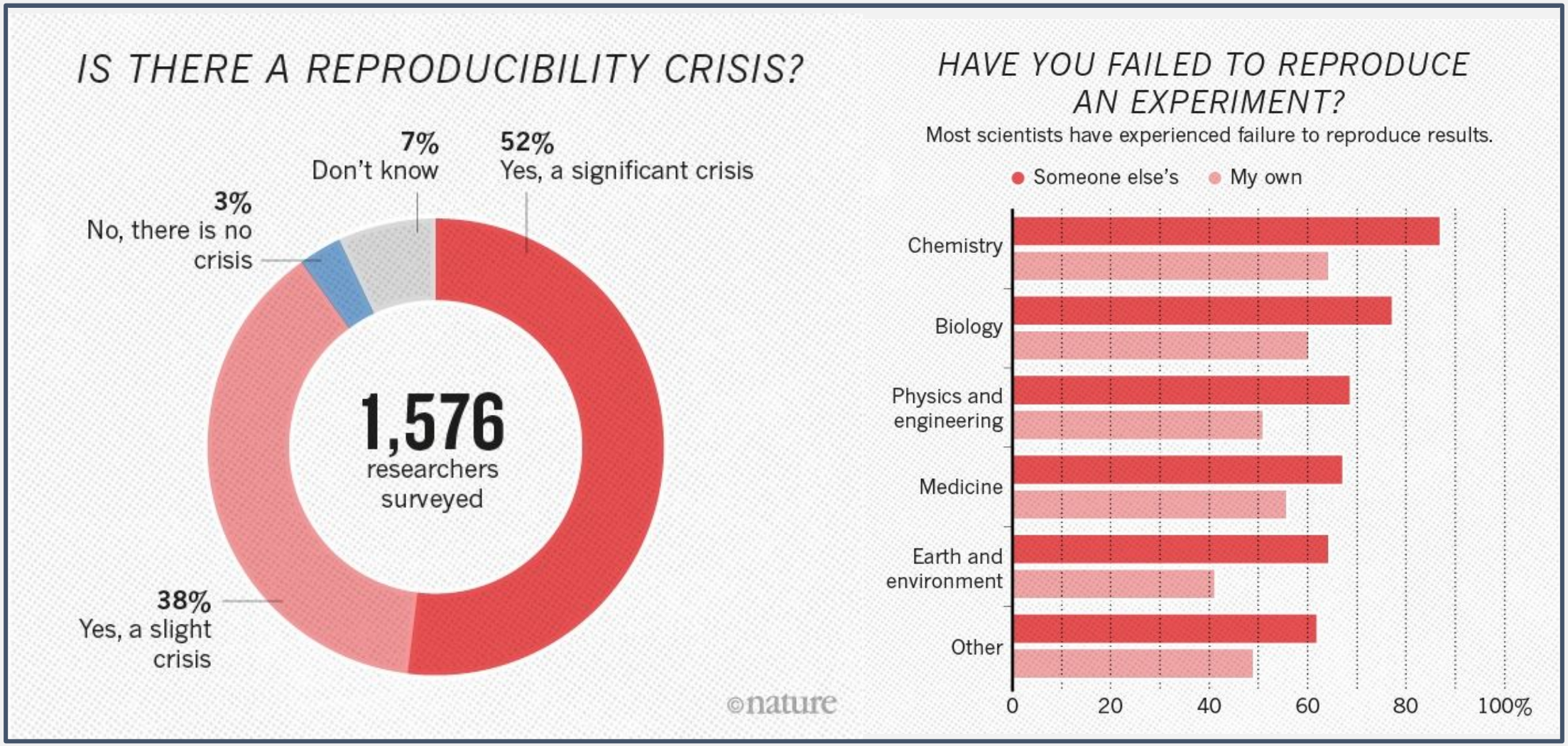
- **Services** — Support for long-running services such as an external database or Spark cluster
- **Caching** — Caches job results so child jobs on the same node can reuse file objects which eliminates wasteful I/O
- **Autoscaling** — Automatic worker provisioning and scaling based on resource demands
- **3rd party language support** — Support for Common Workflow Language (CWL) and Workflow Description Language (WDL)
- Job chaining, preemptable node support, and more...



Spark cluster as a **service**

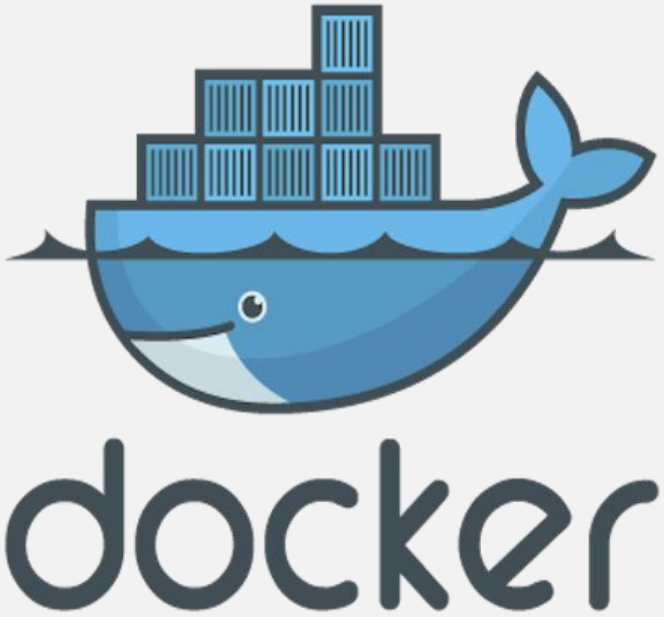
Vivian et al. (2017)

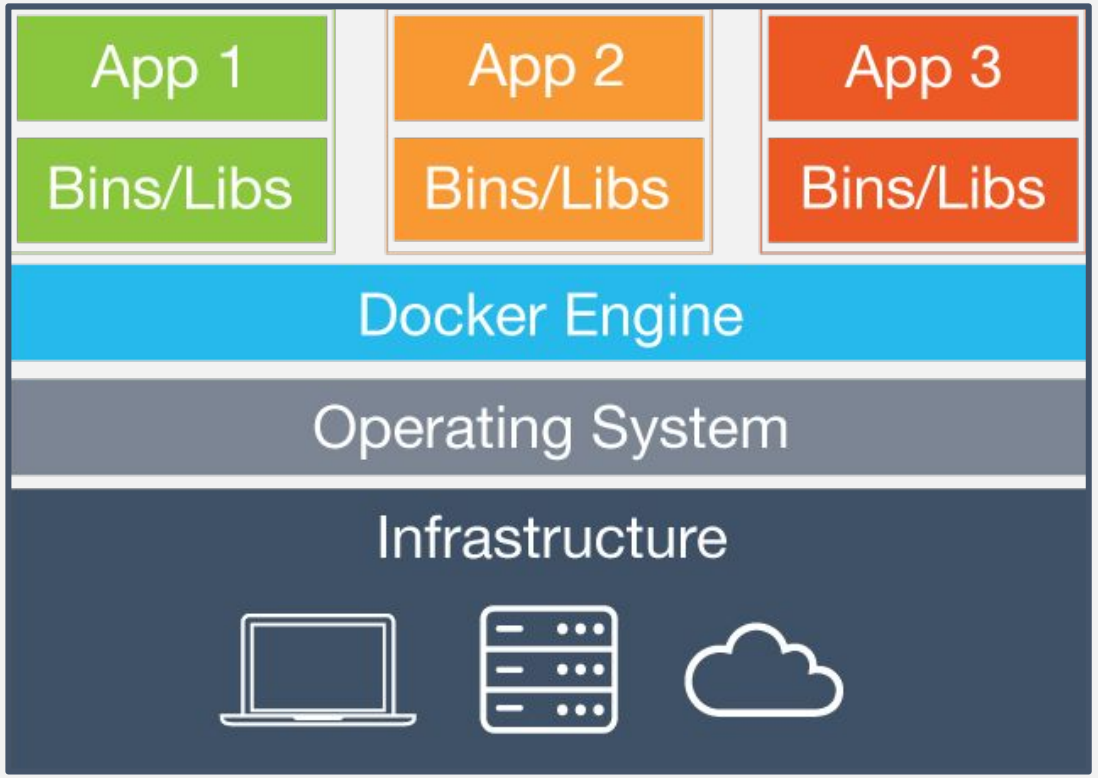
Reproducibility in Science



M. Baker (2016)

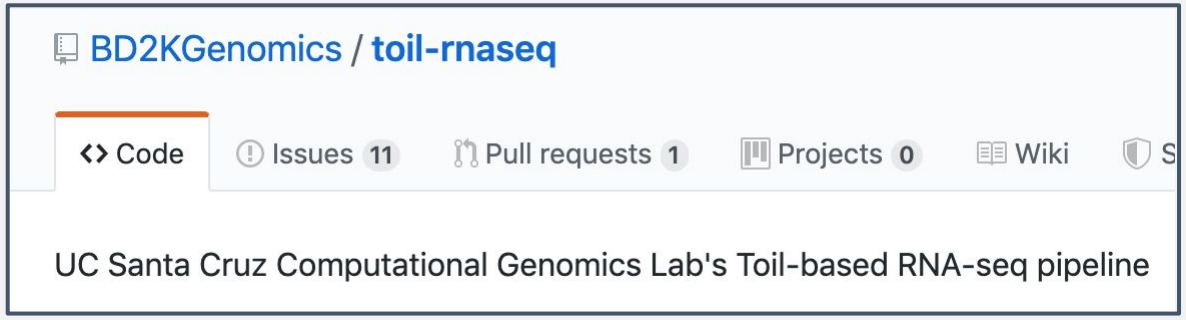
Docker for reproducible environments

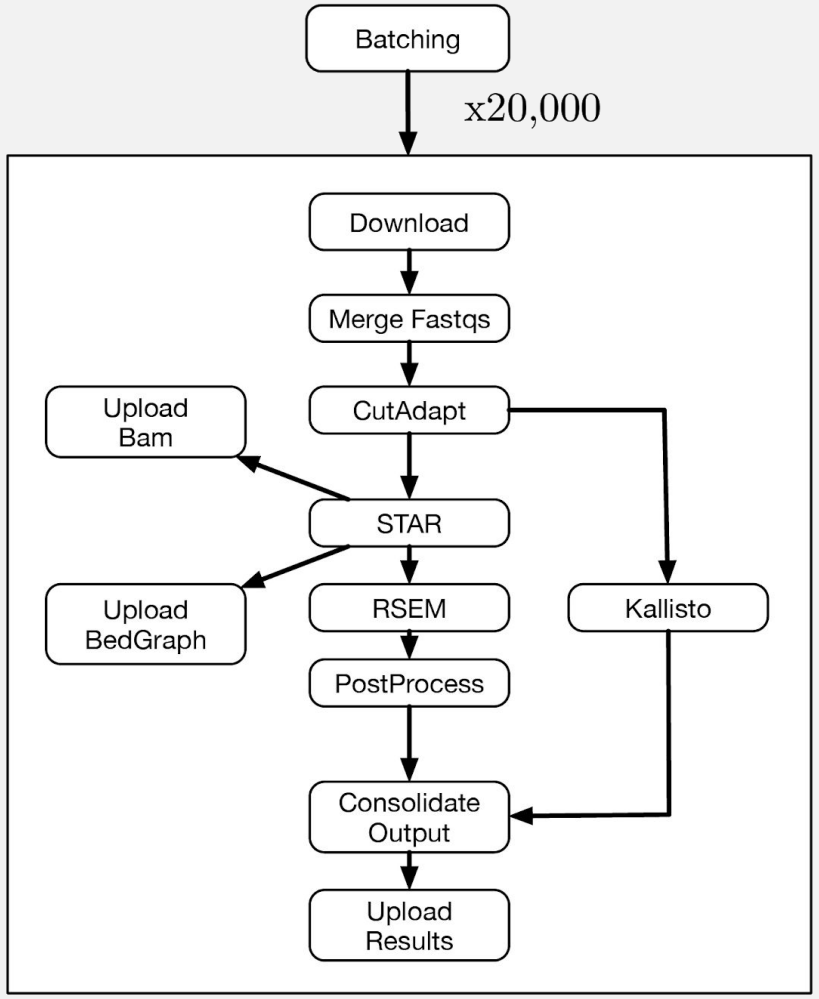




Implementing Toil: RNA-seq workflow

toil-rnaseq
workflow used to process 20,000 samples



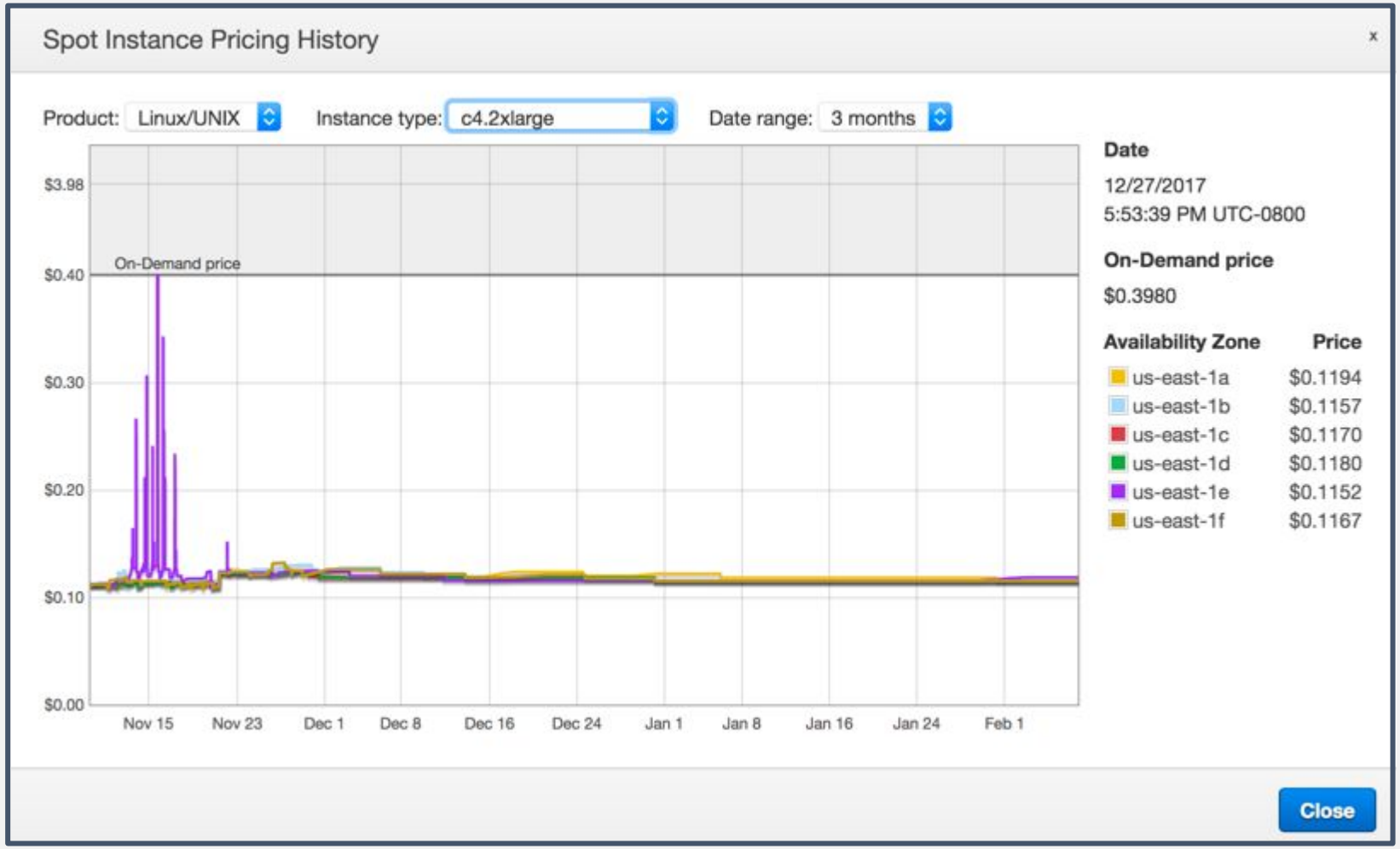


Vivian et al. (2017)

