I. Title: Exome sequencing to discover genetic variants that predispose childhood cancer survivors to the development of subsequent neoplasms

II. CCSS Working groups: Genetics (primary), Second Malignancies (secondary), and Epidemiology/Biostatistics (secondary)

Investigators, by primary area of expertise:

Epidemiology
Lindsay M. Morton, PhD, National Cancer Institute (mortonli@mail.nih.gov)
Smita Bhatia, MD, MPH, University of Alabama at Birmingham (sbhatia@peds.uab.edu)
Gregory T. Armstrong, MD, MSCE, St. Jude Children's Research Hospital (greg.armstrong@stjude.org)
Leslie L. Robison, PhD, St. Jude Children's Research Hospital (les.robison@stjude.org)
Todd M. Gibson, PhD, St. Jude Children's Research Hospital (todd.gibson@stjude.org)
Joseph P. Neglia, MD, MPH, University of Minnesota (jneglia@umn.edu)
Amy Berrington de Gonzalez, PhD, National Cancer Institute (berringtona@mail.nih.gov)

Genetics
Stephen J. Chanock, MD, National Cancer Institute (chanocks@mail.nih.gov)
Margaret A. Tucker, MD, National Cancer Institute (tuckerp@mail.nih.gov)
Louise Strong, PhD, MD Anderson Cancer Center (lstrong@mdanderson.org)

Biostatistics
Joshua N. Sampson, PhD, National Cancer Institute (sampsonjn@mail.nih.gov)
Wendy M. Leisenring, ScD, Fred Hutchinson Cancer Research Center (wleisenr@fhcrc.org)
Yutaka Yasui, PhD, University of Alberta (yyasuiua@gmail.com)
Ting-Huei Chen, PhD, National Cancer Institute (chent5@mail.nih.gov)

Dosimetry
Marilyn Stovall, PhD, M.D. Anderson Cancer Center (mstovall@mdanderson.org)
Susan A. Smith, MPH, M.D. Anderson Cancer Center (sasmith@mdanderson.org)
Rita E. Weathers, MS, M.D. Anderson Cancer Center (rweather@mdanderson.org)

III. Background and rationale:

Although the prognosis for childhood cancer survivors has improved dramatically in the last several decades, at least 20% of childhood cancer survivors develop a subsequent neoplasm within 30 years following diagnosis.[1,2] Radiotherapy and various chemotherapies are major contributors to subsequent neoplasm risk.[3,4] However, some children who receive high treatment doses do not develop subsequent neoplasms, and other children develop multiple subsequent neoplasms even with fewer treatment exposures. These observations suggest that other factors such as genetic susceptibility may play an important role in subsequent neoplasm risk. Individuals with certain hereditary disorders such as ataxia telangiectasia have marked sensitivity to the effects of radiation, but less is known about genetic susceptibility to radiation-related carcinogenesis beyond the context of these rare disorders,[5] and very little is known about genetic susceptibility to chemotherapy-related carcinogenesis.[6]

We have recently completed the first large-scale study of genetic susceptibility to subsequent neoplasms in the Childhood Cancer Survivor Study (CCSS). Using the genome-wide association study (GWAS) approach, we agnostically evaluated over 4.1 million single nucleotide polymorphisms (SNPs) across the genome to identify heretofore unsuspected genomic regions that may predispose childhood cancer survivors to the development of subsequent neoplasms. Genotyping was successfully completed on 5324 childhood cancer survivors of European descent and another 415 of non-European ancestry. Initial

analyses have identified promising SNPs that appear to be associated with risk of specific subsequent neoplasms only in the context of certain treatment exposures, and other SNPs that appear to be associated with risk independent of treatment exposures. Replication of these findings in independent populations is ongoing.

We now propose to expand the genomics data in CCSS using exome sequencing to identify other types of genetic variants (*e.g.*, rare variants, multi-allelic substitutions, insertions, and deletions) that may predispose childhood cancer survivors to the development of subsequent neoplasms. The plausibility that such variants (not detectable using a GWAS array) may be related to subsequent neoplasm risk is supported by their presence in individuals with hereditary disorders that confer radiation sensitivity.[5] Laboratory work will be performed by the Cancer Genomics Research Laboratory (CGR) of the National Cancer Institute's Division of Cancer Epidemiology and Genetics (DCEG) using a reliable, high-throughput pipeline for exome sequencing of germline DNA. The proposed study will further advance understanding of genetic susceptibility to subsequent neoplasms in childhood cancer survivors as well as elucidate potential mechanisms of radiation- and chemotherapy-related carcinogenesis. The study results also have the potential to directly impact clinical decision-making in terms of treatment and long-term follow-up of childhood cancer survivors.

