nature communications

Article

https://doi.org/10.1038/s41467-025-60110-y

Global DNA methylation differences involving germline structural variation impact gene expression in pediatric brain tumors

Received: 18 November 2024

Accepted: 13 May 2025

Published online: 21 May 2025

Check for updates

Fengju Chen[1,7], Yiqun Zhang[1,7], Wei Li[2], Fritz J. Sedlazeck[3,4], Lanlan Shen[5] & Chad J. Creighton[1,3,6]

The extent of genetic variation and its influence on gene expression across multiple tissue and cellular contexts is still being characterized, with germline Structural Variants (SVs) being historically understudied. DNA methylation also represents a component of normal germline variation across individuals. Here, we combine germline SVs (by short-read sequencing) with tumor DNA methylation across 1292 pediatric brain tumor patients. For thousands of methylation probes for CpG Islands (CGIs) or enhancers, rare and common SV breakpoints upstream or downstream associate with differential methylation in tumors spanning various histologic types, a significant subset involving genes with SV-associated differential expression. Cancer predisposition genes involving SV-associated differential methylation and expression include *MSH2*, *RSPA*, and *PALB2*. SV breakpoints falling within CGIs or histone marks H3K36me3 or H3K9me3 associate with differential CGI methylation. Genes with SVs and CGI methylation associated with patient survival include *POLD4*. Our results capture a class of normal phenotypic variation having disease implications.

The extent of genetic variation and its influence on gene expression is still being characterized, as the regulation of specific genes is highly dependent on tissue and cellular context[1,2]. Expression quantitative trait loci (eQTL) analyses aim to associate genotypes with gene expression levels across a sample of individuals, with such studies focusing almost exclusively on germline single nucleotide polymorphisms (SNPs)[3,4]. Gene expression associations with genetic variants are often cell-type-specific and context-dependent[1], requiring multiple studies to systematically catalog these associations across different biological systems, including human disease. In contrast to SNPs, germline Structural Variants (SVs) have historically been understudied in the context of gene expression, partly due to the historical reliance upon SNP genotyping arrays or whole-exome sequencing for eQTL analyses[4]. SVs represent a broad class of variation that includes copy-number variants (CNVs) and balanced rearrangements. Compared to SNPs, germline SVs account for more nucleotide sequence differences and can disproportionately impact genes, through direct deletion or duplication or through impacting non-coding regulatory regions[4–6]. Most disease-associated genetic variants fall within the non-coding regions of the genome[7], where

[1]Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA. [2]Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, CA 92697, USA. [3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA. [4]Department of Computer Science, Rice University, Houston, TX 77005, USA. [5]USDA Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA. [6]Department of Medicine, Baylor College of Medicine, Houston, TX 77030, USA. [7]These authors contributed equally: Fengju Chen, Yiqun Zhang. e-mail: creighto@bcm.edu

major efforts by the ENCODE consortium and others have highlighted the essential role of non-coding DNA in gene regulation[8]. Germline SV eQTLs, using Whole Genome Sequencing (WGS) to call SVs, have been cataloged in at least two studies using the Genotype-Tissue Expression (GTEx) datasets[2,4], with combined WGS and RNA-sequencing (RNA-seq) data, as well as in other studies involving normal cells or tissues[9,10].

The influence of germline SVs on gene expression in human cancers remains to be better characterized. The Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium was perhaps the first major effort to combine WGS and RNA-seq on appreciable numbers of human tumors, involving some 1220 patients with both SV and expression calls[11,12]. Across multiple studies, we and others have explored the global influence of somatic SVs (as opposed to germline SVs) in gene expression in cancer, where we have found widespread associations between SV-mediated cis-regulatory alterations and altered expression after correcting for any related copy number alterations[11,13–17]. Both germline variation and somatic alterations contribute to the molecular profile of cancers. Using the PCAWG datasets, we recently surveyed global associations between germline SVs (as called using a normal blood sample) and tumor-derived gene expression across adult cancers spanning various tissues and cells of origin[18]. Most of the hundreds of germline SV-expression associations identified in the PCAWG cohort would not necessarily have specific roles in cancer but instead reflect a class of normal phenotypic variation across individuals. However, some genes involved had known roles in cancer or other associations that would indicate their potential for disease contributors. Recently, the Children's Brain Tumor Network (CBTN), another multi-institutional effort, has assembled multi-omics data—including short-read WGS, RNA-seq, and DNA methylation—on more than 1000 pediatric brain and Central Nervous System (CNS) tumors[19,20]. Pediatric tumors are molecularly distinct from adult cancers[21], where PCAWG RNA-seq data did not include any pediatric brain tumors[12,18]. Also, we can expect the catalog of germline SV-expression associations in brain and CNS tissues to differ in many respects from the catalog obtained using other tissues represented in PCAWG.