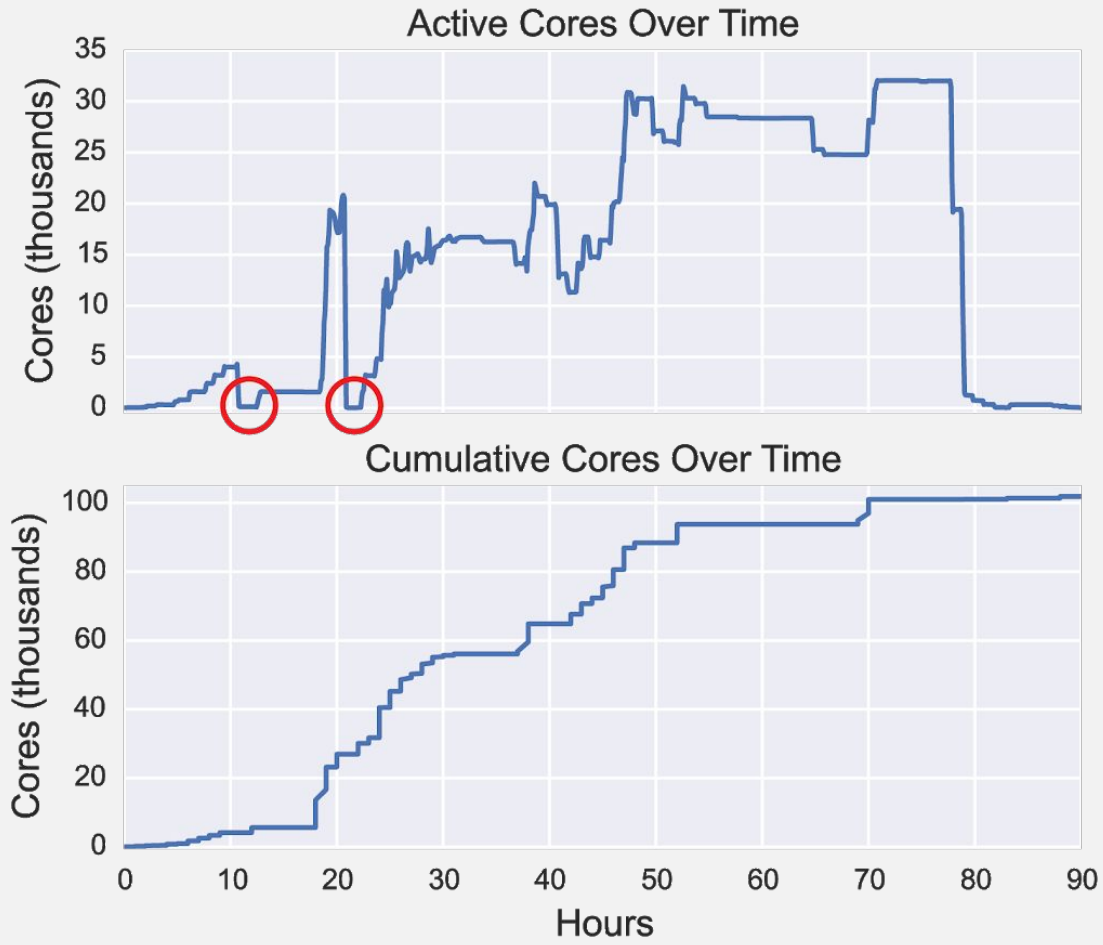
Cost-efficient cloud compute with spot-bidding

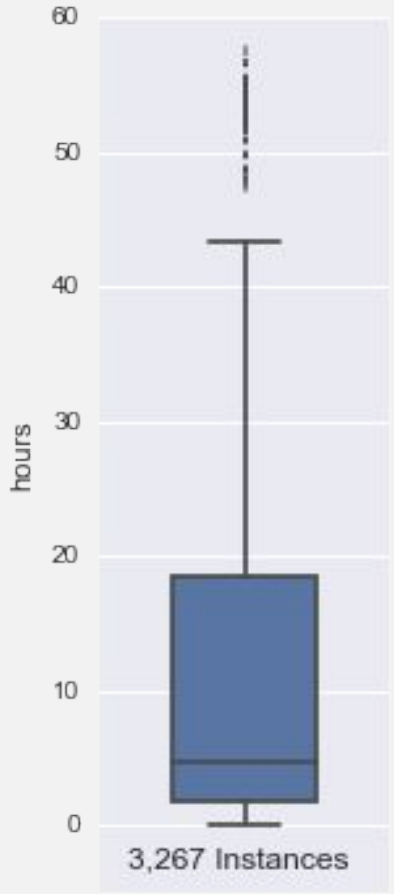
Spot bidding

Paying up to 90% less for spare compute resources, with the caveat that the provisioned machine can be terminated at any time.



Toil demonstration: 20,000 RNA-seq samples





LOG

Slaves

Activated	1,000
Deactivated	0

Tasks

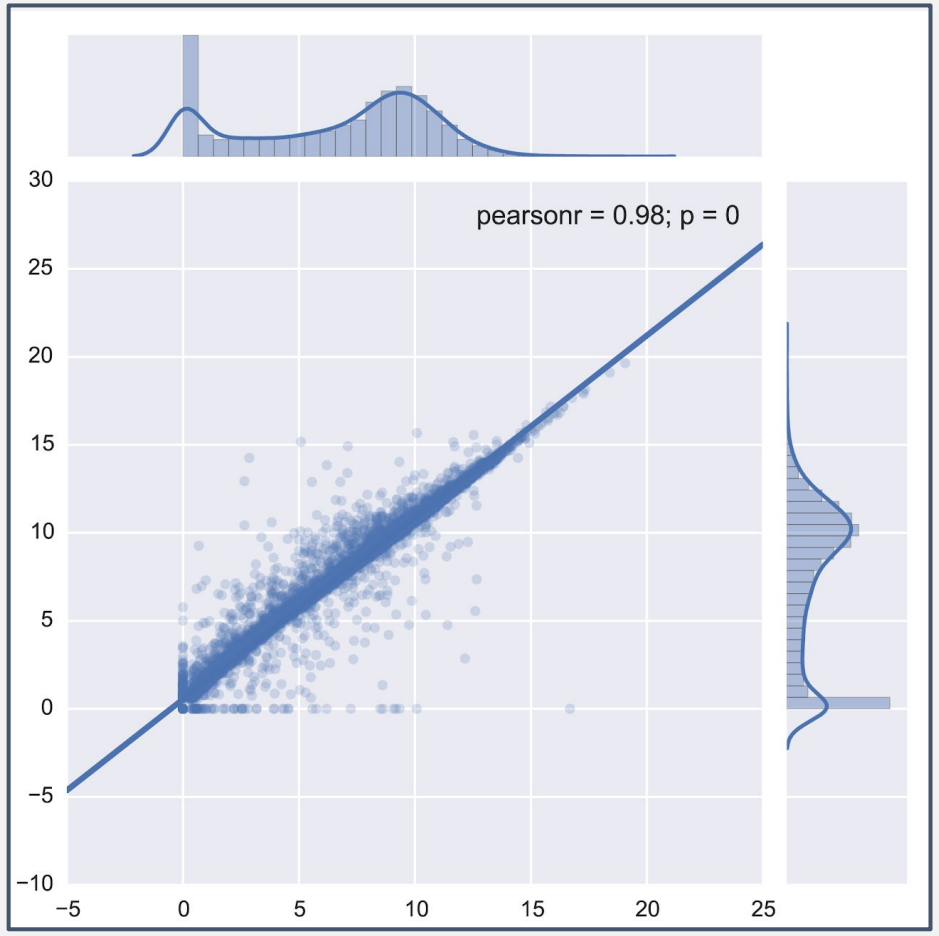
Staged	0
Started	0
Finished	101,146
Killed	0
Failed	36,254
Lost	2,312

Resources

	CPUs	Mem
Total	32000	72291.3 GB
Used	26725	25886.0 GB
Offered	0	0 B
Idle	5275	46405.3 GB

Vivian et al. (2017)

Recompute results

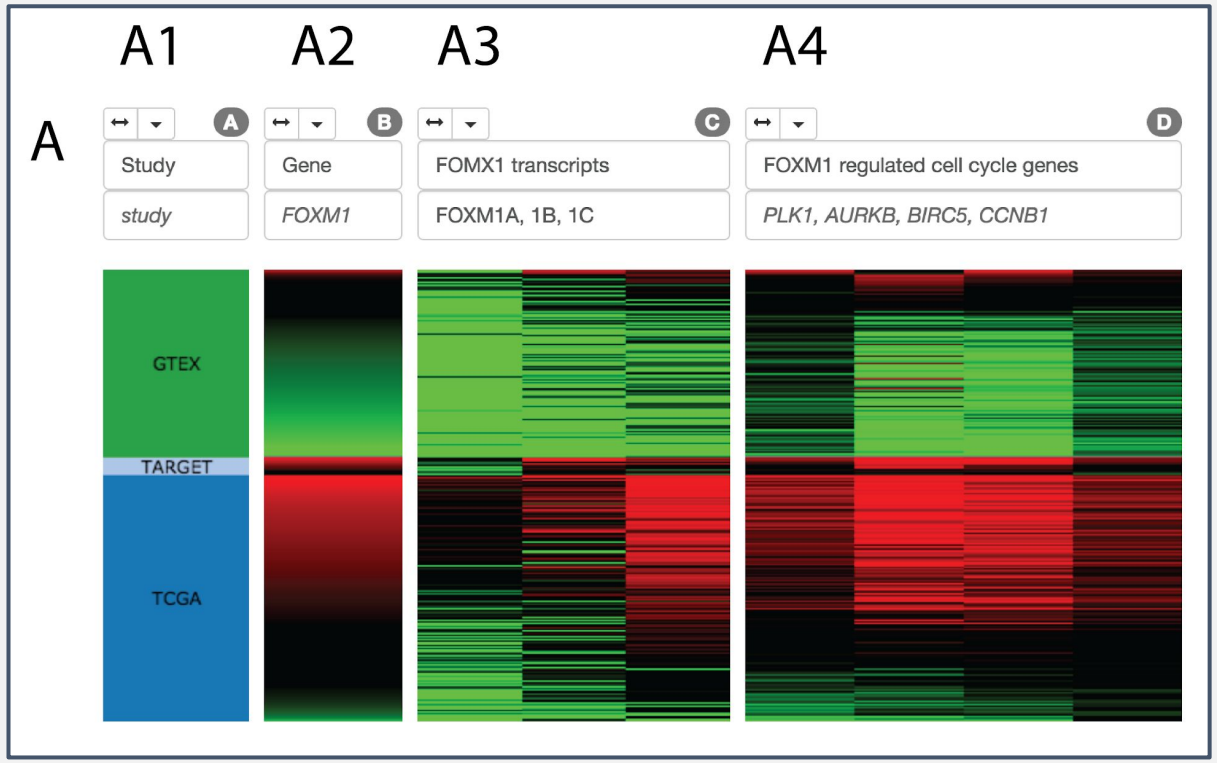


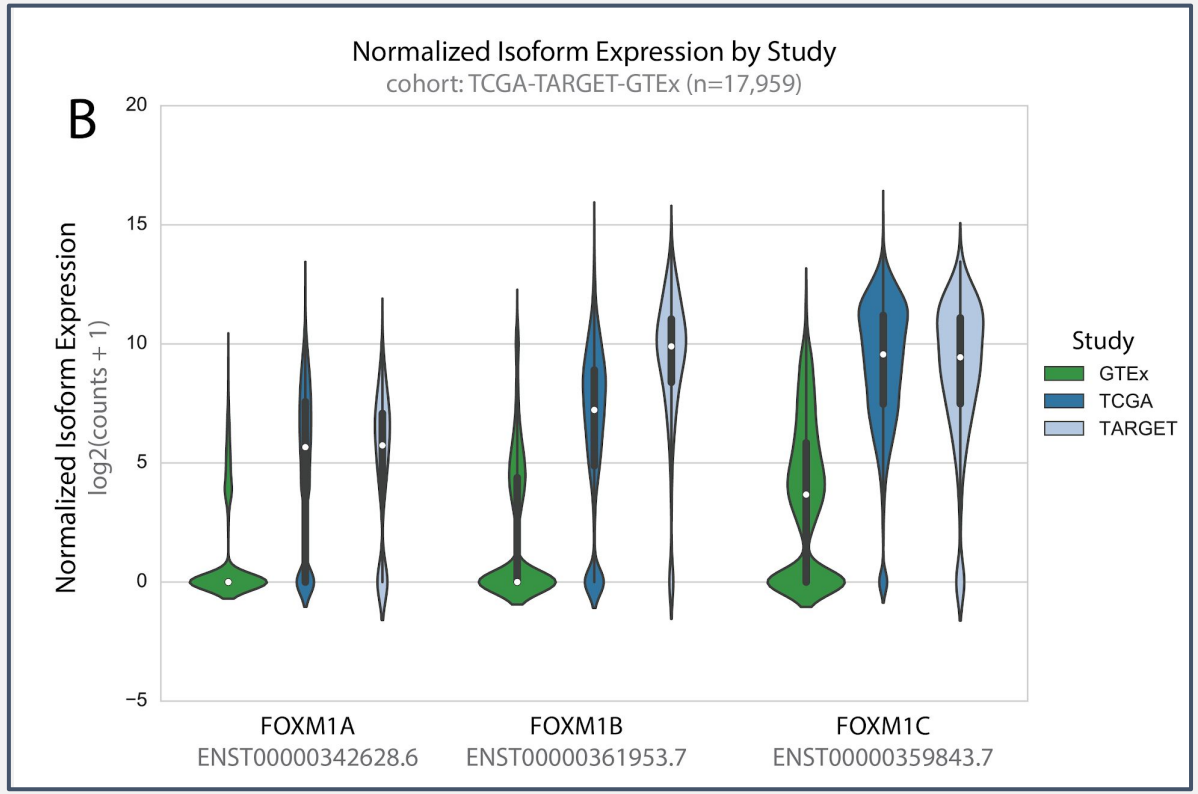
Concordance with TCGA — Despite using an entirely different reference genome and annotation set than TCGA, gene expression results are highly concordant

Cost — $26,071 for the entire run, or $1.30 per sample. If we had only run Kallisto, we estimate the cost would have been $0.19 per sample.

Vivian et al. (2017)

Publicly hosted on UCSC Xena Browser





Vivian et al. (2017)

