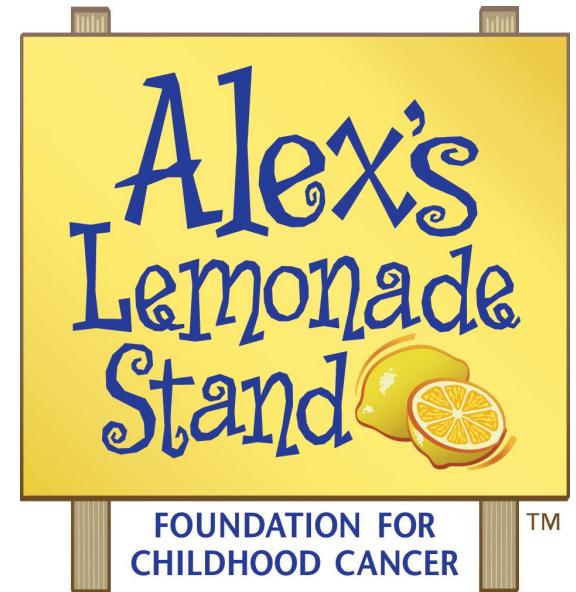


Childhood Cancer Data Lab

Powered By



Jaclyn N. Taroni, PhD
CCDL Data Scientist

Our mission is to find cures for
childhood cancer **sooner** by empowering
scientists and doctors with big data.

**How does open science help
us tackle childhood cancer?**

Childhood cancers are collectively deadly...



Childhood cancer
is the
leading cause
of death
by disease in
children under
the age of 19
in the U.S.

... but individually rare.

No one institution has enough data on its own to solve every problem in childhood cancer.



Slide adapted from ALSF Crazy 8s

**Our mission is to find cures for
childhood cancer sooner by empowering
scientists and doctors with big data.**

Software

Scientific Workflows

Training

Search for harmonized transcriptome data

 Search

Try searching for: [Notch](#) [Medulloblastoma](#) [GSE24528](#)



Find the data you need

Search the multi-organism collection of genome wide gene expression data obtained from publicly available sources like GEO, ArrayExpress, and SRA. The data has been processed uniformly and normalized using a set of standardized pipelines curated by the [Childhood Cancer Data Lab \(CCDL\)](#).