DNA methylation is a chemical as well as an epigenetic modification that is directly associated with gene expression[22]. DNA methylation at CpG sites in CpG islands (CGIs) of promoters typically causes stable silencing of genes, while de-methylation of CGIs may lead to loss of silencing and increased gene expression[23]. DNA methylation at distal enhancer regions has also been implicated in gene regulation[24–27]. DNA methylation defines the cell type and its lineage through gene expression control and genome stability[22], as reflected in pan-cancer surveys of tumors spanning various tissues or cells of origin[28]. DNA methylation also represents a component of normal germline variation across individuals, as evident from numerous SNP-methylation studies to identify methylation quantitative trait loci (mQTL) across different tissue and cellular contexts[29–31]. Most germline SV-expression associations observed in normal or cancer tissues do not have an explanation in terms of functional genomic elements being consistently impacted[2,18], where differential DNA methylation levels may be one plausible mechanism underlying some associations. In contrast to SNPs, germline SVs have been understudied in the context of DNA methylation. One recent study by Shi et al.[32] used paired CNV and CGI methylation data to define mQTLs across lymphoblastoid cell lines from 77 individuals. Recently, we explored the global influence of somatic SVs on DNA methylation across 1343 pediatric brain tumors from the CBTN, involving corresponding changes in gene expression[33]. At the time of our present study, combined WGS-based germline SVs and tumor DNA methylation from the CBTN involved 1292 pediatric brain tumor patients, which might represent the most individuals to date involving a germline SV-methylation association study.

In this work, we explore the influence of germline SVs on tumor DNA methylation across pediatric brain and CNS tumor patients. Our focus here was on DNA methylation patterns involving CGIs or enhancers, including genes for which germline SV-associated differential expression was in the opposite direction of methylation. The overall patterns observed in the CBTN cohort were also observable in an independent adult cancer cohort from The Cancer Genome Atlas (TCGA)[15,18], though involving different genes and CGIs. Most of the significant genes involving differential CGI or enhancer methylation would not necessarily have specific roles in cancer but would instead reflect a class of normal phenotypic variation. However, outside information, including previously established cancer roles for genes or patient survival data, could be leveraged to help prioritize a subset of the top genes arising from our study for future investigation.

Results

Combined germline structural variation and tumor DNA methylation across pediatric brain tumor patients

As an avenue for exploring germline structural variation in the context of DNA methylation, we referred to the CBTN datasets of germline SV calls representing 1292 pediatric brain or CNS tumor patients (Supplementary Data 1 and Fig. 1a, b). For each patient, one tumor was profiled for DNA methylation by Illumina array platform, and germline SV calls were obtained based on whole genome sequencing of the blood normal sample. For 1430 pediatric brain or CNS tumors patients, combined tumor RNA expression and germline SV data were also generated by CBTN, of which 1089 were represented in the 1292-patient dataset. Across the 1658 patients with germline WGS and either RNA-seq or DNA methylation data or both, 306,712 distinct germline SVs were detected after filtering for calls made by both of two algorithms (Supplementary Data 2). Most of these germline SVs had been observed elsewhere, with 88.8% of SV detection events involving SVs being represented in the Database of Genomic Variants (DGV) curated from published studies[34] and 69.9% of events involving SVs being represented in the gnomAD high-quality SV catalog based on WGS analysis of 63,046 individuals[35] (Fig. S1a–c). The frequency of germline SVs in the CBTN cohort broadly correlated with allele frequencies (AFs) as reported in gnomAD (Fig. S1b), where some 96% of gnomAD SVs had AF < 1%. Of the CBTN SVs reported in gnomAD, a very small fraction—about 0.2%—were detected in >10% of CBTN patients but with gnomAD AF < 1% (Supplementary Data 2); however, we make no conclusions here regarding their possible disease relevance in the absence of future case-control studies. Only about 2.7% of CBTN SV detection events had both breakpoints within the same region of segmental duplication (Fig. S1d)[36].

In our present study, we set out to systematically identify CGIs or enhancers for which the proximity of germline SVs was recurrently and significantly associated with differential DNA methylation in the tumor sample across multiple patients (Fig. 1a), focusing primarily on CGIs or enhancers for which the nearby genes also had tumor gene expression differential in association with germline SV (based on the 1430-patient cohort) but in the opposite direction from methylation. For example, the subset of genes having lower expression associated with nearby germline SV breakpoints that also had higher CGI methylation associated with SVs would appear to represent epigenetic gene silencing events mediated by SVs; on the other hand, genes with high expression and corresponding lower CGI methylation association would represent a loss of gene silencing. At the same time, higher methylation of some repressive elements can lead to higher gene expression, e.g., as described for *TERT*[11,37]. However, meaningfully understanding such positive correlations between methylation and expression requires specific domain knowledge regarding the gene and its regulatory region. Therefore, here we focused mainly on SV associations involving inverse methylation and expression relationships when highlighting specific genes, though the larger body of SV-methylation associations also warrants examination (Supplementary Data 3). The pediatric tumors profiled in the datasets represented a range of