Toil in the wild

vg

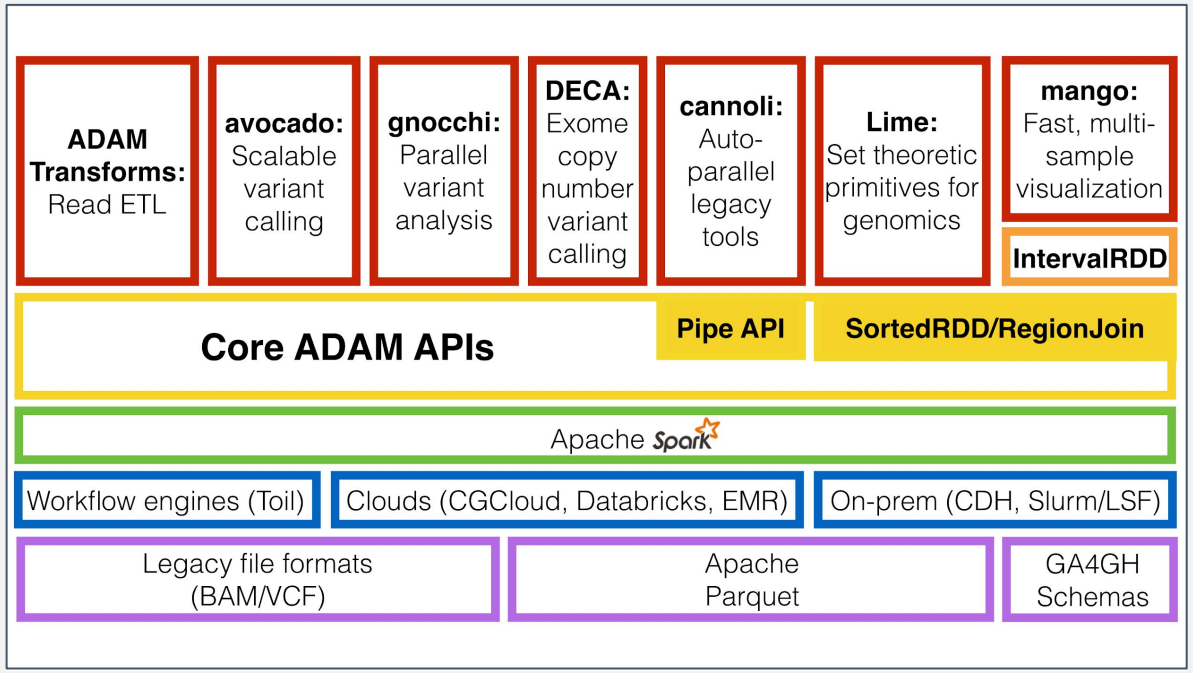
BROAD INSTITUTE

Treehouse CHILDHOOD CANCER INITIATIVE

ComparativeGenomicsToolkit / cactus

<> Code | Issues 39 | Pull requests 4 | Projects 0

Official home of genome aligner based upon notion of Cactus graphs



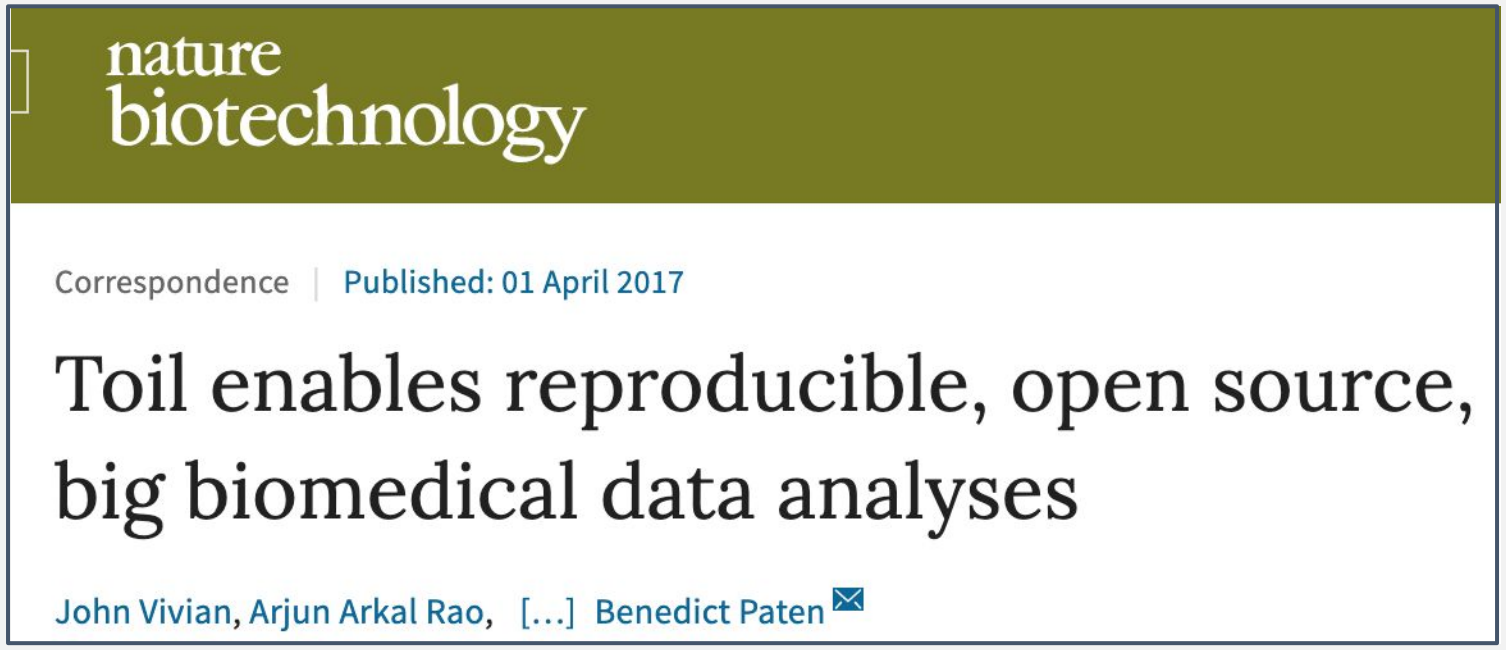
UNIVERSITY OF NORTH CAROLINA CHAPEL HILL — LVX LIBERTAS

Eruditio et Religio

color

Institut Pasteur

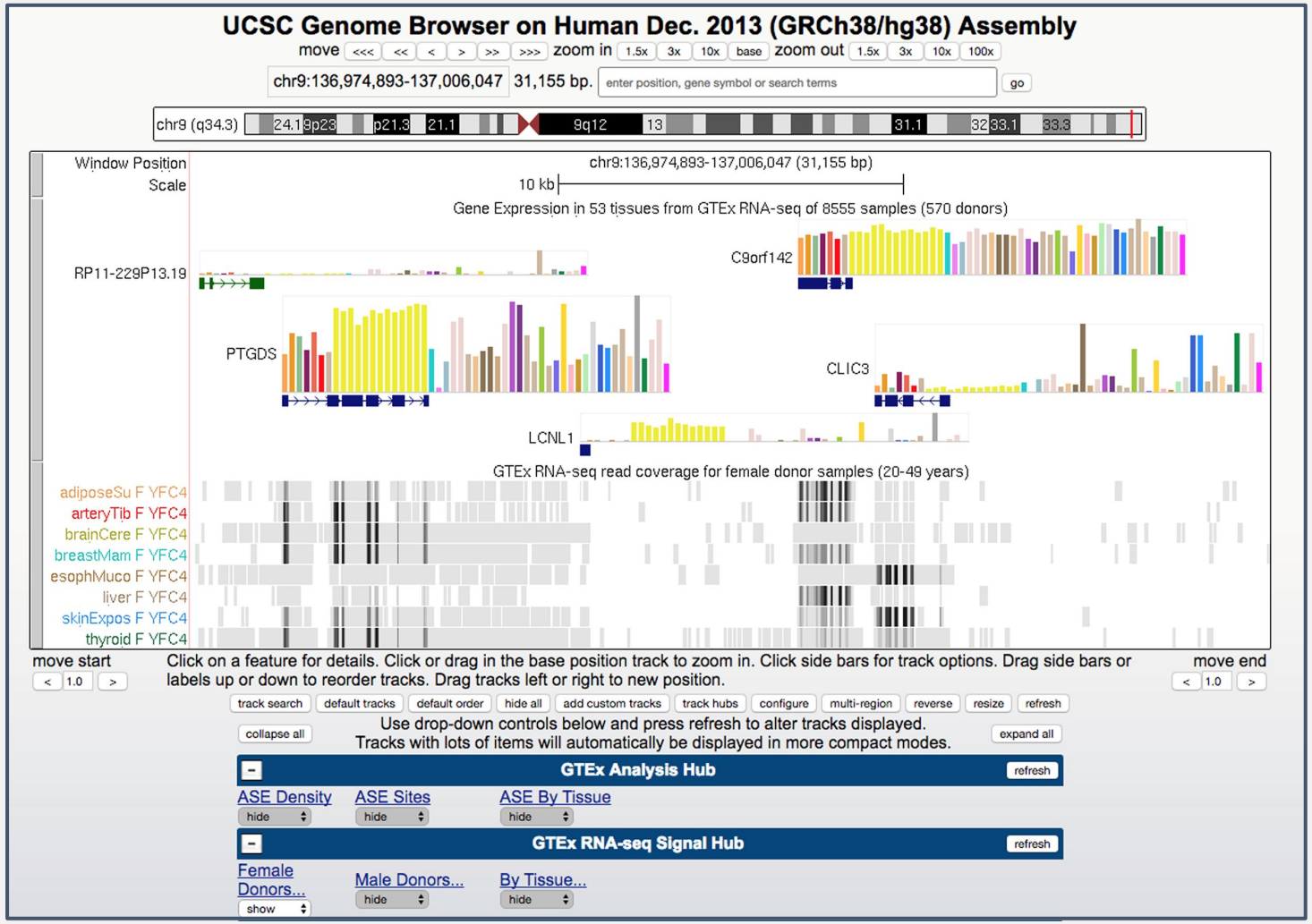
Summary of Toil recompute and collaborations



Summary of Toil recompute and collaborations

GTEx gene expression track for ~8,000 tissue samples obtained from 570 adult postmortem individuals

First track in the browser for gene expression in human tissues since the GNF Atlas microarray tracks from 2004



Tyner et al. (2017)

Summary of Toil recompute and collaborations

Abstract 2466: Identifying confidently measured genes in single pediatric cancer patient samples using RNA sequencing

Holly Beale, Du Linh Lam, John Vivian, Yulia Newton, Avanthi Tayi Shah, Isabel Bjork, Ted Goldstein, Angela N. Brooks, Josh Stuart, Sofie Salama, E. Alejandro Sweet-Cordero, David Haussler1, and Olena Morozova

DOI: 10.1158/1538-7445.AM2017-2466 Published July 2017

Abstract 4890: A pan-cancer analysis framework for incorporating gene expression information into clinical interpretation of pediatric cancer genomic data

Olena Morozova, Yulia Newton, Avanthi Tayi Shah, Holly Beale, Du Linh Lam, John Vivian, Isabel Bjork, Theodore Goldstein, Josh Stuart, Sofie Salama, E. Alejandro Sweet-Cordero, and David Haussler

Treehouse CHILDHOOD CANCER INITIATIVE

AACR American Association for Cancer Research

Beale et al. (2017) and Morozova-Vaske et al. (2017)

Treehouse Childhood Cancer Initiative

Treehouse
CHILDHOOD CANCER INITIATIVE

Motivating question

Can we create a robust statistical framework for identifying gene expression outliers for single patients?

03

Outlier Detection

Motivating example

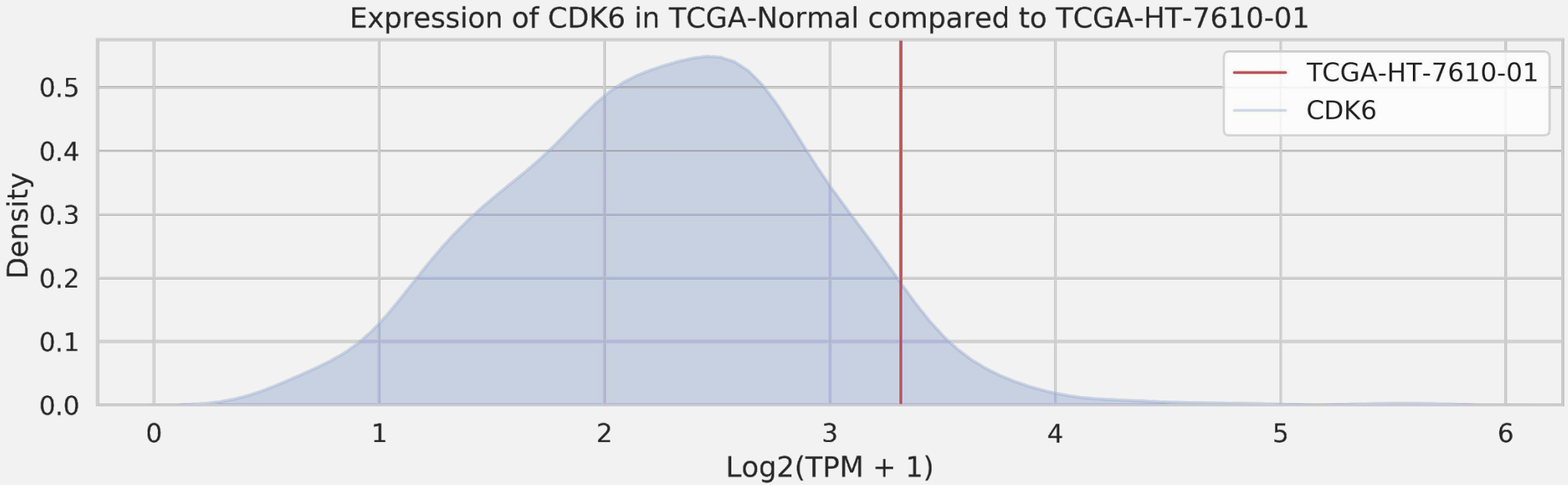
Sample

TCGA-HT-7610-01

Gene

CDK6

Expression of CDK6



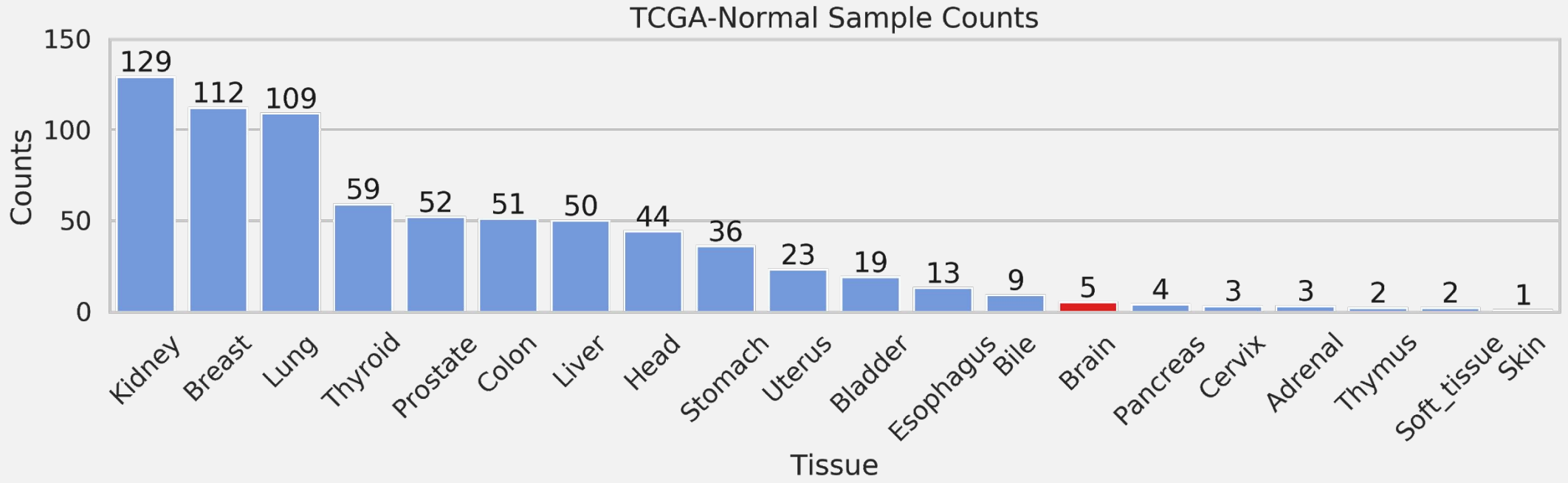
Cancer subtype for sample

Sample
TCGA-HT-7610-01

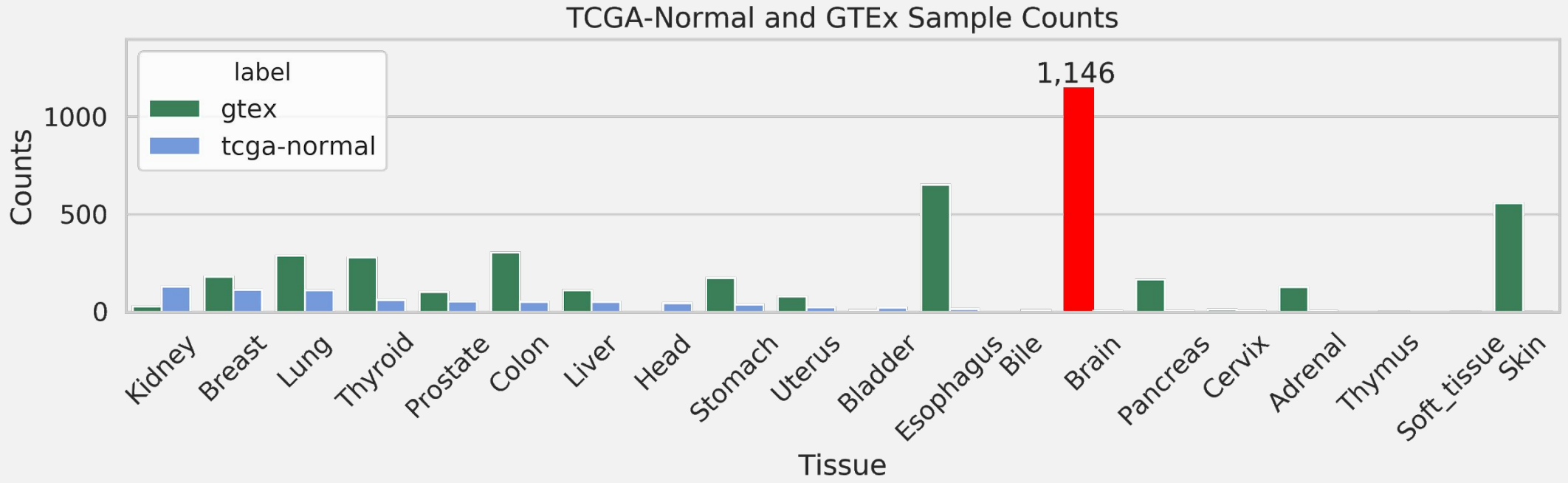
Gene
CDK6

Cancer subtype
Lower grade glioma

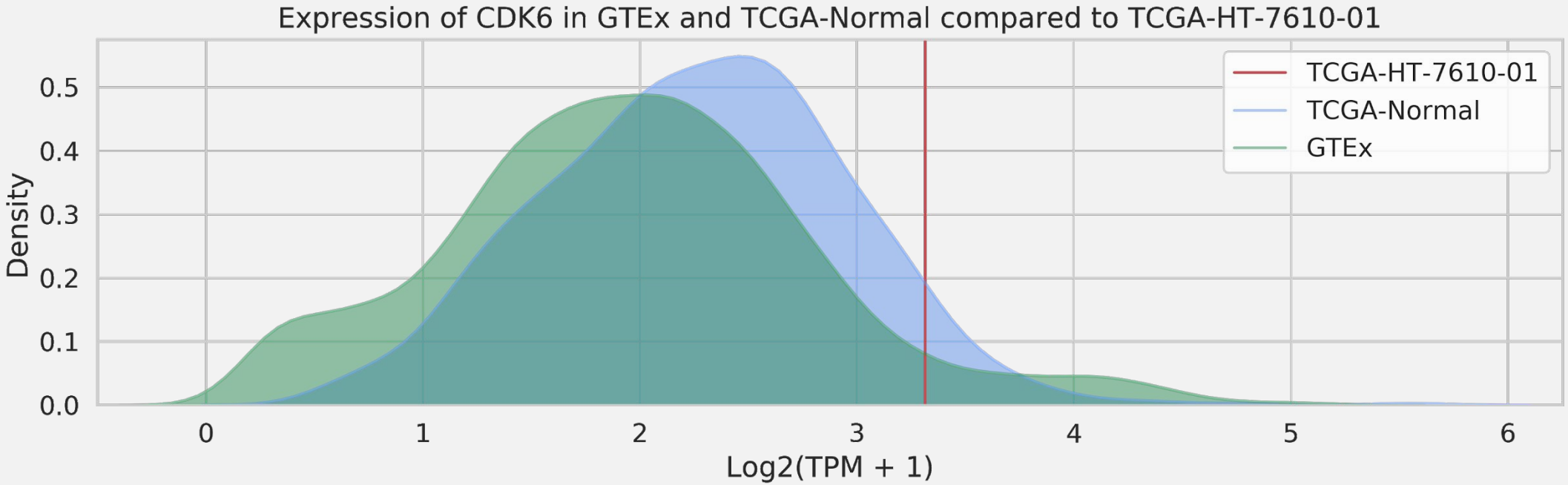
Sample counts for normals in TCGA



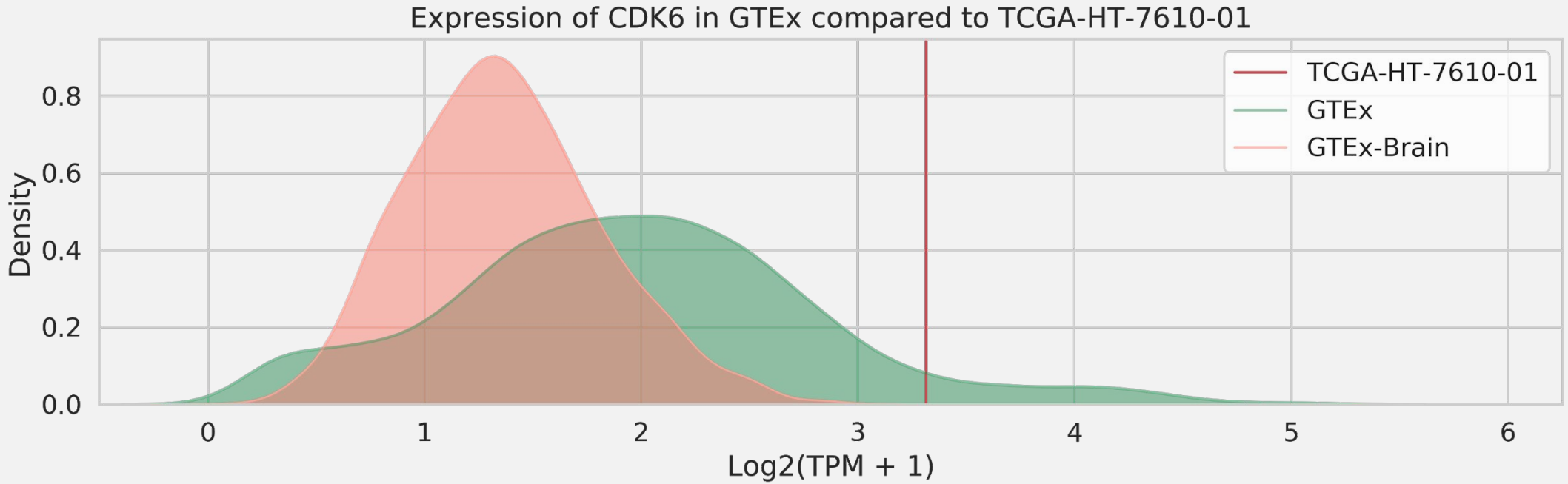
Sample counts for normals in GTEx



Expression of CDK6



Expression of CDK6



Existing methods

- **Differential expression** tools operate best under experimental conditions where both groups consist of several technical replicates, or lacking that, biological replicates

- **IQR outlier** methods ignore sample size contribution when used against a combined dataset and produce binary output which can miss potential leads and are difficult to rank

"Experiments without replicates do not allow for estimation of the dispersion of counts around the expected value for each group, which is critical for differential expression analysis. ... We provide this approach for data exploration only, but for **accurately identifying differential expression, biological replicates are required**."

