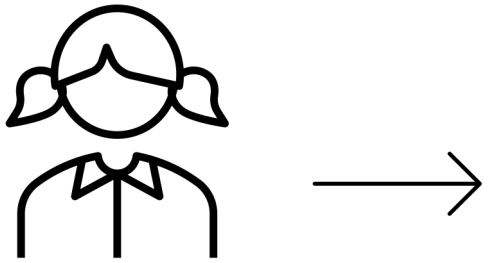
Reproducible cloud-based multi-omic analysis with CAVATICA

Jenea Adams

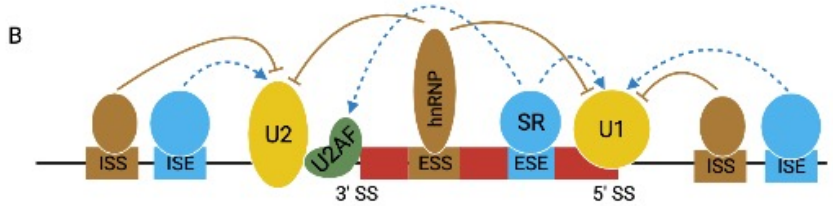
Xing Lab Roundtable

Tuesday, July 13th, 2021

Establishing a multi-omic perspective on the presentation of acute myeloid leukemia (AML) by age group



AML treatments are harsh on young bodies + data based on adults → new therapies needed



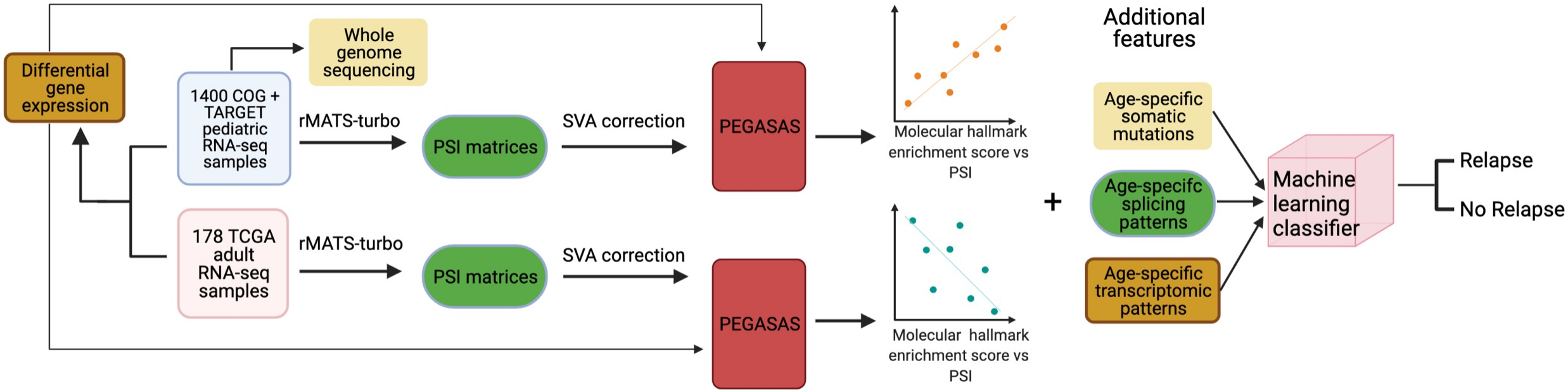
Splicing + molecular pathways = targetable avenue of disease progression

Aim 1: to improve the analysis of splicing in large heterogeneous RNA-seq datasets