https://doi.org/10.1038/s41467-025-60110-y

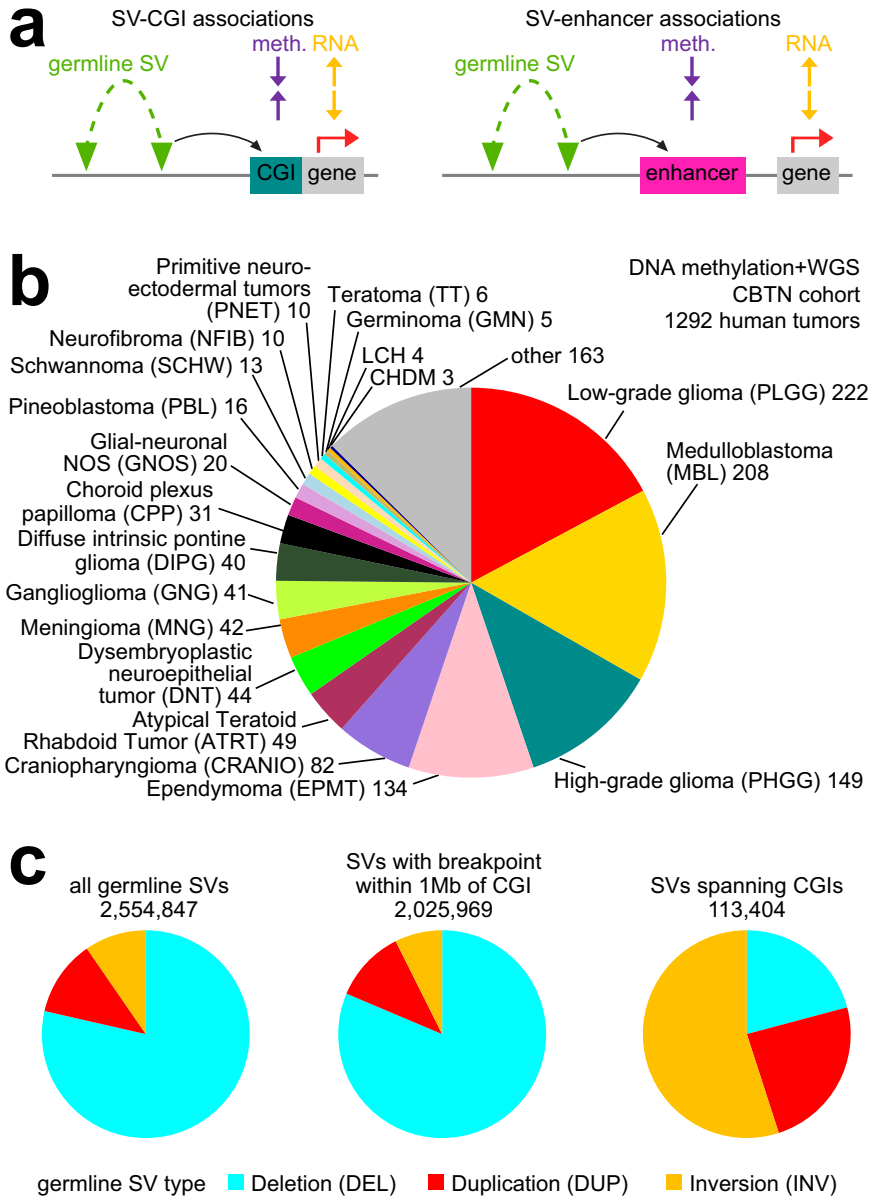


Fig. 1 | Combined whole-genome sequencing and DNA methylation pediatric brain tumor patient datasets to explore methylation differences involving germline SVs. a Schematic of the study approach. We referred to the CBTN multi-omics datasets[20,33] to explore germline structural variation in pediatric brain tumor patients in relation to tumor DNA methylation patterns. Germline SV calls were generated from the blood normal sample (using whole genome sequencing, or WGS, data). Based on the corresponding tumor sample, the CBTN generated both DNA methylation profiles and gene expression profiles. We identified both CpG Islands (CGIs, left) and enhancers (right) with DNA methylation differences associated with nearby germline SV breakpoints, paying particular attention to genes with differential mRNA expression patterns inversely correlated with the methylation patterns. **b** In the CBTN cohort, combined DNA methylation array and germline SV data involved 1292 patients, broken down here by tumor histologic type. **c** For the 1292 patient blood samples, SV class distributions, as observed for all 2,554,847 germline SVs in the dataset (left), for the subset of germline SVs with breakpoint within 1 Mb of a CGI methylation probe (middle), and for the subset of germline SVs with breakpoints spanning a CGI midpoint (right). DEL, deletion; DUP, duplication; INV, inversion (SV types as called by Manta algorithm[67]).

diverse histologic types involving widespread global differences at the DNA methylation level[33,38] (Fig. 1b), which differences our integrative analysis approaches could account for.