Michael Love (DESeq2)

"As the senior author of the edgeR project, I can tell you 100% that edgeR was **always intended to be used with any design that included any degree of replication.**"

Gordon Smyth

Existing methods

- **Differential expression** tools operate best under experimental conditions where both groups consist of several technical replicates, or lacking that, biological replicates

- **IQR outlier** methods ignore sample size contribution when used against a combined dataset and produce binary output which can miss potential leads and are difficult to rank

"Experiments without replicates do not allow for estimation of the dispersion of counts around the expected value for each group, which is critical for differential expression analysis. ... We provide this approach for data exploration only, but for **accurately identifying differential expression, biological replicates are required**."

Michael Love (DESeq2)

"As the senior author of the edgeR project, I can tell you 100% that edgeR was **always intended to be used with any design that included any degree of replication.**"

Gordon Smyth

Existing methods

- **Differential expression** tools operate best under experimental conditions where both groups consist of several technical replicates, or lacking that, biological replicates

- **IQR outlier** methods ignore sample size contribution when used against a combined dataset and produce binary output which can miss potential leads and are difficult to rank

"Experiments without replicates do not allow for estimation of the dispersion of counts around the expected value for each group, which is critical for differential expression analysis. ... We provide this approach for data exploration only, but for **accurately identifying differential expression, biological replicates are required**."

Michael Love (DESeq2)

"As the senior author of the edgeR project, I can tell you 100% that edgeR was **always intended to be used with any design that included any degree of replication.**"

Gordon Smyth

Existing methods

- **Differential expression** tools operate best under experimental conditions where both groups consist of several technical replicates, or lacking that, biological replicates

- **IQR outlier** methods ignore sample size contribution when used against a combined dataset and produce binary output which can miss potential leads and are difficult to rank

"Experiments without replicates do not allow for estimation of the dispersion of counts around the expected value for each group, which is critical for differential expression analysis. ... We provide this approach for data exploration only, but for **accurately identifying differential expression, biological replicates are required**."

Michael Love (DESeq2)

"As the senior author of the edgeR project, I can tell you 100% that edgeR was **always intended to be used with any design that included any degree of replication.**"

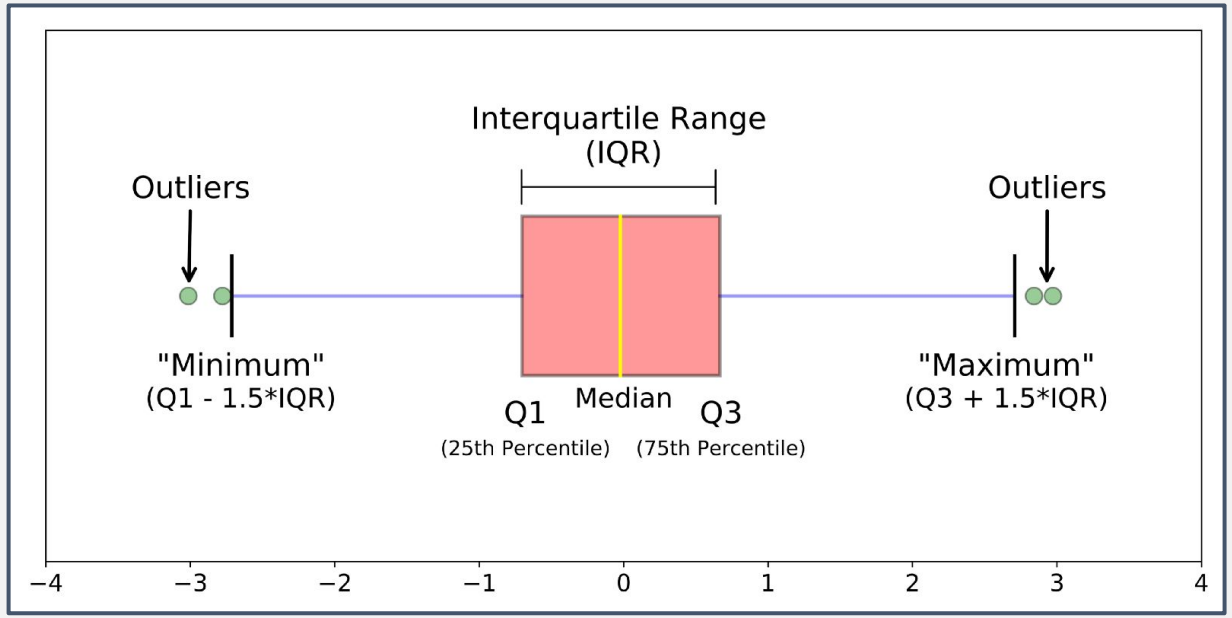
Gordon Smyth

IQR Cutoff Method

Tukey's Outlier Definition

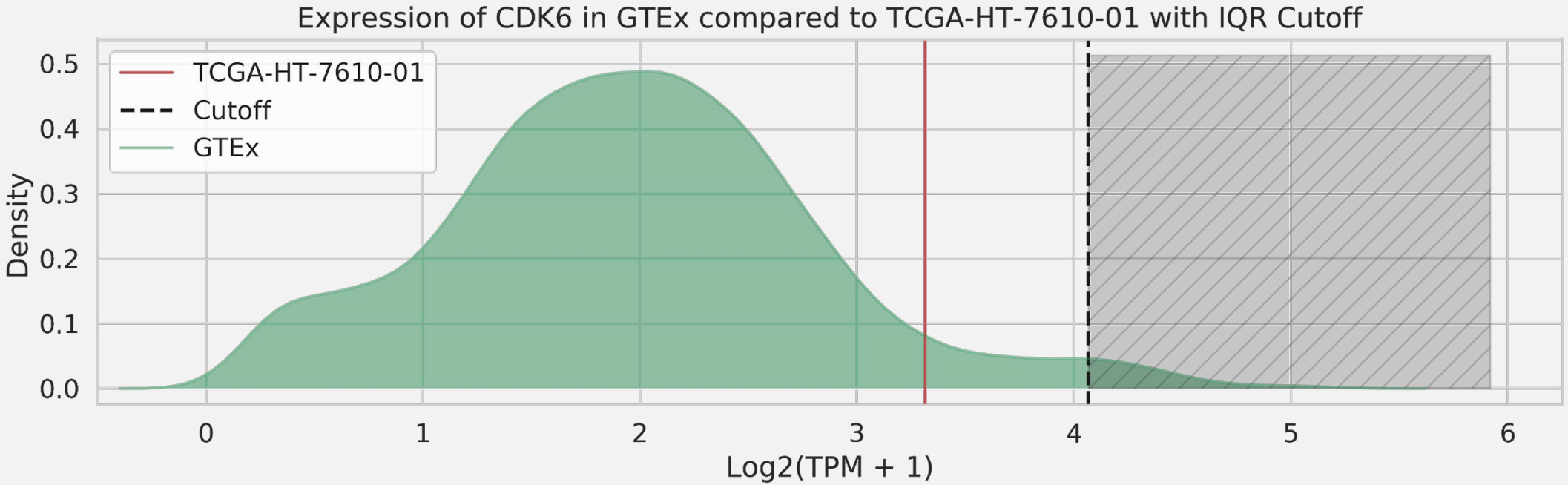
Lower bound: Q1 - (IQR * 1.5)

Upper bound: Q3 + (IQR * 1.5)

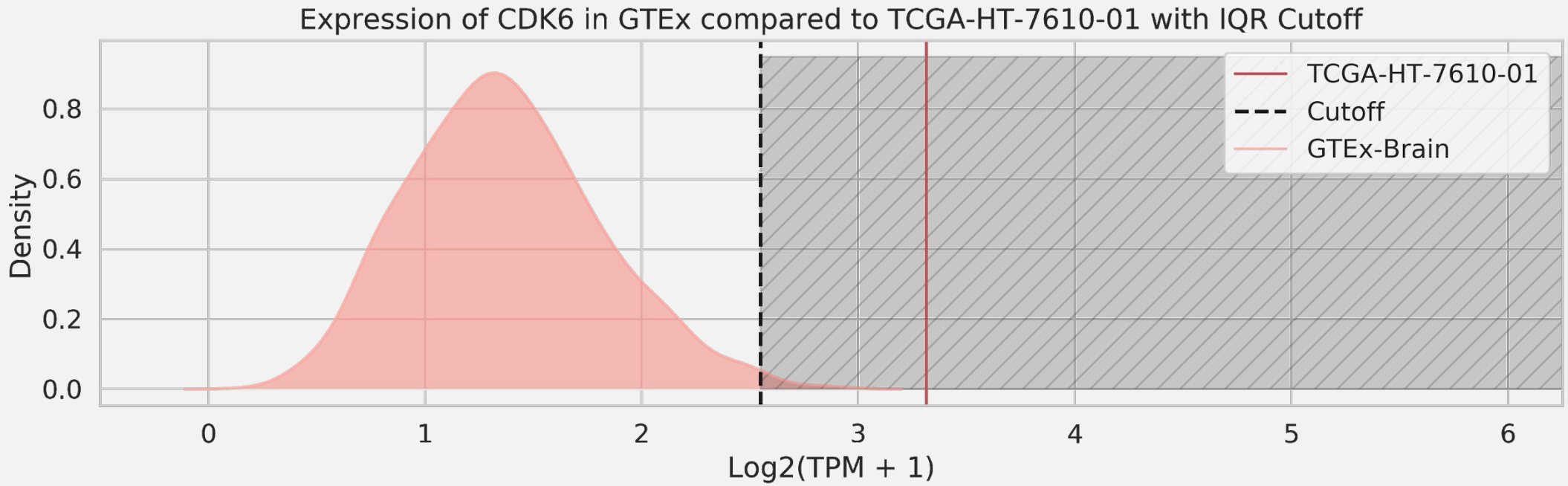


Tukey (1977)

Pan-normal IQR Cutoff approach



Matched Normal Tissue Cutoff



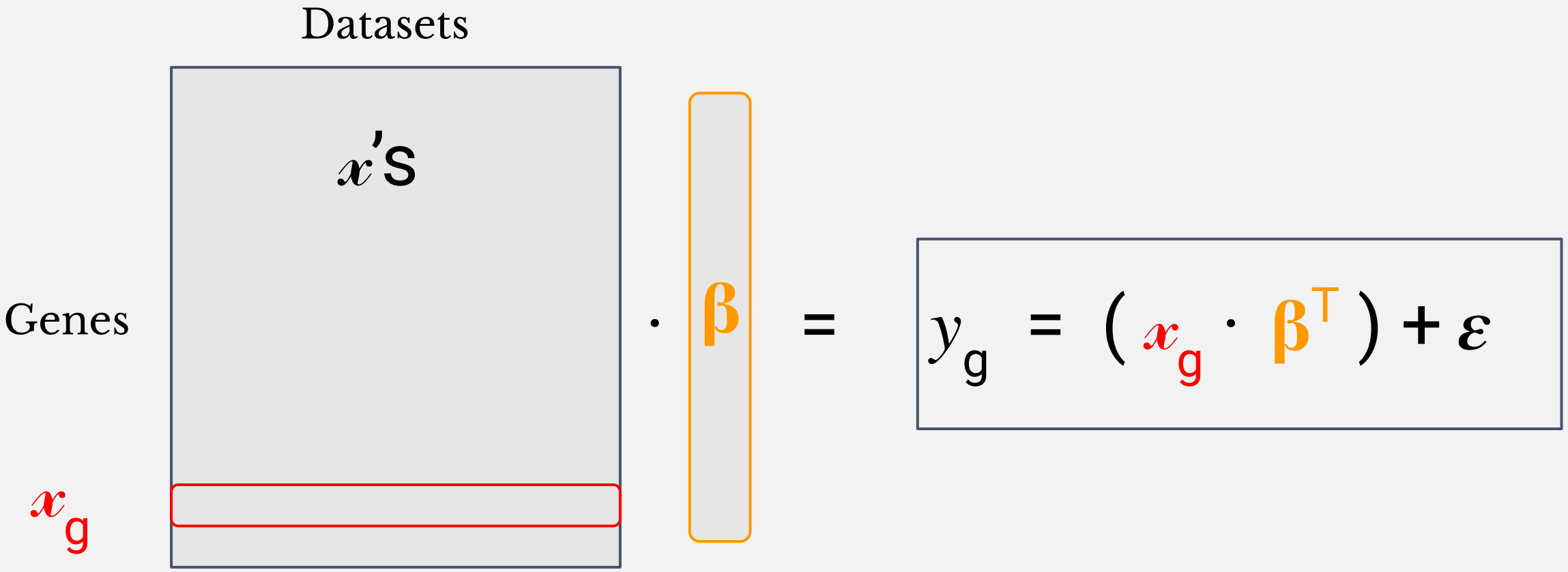
Modeling gene expression from background sets

Assumption: gene expression for a sample can be modeled as a convex mixture of background datasets

$$\text{Gene}_0 = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

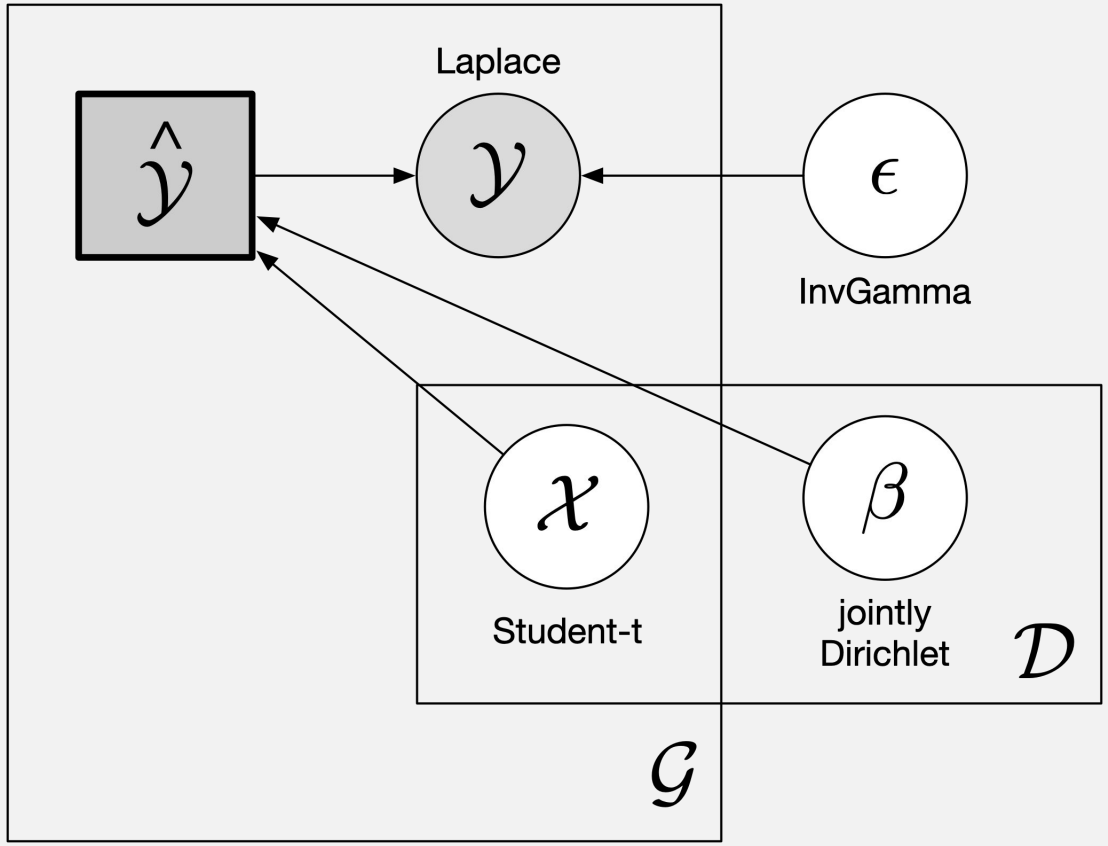
Where x is a random variable representing gene expression for a particular background dataset, e.g. x_0 is *Brain* from GTEx, x_1 is *Pituitary* from GTEx, etc.