Create custom datasets

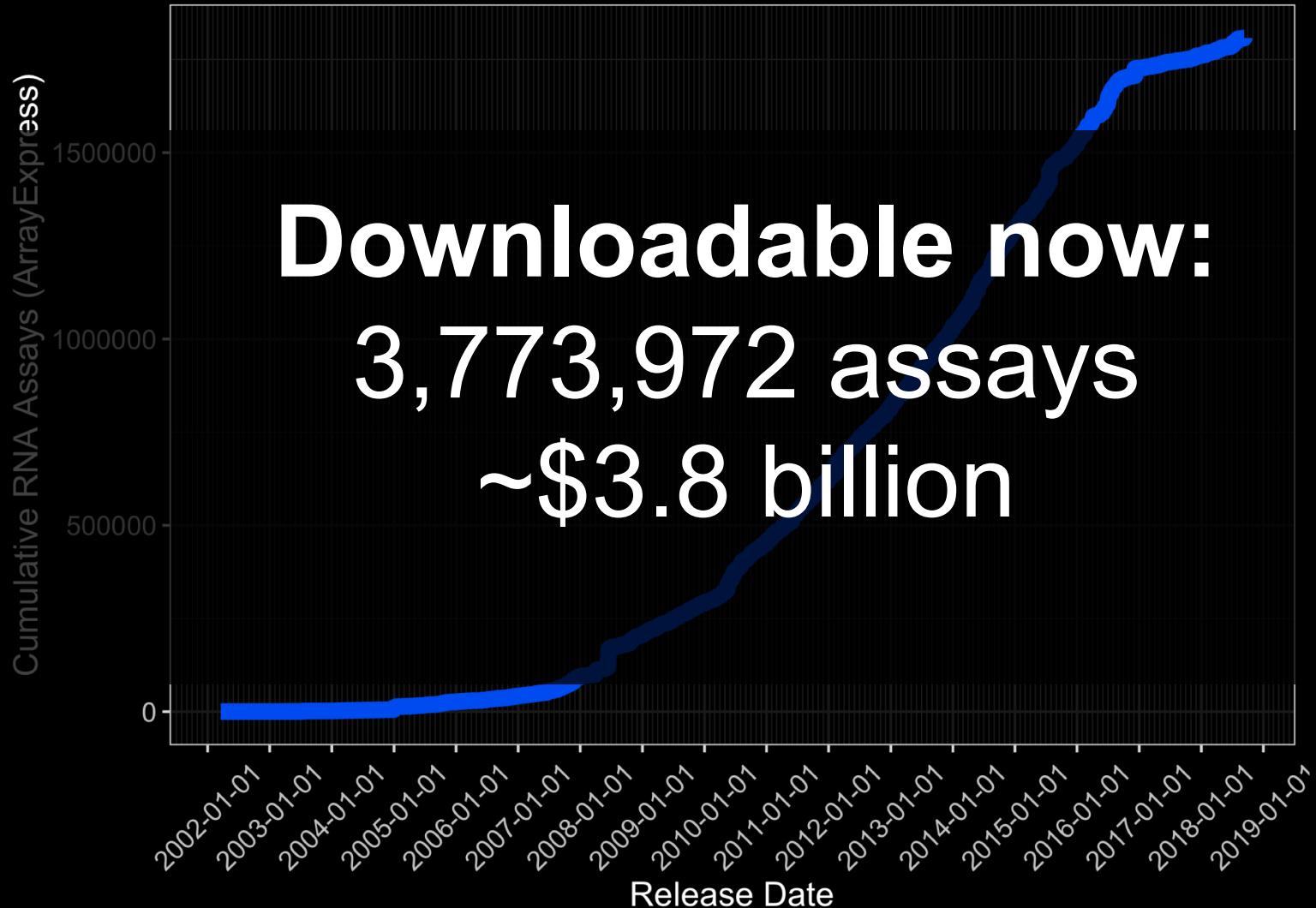
Build and download custom datasets tailored to your needs including gene expression matrices and sample metadata.



