Rare structural variants and their associations with risks of late effects and mortality in long-term survivors of childhood cancer

SJLIFE Working Groups: Genomics/Genetics; Global/Epi/Biostats
CCSS Working Groups: Genetics; Chronic Diseases; Subsequent neoplasm

Investigators

Jennifer French	jenn.french@stjude.org
Kendrick Li	kendrick.li@stjude.org
Achal Neupane	achal.neupane@stjude.org
Chang Li	chang.li@stjude.org
Michael Betti	michael.betti@stjude.org
Zimo Zhang	zimo.zhang@stjude.org
Monica Gramatges	gramatge@bcm.edu
Lucie Turcotte	turc0023@umn.edu
Kumar Srivastava	kumar.srivastava@stjude.org
Eric Chow	ericchow@uw.edu
Kevin Oeffinger	kevin.oeffinger@duke.edu
Cindy Im	imcindy@umn.edu
Wendy Leisenring	wleisenr@fredhutch.org
Matt Ehrhardt	matt.ehrhardt@stjude.org
Yutaka Yasui	yutaka.yasui@stjude.org
Kevin Krull	kevin.krull@stjude.org
Kirsten K. Ness	kiri.ness@stjude.org
Melissa M. Hudson	melissa.hudson@stjude.org
Gregory T. Armstrong	greg.armstrong@stjude.org
Yadav Sapkota	yadav.sapkota@stjude.org

Background

Due to advances in treatment of childhood cancers, today the five-year survival for childhood cancer in the US exceeds 85%[1; 2]. However, these survivors are at an increased risk for a variety of late effects of childhood cancer, including subsequent neoplasms and pulmonary, auditory, endocrine or reproductive, cardiac and neurocognitive outcomes[3]. Studies have found that 62% - 75% survivors of childhood cancer have at least one chronic health condition (CHC) or adverse event, with roughly 24.6% having 5 or more adverse events and 27.5% having severe or life-threatening conditions[4-6]. Childhood cancer survivors also have an increased risk of premature death due to primary cancer recurrence, subsequent neoplasms, and cardiovascular disease[7].There has been a call from researchers to better identify high-risk survivors who may benefit from interventions to reduce the burden of late effects in survivors of childhood cancer[8].

While significant progress has been made in identifying risk factors for late effects of childhood cancer, such as treatment exposures and genetic predisposition, much remains to be learned. Genetic studies conducted thus far have identified single nucleotide variants (SNVs) associated with risk of multiple late effects[9-11]. However, to our knowledge, no studies have investigated the role of germline structural variants in risk of late effects in long-term survivors of childhood cancer. Structural variants (SVs) are 50 bp in size or larger and include inversions, translocations, deletions, insertions, interspersed elements, tandem repeats and duplications[12-15]. While the functional effect of most SVs remains unknown, they do have a large impact on phenotypes and variation in phenotypes[12]. It is suggested that SVs play a role in many health conditions, such as cognition, obesity, and cancer. SVs can have a greater impact than SNVs on gene expression and also affect genome stability and oncogenesis[12; 15]. SVs can alter driver genes, tumor suppressors, and regulatory elements are responsible for many driver mutations in cancer[13; 16]. This proposal aims to examine the association between rare germline SVs and risk of subsequent neoplasms, cardiovascular disease, neurocognition, and mortality among survivors of childhood cancer from the Childhood Cancer Survivor Study (CCSS) and St. Jude Lifetime Cohort (SJLIFE).

Specific Aims

Aim 1a: To characterize rare germline SVs (large indels, copy number gains, losses, copy neutral loss of heterozygosity, etc.) among five-year survivors of childhood cancer.

Aim 1b: To compare prevalence of the SVs in survivors [N~5,000 for SJLIFE and N=2,939 for CCSS] with prevalence from apparently healthy controls from the general population (e.g., gnomAD non-cancer controls, and All of Us controls, and SJLIFE community controls [N~450]).

Aim 1c: To assess if the prevalence of specific SVs (carrier status) differs in survivors with corresponding late effects compared to survivors without these late effects.

- As a proof of concept, we will preliminarily focus on subsequent cancers, cardiovascular conditions (specifically atrial fibrillation, coronary artery disease [CAD], and cardiomyopathy), and neurocognitive conditions.
- Since we are considering rare SVs, we will first run analysis broadly on these conditions as a group (e.g., organ system). For any grouped condition in which a pattern is observed, we will run analysis on individual late effects within the group (e.g., CAD, cardiomyopathy).
- We will also consider prioritizing rare SVs overlapping protein-coding genes, regulatory elements, or constrained genomic regions to improve interpretability and power.

