Connecting clinical and genetic data on the St. Jude Survivorship Portal

Xin Zhou, Nickhill Bhakta, Jian Wang, Edgar Sioson, Jaimin Patel, Kyla Shelton, Zhaoming Wang, Shaohua Lei, Alexander M. Gout, Carmen L. Wilson, Wendy Leisenring, Smita Bhatia, Yutaka Yasui, Melissa M. Hudson, Gregory T. Armstrong, Leslie L. Robison, Jinghui Zhang

CCSS

Childhood Cancer
Survivor Study

Introduction

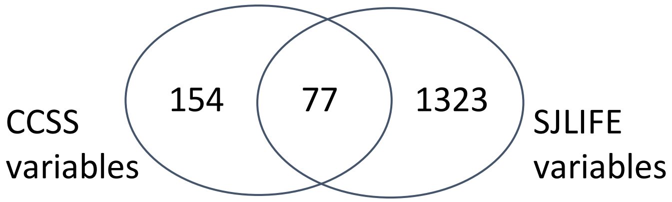
- Inter-departmental collaboration in 2019
  - Dr. Les Robison (Epidemiology and Cancer Control, SJLIFE)
  - Dr. Jinghui Zhang (Computational Biology)
- Built by the Comp Bio data visualization team led by Xin Zhou
  - 2019.10 ASHG → SJLIFE
  - 2020.7 AACR → SJLIFE and CCSS
  - Portal is under active development
- Design goals:
  - Curate and share clinical and genetic data from pediatric cancer survivors
  - Supports interactive real-time data analysis to promote genetic and epidemiology research

Clinical and genetic data

Clinical data

1531 variables about baseline assessments

- Cancer-related Variables
- Demographic Variables
- Self-reported Behavior and Outcome Variables
- Clinically-assessed Variables