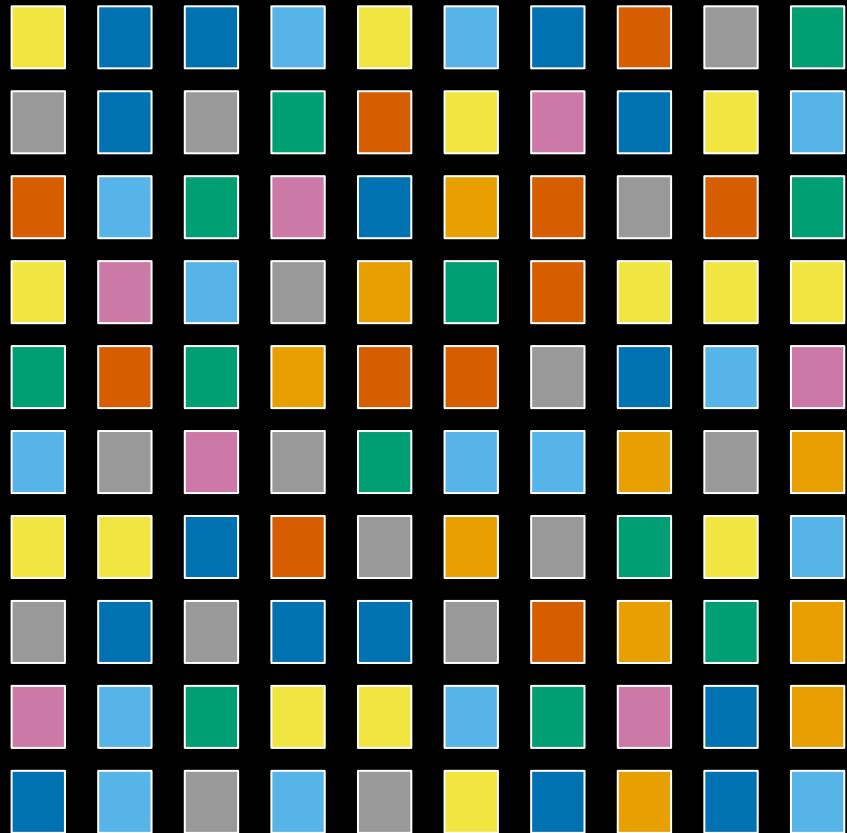
Explore the docs

Learn more about our [pipelines](#) or explore [example workflows](#) to see how data from refine.bio can help with your analysis.

Transcriptome data growth over time



Why can't people use it?



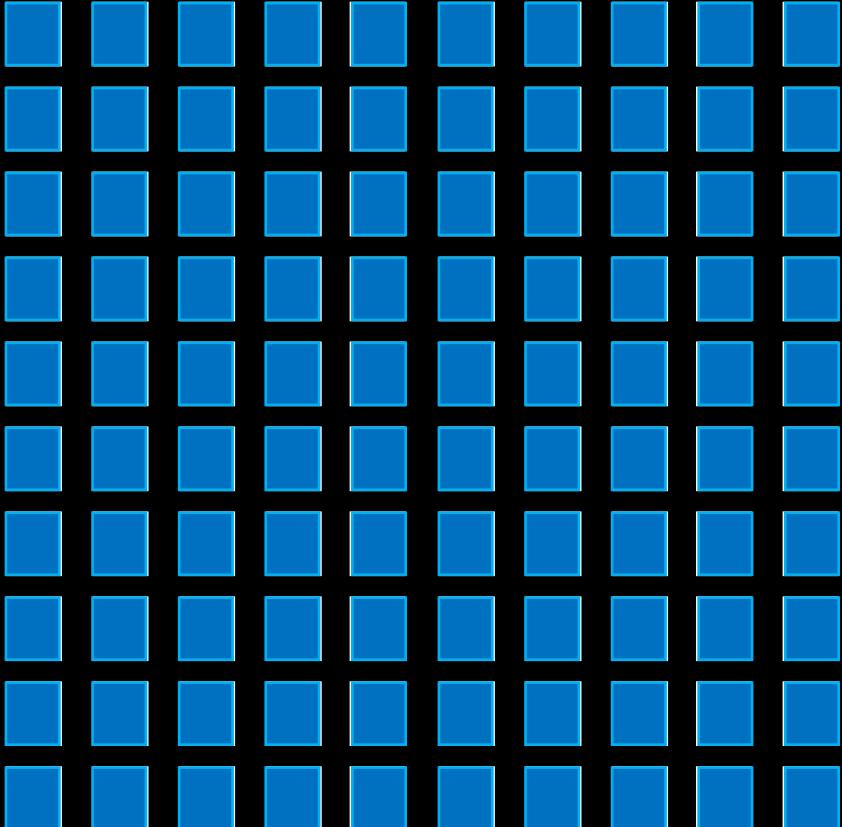
Poorly labeled,
Unclear how it was
processed