Aim 2a: To evaluate the association between carrying individual SVs (and the overall burden) and risk of incident corresponding CHCs (see Aim 1c), accounting for demographic and cancer treatment exposures.

Aim 2b: To evaluate the association between carrying individual SVs (and the overall burden) and laboratory measured continuous variables related to CHCs evaluated in Aims 1c and 2a (e.g., ejection fraction, systolic blood pressure) in SJLIFE.

Aim 3: To examine association between structural variants (carrier status or global burden) with all-cause and cause-specific (subsequent cancer, cardiovascular disease [i.e., those conditions specified in Aim 1c]) mortality risks.

Methods

Study Population

We will include all SJLIFE (N~5,000) and CCSS (N=2,939) 5-year survivors of childhood cancer with whole genome sequencing (WGS) data available. All analyses will be conducted independently in each cohort and also within subgroups stratified by sex, cancer treatment exposures, and genetic ancestry.

In SJLIFE, for Aims 1a, 1b and 3, we will include all survivors with WGS data available; for Aims 1c, 2a and 2b, we will include survivors who underwent both SJLIFE core evaluation and whole genome sequencing.

Outcome variables

Since this is an unbiased analysis of all potential phenotypes, similar to a PheWAS approach, we will consider all continuous variables and categorical outcomes in the SJLIFE. Analyses in CCSS will only include categorical outcomes.

- Subsequent neoplasms (any and specific neoplasm, benign and malignant)
  - Limited to those with greater than 50 cases
- Cardiovascular chronic health conditions (CHCs), specifically atrial fibrillation, CAD, and cardiomyopathy) as determined by CTCAE v5 grading system[17]
  - Limited to CHCs with greater than 50 cases
  - Grade 0 vs 2+; grade 0 vs 3+
- Neurocognitive measures (only in SJLIFE)
- All continuous measured variables (first non-missing measurement in SJLIFE only) related to CHCs
  - Limited to those with <50% missingness

Sociodemographic/clinical variables

- Age at last contact

- Sex
- Primary cancer diagnosis
- Age at primary cancer diagnosis
- Cancer treatment exposures within 5 years of primary cancer diagnosis
    - Any RT (yes/no)
    - Field-specific RT with total body irradiation (TBI) (yes/no) and dose (cranial, neck, chest, abdomen, pelvic, arm, leg)
    - Chemotherapy
        - Any chemotherapy (yes/no)
        - Alkylating agents: cyclophosphamide-equivalent dose
        - Anthracyclines: doxorubicin-equivalent dose
        - Epipodophyllotoxins dose
        - Methotrexate (by route) dose
        - Platinum dose

Genetic data

For SJLIFE survivors and CCSS survivors with primary cancers diagnosed between 1987-1999, we will use quality-controlled WGS data completed with Illumina HiSeq x10 or NovaSeq platforms (30x mean coverage)[18]. In both CCSS and SJLIFE, genetic ancestry (European, African and East Asian) of survivors will be determined using a K-means clustering approach implemented in Admixture[19], on the basis of genotype data of an independent set of common autosomal SNVs and the 1000 Genomes Project samples as ancestral populations. Survivors will then be grouped into European (%European >80%), African (%African >60%) and Others based on the estimated ancestry proportions.

Structural variants

It is estimated that every individual on average carries up to 27,000 SVs or 16Mb of SVs[12]. However, identifying these variants can be challenging. Here we will use WGS, which will allow for identification of SVs outside of the exome. One important limitation we will face is the use of short-read sequencing, which will result in missing some potentially important SVs. Previous studies have found that 68% of SVs identified using long-read sequencing would have been missed in short-read sequencing[20]. This will be the first study investigating SVs and any findings will provide an important insight into the impact of SVs and improved understanding of how they contribute to the risk of late effects in survivors of childhood cancer.