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

Created with BioRender.com

Analysis across data types and datasets is imperative



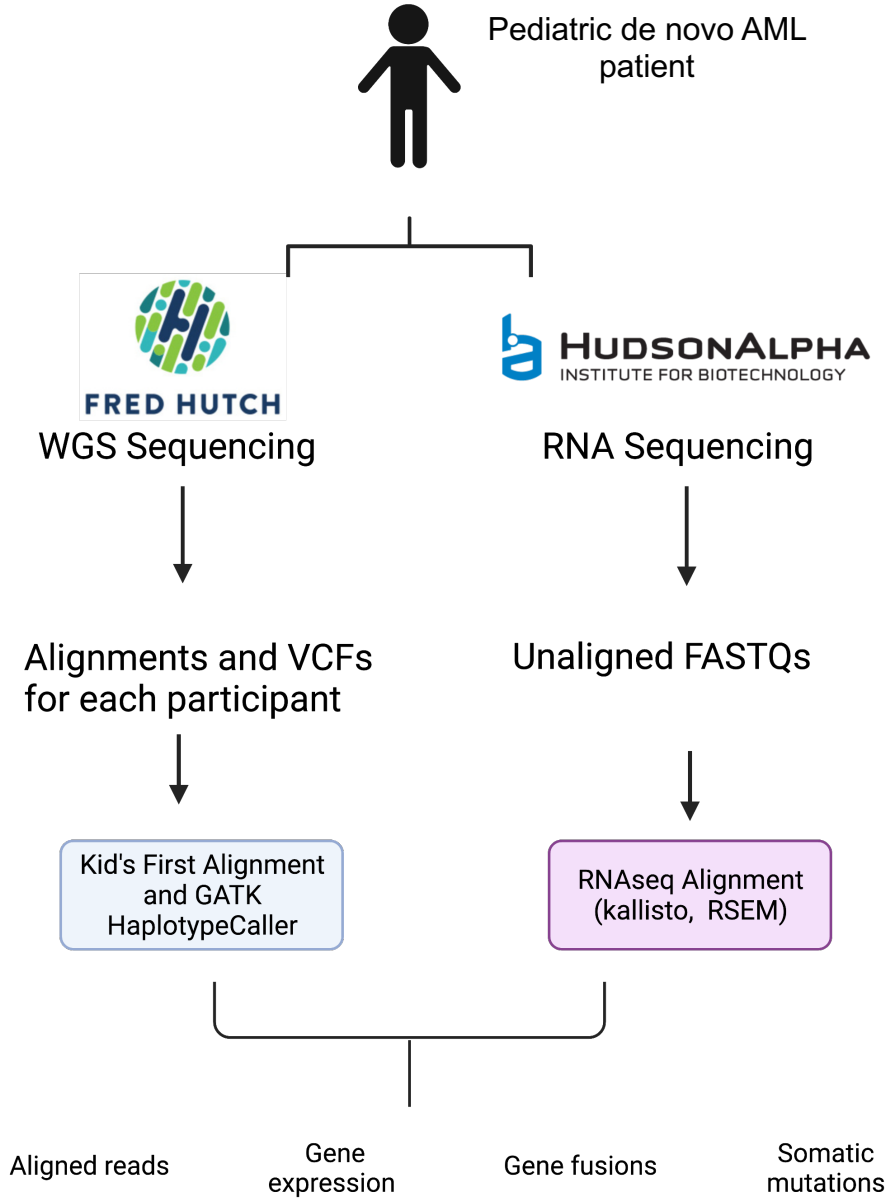
CAVATICA: cloud-based collaboration platform for pediatric cancer data

- Launched in 2016
  - Child of Seven Bridges Genomics x CHOP x Cancer Moonshot *et al*
- 2017 → integrated into NIH Common Fund's Kid's First DRC
- Portable, shareable, reproducible workflows that save time
- Public and private workspaces
- Kid's First DRC
  - > 20k samples, >16k participants, 23 studies
  - WGS, RNA-seq
  - Cancer and structural birth defect studies