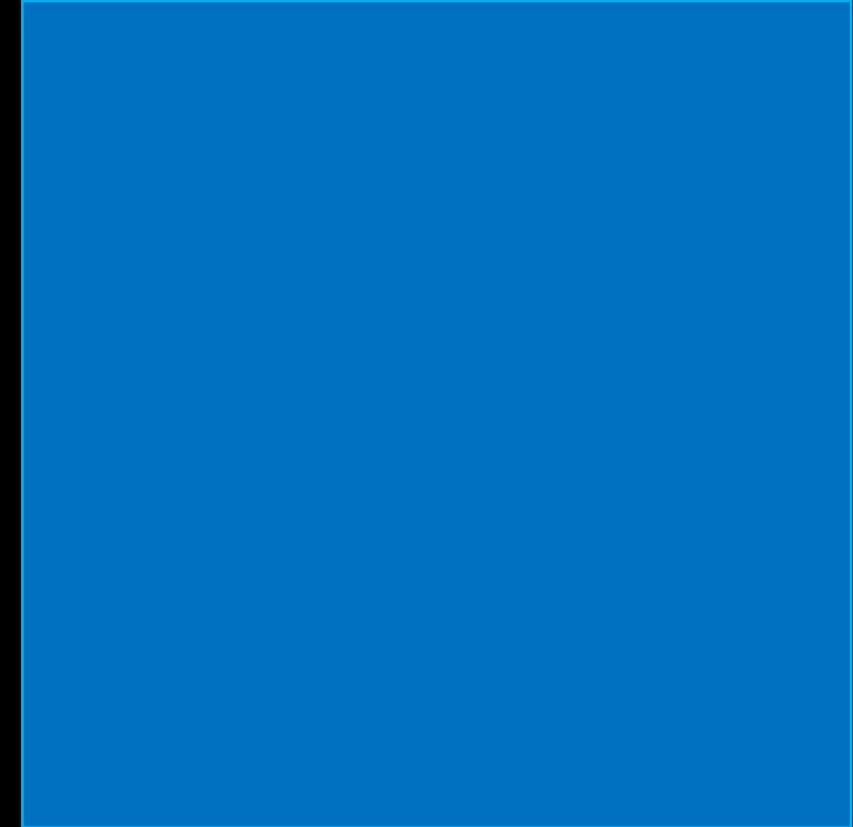
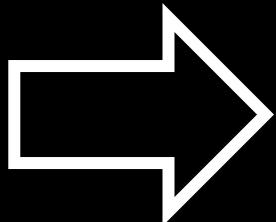
**A collection of harmonized genome-wide
gene expression data readily available for
use**

