Developing an age-specific transcriptomic model of pediatric AML through robust large-scale multi-omic data analysis

Jenea I. Adams

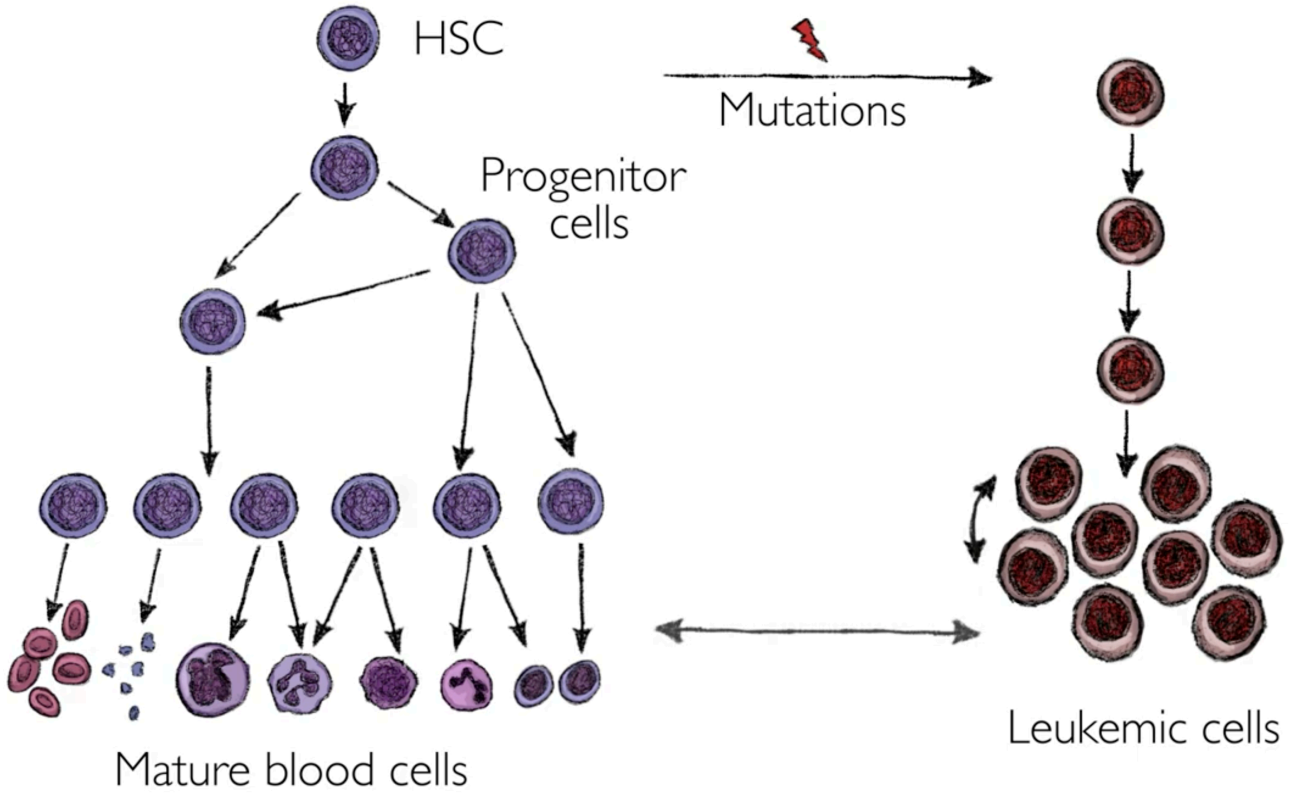
Ph.D. Student (GCB)

Xing Lab Roundtable

Tuesday, April 13th, 2021

Acute myeloid leukemia (AML) is the most fatal of childhood cancers with no good treatments

- Caused by accumulation of immature myeloid cells in bone marrow
- Affects 25% of children with leukemia
- Chemotherapy for acute lymphoblastic leukemia (ALL) is generally successful
  - Kills healthy and aberrant B-cells → stable prognosis