The 1292 patient samples involved 2,554,847 germline SV events (minimum size ten bp), with 78.5% deletions, 11.8% duplications, and 9.6% inversions (Fig. 1c). A tiny fraction of germline SVs—1835 out of 2,554,84 (0.07%)—were insertions. The 2,554,847 germline SVs did not include translocations, due to concerns of such calls in germline data largely representing artifacts[18,39]. Most (79%) of germline SVs had a breakpoint within 1 Mb of a CGI, which calls were involved in our analyses associating SVs with differential CGI methylation (see below), as germline SVs are understood to be involved in long-range gene regulation and DNA methylation variation[18,32,40]. Interestingly, of the 4.5% of germline SVs for which both breakpoints spanned a CGI, over half of these (55%) were inversions, and only 21% were deletions (Fig. 1c). Overall, inversion SVs tended to be larger and deletion SVs smaller (Fig. S1e), though the above patterns were observed when breaking down the SV call sets by size (Fig. S1f). The significant anti-enrichment patterns of deletion SVs surrounding CGIs suggest that germline deletion of CGIs resulting in their lower methylation is uncommon, consistent with our observations below, although deletion of genes can result in observed lower expression.

CGI-level germline SV-associated tumor DNA methylation differences

Across the CBTN cohort, a set of 6035 CGI array probes (out of 133,345 examined) showed germline SV-associated differential DNA methylation at False Discovery Rate (FDR) < 10% significance level (linear modeling correcting for covariates) for any one of four gene region windows examined for SV breakpoints: 100 kb upstream (1257 probes with FDR < 10%), 100 kb downstream (1313 probes), within the gene (846 probes), and 1 Mb upstream or downstream (4484 probes, Fig. 2a and S2a and Supplementary Data 3). CGI probes with SV-associated higher methylation were highly enriched for promoter-associated CGIs, while CGI probes with SV-associated lower methylation were highly enriched for gene body CGIs (Fig. 2b, $p < 1E\text{-}80$ and $p < 1E\text{-}6$, respectively, chi-squared test). All major SV classes (duplication, inversion, and deletion) and sizes were involved in the significant germline SV-CGI methylation associations (Fig. S2b, c). The above associations corrected for tumor histologic type, with the methylation differences remaining significant, even accounting for widespread differences in DNA methylation according to tumor histology. Without correcting for tissue histology, we found even greater numbers of CGIs with germline SVs associated with differential methylation (Fig. S2a), suggesting that for many of the uncorrected associations, the differences may involve methylation patterns specific to some but not all brain or CNS cell types. Previously, we associated somatic SVs with differential CGI methylation patterns in the CBTN cohort[33]. However, we observed relatively little overlap, close to chance expected, between the CGI-level germline SV-methylation and somatic SV-methylation associations (Fig. S2d), consistent with previous observations involving SV-expression associations[18] and reflecting how germline SVs primarily represent normal phenotypic variation rather than disease contributors.

Significant fractions of the SV-CGI methylation associations present in the CBTN cohort also involved SV-expression associations for the gene near the CGI (Fig. 2c). Using a p-value cutoff of <0.01 (linear modeling, any of regions 100 kb up- or downstream or within gene), 4428 CGI methylation probes had higher methylation associated with nearby SV breakpoints, of which 88 probes involved genes with lower mRNA expression associated with SV breakpoints ($p < 0.01$), a significant overlap ($p < 1E\text{-}13$, chi-squared test). Similarly, of the top 2011 CGI probes with lower methylation associated with SV breakpoints, 46 involved genes with higher expression associated with SVs (overlap $p < 1E\text{-}15$). There were also significant overlaps between CGI probes and genes having SV-associated higher methylation and expression, respectively (overlap $p = 0.004$), and between CGI probes and genes having SV-associated lower methylation and expression, respectively (overlap $p < 1E\text{-}20$). Regarding the genes with both SV-expression and SV-methylation associations (Fig. 2c), these shared significant overlaps with SV-eQTLs previously cataloged using normal brain tissues from the Genotype-Tissue Expression project (Fig. S2e)[2] and tended to show the same expression-methylation relationships across normal, non-brain tissues (Fig. S2f)[41].