We will follow the established methods used by the NHLBI TopMed Program and outlined in previous research[15] for detecting SVs, starting with use of the Parliament2 pipeline[21]. Briefly, Parliament2 pipeline executes any combination of six different SV detection methods, such as Breakdancer, Breakseq, CNVnator, Delly, Lumpy, and Manta, to generate candidate SV events. The SV calls from these programs can then be joined together and filtered to include only SVs with a quality score ≥3 and minor allele count ≥5. SV genotypes will be converted to a biallelic format for ease of structural variant association analyses. Rare SVs will be defined as those occurring with allele frequency <0.01 in gnomAD[22] population data (overall and European and African ancestry-specific). We will use AnnotSV[23], SVScore[24], and StrVCTVRE[25] to prioritize SVs with functional or biological plausibility.

Statistical analysis

Analysis will be conducted separately in SJLIFE and CCSS and stratified by ancestry groups outlined above. Results will be combined using meta-analysis approaches. Survivors participating in both studies will be included in SJLIFE and will be excluded from CCSS.

For Aim 1a, we will summarize the prevalence of rare SVs, overall and by SV type in survivors of childhood cancer. Specific SVs that will be evaluated include deletions, duplications, copy number variants, insertions, inversions, translocations, copy-neutral loss of heterozygosity and complex SVs. These SVs will also be compared in SJLIFE community controls. For Aim 1b, we will calculate the prevalence of SVs in survivors from CCSS and SJLIFE, SJLIFE community controls, All of Us controls and gnomAD non-cancer controls (available

at https://databrowser.researchallofus.org/ and https://gnomad.broadinstitute.org/). Average proportions and standard errors across population in cancer survivors and non-cancer controls will be calculated based on inverse-variance weighted log odds (a small number, e.g., 0.001 will be added to all cells for zero-cell adjustment). They will further be compared using chi-squared tests. For Aim 1c, the prevalence of rare SVs will be summarized in CCSS and SJLIFE childhood cancer survivors with or without late effects. Unadjusted effect sizes of SVs on the outcomes (mean difference for continuous outcomes and log odds ratios for binary outcomes) in CCSS and SJLIFE will be summarized separately and meta-analyzed using the inverse-variance weighted estimates (for binary outcomes, a small number e.g., 0.001 will be added to all cells for zero-cell adjustment). The meta-analyzed effect size estimates will be compared to zero using chi-squared test. We will also conduct similar analysis for specific categories of late effects, e.g., cardiovascular disease, pulmonary diseases, secondary neoplasms and early mortality.

For Aims 2 and 3, overall survival will be estimated using the Kaplan-Meier method where comparison of survival by SV carrier status will be based on the log-rank test, using time since 5-year post primary cancer diagnosis as the time scale. The cumulative incidence of CHCs in each disease group of interest and cause-specific mortality accounting for competing risks will be estimated, where differences by SV carrier status will be assessed using the Gray K-sample test[26]. We will use Cox proportional hazard regression models[27] to evaluate associations between SV carrier status and all-cause mortality, and to evaluate the associations between SV carrier status and time-to-event outcomes of interest (CHCs for Aim 2 and all-cause or cause-specific deaths for Aim 3), adjusting for sex, age at primary cancer diagnosis, cancer treatment decade, cancer treatments and genetic ancestry. We will use a multivariable linear model to evaluate associations between SV carrier status and first available continuous outcomes of interest, adjusting for sex, age at childhood cancer diagnosis, age at measurement, cancer treatment decade, cancer treatments and genetic ancestry. Multivariable linear mixed-effects model with survivor-specific random intercepts will also be performed to take account of the repeated measures. For initial analyses (screening purpose), we will use any chemotherapy, any radiation exposure and any surgery, as cancer treatments covariates. For any outcomes with preliminary evidence of association, we will more carefully adjust for specific cancer treatment exposures, either based on previous studies or consultation with clinical scientists.

Analyses described above will also be conducted in subgroups stratified by sex, genetic ancestry and cancer treatment exposures, as appropriate. For ancestry-specific analysis, the regression models will be further adjusted for the first three genotype-based principal components derived of an independent set of common autosomal SNVs. Separate analyses will be conducted in SJLIFE and CCSS. Meta-analysis of results across cohorts/ancestry groups/studies will use fixed-effects inverse variance-weighted method and heterogeneity will be examined using the Cochran's Q test[28] and the I^2 index[29].