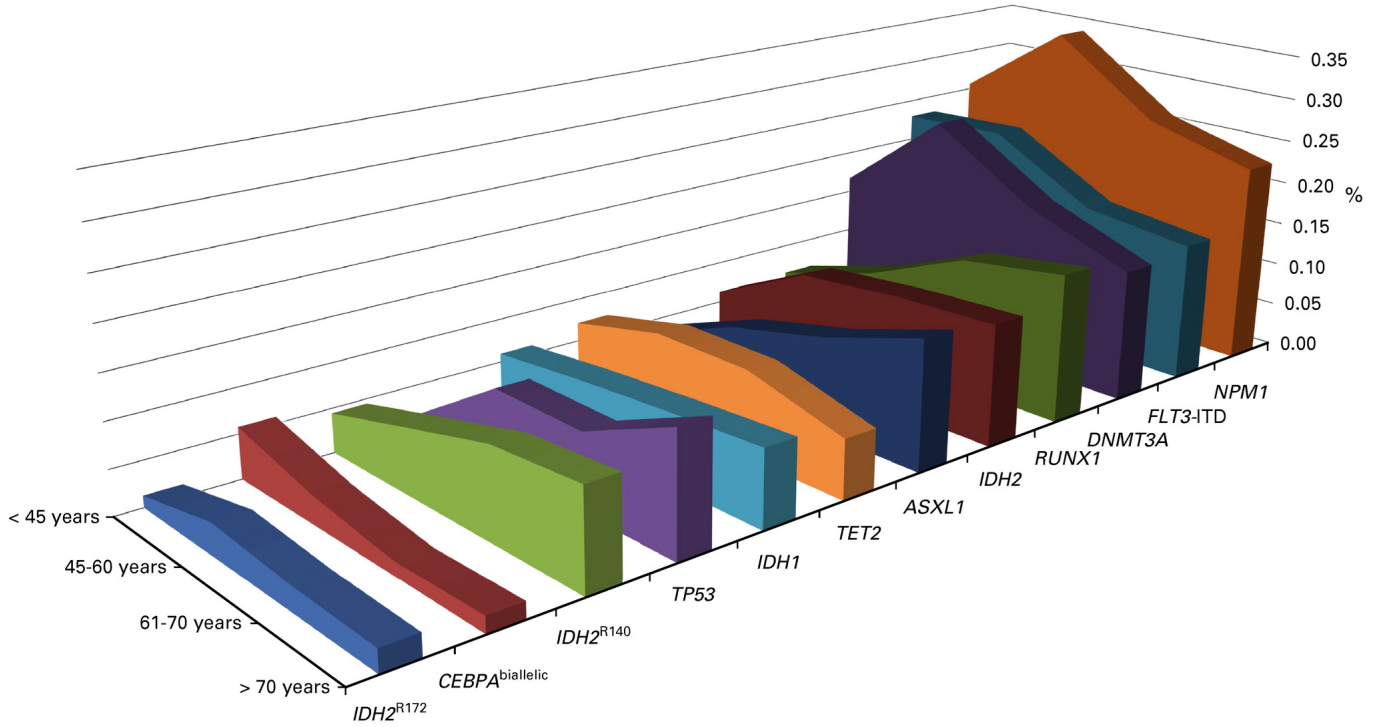


van Galen *et al.*, 2019

Current treatment strategies for pediatric AML are based adult genomes/transcriptomes