IV. Objective: Conduct exome sequencing of childhood cancer survivors from CCSS.

Specific aims:

1) Identify genetic variants associated with the development of subsequent neoplasms among childhood cancer survivors. We aim to identify variants that modify the effects of radiotherapy and chemotherapy on risk of subsequent neoplasms as well as those that are independent of treatment exposures.
2) Identify genetic variants associated with the risk of childhood cancer.
3) Develop a resource of genetic data that can be used by investigators to conduct secondary analyses of more specific hypotheses related to the aims listed above or to conduct analyses of other outcomes (*e.g.*, cardiovascular events, other sequelae).

Research hypotheses (corresponding to the aims above):

We hypothesize that:

1) Children who develop multiple primary cancers have a high probability of an inherited predisposition to cancer. This inherited variation may contribute to carcinogenesis directly, or it may alter the biological response of normal cells to DNA damage and immunosuppression from ionizing radiation and/or chemotherapeutic agents and thereby alter cancer susceptibility.
2) Children who develop a childhood cancer have an inherited predisposition to cancer.
3) Inherited genetic variation also contributes to the spectrum of adverse outcomes observed among childhood cancer survivors.

V. Analysis Framework:

Outcomes of interest

The primary outcome of interest for Aim #1 is the occurrence of subsequent neoplasms, including malignant (invasive and *in situ* cancers) and certain benign tumors (*e.g.*, meningioma). Certain analyses will combine all radiation-related subsequent neoplasms, defined as neoplasms that have been shown to be highly radiosensitive in this and other study populations (including cancers of the breast, central nervous system [CNS], thyroid, gastrointestinal tract; sarcoma; acute leukemia; and non-melanoma skin cancer [NMSC]) to increase statistical power to evaluate radiation-specific genetic variants. To provide approximate counts for this analysis, characteristics of the individuals of European descent from the GWAS are presented in Tables 1-2.

Subject population

For the present study, eligible individuals from the CCSS population must:

- have available (non-missing) information on radiotherapy and chemotherapy.
- not have a history of allogeneic bone marrow transplantation.
- have at least 500 ng of DNA. Samples will be selected in the following order to maximize data quality (based on pilot data from other studies using various source materials):
    - genomic DNA (gDNA) derived from blood or Oragene samples
    - gDNA derived from buccal mouthwash samples
    - whole-genome amplified DNA (wgaDNA) derived from blood or Oragene samples
    - wgaDNA derived from buccal mouthwash samples.

As described below, primary analyses will be based on the CCSS cohort, comparing individuals who develop subsequent neoplasms to those who do not. Analyses also may utilize a control set of individuals known to be cancer-free as of age 55 years. These individuals will be identified from DCEG's Population Exome Controls, a set of primarily Caucasian individuals who are already being sequenced in the same laboratory as part of parallel studies.

Key variables

Requested data include basic demographic information (*e.g.*, sex, race/ethnicity, date of birth), information on all primary neoplasm diagnoses (including site, histology, date of diagnosis, and microscopic confirmation), and all available treatment data (for the first primary cancer as well as any subsequent neoplasms, recognizing that data were collected systematically only for those treatments occurring within 5 years of the first primary cancer). For subsequent neoplasms, additional detailed information on tumor location is requested as well. Finally, data on body-mass index, hormonal factors, and family history are requested for consideration of potential confounding (but will not be used as outcomes).

Analyses that consider radiotherapy dose will include more detailed radiation dosimetry data (see below). Analyses focused on chemotherapy will primarily consider broad classes of chemotherapy, including alkylating agent (including platinum), anthracycline, antimetabolite, or epipodophyllotoxin-based chemotherapy. Chemotherapy analyses that consider doses within a class of chemotherapeutic agents will use a scored variable that reflects the dose distributions across multiple agents within that class.[7] Secondary analyses of chemotherapy will consider other approaches for combining agents within a particular class (*e.g.*, summing the quartile of the dose distribution for each agent) or will consider actual doses of specific agents (mg/m^2).

Radiation dosimetry

Radiation doses for analysis will rely on the region-based dose estimates generated by collaborating physicists at M.D. Anderson Cancer Center, using standard methodology.[8] Briefly, data on individual patients' radiotherapy fields and tumor dose already have been collected from radiotherapy and other medical records. These data form the foundation for estimating doses to specific locations in the body using a custom-designed dose program, based on measurements in water and anthropomorphic phantoms constructed of tissue-equivalent material.