CAVATICA

CAVATICA houses raw and harmonized multi-omic data

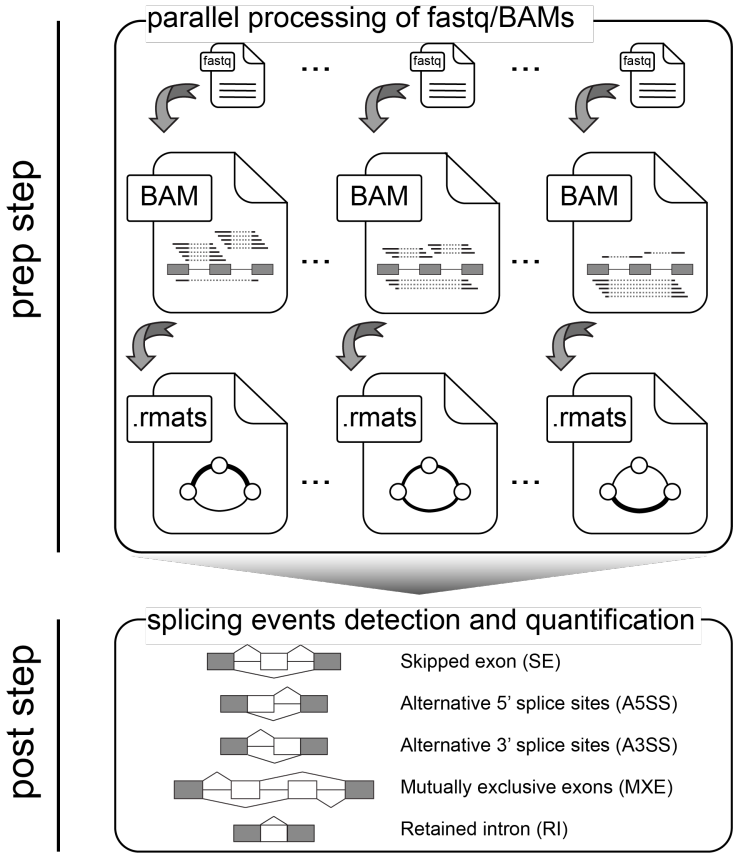
- De novo AML
- Children's Oncology Group Clinical Trial (AAML1031)
- 1113 RNA-seq (aligned and quantified) files
  - Both kallisto and RSEM were used
  - Gene fusions also quantified
- 408 WGS files