Genes involving SV-associated differential CGI methylation with concordant mRNA changes included *NBPF3* and *DHCR7* (Fig. 2d, e). *NBPF3* represents Neuroblastoma breakpoint family, member 3, named so for belonging to a family of genes located primarily on duplicated regions of chromosome 1, with one gene in the family identified by the positional cloning of a translocation breakpoint from a neuroblastoma patient[42]. The NBPF genes involve high intragenic and intergenic sequence similarities that can expose these to illegitimate recombination, resulting in structural variation in the NBPF genes[42]. Inversion SVs with breakpoints respectively upstream of *NBPF3* and within the gene were associated with higher CGI methylation and lower expression, where *NBPF3* showed similar germline SV-expression associations and SV-methylation in both the PCAWG and CPTAC cohorts of adult cancers[13,18] (Supplementary Data 4 and Fig. S2g). The within-gene breakpoints of the inversion SV spanned a region of ~760 bp, involving very few recurrent SVs across patients (Supplementary Data 2). In contrast, a separate, highly-recurrent deletion SV within *NBPF3* was not strongly associated with differential methylation or expression (Fig. 2e). For *DHCR7*, SVs with downstream breakpoints within 100 kb of the gene (mostly duplications) were associated with lower CGI methylation and higher expression. *DHCR7* encodes 7-dehydrocholesterol reductase, and mutations in the gene can cause Smith-Lemli-Opitz syndrome, which disease involves low serum cholesterol levels and impaired brain development[43]. Overall, for ~75% of CGI probes involving both significant SV-methylation and SV-expression associations (Fig. 2c), SV breakpoint patterns significantly contributed information on expression in multivariate models incorporating combined SV and methylation patterns (Fig. S3), consistent with differential expression being uniquely associated with SVs.

In addition to the CBTN pediatric brain tumor patient cohort, we examined a combined dataset of germline SVs and tumor DNA methylation for 633 cancer patients from The Cancer Genome Atlas (TCGA, Fig. S4a). The overall patterns of CGI methylation differences associated with germline SVs observed above for the CBTN cohort were also observed in TCGA cohort, even if the respective top CGIs involved differed between the two (Fig. S4b–e and Supplementary Data 4). Notably, we found widespread germline SV-associated differential CGI methylation differences over chance expected in TCGA cohort (with even more differences observed when not correcting for tumor tissue of origin, Fig. S4b), CGI probes with SV-associated higher methylation and SV-associated lower methylation were respectively enriched for promoter-associated CGIs and gene body CGIs (Fig. S4c), and significant fractions of the SV-CGI methylation associations also involved SV-expression associations for the nearby gene (Fig. S4d). Of the 7725 CGI probes having higher methylation associated with germline SV breakpoints in the CBTN cohort ($p < 0.01$, 1 Mb region window), a significant number—175—were similarly associated between germline SVs and higher methylation in TCGA cohort (overlap $p < 1E\text{-}19$, one-sided Fisher's exact test; Fig. S4e). Similarly, 73 of the 2718 CGIs with mRNAs having lower methylation associated with germline SV breakpoints in the CBTN cohort ($p < 0.01$) were also similarly associated in TCGA cohort (overlap $p < 1E\text{-}20$). For TCGA cohort, there were fewer top CGIs with significant SV-methylation associations, due likely in part to the lower patient numbers involved.

Cancer-associated genes with SV-associated CGI methylation differences and concordant expression

In the CBTN cohort, of the top CGI probes with significant germline SV-methylation associations and with genes having SV-expression associations in the opposite direction from methylation, significant numbers involved either established cancer-related genes by the Catalogue of Somatic Mutations in Cancer (COSMIC)[44] or known cancer susceptibility genes from the literature[45] (Fig. 3a). Of the 134 CGI probes with methylation and gene mRNA associated with germline SV in inverse directions with respect to each other (from Fig. 2c, 88 + 46), 13 probes represent COSMIC genes including *TFG*, *MSH2*, *PAFAH1B2*, *PALB2*, *PPARG*, *MAP2K1* (enrichment $p = 0.04$, one-sided Fisher's exact test). Of the 88 CGI probes with positive methylation and negative expression associations (Fig. 2c), eight represented cancer susceptibility genes for which familial mutation or a loss of expression would associate with higher cancer risk (enrichment $p = 0.04$, one-sided Fisher's exact test). The analysis of COSMIC genes included those with well-established disease associations in pediatric brain tumors, including *MYC* and *MYCN*. Although we had previously found that these two genes involved somatic SV breakpoints associated with lower DNA methylation in pediatric tumors[33], analogous germline SV associations were not identified here. Few genes were involved in both germline SV-methylation and somatic SV-methylation associations across the CBTN cohort (Fig. S2d), though notable genes include *MSH2*, *BCOR*, and *HLA-DRB5*.

