Check for updates

RESEARCH ARTICLE SUMMARY

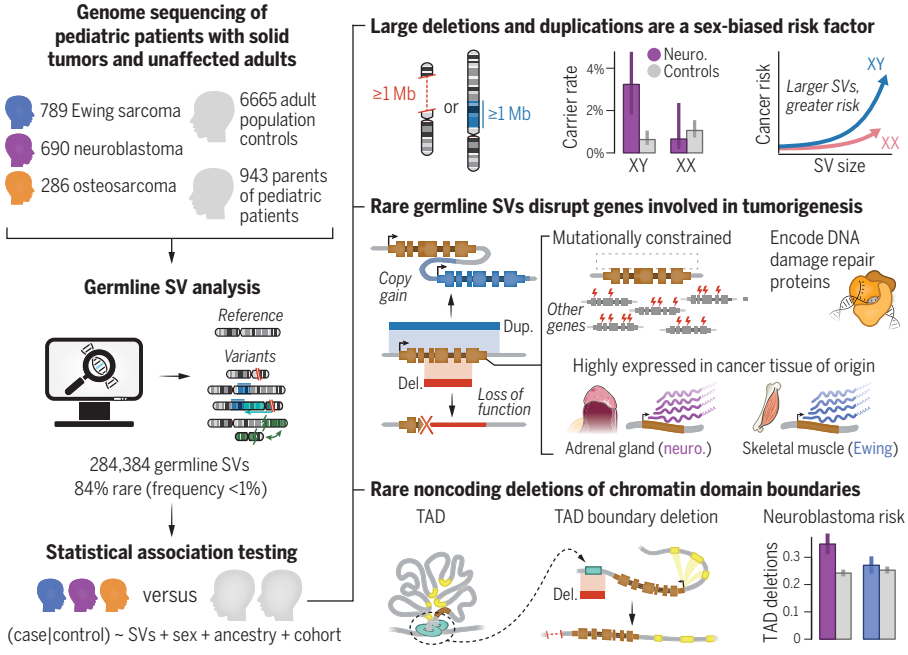
CANCER GENOMICS

Rare germline structural variants increase risk for pediatric solid tumors

Riaz Gillani*†, Ryan L. Collins†, Jett Crowdis, Amanda Garza, Jill K. Jones, Mark Walker, Alba Sanchis-Juan, Christopher W. Whelan, Emma Pierce-Hoffman, Michael E. Talkowski, Harrison Brand, Kevin Haigis, Jaclyn LoPiccolo, Saud H. AlDubayan, Alexander Gusev, Brian D. Crompton, Katherine A. Janeway, Eliezer M. Van Allen*

INTRODUCTION: Extracranial solid tumors make up one-third of all new pediatric cancer diagnoses annually and are a leading cause of childhood morbidity and mortality. The somatic drivers of these cancers have been meticulously characterized by prior studies, but the biomedical community lacks a comparable understanding of the earliest risk factors that lead to tumor development. Germline genetic factors are thought to play an important role in disease pathogenesis owing to the early age of onset and lower tumor mutational burden in these cancers, but presently recognized predisposition genes only account for 10 to 15% of all diagnoses. Identifying more germline risk factors in pediatric solid tumors might unlock a deeper understanding of the central tumorigenic biology and ultimately contribute to more informed approaches to clinical management and treatment.

RATIONALE: Most prior germline genetic studies in pediatric solid tumors have focused on short variants, specifically single-nucleotide variants and small insertions or deletions in the protein-coding regions of focused sets of genes. Structural variants (SVs) are a relatively understudied class of germline variation that involve rearrangements of ≥50 nucleotides. We hypothesized that rare germline SVs contribute to pediatric solid tumor predisposition and therefore undertook a study to evaluate their impact across coding and noncoding genomic regions in neuroblastoma, Ewing sarcoma, and osteosarcoma. We analyzed germline genome sequencing from 1765 affected patients, their 943 unaffected parents, and 6665 adult controls and statistically compared patterns of rare germline SVs in pediatric patients versus adult controls.



Rare germline SVs are risk factors for pediatric solid tumors. We analyzed germline genome sequencing from 1765 patients with solid tumors, their 943 unaffected parents, and 6665 adult controls. Focusing on rare SVs, we found that large events, events that disrupted genes from biologically relevant pathways, and events that affected chromatin domain boundaries, were risk factors for pediatric solid tumors. Del., deletion; Dup., duplication; neuro., neuroblastoma; TAD, topologically associated domain.

RESULTS: We identified a median of 7275 germline SVs per genome. Overall, de novo germline SV rates in neuroblastoma and Ewing sarcoma trios were comparable to those of the general population (~0.2 per genome), and de novo SVs were significantly larger and more likely to be gene-disruptive than inherited SVs. We discovered that ultrarare germline deletions or duplications of more than 1 million nucleotides were a sex-biased risk factor for pediatric solid tumors, exhibiting a fourfold enrichment in male patients relative to cancer-free controls. By implementing a category-wide association study framework, we found that germline SVs in mutationally constrained genes were enriched in neuroblastoma patients, and SVs that affect DNA damage repair genes were enriched in both Ewing sarcoma and neuroblastoma. Cross-examination of these gene-disruptive SVs with gene expression data revealed that the SVs in patients tended to disrupt genes that are highly expressed in their cancer's tissue of origin, and the dysregulated expression caused by these SVs manifested in the patients' tumors. Across our cohort, we also identified dozens of germline SVs that affect established cancer predisposition genes, such as *PHOX2B* and *BARD1* in neuroblastoma and Fanconi anemia genes in Ewing sarcoma, demonstrating that germline SV analyses can uncover new clinically relevant findings in individual patient genomes. Finally, we discovered that ultrarare SVs that overlap chromatin domain boundaries were enriched in neuroblastoma versus adult controls, which was not the case for Ewing sarcoma. This implied that subsets of rare noncoding SVs outside of protein-coding genes likely contribute to the pathogenesis of some pediatric solid tumors.

CONCLUSION: Overall, our results demonstrate that rare germline SVs that span coding and noncoding loci are important risk factors for pediatric solid tumors. We observed stronger SV effects in neuroblastoma relative to Ewing sarcoma, which is consistent with the markedly earlier age of onset in neuroblastoma. We estimate that rare SVs contribute 1.1 to 5.6% of liability for these cancers, which underscores a modest yet meaningful role for rare germline SVs in the missing heritability of pediatric solid tumors. Our work supports the integration of germline SVs into clinical practice and research efforts related to pediatric solid tumor predisposition. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: eliezerm_vanallen@dfci.harvard.edu (E.M.V.A.); riaz_gillani@dfci.harvard.edu (R.G.)
†These authors contributed equally to this work.
Cite this article as R. Gillani *et al.*, *Science* **387**, eadq0071 (2025). DOI: 10.1126/science.adq0071

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adq0071

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

RESEARCH ARTICLE

CANCER GENOMICS

Rare germline structural variants increase risk for pediatric solid tumors

Riaz Gillani[1,2,3,4]*†, Ryan L. Collins[2,3,5]†, Jett Crowdis[3], Amanda Garza[2,5], Jill K. Jones[3,4,6], Mark Walker[7,8,9], Alba Sanchis-Juan[8,9,10], Christopher W. Whelan[7,8,9], Emma Pierce-Hoffman[7,8,9], Michael E. Talkowski[3,8,9,10], Harrison Brand[3,8,9,10], Kevin Haigis[2,3,11,12], Jaclyn LoPiccolo[2,3,5], Saud H. AlDubayan[2,5,13,14], Alexander Gusev[2,3,5,15], Brian D. Crompton[1,2,3,4], Katherine A. Janeway[1,2,3,4], Eliezer M. Van Allen[2,3,5,16]*

Pediatric solid tumors are a leading cause of childhood disease mortality. In this work, we examined germline structural variants (SVs) as risk factors for pediatric extracranial solid tumors using germline genome sequencing of 1765 affected children, their 943 unaffected parents, and 6665 adult controls. We discovered a sex-biased association between very large (>1 megabase) germline chromosomal abnormalities and increased risk of solid tumors in male children. The overall impact of germline SVs was greatest in neuroblastoma, where we uncovered burdens of ultrarare SVs that cause loss of function of highly expressed, mutationally constrained genes, as well as noncoding SVs predicted to disrupt chromatin domain boundaries. Collectively, we estimate that rare germline SVs explain 1.1 to 5.6% of pediatric cancer liability, establishing them as an important component of disease predisposition.

Extracranial pediatric solid tumors make up approximately one-third of all new pediatric cancer diagnoses per year and are a leading cause of childhood mortality (*1*). These cancers are often aggressive and require multimodal treatment with cytotoxic chemotherapy, surgery, and radiation. Cure rates have remained modest owing to a reliance on nontargeted therapies, and patients who are cured often face substantial lifelong morbidity (*2*, *3*). Progressive improvements in genomics over the past two decades have greatly expanded our understanding of the somatic and germline molecular features of these rare diseases (*4*–*7*). In particular, the study of germline genetics—especially rare variants—has provided insight into the biological pathways that are responsible for the earliest stages of tumorigenesis (*8*).

Germline genetic factors are especially important in pediatric cancers, where patients have not lived long enough to accrue substantial exposure to environmental carcinogens such as tobacco smoke or ultraviolet radiation (*9*). Epidemiological studies have estimated a 4.5-fold familial relative risk for pediatric solid tumors (*10*), yet only 10 to 15% of pediatric cancer cases can be attributed to presently recognizable germline risk factors (*11*, *12*). This observation implies that a large portion of the heritability for pediatric cancers is still "missing" (undiscovered) (*13*). However, conventional germline genetic analyses in these diseases have often focused on small subsets of variants, specifically single-nucleotide variants (SNVs) and small insertions or deletions (indels) in a narrow list of genes with preexisting evidence as risk factors in other cancers (*5*, *14*–*16*). Identifying missing risk factors beyond conventional SNVs and indels is thus crucial to developing rational screening, diagnostic testing, and treatment strategies for these malignancies as well as revealing their central oncogenic programs.

Germline structural variants (SVs) are an understudied family of genomic alterations that includes any rearrangement that affects ≥50 nucleotides. By convention, SVs are classified into distinct mutational classes based on their alternative allele structures, such as deletions or duplications [collectively, copy-number variants (CNVs)], insertions, inversions, translocations, and more complex rearrangements (*17*, *18*). Somatically acquired SVs have been widely implicated in oncogenesis and tumor evolution (*19*, *20*). In addition, germline SVs are now recognized as critical factors in human population genetics, gene regulation, and numerous diseases (*21*, *22*). Prior focused evaluations of rare germline SVs in patients with pediatric solid cancers have revealed hints of possible germline SV risk factors (*23*, *24*), but these studies have been limited by small sample sizes, a lack of matched healthy controls, and reliance on low-resolution genomic technologies.

We hypothesized that germline SVs might represent a source of disease predisposition and oncogenesis with potential clinical impact. We therefore investigated rare germline SVs in three pediatric extracranial solid tumors—Ewing sarcoma, neuroblastoma, and osteosarcoma—using germline whole-genome sequencing (WGS) from pediatric patients and ancestry-matched adult controls. We sequentially dissected five aspects of rare germline SVs in disease predisposition (Fig. 1A), which collectively demonstrated that rare SVs are an important component of predisposition to pediatric solid tumors and should be considered in both translational research and clinical diagnostic testing.

Results

Discovery and characterization of germline SVs in 9373 genomes

We aggregated and harmonized short-read germline WGS data (median coverage = 24-fold) generated from DNA derived mostly from peripheral blood or saliva of 10,590 individuals recruited by six independent research studies (fig. S1 and table S1). Effectively all (99%) pediatric cancer patient sequencing data were sourced from the Gabriella Miller Kids First (GMKF) Pediatric Research Program (GMKF) or St. Jude Children's Research Hospital (*25*–*29*). We combined these patients with ancestry-matched adult controls whose genomes were sequenced by two cardiovascular disease studies: the Multiethnic Study of Atherosclerosis (MESA) and the Mt. Sinai BioMe Biobank (BioMe) (*30*). We also included an ancestrally diverse panel of parent-child trios from the 1000 Genomes Project for ancestry inference and technical benchmarking (*31*). We performed genome-wide SV discovery, genotyping, and annotation jointly across all samples using GATK-SV, an ensemble cloud-based pipeline for population-scale WGS studies of germline SVs (*17*). After extensive quality control, we retained a high-confidence SV dataset comprising 9373 individuals who passed all data quality filters, including 1765 children diagnosed with one of three pediatric cancers—Ewing sarcoma (n = 789), neuroblastoma (n = 690), or osteosarcoma (n = 286)—as well as their 943 unaffected parents and 6665 adult

[1]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Boston Children's Hospital, Boston, MA, USA. [5]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [6]Harvard/MIT MD-PhD Program, Harvard Medical School, Boston, MA, USA. [7]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [9]Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [10]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [11]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [12]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [13]Division of Genetics, Brigham and Women's Hospital, Boston, MA, USA. [14]College of Medicine, King Saudi bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia. [15]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [16]Center for Cancer Genomics, Dana-Farber Cancer Institute, Boston, MA, USA.
***Corresponding author. Email: eliezerm_vanallen@dfci.harvard.edu (E.M.V.A.); riaz_gillani@dfci.harvard.edu (R.G.)**
†These authors contributed equally to this work.

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

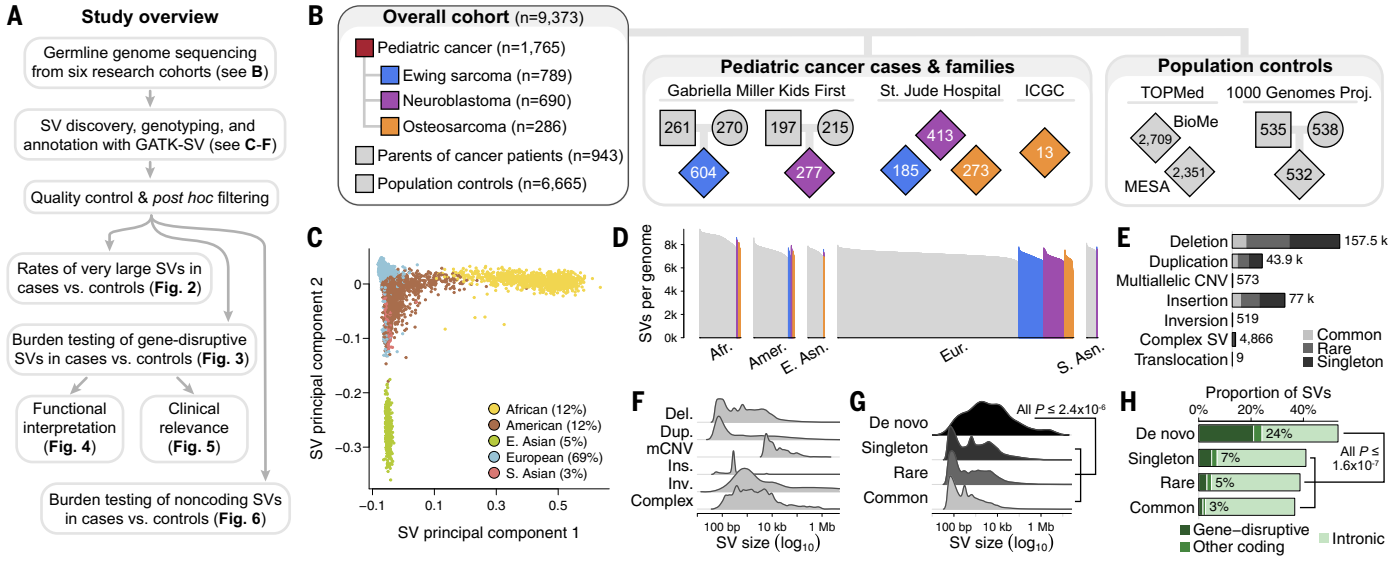


Fig. 1. The landscape of germline SVs in 9373 pediatric cancer patients and controls. (**A**) Roadmap of key datasets and analyses in this study. (**B**) Summary of study cohort structure after excluding low-quality WGS samples. ICGC, International Cancer Genome Consortium; TopMed, Trans-Omics for Precision Medicine. (**C**) Individuals clustered by continental genetic ancestry groups based on the top two principal components computed from common, high-quality SVs. See also fig. S3. (**D**) We detected an average of 7275 high-quality SVs per genome, which was correlated with sample ancestry as expected based on human demographic history (*18*). Afr., African or African-American; Amer., American; E. Asn., East Asian; Eur., European; S. Asn., South Asian. (**E**) Most SVs were rare (AF <1%). For common SVs, AF ≥1%; for rare SVs, AF <1% and allele count (AC) >1; and for singleton SVs, AC = 1. (**F**) Most SVs were small (<1 kb). Del., deletion; Dup., duplication; mCNV, multiallelic CNV; Ins., insertion; Inv., inversion. (**G**) De novo SVs were significantly larger on average than singleton, rare, or common inherited SVs. See also fig. S5. (**H**) De novo SVs were significantly more likely to result in gene-disruptive consequences than singleton, rare, or common inherited SVs.

population controls (Fig. 1B and fig. S2). The top six principal components (PCs) derived from common SVs tracked with genetic ancestry, which revealed that 31.4% of individuals clustered with non-European populations from the 1000 Genomes Project in six-dimensional genetic PC space (11.8% admixed American, 11.8% African or African-American, 4.7% East Asian, 3.1% South Asian; Fig. 1C and fig. S3). In parallel, we inferred genetic relatedness between all pairs of samples based on their germline SV genotypes, which defined a subset of 6727 strictly unrelated and ancestry-matched patients and controls for subsequent association testing.

We identified 284,384 germline SVs present in at least one genome, with 7275 as the median number of SVs per individual in this study (range = 5486 to 9324 SVs; Fig. 1D). We rigorously assessed the quality of our SV dataset against multiple benchmarks, which confirmed that the properties and distributions of SVs in this study mirrored expectations set by prior noncancer WGS initiatives (*17*, *31*) (supplementary text and fig. S4). Most SVs in our dataset (84.0%) appeared as rare alleles in the overall cohort [allele frequency (AF) <1%; Fig. 1E], and the average SV was relatively small [median size = 281 base pairs (bp); Fig. 1F]. Consistent with prior reports, the total number of SVs per genome correlated with continental ancestry, with individuals of primarily African ancestry exhibiting the greatest genetic diversity (median = 8704 SVs per genome) (*17*, *18*, *32*). In summary, this dataset represented a high-confidence atlas of germline SVs from patients and ancestry-matched controls that allowed us to interrogate the role of germline SVs in predisposition to the three pediatric tumor histologies.