7169 participants with clinical data

- CCSS n=2641, SJLIFE n=4528

Germline SNVs from whole-genome sequencing

90 million single nucleotide variations in total

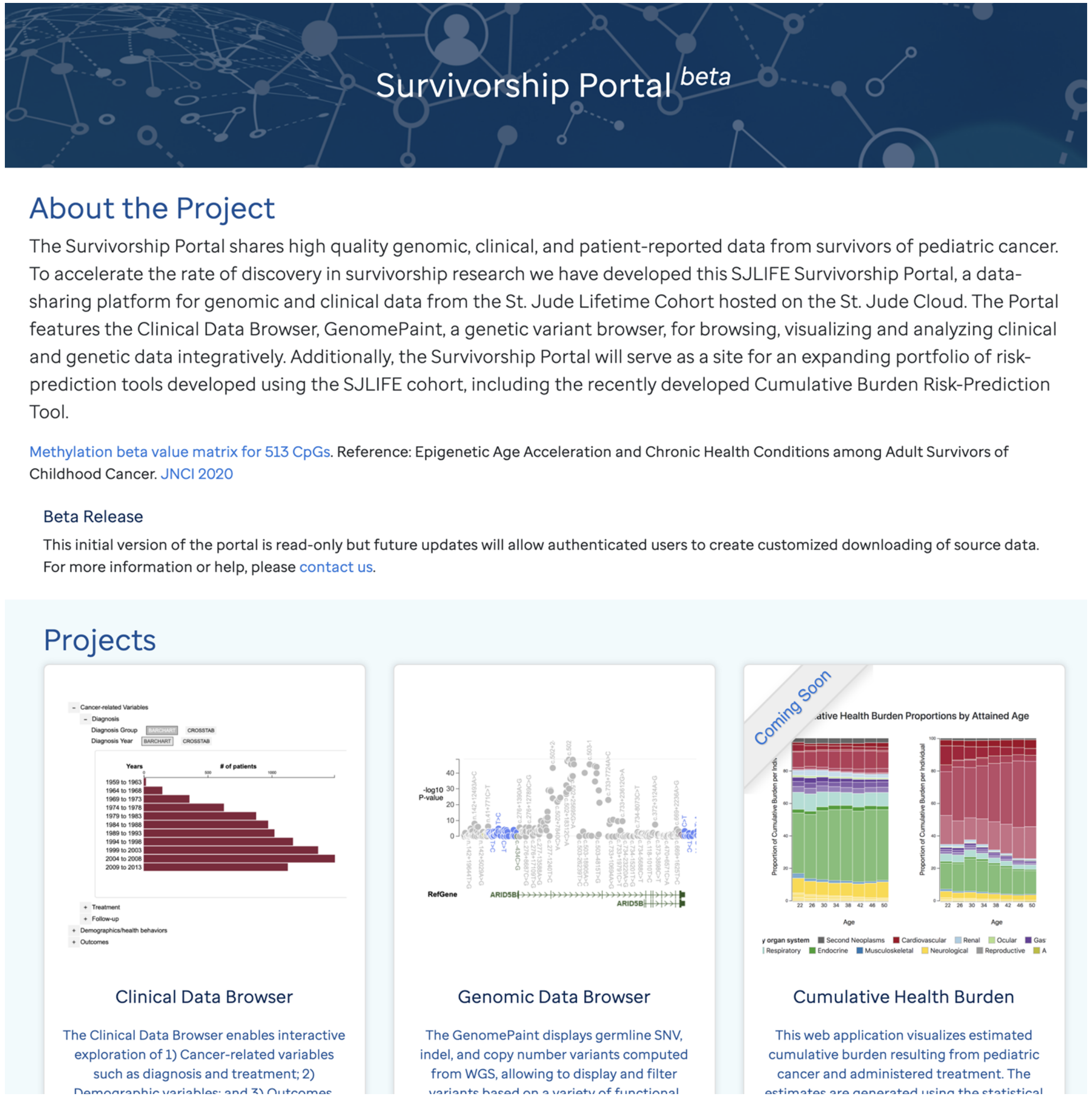
5773 participants with WGS data

- CCSS n=2641, SJLIFE n=3132

The Survivorship Portal

ccss

https://survivorship.stjude.cloud/



Survivorship Portal *beta*

About the Project

The Survivorship Portal shares high quality genomic, clinical, and patient-reported data from survivors of pediatric cancer. To accelerate the rate of discovery in survivorship research we have developed this SJLIFE Survivorship Portal, a data-sharing platform for genomic and clinical data from the St. Jude Lifetime Cohort hosted on the St. Jude Cloud. The Portal features the Clinical Data Browser, GenomePaint, a genetic variant browser, for browsing, visualizing and analyzing clinical and genetic data integratively. Additionally, the Survivorship Portal will serve as a site for an expanding portfolio of risk-prediction tools developed using the SJLIFE cohort, including the recently developed Cumulative Burden Risk-Prediction Tool.

Methylation beta value matrix for 513 CpGs. Reference: Epigenetic Age Acceleration and Chronic Health Conditions among Adult Survivors of Childhood Cancer. JNCI 2020

Beta Release

This initial version of the portal is read-only but future updates will allow authenticated users to create customized downloading of source data. For more information or help, please contact us.

Projects

Clinical Data Browser

The Clinical Data Browser enables interactive exploration of 1) Cancer-related variables such as diagnosis and treatment; 2) Demographic variables; and 3) Outcomes

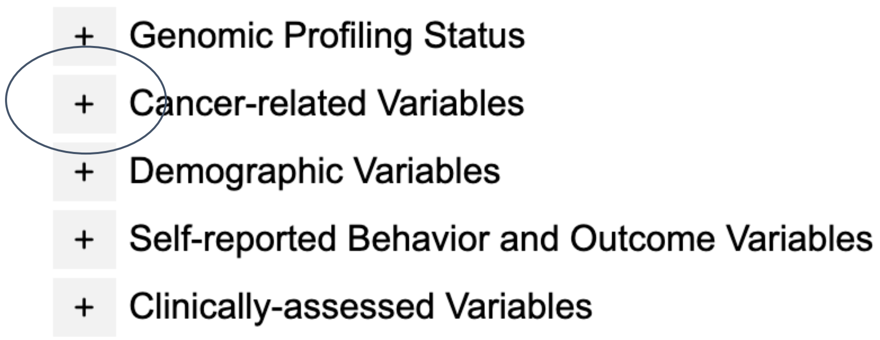
Genomic Data Browser

The GenomePaint displays germline SNV, indel, and copy number variants computed from WGS, allowing to display and filter variants based on a variety of functional