For these analyses, radiation exposure of interest will be the dose to the body region where each subsequent neoplasm occurred, including: four segments of the brain, pituitary, other head, neck, thyroid, chest, four quadrants of the breast (only for patients who received a mantle field treatment), abdomen, kidney, pelvis, ovary, testes, whole spine, arms, legs, and bone marrow. Doses take into account direct in-beam contributions to that region based on field type and location (assuming standard blocking to protect normal tissue), beam energy, and prescribed dose.

Exome sequencing:

We aim to sequence all individuals in CCSS with an available biospecimen. Most samples will be sequenced using 200ng of input DNA, but for samples with lower available DNA quantities, we can sequence with lower input (50ng) to conserve the specimen. For many individuals, samples of appropriate concentration are already stored at CGR following the GWAS. We will work closely with CCSS investigators to ensure that the project includes as many individuals from the cohort as possible to facilitate the creation of a resource of genomic data. Any new samples that are received will undergo standard sample handling procedures, including evaluation of DNA quantity and quality. Quality will be assessed by the Applied Biosystems Identifiler® assay, which performs multiplex PCR (16 amplicons) in a single tube and requires high-quality double-stranded DNA (though the DNA may be fragmented). Availability of good data for at least 13 of these 16 markers is correlated with adequate sequencing performance.

DNA Preparation: For each sample, 200 ng genomic DNA will be sheared with a Covaris E210 Sonicator (Covaris, Inc., Woburn, MA, USA) to an average size of 300 bp. An adapter-ligated library will be prepared with the KAPA Hyper Prep Kit (KAPA Biosystems, Wilmington, MA) using Bioo Scientific NEXTflex™ DNA Barcoded Adapters (Bioo Scientific, Austin, TX, USA) according to KAPA-provided protocol.

Pre-Hybridization LM-PCR: Genomic DNA sample libraries will be amplified pre-hybridization by ligation-mediated PCR consisting of one reaction containing 20 µL library DNA, , 25 uL 2x KAPA HiFi HotStart ReadyMix, and 5µL 10x Library Amplification Primer Mix(includes two primers whose sequences are: 5'-AATGATACGGCGACCACCGA-3' and 5'-CAAGCAGAAGACGGCATACGA-3'). PCR cycling conditions will be as follows: 98˚C for 45 seconds (s), followed by 6 cycles of 98˚C for 15 s, 60˚C for 30 s, 72˚C for 30 s. The last step will be an extension at 72˚C for 1 minute. The reaction will be kept at 4˚C until further processing. The amplified material will be cleaned with Agencourt AMPure XP Reagent (Beckman Coulter Inc, Brea, CA, USA) according to the KAPA-provided protocol. Amplified sample libraries will be quantified using Quant-iT™ PicoGreen dsDNA Reagent (Life Technologies, Carlsbad, CA, USA).

Liquid Phase Sequence Capture: Prior to hybridization, amplified sample libraries with unique barcoded adapters will be combined in equal amounts into 1.1 µg pools for multiplex sequence capture. Exome sequence capture will be performed with NimbleGen's SeqCap EZ Human Exome Library v3.0+UTR with 64 Mb of exonic sequence targeted (Roche NimbleGen, Inc., Madison, WI, USA). Prior to hybridization the following components will be added to the 1.1 ug pooled sample library: 4 µL of NEXTflex HE Universal Oligo 1, 250 µM (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'), 40 µL total 25 µM NEXTflex INV-HE blocking oligos, equal volumes of each blocking oligo complementary to the barcodes in the pool (5'-CAAGCAGAAGACGGCATACGAGATXGTGACT GGAGTTCAGACGTGTGCTCTTCCGATCT/C3 Spacer/-3', where X is 8-bases of sequence specific to adapter barcode used for library construction), and 5 µL of 1 mg/mL COT-1 DNA (Invitrogen, Inc., Carlsbad, CA, USA). Samples will be dried down by puncturing a hole in the plate seal and processing in an Eppendorf 5301 Vacuum Concentrator (Eppendorf, Hauppauge, NY, USA) set to 60˚C for approximately 1 hour. To each dried pool, 7.5 µL of NimbleGen Hybridization Buffer and 3.0 µL of NimbleGen Hybridization Component A will be added, and placed in a heating block for 10 minutes at 95˚C. The mixture will then be transferred to 4.5 µL of EZ Exome Probe Library and hybridized at 47˚C for 64 to 72 hours. Washing and recovery of captured DNA will be performed as described in NimbleGen SeqCap EZ Library SR Protocol.