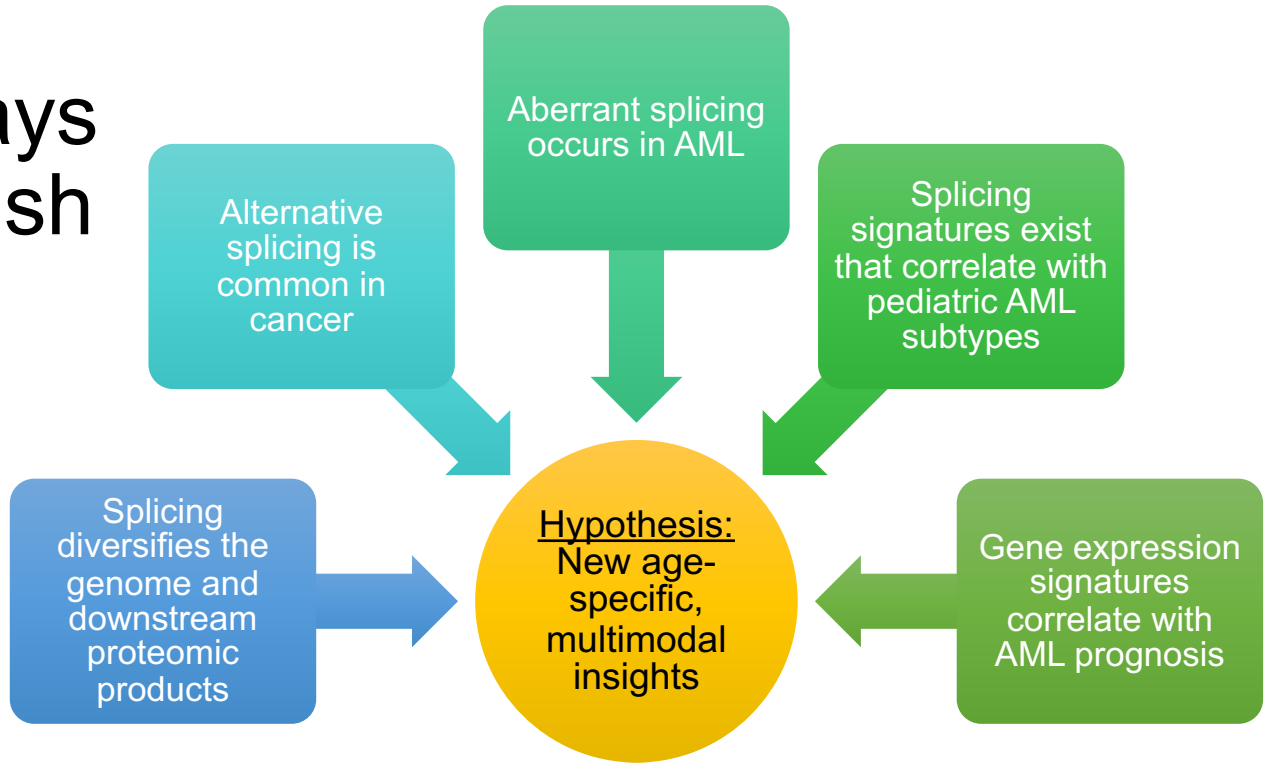
- Mean age of AML onset is 45 years
- ”Pediatric” considered 0 years to ~30 years old
- Age-specific genomic signatures have been linked to prognosis (Bullinger *et al.*, 2017)
- Hematopoietic expansion has age-specific regulation (Bullinger *et al.*, 2017)

What about splicing?

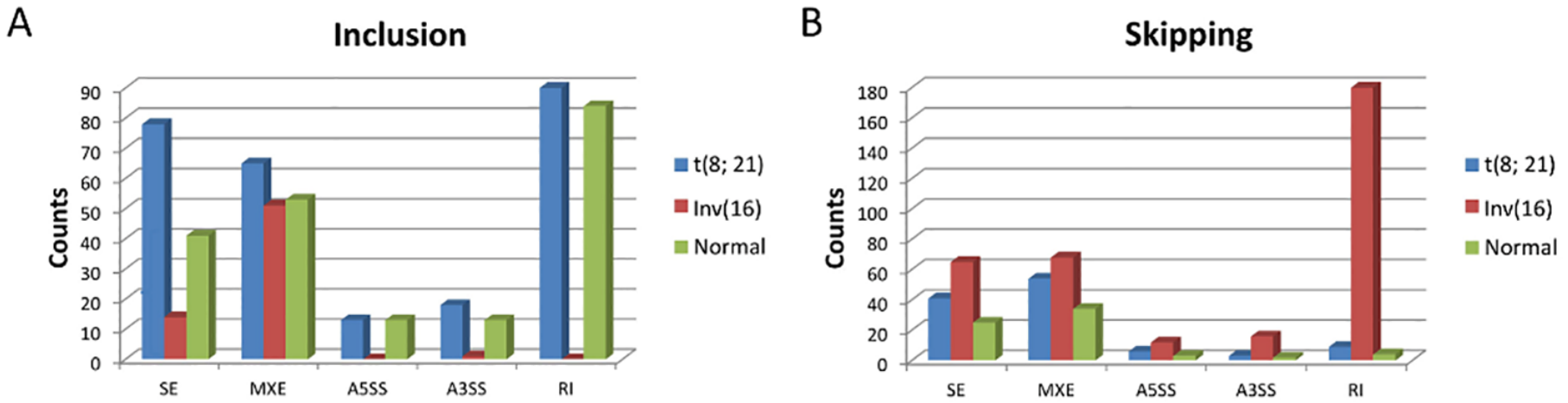


Bullinger *et al.*, 2017

Alternative splicing and correlated molecular pathways could help to better distinguish pediatric vs adult AML



