The St. Jude Survivorship Portal Links Whole-Genome Genetic Data with Clinical Therapy and Outcome Phenotypes for 7302 Pediatric Cancer Survivors

Clinical and basic research addressing the long-term outcomes of the increasing and high-risk population of pediatric cancer survivors requires large cohorts with high quality information. We have assembled the largest group of pediatric cancer survivors to date with comprehensive clinical characterization and germline whole-genome sequencing (WGS), and made these data available through the Survivorship Portal on St. Jude Cloud (https://survivorship.stjude.cloud).

The Portal contains data from 7302 survivors from two large studies, including 4402 from the St. Jude Lifetime Cohort Study (SJLIFE) with clinically-assessed outcome phenotypes, and 2900 from the Childhood Cancer Survivor Study (CCSS) with self-reported outcome phenotypes. High-quality germline variants and genotype calls from WGS were generated and curated by an in-house pipeline with corrected indel alleles and read counts. Germline variants of individual survivors are linked with standardized phenotypes described by >300 variables, spanning cancer-related data including diagnosis, length of follow-up, treatment (cumulative doses of chemotherapy, region-specific radiation therapy doses, surgery), demographic characteristics, selected health behaviors, and long-term outcomes (severity-graded chronic health conditions including second cancers). In addition, clinically-relevant genetic variables including ancestry admixture, HLA alleles, leukocyte telomere length, and blood type have been computed from WGS. Both phenotypic and genetic variables are represented by the Dictionary Browser that allows to quickly identify variables of interest, view summary graphics, cross-tabulate variables by categories and test for significant frequency deviation with chi-squared tests. Using GenomePaint on the Portal, investigators can navigate to a locus of interest and explore the presence and frequencies of variants in the cohort, filter variants with multiple criteria including LD r^2 values, and identify DNA binding motif change for non-coding variants. Combining GenomePaint with Dictionary Browser, the real-time association analysis allows to identify trait-associated variants at a locus, through the definition of traits, covariates, and inclusion/exclusion criteria using the Dictionary Browser. Future implementation includes supporting copy number and structural variations, characterization of pharmacogenetic diplotypes, gene-level rare variant analysis, polygenic score, survival analysis, and data download and session management enabled by user login. We envision this cohort with high quality phenotypic and genetic information, together with an enabling software platform co-developed with multidisciplinary principal investigators, will accelerate the discovery in both survivorship research and human genetics in general.

2583 characters excluding spaces.

Title and body combined cannot exceed 2600 characters excluding spaces.

List of authors:

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