Post-Hybridization LM-PCR: Pools of captured DNA will be amplified by ligation-mediated PCR consisting of one reaction for each pool containing 20µl captured library DNA, 25 uL 2x KAPA HiFi HotStart ReadyMix, and 5µL 10x Library Amplification Primer Mix(includes two primers whose sequences are: 5'-AATGATACGGCGACCACCGA-3' and 5'-CAAGCAGAAGACGGCATACGA-3'). PCR cycling conditions were as follows: 98˚C for 45 seconds, followed by 8 cycles of 98˚C for 15 s, 60˚C for 30 s, 72˚C for 30 s. The last step will be an extension at 72˚C for 1 minute. The reaction will be kept at 4˚C until further processing. The amplified material will be cleaned with Agencourt AMPure XP Reagent (Beckman Coulter Inc, Brea, CA, USA) according to NimbleGen SeqCap EZ Library SR Protocol. Pools of amplified captured DNA will then be quantified via Kapa's Library Quantification Kit for Illumina (Kapa Biosystems, Woburn, MA, USA) on the LightCycler 480 (Roche, Indianapolis, IN, USA).

Sequencing: The resulting post-capture enriched multiplexed sequencing libraries will be diluted to 15 pM and used in cluster formation on an Illumina cBOT (Illumina, San Diego, CA, USA), and paired-end sequencing will be performed using an Illumina HiSeq following Illumina-provided protocols for 2x125bp paired-end sequencing. Each exome will be sequenced to high depth in order to achieve 40x average coverage of the coding sequence, based on the UCSC hg19 "known gene" transcripts (http://genome.ucsc.edu/).

Bioinformatics analysis

The goal of the bioinformatics analysis is to yield reliable variant calls and to subsequently filter the variants to focus on those that are most likely to be deleterious.

The human reference genome and the "known gene" transcript annotation have been downloaded from the UCSC database (http://genome.ucsc.edu/), version hg19 (corresponding to Genome Reference Consortium assembly GRCh37). Sequencing reads are first trimmed using the Trimmomatic program (v0.32), which marks all low-quality stretches (average quality score < Q15 in a 4-bp sliding window) and reports the longest high-quality stretch of each read. Only read pairs with both ends no shorter than 36 bp are used. Reads are then aligned to the hg19 reference genome using the Novoalign software (v3.00.05) (http://www.novocraft.com). Duplicate reads due to either optical or PCR artifacts are removed from further analysis using the MarkDuplicates module of the Picard software (v1.126) (http://picard.sourceforge.net/). Additionally, our analysis uses only properly aligned read pairs, in the sense that the two ends of each pair must be mapped to the reference genome in complementary directions and must reflect a reasonable fragment length (300+/-100 bp). These high-quality alignments for each individual are further refined according to a local realignment strategy around known and novel sites of insertion and deletion polymorphisms using the RealignerTargetCreator and IndelRealigner modules from the Genome Analysis Toolkit (GATK v3.1). Bam file level recalibration is also performed using BaseRecalibrator module from GATK.

Variant discovery and genotype calling of multi-allelic substitutions, insertions and deletions are performed on all individuals globally using the UnifiedGenotyper and HaplotypeCaller modules from Genome Analysis Toolkit (GATK v3.1) as well as the FreeBayes variant caller (v9.9.2). The Ensemble variant calling pipeline (v0.2.2 http://bcb.io/2013/02/06/an-automated-ensemble-method-for-combining-and-evaluating-genomic-variants-from-multiple-callers/) is then implemented to integrate analysis results from above three callers. Then Ensemble variant calling pipeline applies a machine learning algorithm called Support Vector Machine (SVM) to identify an optimal decision boundary based on the variant calling results out of multiple variant callers, with an aim to improve the caller's ROC (*i.e.*, a more balanced decision between false positives and true positives).

In addition, insertions and deletions are left-aligned at both post-alignment (BAM) and post-variant-calling (VCF) levels using GATK's LeftAlignIndels and LeftAlignVariants modules, respectively. Annotation and variants dissemination (optional) are performed using our in-house custom software annotation pipeline. This pipeline adds different types of functional annotations that range from DNA and RNA level to protein/histone level through integration of multiple public-domain applications including SnpEff/SnpSift (http://snpeff.sourceforge.net/) ANNOVAR (http://www.openbioinformatics.org/annovar/), etc. and public databases such as UCSC GoldenPath database (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/), ESP6500 dataset from University of Washington's Exome Sequencing Project (http://evs.gs.washington.edu/EVS/), dbNSFP - database of human nonsynonymous SNPs and function predictions (https://sites.google.com/site/jpopgen/dbNSFP), the Molecular Signatures Database - MSigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp), National Center for Biotechnology Information dbSNP database build 137, and 1000 Genomes Project.