Modeling gene expression from background sets



Bayesian inference: plate notation

Plate notation is a method of representing random variables that repeat in a graphical model.



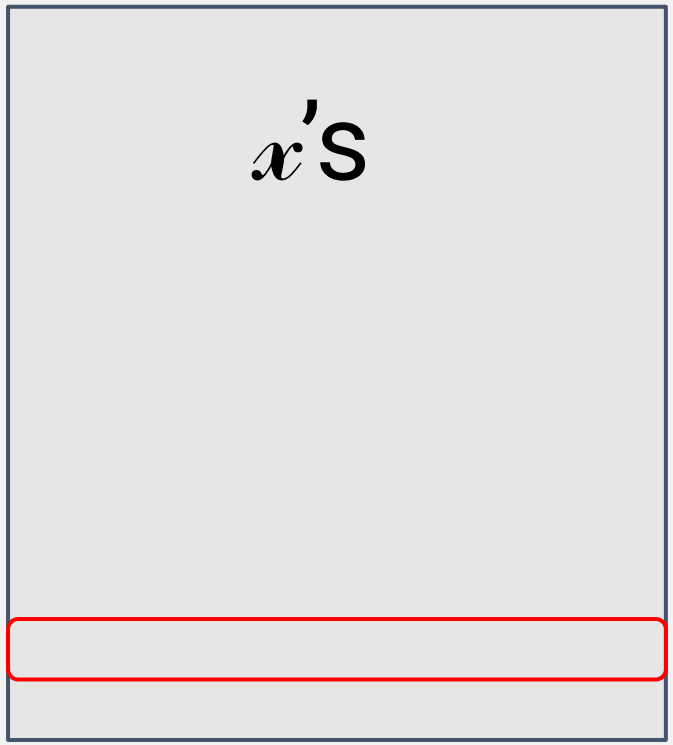
- G represents Genes
- D represents Datasets
- ϵ represents model error
- x represents expression for that gene / dataset
- $\hat{y}$ represents a convex model for one gene
- y is the posterior distribution

Normalizing for large variances

Datasets

Genes

x_g



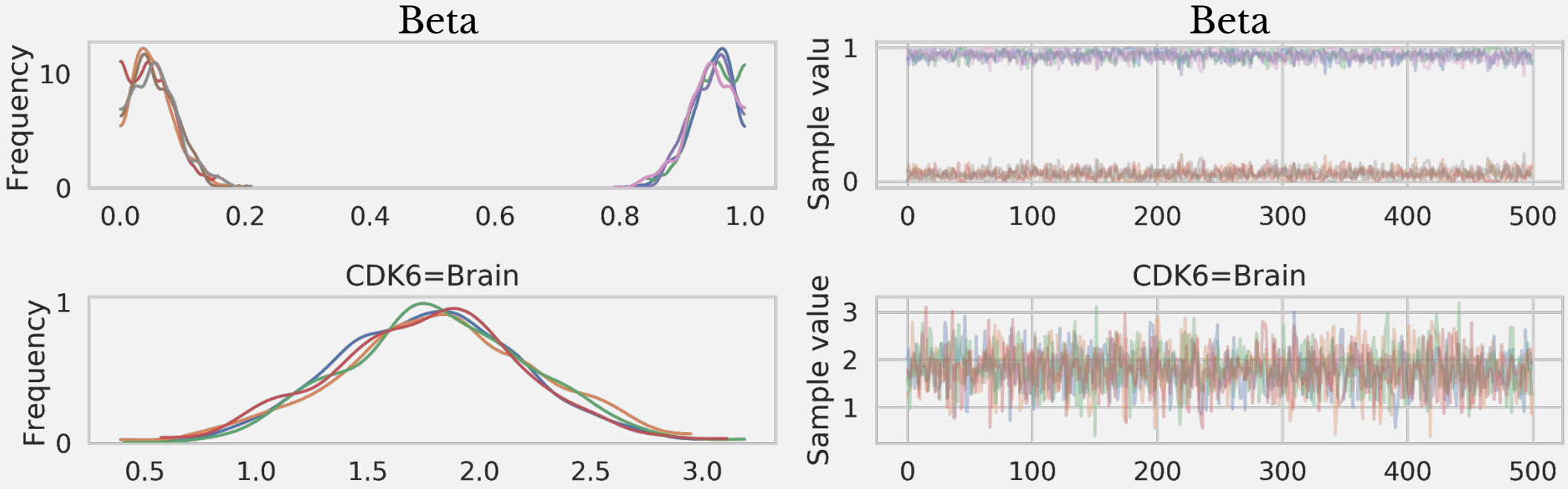
$\cdot$ β =

$$y_g = (x_g \cdot \beta^T) + \varepsilon$$

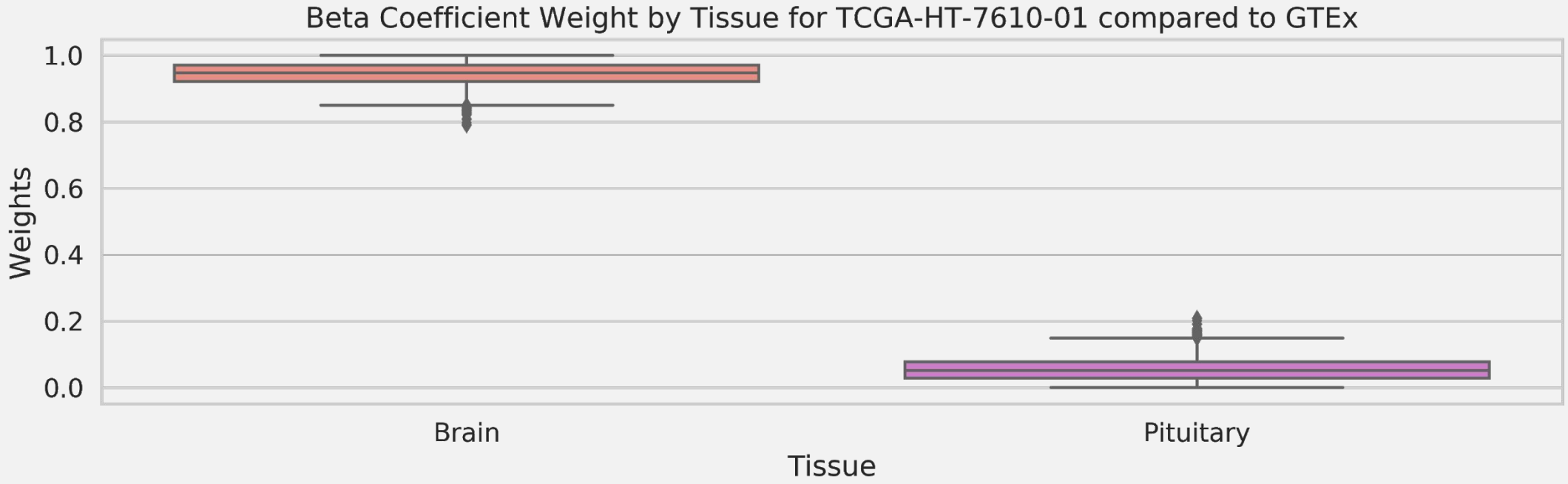
$$y_g = \frac{\sum_d \frac{x\beta^T}{\sigma} + \epsilon}{\sum_d \frac{\beta}{\sigma}}$$

How do we evaluate the model training?

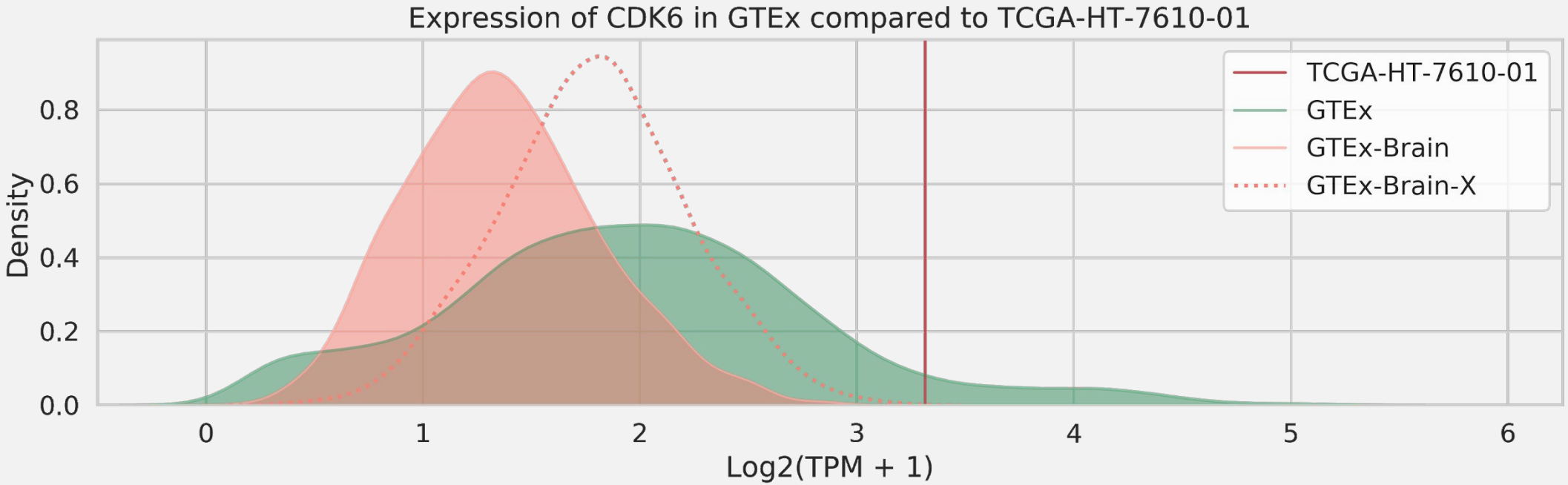
Trace plot — A tool for visualizing the convergence of a Markov chain is the trace plot: the plot of the values generated from the Markov chain versus the iteration number.