Effectively all (>99.9%) germline SVs identified in pediatric cancer patients were stably inherited from unaffected parents. However, given that de novo germline SVs are enriched for pathogenic effects in other pediatric diseases (*33*–*36*), we next wanted to quantify the rate of de novo SVs among the children with pediatric solid tumors in our study. We applied strict post hoc filtering (*37*–*39*) to identify 63 high-confidence de novo SVs (54 predicted germline and 9 predicted postzygotic mosaic) carried by pediatric cancer probands from 330 complete parent-child trios in our final dataset. The rate of de novo SVs in neuroblastoma [24 de novo SVs in 103 trios; 0.233 per child; 95% confidence interval (CI): 0.149 to 0.347] was not significantly higher than in Ewing sarcoma (39 de novo SVs in 227 trios; 0.172 per child; 95% CI: 0.122 to 0.235, P = 0.276; Przyborowski-Wilenski C test). Furthermore, the rate of de novo SVs in both diseases was not significantly different from the previously reported rate in 2372 control trios (0.160; 95% CI: 0.144 to 0.177) detected from blood-derived WGS with the same methods used here (neuroblastoma P = 0.079; Ewing sarcoma P = 0.483; fig. S5A). Although the overall rate of de novo SVs did not differ between patients and controls, we noted subtle differences in the mutational spectra of de novo SVs that arose in patients versus controls: Neuroblastoma patients had a higher rate of de novo duplication CNVs (P = 0.002), whereas Ewing sarcoma patients exhibited elevated rates of de novo complex and inversion-mediated SVs ($P = 4.6 \times 10^{-6}$; fig. S5B). Collectively, the 58 de novo SVs we identified in pediatric cancer patients exhibited clear characteristics of strongly deleterious variation because they were 4.6 times larger on average than all other rare or common SVs (median size = 3143 bp; $P \leq 2.4 \times 10^{-6}$; Fig. 1G) and were 4.2-fold enriched for gene-disruptive consequences compared with inherited SVs ($P = 1.6 \times 10^{-7}$; Fig. 1H).

Germline chromosomal dosage alterations are a sex-biased risk factor for pediatric solid tumors

We first hypothesized that large germline SVs might predispose to pediatric cancers given that large SVs are prominent risk factors for many other severe pediatric disorders (*40*–*43*). We systematically documented all large (>1 Mb), rare (AF <1%) germline SVs present in the subset of 6727 ancestry-matched patients and controls (Fig. 2A). We identified 143 large, rare SVs carried by 160 samples (50 patients and 110 controls; four samples carried two SVs each). These 143 germline SVs spanned a diverse mutational spectrum, including 71 CNVs (27 deletions and 44 duplications), 45 balanced inversions, 12 complex SVs, 10 aneuploidies

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

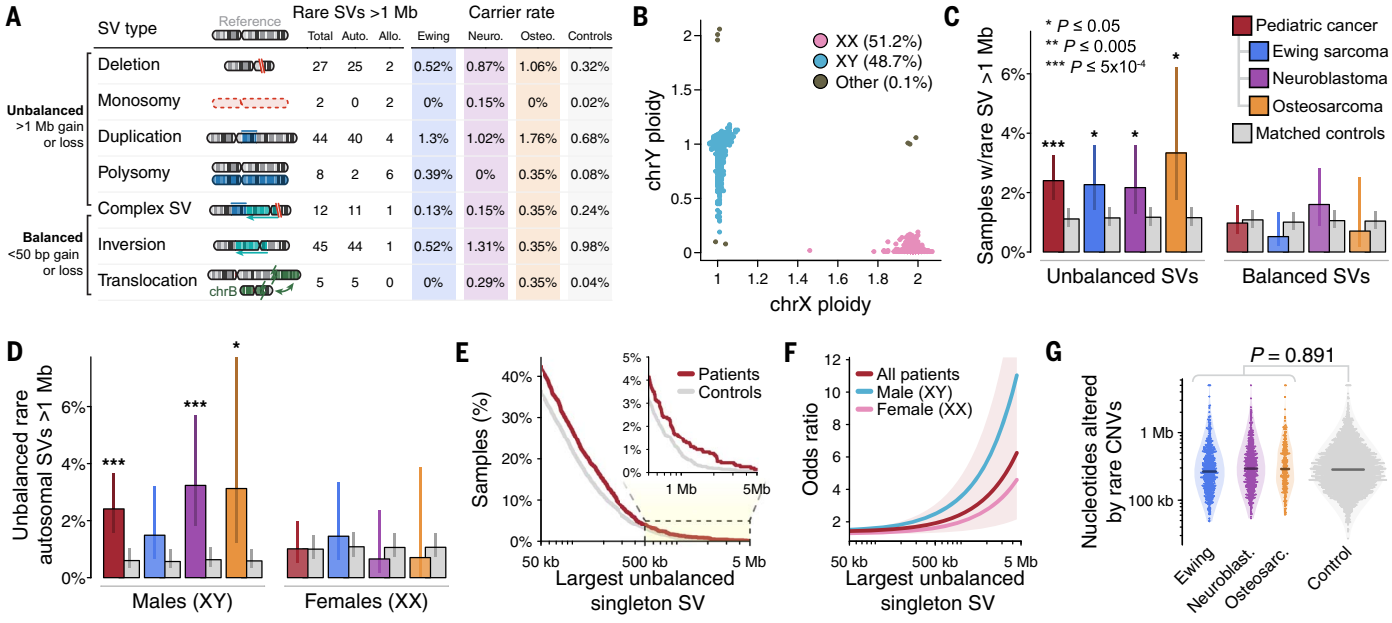


Fig. 2. Large germline CNVs increase risk for pediatric solid tumors in a sex-specific manner. (**A**) We identified 143 rare (AF <1%), large (>1 Mb) germline SVs in a set of 1744 pediatric cancer patients and 4983 ancestry-matched adult controls. Auto., autosomal; Allo., allosomal; Neuro., neuroblastoma; Osteo., osteosarcoma. (**B**) WGS delineated sample sex based on ploidy (copy number) estimates of X and Y chromosomes. (**C**) We discovered an association between pediatric solid tumors and the 84 unbalanced SVs from (A). We did not observe any significant associations with comparably large but balanced rare SVs. Error bars indicate the 95% CI; *P* values are derived from logistic regression adjusted for sex, cohort, and ancestry. (**D**) Results from (C) restricted to autosomes and stratified by male (karyotypic XY) versus female (karyotypic XX) samples. Error bars indicate 95% CIs. (**E**) Proportion of patients and controls who carried at least one singleton unbalanced germline SV that was larger than the size specified on the *x* axis. (**F**) Relationship between singleton SV size and the corresponding OR for pediatric cancer based on the data from (D) presented as a cubic smoothing spline. The shaded area is the 95% CI for a pooled model of both sexes from all histologies. (**G**) The total sum of nucleotides altered by all autosomal, rare, unbalanced SVs did not significantly differ between patients and controls. Each point represents a single individual's genome, the horizontal bars represent group-wise medians, and the shaded area is the kernel density estimate of each group's underlying distribution. See also figs. S6 to S8.

(three XXY, three XYY, two monosomy X, and two trisomy 21; Fig. 2B and fig. S6), and five reciprocal translocations. All three pediatric cancer histologies exhibited elevated rates of unbalanced (net DNA gain or loss >1 Mb) rare SVs compared with ancestry-matched controls [pan-cancer $P = 2.0 \times 10^{-4}$; odds ratio (OR) = 2.31; 95% CI: 1.48 to 3.59; $n = 84$ SVs; logistic regression adjusted for sex, cohort, and genetic ancestry; Fig. 2C and table S2]. To ensure that this observed association was not driven by technical false positives in patients, we scrutinized the WGS evidence for all 84 unbalanced SVs and found unambiguous support in the raw WGS depth of coverage for 97.6% (82/84) of SVs (fig. S7). Moreover, 11 of the 84 SVs were carried by patients whose parents were also in our final dataset, and we found that 9 of 11 (82%) of those large SVs were transmitted from an unaffected parent (5/9 paternal; 4/9 maternal), whereas the remaining 2 of 11 (18%) likely arose de novo. These de novo unbalanced SVs were duplications of 2.1 Mb on chromosome 6 in a neuroblastoma patient and 2.3 Mb on chromosome 7 in an Ewing sarcoma patient, neither of which overlapped any genes with established roles in cancer (*5*, *14*–*16*). We did not observe a similar association for the large, rare SVs that did not involve altered DNA dosage, such as balanced inversions and reciprocal translocations (pan-cancer $P = 0.60$; OR = 0.86), which suggests that the relationship between large, rare SVs and pediatric cancer predisposition is specific to alterations in germline copy number.

Further dissection of all large, unbalanced SVs revealed that the observed excess in patients was almost entirely concentrated in males (karyotypic XY only; pan-cancer $P = 9.7 \times 10^{-5}$; OR = 3.65), with no significant enrichment observed in females (karyotypic XX only; pan-cancer $P = 0.49$; OR = 1.27; fig. S8A). This difference in effect sizes between sexes was significant ($P = 0.028$; Welch's *t* test), and the rates of these SVs did not significantly differ between male and female controls (25/2245 male controls versus 29/2472 female controls; $P = 0.939$). These sex biases were most pronounced in neuroblastoma and persisted even after (i) excluding all SVs on sex chromosomes (Fig. 2D), (ii) restricting to individuals of European ancestry (fig. S8B), and (iii) analyzing independent cohorts of patients separately (fig. S8C). We next assessed whether the enrichment of large, unbalanced SVs in patients versus controls was also present for smaller unbalanced SVs. The elevated rates of rare, unbalanced SVs in patients extended to relatively smaller SVs, albeit with markedly weaker effects (pan-cancer OR = 1.37 versus OR = 2.50 for singleton SVs >50 kb versus >1 Mb in both sexes, respectively; Fig. 2, E and F, and fig. S8, D to F). Given that the median size of all unbalanced SVs in our dataset was 529 bp, we tested whether these findings were simply proxies for a general genome-wide excess of rare, unbalanced SVs of all sizes in patients, but we found no significant difference in the sum of all nucleotides gained and lost due to all rare SVs between patients and controls (Fig. 2G and fig. S8G). Collectively, these results established that children—especially males—are at substantially greater risk for pediatric solid cancers if their genomes harbor extensive DNA dosage imbalance because of a single large germline SV.

We reasoned that this enrichment of large, unbalanced SVs in patients could either be driven by (i) focal excesses of SVs concentrated at specific pathogenic loci or (ii) a nonspecific excess of SVs distributed uniformly throughout the genome; the former is observed in many developmental disorders (*43*, *44*), whereas the latter might be indicative of mechanisms beyond direct gene disruption, such as mitotic chromosomal instability. To distinguish between these possibilities, we first excluded all SVs

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

that overlapped any of 621 genes with a previously established role in cancer (*5*, *14*–*16*) and found that the original enrichment in patients remained largely undiminished (pan-cancer $P = 0.007$; OR = 2.10; $n = 62$ autosomal SVs; fig. S8H). Second, we performed case-control burden tests for all rare, unbalanced SVs ≥100 kb in 1-Mb sliding windows across all autosomes, which did not identify any loci surpassing Bonferroni-corrected significance ($P < 1.9 \times 10^{-5}$) (fig. S9). The lack of overlap between obvious pathogenic loci and large, unbalanced SVs in patients suggests that these SVs may impart disease risk through undiscovered pathogenic genes or extragenic mechanisms, such as an effect on somatic genome stability.

Gene-disruptive germline SVs are enriched in pediatric cancers

We next turned our attention to a different subset of SVs: focal gene-disruptive SVs, which, on average, were ~250 times smaller (median size = 8.0 kb) but ~60 times more abundant ($n = 8636$ SVs) than the large, rare SVs that were the focus of our earlier analyses. We specifically hypothesized that gene-disruptive germline SVs might also mediate risk for pediatric solid tumors. We found that patients across all three histologies had an average of 6 to 10 more autosomal genes disrupted by germline SVs than controls after accounting for sex, ancestry, and cohort (pan-cancer patient mean = 127.5 genes; control mean = 120.9 genes; $P < 10^{-30}$; Fig. 3A). This result was not explained by outsized contributions from common SVs or large, multigene SVs: Patients carried a significantly greater total number of rare gene-disruptive SVs (pan-cancer $P = 1.7 \times 10^{-5}$; OR = 1.07), which remained significant even when restricting to small SVs that affect single genes ($P = 0.002$; OR = 1.06; fig. S10A). In contrast to the sex-biased effects of large, unbalanced SVs, we did not observe a significant difference in effect sizes of all rare gene-disruptive SVs between males and females (pan-cancer male OR = 1.09; female OR = 1.05; $P = 0.27$, Welch's t test; fig. S10B). The enrichment for gene-disruptive SVs in patients was twice as strong when restricted to singleton SVs (Fig. 3B) and exhibited comparable effect sizes in samples from GMKF and St. Jude cohorts when analyzed separately (fig. S10C). We also noted independent enrichments of two distinct subsets of SVs predicted to have opposite consequences on gene function (Fig. 3C): loss of function (LoF), usually caused by exonic deletion (for example, singleton LoF pan-cancer $P = 2.4 \times 10^{-7}$; OR = 1.19), and whole-gene copy-gain (CG) caused by duplication SVs (for example, singleton CG pan-cancer $P = 0.001$; OR = 1.32); however, we did not find that singleton LoF and CG SVs were mutually exclusive within individual genomes ($P = 0.57$, chi-square test). Taken together, these results demonstrated that the genomes of children with pediatric solid tumors tend to harbor more rare germline SVs that duplicate and/or inactivate protein-coding genes than the average cancer-free adult.

To gain insight into the genes driving this global enrichment of rare LoF and CG SVs in pediatric cancer patients, we next implemented a systematic category-wide association study (CWAS) approach for Ewing sarcoma and neuroblastoma (*38*, *45*), the two more highly powered histologies in our study. Briefly, we enumerated all possible combinations of six filters for our SV dataset, including filters for AF, predicted coding consequence, SV type, genic mutational constraint, gene set membership, and gene expression in the adult tissue from the Genotype-Tissue Expression (GTEx) project that was the most proximal to the putative cancer tissue of origin (adrenal gland for neuroblastoma and skeletal muscle for Ewing sarcoma; tables S3 to S5) (*46*). This strategy resulted in 4800 possible categories of SVs per disease. To focus on categories with sufficiently dense SV data, we required >10 SVs per category when summed across patients and controls; this yielded 718 and 682 "testable" categories in neuroblastoma and Ewing sarcoma, respectively (Fig. 3D). Within each category, we evaluated the burden of autosomal SVs in patients relative to controls after adjusting for sex, cohort, and genetic ancestry and identified categories that met a Bonferroni significance threshold ($P \leq 1.5 \times 10^{-4}$) after estimating the number of effective tests using permutation to account for correlation between similar categories, as has been previously proposed (*38*, *45*).

In neuroblastoma, we identified 29 categories of SVs that surpassed our conservative Bonferroni significance threshold. Most significant categories involved LoF SVs associated with increased disease risk (for example, singleton LoF deletions, $P = 4.6 \times 10^{-10}$; OR = 1.34; Fig. 3E and table S6). Among all significant categories, we observed the strongest effects for SVs that cause loss of function of mutationally constrained genes (for example, singleton LoF deletions of constrained genes; $P = 6.2 \times 10^{-6}$; OR = 1.81; $n = 67$ SVs), which mirrored patterns reported in pediatric neurodevelopmental disorders (*47*, *48*). The multilayer design of the CWAS allowed us to further dissect these signals and identify subsets of SVs driving these broad enrichments with comparatively greater effect sizes, albeit at reduced power. We identified several smaller categories that exhibited especially strong effect sizes, such as singleton complex SVs resulting in copy gain of one or more protein-coding genes ($P = 4.0 \times 10^{-4}$; OR = 6.27; $n = 7$ SVs), rare duplications resulting in copy gain of cancer genes from the Catalogue of Somatic Mutations in Cancer (COSMIC) Project ($P = 0.0024$; OR = 2.64; $n = 13$ SVs), and rare deletions resulting in loss of function of DNA double-strand break repair genes ($P = 6.5 \times 10^{-4}$; OR = 2.88; $n = 14$ SVs), all of which may provide insights into the types of germline SVs that confer relatively greater risk for neuroblastoma. By contrast, we found no categories that surpassed Bonferroni significance in Ewing sarcoma (Fig. 3F and table S7), although a handful of categories were nominally significant, such as singleton deletions resulting in loss of function of DNA damage repair genes ($P = 0.0047$; OR = 3.91; $n = 8$ SVs) and rare complex SVs resulting in copy gain of one or more genes ($P = 3.2 \times 10^{-4}$; OR = 2.27; $n = 22$ SVs). In aggregate, this systematic CWAS of coding SVs in neuroblastoma and Ewing sarcoma supported our hypothesis that specific subsets of gene-disruptive rare germline SVs predispose to these diseases and suggested that rare germline SVs play a relatively larger role in the pathogenesis of neuroblastoma than Ewing sarcoma.