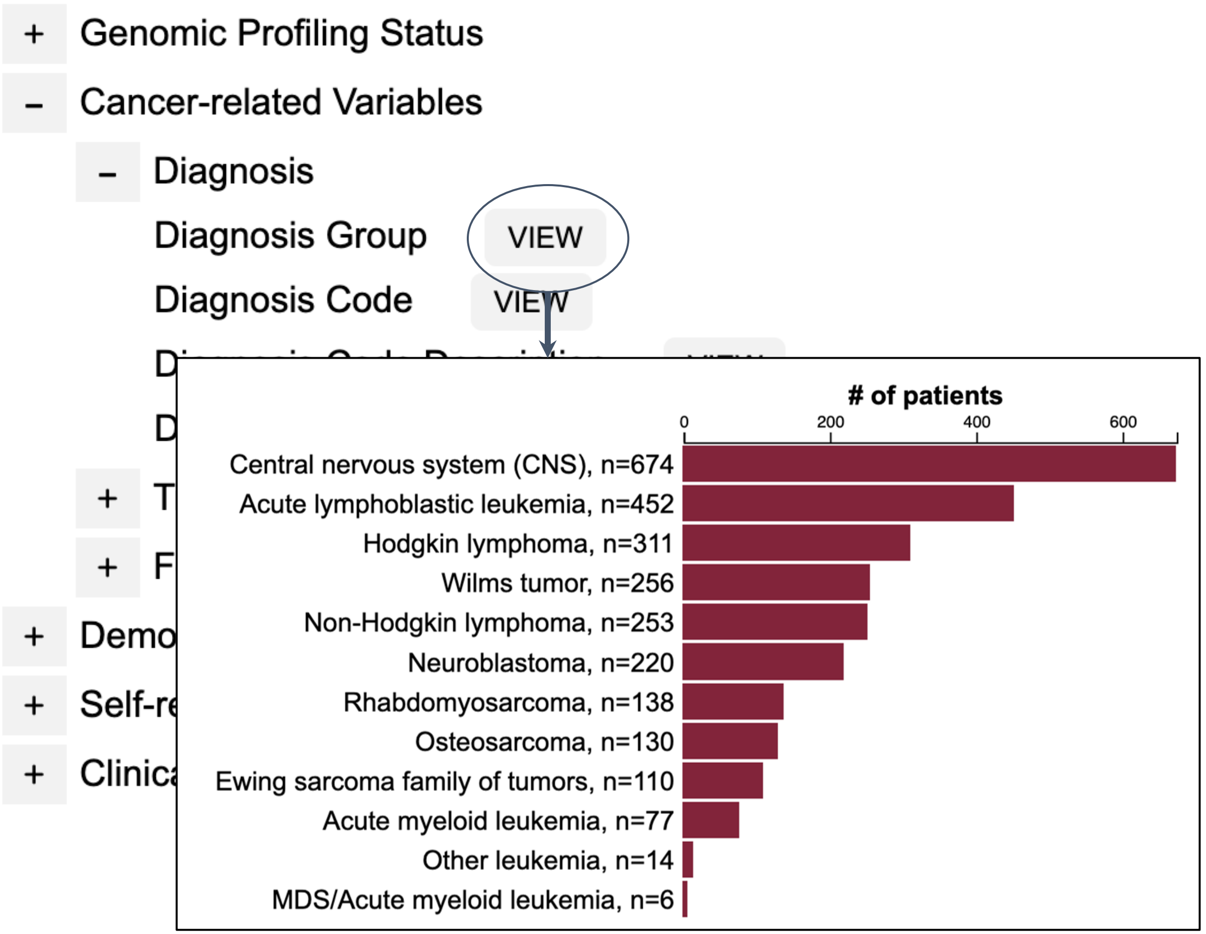
Cumulative Health Burden

This web application visualizes estimated cumulative burden resulting from pediatric cancer and administered treatment. The estimates are generated using the statistical