PEGASAS correlates gene ontology and alternative splicing events



Hsu *et al.*, 2016

Hypothesis: Alternative splicing and correlated molecular pathways could play important roles in distinguishing pediatric vs adult AML

Aim 1

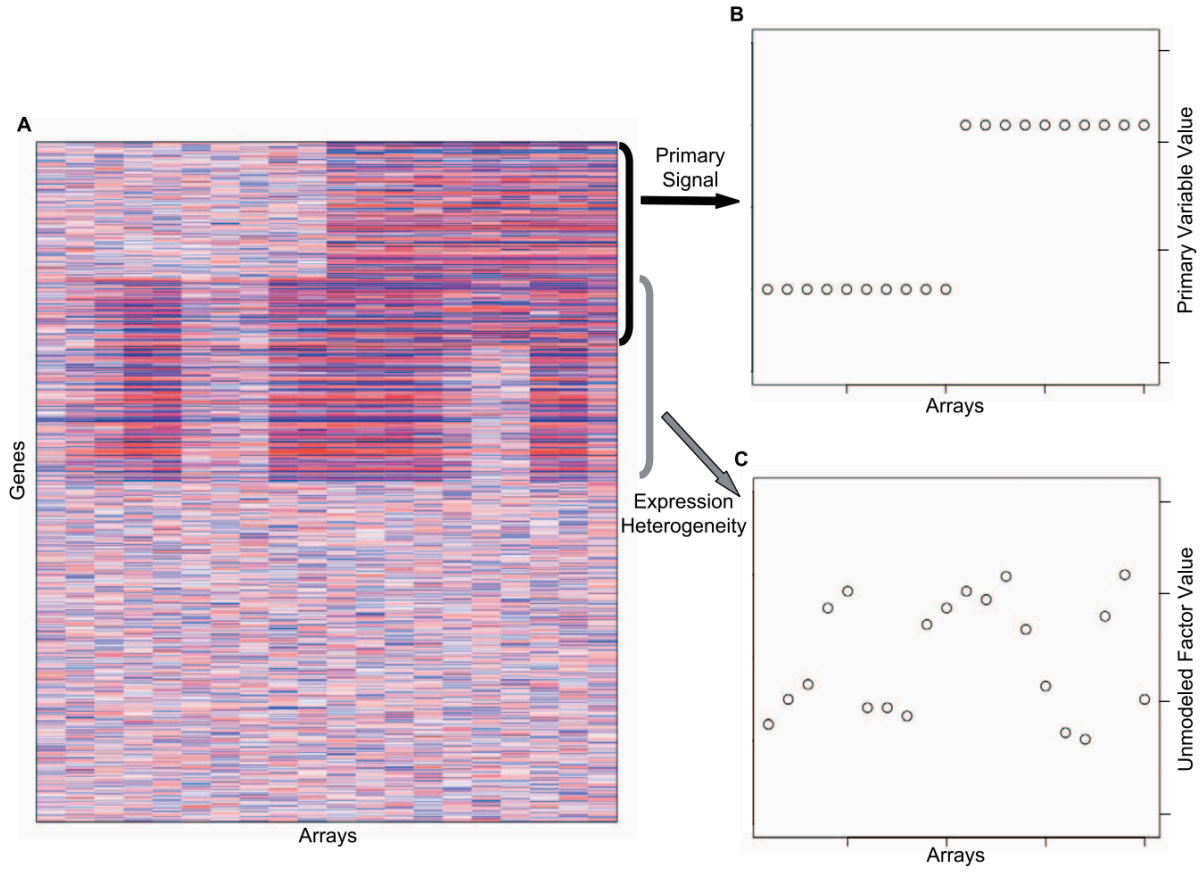
Improve analysis of splicing in large heterogeneous RNA-seq datasets

Aim 2

Discover age-specific, pathway-dependent alternative splicing patterns in pediatric AML RNA-seq data