Ultrarare germline SVs disrupt functionally relevant genes and pathways in pediatric tumors

Given that singleton LoF SVs were among the most statistically significant risk factors identified by the CWAS in both neuroblastoma and Ewing sarcoma, we next sought to clarify their functional impact on disease pathogenesis. Particularly in neuroblastoma, sequentially filtering singleton SVs to isolate LoF deletions in constrained, tissue-expressed genes increased SV effect sizes (from OR = 1.31 to 1.81 in neuroblastoma; Fig. 4A). This finding reflected the guided trade-off between power and interpretability in the CWAS framework and also implied that specific subsets of singleton SVs increased disease risk more than others. To understand biological processes that might be affected by these genic singleton SVs, we conducted gene set enrichment analyses to identify gene ontology (GO) terms whose constituent genes were disrupted by singleton SVs in patients more than would be expected by chance ($n = 2725$ gene sets tested). In neuroblastoma, which arises from the sympathetic nervous system (*49*), we found that gene sets involved in neurogenesis and synaptic transmission were significantly enriched for singleton gene-disruptive SVs [for example, central nervous system development (GO:0007417); OR = 3.1; $P = 2.2 \times 10^{-6}$; Fisher's exact test; Fig. 4B]. Although less notable, we also identified several gene sets enriched for singleton gene-disruptive SVs in Ewing sarcoma that could be linked to its mesenchymal origin (*50*), including calcium homeostasis and peptide synthesis [for example, peptide biosynthetic process (GO:0043043); OR = 4.5; $P = 7.8 \times 10^{-7}$]. For both diseases, we did not observe as strong an enrichment for SVs found in controls in these gene sets, which suggests that disease-specific

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

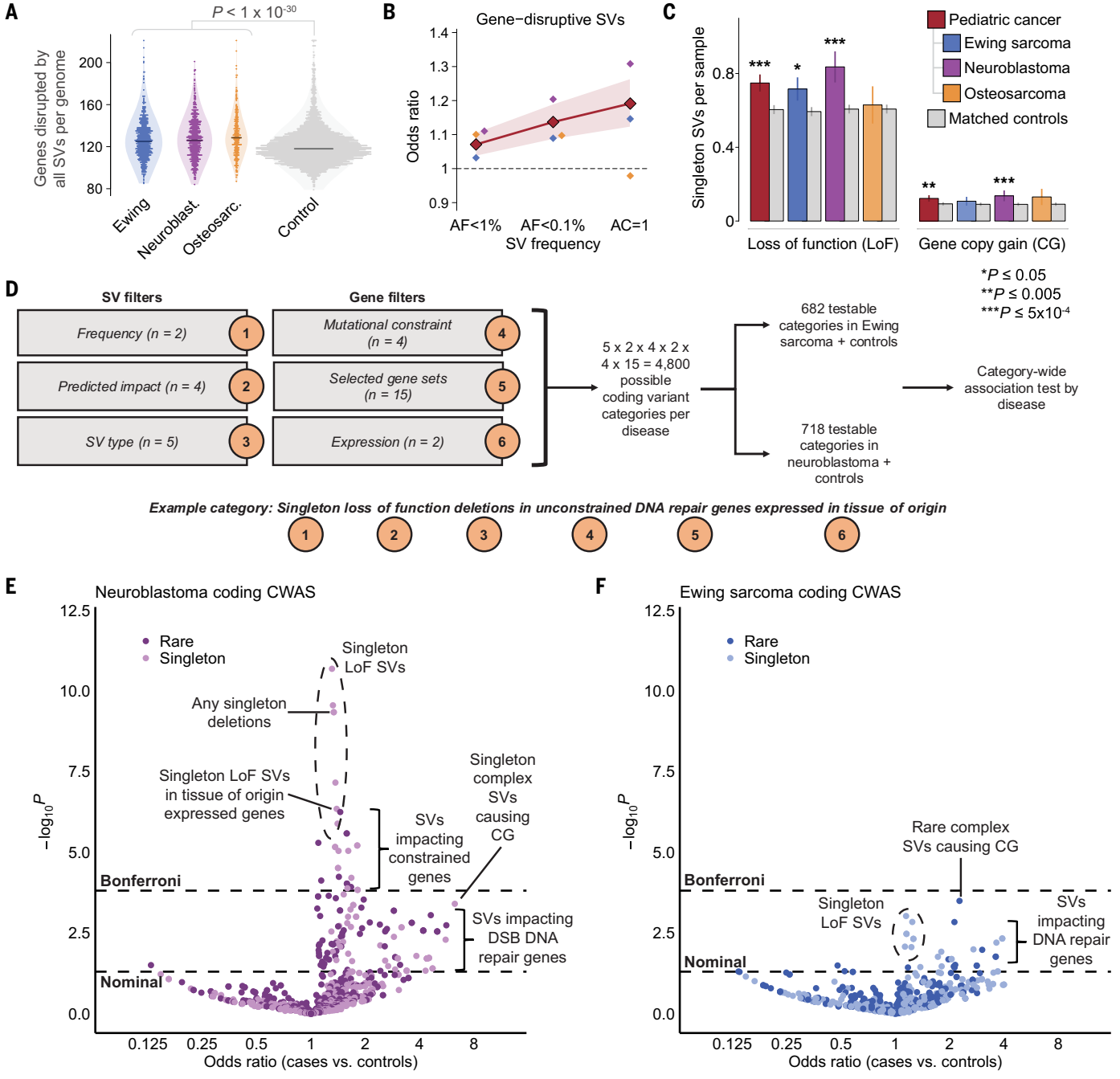


Fig. 3. Pediatric cancer patients carry an excess of rare germline SVs that affect disease-relevant genes. (**A**) On average, the number of protein-coding genes disrupted by all germline SVs was significantly greater among pediatric cancer patients relative to adult controls. Each point represents a single individual's genome, the horizontal bars represent group-wise medians, and the shaded area is the kernel density estimate of each group's underlying distribution. (**B**) Rare gene-disruptive SVs were enriched in patients versus controls, and this enrichment was inversely correlated with SV frequency. The shaded area is the 95% CI for all histologies. See also fig. S10. (**C**) We found comparable enrichments in two independent subsets of singleton SVs with opposing predicted consequences on gene function: loss of function and whole-gene copy gain. (**D**) We carried out a CWAS in neuroblastoma and Ewing sarcoma, combining six layers of filters to categorize types of coding SVs for burden testing (*38*, *45*). We evaluated 682 and 718 categories of gene-disruptive germline SVs in Ewing sarcoma and neuroblastoma, respectively. (**E**) In neuroblastoma, 29 categories of SVs exceeded Bonferroni significance, including singleton LoF SVs affecting mutationally constrained genes and genes expressed in the adult adrenal gland. DSB, double-strand break. (**F**) In Ewing sarcoma, no single category of SVs was enriched in patients or controls at Bonferroni significance, although multiple categories of potential biological interest were nominally significant.

factors, rather than just genomic factors such as gene size, mediate their enrichment. Furthermore, we found that more specific categories with relatively greater effects also exhibited enrichment for these gene sets, which highlights that the subsets of SVs that most strongly increase risk likely also affect these pathways (Fig. 4C and fig. S11, A and B).

We next reasoned that singleton germline SVs may contribute to disease risk by directly altering gene expression in the tissue of origin for each histology. For both neuroblastoma

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

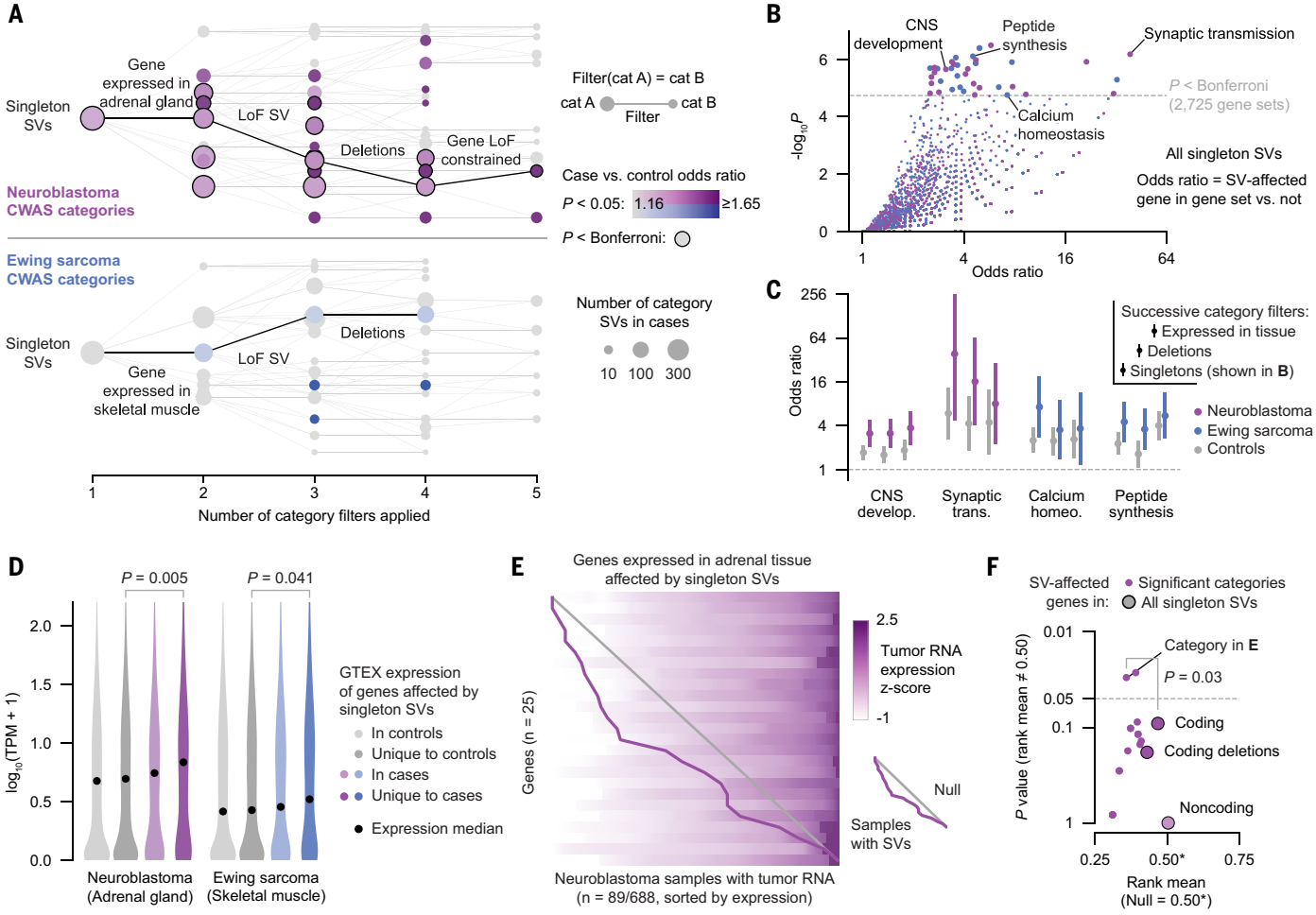


Fig. 4. Ultrarare germline SVs in pediatric cancer patients dysregulate key pathways and gene expression in premalignant tissues and in tumors. (**A**) Diagrams of CWAS category relationships. Singleton SV categories are depicted as nodes and filters as edges. The *x* axis represents the number of filters applied to each category, and the vertical space between categories is proportional to their similarity (Jaccard index). Example filter paths containing at least nominally significant categories are shown in black. (**B**) Gene set enrichment results for risk-carrying singleton SVs and GO biological processes. ORs were computed by a Fisher's exact test of an SV-affected gene being in the category and also in the gene set. (**C**) Gene set enrichment results for select gene sets and categories in patients and controls. The three categories shown for each gene set represent singleton SVs, singleton deletions, and singleton deletions affecting genes expressed in the tissue of origin (successively applied filters). Bars represent the 95% CI. (**D**) Expression in GTEx v8 for genes affected by singleton SVs in controls, exclusive to controls, in patients, and exclusive to patients (*46*). *P* values are from a Wilcoxon test. (**E**) Expression heatmap of genes expressed in adrenal tissue that were affected by singleton SVs in the subset of neuroblastoma patients with tumor RNA data available. Each row represents a gene, and each column represents a sample; samples within each row are ordered by expression z-score. Samples with SVs (one per row) are connected by a purple line. The diagonal gray line indicates expectation under a uniform null. (**F**) Effect on tumor expression of SVs in significant neuroblastoma categories. The *y* axis represents the *P* value of comparing to the null uniform distribution (no effect on expression). Larger points represent higher-level groupings of SVs. The indicated *P* value is from bootstrap-comparing the category from (E) to the background of all coding SVs. See also fig. S11.

and Ewing sarcoma, we found that the genes disrupted by singleton SVs exclusively in affected patients were expressed at higher levels in normal tissue than those exclusively disrupted in controls, suggesting that risk-carrying SVs preferentially affect genes that tended to be more highly expressed than other genes in the cancer's tissue of origin (neuroblastoma patient–specific versus control-specific median transcripts per million (TPMs) = 5.8 versus 3.9; Wilcoxon $P = 0.005$; Fig. 4D). Most genes disrupted exclusively by singleton SVs in patients were also broadly expressed in other tissues, mirroring observations of pathogenic germline variants in severe developmental disorders (*48*) (fig. S11C). Although RNA from normal tissue was not available from patients, a subset of neuroblastoma tumors from GMKF patients ($n = 89$ of 688) had previously undergone transcriptome profiling, allowing us to assess the effect of SVs on tumor expression (*28*). To assess the overall impact of singleton SVs on genes with varied degrees of expression, we converted gene expression values to a sample-normalized percentile rank (0 to 1) across all 89 samples and assessed the average rank of samples with SVs. We found that singleton SVs tended to decrease expression of their affected genes in the patients' corresponding tumor, a result that was significant for two categories (for example, singleton SVs affecting genes expressed in adrenal tissue had a rank mean of 0.36 versus 0.50 for the uniform null, $P = 0.03$; Fig. 4E and fig. S11, D and E). We also found that singleton SVs that affected genes expressed in adrenal tissue significantly decreased expression in tumors more than coding SVs generally (rank mean = 0.35 versus 0.47; permutation test $P = 0.034$). Noncoding SVs, by contrast, did not exhibit an

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

effect on expression of their closest gene (rank mean = 0.50; $P \sim 1$; Fig. 4F). Taken together, these results showed that singleton SVs contribute important disease risk and that these SVs affect genes expressed in the disease tissue of origin, dysregulate relevant biological processes, and may even affect eventual tumor gene expression.

Prevalence of germline SVs that disrupt known cancer genes

A growing body of literature has called for WGS to become the first-tier diagnostic screen in diseases with suspected genetic etiologies (*37*, *51*, *52*). The clinical utility of WGS will hinge on its ability to detect pathogenic variants that are inaccessible to technologies presently used in clinical diagnostics. Because germline SVs are not routinely examined in genetic testing for pediatric solid tumors and their prevalence has not been systematically examined at the resolution of WGS, we sought to understand the spectrum of SVs in our dataset that affected recognized germline cancer predisposition genes (CPGs) and COSMIC cancer genes (*5*, *14*–*16*). We identified 69 patients (3.5%) who carried SVs with AF <0.1% that affected genes previously implicated in cancer predisposition and oncogenesis (fig. S12), including LoF SVs in DNA damage repair genes such as *PALB2* and *BARD1* (Fig. 5A), which mirrors previous reports of pathogenic short variants in similar genes and pathways (*53*–*55*) (fig. S13). We additionally found that several of these SVs that affected established CPGs, such as *PHOX2B* and *FANCA* (*53*, *56*), were inherited from unaffected parents and therefore likely represent incompletely penetrant risk factors (Fig. 5B). We also noted duplications and complex SVs converging on key members of oncogenic pathways in pediatric solid tumors, such as RAS-MAPK (Fig. 5C) (*57*), which correlated with altered gene expression in patients where tumor RNA was available (Fig. 5D). A subset of the rare SVs that affected CPGs or COSMIC cancer genes were low-frequency polymorphisms in the general population, including duplications of *RAF1*, *CHEK2*, *BRCA1*, and *ERCC2*, all of which were observed at greater frequencies in patients than in controls (for example, the *ERCC2* duplication was present in 5 of 772 Ewing sarcoma patients versus 7 of 4575 controls; $P = 0.051$; OR = 3.35; Fig. 5E).

Many pediatric tumors are associated with recurrent somatic genome rearrangements, such as *EWSR1-FLI1* fusions in Ewing sarcoma and *MYCN* amplifications in high-risk neuroblastoma (*50*, *58*, *59*). We asked whether rare germline SVs might act as precursor lesions to these recurrent somatic rearrangements, but we found no evidence of germline SVs enriched

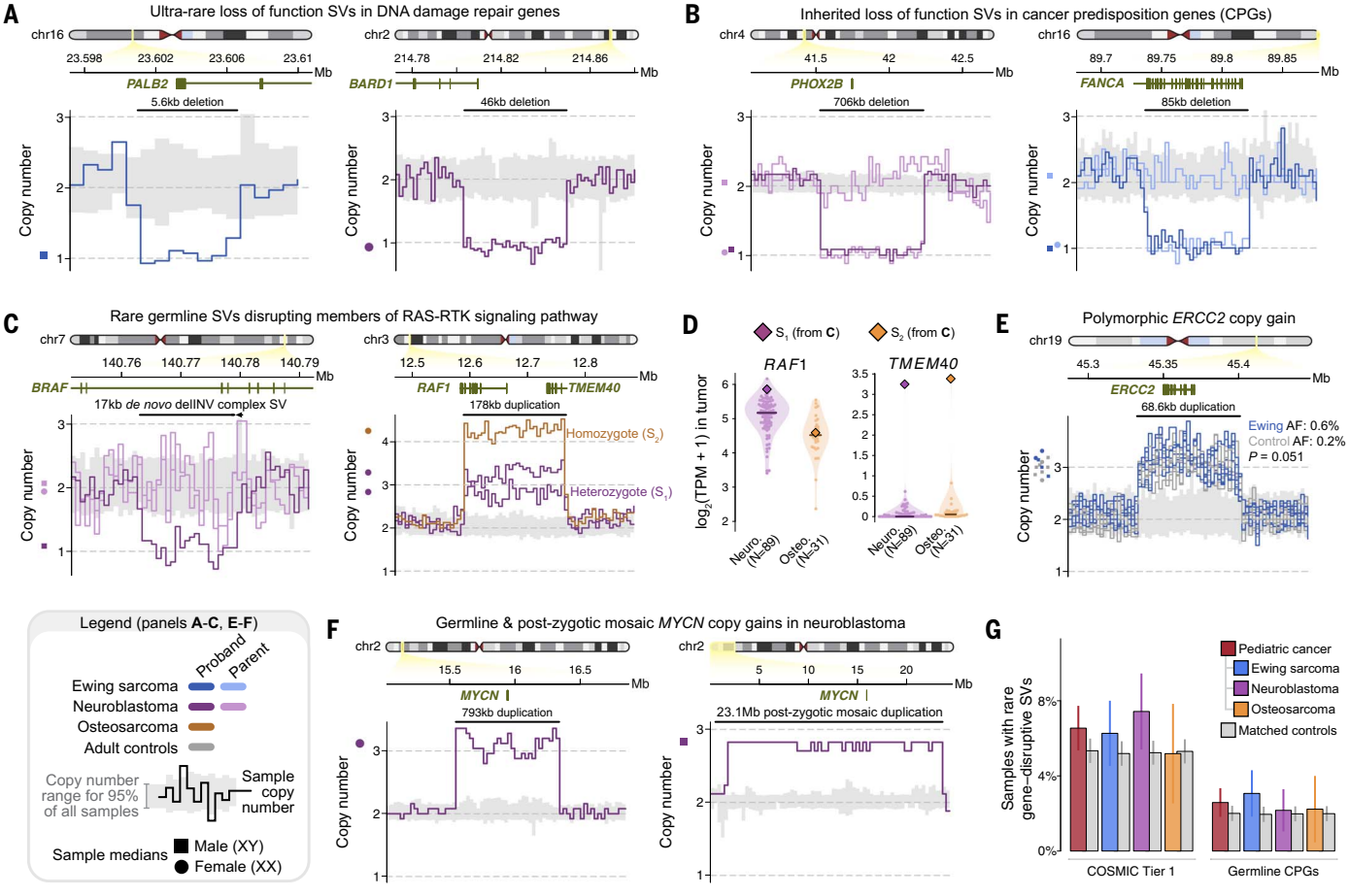


Fig. 5. Identification of gene-disruptive germline SVs in CPGs and COSMIC cancer genes in pediatric patients with solid tumors. (**A**) We found germline LoF deletions in DNA damage repair genes, such as *PALB2* in Ewing sarcoma and *BARD1* in neuroblastoma. (**B**) We also observed germline LoF SVs of known CPGs, such as *PHOX2B* in neuroblastoma and *FANCA* in Ewing sarcoma, that were carried by affected children but were inherited from unaffected parents. (**C**) RAS-MAPK genes were affected by germline SVs, including a singleton de novo complex SV resulting in a two-exon deletion of *BRAF* in one neuroblastoma patient and an ultrarare polymorphic duplication over *RAF1* in three unrelated patients. delINV, deletion-flanked inversion. (**D**) The germline *RAF1* duplication from (C) was associated with high *RAF1* expression in a neuroblastoma tumor and increased expression of *TMEM40*, the gene upstream of the *RAF1* promoter, in neuroblastoma and osteosarcoma tumors. (**E**) A rare polymorphic duplication predicted to result in copy gain of *ERCC2*, a DNA damage repair gene, was found at a threefold higher frequency in Ewing sarcoma patients than in controls. (**F**) We discovered a de novo germline *MYCN* duplication and a likely postzygotic *MYCN* duplication in two unrelated neuroblastoma patients. (**G**) The overall rates of gene-disruptive rare SVs in CPGs and COSMIC cancer genes were not significantly higher in patients relative to controls. See also figs. S12 to S15.