Impact statement:

While more than 80% of childhood cancer survivors achieve 5-year survival, survivors face increased risks for late effects of childhood cancer that contribute to long-term morbidity and premature mortality. The proportion of the risk attributable to genetic predispositions remains unclear, but by studying the contribution of germline structural variants, we can gain a deeper understanding of genetic predispositions. We will be describing the prevalence and risk of late effects for individuals carrying germline structural variants for the first time in childhood cancer survivors, and in one of the largest cohorts of childhood cancer survivors. Study findings may guide follow-up care, helping to identify survivors most susceptible to late effects in an effort to prevent or delay onset.

Example tables and figures

Table 1. Characteristics of childhood cancer survivors from SJLIFE and CCSS.

Characteristics	SJLIFE	CCSS
Age at primary cancer diagnosis (years)		
Sex		
Male		

Female		
Radiation Therapy		
Any radiation		
Cranial		
Neck		
Chest		
Abdomen		
Pelvis		
Extremity		
Total body irradiation		
Chemotherapy		
Any		
Anthracyclines		
Platinum		
Epipodophyllotoxins		
Methotrexate		
High dose		
Intrathecal		

Table 2. Prevalence of structural variants among survivors participating in SJLIFE compared to community controls and the general population.

Class	**SJLIFE**	**SJLIFE community controls**	**TOPMed**	**gnomad**	**All of Us**
Large indels	Prev (std er)				
Duplications					
CNV					
Loss of heterozygosity					
Inversions					
Translocations					

Table 3. Structural variants significantly associated with CHCs.

Variant	**Gene**	**AF**	**HR**	***P***

Figure 1. Number of structural variants identified among childhood cancer survivors in SJLIFE and CCSS. (boxplot)

Figure 2. Cumulative incidence of various outcomes in carriers and non-carriers of structural variants, organized by class of structural variants.

Figure 3. Manhattan plot of structural variants associated with CHCs.

Figure 4. Forest plots of adjusted hazard ratios of various outcomes in carriers of structural variants compared to non-carriers, stratified by class of structural variants.

- Panels stratified by sex, ancestry, and cancer treatment exposures

References

1. Armstrong, G. T., Chen, Y., Yasui, Y., et al. (2016). Reduction in Late Mortality among 5-Year Survivors of Childhood Cancer. *N Engl J Med*, *374*(9), 833-842. https://doi.org/10.1056/NEJMoa1510795
2. Howlader, N., Noone, M., Krapcho, M., et al. (2021). *SEER Cancer Statistics Review, 1975-2018, National Cancer Institute*. https://seer.cancer.gov/csr/1975_2018/
3. Hudson, M. M., Ness, K. K., Gurney, J. G., et al. (2013). Clinical ascertainment of health outcomes among adults treated for childhood cancer. *JAMA*, *309*(22), 2371-2381. https://doi.org/10.1001/jama.2013.6296
4. Robison, L. L., & Hudson, M. M. (2014). Survivors of childhood and adolescent cancer: life-long risks and responsibilities. *Nat Rev Cancer*, *14*(1), 61-70. https://doi.org/10.1038/nrc3634
5. Oeffinger, K. C., Mertens, A. C., Sklar, C. A., et al. (2006). Chronic health conditions in adult survivors of childhood cancer. *N Engl J Med*, *355*(15), 1572-1582. https://doi.org/10.1056/NEJMsa060185
6. Geenen, M. M., Cardous-Ubbink, M. C., Kremer, L. C., et al. (2007). Medical assessment of adverse health outcomes in long-term survivors of childhood cancer. *JAMA*, *297*(24), 2705-2715. https://doi.org/10.1001/jama.297.24.2705
7. Fidler, M. M., Reulen, R. C., Winter, D. L., et al. (2016). Long term cause specific mortality among 34 489 five year survivors of childhood cancer in Great Britain: population based cohort study. *BMJ*, *354*, i4351. https://doi.org/10.1136/bmj.i4351
8. Bhatia, S., Armenian, S. H., Armstrong, G. T., et al. (2015). Collaborative Research in Childhood Cancer Survivorship: The Current Landscape. *J Clin Oncol*, *33*(27), 3055-3064. https://doi.org/10.1200/JCO.2014.59.8052
9. Chen, C., Qin, N., Wang, M., et al. (2023). Cancer germline predisposing variants and late mortality from subsequent malignant neoplasms among long-term childhood cancer survivors: a report from the St Jude Lifetime Cohort and the Childhood Cancer Survivor Study. *Lancet Oncol*, *24*(10), 1147-1156. https://doi.org/10.1016/S1470-2045(23)00403-5
10. Richard, M. A., Mostoufi-Moab, S., Rathore, N., et al. (2022). Germline Genetic and Treatment-Related Risk Factors for Diabetes Mellitus in Survivors of Childhood Cancer: A Report From the Childhood Cancer Survivor Study and St Jude Lifetime Cohorts. *JCO Precis Oncol*, *6*, e2200239. https://doi.org/10.1200/PO.22.00239
11. Sapkota, Y., Ehrhardt, M. J., Qin, N., et al. (2022). A Novel Locus on 6p21.2 for Cancer Treatment-Induced Cardiac Dysfunction Among Childhood Cancer Survivors. *J Natl Cancer Inst*, *114*(8), 1109-1116. https://doi.org/10.1093/jnci/djac115
12. Pokrovac, I., & Pezer, Z. (2022). Recent advances and current challenges in population genomics of structural variation in animals and plants. *Front Genet*, *13*, 1060898. https://doi.org/10.3389/fgene.2022.1060898
13. Quigley, D. A., Dang, H. X., Zhao, S. G., et al. (2018). Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell*, *174*(3), 758-769 e759. https://doi.org/10.1016/j.cell.2018.06.039
14. Sudmant, P. H., Rausch, T., Gardner, E. J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75-81. https://doi.org/10.1038/nature15394