https://doi.org/10.1038/s41467-025-60110-y

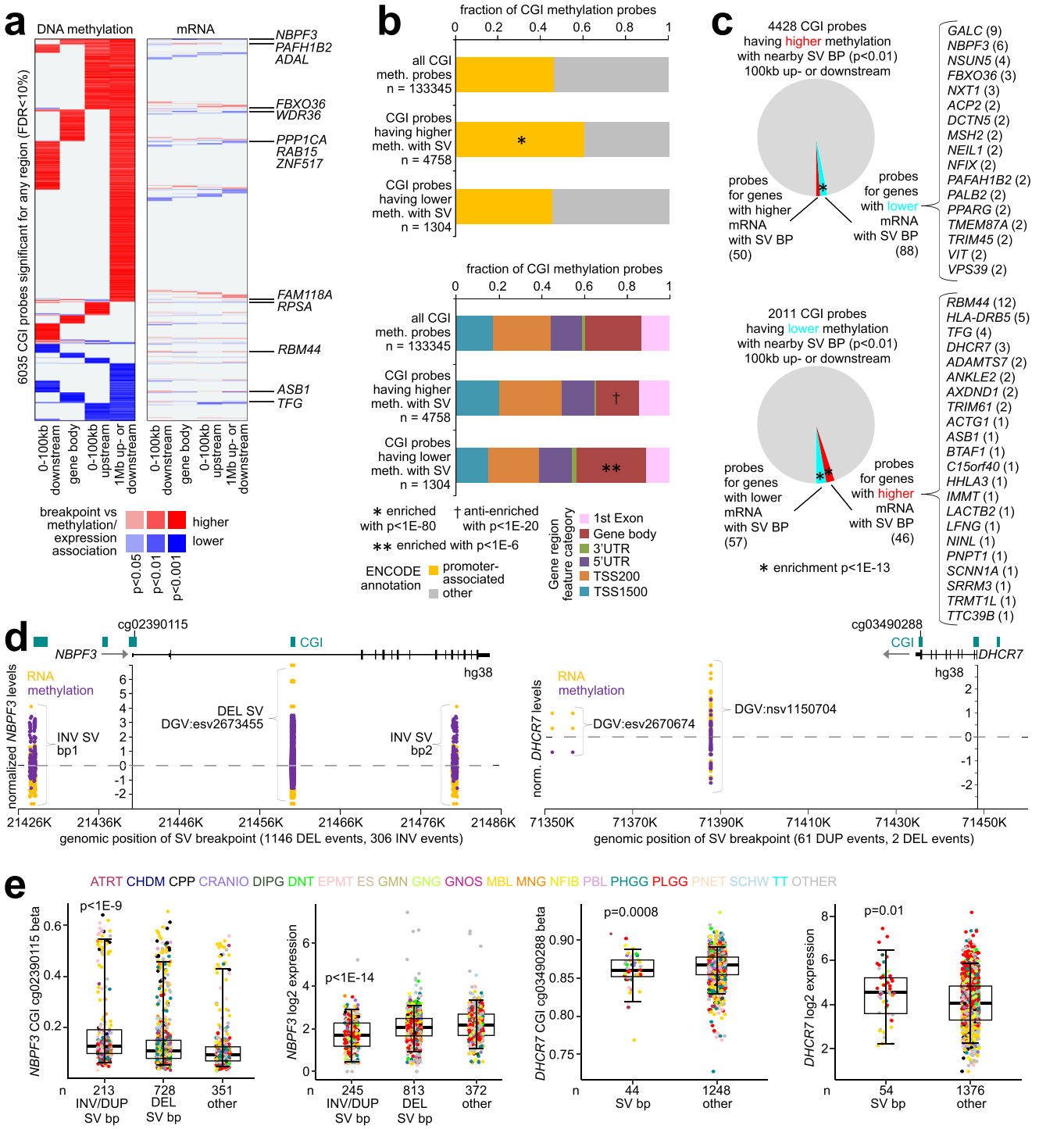
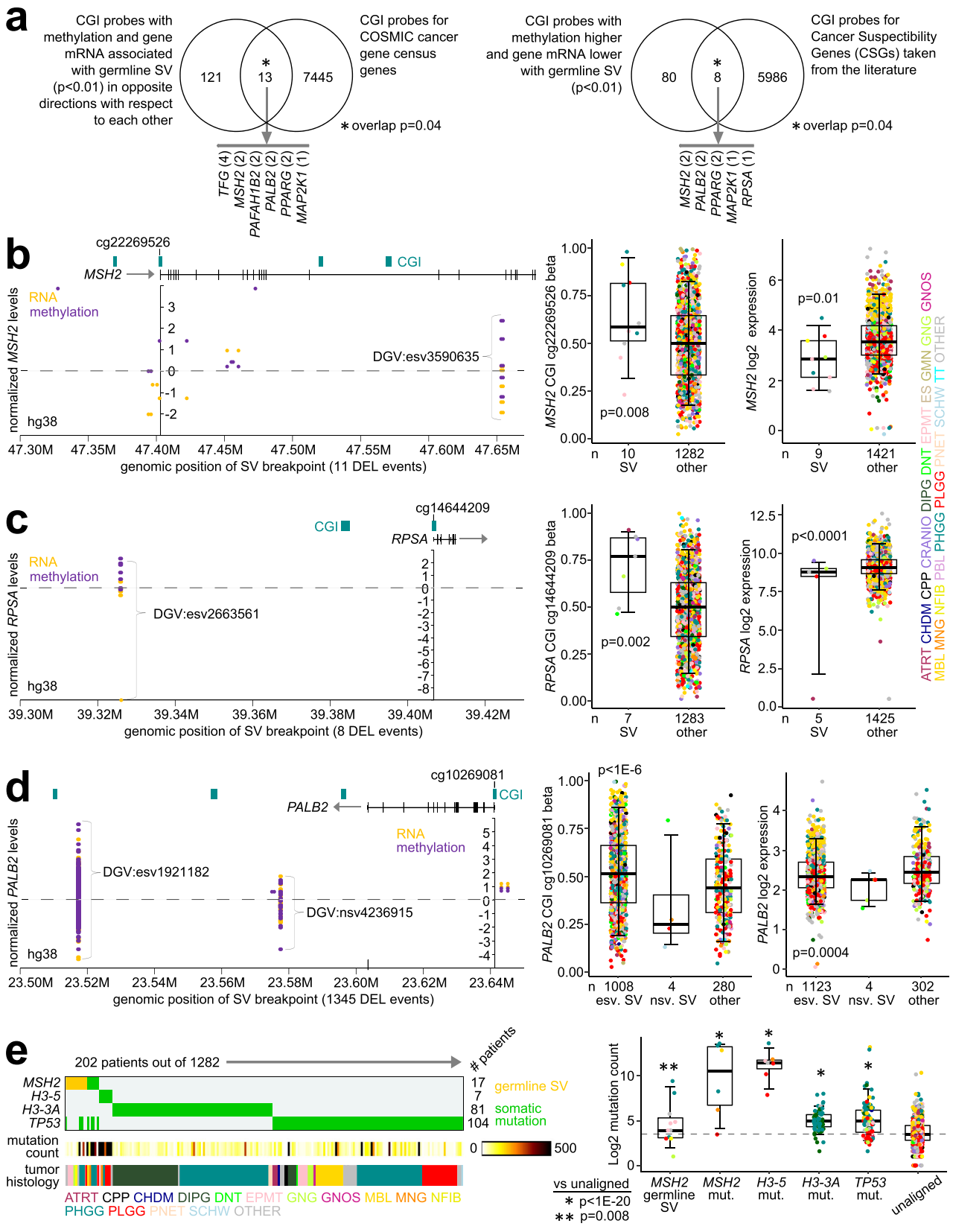