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

in patients versus controls at these loci (fig. S14). However, we did identify a single neuroblastoma patient who carried an apparently de novo 793-kb duplication that resulted in copy gain of *MYCN* that was statistically consistent with a full germline event (estimated copy number = 3.09; 95% CI: 2.78 to 3.41; Fig. 5F). Joint analysis with the patient's matched tumor demonstrated a high amplification (>45 copies) of *MYCN* in the tumor and no detectable tumor-in-normal contamination (fig. S15A). While scrutinizing the *MYCN* locus, we also uncovered an apparently mosaic 23.2-Mb CG duplication of *MYCN* in a second, unrelated neuroblastoma patient; this very large event was initially excluded by our GATK-SV pipeline because it was statistically inconsistent with a true germline event (estimated copy number = 2.79; 95% CI: 2.60 to 2.98), but the near-integer copy number might imply an early postzygotic mutational origin because tumor-in-normal contamination was unlikely in this saliva-derived normal sample. Although we were unable to molecularly validate the presence of these *MYCN* duplications owing to the lack of additional germline DNA, we note that one other ultrarare *MYCN* germline duplication had been reported in a neuroblastoma patient in a prior study (*23*), which supports the relevance of germline or early postzygotic mosaic *MYCN* duplications in neuroblastoma pathogenesis.

Given the numerous examples of rare SVs predicted to disrupt established cancer genes in affected patients, we anticipated that the collective rates of rare SVs in CPGs and COSMIC cancer genes would be elevated in patients compared with controls. However, we did not find this to be the case (CPGs, pan-cancer OR = 1.26; P = 0.24; COSMIC cancer genes, pan-cancer OR = 1.15; P = 0.27; Fig. 5G and fig. S15B), which may imply that the totality of germline SVs across all presently recognized CPGs and COSMIC cancer genes will not collectively explain a large fraction of the missing heritability of these pediatric solid tumors. Taken together with the enrichment of rare gene-disruptive SVs in constrained genes beyond CPGs and COSMIC genes (fig. S15C), we concluded that there are likely undiscovered CPGs that confer disease risk specifically for these tumors. We systematically searched for candidate CPGs by conducting gene-based association tests for 15,544 autosomal, protein-coding genes and identified just one candidate that exceeded exome-wide significance ($P < 3.2 \times 10^{-6}$; fig. S16): Rare LoF deletions that affected the promoter and first exon of *KL*, a gene previously implicated in aging that is silenced in some adult cancers (*60*, *61*), were associated with sharply increased risk for neuroblastoma ($P = 2.2 \times 10^{-8}$; OR = 123.5; 95% CI lower bound = 5.9). However, this candidate CPG should be interpreted cautiously in the absence of independent replication in an external dataset, and larger sample sizes will be needed to discover CPGs in these rare histologies.

The contribution of noncoding germline SVs to pediatric cancer predisposition

Our analyses thus far had established that very large and gene-disruptive rare germline SVs were enriched in pediatric cancer patients compared with controls. However, we reasoned that yet more of the missing heritability of pediatric solid tumors may be hidden in the ~98% of the genome that does not encode proteins. To test this hypothesis, we adapted the CWAS framework to focus exclusively on SVs with no predicted coding consequence and added a functional filter that captured epigenetic and other regulatory genomic annotations (fig. S17 and tables S9 to S11), which allowed us to systematically categorize noncoding SVs in each histology. We enumerated 26,400 possible categories and found that 7901 categories in neuroblastoma and 6233 categories in Ewing sarcoma met our minimum "testable" threshold of >10 SVs in total across patients and controls. We once again evaluated the burden of autosomal SVs in patients relative to controls for all categories and identified noncoding categories that met Bonferroni significance ($P \leq 5.4 \times 10^{-5}$) after estimating the number of effective tests per disease with the same permutation strategy as in our earlier coding CWAS (*38*, *45*).

We discovered multiple Bonferroni-significant categories associated with increased neuroblastoma risk, all of which involved noncoding singleton SVs overlapping topologically associating domain (TAD) boundaries defined from healthy adult adrenal gland tissue (Fig. 6A and table S12) (*62*). By contrast, no categories reached Bonferroni significance in Ewing sarcoma despite our CWAS test statistics being well calibrated (Fig. 6, B and C, and table S13). In neuroblastoma, we observed a significantly stronger effect size for noncoding singleton SVs intersecting TAD boundaries relative to all noncoding singletons ($P = 4.3 \times 10^{-4}$, Welch's t test), whereas this difference was not seen in Ewing sarcoma (P = 0.39) (Fig. 6D). Taken together, the enrichment of noncoding germline SVs overlapping tissue-specific TAD boundaries supported the potential disruption of these domains—and presumably the dysregulation of the genes residing within these TADs—as a germline risk factor for the development of neuroblastoma.

We further scrutinized the noncoding TAD boundary SV categories in neuroblastoma and found that their effects were consistent in both the St. Jude (P = 0.004; OR = 1.25) and GMKF (P = 0.001; OR = 1.34) cohorts when analyzed separately (Fig. 6E). We examined which noncoding filters resulted in the categories that carried the strongest disease risk and found that deletions of TAD boundaries near adrenally expressed genes led to the greatest increase in SV effect sizes (Fig. 6F). However, we did not find any GO terms with genes that were significantly overrepresented near noncoding SVs overlapping TAD boundaries, nor did these SVs have an appreciable effect on nearby gene expression at our sample size, which reflects the major challenges of predicting the functional consequences of noncoding germline SVs on gene regulation mediated by alterations to three-dimensional chromatin domains (*63*–*66*).

Disease heritability attributable to rare germline SVs

Lastly, we sought to quantify the aggregate contribution of rare germline SVs to pediatric solid tumor predisposition by estimating heritability—that is, phenotypic variance—attributable to this variant class (*67*–*71*). We estimated heritability separately for neuroblastoma and Ewing sarcoma among individuals of European ancestry in our study. For each disease, we fit an L_2-penalized regression model with fivefold cross-validation to predict disease status from the totality of SV categories tested in the coding and noncoding CWAS for each disease plus two separate features for sex and large, rare, unbalanced SVs, as well as an interaction term to capture the apparent sex-bias of these large SVs that we discovered earlier. We transformed in-sample estimates of variance explained to liability scale heritability using established statistical procedures and published disease prevalence estimates (*72*–*74*). When taking the best-fit held-out test fold from cross-validation as a likely underpowered lower bound and the full in-sample estimate as a likely overfit upper bound, we estimated that rare germline SVs collectively account for 1.5 to 5.6% of heritability for neuroblastoma and 1.1 to 4.6% of heritability for Ewing sarcoma (table S14). These estimates are inherently limited by relatively small sample sizes and do not account for rare germline SV types and categories not accessible to the short-read WGS approach used in our present study. Nonetheless, these estimates firmly establish rare germline SVs as important risk factors for pediatric extracranial solid tumors but also place an upper bound on realistic expectations for the proportion of "missing" heritability that is likely to be explained by rare germline SVs in these diseases.

Discussion

In this study, we systematically dissected the wide-ranging influences of rare germline SVs on pediatric solid tumor disease risk. Prior studies have implicated rare SNVs and/or indels in known cancer genes in the pathogenesis of these aggressive malignancies (*5*, *53*, *55*), but much has remained elusive about the breadth of other germline risk factors. By harmonizing

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

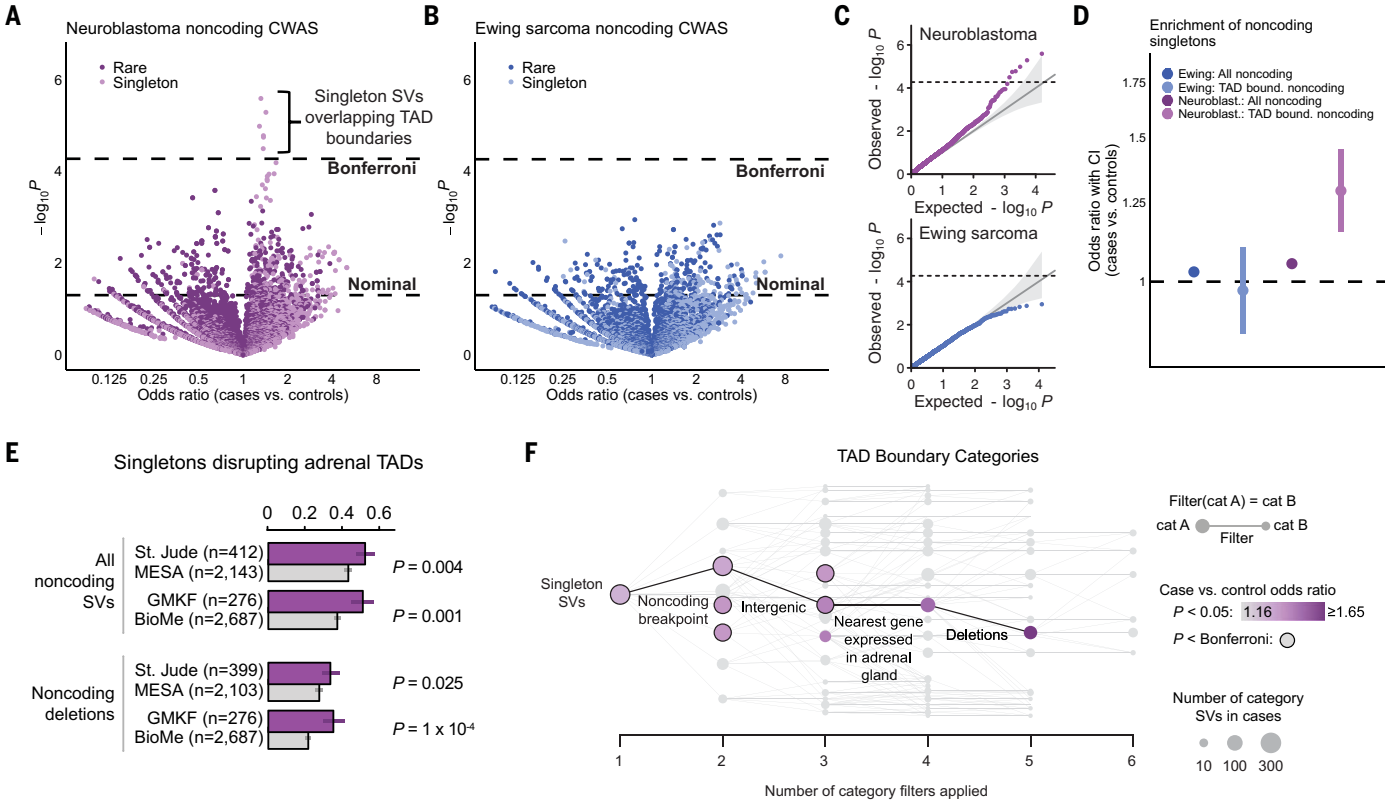


Fig. 6. Noncoding germline SVs that affect TAD boundaries in disease-relevant tissues increase risk for neuroblastoma. (**A**) We performed a CWAS for rare, noncoding SVs in neuroblastoma patients versus controls and found that singleton germline SVs overlapping adrenal gland–derived TAD boundaries were significantly enriched in patients after correcting for the estimated number of effective CWAS tests (Bonferroni). See also fig. S17. (**B**) No noncoding SV categories reached the threshold of Bonferroni significance for enrichment in Ewing sarcoma patients relative to controls. (**C**) Quantile-quantile plots for all noncoding CWAS categories for neuroblastoma (top) and Ewing sarcoma (bottom) demonstrated good calibration of CWAS test statistics. The dashed horizontal line corresponds to Bonferroni significance. The shaded area is the 95% CI of expected *P* values under a uniform null. (**D**) We observed a significantly stronger effect size for noncoding singleton SVs intersecting TAD boundaries defined in the putative tissue of origin (adrenal gland) relative to all noncoding singletons in neuroblastoma ($P = 4.3 \times 10^{-4}$, Welch's *t* test), whereas this difference was not seen in Ewing sarcoma ($P = 0.39$), consistent with the lack of a TAD-related association in the Ewing sarcoma CWAS. (**E**) The enrichment in neuroblastoma patients for singleton germline SVs intersecting TAD boundaries was significant in the St. Jude and GMKF cohorts when analyzed separately. Error bars indicate 95% CIs. (**F**) Effect sizes for singleton SVs intersecting TAD boundaries in neuroblastoma increased monotonically when subsetting to deletions in proximity to genes expressed in adrenal tissue.

10,590 germline genomes, applying GATK-SV for comprehensive genome-wide SV discovery, implementing rigorous quality control, and leveraging an ancestry-aware case-control association framework, we show that rare germline SVs across the coding and noncoding genome are important risk factors for neuroblastoma, Ewing sarcoma, and osteosarcoma.

We found that large, unbalanced germline SVs were risk factors for all three pediatric solid tumor histologies. This signal was stratified by sex, with male patients having a four-fold higher incidence of these SVs than male controls, whereas female patients exhibited rates comparable to those of female controls. Previous epidemiological studies have reported a slight male predominance across pediatric cancers, including the three histologies in our study (male-to-female incidence relative rates: 1.1 to 1.7) (*75*, *76*), but the reasons for this sex bias have remained opaque. The sex-specific pattern of large, unbalanced SVs that we discovered may imply that there are specific categories of autosomal germline variants that differentially increase cancer risk in males and females, although additional studies are needed to unambiguously confirm this trend. In contrast to developmental disorders, where large CNVs at dozens of established loci are associated with disease risk (*43*, *44*), the large, unbalanced SVs observed in pediatric cancer patients did not coalesce at specific loci, and 91% did not overlap any genes with known roles in cancer. Instead, these SVs were scattered throughout the genome without a discernible pattern, hinting at a possible role in genomic instability during tumor initiation that echoes the previously reported associations of pathogenic germline rare variants in DNA damage repair genes in these same histologies (*53*). Similarly, the lack of any association with equally large but balanced SVs (such as inversions) was curious because large inversions and translocations are strongly associated with other pediatric congenital anomalies (*42*). This discordance suggests that the pathogenic effects of large SVs in pediatric cancers are specific to dosage imbalances caused by exceptionally large germline CNVs.

Rare gene-disruptive germline SVs were also enriched in pediatric solid tumor patients relative to adult controls, and this effect was most pronounced for ultrarare SVs. We implemented a comprehensive CWAS framework to identify specific subsets of SVs contributing to disease risk, which pinpointed LoF singleton SVs in mutationally constrained genes as an especially strong category of risk-carrying SVs in neuroblastoma. Further dissection of the CWAS results revealed nominally significant subsets of SVs with even greater effect sizes and higher biological specificity, such as SVs predicted to cause loss of function of DNA damage repair

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

genes in neuroblastoma and Ewing sarcoma. Mutationally constrained and DNA damage repair gene sets are therefore promising starting points for future discoveries of CPGs specific to pediatric solid tumors. We also found that risk-carrying SVs appeared to affect cellular functions with plausible connections to disease initiation, such as neurogenesis and synaptic transmission in neuroblastoma. By integrating our results with RNA data from normal tissues of adult donors and matched patient tumors, we further found that these risk-carrying SVs tended to affect genes that were highly expressed in the disease's putative tissue of origin and even led to altered expression in the eventual tumors from SV-carrying patients.

We identified dozens of germline SVs that affected established cancer genes and pathways, such as *PHOX2B*, *BARD1*, *MYCN*, and RAS-MAPK genes in neuroblastoma or DNA damage repair and Fanconi Anemia genes in Ewing sarcoma, which demonstrates that germline SVs converge onto many of the same genes and pathways that have been implicated in prior germline SNV and/or indel studies of these same diseases. However, the rate of rare gene-disruptive germline SVs across all known CPGs and COSMIC cancer genes was not significantly higher in patients relative to controls, bolstering the argument along with our CWAS results that there are likely CPGs specific to these pediatric malignancies that have not yet been discovered. Our gene-based rare SV association tests identified just one candidate CPG: loss of function of *KL* in neuroblastoma, which may plausibly function as a tumor suppressor on the basis of existing experimental evidence (*60*, *61*) but requires replication and further functional evaluation in these contexts. The robust identification of yet more undiscovered CPGs will likely require greater power afforded by the integration of SNVs, indels, and SVs in larger cohorts.