Analytic approach:

Aim 1: Primary analyses will use Cox regression to test for associations with subsequent neoplasm development, considering both single variants (*i.e.*, recurrent exonic variants) and gene-based tests (*e.g.*, burden tests). The burden test collapses rare variants in a gene region into a single variable.[9,10] Outcomes of interest (with N≥50) include: any subsequent neoplasm, any radiation-related subsequent neoplasm, NMSC (restricted to basal cell carcinoma), female breast cancer, meningioma, thyroid cancer, sarcoma, and other (as allowed by sample size).

Models will be based on Cox proportional hazards regression with age as the underlying time variable, adjusted for sex and DNA type (gDNA vs. wgaDNA). We will include radiation as a "location-dependent" variable. For example, if the event at time t was a breast cancer occurring in the upper outer quadrant, we will set everyone's radiation dose at time t to be the dose to the upper outer quadrant. Doses will be considered in both continuous and categorical (<1, 1-9.9, or 10+ Gy) form. Note that this approach requires that each subsequent neoplasm is diagnosed at a unique age (*i.e.*, no ties). If there are multiple subsequent neoplasms occurring at the same age, the ages will be assigned randomly for each subsequent neoplasm by subtracting 0.001 years from the actual age at diagnosis. Analyses will focus on the first occurrence of the subsequent neoplasm type of interest, and the control group for each analysis will be only those individuals who were never diagnosed with any subsequent neoplasm. Primary childhood cancer type will be adjusted for indirectly by use of age as the time scale and adjustment for treatment exposures. Secondary analyses will address the role of first primary childhood cancer type by 1) adjusting for it in the model, and 2) restricting analyses to individuals with the most common primary childhood cancer types (acute lymphoblastic leukemia, CNS tumors, and Hodgkin lymphoma). We also will conduct secondary analyses of specific subsets of childhood cancer survivors with unusual characteristics, such as individuals who developed multiple subsequent neoplasms or who developed numerous basal cell carcinomas.[1,11]

Aim 2: A standard case/control logistic regression approach will be used to discover genetic variants that are associated with development of childhood cancer, using cancer-free individuals as of age 55 years as controls, matched on ancestry.

Statistical power

For each gene, we define the aggregated rare allele frequency (ARAF) as half the ratio of the number of these rare alleles carried by controls to the number of controls. When only one rare variant exists, ARAF is equivalent to minor allele frequency. For each given gene, the power is calculated based on two assumptions: only a proportion (λ) of rare variants are disease causing, and the disease-causing SNPs have the same risk estimate. When λ is smaller, smaller power is expected. The power for detecting genome-wide significant associations ($\alpha<0.05/20{,}000$ genes$=2.5\times10^{-6}$) is in Table 3, including information for analyses of all subsequent neoplasms combined, any radiation-related subsequent

neoplasm, basal cell carcinoma of the skin, and female breast cancer. Power will be more limited for other rarer subsequent neoplasms.

Replication

Aim 1: Childhood cancer survivors with available exome sequence data from the St. Jude Life cohort study will serve as replication for analyses of any subsequent neoplasm, any radiation-related subsequent neoplasm, NMSC (restricted to basal cell carcinoma), female breast cancer, meningioma, and thyroid cancer (laboratory work pending). Hereditary retinoblastoma survivors with available exome sequence data from the NCI Long-Term Follow-Up Study of Retinoblastoma will serve as replication for analyses of subsequent sarcoma (sequencing completed).

Aim 2: Data from this study will be combined with other ongoing studies (*e.g.*, an ongoing exome sequencing project in 700 children with osteosarcoma at NCI).

Contact with participants

The CCSS has used several different consent forms for collecting biologic specimens. Individuals consenting to collection of blood or buccal mouthwash signed a consent form that stated, "Even if your tissue is used for this kind of research, the results will not be put in your health records. Reports about research done with your tissue will not be given to you or your doctor. The research will not have an effect on your care." The Oragene consent form was silent with respect to the return of findings to participants. Because this sequencing is being done for research (not clinical) purposes and is not being conducted in a Clinical Laboratory Improvement Amendments (CLIA)-certified laboratory, and because the specimen "chain of custody" also has not been CLIA certified, we propose that the results will not be returned to participants. This includes research findings (any variants found to be associated in our primary analyses), secondary findings, and incidental findings.