Browsing the clinical dictionary

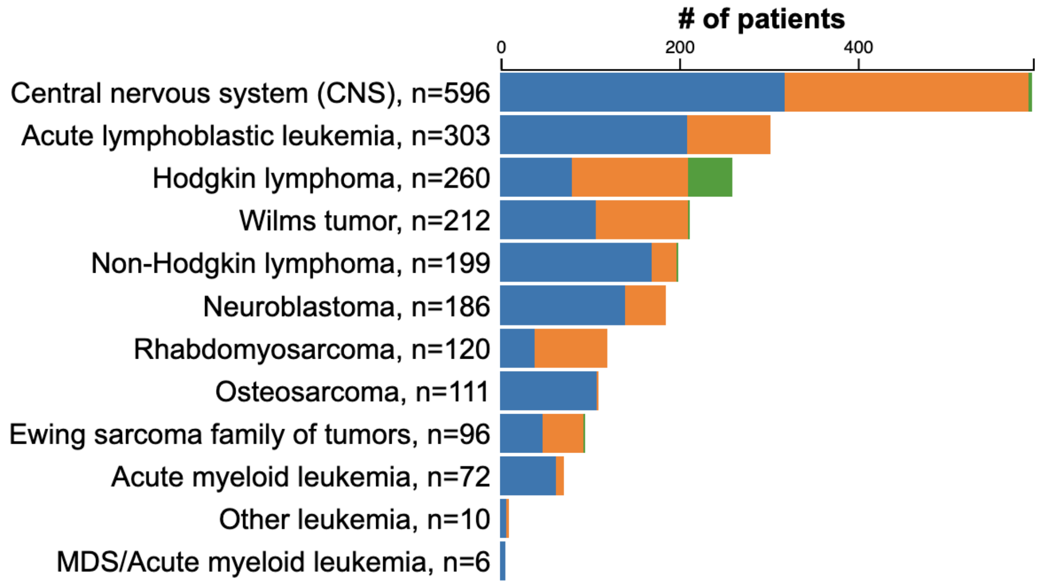


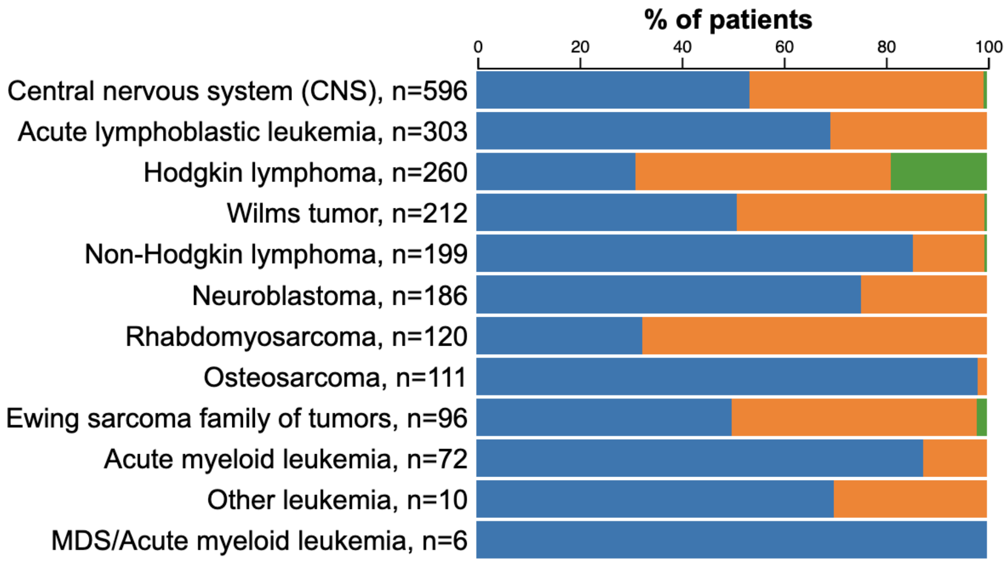
- \+ Genomic Profiling Status
- \+ Cancer-related Variables
- \+ Demographic Variables
- \+ Self-reported Behavior and Outcome Variables
- \+ Clinically-assessed Variables



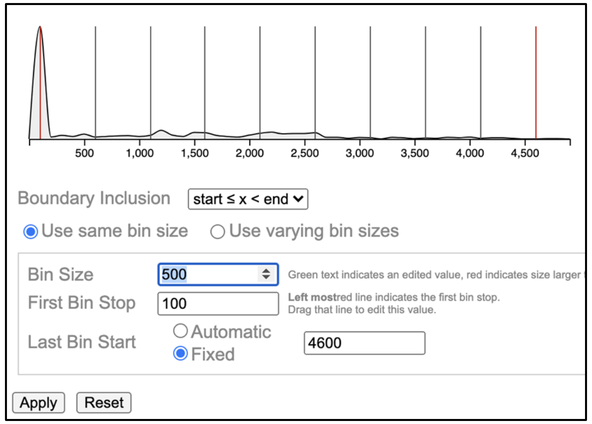
- \+ Genomic Profiling Status
- – Cancer-related Variables
  - – Diagnosis
    - Diagnosis Group VIEW
    - Diagnosis Code VIEW
    - D
    - D
  - \+ T
  - \+ F
- \+ Demo
- \+ Self-re
- \+ Clinica

Cross-tabulating two variables





Average dose to heart + TBI, cGy
🟦 not treated, n=1300
🟥 unknown exposure, n=464
🟪 exposed, dose unknown, n=6
🟧 ≤3000, n=814
🟩 >3000, n=57