Lastly, we tackled the open question of how noncoding germline SVs may confer disease risk. By incorporating tissue-of-origin epigenetic annotations into our CWAS framework, we discovered that rare noncoding SVs overlapped TAD boundaries in adrenal gland tissue more frequently in neuroblastoma patients than in controls. Deletions of TAD boundaries close to adrenally expressed constrained genes had larger effect sizes than TAD-boundary SVs generally, suggesting that dysregulation of TADs specifically affecting developmental genes expressed in the putative tissue of origin may be important in neuroblastoma pathogenesis. The lack of significant noncoding SV categories in Ewing sarcoma was intriguing because comparable sample sizes implies that rare noncoding germline SVs play a larger role in neuroblastoma oncogenesis, but we cannot rule out effects specific to certain cell types or developmental time points that are not represented in existing public datasets.

Our study had notable limitations. As with most rare disease research, the scarcity of large patient datasets limited our power for gene- and locus-level association tests and prevented the replication of our findings in an independent dataset. We also did not focus on common SVs in this study because their effect sizes are expected to be much smaller and require greater sample sizes than rare variant analyses. Lastly, our noncoding and functional analyses relied on existing datasets derived from healthy adult tissues or matched tumors of patients from this study, neither of which are perfect proxies for healthy fetal or pediatric tissues matched on developmental timing. Ongoing germline WGS efforts in other pediatric cancer populations will enable further generalization and replication of our findings, as well as the discovery of additional disease-specific risk factors.

This work represents a substantial step forward in explaining the missing heritability of pediatric solid tumors through the systematic characterization of germline SVs at nucleotide resolution. Our findings indicate that rare germline SVs explain a modest but meaningful component (1.1 to 5.6%) of an individual's total liability for childhood cancer, which lays the groundwork for the eventual incorporation of germline SV analyses into routine diagnostic screening and clinical practice. Future work focused on integrated genome-wide evaluation of SVs with SNVs and indels in the same patients will enable a more complete understanding of germline factors that contribute to the development of pediatric cancer and, in turn, a better understanding of the central oncogenic programs that give rise to these malignancies.

Materials and methods

Ethics approval and consent to participate

Written informed consent from individuals and institutional review board approval, allowing comprehensive genetic analysis of germline samples, were obtained by the original studies that enrolled individuals. The secondary genomic analyses performed for this study were approved under Dana-Farber Cancer Institute institutional review board protocols 21-143 and 20-691. This study conforms to the Declaration of Helsinki.

Sample selection

Germline WGS data from neuroblastoma (NBL), Ewing sarcoma (EWS), and osteosarcoma (OS) patients were aggregated across the Gabriella Miller Kids First (GMKF; dbGaP study "Genomic Sequencing of Ewing Sarcoma" phs000804.v1.p1 with individuals enrolled through Project GENESIS at the Huntsman Cancer Institute and Children's Oncology Group protocol AEPI10N5; dbGaP study "Discovering the Genetic Basis of Human Neuroblastoma: A Kids First Project" phs001436.v1.p1), St. Jude Cloud (Pediatric Cancer Genome Project, St. Jude Lifetime, Genomes for Kids, and Childhood Cancer Survivor Study), and International Cancer Genome Consortium (ICGC; "Bone Cancer – UK", BOCA-UK) studies. Germline WGS for a total of 2277 patients (996 NBL, 887 EWS, 394 OS) were aggregated before quality control, outlier exclusion, and relatedness inference (*25*–*29*). Proband patients from the GMKF study were augmented with matched germline WGS from parents where complete parent-child trios were available (405 trios in NBL, 283 trios in EWS). We selected 5320 adult control samples from the TOPMed BioMe (2842; dbGaP study "NHLBI TOPMed - NHGRI CCDG: The BioMe Biobank at Mount Sinai" phs001644.v3.p2) and MESA studies (2478; dbGaP study "NHLBI TOPMed: MESA and MESA Family AA-CAC" phs001416.v3.p1) (*30*). Controls were selected based on reported ancestry to approximately match the proportions of genetic ancestry among NBL, EWS, and OS patients to enable maximally powered ancestry-matched association analyses. Finally, 539 parent-proband WGS trios (1617 individuals) from the 1000 Genomes Project (1000G) were included for batched analysis with the selected patients and controls to enable downstream quality control, SV benchmarking, and robust ancestry inference (*31*). Across these cohorts, WGS was performed on DNA extracted from predominantly three main sources, including peripheral blood, saliva, or—in the case of 1000G samples only—lymphoblastoid cell lines (fig. S1B).

Single-sample SV preprocessing

All WGS samples were processed according to the recommended default configuration of GATK-SV v0.24.3-beta, which has been described in detail in previous publications and is available from GitHub (see Acknowledgments section Data and materials availability) (*17*). The initial single-sample step of GATK-SV executes three complementary SV discovery algorithms in parallel for each WGS sample: Manta v1.5.0 (*77*), MELT v2.0.5 (*78*), and Wham v1.7.0-311-g4e8c (*79*). Additionally, raw SV evidence metrics were collected genome-wide for each WGS sample across four evidence classes, including anomalous paired-end (PE) reads, split reads (SR), read depth (RD), and B-allele frequencies (BAF) of SNVs and indels.

Sample quality control, batching, sex inference, and read-depth CNV discovery

The first half of the GATK-SV pipeline operates on batches of 100 to 1000 samples. We therefore undertook a joint batching and sample quality control process to simultaneously exclude low-quality WGS samples while dividing our cohort into homogeneous batches of samples for

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

subsequent SV discovery and genotyping. This procedure roughly follows the protocol used for the most recent release of the Genome Aggregation Database (gnomAD) (*17*). We began by excluding extreme outlier WGS samples whose technical qualities deviated markedly from all other samples. We defined these "global" outlier samples as those meeting any of the following criteria:

1) Dosage bias (∂; coverage nonuniformity) less than the first quartile (Q1) minus eight times the median absolute deviation (MAD) across all samples;

2) ∂ greater than the third quartile (Q3) plus eight MAD across all samples;

3) median WGS coverage <15 or >60;

4) WGS insert size < Q1 − 8*MAD;

5) WGS insert size > Q3 + 8*MAD;

6) highly variable autosomal ploidy estimates, defined as the median absolute copy-number deviation from diploid exceeding >0.1 across all 22 autosomes;

7) ambiguous sex chromosome ploidy; and

8) inferred XX or XY karyotype that disagreed with a self-reported binary sex (inferred XX but reported "male" or inferred XY but reported "female").

In total, 364 samples (3.4% of total cohort) failed one or more of the criteria above and were excluded outright from the study. For clarity, we provide a REMARK diagram of sample attrition at all stages of post hoc filtering in fig. S2.

After global outlier exclusion, we inferred sex chromosome ploidy for all samples using GATK-SV's ploidy inference functionality (Fig. 2B and fig. S6) before dividing the remaining 10,226 samples into 20 batches of 509 to 513 samples per batch. We assigned samples to batches roughly matched on ∂ and genome-wide median coverage while ensuring that all 20 batches also had the same proportions of male patients, female patients, male controls, and female controls. Once all samples were assigned to a batch, we performed a second layer of outlier exclusion within each batch. We excluded samples from each batch if they violated any of the following criteria, all of which were computed while restricting to samples in a given batch:

1) ∂ < Q1 - 7*MAD;

2) ∂ > Q3 + 7*MAD;

3) WGS insert size < Q1 − 7*MAD; and

4) WGS insert size > Q3 + 7*MAD.

A total of 77 additional samples failed these criteria; these samples were removed from their batch and excluded outright from the remainder of the study (fig. S2).

This joint batching and outlier exclusion procedure retained a total of 10,149 samples across 20 batches of 491 to 513 samples per batch. After batch assignment, we conducted RD-based CNV discovery for all samples per batch using GATK-gCNV v4.2.6.1 (*80*) and cn.MOPS v1.40.0 (*81*), both of which were executed using GATK-SV recommended default parameters. The CNV segments predicted by GATK-gCNV and GATK-SV were merged per sample using GATK-SV default parameters to generate a single set of RD-based CNV predictions for each sample.

Joint SV filtering, clustering, and genotyping

We used GATK-SV to perform joint SV filtering and genotyping across all 10,149 samples passing initial quality thresholds. This pipeline has been extensively detailed in prior studies (*17*) and all code and parameters are available from GitHub. Accordingly, only an overview is outlined below; please refer to the GATK-SV documentation and source code for exact technical specifics. The GATK-SV pipeline involves a series of operations for each batch of samples before integrating the results per batch across all batches in a cohort. We performed the following operations per batch in the order listed below:

1) SV calls from each algorithm were clustered across all samples in a batch using default clustering GATK-SV parameters.

2) Raw WGS evidence (PE, SR, RD, BAF) was collected for all samples for each candidate SV from step 1.

3) A series of random forest classifiers were trained on the output of step 2 to predict true positive SVs and exclude technical false positives. After training, these classifiers were applied to all SVs from step 1 and only predicted true positive SVs were retained for subsequent steps.

4) Outlier samples were excluded from each batch based on SV counts after SV filtering from step 3. Outliers were defined as samples with SV counts either <Q1 − (10*MAD) or >Q3 + (10*MAD). Outliers were assessed separately for every combination of algorithm (e.g., Manta) and SV type (e.g., deletion), and any sample flagged as an outlier for any of these strata was excluded from the batch and from the remainder of the study. In total, 215 (2.1%) of all samples were excluded at this stage; the remaining 9934 samples were retained for subsequent processing and analysis (fig. S2).

5) SV calls from the three PE/SR-based algorithms—Manta, MELT, and Wham—were clustered within each batch to produce a single set of PE/SR-based SV calls for each batch.

After completing the above operations per batch, we next unified the SV predictions from each batch across all samples in the cohort to produce a uniformly genotyped SV callset for the entire cohort. This required several sequential operations performed cohort-wide, as follows:

6) A nonredundant list of all candidate SVs across all batches was constructed separately for PE/SR- and RD-based SV calls. No clustering was performed in this step.

7) Each sample was assigned a genotype and genotype quality (GQ) score for every candidate SV present in the output from step 6. These genotypes were assigned using GATK-SV's default SV genotyping module, which reuses the SV evidence thresholds learned by the random forest classifiers from step 3 to predict biallelic genotypes based on the strength of evidence supporting each SV.

8) After genotyping, SVs were clustered across all batches using default GATK-SV parameters. This algorithm first clusters PE/SR- and RD-based SVs separately across all batches before subsequently clustering the cohort-wide PE/SR- and RD-based SVs into a single unified file of nonredundant, clustered SVs with genotypes for every sample.

9) Complex SVs were resolved using the standard GATK-SV approach, which involves first identifying clusters of proximal SV breakpoints in the same sample(s) that may represent complex SVs before comparing the coordinates of these breakpoints to a reference taxonomy of 16 possible germline complex SV allele configurations using svtk resolve (*17*, *38*, *82*). Candidate complex SVs involving predicted segments of deletion or duplication were further interrogated by regenotyping the predicted CNV interval(s) based on RD evidence for all predicted carrier samples; complex SVs that were not orthogonally confirmed by RD regenotyping were excluded as "unresolved" variants.

10) The final module of the core GATK-SV pipeline, "CleanVCF," applied a series of minor adjustments to handle myriad edge cases, such as adjusting genotypes for male and female samples on allosomes, detecting and regenotyping multiallelic CNVs, and assessing predicted copy numbers for samples with two independent overlapping CNVs (i.e., predicted compound heterozygotes).

The final output of the core GATK-SV pipeline was a single VCF comprising all SVs detected in the cohort with predicted genotypes for all 9934 samples.

Post hoc SV callset refinement

After joint SV discovery and genotyping across all samples, we performed a series of post hoc cleanup steps to further improve the specificity of our SV calls and genotypes. We largely followed examples set by gold-standard GATK-SV callsets generated by gnomAD and the NIH *All of Us* project (https://support.researchallofus.org/hc/en-us/articles/14941865780500-Benchmarking-and-quality-analyses-on-the-All-of-Us-short-read-structural-variant-calls) (*17*, *83*); we deviated slightly in some respects to emphasize specificity at the expense of sensitivity, reasoning that spurious false positives were less desirable than false negatives given our focus on rare variant associations in rare diseases. We evaluated numerous permutations and alternative approaches, finding the below strategy to achieve a relatively optimal balance of specificity and sensitivity (table S15). In total, these

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

cleanup steps involved a mixture of established and new methods applied in the following order:

1) We first pruned low-quality genotypes using a pre-trained model, FilterGenotypes (v7), that was developed as an extension for GATK-SV in the context of the NIH *All Of Us* WGS dataset (*83*). Detailed documentation for the FilterGenotypes model is available from the All of Us genomic data quality report available on the All of Us User Support Hub (https://support.researchallofus.org/hc/en-us/articles/4617899955092-All-of-Us-Genomic-Quality-Report; see pp. 34 to 37). In brief, FilterGenotypes computes a scaled likelihood (SL) for each genotype, which is proportional to the log-odds that the genotype is a true positive versus a false positive, then optimizes minimum SL thresholds below which genotypes are deemed low-confidence. Low-confidence genotypes are erased and reassigned to no-calls (i.e., "./." in VCF). The FilterGenotypes model was trained on GATK-SV data for 2000 All of Us participants to optimize concordance with SNP microarrays (on all 2000) and long-read WGS (available on a subset of 606 out of 2000 participants. We applied the pretrained "high sensitivity" FilterGenotypes configuration from All of Us directly to the SV callset we generated in the present study. The SL cutoffs from the "high sensitivity" configuration corresponded to an approximate 15% false discovery rate (FDR) in All of Us according to publicly available All of Us benchmarking information (see URL above). We evaluated both "high sensitivity" and "high specificity" configurations, finding that the "high sensitivity" model was optimal when coupled with the additional filtering steps below (table S15).

2) We further excluded residual low-quality genotypes not captured in step 1 by applying a second method, minGQ, which was developed by gnomAD to optimize genotype quality thresholds based on empirical rates of SV transmission between parents and children in complete trio pedigrees (*17*). The minGQ method learns a decision tree of minimum GQ thresholds to retain genotypes given an SV's size, frequency, type, and other technical features. We trained a minGQ filtering model to optimize cutoffs for a FDR of 5% for the pediatric cancer parent-child trios from GMKF that were present in our dataset at this stage. As with the FilterGenotypes model, we also evaluated minGQ in several different configurations, finding this combination of high-sensitivity FilterGenotypes daisy chained to minGQ to be relatively optimal (table S15).

3) All multiallelic CNVs (mCNVs) smaller than 5 kb were excluded from our final callset. We have previously observed that mCNV predictions from GATK-SV below this size threshold are often unreliable (*17*).

4) We implemented a procedure to identify splice junctions from genic retrocopy insertion polymorphisms (GRIPs), which can appear in WGS as deletion SVs precisely spanning exon-exon junctions despite not corresponding to real deletions of endogenous DNA (*84*). We systematically detected these spurious deletion calls by intersecting all deletions lacking RD support with introns defined in the MANE Select v1.2 human gene reference catalog (*85*). Any such deletion with ≥95% reciprocal overlap with an annotated intron was tagged as a likely GRIP splice junction and not included for any downstream analyses.

5) Common SVs (AF >5%) that were predicted to be >500 kb in size were marked as "unresolved" variants; these edge-case SV predictions usually correspond to small polymorphic insertions that were incorrectly resolved as large SVs by GATK-SV.

6) Common reciprocal translocations (AF >1%) were marked as "unresolved" for the same reason as step 5.

7) Any SV that exhibited >20% coverage by loci that had at least one alternate haplotype in the hg38 reference assembly were excluded to forestall possible issues related to variant callers handling multiply mapped reads corresponding to alternative haplotypes.

8) Genotypes for each sample were masked if they had GQ of zero before step 1 (FilterGenotypes) and corresponded to uncommon (AF <5%), biallelic SVs called exclusively by Manta or Wham.

9) SVs were excluded if they exhibited a mean SL <0 and maximum SL <75 across all nonreference samples, were called exclusively by Wham or Manta, were smaller than 1 kb in size, did not exhibit BAF evidence, were uncommon (AF <5%), had a no-call genotype rate (NCR; i.e., rate of pruned or missing genotypes) >0.1%, and did not exhibit overdispersed PE/SR genotyping evidence.

10) We applied a final layer of systematic outlier sample exclusion. For each sample, we collected counts of (i) all SVs, (ii) all SVs per SV type, (iii) all rare SVs, (iv) rare SVs per SV type, (v) deletions with AF <5% between 400 bp and 1 kb in size, (vi) CNVs overlapping transcription start sites, and (vii) CNVs overlapping protein-coding exons. For each strata above, we defined outlier samples on a population-specific basis as samples with SV counts less than Q1 − 5*MAD or greater than Q3 + 5*MAD within each population based on preexisting ancestry estimates (*53*, *86*) and/or self-reported race and ethnicity information if necessary. Samples with unknown or nonspecific ancestry or self-reported race and ethnicity were grouped with European samples for the purposes of this analysis. All SV count strata were treated equally when determining outliers except for strata where all samples had <10 SVs or the population-wide MAD of SV counts was zero; such especially sparse strata were not considered for this outlier definition. Any sample labeled as an outlier in any of the strata or populations above was excluded outright from the study and not included in any subsequent analyses. In total, this process excluded 557 samples, leaving 9377 nonoutlier samples in our overall cohort for downstream steps (fig. S2).

11) Any SV with NCR ≥4% was excluded. This threshold was decreased to NCR ≥0.4% for deletions between 400 bp and 1 kb in size with AF <5%. All samples were included for NCR calculations on autosomes, whereas we only considered female or male samples when defining NCRs on chromosome X or Y, respectively.