Fig. 2 | Genes with CGI methylation differences and concordant expression involving nearby germline SV breakpoints. a Heatmap of significance patterns for 6035 CGI probes with SV-associated DNA methylation differences in the CBTN cohort (FDR < 10%, linear model) for any genomic region window examined (SV breakpoints 100 kb upstream of the gene, 100 kb downstream, within the gene, or 1 Mb upstream or downstream, as indicated). Red denotes a significant association with higher methylation; blue, with lower methylation. The corresponding significance results for the CGI-associated genes are represented at the mRNA level. Genes listed to the right were significant for both methylation and expression at p < 0.001 (linear model) in opposite directions. **b** Top: Fraction of promoter-associated CGIs for those respectively associated (from part a) with higher or lower methylation (meth.) with SV breakpoints (BP). Bottom: Breakdown by probe position relative to the gene for CGIs associated with higher or lower methylation, respectively. TSS, transcription start site; UTR, untranslated region; n, CGI methylation probes. **c** Overlap between CGI probes with SV-associated methylation differences and nearby genes with corresponding SV-associated expression differences ($p < 0.01$ for both by linear modeling). Genes with mRNA- and methylation-SV breakpoint associations inverse to each other are listed by name with CGI probe numbers involved. For b and c, enrichment p-values by chi-squared test. In parts a-c, SV associations correct for histology (and gene copy for mRNA, as well as patient sex for X or Y loci). **d** CGI methylation (purple) and mRNA (orange) levels of *NBPF3* (left) and *DHCR7* (right), corresponding to SV breakpoints in the genomic region surrounding the respective genes. Each point represents a patient. Methylation and mRNA values are respectively normalized to standard deviations from the median (using logit- and log2-transformed values, respectively). **e** Boxplots of *NBPF3* and *DHCR7* CGI methylation (beta) and log2 expression by germline SV class. *P*-values comparing the first group versus all other patients by *t*-test on logit-transformed (methylation) or log2-transformed values (RNA). Boxplots represent 5%, 25%, 50%, 75%, and 95%; n, numbers of patients.