refine.bio in two steps

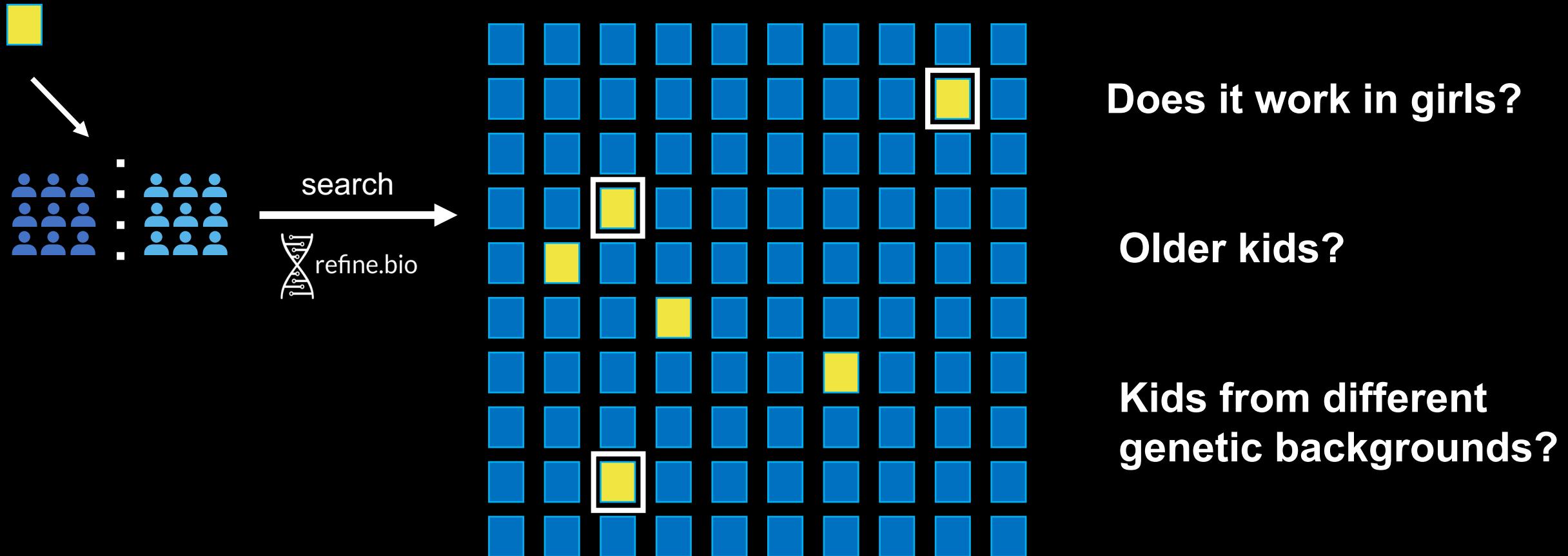
uniformly process



aggregate and normalize



refine.bio accelerates the cycle of experimentation and validation



**Our mission is to find cures for
childhood cancer sooner by empowering
scientists and doctors with big data.**