12) We excluded 15 CNVs ≥1 Mb with AF <2% after manual review of RD evidence for each CNV as follows. We visualized RD profiles for all 121 biallelic CNVs ≥1 Mb with AF <2% using standard visualization approaches implemented in GATK-SV (*38*, *87*). We found strong visual support in RD profiles for 106 CNVs that were consistent with a germline gain or loss of a full copy (fig. S7). We noted an additional two CNVs that had clear visual RD support but at a sub-integer copy number, potentially indicative of postzygotic mosaic mutation origins and/or sample contamination. Finally, we found no clear RD support for 13 CNVs, concluding that these were technical false positives that had been incorrectly resolved as very large CNVs by GATK-SV. To remain conservative, we retained only the 106 CNVs that exhibited strong integer copy states consistent with germline mutational origins based on RD evidence; the other 15 of 121 CNVs were excluded from the remainder of the study.

13) We excluded four complex SVs based on a manual review of RD profiles similar to step 11. We reviewed all 19 complex SVs that were ≥1 Mb in size, AF <2%, and had at least one predicted CNV segment ≥5 kb. We found strong integer copy-number support for 15 of 19 and found little-to-no convincing evidence for 4 of 19. We did not observe any candidate mosaic SVs in this review.

14) SVs were excluded if they exhibited significant AF bias between pairs of patient or control cohorts (e.g., St. Jude versus GMKF or MESA versus BioMe). We performed a Fisher's exact test of allele counts for pairs of cohorts for each SV with AF <5% in our overall dataset and excluded any SVs with $P < 0.05$. This test was further controlled for genetic ancestry and population stratification on a variant-by-variant basis by restricting each test to the continental ancestry group with the largest number of variant carriers as has been suggested previously (*17*). We did not at any point compare patients and controls in this analysis to avoid confounding technical batch effects with potentially disease-associated SVs.

15) We masked all genotypes for CNVs exclusively with RD evidence ("RD-only CNVs") in samples with unusually high RD-only CNV counts. We first collected counts of RD-only biallelic deletions and duplications in all samples and defined a sample as an RD-specific

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

outlier if their count of RD-only deletions or duplications was >Q3 + 5*MAD within their respective continental ancestry group. These samples were retained in our overall dataset as their counts of non-RD-only SVs were well behaved, but we masked the genotypes of RD-only samples to prevent these samples from disproportionately contributing noise to analyses including RD-only CNVs.

16) Lastly, three samples with canonical oncogenic translocations (one control with *BCR-ABL1*, one Ewing sarcoma patient with *NUP98-PRRX1*, one Ewing sarcoma patient with *EWSR1-FLI1*) and one Ewing sarcoma patient with an excess of apparently mosaic large (>500 kb) CNVs were assumed to be either tumor-normal sample swaps or substantial tumor-in-normal contamination, and therefore excluded outright (fig. S2). To confirm that the *EWSR1-FLI1* translocation detected in the Ewing sarcoma patient was not a germline event, we acquired a banked germline sample of white blood cells from the Children's Oncology Group and performed polymerase chain reaction (PCR) to test for the predicted translocation breakpoints. Endpoint PCR was performed using Q5 High Fidelity DNA polymerase 2X Master Mix (New England Biolabs) in a 20-μl reaction volume consisting of 20 ng of DNA and 200 μM of each forward and reverse primer (IDT). The primers we used are detailed in table S16.

The PCR reactions were performed in a C1000 Touch Thermal Cycler (Bio-Rad): initial denaturation at 98°C for 30 s, followed by 30 cycles of 98°C for 5 s, 64.3°C for 30 s, 72°C for 30 s, and a final extension at 72°C for 2 min. PCR endpoint products were loaded in a 2% agarose gel with 1X TBE Buffer (ThermoScientific) at 75 V for 1 hour. GelRed Nucleic Acid Stain (Biotium) was used to visualize the PCR product on Biorad Gel Dox XR+. Wild-type *EWSR1* and *FLI1* sequences were amplified, but not the *EWSR1-FLI1* translocation sequence spanning the reported breakpoint (fig. S18).

After applying all the 16 criteria above, we retained a total of 284,384 high-quality SVs carried by at least one of 9373 high-quality samples. All subsequent analyses were restricted to this dataset. The code required to apply each of the steps above is freely provided on GitHub (see Acknowledgments section Data and materials availability).

Functional consequence annotation

We next annotated all predicted genic effects for each SV using GATK-SV's inbuilt annotation toolkit SVAnnotate. We used the MANE Select v1.2 gene reference dataset for all genic predictions (*85*). The overall approach used by GATK-SV to annotate genic consequences has been described previously (*17*). The method has since been improved since its original description; full details of the SVAnnotate method are provided online (https://gatk.broadinstitute.org/hc/en-us/articles/9570422529179-SVAnnotate). This method assigns one of 11 possible predicted consequences to each gene overlapped by each SV. For genes where more than one consequence could be assigned, more "severe" consequences are prioritized in a conceptually similar manner to other annotation frameworks for SNVs and indels (*88*) in the following order: LoF, CG, intragenic exonic duplication (IED), partial exon duplication (PED), transcription start site (TSS) duplication, partial gene duplication, spanning inversion, multiallelic SV exon overlap, untranslated region (UTR) overlap, intronic, or promoter overlap. For the purposes of this study, we defined an SV as "gene-disruptive" if it was assigned a LoF, CG, IED, or PED consequence for any protein-coding gene.

Noncoding functional annotations were derived based on prior assays in the putative tissue of origin where available: adrenal gland in NBL and skeletal muscle in EWS. ATAC peaks, consensus enhancer tracks, TAD boundary definitions from HiC data, and H3K27Ac peaks were derived from ENCODE (*62*). Enhancer, genic enhancer, bivalent enhancer, and flanking active transcription start site annotations were used from the Roadmap Epigenomics Project 15-state ChromHMM model (*89*). Additional noncoding annotations included Activity-by-Contact model enhancers (Engreitz Lab) (*90*), ultraconserved noncoding elements (UCNE base) (*91*), recombination hotspots (deCODE from the UCSC Genome Browser) (*92*), and fragile sites (PCAWG consortium) (*93*, *94*). Sources of noncoding annotations are detailed in table S9. We annotated overlap between all noncoding annotation tracks and every SV in our dataset using GATK-SV SVAnnotate, which considers two possible relationships between SVs and noncoding elements: "breakpoint," wherein an SV breakpoint falls directly within the noncoding element, and "span," wherein the SV full spans the entirety of the noncoding element.

Ancestry and relatedness inference

We inferred genetic ancestry for all samples and relatedness between all pairs of samples using Hail v0.2.119. We first restricted to well-genotyped polymorphic autosomal SVs, which we defined here as autosomal SVs that had QUAL >10, NCR ≤0.1%, 0.1% ≤ AF ≤ 99.9%, and Hardy-Weinberg equilibrium (HWE) chi-square $P < 10^{-5}$. We then computed the top 20 HWE-normalized PCs (fig. S2) before inferring kinship coefficients and identity-by-descent proportions for all pairs of samples using Hail's implementation of PCRelate (*95*).

We assigned samples to one of five continental ancestry groups for the purposes of variant frequency annotation. To achieve this, we trained a linear support vector machine (SVM) on the top 10 genetic PCs generated by Hail to cluster samples into the following five populations: African or African-American (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). We trained the SVM classifier on ground-truth population assignments available for samples from the 1000 Genomes project samples and supplemented with samples from the TOPMed MESA cohort that had previously undergone WGS SNV-based ancestry assignment as part of a prior study (*31*, *86*). We validated the performance of our ancestry classifier by comparing its predicted labels to ancestry labels previously inferred from SNV and indel analyses for a subset of these samples not included in the training set (*53*). This ancestry classifier achieved accuracies of 98.5 and 97.1% on the training and validation sets, respectively, with classification errors primarily corresponding to admixed individuals that did not clearly cluster with any one ancestry group in PC1-10. Following training and validation, we applied this classifier to assign predicted genetic ancestry memberships to all samples in our cohort.

Ancestry rebalancing and sample pruning

Before undertaking SV-disease association analyses, we wanted to isolate a strictly unrelated set of ancestry-matched patients and controls for all association tests. We began by excluding all parents from known parent-child trios and all samples from the 1000 Genomes Project. We further pruned all predicted pairs of cryptic (i.e., unreported) first- or second-degree relatives from our dataset by (i) identifying all pairs of individuals with kinship coefficient (Φ) ≥0.1, (ii) tallying the number of times each sample appeared in pairs of apparently related samples from step (i), (iii) excluding the sample with the largest count of apparent relatives from step (ii), and (iv) repeating steps (ii) to (iii) until no more predicted pairs of samples with Φ ≥ 0.1 remained in our cohort. In the case of ties between two or more samples in step (ii), we prioritized retaining affected patients and therefore preferentially removed controls where possible.

Next, within each disease (NBL, EWS, OS, or all histologies combined), we rebalanced the proportions of inferred genetic ancestries between patients and controls by iteratively downsampling controls from the ancestry groups with the largest absolute fold-difference in sample size between patients and controls until a chi-square test of sample counts per ancestry between patients and controls converged to $P < 0.01$. This process resulted in a set of 6727 unrelated patients and controls with approximately equal representation from four major continental genetic ancestries, which we used for all subsequent disease association analyses in this study. Note that slight differences in ancestry composition between the three diseases

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

required very slight deviations in sets of controls used for association testing per ancestry (control counts: NBL = 4830; EWS = 4574; OS = 4805; pan-cancer = 4717), although any sample used as a control for any disease was retained in our final dataset.

SV frequency annotation and definitions for rare SVs

As the final step before downstream association analyses, we used GATK-SV's annotation module to log the AFs of each SV in each of the five continental populations (AFR, AMR, EAS, EUR, SAS) based on the SV genotypes in our cohort, as well as a pooled estimate across all populations. All samples were included for autosomes; samples with sex chromosome abnormalities were excluded when calculating frequencies of SVs on allosomes. We further defined the cross-population maximum ("PopMax") AF as the maximum AF observed in any of the four populations with at least 100 samples in our final dataset (AFR, AMR, EAS, EUR). Additionally, we annotated the frequencies of the SVs detected in our cohort with the AFs of their corresponding SVs in gnomAD v4.1, if any (*17*). We recorded the SV AF from gnomAD for all major populations and further computed a gnomAD-specific PopMax AF across all non-bottlenecked populations with ≥100 samples included in gnomAD v4.1; these populations included AFR, AMR, EAS, non-Finnish EUR, and SAS.

Throughout the downstream analyses in this study, we frequently referred to two different sets of SVs defined by their frequency: "rare" and "singleton." The exact criteria for each of these SV sets is defined as follows:

1) Rare: Biallelic SVs with an in-cohort PopMax AF <1% after excluding first- and second-degree relatives, as well as gnomAD v4.1 PopMax AF <1% (*17*).

2) Singleton: Biallelic SVs with exactly one alternative allele observed across all unrelated samples in our dataset (i.e., AC = 1 in the VCF generated in this study). We further required PopMax AF <1% in gnomAD v4.1 as for rare SVs, reasoning that many singletons in our cohort would likely represent ultrarare polymorphisms in the general population and therefore we were intentionally more permissive with gnomAD frequency criteria for singletons in our cohort.

Final SV callset benchmarking and quality assessments

We performed four analyses to assess the technical quality of our final SV dataset (supplementary text and fig. S4). The specifics of each analysis are provided here:

Frequency comparisons to gnomAD

We assessed the concordance between the AFs of SVs detected in our study and their corresponding AFs in gnomAD v4.1 (*17*) using the same gnomAD AF annotations described above. We performed Pearson correlation tests for SV AFs within each of five continental ancestries that were present in gnomAD and in our study. The one exception was Europeans: given that gnomAD annotates Finnish and non-Finnish Europeans (NFE) separately, we compared gnomAD NFE AFs with the EUR AFs in our study.

Hardy-Weinberg equilibrium

We tested whether the genotype distribution for each autosomal, biallelic SV was consistent with expectations under HWE. For each SV, we performed a chi-square test of genotype counts for each of the five major continental ancestry groups considered in our study using the R HardyWeinberg package v1.7.5 (*96*). We considered an SV to be significantly deviant from HWE if the chi-square *P* value surpassed a Bonferroni significance threshold after accounting for all SVs tested within a given continental ancestry group.

Mendelian transmission

We evaluated the rate of SV inheritance in 859 parent-child trios where all three family members had passed all filters and were present in our final SV callset. We first documented the biallelic SVs with nonreference genotypes in each child's genome and annotated each SV with the genotypes for both parents, only retaining sites with informative (i.e., nonmissing) genotypes in all three family members. Next, for each family, we computed the fraction of SVs present in the child genome that were not found in either parent; this fraction reflects a combination of false positive genotypes in the child and false negative genotypes in one or both parents. Lastly, we computed the median of these fractions across all 859 trios as the overall assessment for the cohort.

Comparisons to long-read WGS

A subset of 24 samples from the 1000 Genomes project included in our study also had germline SV calls based on long-read genome assemblies from the Human Genome Structural Variation Consortium (HGSVC) (*32*). We leveraged these data to assess the precision of the short-read-based SV calls in our study. We applied GATK-SV module 09 to identify matching SVs in our short-read dataset and the HGSVC long-read dataset. This approach has been described elsewhere (*17*, *31*), but in brief it searches for overlapping SVs present in the same samples and applies additional criteria based on SV size and type; generally, smaller SVs (<5 kb) are considered a match based on single-linkage clustering of their breakpoints (±250 bp), whereas larger SVs (≥5 kb) are not required to have breakpoint clustering but instead must exhibit 50% reciprocal overlap. Given that short-read SV calls are less accurate in the most highly repetitive genomic regions (*97*), we restricted this analysis to SVs that did not have more than 10% overlap with annotated segmental duplications or simple repeats in the hg38 primary assembly. After excluding SVs in repetitive loci, we computed the fraction of short-read SVs per genome that were supported by long-read SVs in that same sample and reported the median across all samples as the overall summary metric for the cohort.

Identification of high-confidence de novo SVs

Our Mendelian inheritance benchmarking analyses (see above) confirmed that our final SV callset achieved a genotype FDR on a par with gold standards set by flagship population genomics efforts like the 1000 Genomes Project and gnomAD (*17*, *18*). Although an FDR of ~5% is suitable for robust population-based case-control association analyses, these rates still retain several hundred putative de novo SVs per genome, essentially all of which are likely false positives localized to low-complexity repetitive loci. Thus, we followed a comprehensive post hoc filtering approach established in a series of recent family-based WGS studies that used GATK-SV for genome-wide analyses of de novo SVs (*37*–*39*), with the first of these studies achieving a 99.7% molecular validation rate for all 171 predicted de novo SVs (*38*). These methods are described in detail in their original publications (*37*–*39*); briefly, they seek to substantially reduce the number of candidate de novo SVs per genome through a combination of strict frequency filtering, automated readjudication of raw WGS evidence, and manual visualization of the read-level evidence supporting each candidate.

We conducted a targeted de novo SV analysis on the 330 pediatric cancer trios where all three members passed all quality filters and were included in our final SV dataset (n = 227 Ewing sarcoma and n = 103 neuroblastoma). We first identified all potential de novo SV alleles per proband in each trio before applying AF-based filters to retain (i) singletons in our cohort irrespective of their gnomAD AF and (ii) doubletons (AC = 2) in our cohort that had POPMAX AF <1% in gnomAD v4.1 (*17*). We further filtered these candidate de novo alleles to require (i) fully resolved alternate allele structures, (ii) symmetrical split-read support for all breakpoints (where relevant), and (iii) genotypes that were in agreement with the raw SV calls produced by Manta, Wham, MELT, cn.MOPS, or GATK-gCNV. CNVs from this filtered list of candidate de novo alleles were further scrutinized for overlapping segments of copy-number changes in parental genomes that may not have been correctly clustered or resolved initially by GATK-SV. The final automated filtering step of this de novo pipeline excluded low-quality genotypes for de novo candidates using thresholds as previously defined (*38*, *39*). Evidence

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

supporting candidate de novo alleles passing all of the above in silico filters were manually visualized using the Integrative Genomics Viewer (*98*), a custom implementation of CNView (*87*), and RDTest (*38*). Only SVs with clear evidence supporting a de novo event with no evidence of parental mosaicism were retained as final de novo predictions; these final de novo predictions were further stratified by whether they appeared truly germline heterozygous at integer copy states or whether there was any evidence of possible postzygotic mosaicism at noninteger copy states.

Case-control association testing

Unless otherwise specified, all case-control association tests reported in this study correspond to a logistic general linear model (GLM) with affected status (patient|control) as the outcome variable, count of SVs as the primary predictor variable, and covariates for sample sex, cohort, and genetic PCs 1-3. All patients and their corresponding ancestry-matched controls were included except for samples that were missing >5% of SV genotypes for that test (e.g., samples missing genotypes for >2 SVs out of a set of 40 SVs would be excluded). For tests with extremely sparse SV data causing the GLM to fail to converge or produce coefficients with unbelievably large standard errors (log-odds standard error >10), we reran the association test using Firth's bias-reduced penalized logistic regression as has been implemented in many gold-standard tools for conventional genetic association studies of SNVs and indels (*99*, *100*).

Genome-wide sliding window association tests of large, rare CNVs

We scanned all autosomes for loci that exhibited excesses of large (≥100 kb), rare germline SVs in pediatric cancer patients versus adult controls. We conducted these tests using a sliding window approach as has been recently proposed in similar studies of large, rare CNVs in developmental disorders (*43*). We first segmented all 22 autosomes into 1-Mb windows sliding across each contig in 250-kb steps before excluding any windows with ≥30% coverage by segmental duplications, simple repeats, gaps in the reference assembly, loci with alternative haplotype contigs, and loci with reference patches. For each window, we next identified all rare, large SVs involving at least ≥100 kb of deletion or duplication that overlapped the window. We performed one case-control association test per window for each combination phenotype (NBL, EWS, OS, or pan-cancer) and CNV type (deletions only, duplications only, or all CNVs), resulting in a total of 12 tests per window (fig. S9). We assessed significance after correcting for the total number of nonoverlapping 1-Mb windows tested (n = 2623.5 independent windows; $P < 1.9 \times 10^{-5}$).