CAVATICA

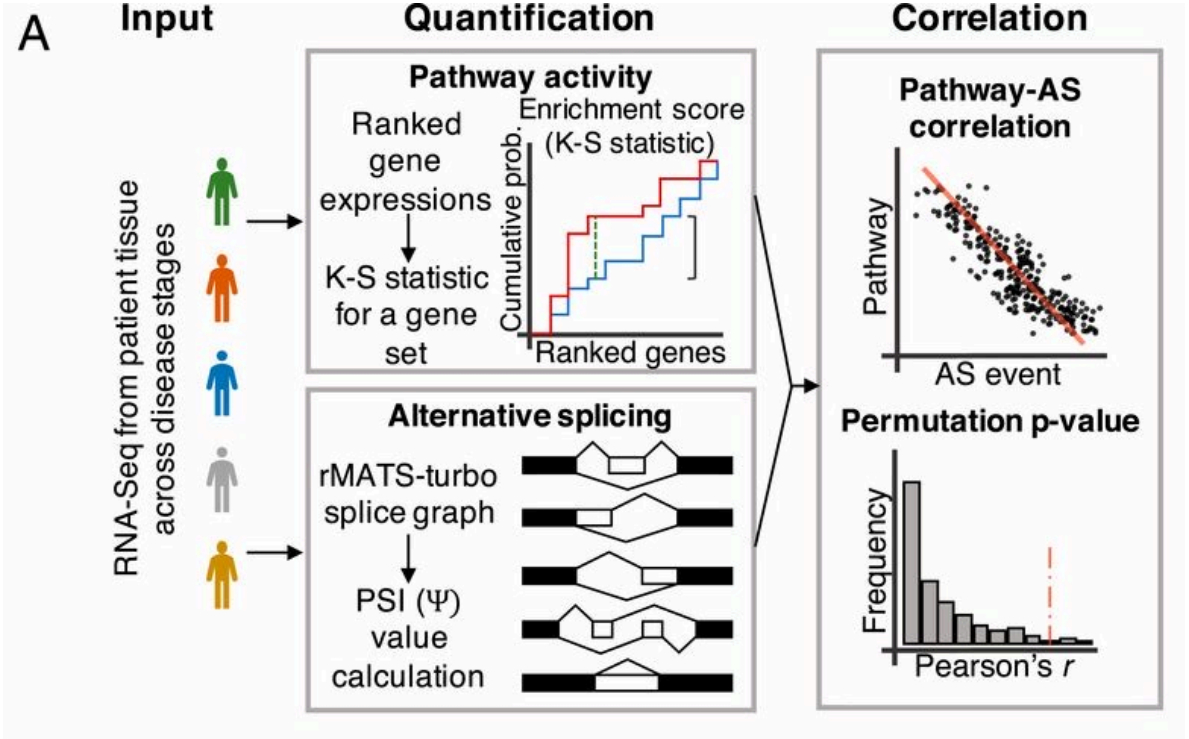
Approach: Xing Lab Software

rMATS-turbo



Yuanyuan Wang

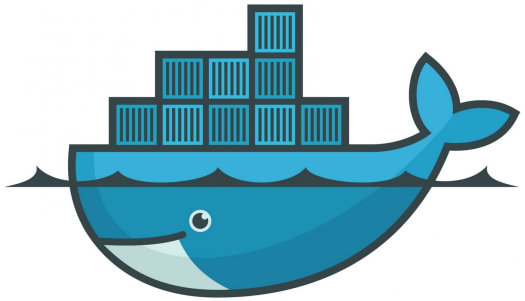
Pathway Enrichment-Guided Activity Study of Alternative Splicing (PEGASAS)



JW Phillips, Yang Pan *et al* 2020

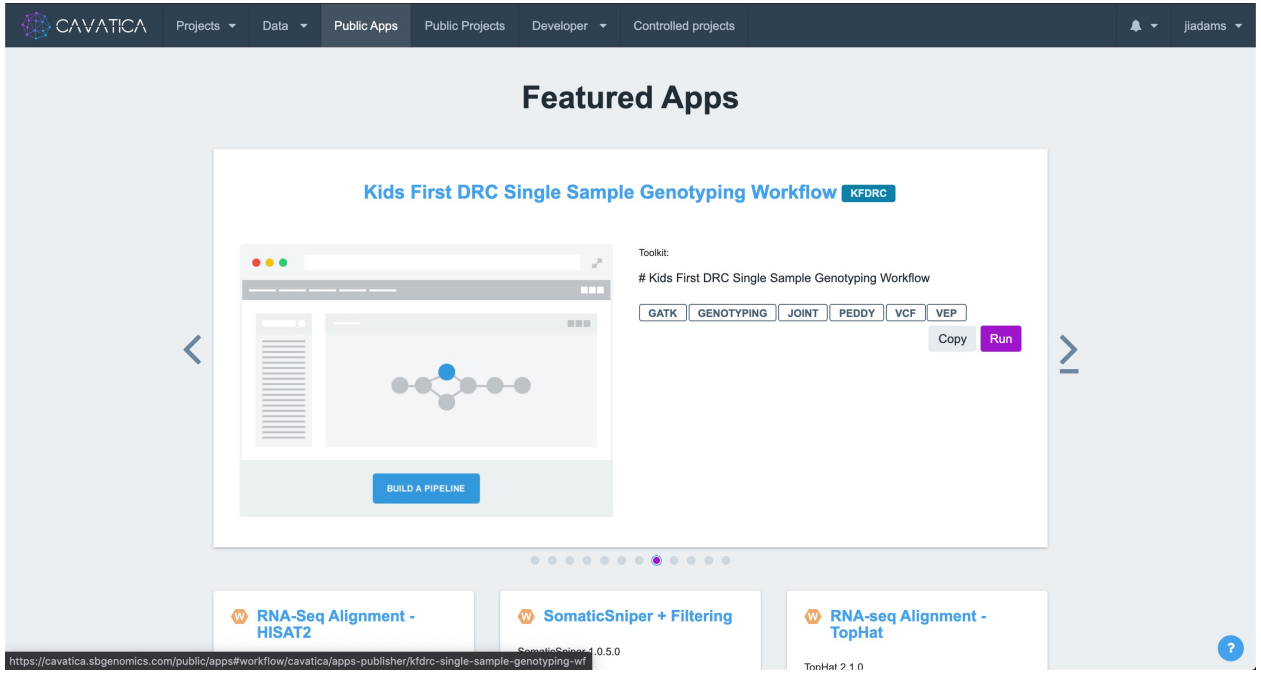
rMATS-cloud?