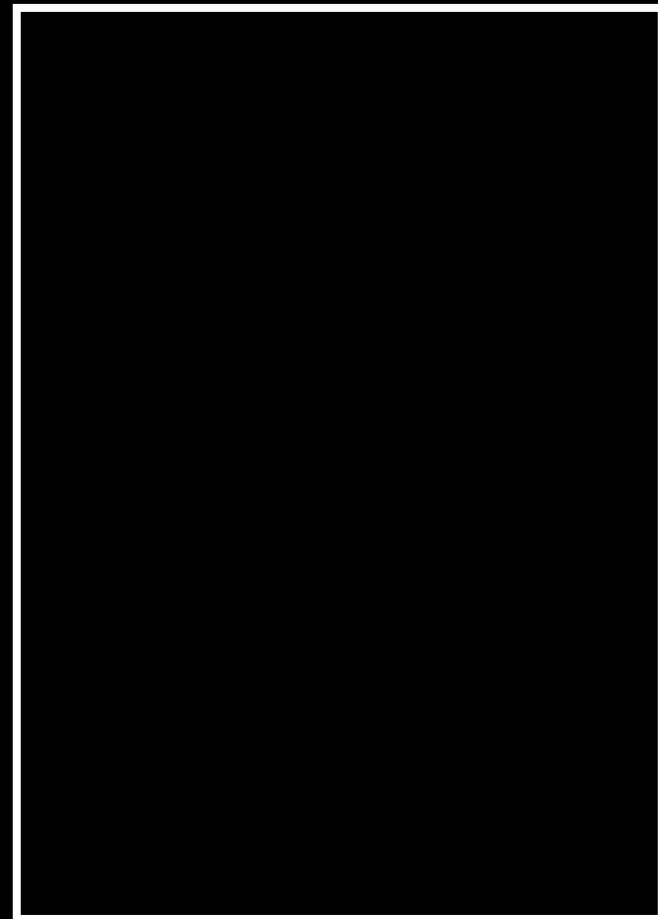
Software

Scientific Workflows

Training

Machine learning usually needs many examples

We can know some things



About many people

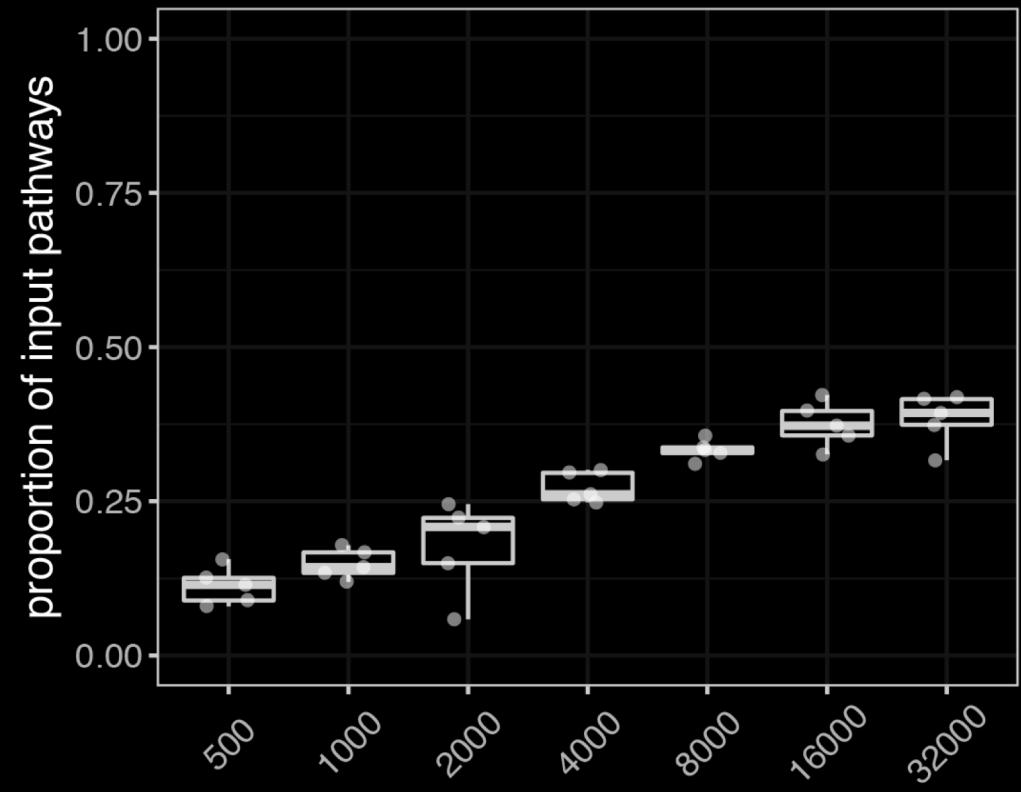
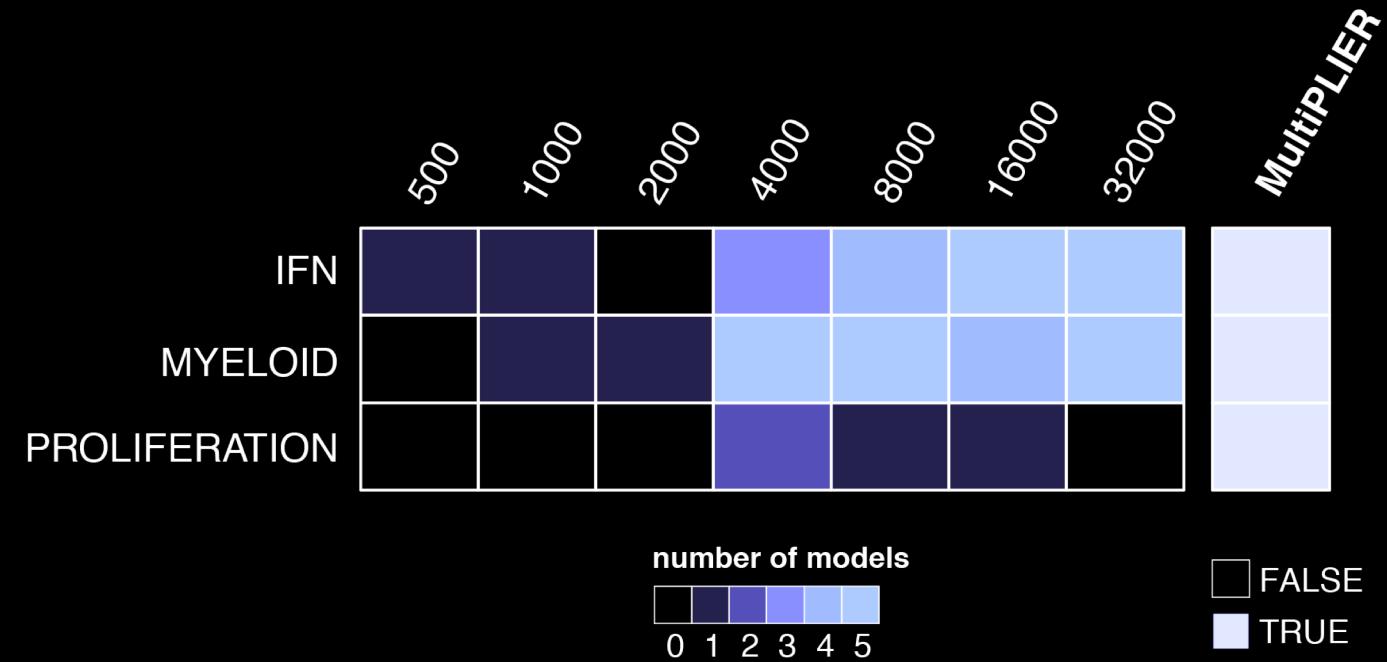
Rare disease datasets have many features, but few examples