Aim 1: Improve analysis of splicing in large heterogenous RNA-seq datasets

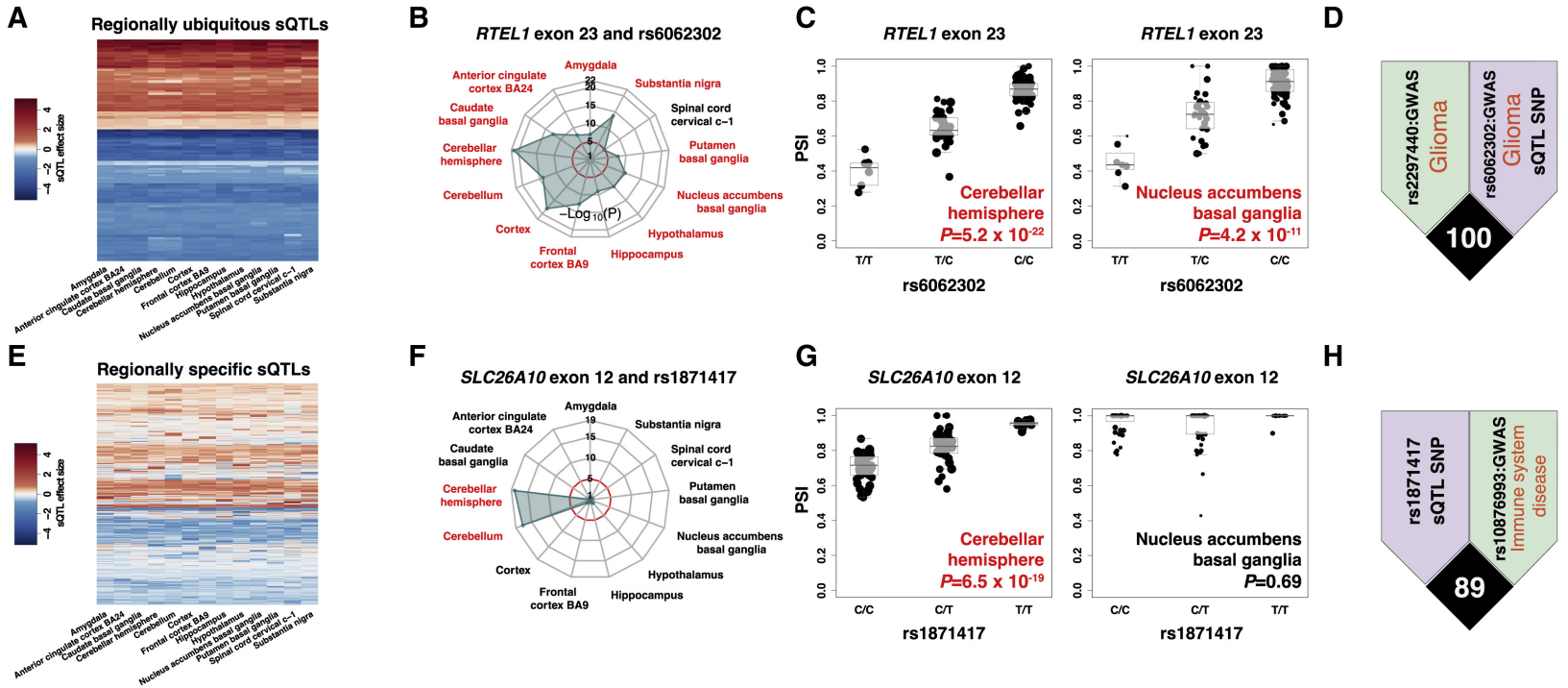
- PEGASAS lacks an approach to make confounding factor-informed correlations
  - Batch effects in RNA quality
  - Gender
- Capture and use expression heterogeneity to mitigate batch effects → Surrogate variable analysis (SVA)
- Using SVA could improve reproducibility and downstream accuracy
- Approach
  - Compare pathway-relevant exons detected with/without SVA
  - SVA → more significant splicing events?



Leek & Storey, 2007

A consistent approach for evaluating SVA could refine downstream analysis

- Zhang et al passes logit-transformed PSI values to SVA
  - Goal: identify sQTLs across different brain regions
  - How might the identified ubiquitous vs regionally specific sQTLs vary with different ”flavors” of SVA evaluated?