- Cost effective
- Shareable, reproducible
- Scalable (as more data is uploaded)



ULTRA-FAST SPLICING QUANTIFICATION

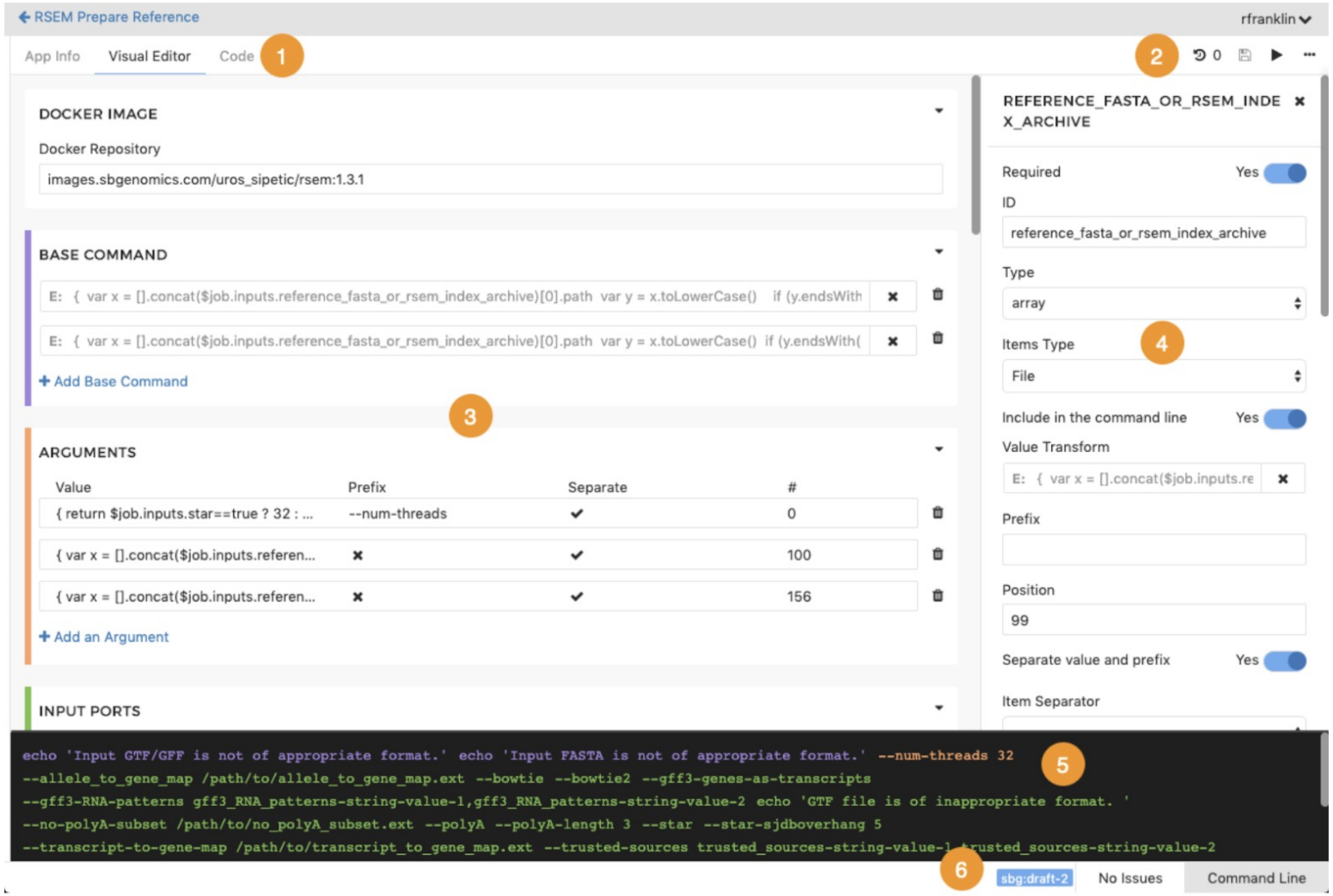