We can know many things

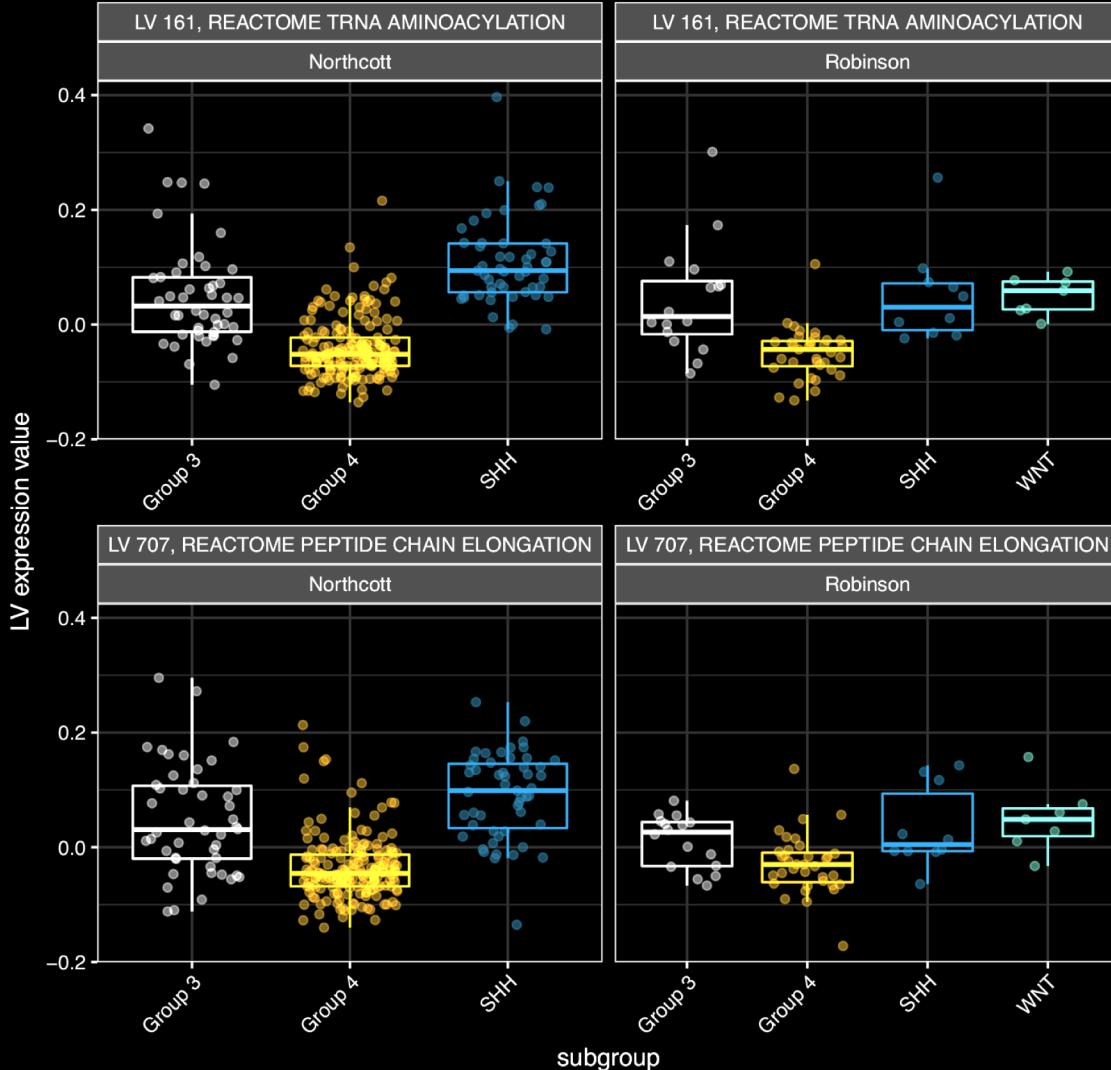


About few people

Learning from larger collection of random samples allows us to learn more biology!