Higher resolution in batch correction may account for more within- and between-group heterogeneity

ELSEVIER

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing

Wenan Chen [a,1], Silu Zhang [b,1], Justin Williams [c], Bensheng Ju [c], Bridget Shaner [c], John Easton [c], Gang Wu [a], Xiang Chen [c,*]

[a] Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN, United States
[b] Department of Diagnostic Imaging, St. Jude Children's Research Hospital, Memphis, TN, United States
[c] Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, United States

The “flavors” of surrogate variable based methods

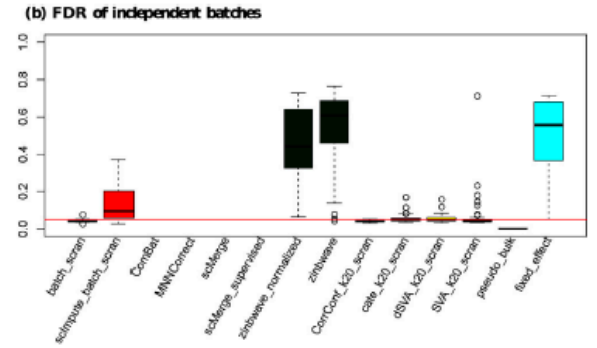
- SVA
  - Iteratively estimates probability of each gene being affected only by batch (not group)
  - Weighted SVD to estimate surrogate variables (SVs)
- cate
  - Estimates coefficients of batch effects w/factor analysis
  - Estimates batch variables w/regression
- dSVA
  - Regress out variables of interest
  - SVD on residual matrix
  - Estimates batch variables w/improved interpretability through least squares regression
- CorrConf
  - Corrects bias produced by cate when the confounding batch effect is weak
  - Fast estimation of SVs

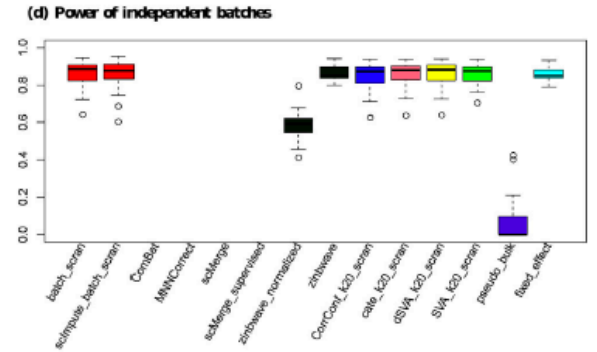
Characteristics of surrogate variable based methods

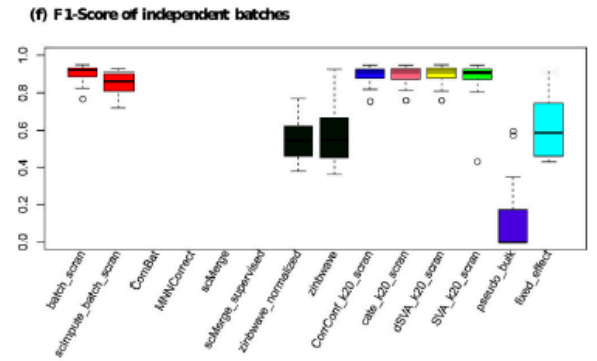
- Assumes a Gaussian distributions for the data matrix
  - In this paper, a $\log_2$(TPM+0.1) transformation before applying the methods
- More FDR control achieved with more cells
  - In this paper, they aggregated sorted cells into “pseudo-cells” to achieve this
  - Hypothesized to be connected to the Normal distribution connection

Each of the surrogate variable based methods produced variable SVs

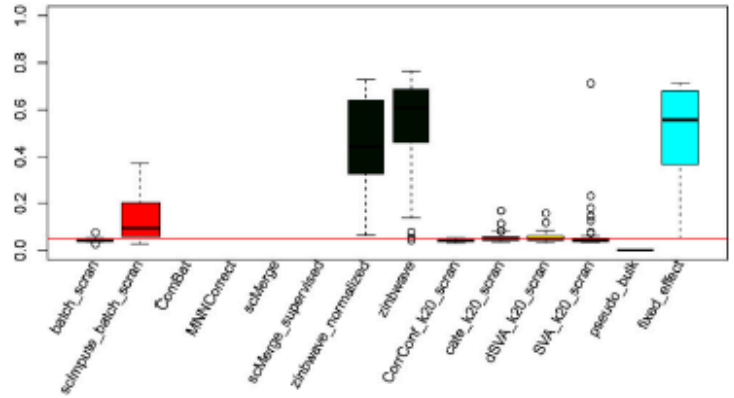
- Automatically inferred SVs → poor performance
  - Inflated FDR
  - Low power
- In scenarios with independent, latent batches, SV based methods performed the best with minor exceptions
-