Boundary Inclusion: start ≤ x < end

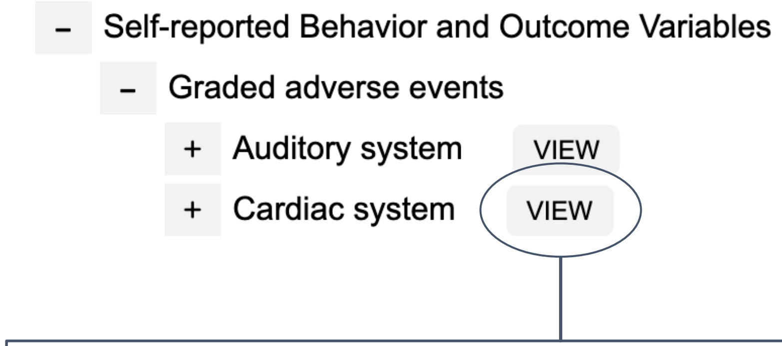
● Use same bin size ○ Use varying bin sizes

Bin Size: 500 — Green text indicates an edited value, red indicates size larger

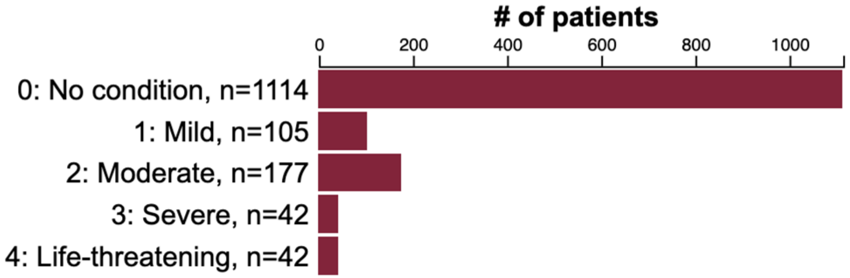
First Bin Stop: 100 — **Left most**red line indicates the first bin stop. Drag that line to edit this value.