Genomic data sharing

In accordance with NIH policy, the exome sequence data will be submitted to the database of Genotypes and Phenotypes (dbGaP). Access to controlled data will be granted by the intramural Data Access Committee (iDAC) of the NCI. Users requesting access to controlled data must submit a Data Access Request (DAR) to the iDAC for approval, which will be dependent upon completion of the DAR, agreeing to the terms and conditions in the Data Use Certification (DUC), and confirmation that the proposed research use is consistent with any restrictions on data use identified by the institutions that submitted the dataset to dbGaP. A biomedical research scientist from a recognized research institution can access both the genotype data and the executive summaries. All identifiers will be removed and only limited covariate data (case/control status, age group, and sex) will be available so as to prevent identification of subjects. Any other data (*i.e.*, other covariates) will only be accessible through the CCSS, who will oversee linkage of covariates with dbGaP datasets, restricted to only approved users.

Potential tables and figures

We anticipate publishing manuscripts focused on specific subsequent neoplasms as well as groups of therapy-related neoplasms. The tables and figures for each manuscript likely will be similar:

Tables

1. Selected characteristics of cases and controls from the Childhood Cancer Survivor Study and other study populations used in replication analyses (*e.g.*, sex, race, first primary cancer diagnosis, calendar year of first primary cancer, age at first primary cancer, treatments for first primary cancer, diagnoses of subsequent neoplasms, time from first primary cancer to subsequent neoplasms).
2. Genomic regions identified in exome sequencing (variant type, location, number cases/controls, risk estimate, 95% confidence interval, p-value, predicted effect).

Figures

1. Sequence conservation of variants of interest among multiple species.
2. Possible structural implications of exonic variants.

VI. Timeframe:

Required approvals

April-May 2015	CCSS Steering Committee, Genetics Working Group, and Research & Publications Committee; DCEG Radiation Epidemiology Branch, Biomarker Conceptual Review Group, and Technical Review Committee
March-May 2015	St. Jude and NCI IRBs

Research timeline

Spring 2015	Sample handling
Summer 2015-Winter 2016	Exome sequencing of discovery set
Winter-Spring 2016 +	Analysis of discovery set, replication, and publication

VII. References:

1. Armstrong, G.T., *et al.* Occurrence of multiple subsequent neoplasms in long-term survivors of childhood cancer: a report from the Childhood Cancer Survivor Study. *J Clin Oncol* **29**, 3056-3064 (2011).
2. Friedman, D.L., *et al.* Subsequent neoplasms in 5-year survivors of childhood cancer: The Childhood Cancer Survivor Study. *J Natl Cancer Inst* **102**, 1083-1095 (2010).
3. Morton, L.M., Onel, K., Curtis, R.E., Hungate, E.A. & Armstrong, G.T. The rising incidence of second cancers: patterns of occurrence and identification of risk factors for children and adults. *American Society of Clinical Oncology educational book / ASCO. American Society of Clinical Oncology. Meeting*, e57-67 (2014).
4. Berrington de Gonzalez, A., *et al.* Second solid cancers after radiation therapy: a systematic review of the epidemiologic studies of the radiation dose-response relationship. *Int J Radiat Oncol Biol Phys* **86**, 224-233 (2013).
5. Barnett, G.C., *et al.* Normal tissue reactions to radiotherapy: towards tailoring treatment dose by genotype. *Nat Rev Cancer* **9**, 134-142 (2009).
6. Knight, J.A., *et al.* Genome-wide association study to identify novel loci associated with therapy-related myeloid leukemia susceptibility. *Blood* **113**, 5575-5582 (2009).
7. Green, D.M., *et al.* The cyclophosphamide equivalent dose as an approach for quantifying alkylating agent exposure: a report from the Childhood Cancer Survivor Study. *Pediatr Blood Cancer* **61**, 53-67 (2014).
8. Stovall, M., *et al.* Dose reconstruction for therapeutic and diagnostic radiation exposures: use in epidemiological studies. *Radiation Research* **166**, 141-157 (2006).
9. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311-321 (2008).
10. Liu, D.J. & Leal, S.M. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* **6**, e1001156 (2010).
11. Athar, M., Li, C., Kim, A.L., Spiegelman, V.S. & Bickers, D.R. Sonic hedgehog signaling in Basal cell nevus syndrome. *Cancer Res* **74**, 4967-4975 (2014).