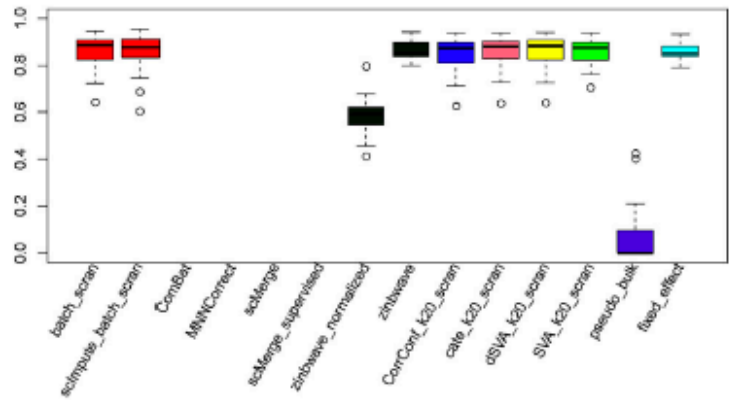




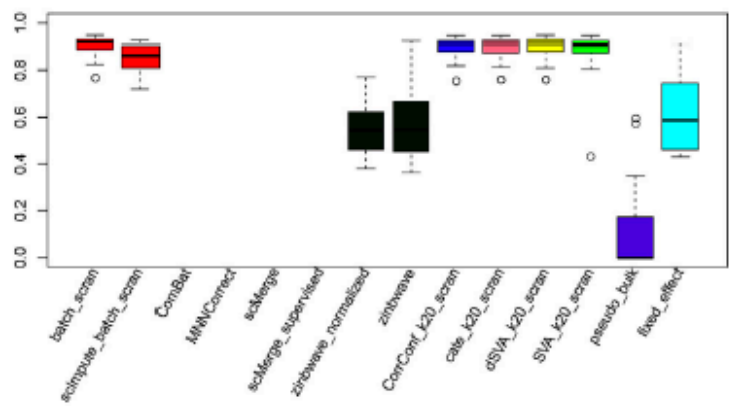
(b) FDR of independent batches



(d) Power of independent batches



(f) F1-Score of independent batches



Summarizing SV based method performance on simulated scRNA seq data

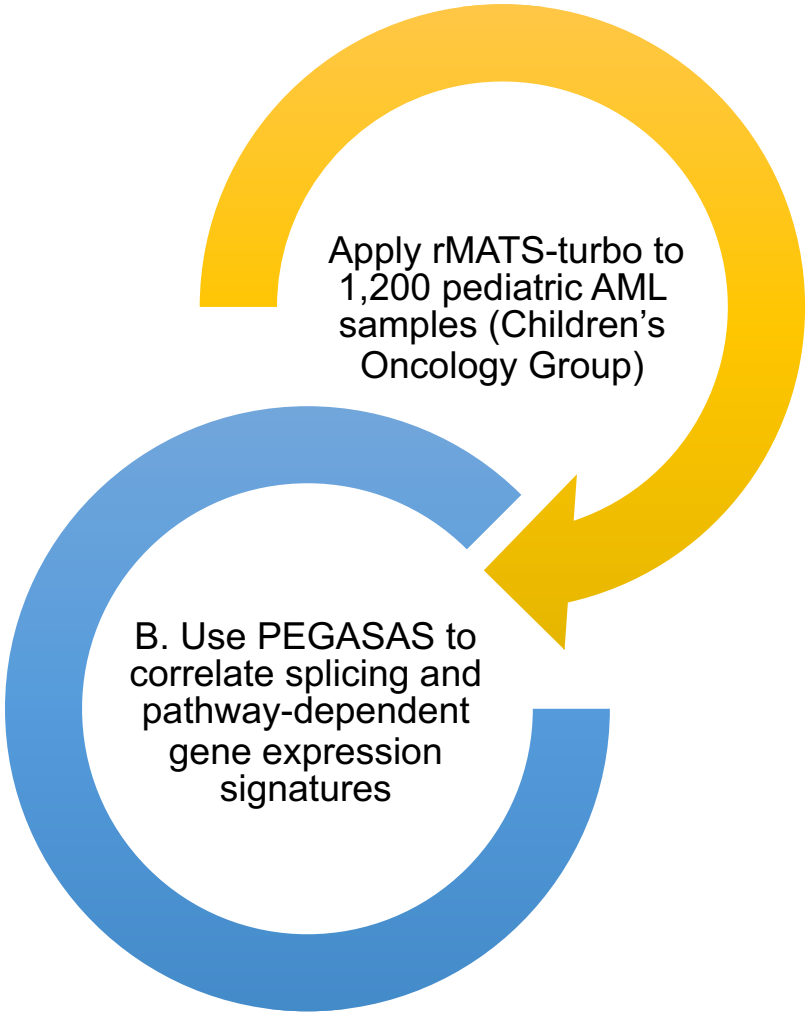
Methods	Advantage	Limitation	Recommend application
CorrConf	Good control of FDR and high power when the group effects are small	Inflated FDR or reduced power when the group effects are large or the group is impure	DE analysis for moderate effects or when the group information is highly accurate. Can be used together with SVA for a robust check
cate	Good or slightly inflated FDR and high power when the group effects are small	Inflated FDR or reduced power when the group effects are large or the group is impure	DE analysis when the group information is highly accurate. Can be used together with SVA for a robust check
dSVA	Good or slightly inflated FDR and high power when the group effects are small	Inflated FDR or reduced power when the group effects are large or the group is impure	DE analysis for moderate effects or the group information is highly accurate. Can be used together with SVA for a robust check
SVA	Good control of FDR and high power when the group effects are small; it is also little affected by the group label purity	Occasionally not very stable	Good candidate for DE analysis. Can be used together with cate/CorrConf /dSVA for a robust check