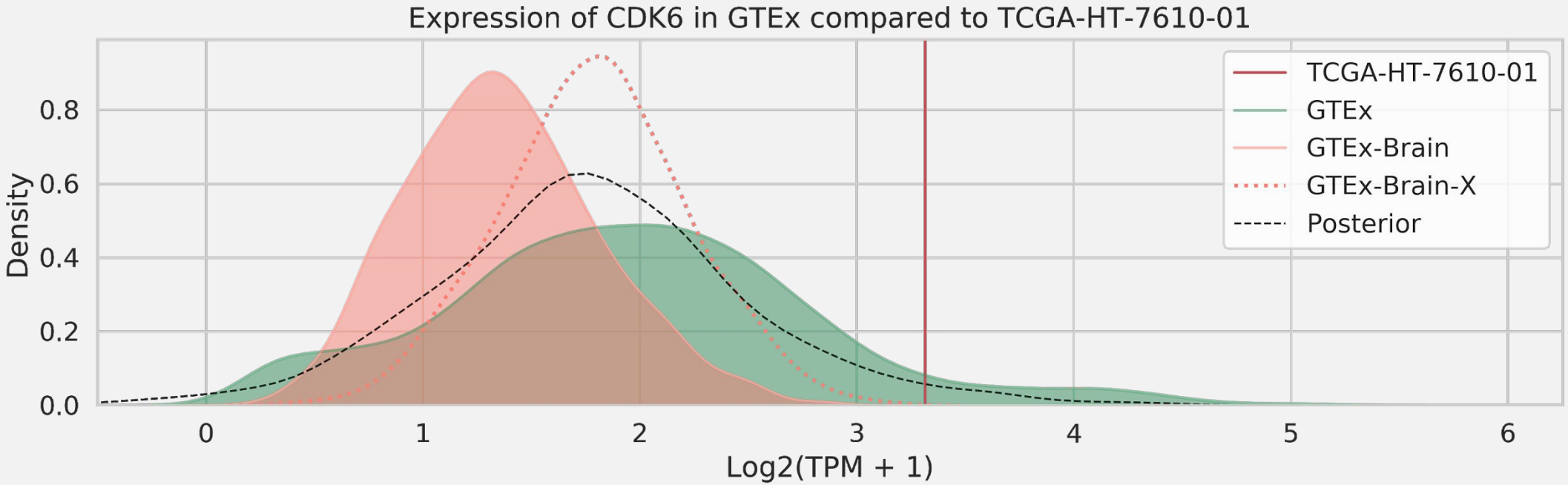
β weights of background datasets



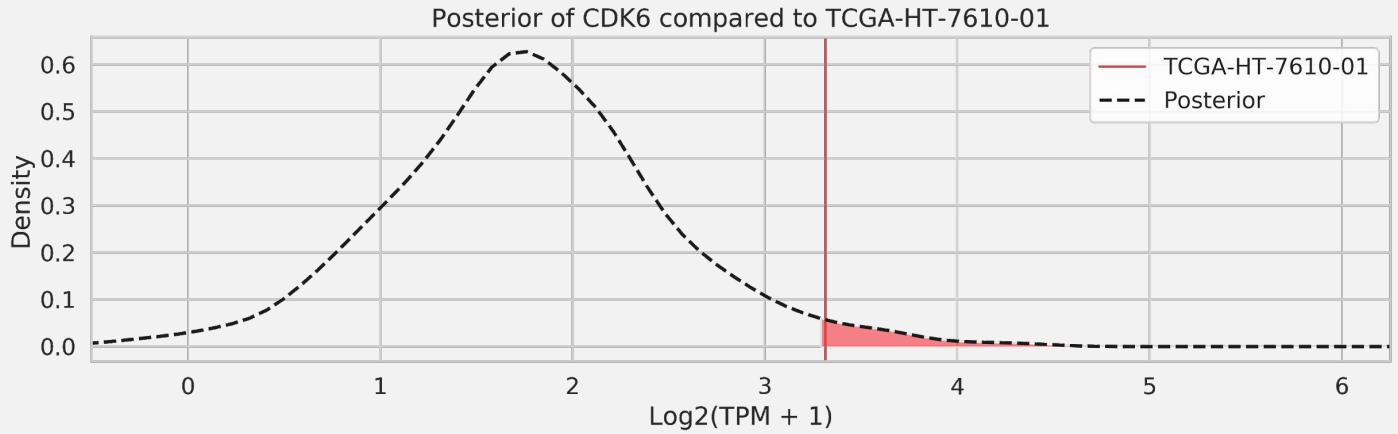
x distribution

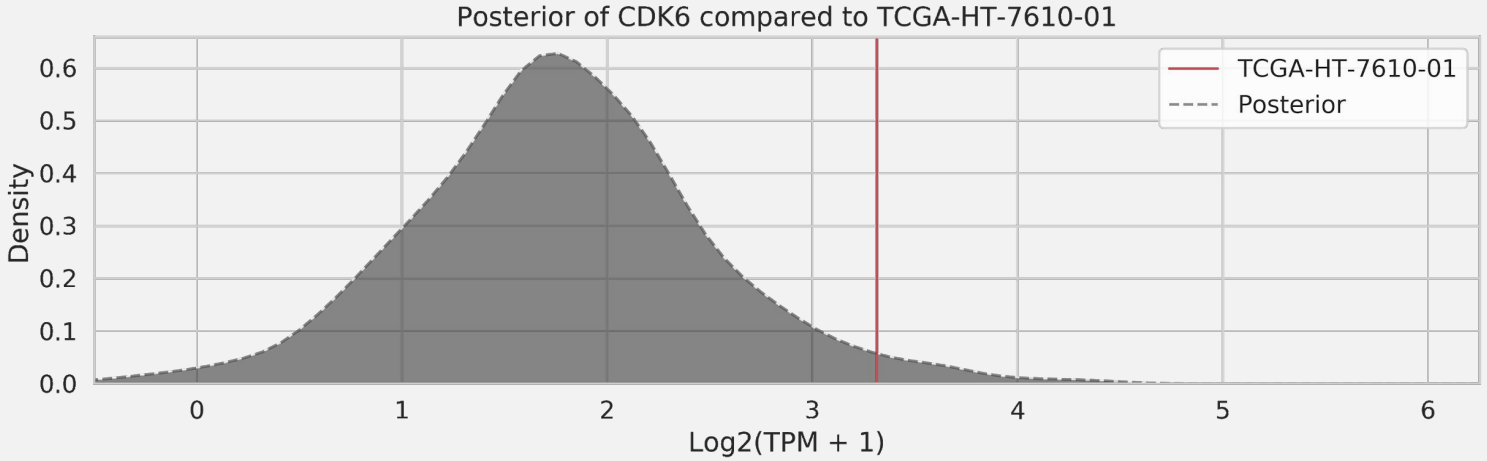


Finally, the posterior distribution



Posterior predictive p-values





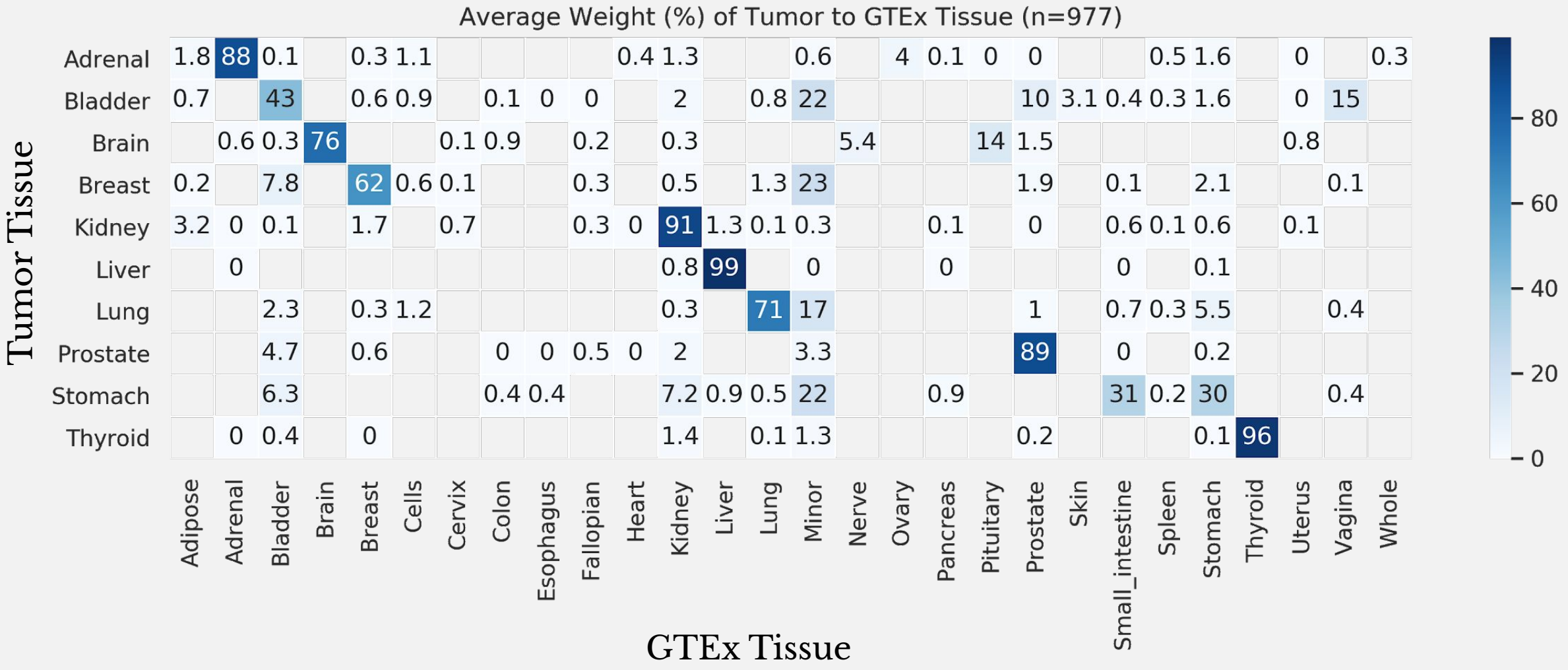
P-value of **CDK6** for **TCGA-HT-7610-01** when compared to **GTEx**

$$\frac{3{,}289}{100{,}000} = 0.033$$

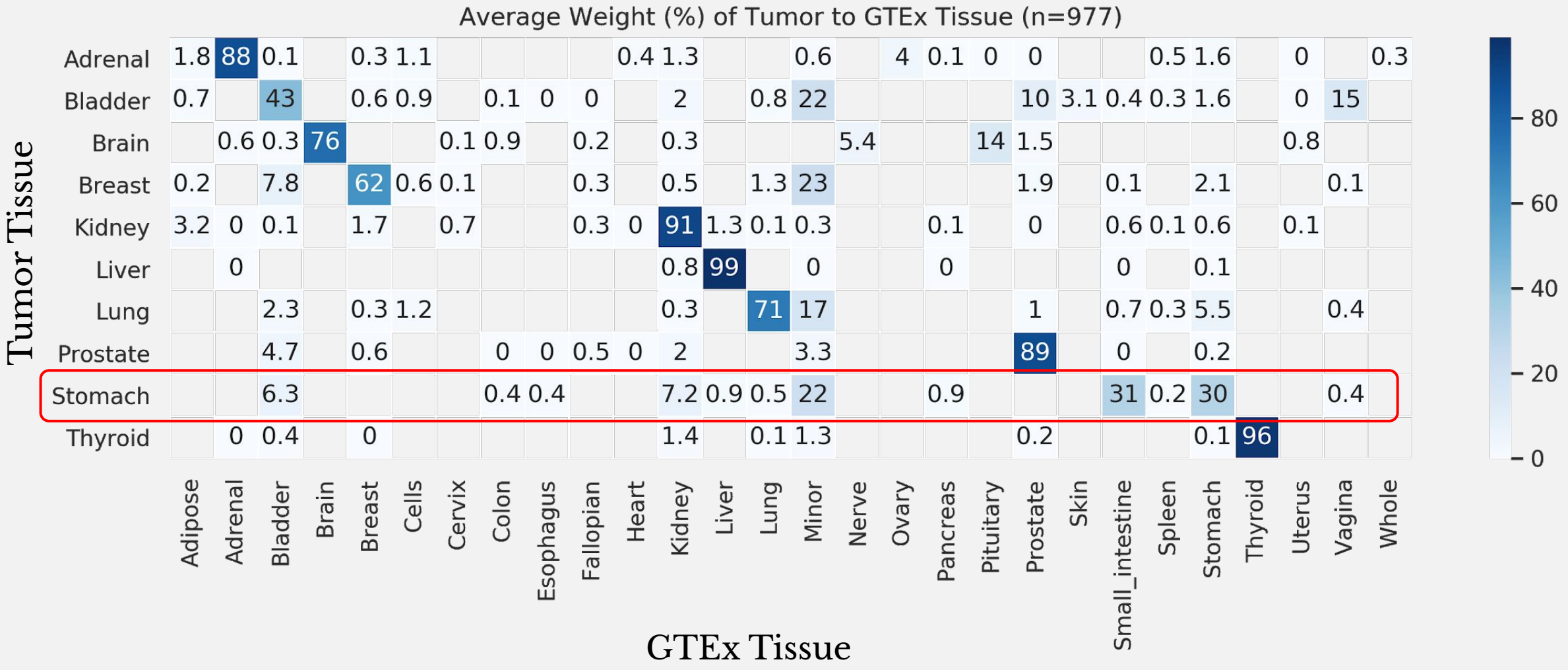
Summary: improvements over existing methods

- P-values over binary output
- Automatically selects appropriate background dataset
- Confidence in background selection by examining weight variance
- And more! (if we have time)

Model weight across tumor subtypes



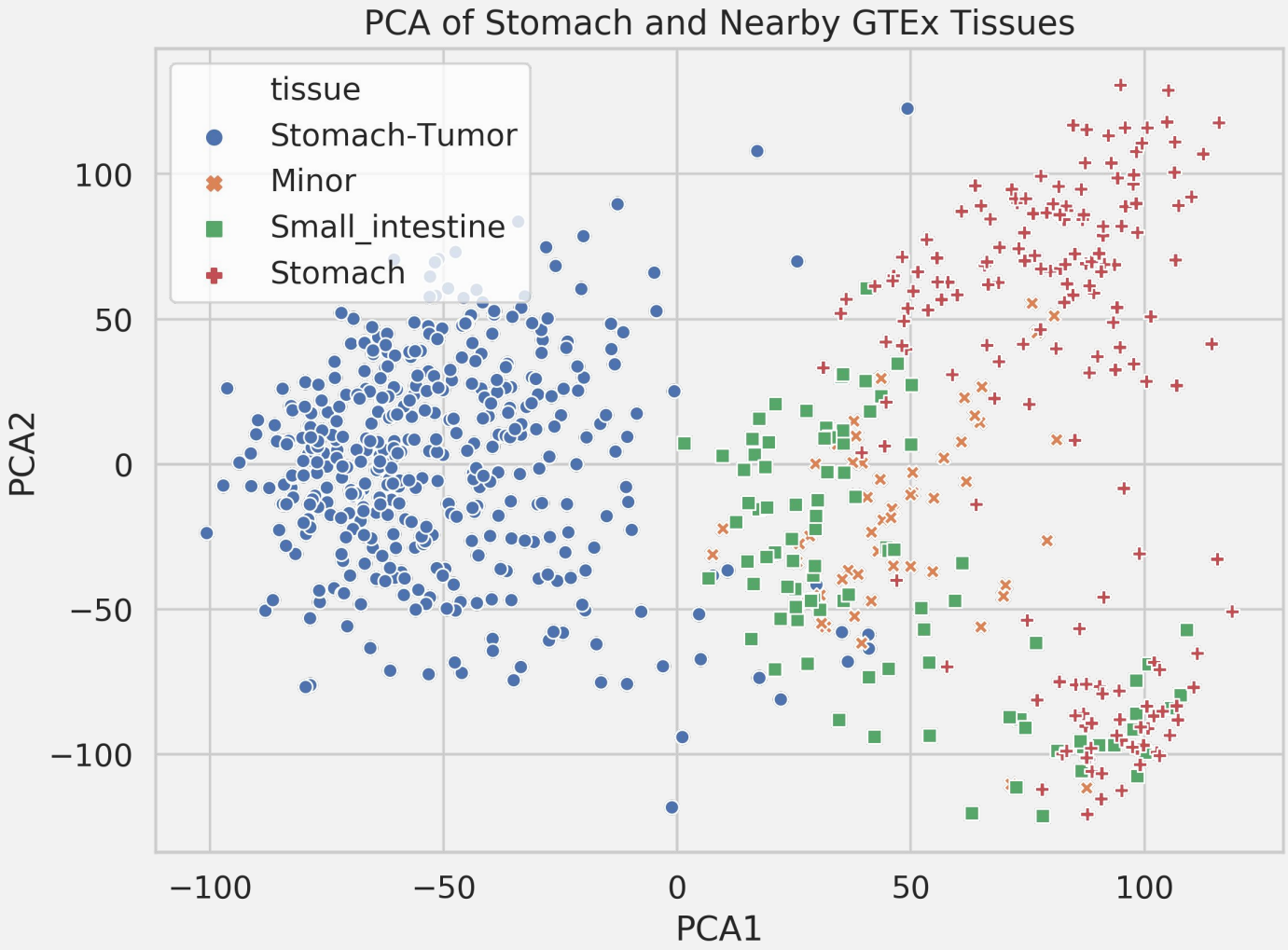
Model weight across tumor subtypes



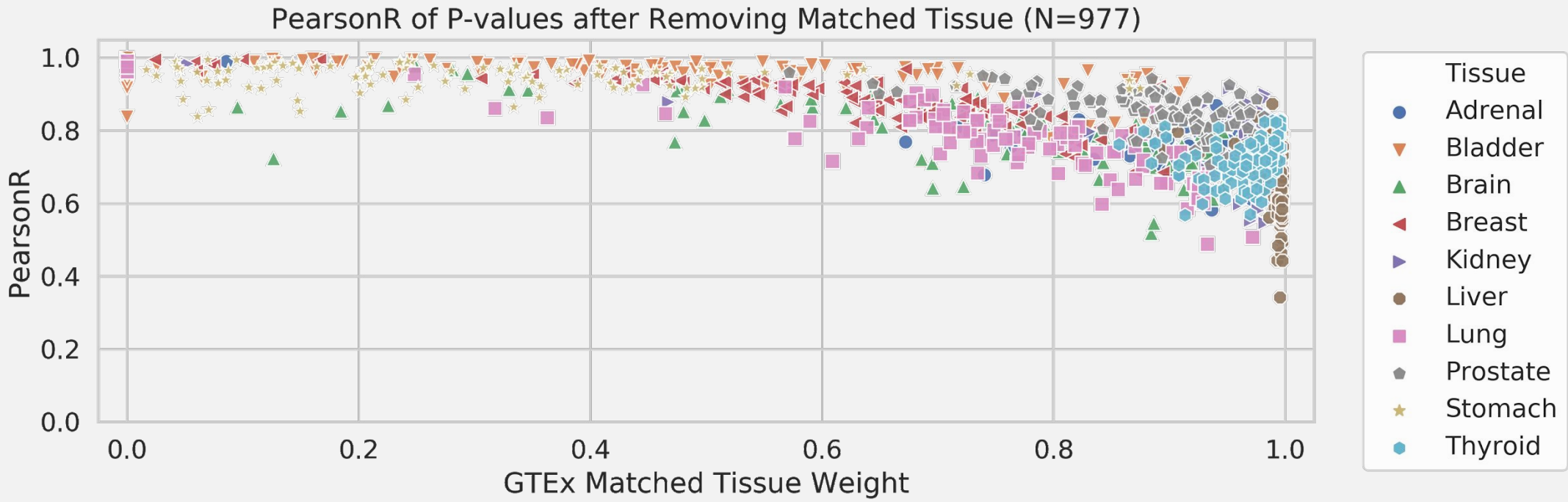
Dimensionality reduction of low-weight sample

Stomach Weight Assignment

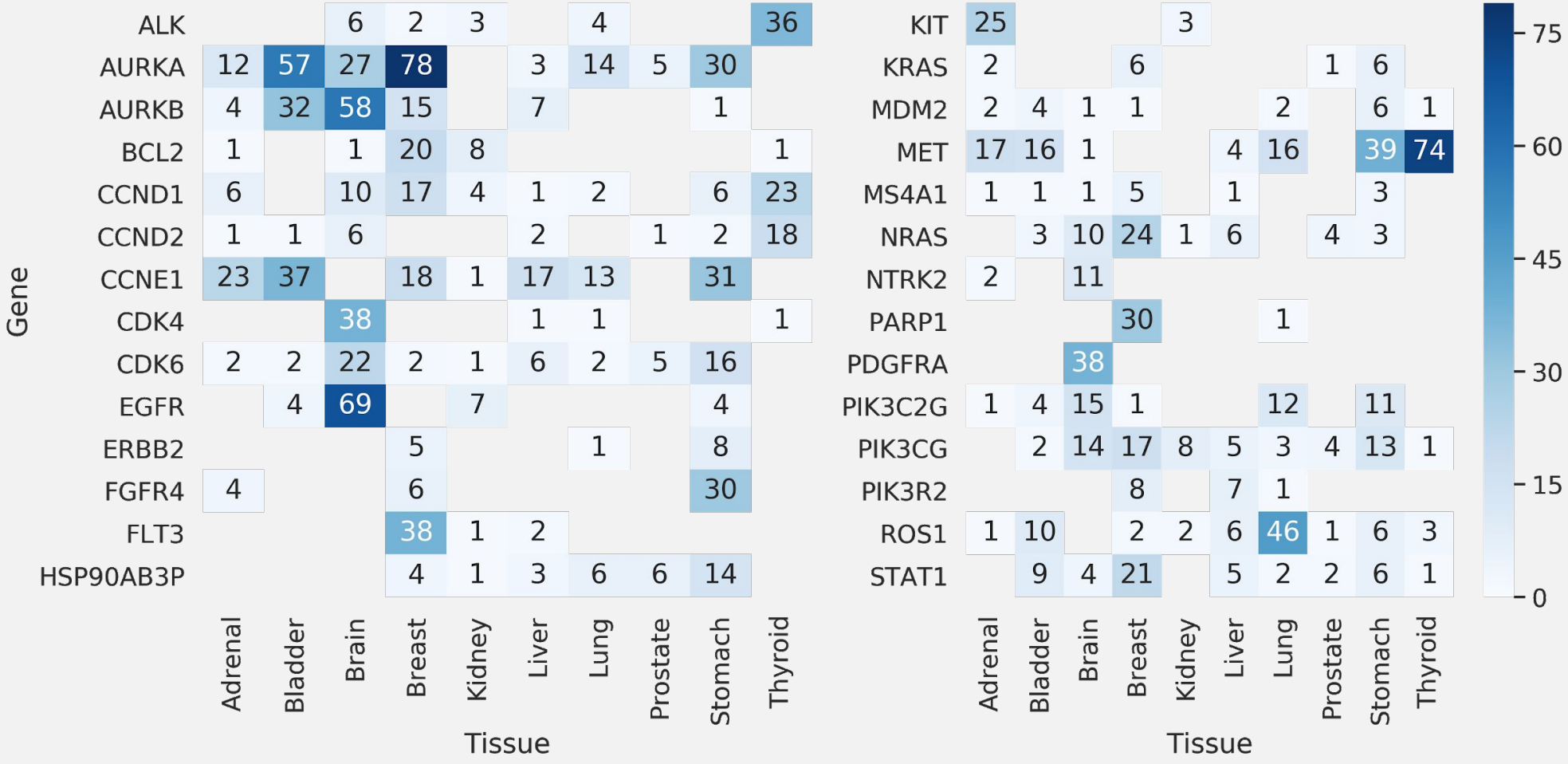
Tissue	Weight (%)
Small Intestine	31
Stomach	30
Minor Salivary Gland	22
Kidney	7
Bladder	6.3
Other	~4