Table 1. Selected characteristics for the primary analytic study population from the CCSS GWAS (N=5324)

	Total		By subsequent neoplasm status			
			No		Yes	
Characteristics	N	(%)	N	(%)	N	(%)
Total	5324	(100)	4445	(100)	879	(100)
Sex						
Male	2585	(49)	2245	(51)	340	(39)
Female	2739	(51)	2200	(49)	539	(61)
Race (self-reported)						
White, non-Hispanic	5195	(98)	4332	(97)	863	(98)
Black, non-Hispanic	2	(0)	2	(0)	0	(0)
Hispanic	77	(1)	72	(2)	5	(1)
Other	50	(1)	39	(1)	11	(1)
Primary cancer type						
Hodgkin lymphoma	724	(14)	431	(10)	293	(33)
NHL	437	(8)	384	(9)	53	(6)
Neuroblastoma	390	(7)	365	(8)	25	(3)
Soft tissue sarcoma	475	(9)	408	(9)	67	(8)
ALL	1572	(30)	1358	(31)	214	(24)
AML	95	(2)	84	(2)	11	(1)
Leukemia, other	27	(1)	19	(0)	8	(1)
CNS	643	(12)	542	(12)	101	(11)
Ewing sarcoma	162	(3)	129	(3)	33	(4)
Osteosarcoma	283	(5)	245	(6)	38	(4)
Bone, other/NOS	18	(0)	15	(0)	3	(0)
Kidney (Wilms)	498	(9)	465	(10)	33	(4)
Year of primary cancer diagnosis						
1970-1975	1240	(23)	920	(21)	320	(36)
1976-1981	1951	(37)	1586	(36)	365	(42)
1982-1986	2133	(40)	1939	(44)	194	(22)
Age at primary cancer diagnosis (years)						
< 5	2076	(39)	1872	(42)	204	(23)
5 - < 10	1162	(22)	1019	(23)	143	(16)
10 - < 15	1109	(21)	867	(20)	242	(28)
15+	977	(18)	687	(15)	290	(33)
Radiotherapy for primary cancer within 5 years						
No	1737	(33)	1623	(37)	114	(13)
Yes	3316	(62)	2579	(58)	737	(84)
Unknown	271	(5)	243	(5)	28	(3)
Chemotherapy for primary cancer within 5 years						
No	1063	(20)	836	(19)	227	(26)
Yes	3954	(74)	3340	(75)	614	(70)
Unknown	307	(6)	269	(6)	38	(4)

Abbreviations: acute lymphocytic leukemia (ALL), acute myeloid leukemia (AML), basal cell carcinoma (BCC), central nervous system (CNS), Childhood Cancer Survivor Study (CCSS), genome-wide association study (GWAS), non-Hodgkin lymphoma (NHL), not otherwise specified (NOS).

Table 2. Occurrence of subsequent neoplasms by primary cancer type for the primary analytic study population from the CCSS GWAS (N=5324)

			By SN type (One each type/person)				
	Total	Ever SN	Breast	Skin, BCC	Thyroid	Meningioma	Sarcoma
Primary cancer	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
Hodgkin lymphoma	724 (14)	293 (33)	115 (64)	152 (42)	31 (34)	3 (2)	15 (26)
NHL	437 (8)	53 (6)	11 (6)	15 (4)	6 (7)	4 (3)	2 (3)
Neuroblastoma	390 (7)	25 (3)	1 (1)	2 (1)	5 (6)	1 (1)	3 (5)
Soft tissue sarcoma	475 (9)	67 (8)	10 (6)	17 (5)	4 (4)	3 (2)	13 (22)
ALL	1572 (30)	214 (24)	11 (6)	107 (30)	18 (20)	66 (53)	4 (7)
AML	95 (2)	11 (1)	4 (2)	4 (1)	0	1 (1)	3 (5)
Leukemia, other	27 (1)	8 (1)	0	1 (0)	4 (4)	2 (2)	0
CNS	643 (12)	101 (11)	2 (1)	31 (9)	11 (12)	44 (35)	5 (9)
Ewing sarcoma	162 (3)	33 (4)	10 (6)	10 (3)	4 (4)	0 (0)	3 (5)
Osteosarcoma	283 (5)	38 (4)	13 (7)	6 (2)	5 (6)	1 (1)	4 (7)
Bone, other/NOS	18 (0)	3 (0)	0	1 (0)	0	0	0
Kidney (Wilms)	498 (9)	33 (4)	3 (2)	16 (4)	2 (2)	0	6 (10)
Total	5324 (100)	879 (100)	180 (100)	362 (100)	90 (100)	125 (100)	58 (100)

Abbreviations: acute lymphocytic leukemia (ALL), acute myeloid leukemia (AML), basal cell carcinoma (BCC), central nervous system (CNS), Childhood Cancer Survivor Study (CCSS), genome-wide association study (GWAS), non-Hodgkin lymphoma (NHL), not otherwise specified (NOS), subsequent neoplasm (SN).