Coding category-wide SV association study framework

To systematically evaluate the differential burden of groups of SVs affecting genes in patients relative to controls, we carried out a CWAS of gene-disruptive SVs in EWS and NBL (*38*, *45*). We enumerated 4800 possible coding SV categories per disease based on six combinatorial layers of filters: SV AF, predicted coding consequence, SV type, genic mutational constraint (*48*), gene set membership (*101*), and tissue of origin gene expression (*46*) (tables S3 to S5). For each filter layer, we also included a catch-all "any" filter value that included all SVs. Loss of function constrained genes were defined as genes in the top sextile of the LOEUF metric as reported in gnomAD, whereas missense constrained genes were defined as the top sextile of missense observed/expected (*48*). All protein-coding genes were defined according to MANE Select v1.2 (*85*), and functional gene sets were derived from the Reactome Pathway Database. Tissue-of-origin expressed genes were defined as the subset of genes with a minimum expression of TPM >5 in adrenal gland and skeletal muscle from the GTEx project for NBL and EWS, respectively (*46*). Deletion, duplication, inversion, complex, and insertion SVs with predicted gene-disruptive consequences (copy gain, loss of function, intragenic exon duplication, and partial exon duplication) located on autosomes were assigned to coding variant categories based on SV and gene properties. To focus on categories for which we had sufficiently dense SV counts, we restricted all analyses to categories meeting a minimum threshold of >10 SVs across patients and controls; this yielded 718 and 682 "testable" categories in NBL and EWS, respectively. Case-control burden tests for each category were performed using the logistic GLM as described above.

Many categories included overlapping sets of SVs and were thus highly correlated. To estimate the effective number of independent tests performed in each histology, sample IDs were randomly permuted relative to SV counts 1000 times for each category, and burden testing was repeated for each permutation to generate simulated Z scores. A Pearson correlation matrix between categories based on permuted Z scores was generated, and this correlation matrix was decomposed into eigenvalues and eigenvectors. We estimated the effective number of independent tests conducted for each disease to be the number of eigenvalues that captured 99% of variance in the intercategory correlation matrix. This estimated number of effective tests was used for the Bonferroni correction in each disease, resulting in a multiple hypothesis testing threshold of ~1.5×10^{-4} for NBL and EWS.

To visualize the relationships between categories, we depicted them as a graph with nodes representing categories and edges representing filters (e.g., Figs. 4A and 6F). Categories were connected if the application of a filter to one category resulted in the other (e.g., a "deletion" filter to "all singletons" yields "all singleton deletions"). To depict the high degree of similarity between some categories, the y positions of the categories were then adjusted using a force-directed approach in which the force was proportional to their Jaccard similarity via SV overlap, such that similar categories were pulled together. To reduce the number of categories depicted, categories with Jaccard similarity >0.40 were collapsed, keeping the one with the lower P value and inheriting the other's edges.

Gene-disruptive events in germline CPGs and COSMIC cancer genes

We undertook manual review of a focused subset of ultrarare (AF <0.1%) gene-disruptive SVs in select germline CPGs and COSMIC cancer genes (148 established CPGs and seven additional somatic drivers in NBL, EWS, and OS; total n = 155 genes; table S3). To maximize sensitivity, we began by identifying SVs intersecting the exons of the gene set above across all 1992 patients in our intermediate SV callset before post hoc SV callset refinement. Candidate SVs were visualized and manually reviewed using a combination of the Integrative Genomics Viewer (*98*) and GATK-SV's native RD visualization functionality used in the same manner as for the large CNV review during SV callset quality control (see above) (*17*, *87*). SVs with a visually unambiguous accumulation of evidence upon manual review were deemed to be true positives for the purposes of this clinically focused analysis. Matched tumor copy-number analysis was carried out using GATK v4.0.5.1 for the patient identified as having a de novo germline *MYCN* duplication (*102*). Tumor-in-normal contamination was assessed using deTiN v1.8.5 (*103*). Separately, we quantified overall rates of rare and singleton gene-disruptive SVs in our final high-quality SV dataset (i.e., not only restricted to manually reviewed variants described above) that affected a set of CPGs (n = 148) or COSMIC Cancer Census Tier 1 genes (n = 566; table S3) in NBL, EWS, and osteosarcoma patients compared with controls.

Gene-based rare SV association testing

We performed case-control rare SV association tests for 15,544 autosomal protein-coding genes annotated in MANE Select v1.2 whose gene bodies had <10% overlap with segmental duplications, simple repeats, reference patch loci, unalignable reference loci, and loci with alternative haplotype contigs in the hg38 assembly. For each disease, we separately evaluated rare gene-disruptive SVs (defined above), rare LoF SVs, and rare CG SVs using the logistic GLM approach as in all other analyses. Genes were

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

considered significant if they reached a Bonferroni-adjusted threshold accounting for the number of genes tested ($P < 3.2 \times 10^{-6}$).

Approach to noncoding CWAS

We extended our coding CWAS framework described above to focus exclusively on SVs with no predicted coding consequence with an additional functional layer capturing epigenetic and other noncoding genomic annotations (tables S9 to S11). We enumerated 26,400 possible noncoding SV categories per disease based on eight combinatorial layers of filters: SV AF, genic relationship (e.g., intronic, promoter, UTR, intergenic), SV type, functional element overlap, nature of intersection with the functional element, mutational constraint of the nearest gene, gene set membership of the nearest gene, and tissue-of-origin expression of the nearest gene. Deletion, duplication, inversion, complex, and insertion SVs with no known coding consequences and locations on autosomes were assigned to noncoding SV categories accordingly. Consistent with our previous coding CWAS, we again restricted to categories meeting a minimum threshold of >10 SVs across patients and controls; this yielded 7901 and 6233 "testable" categories in NBL and EWS, respectively. We once again evaluated the burden of autosomal SVs in patients relative to controls and identified noncoding categories meeting Bonferroni significance thresholds as described above, with a multiple hypothesis testing threshold of ~5.4×10^{-5} for NBL and EWS. Finally, saddlepoint approximation was used to recalibrate the null distribution used for inference, from which we subsequently recalibrated P values and effect size point estimates (*104*, *105*).

CWAS gene set enrichment

We used an approach similar to GO-term analysis to identify associations between CWAS categories and gene sets, modified to use SV counts. We first downloaded biological process gene sets from GO (*106*). Recognizing that we would have little power to detect enrichment for small gene sets and that large gene sets are usually nonspecific, we included gene sets containing between 30 and 1000 genes (n = 2,725). The association between category and gene set based on SV counts was calculated using a Fisher's exact test with the following contingency table: (count of gene-set genes affected by category SVs/count of gene-set genes affected by non-category SVs)/(count of non-gene-set genes affected by category SVs/count of non-gene-set genes affected by non-category SVs).

A continuity correction of one was added to all cells, and the "background" of noncategory SVs was defined as all rare SVs that affected genes. For each analysis, this background was further filtered such that (i) the background SV class matched the category of interest (coding versus noncoding), (ii) SVs affecting more than 10 genes were removed, (iii) SVs that did not have a coding or noncoding consequence detailed in the CWAS were removed, and (iv) noncoding SVs farther than 500 kb from their nearest gene for the noncoding consequence "INTERGENIC" were removed.

This Fisher's exact test was calculated for all gene sets, all Bonferroni-significant categories, and for patients and controls separately. We note that some categories—particularly those with functional filters like "affected gene expressed in adrenal gland"—can have substantial enrichment for certain gene sets even under a null of random SV locations because these categories are filtered to specific subsets of the genome. We therefore focused most of our discussion on "high-level categories" without functional filters, like "all singleton deletions." We also note that this contingency table ignores genomic covariates that influence which genes are affected by SVs (like gene size), so we relied on the control cohort, which will share the same technical confounders, for comparison to patients.

Expression dysregulation association with germline SVs

We used GTEx v8 to evaluate the normal tissue expression of genes affected by rare germline SVs in our cohort (*46*). We used the matrix of median TPMs by tissues to define expression of each gene in each tissue, focusing on adrenal gland and skeletal muscle for NBL and EWS.

To evaluate the effect of rare germline SVs on tumor expression, we leveraged RNA sequencing from NBL tumors available for a subset of the GMKF cohort (n = 89 of 688) (dbGaP study "Discovering the Genetic Basis of Human Neuroblastoma: A Kids First Project" phs001436.v1.p1). We assessed the effect on expression of a group of SVs by category. Similar to traditional tumor differential expression analysis, we first identified genes with sufficient expression to detect an effect. We required that genes have at least 10 counts in at least five samples to be included. We also required fewer than 20% of samples to have 0 expression for that gene.

We then converted sample TPMs to z-scores (for visualization in Fig. 4E), and then to sample ranks (0 to 88, normalized 0 to 1). Each sample tied in expression (usually zero TPMs) was assigned the average rank of those samples, although this occurrence was rare with the above filtering. If a group of SVs does not affect expression of their associated genes, sample-normalized expression ranks will be uniformly distributed between 0 and 1. To account for positive or negative effects on expression (e.g., copy gain or deletions), we instead considered the absolute rank deviation from 0.5 (i.e., |0.5 - normalized rank|), in which case the null is uniform(0, 0.5), which has mean 0.25 and variance 1/48. We defined our test statistic as the mean absolute rank deviation over the samples with SVs in a group. Because the mean of n samples from any distribution is normal by the central limit theorem, we could directly calculate a P value for an observed mean absolute rank deviation r with n samples with SVs using the following formula:

$$\text{two-sided } P \text{ value} = 2\mathrm{P}[\mathrm{X} < r | \mathrm{X} \sim \mathrm{norm}(0.25, 1/48n)]$$

We assessed the Bonferroni-significant categories for NBL as well as larger groups like "all coding SVs." To compare the effect of singleton SVs affecting genes expressed in adrenal tissue (n = 25 affected genes) to any singleton coding SV (n = 233 affected genes), we bootstrapped a P value by sampling 25 values from the "any coding SV" group 10 million times and calculated how often the observed rank mean was less than that of the singleton SVs affecting adrenal genes group.

Heritability estimation

We estimated heritability attributable to rare SVs using L_2-penalized regression models fit separately for Ewing sarcoma and neuroblastoma. For each disease, we first constructed a feature matrix encompassing all patients and their matched controls. We only considered individuals of European genetic ancestry in this analysis to protect against potential confounding by population stratification. We populated this feature matrix with sample sex, a binary indicator for whether the sample carried a large (>1 Mb), rare, unbalanced SV, and an interaction term to capture the apparent sex-dependent effects of these large, rare, unbalanced SVs. We further included all coding and noncoding SV categories considered in each disease's respective CWAS analyses (described above); however, given the inherent correlation between hierarchically related CWAS categories, we first performed unsupervised feature decorrelation on the union of all coding and noncoding CWAS categories by (i) computing the Pearson correlation coefficient, r, between SV count vectors for all pairs of categories, (ii) identifying all pairs of categories with $r > 0.8$, and (iii) iteratively pruning categories until no pairs of categories remained with $r > 0.8$. For step (iii), we pruned categories in descending order based on the total number of other categories with $r > 0.8$. We added these decorrelated CWAS categories to form our final feature matrix for each disease before fitting an L_2-penalized regression to predict binarized case|control status as a linear combination of all features. We fit these regression models with fivefold cross-validation to protect against overfitting and retained the fit with the greatest variance explained (R^2) in the held-out test fold as our

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

final model for each disease. We considered the R^2 obtained in the held-out test fold to be a lower bound of our overall heritability estimate owing to sparse SV count data and small sample size of each test fold. To obtain an upper bound for our estimate, we computed the R^2 for the best-fit cross-validated model on the full dataset (including all training and testing folds), acknowledging that this approach does not maintain independence between testing and training and may be overfit despite the L_2 penalty applied to feature coefficients. Finally, we transformed these in-sample R^2 values to liability-scale heritability estimates using the established transformation proposed by Lee and colleagues (*72*). The Lee transformation requires a prespecified population prevalence parameter, which we estimated using data obtained from the National Cancer Institute (NCI) PDQ for neuroblastoma and Ewing sarcoma, respectively (*73*, *74*).

REFERENCES AND NOTES

1. S. M. Pfister *et al*., A summary of the inaugural WHO classification of pediatric tumors: Transitioning from the optical into the molecular era. *Cancer Discov.* **12**, 331–355 (2022). doi: 10.1158/2159-8290.CD-21-1094; pmid: 34921008
2. N. J. Balamuth, R. B. Womer, Ewing's sarcoma. *Lancet Oncol.* **11**, 184–192 (2010). doi: 10.1016/S1470-2045(09)70286-4; pmid: 20152770
3. K. K. Matthay *et al*., Neuroblastoma. *Nat. Rev. Dis. Primers* **2**, 16078 (2016). doi: 10.1038/nrdp.2016.78; pmid: 27830764
4. N. D. Anderson *et al*., Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science* **361**, eaam8419 (2018). doi: 10.1126/science.aam8419; pmid: 30166462
5. L. Mirabello *et al*., Frequency of pathogenic germline variants in cancer-susceptibility genes in patients with osteosarcoma. *JAMA Oncol.* **6**, 724–734 (2020). doi: 10.1001/jamaoncol.2020.0197; pmid: 32191290
6. T. J. Pugh *et al*., The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45**, 279–284 (2013). doi: 10.1038/ng.2529; pmid: 23334666
7. S. Behjati *et al*., Recurrent mutation of IGF signalling genes and distinct patterns of genomic rearrangement in osteosarcoma. *Nat. Commun.* **8**, 15936 (2017). doi: 10.1038/ncomms15936; pmid: 28643781
8. J. E. Garber, K. Offit, Hereditary cancer predisposition syndromes. *J. Clin. Oncol.* **23**, 276–292 (2005). doi: 10.1200/JCO.2005.10.042; pmid: 15637391
9. L. G. Spector, N. Pankratz, E. L. Marcotte, Genetic and nongenetic risk factors for childhood cancer. *Pediatr. Clin. North Am.* **62**, 11–25 (2015). doi: 10.1016/j.pcl.2014.09.013; pmid: 25435109
10. Q. Feng *et al*., Increased burden of familial-associated early-onset cancer risk among minority Americans compared to non-Latino Whites. *eLife* **10**, e64793 (2021). doi: 10.7554/eLife.64793; pmid: 34155975
11. J. Zhang, K. E. Nichols, J. R. Downing, Germline mutations in predisposition genes in pediatric cancer. *N. Engl. J. Med.* **374**, 1391 (2016). pmid: 27050224
12. S. N. Gröbner *et al*., The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327 (2018). doi: 10.1038/nature25480; pmid: 29489754
13. T. A. Manolio *et al*., Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009). doi: 10.1038/nature08494; pmid: 19812666
14. N. Rahman, Realizing the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014). doi: 10.1038/nature12981; pmid: 24429628
15. K. L. Huang *et al*., Pathogenic germline variants in 10,389 adult cancers. *Cell* **173**, 355–370.e14 (2018). doi: 10.1016/j.cell.2018.03.039; pmid: 29625052
16. Z. Sondka *et al*., The COSMIC Cancer Gene Census: Describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018). doi: 10.1038/s41568-018-0060-1; pmid: 30293088
17. R. L. Collins *et al*., A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020). doi: 10.1038/s41586-020-2287-8; pmid: 32461652
18. P. H. Sudmant *et al*., An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015). doi: 10.1038/nature15394; pmid: 26432246
19. L. A. Garraway, E. S. Lander, Lessons from the cancer genome. *Cell* **153**, 17–37 (2013). doi: 10.1016/j.cell.2013.03.002; pmid: 23540688
20. R. Beroukhim *et al*., The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010). doi: 10.1038/nature08822; pmid: 20164920
21. G. H. Perry *et al*., Copy number variation and evolution in humans and chimpanzees. *Genome Res.* **18**, 1698–1710 (2008). doi: 10.1101/gr.082016.108; pmid: 18775914
22. J. Weischenfeldt, O. Symmons, F. Spitz, J. O. Korbel, Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013). doi: 10.1038/nrg3373; pmid: 23329113
23. L. E. Egolf *et al*., Germline 16p11.2 Microdeletion Predisposes to Neuroblastoma. *Am. J. Hum. Genet.* **105**, 658–668 (2019). doi: 10.1016/j.ajhg.2019.07.020; pmid: 31474320
24. J. Kim *et al*., Pathogenic germline variants in cancer susceptibility genes in children and young adults with rhabdomyosarcoma. *JCO Precis. Oncol.* **5**, 75–87 (2021). doi: 10.1200/PO.20.00218; pmid: 34095712
25. C. McLeod *et al*., St. Jude Cloud: A pediatric cancer genomic data-sharing ecosystem. *Cancer Discov.* **11**, 1082–1099 (2021). doi: 10.1158/2159-8290.CD-20-1230; pmid: 33408242
26. J. R. Downing *et al*., The Pediatric Cancer Genome Project. *Nat. Genet.* **44**, 619–622 (2012). doi: 10.1038/ng.2287; pmid: 22641210
27. L. L. Robison *et al*., Study design and cohort characteristics of the Childhood Cancer Survivor Study: A multi-institutional collaborative project. *Med. Pediatr. Oncol.* **38**, 229–239 (2002). doi: 10.1002/mpo.1316; pmid: 11920786
28. A. Hudson, M. Fournier, J. Coulombe, D. Daee, Using existing pediatric cancer data from the Gabriella Miller Kids First Data Resource Program. *JNCI Cancer Spectr.* **7**, pkad079 (2023). doi: 10.1093/jncics/pkad079; pmid: 37788089
29. J. Zhang *et al*., The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019). doi: 10.1038/s41587-019-0055-9; pmid: 30877282
30. D. Taliun *et al*., Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021). doi: 10.1038/s41586-021-03205-y; pmid: 33568819
31. M. Byrska-Bishop *et al*., High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022). doi: 10.1016/j.cell.2022.08.004; pmid: 36055201
32. P. Ebert *et al*., Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021). doi: 10.1126/science.abf7117; pmid: 33632895
33. J. Sebat *et al*., Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007). doi: 10.1126/science.1138659; pmid: 17363630
34. S. J. Sanders *et al*., Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011). doi: 10.1016/j.neuron.2011.05.002; pmid: 21658581
35. D. Warburton, Outcome of cases of de novo structural rearrangements diagnosed at amniocentesis. *Prenat. Diagn.* **4**, 69–80 (1984). doi: 10.1002/pd.1970040706; pmid: 6463034
36. C. Gilissen *et al*., Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014). doi: 10.1038/nature13394; pmid: 24896178
37. C. Lowther *et al*., Systematic evaluation of genome sequencing for the diagnostic assessment of autism spectrum disorder and fetal structural anomalies. *Am. J. Hum. Genet.* **110**, 1454–1469 (2023). doi: 10.1016/j.ajhg.2023.07.010; pmid: 37595579
38. D. M. Werling *et al*., An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018). doi: 10.1038/s41588-018-0107-y; pmid: 29700473
39. J. R. Belyeu *et al*., De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am. J. Hum. Genet.* **108**, 597–607 (2021). doi: 10.1016/j.ajhg.2021.02.012; pmid: 33675682
40. G. M. Cooper *et al*., A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011). doi: 10.1038/ng.909; pmid: 21841781
41. C. R. Marshall *et al*., Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008). doi: 10.1016/j.ajhg.2007.12.009; pmid: 18252227
42. C. Redin *et al*., The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* **49**, 36–45 (2017). doi: 10.1038/ng.3720; pmid: 27841880
43. R. L. Collins *et al*., A cross-disorder dosage sensitivity map of the human genome. *Cell* **185**, 3041–3055.e25 (2022). doi: 10.1016/j.cell.2022.06.036; pmid: 35917817
44. S. Girirajan *et al*., Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *N. Engl. J. Med.* **367**, 1321–1331 (2012). doi: 10.1056/NEJMoa1200395; pmid: 22970919
45. J. Y. An *et al*., Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018). doi: 10.1126/science.aat6576; pmid: 30545852
46. F. Aguet *et al*., The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020). doi: 10.1126/science.aaz1776; pmid: 32913098
47. J. A. Kosmicki *et al*., Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017). doi: 10.1038/ng.3789; pmid: 28191890
48. K. J. Karczewski *et al*., The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi: 10.1038/s41586-020-2308-7; pmid: 32461654
49. J. M. Maris, Recent advances in neuroblastoma. *N. Engl. J. Med.* **362**, 2202–2211 (2010). doi: 10.1056/NEJMra0804577; pmid: 20558371
50. T. G. P. Grünewald *et al*., Ewing sarcoma. *Nat. Rev. Dis. Primers* **4**, 5 (2018). doi: 10.1038/s41572-018-0003-x; pmid: 29977059
51. L. Farnaes *et al*., Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ Genom. Med.* **3**, 10 (2018). doi: 10.1038/s41525-018-0049-4; pmid: 29644095
52. K. B. Brothers, J. L. Vassy, R. C. Green, Reconciling Opportunistic and Population Screening in Clinical Genomics. *Mayo Clin. Proc.* **94**, 103–109 (2019). doi: 10.1016/j.mayocp.2018.08.028; pmid: 30611438
53. R. Gillani *et al*., Germline predisposition to pediatric Ewing sarcoma is characterized by inherited pathogenic variants in DNA damage repair genes. *Am. J. Hum. Genet.* **109**, 1026–1037 (2022). doi: 10.1016/j.ajhg.2022.04.007; pmid: 35512711
54. M. P. Randall *et al*., *BARD1* germline variants induce haploinsufficiency and DNA repair defects in neuroblastoma. *J. Natl. Cancer Inst.* **116**, 138–148 (2024). doi: 10.1093/jnci/djad182; pmid: 37688570
55. J. Kim *et al*., Germline pathogenic variants in neuroblastoma patients are enriched in *BARD1* and predict worse survival. *J. Natl. Cancer Inst.* **116**, 149–159 (2024). doi: 10.1093/jnci/djad183; pmid: 37688579
56. D. Trochet *et al*., Germline mutations of the paired-like homeobox 2B (*PHOX2B*) gene in neuroblastoma. *Am. J. Hum. Genet.* **74**, 761–764 (2004). doi: 10.1086/383253; pmid: 15024693
57. J. Kamihara *et al*., Retinoblastoma and neuroblastoma predisposition and surveillance. *Clin. Cancer Res.* **23**, e98–e106 (2017). doi: 10.1158/1078-0432.CCR-17-0652; pmid: 28674118
58. J. J. Molenaar *et al*., Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483**, 589–593 (2012). doi: 10.1038/nature10910; pmid: 22367537
59. R. C. Seeger *et al*., Association of multiple copies of the N-*myc* oncogene with rapid progression of neuroblastomas. *N. Engl. J. Med.* **313**, 1111–1116 (1985). doi: 10.1056/NEJM198510313131802; pmid: 4047115
60. G. J. Prud'homme, M. Kurt, Q. Wang, Pathobiology of the Klotho antiaging protein and therapeutic considerations. *Front. Aging* **3**, 931331 (2022). doi: 10.3389/fragi.2022.931331; pmid: 35903083
61. H. Ligumsky, K. Merenbakh-Lamin, N. Keren-Khadmy, I. Wolf, T. Rubinek, The role of α-klotho in human cancer: Molecular