Performance when matched tissue is removed

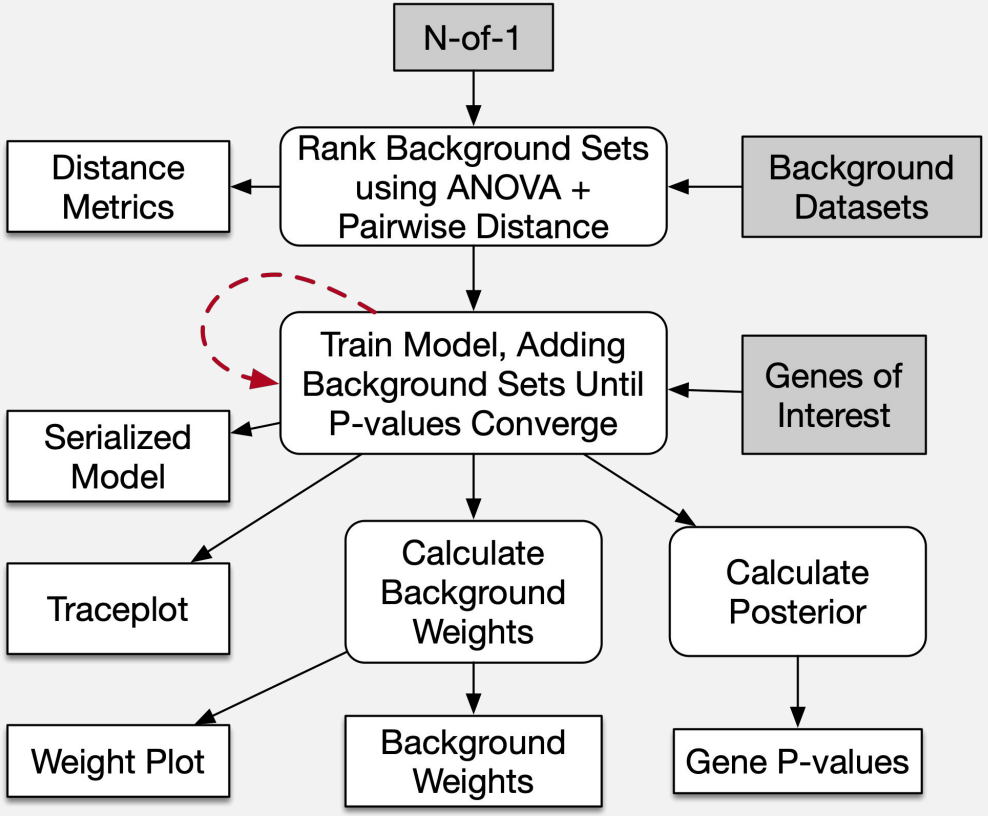


Outlier counts for Treehouse genes (n>10)



Software engineering

Identifying Gene Expression Outliers for an N-of-1 Patient



N-of-1 Gene Outlier Detection

For RNA-seq Expression Data

code style black | build passing | coverage 99%

Acknowledgements

Committee
David Haussler
Benedict Paten
Olena Morozova-Vaske

Lab Colleagues
Jordan Eizenga
Joel Armstrong
Ryan Lorig-Roach
Colleen Bosworth
Charles Markello
Trevor Pesout
Andrew Bailey
Xian Chang
Kishwar Shafin
Marina Haukness
Yohei Rosen
Alden Deran

Toil Team
Hannes Schmidt
Jake Narkizian
CJ Ketchum
Natan Lao
Alex Hancock
Lon Blauvelt

O.G. Cohort
Ian Fiddes
Arjun Rao
Nathan Schaefer
Edward Rice

Family
David, Meredith, Brian, Mom, and Dad

CGL Scientists
Jingchun Zhu
Mark Diekhans
Melissa Cline
Max Haeussler
Adam Novak
Jean Monlong
Hongxu Ding

CGL Engineers
Brian Craft
Jesse Brennan
Abraham Chavez
David Steinberg
Mary Goldman
Kevin Osborn
Walt Shands

CGL Admin
Kelly Sauder
Ben Coffey
Tina Bernard
Lynn Brazil
Nadine Gassner

Treehouse
Holly Beale
Lauren Sanders
Rob Currie
Sofie Salama
Isabel Bjork
Katrina Learned
Jacob Pfeil
Jackie Rogers

Grad Students
Jacob Schreiber
Miten Jain
Verena Friedl
Alana Weinstein
David Haan
Jessie Lopez
Kevin Hagy
James Francis
Jolene Draper
Arthur Rand

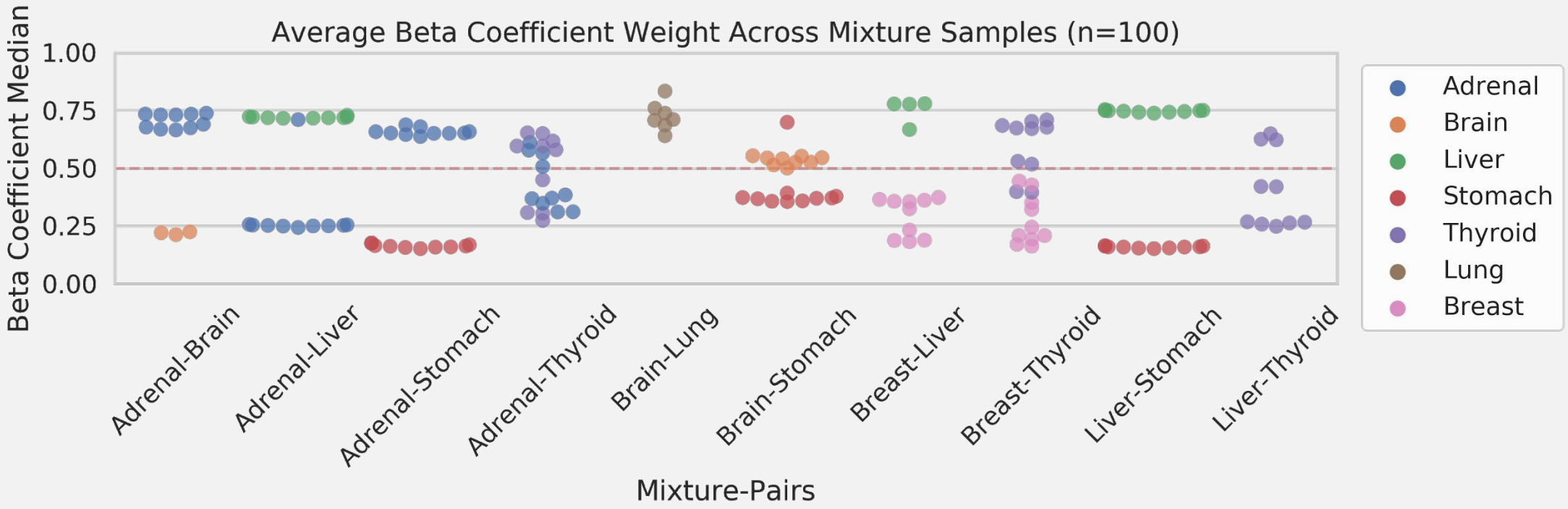
Other Staff
Erich Weiler
Kate Rosenbloom
Hiram Clawson
Kristof Tigyi
Tracie Tucker

Friends
Kaia Jystad
Aaron Cooke
Lacy Cooke
Sam Cloud
Scott Morley
Tyler Reed
Steven Santana
Grace Kistler-Fair
Nathan Peterson
Reese Robertson
Robert Hoffman
Jack Milner

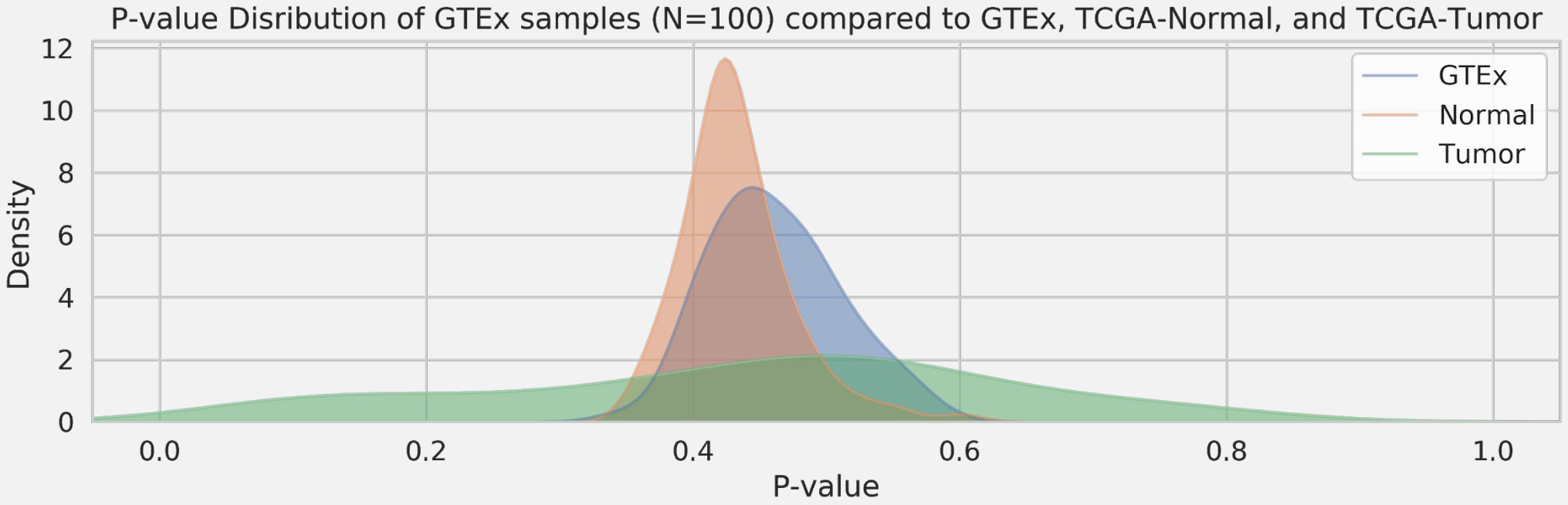
And many more...

Supplementary Slides

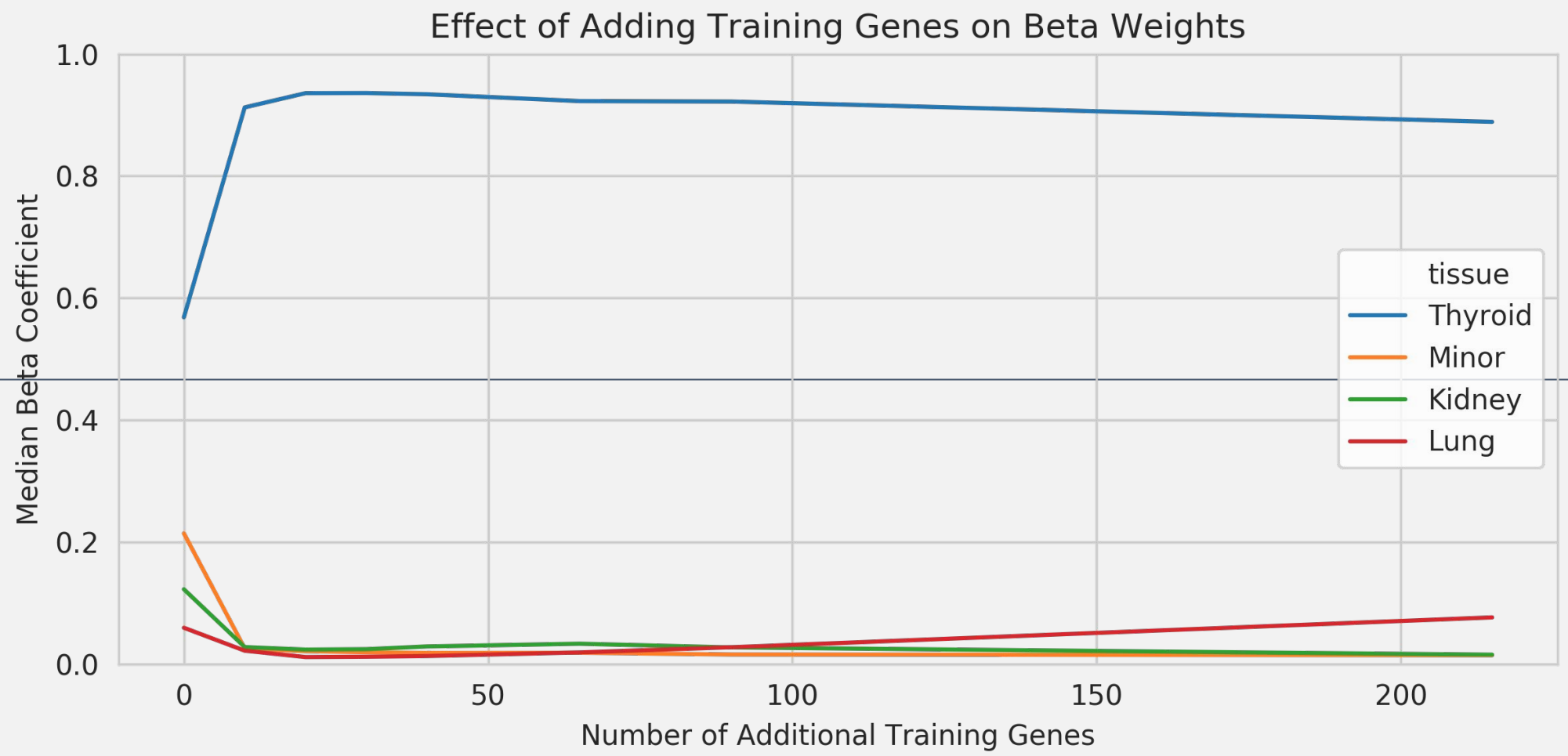
Mixture simulation



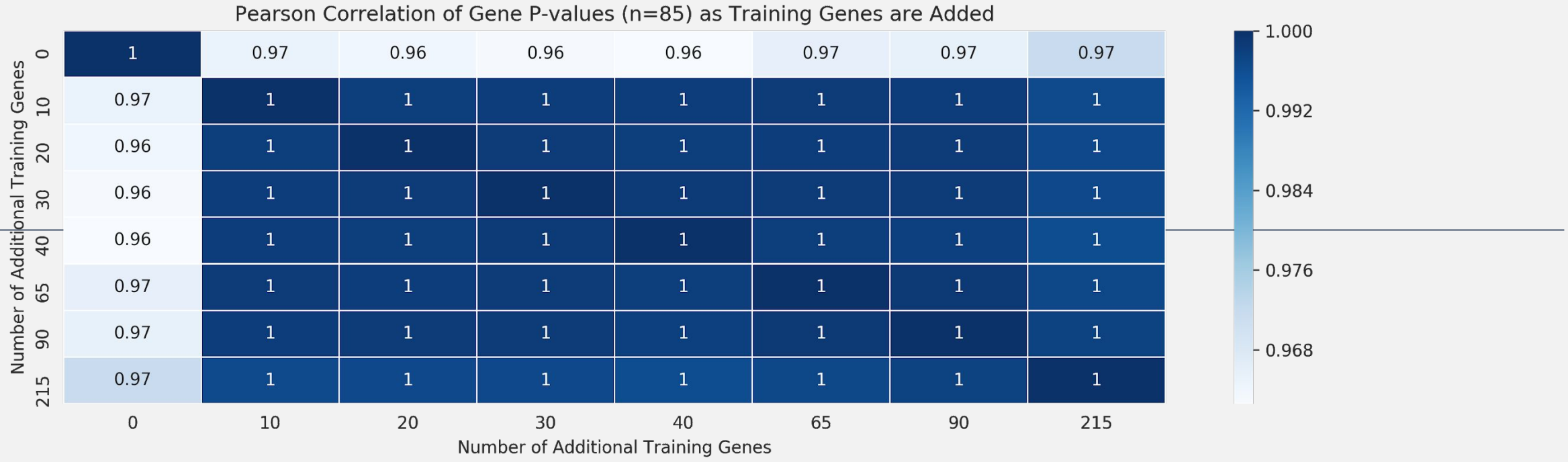
Negative control



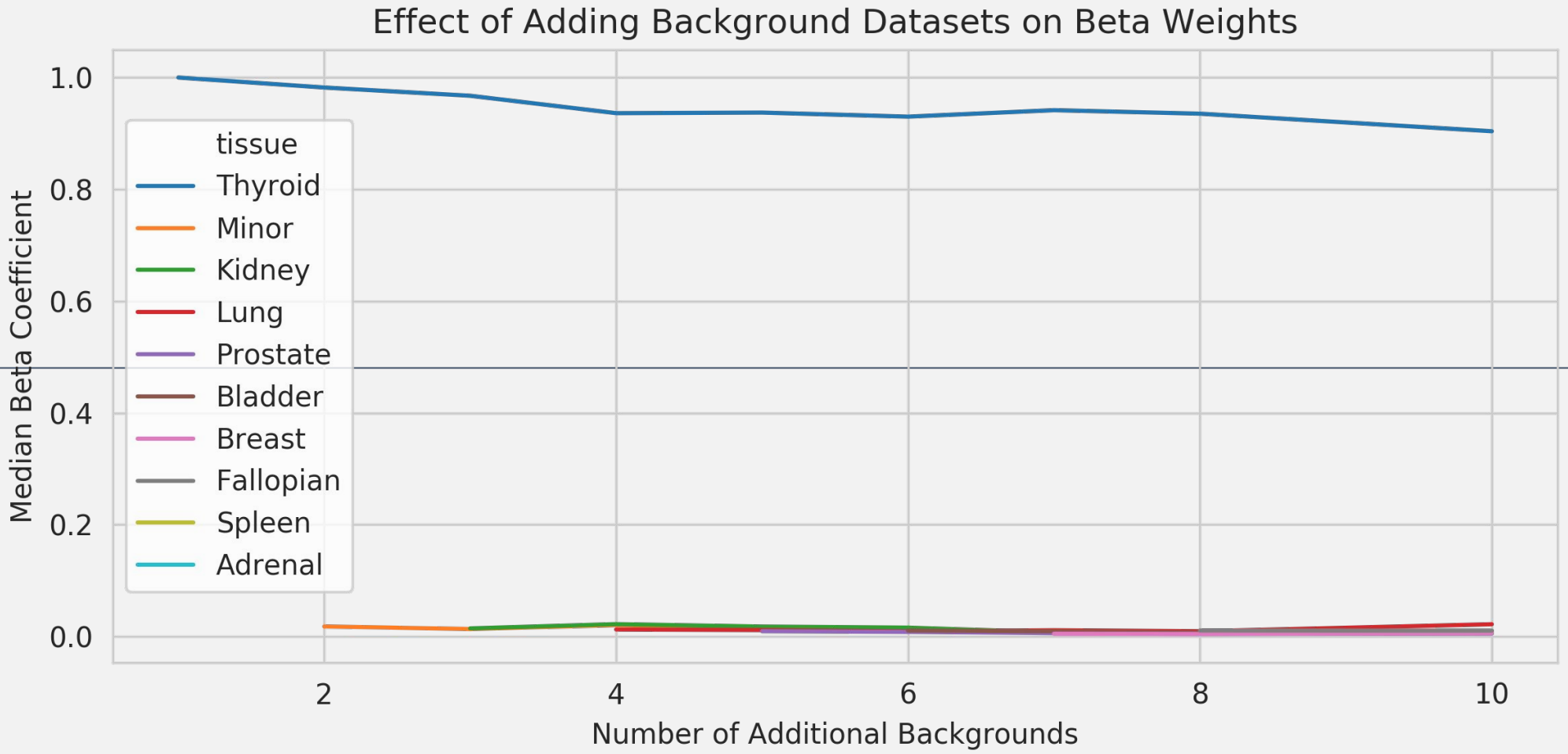
Supplementary Slides: n genes on parameters