15. Wheeler, M. M., Stilp, A. M., Rao, S., et al. (2022). Whole genome sequencing identifies structural variants contributing to hematologic traits in the NHLBI TOPMed program. *Nat Commun*, *13*(1), 7592. https://doi.org/10.1038/s41467-022-35354-7
16. Cosenza, M. R., Rodriguez-Martin, B., & Korbel, J. O. (2022). Structural Variation in Cancer: Role, Prevalence, and Mechanisms. *Annu Rev Genomics Hum Genet*, *23*, 123-152. https://doi.org/10.1146/annurev-genom-120121-101149
17. Hudson, M. M., Ehrhardt, M. J., Bhakta, N., et al. (2017). Approach for Classification and Severity Grading of Long-term and Late-Onset Health Events among Childhood Cancer Survivors in the St. Jude Lifetime Cohort. *Cancer Epidemiol Biomarkers Prev*, *26*(5), 666-674. https://doi.org/10.1158/1055-9965.EPI-16-0812
18. Sapkota, Y., Cheung, Y. T., Moon, W., et al. (2019). Whole-Genome Sequencing of Childhood Cancer Survivors Treated with Cranial Radiation Therapy Identifies 5p15.33 Locus for Stroke: A Report from the St. Jude Lifetime Cohort Study. *Clin Cancer Res*, *25*(22), 6700-6708. https://doi.org/10.1158/1078-0432.CCR-19-1231
19. Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, *19*(9), 1655-1664. https://doi.org/10.1101/gr.094052.109
20. Ebert, P., Audano, P. A., Zhu, Q., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, *372*(6537). https://doi.org/10.1126/science.abf7117
21. Zarate, S., Carroll, A., Mahmoud, M., et al. (2020). Parliament2: Accurate structural variant calling at scale. *Gigascience*, *9*(12). https://doi.org/10.1093/gigascience/giaa145
22. Karczewski, K. J., Francioli, L. C., Tiao, G., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434-443. https://doi.org/10.1038/s41586-020-2308-7
23. Geoffroy, V., Herenger, Y., Kress, A., et al. (2018). AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, *34*(20), 3572-3574. https://doi.org/10.1093/bioinformatics/bty304
24. Ganel, L., Abel, H. J., FinMetSeq, C., et al. (2017). SVScore: an impact prediction tool for structural variation. *Bioinformatics*, *33*(7), 1083-1085. https://doi.org/10.1093/bioinformatics/btw789
25. Sharo, A. G., Hu, Z., Sunyaev, S. R., et al. (2022). StrVCTVRE: A supervised learning method to predict the pathogenicity of human genome structural variants. *Am J Hum Genet*, *109*(2), 195-209. https://doi.org/10.1016/j.ajhg.2021.12.007
26. Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 1141-1154.
27. Kom, E. L., Graubard, B. I., & Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *American Journal of Epidemiology*, *145*(1), 72-80.
28. Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101-129.
29. Ioannidis, J. P., Patsopoulos, N. A., & Evangelou, E. (2007). Heterogeneity in meta-analyses of genome-wide association investigations. *Plos one*, *2*(9), e841.