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

and clinical aspects. *Oncogene* **41**, 4487–4497 (2022). doi: 10.1038/s41388-022-02440-5; pmid: 36038662
62. J. E. Moore *et al*., Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). doi: 10.1038/s41586-020-2493-4; pmid: 32728249
63. C. Lowther *et al*., Balanced chromosomal rearrangements offer insights into coding and noncoding genomic features associated with developmental disorders. medRxiv 2022.2002.2015.22270795 [Preprint] (2022); https://doi.org/10.1101/2022.02.15.22270795.
64. K. Mohajeri *et al*., Transcriptional and functional consequences of alterations to *MEF2C* and its topological organization in neuronal models. *Am. J. Hum. Genet.* **109**, 2049–2067 (2022). doi: 10.1016/j.ajhg.2022.09.015; pmid: 36283406
65. D. G. Lupiáñez, M. Spielmann, S. Mundlos, Breaking TADs: How alterations of chromatin domains result in disease. *Trends Genet.* **32**, 225–237 (2016). doi: 10.1016/j.tig.2016.01.003; pmid: 26862051
66. M. Spielmann, D. G. Lupiáñez, S. Mundlos, Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018). doi: 10.1038/s41576-018-0007-0; pmid: 29692413
67. S. Jiao *et al*., Estimating the heritability of colorectal cancer. *Hum. Mol. Genet.* **23**, 3898–3905 (2014). doi: 10.1093/hmg/ddu087; pmid: 24562164
68. A. Gusev *et al*., Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. *Nat. Commun.* **7**, 10979 (2016). doi: 10.1038/ncomms10979; pmid: 27052111
69. J. Dai *et al*., Estimation of heritability for nine common cancers using data from genome-wide association studies in Chinese population. *Int. J. Cancer* **140**, 329–336 (2017). doi: 10.1002/ijc.30447; pmid: 27668986
70. X. Jiang *et al*., Shared heritability and functional enrichment across six solid cancers. *Nat. Commun.* **10**, 431 (2019). doi: 10.1038/s41467-018-08054-4; pmid: 30683880
71. L. A. Mucci *et al*., Familial risk and heritability of cancer among twins in Nordic countries. *JAMA* **315**, 68–76 (2016). doi: 10.1001/jama.2015.17703; pmid: 26746459
72. S. H. Lee, N. R. Wray, M. E. Goddard, P. M. Visscher, Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011). doi: 10.1016/j.ajhg.2011.02.002; pmid: 21376301
73. National Cancer Institute, Neuroblastoma Treatment (PDQ)–Health Professional Version; https://www.cancer.gov/types/neuroblastoma/hp/neuroblastoma-treatment-pdq.
74. National Cancer Institute, Ewing Sarcoma and Undifferentiated Small Round Cell Sarcomas of Bone and Soft Tissue Treatment (PDQ)–Health Professional Version; https://www.cancer.gov/types/bone/hp/ewing-treatment-pdq.
75. L. A. Williams, M. Richardson, E. L. Marcotte, J. N. Poynter, L. G. Spector, Sex ratio among childhood cancers by single year of age. *Pediatr. Blood Cancer* **66**, e27620 (2019). doi: 10.1002/pbc.27620; pmid: 30815990
76. L. A. Williams *et al*., Sex differences in associations between birth characteristics and childhood cancers: A five-state registry-linkage study. *Cancer Causes Control* **32**, 1289–1298 (2021). doi: 10.1007/s10552-021-01479-1; pmid: 34297242
77. X. Chen *et al*., Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016). doi: 10.1093/bioinformatics/btv710; pmid: 26647377
78. E. J. Gardner *et al*., The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017). doi: 10.1101/gr.218032.116; pmid: 28855259
79. Z. N. Kronenberg *et al*., Wham: Identifying structural variants of biological consequence. *PLOS Comput. Biol.* **11**, e1004572 (2015). doi: 10.1371/journal.pcbi.1004572; pmid: 26625158
80. M. Babadi *et al*., GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nat. Genet.* **55**, 1589–1597 (2023). doi: 10.1038/s41588-023-01449-0; pmid: 37604963
81. G. Klambauer *et al*., cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40**, e69 (2012). doi: 10.1093/nar/gks003; pmid: 22302147
82. R. L. Collins *et al*., Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* **18**, 36 (2017). doi: 10.1186/s13059-017-1158-6; pmid: 28260531
83. The All of Us Research Program Genomics Investigators, Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024). doi: 10.1038/s41586-023-06957-x; pmid: 38374255
84. A. D. Ewing *et al*., Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.* **14**, R22 (2013). doi: 10.1186/gb-2013-14-3-r22; pmid: 23497673
85. J. Morales *et al*., A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022). doi: 10.1038/s41586-022-04558-8; pmid: 35388217
86. A. V. Khera *et al*., Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation* **139**, 1593–1602 (2019). doi: 10.1161/CIRCULATIONAHA.118.035658; pmid: 30586733
87. R. L. Collins, M. R. Stone, H. Brand, J. T. Glessner, M. E. Talkowski, CNView: a visualization and annotation tool for copy number variation from whole-genome sequencing. bioRxiv 049536 [Preprint] (2016); https://doi.org/10.1101/049536.
88. W. McLaren *et al*., The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016). doi: 10.1186/s13059-016-0974-4; pmid: 27268795
89. A. Kundaje *et al*., Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). doi: 10.1038/nature14248; pmid: 25693563
90. J. Nasser *et al*., Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021). doi: 10.1038/s41586-021-03446-x; pmid: 33828297
91. S. Dimitrieva, P. Bucher, UCNEbase—A database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.* **41**, D101–D109 (2013). doi: 10.1093/nar/gks1092; pmid: 23193254
92. A. Kong *et al*., Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010). doi: 10.1038/nature09525; pmid: 20981099
93. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020). doi: 10.1038/s41586-020-1969-6; pmid: 32025007
94. Y. Li *et al*., Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020). doi: 10.1038/s41586-019-1913-9; pmid: 32025012
95. M. P. Conomos, A. P. Reiner, B. S. Weir, T. A. Thornton, Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016). doi: 10.1016/j.ajhg.2015.11.022; pmid: 26748516
96. J. Graffelman, J. M. Camarena, Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Hum. Hered.* **65**, 77–84 (2008). doi: 10.1159/000108939; pmid: 17898538
97. X. Zhao *et al*., Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.* **108**, 919–928 (2021). doi: 10.1016/j.ajhg.2021.03.014; pmid: 33789087
98. J. T. Robinson *et al*., Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011). doi: 10.1038/nbt.1754; pmid: 21221095
99. J. Mbatchou *et al*., Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021). doi: 10.1038/s41588-021-00870-7; pmid: 34017140
100. S. Purcell *et al*., PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). doi: 10.1086/519795; pmid: 17701901
101. M. Gillespie *et al*., The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022). doi: 10.1093/nar/gkab1028; pmid: 34788843
102. M. A. DePristo *et al*., A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011). doi: 10.1038/ng.806; pmid: 21478889
103. A. Taylor-Weiner *et al*., DeTiN: Overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018). doi: 10.1038/s41592-018-0036-9; pmid: 29941871
104. R. Dey, E. M. Schmidt, G. R. Abecasis, S. Lee, A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017). doi: 10.1016/j.ajhg.2017.05.014; pmid: 28602423
105. W. Zhou *et al*., Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018). doi: 10.1038/s41588-018-0184-y; pmid: 30104761
106. P. D. Thomas *et al*., PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022). doi: 10.1002/pro.4218; pmid: 34717010
107. R. GIllani *et al*., vanallenlab/ped_germline_SV: Manuscript revision 1. Zenodo (2024); https://10.5281/zenodo.13773757.

ACKNOWLEDGMENTS

We acknowledge the participants and studies who provided biological samples to enable this research. We thank E. Pimenta for her thoughtful and critical review of this manuscript. We thank J. Schiffman, S. O'Brien, E. L. Young, L. Hayes, G. Mitchell, T. Fowler, J. Anson, and the team at the Huntsman Cancer Institute for their foundational work in establishing the GMKF Ewing sarcoma dataset. We thank the Children's Oncology Group, QuadW Foundation, M. D. Krailo, and D. A. Barkauskas for the provision of banked normal tissue from an Ewing sarcoma patient to enable PCR evaluation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health (NIH). **Funding:** This work was funded by a Alex's Lemonade Stand Foundation young investigator grant (R.G.); an American Society of Clinical Oncology career development award (R.G.); an American Society of Clinical Oncology–Conquer Cancer Sarcoma Foundation of America young investigator award (R.G.); a Boston Children's Hospital translational research program mentored translational investigator service award (R.G.); a Dana-Farber Cancer Institute Wong Family Award in Translational Oncology (R.G.); a Dana-Farber Cancer Institute–Boston Children's Hospital Pedals for Pediatrics grant (R.G., K.A.J.); Department of Defense idea development award KC210042/W81XWH-22-1-0455 (S.H.A.); Department of Defense peer-reviewed cancer research program fellow career development award CA220721 (R.G.); Department of Defense physician research award W81XWH-21-1-0084, PC200150 (S.H.A.); an Innovation in Cancer Informatics award (E.M.V.A.); NIH grants 1X01HL140547 (Children's Oncology Group), K08CA276701 (R.G.), K99CA286805 (R.L.C.), R01MH115957 (M.E.T.), R01CA227388 (E.M.V.A.), R37CA222574 (E.M.V.A.), R37CA222574_S1 (EMVA), T32GM007753 (J.K.J.), T32GM144273 (J.K.J.), U01CA233100 (E.M.V.A.), U10CA180886 (Children's Oncology Group), U10CA180899 (Children's Oncology Group), and U24CA196173 (Children's Oncology Group); a Rally Foundation career development Award (R.G.); Rossy Foundation Fund at KBF Canada (R.L.C.); and a Sarcoma Foundation of America Jay Vernon Jackson memorial research award (R.G.). **Author contributions:** Conceptualization: R.G., R.L.C., E.M.V.A.; Formal analysis: R.L.C., R.G., J.C., J.K.J.; Funding acquisition: R.G., K.A.J., E.M.V.A., K.H.; Investigation: R.L.C., R.G., J.C.; Methodology: R.L.C., M.W., A.-S.J., C.W.W., E.P.-H., M.E.T., H.B.; Resources: K.A.J., B.D.C., J.L.; Supervision: E.M.V.A.; Validation: A.Ga.; Visualization: R.L.C., J.C., R.G.; Writing – original draft: R.G., R.L.C., J.C.; Writing – review and editing: E.M.V.A., H.B., A.Gu., S.H.A., K.A.J., B.D.C. **Competing interests:** R.G. has equity in Google, Microsoft, Amazon, Apple, Moderna, Pfizer, and Vertex Pharmaceuticals; his spouse is employed by Apree Health. R.L.C. has equity in Microsoft, Natera Inc., Vertex, Beam Therapeutics, Verve Therapeutics, Grail Inc., Illumina Inc., and Pacific Biosciences. C.W.W., M.E.T., and H.B. have filed an institutional patent on methods for noninvasive prenatal sequencing. M.E.T. receives research funding and/or reagents from Illumina Inc., Pacific Biosciences, Microsoft Inc., Ionis Therapeutics, and Levo Therapeutics and is a scientific advisory board member of BridgeBio. K.H. holds a consulting role with Aclaris Pharmaceuticals. J.L. holds a consulting role with 23andMe and the Troper Wojcicki Foundation. B.D.C. has equity in Acceleron Pharma; holds consulting roles with PetDx, Animal Cancer Foundation, and the Osteosarcoma Institute; and has received research funding from Gradalis; his spouse is employed by and receives equity from Generate Biomedicines. K.A.J. has received honoraria from Foundation Medicine and Takeda, holds consulting roles with Bayer and Ipsen, and has received travel reimbursement from Bayer. E.M.V.A. holds consulting roles with Tango Therapeutics, Genome Medical, Genomic Life, Enara Bio, Janssen, Manifold Bio, Serinus Bio, and TracerDx; receives research support from Bristol-Myers Squibb and Novartis; has equity in Tango Therapeutics, Genome Medical, Genomic Life, Syapse, Enara Bio, Manifold Bio, Microsoft, Riva Therapeutics, Serinus Bio, and TracerDx; has received travel reimbursement from Roche and Genentech; and has filed institutional patents on

Downloaded from https://www.science.org at Broad Institute on January 03, 2025

chromatin mutations, immunotherapy response, and methods for clinical interpretation. The other authors declare that they have no competing interests. **Data and materials availability:** All raw genome sequencing data analyzed in this study are hosted by dbGaP (https://dbgap.ncbi.nlm.nih.gov/; cohorts: MESA, BioMe, GMKF), St. Jude Cloud (https://www.stjude.cloud/), the International Cancer Genome Consortium (https://dcc.icgc.org/; https://ega-archive.org/), or the International Genome Sample Resource (1000 Genomes Project; https://www.internationalgenome.org). Accession numbers for each cohort are provided in the materials and methods and table S1. The GATK-SV pipeline is publicly available from GitHub (https://github.com/broadinstitute/gatk-sv). Given that this study combines germline data from multiple independent controlled-access data repositories and consortia, it is not feasible to deposit the SV callset in preexisting or newly created centralized repositories. The corresponding authors will facilitate access to the germline SV calls for qualified follow-up research upon request to those who have been approved to access the corresponding controlled-access studies in the materials and methods section. All other software and code used to generate the results in this study are publicly available from GitHub (https://github.com/vanallenlab/ped_germline_SV), and the version corresponding to this publication has been archived on Zenodo (*107*). **License information:** Copyright © 2025 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adq0071
Supplementary Text
Figs. S1 to S18
Tables S1 to S16
References (*108*, *109*)
MDAR Reproducibility Checklist

Submitted 24 April 2024; accepted 25 October 2024
10.1126/science.adq0071

Downloaded from https://www.science.org at Broad Institute on January 03, 2025