MultiPLIER tells us more about medulloblastoma than we could learn from medulloblastoma alone



**Our mission is to find cures for
childhood cancer sooner by empowering
scientists and doctors with big data.**

Software

Scientific Workflows

Training

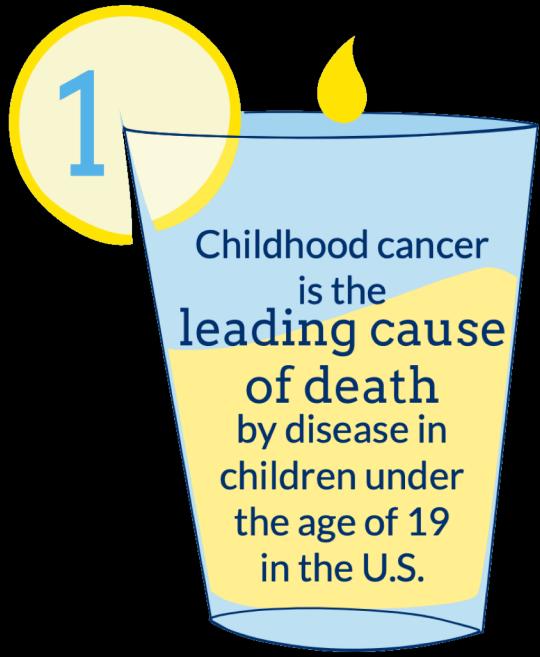
Training material is available on GitHub

Houston CCDL Training

Location: Northwestern Mutual. 1 Riverway, Suite 900, Houston TX Dates: 3/27/2019-3/29/2019

Time	Topic
### Day 1	
9:00 AM	Intro & Welcome (Install Docker ; Prepare the RNA-seq Environment , Introduction to RStudio)
10:00	Break
10:15	Introduction to R
11:00	R Intermediate Topics (including tidyverse)
12:00 PM	Lunch
1:00	Run Salmon for RNA-seq Salmon and QC, Shell scripting for reproducibility
2:00	Data Diagnostics w/ a Gastric Cancer Dataset tximport , Exploratory Analyses
3:30	Break
4:00	Differential Expression Analysis tximport , Differential Expression
5:00	End
### Day 2	
9:00	scRNA-seq I: Normalizing a count matrix
10:15 AM	Break
10:30	scRNA-seq II Processing 10X data from raw
12:30 PM	Lunch
2:00	scRNA-seq III: Dimensionality reduction
3:00	Break
3:30	Machine Learning Data Preparation , Clustering , PLIER , Differential PLIER LVs
5:00	End
6:00	Dinner at El Tiempo Washington
### Day 3	
9:00 AM	Analyze your own data (or some stuff we have)!
12:30 PM	Lunch
2:00	Presentations, interpretations, and discussion
3:30	Break
4:00	Talk from CCDL data science team and wrap-up
5:00	Adjourn





Breaking down barriers, in access and in knowledge, and put these data and powerful methods in the hands of pediatric cancer experts poised for the next big discovery



Casey Greene,
PhD
Director



Deepa Prasad
UX Designer



Dongbo Hu
Software
Engineer



Rich Jones
Software
Engineer



Ariel Rodriguez
Software
Engineer



Kurt Wheeler
Software
Engineer



Jaclyn Taroni,
PhD
Data Scientist



Candace Savonen
Biological Data
Analyst

ccdatalab.org
github.com/AlexsLemonade
[@CancerDataLab](https://twitter.com/CancerDataLab)

The CCDL is hiring!

<https://www.ccdatalab.org/careers>