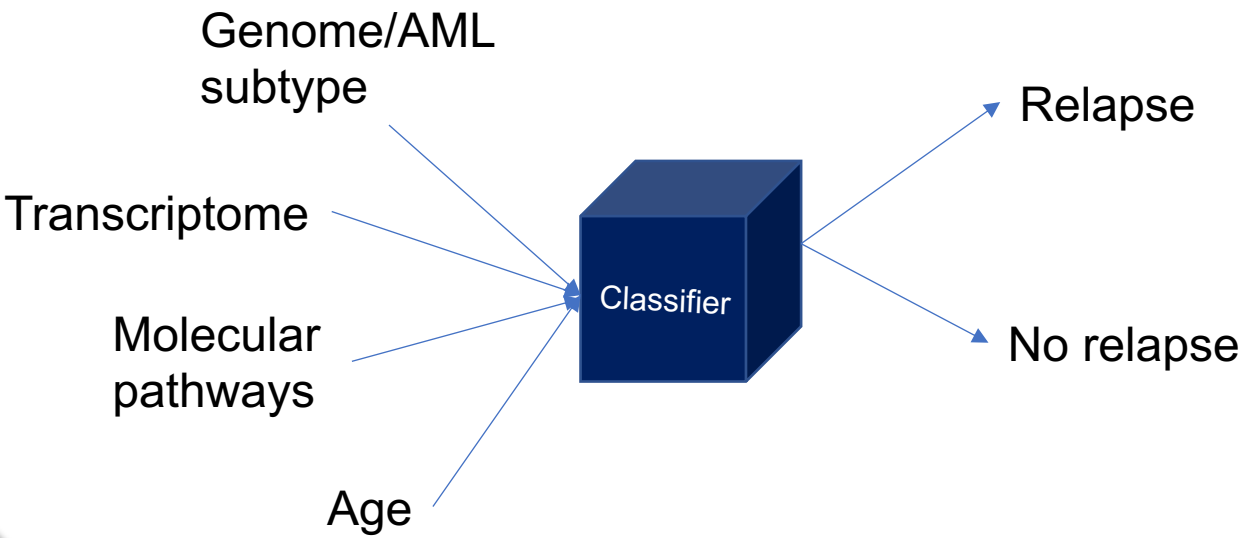
Lessons from this paper

- Evaluating SVA's performance can be a solid investment on incorporating it into existing and future computational pipelines for large-scale RNA data
- I want to use a "flavor" of SVA that could sharpen the contrast in age-specific splicing events +molecular pathways in pediatric AML
- Further defining interpretable features of SVA outputs for multi-omic cancer data is important

Thank you

Aim 2: Discover age-specific, pathway-dependent alternative splicing patterns in pediatric AML RNA-seq data





Age and genetics: how do prognostic factors at diagnosis explain disparities in acute myeloid leukemia?

Manali I Patel [1], Yifei Ma, Beverly S Mitchell, Kim F Rhoads

Affiliations + expand

PMID: 23608826 DOI: 10.1097/COC.0b013e31828d7536