Last Bin Start: ○ Automatic ● Fixed 4600

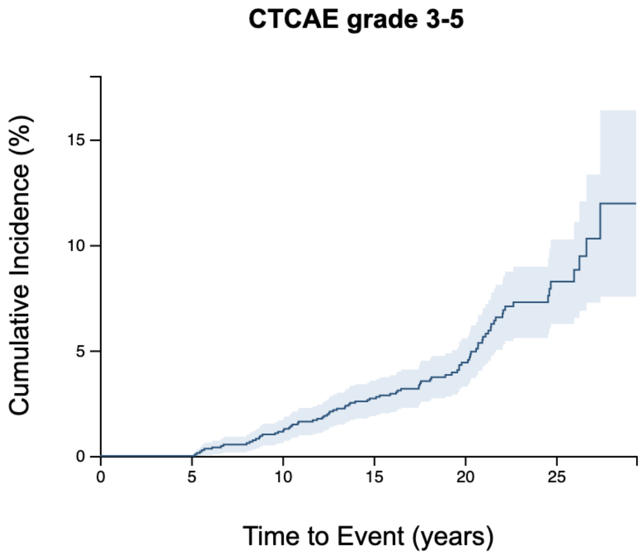
Graded adverse events



Maximum grade of each patient



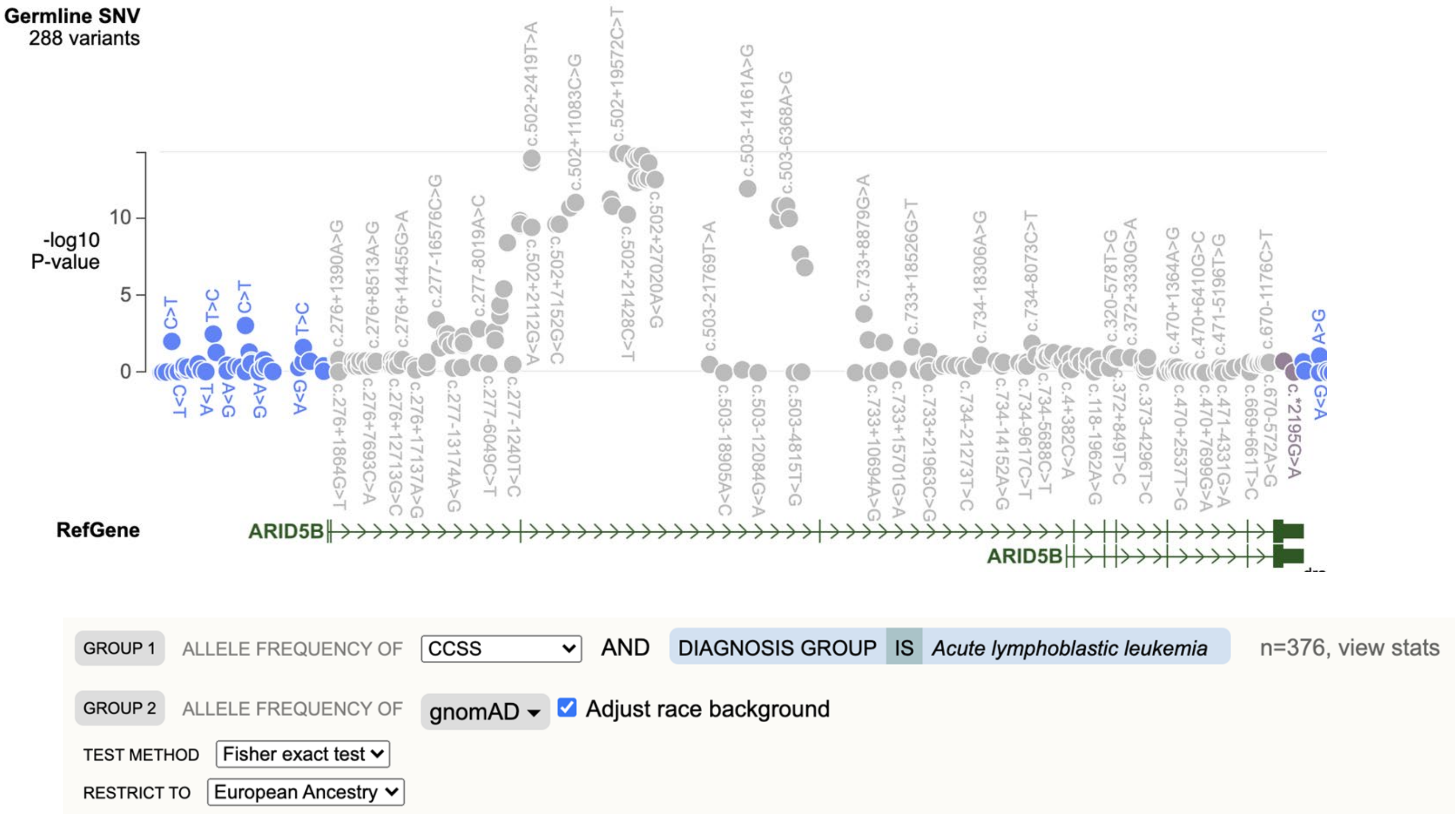
Cumulative incidence



Genetic data browser

ccss

Locus-specific association study



GROUP 1 ALLELE FREQUENCY OF CCSS AND DIAGNOSIS GROUP IS *Acute lymphoblastic leukemia* n=376, view stats

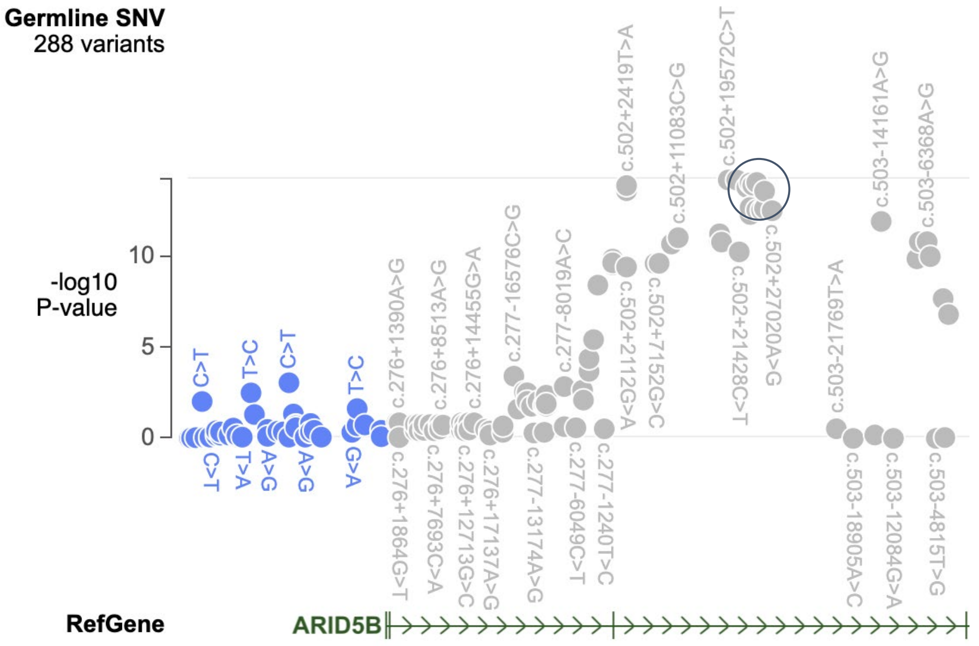
GROUP 2 ALLELE FREQUENCY OF gnomAD ☑ Adjust race background