rMATS-TURBO



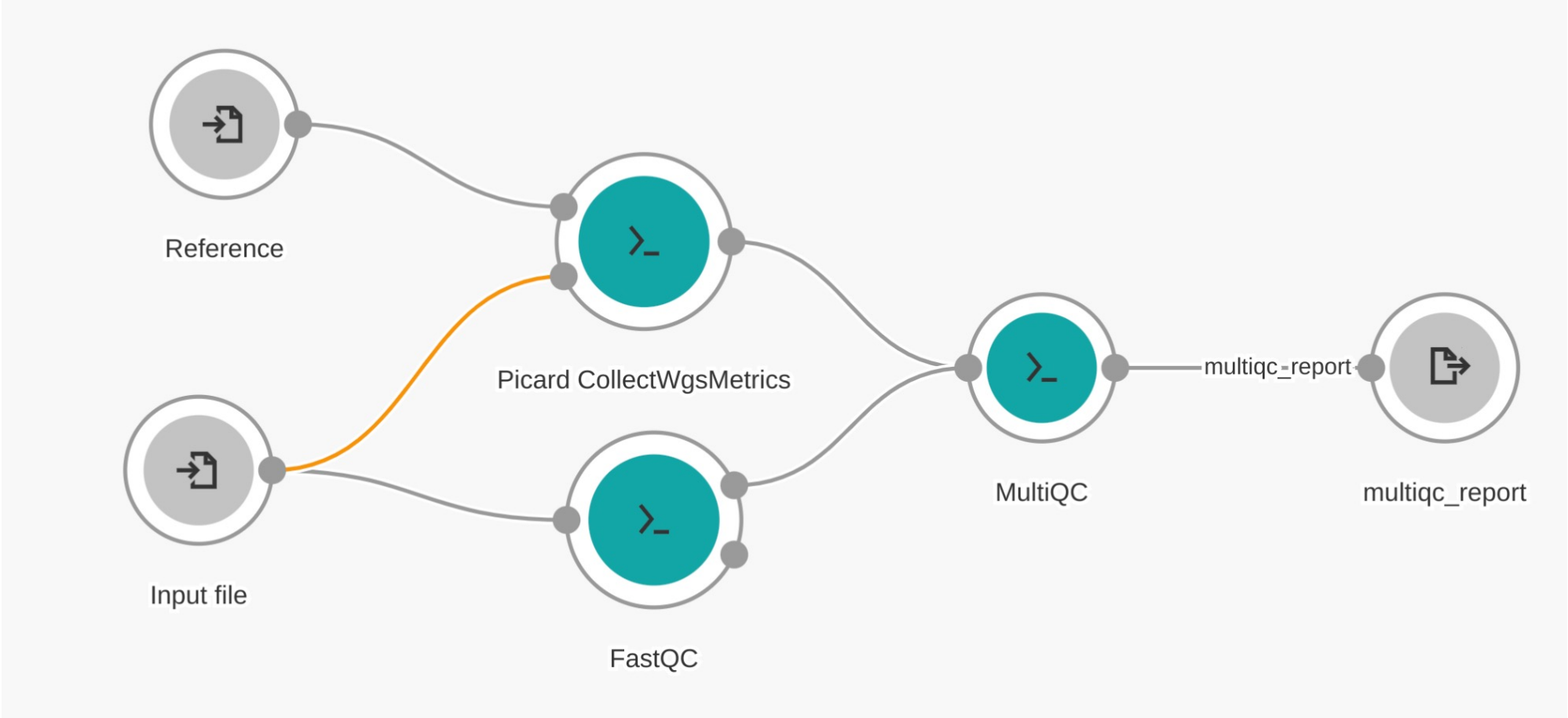
CAVATICA APP

- Apps = collections of tools and workflows
- Housed within a “project”