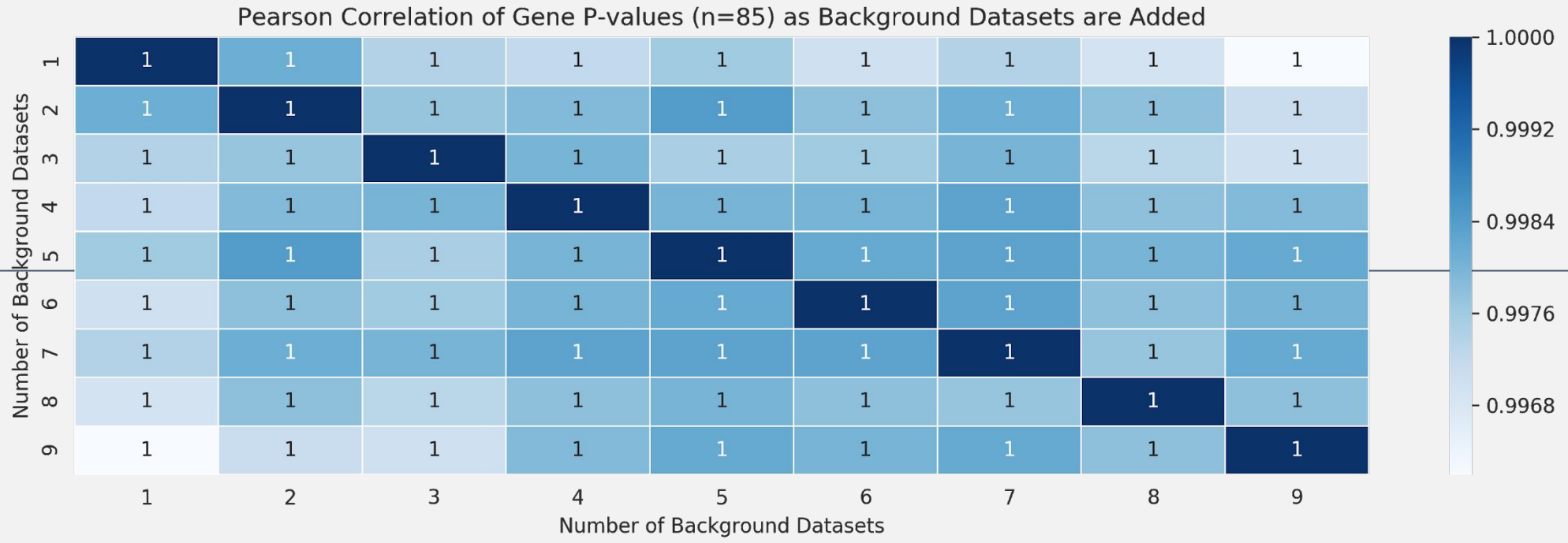
Supplementary Slides: n genes on parameters



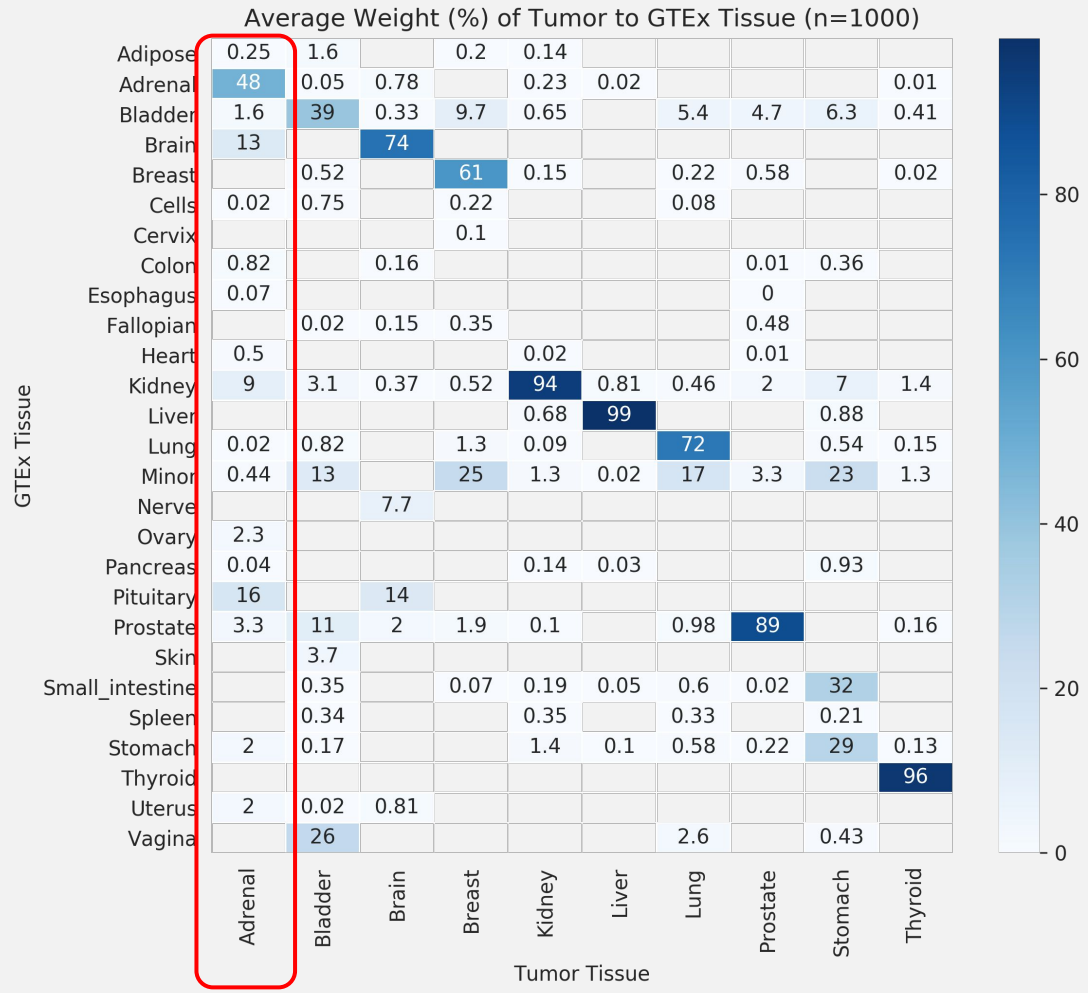
Supplementary Slides: n backgrounds on parameters

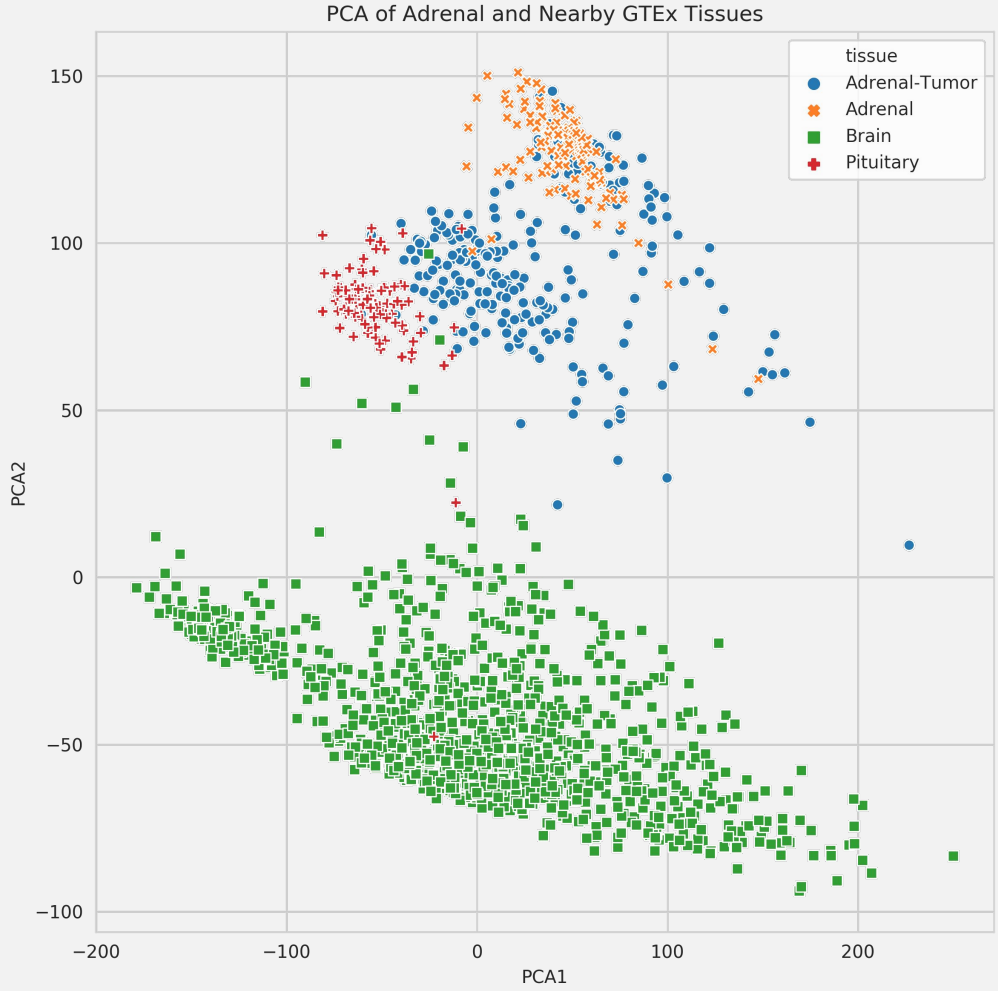


Supplementary Slides: n backgrounds on parameters

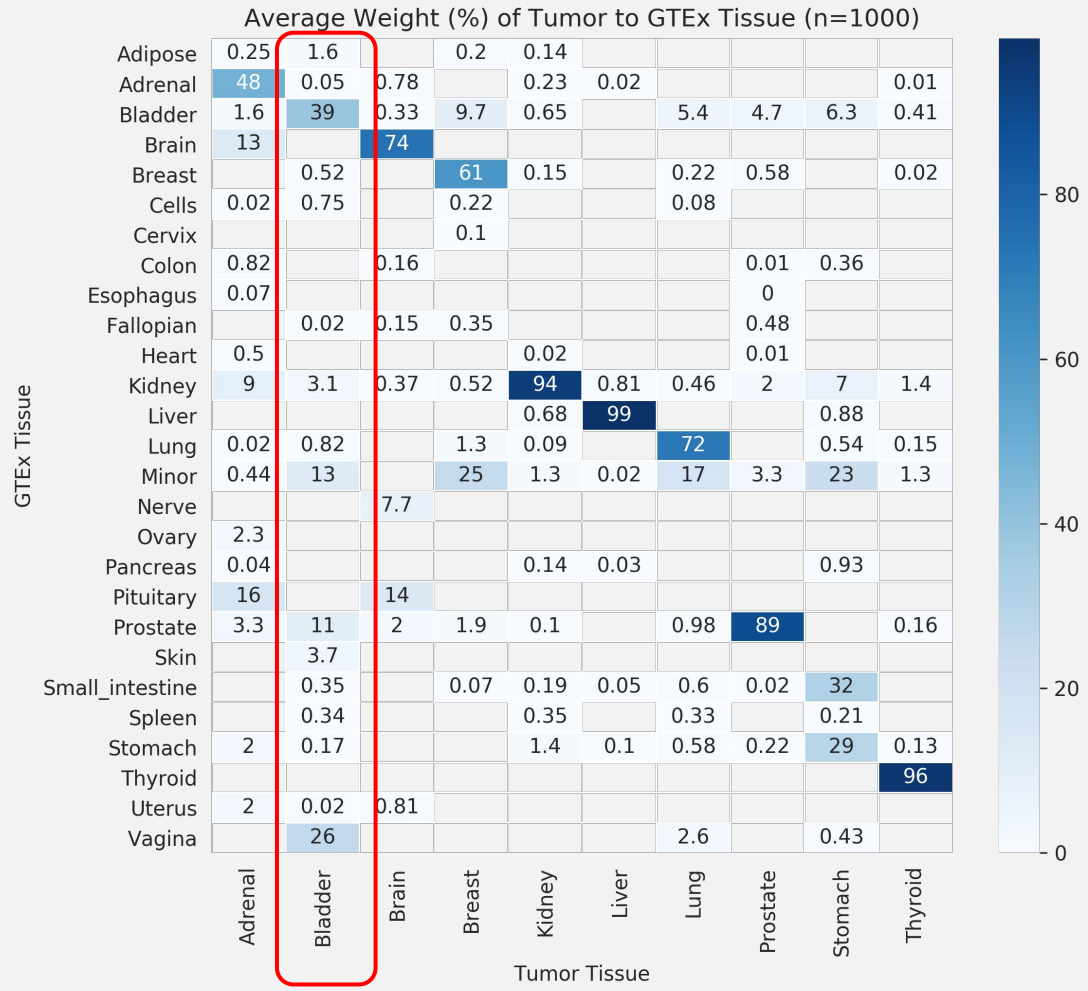


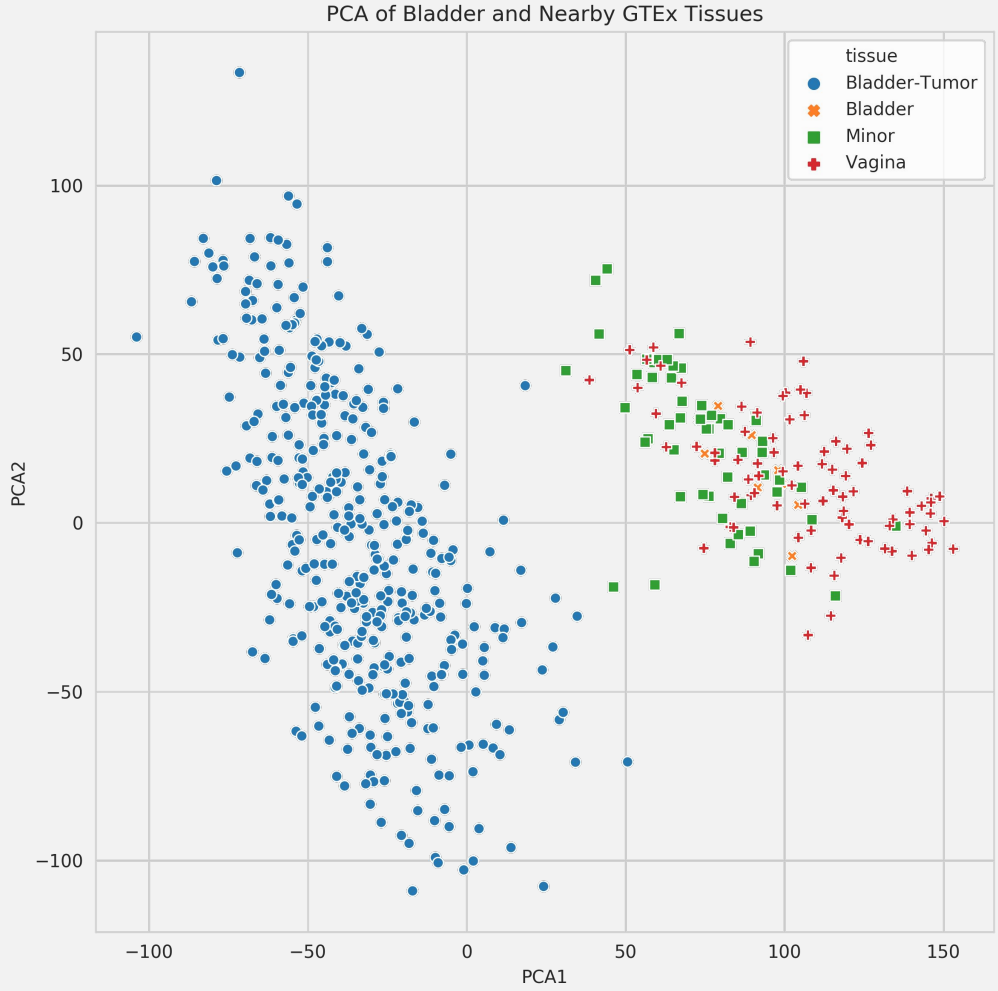
Dimensionality Reduction of Low-Weight Samples





Dimensionality Reduction of Low-Weight Samples





Supplementary Slides

Template