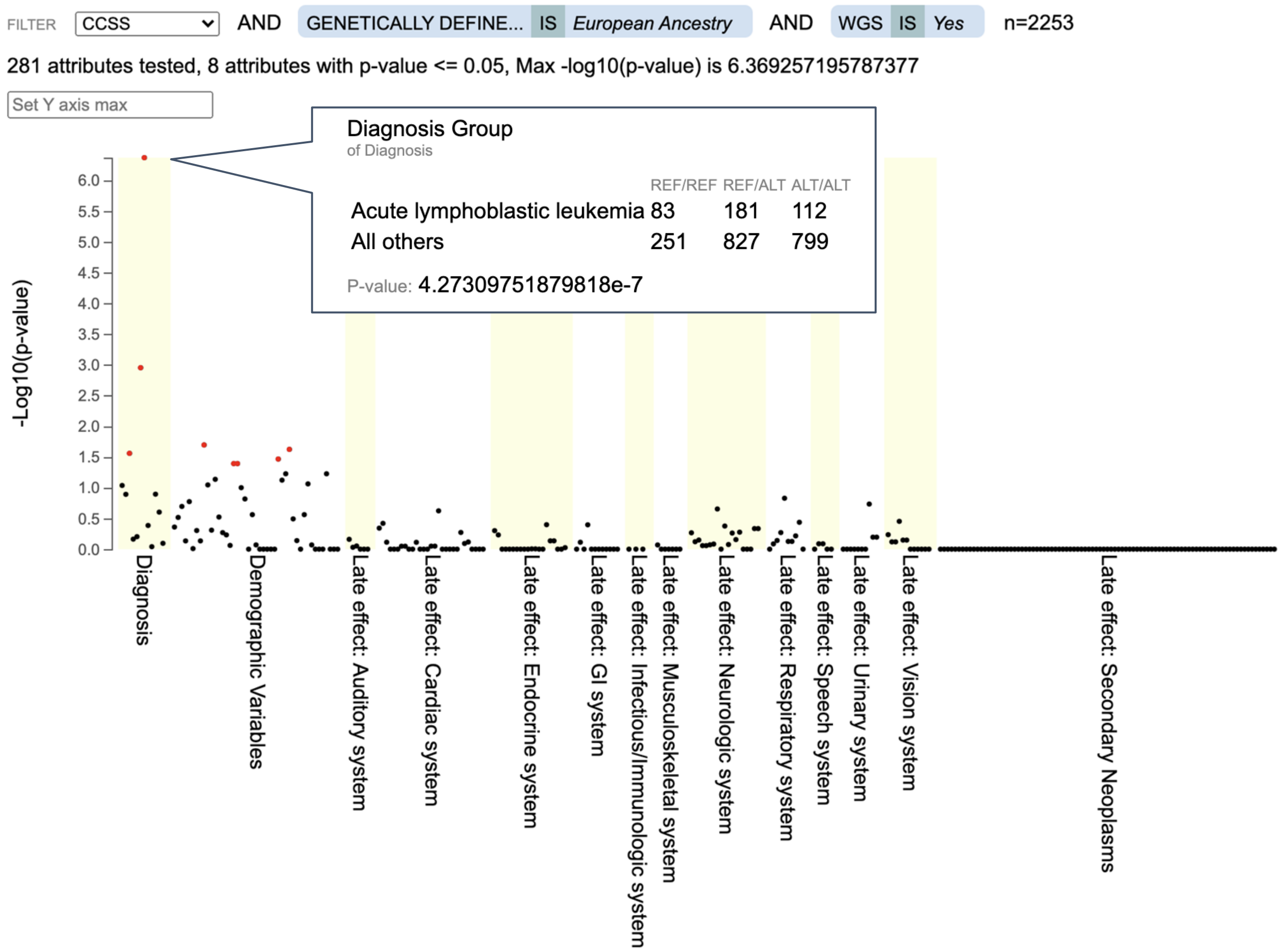
TEST METHOD Fisher exact test

RESTRICT TO European Ancestry

Phewas

CCSS





FILTER CCSS AND GENETICALLY DEFINE... IS *European Ancestry* AND WGS IS *Yes* n=2253