CAVATICA houses a visual CWL-based tool editor …



… and tools can be connected into workflows



Next Steps

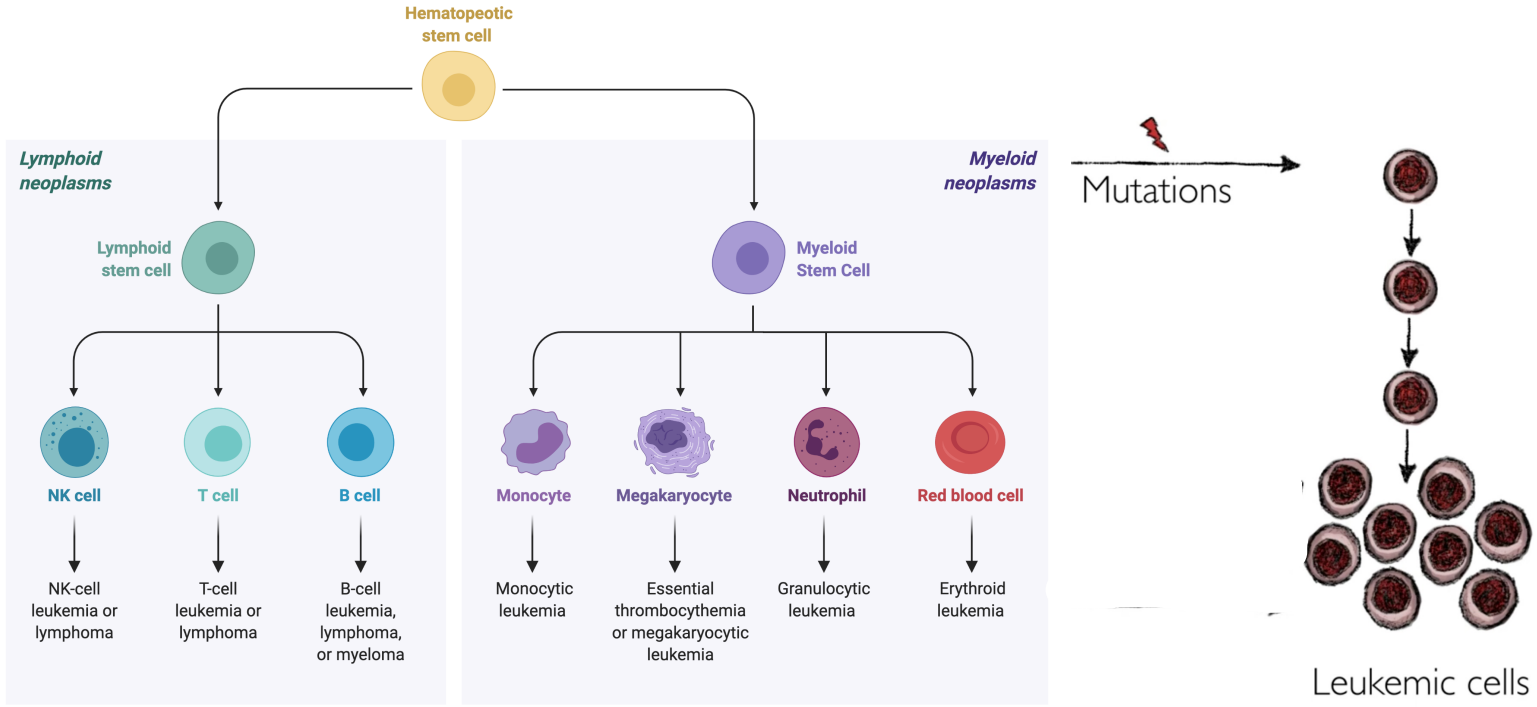
- Testing the CAVATICA tool editor with rMATS-turbo Docker files
- Comprehensive cost analysis of cloud-based vs HPC-based computing for rMATS
- Adaption of other relevant XING lab tools to cloud