Table 3: Power for detecting genome-wide significant associations at $\alpha<2.5\times10^{-6}$

Study with controls (n=4445) and cases (n=879 for any subsequent neoplasm).

	λ[b]=1				λ=0.8				λ =0.5			
ARAF[a]	OR=2	OR=3	OR=4	OR=5	OR=2	OR=3	OR=4	OR=5	OR=2	OR=3	OR=4	OR=5
0.5%	0.00	0.00	0.19	0.67	0.00	0.00	0.04	0.28	0.00	0.00	0.00	0.01
1%	0.00	0.40	0.96	1.00	0.00	0.12	0.72	0.98	0.00	0.00	0.08	0.43
2%	0.14	0.99	1.00	1.00	0.03	0.85	1.00	1.00	0.00	0.15	0.77	0.99
3%	0.52	1.00	1.00	1.00	0.20	0.99	1.00	1.00	0.01	0.56	0.98	1.00
4%	0.83	1.00	1.00	1.00	0.48	1.00	1.00	1.00	0.04	0.85	1.00	1.00

Study with controls (n=4445) and cases (n=797 for any radiation-related subsequent neoplasm).

	λ=1				λ=0.8				λ =0.5			
ARAF	OR=2	OR=3	OR=4	OR=5	OR=2	OR=3	OR=4	OR=5	OR=2	OR=3	OR=4	OR=5
0.5%	0.00	0.00	0.12	0.54	0.00	0.00	0.03	0.19	0.00	0.00	0.00	0.00
1%	0.00	0.31	0.93	1.00	0.00	0.08	0.61	0.96	0.00	0.00	0.05	0.31
2%	0.10	0.97	1.00	1.00	0.02	0.78	1.00	1.00	0.00	0.11	0.67	0.97
3%	0.43	1.00	1.00	1.00	0.15	0.99	1.00	1.00	0.01	0.44	0.97	1.00
4%	0.75	1.00	1.00	1.00	0.37	1.00	1.00	1.00	0.03	0.77	1.00	1.00

Study with controls (n=4445) and cases (n=362 for basal cell carcinoma).

	λ=1				λ=0.8				λ =0.5			
ARAF	OR=2	OR=3	OR=4	OR=5	OR=2	OR=3	OR=4	OR=5	OR=2	OR=3	OR=4	OR=5
0.5%	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.00	0.00	0.13	0.51	0.00	0.00	0.03	0.18	0.00	0.00	0.00	0.01
2%	0.00	0.28	0.91	1.00	0.00	0.08	0.59	0.94	0.00	0.00	0.06	0.31
3%	0.02	0.77	1.00	1.00	0.00	0.41	0.94	1.00	0.00	0.02	0.32	0.79
4%	0.10	0.96	1.00	1.00	0.02	0.73	1.00	1.00	0.00	0.11	0.66	0.97

Study with controls (n=2200) and cases (n=180 for female breast cancer).

	λ=1				λ=0.8				λ =0.5			
ARAF	OR=2	OR=3	OR=4	OR=5	OR=2	OR=3	OR=4	OR=5	OR=2	OR=3	OR=4	OR=5
0.5%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2%	0.00	0.01	0.12	0.49	0.00	0.00	0.03	0.19	0.00	0.00	0.00	0.00
3%	0.00	0.08	0.55	0.93	0.00	0.02	0.23	0.67	0.00	0.00	0.01	0.08
4%	0.00	0.26	0.86	0.99	0.00	0.09	0.57	0.92	0.00	0.00	0.06	0.29

[a] Aggregated rare allele frequency (ARAF) for a gene. If 1000 controls carry 20 "qualified" rare alleles in the gene, ARAF=20/1000/2=1%.

[b] λ is the proportion of rare variants that are disease causing. This proportion depends on the specific characteristics of the included variants (*e.g.*, a high proportion of nonsense mutations are expected to be disease causing).