281 attributes tested, 8 attributes with p-value <= 0.05, Max -log10(p-value) is 6.369257195787377

Diagnosis Group
of Diagnosis

	REF/REF	REF/ALT	ALT/ALT
Acute lymphoblastic leukemia	83	181	112
All others	251	827	799

P-value: 4.27309751879818e-7

Planned work

Enhance Opportunities for Research by Providing Access to CCSS Phenotype, Genotype and Outcomes Data Through Development of a Cloud-based Data Analysis Ecosystem

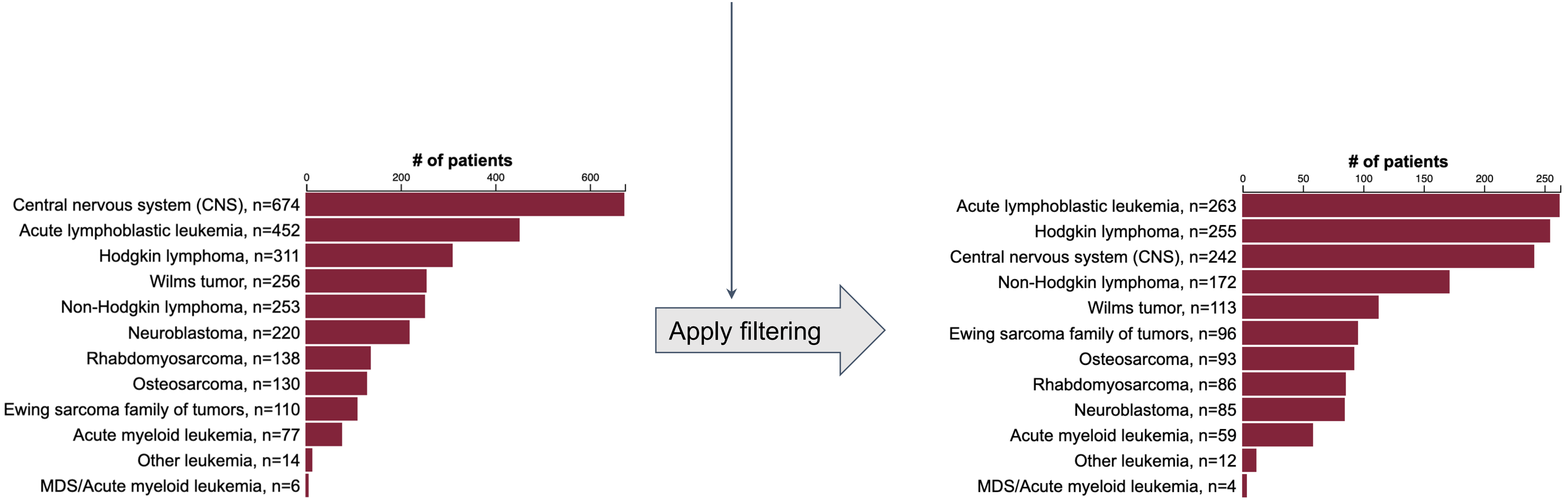
-Includes adding phenotype data for all 25,665 participants and genetic data on an addition 5,900 participants.

- Access to both raw and summary data supported by a searchable ontology-based data dictionary
  - Longitudinal follow-up and detailed lab results
- Conduct cloud-based computing to analyze genomic and clinical data using sophisticated computational/statistical pipelines in real time
  - Polygenic risk score
  - Adjust for covariates in association/regression analysis
- Download customized data sets through controlled access
- Access to a portfolio of intuitive, field-tested visualization tools to explore data and apply complex risk prediction modelling in a secure cloud environment

Filtering

CCSS

GENETICALLY DEFINE... IS *European Ancestry* AND (AVERAGE DOSE TO HEART + ... IS $x \geq 0$ OR CUMULATIVE ANTHRACYCLINE... IS $x \geq 0$)



Planned work

CCSS

- **Bring the complete CCSS cohort to St. Jude Survivorship portal**
- Clinical data
    - Longitudinal follow-up and detailed lab results
- Genetic data
    - WGS indel genotype calls, HLA typing, telomere length, haplotype
- Analysis features
    - Polygenic risk score
    - Adjust for covariates in association/regression analysis
    - Data download