- Special thanks to Eric for software development expertise

Supplementary

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

- Affects 25% of children with leukemia
- Accumulation of immature myeloid cells in bone marrow
- Lacks treatment options comparable to ALL
- Treatments based on adult data
  - Median age of AML onset is > 60 years
  - Pediatric = 0-30 y/o

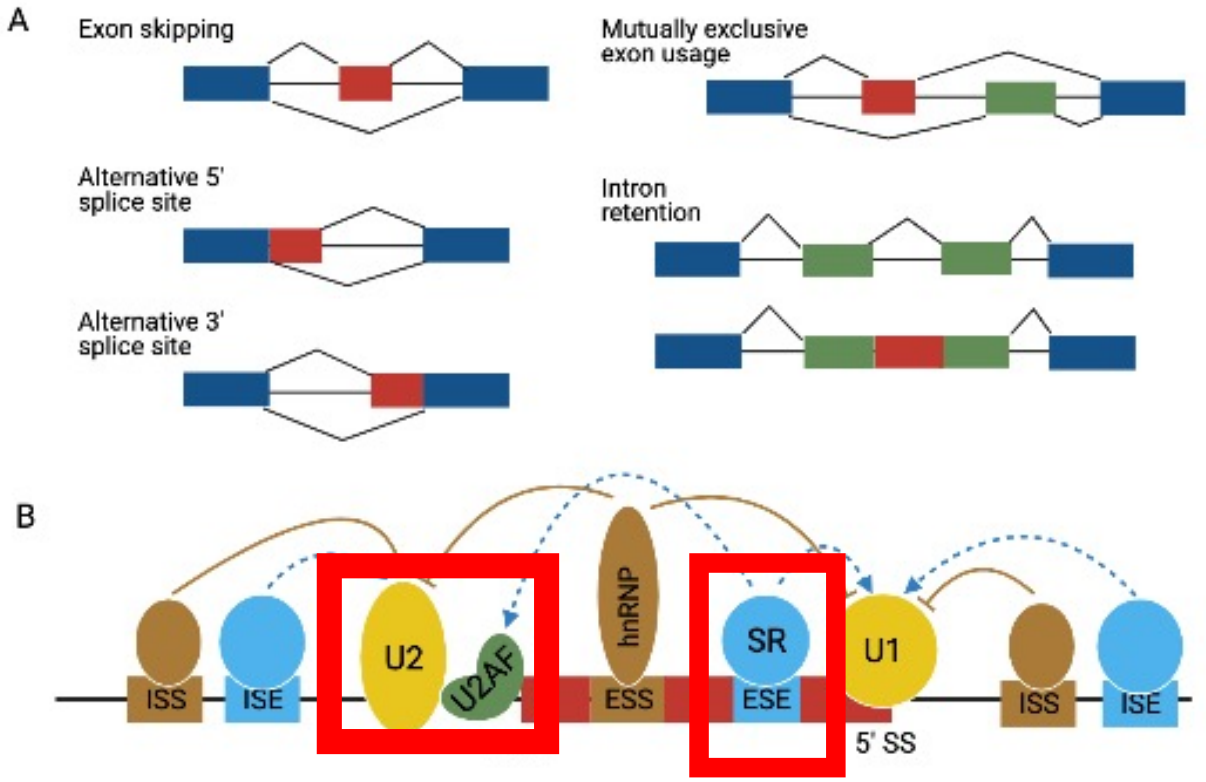


van Galen *et al.,* 2019, Created with BioRender.com

AML treatments needed that preserve healthy myeloid cells

Alternative splicing and correlated molecular pathways could play important roles in distinguishing pediatric from adult AML

- >90% of human genes are alternatively spliced
  - Measure PSI values
- Splicing is associated with every hallmark in cancer
- Splicing signatures correlate AML subtypes (Hsu et al, 2016)
- Splicing-focused therapies have vast potential in AML



U2AF and other splicing factors are often mutated in AML

Navigating a reproducible batch correction pipeline for splicing data

- Transforming PSI values(?)
- Interpretability between expression data
- Compared to other non-SV based approaches
- Visualizing modified variance between corrections
- Consistent splicing event detection metrics
- How robust is splicing to batch effects in general?