https://doi.org/10.1038/s41467-025-60110-y



Cancer susceptibility genes involving SV-associated higher CGI methylation with lower expression included *MSH2*[46], *RPSA*[47], and *PALB2*[48] (Fig. 3b–d). Deletion SVs within or immediately upstream of *MSH2* involved ten of the 1292 patients with methylation data (Fig. 3b). A recurrent deletion SV cataloged in DGV[34] (DGV, entry esv2663561) 100 kb upstream of *RPSA* involved seven of the 1292 patients (Fig. 3c). Regarding *PALB2*, a common and recurrent deletion SV 100 kb downstream of the gene, with DGV entry esv1921182, involved 1008 of the 1292 patients (Fig. 3d). Regarding the above cancer-associated genes, SV-expression associations were not present in TCGA cohort, and SV-methylation associations were found for only *TFG* (Fig. S4d,e and Supplementary Data 4), suggesting that the associations observed in

Fig. 3 | Canonical cancer-associated genes with CGI methylation differences and concordant expression involving germline SV breakpoints. a Venn diagrams representing overlapping top CGI probes between the CBTN-based germline SV-methylation-expression associations (from Fig. 2c) and probes for genes with established cancer association by COSMIC[44] (left) or for known cancer susceptibility genes from the literature[45]. Left diagram is for genes positively or negatively associated with SV breakpoints, with corresponding CGI methylation association in the opposite direction; right diagram, for negatively associated genes with positive CGI methylation association. Significance of overlap by one-sided Fisher's exact test. **b** Left: CGI methylation (purple) and mRNA (orange) levels of *MSH2*, corresponding to SV breakpoints in the genomic region surrounding the gene. Each point represents a patient. Methylation and mRNA values are respectively normalized to standard deviations from the median (using logit- and log2-transformed values, respectively). Right: Boxplots of *MSH2* CGI methylation (beta) and log2 expression by germline SV class. *P*-values versus other by *t*-test on logit-transformed (methylation) or log2-transformed values (RNA). **c** Similar to part b, but for *RPSA*-associated germline SVs and *RPSA* differential CGI methylation and expression. **d** Similar to part b, but for *PALB2*-associated germline SVs and *PALB2* differential CGI methylation and expression (*p*-values compare DGV:esv1921182[34] SV, or "esv. SV", versus the rest of patients). DGV:nsv4236915 ($n = 4$ patients), or "nsv. SV", in the same *PALB2* downstream region, was not associated with differential methylation. **e** Across 202 impacted patients, *MSH2* germline SV events versus somatic mutation events involving *MSH2*, *H3-5*, *H3-3A*, and *TP53* (left), along with their respective associations with overall tumor somatic mutation burden (right); *p*-values by *t*-test. All boxplots represent 5%, 25%, 50%, 75%, and 95%, with n representing the numbers of patients.

the CBTN cohort would be specific to brain tissues or tumors. Consistent with the well-known role of *MSH2* in DNA mismatch repair, patients harboring *MSH2* germline SVs tended to have a higher tumor mutation burden, though not as dramatic as observed for other patients harboring somatic *MSH2* mutations (Fig. 3e).

Germline SV breakpoints falling within epigenetic elements are associated with differential CGI methylation

DNA and histone methylation cooperate to maintain the cellular epigenomic landscape[49]. We focused here on the subset of germline SVs with breakpoints falling within any one of the following epigenetic elements: CGIs, H3K36me3 (enriched at the gene bodies and playing a role in transcriptional activation[50]), H3K9me3 (often associated with heterochromatin and involved with transcriptional silencing[51]), and H3K27me3 (involved with transcriptional silencing[52]). We first tabulated germline SV breakpoint-to-gene associations for all genes and samples represented in the combined CBTN CGI methylation and germline SV datasets, taking the closest SV with a breakpoint within 1 Mb of the gene. For each patient-to-gene association, we then determined if that association involved higher or lower CGI methylation near the gene in conjunction with the SV breakpoint ($p < 0.01$ across patients by linear modeling with covariates, with methylation levels higher or lower in the patient consistent with the global association). Compared to all SV-gene associations, the subset involving genes with higher CGI methylation was highly enriched for SV breakpoints falling within CGIs or H3K36me3 marks (Fig. 4a, $p <$ 1E-200, chi-squared test). In contrast, SV-gene associations involving lower CGI methylation were highly enriched for SV breakpoints falling within H3K9me3 marks (Fig. 4a). SV-gene associations involving higher CGI methylation were also moderately enriched for SV breakpoints within H3K27me3 marks. In addition to the results from the CBTN cohort reported above, we observed the same overall enrichment patterns involving epigenetic elements in the independent TCGA cohort (Fig. S5a), though for TCGA there was also a strong enrichment pattern involving higher CGI methylation and SV breakpoints falling within H3K27me3. The above associations suggest interrelated functional relationships involving germline SVs, CGI methylation, and histone methylation.

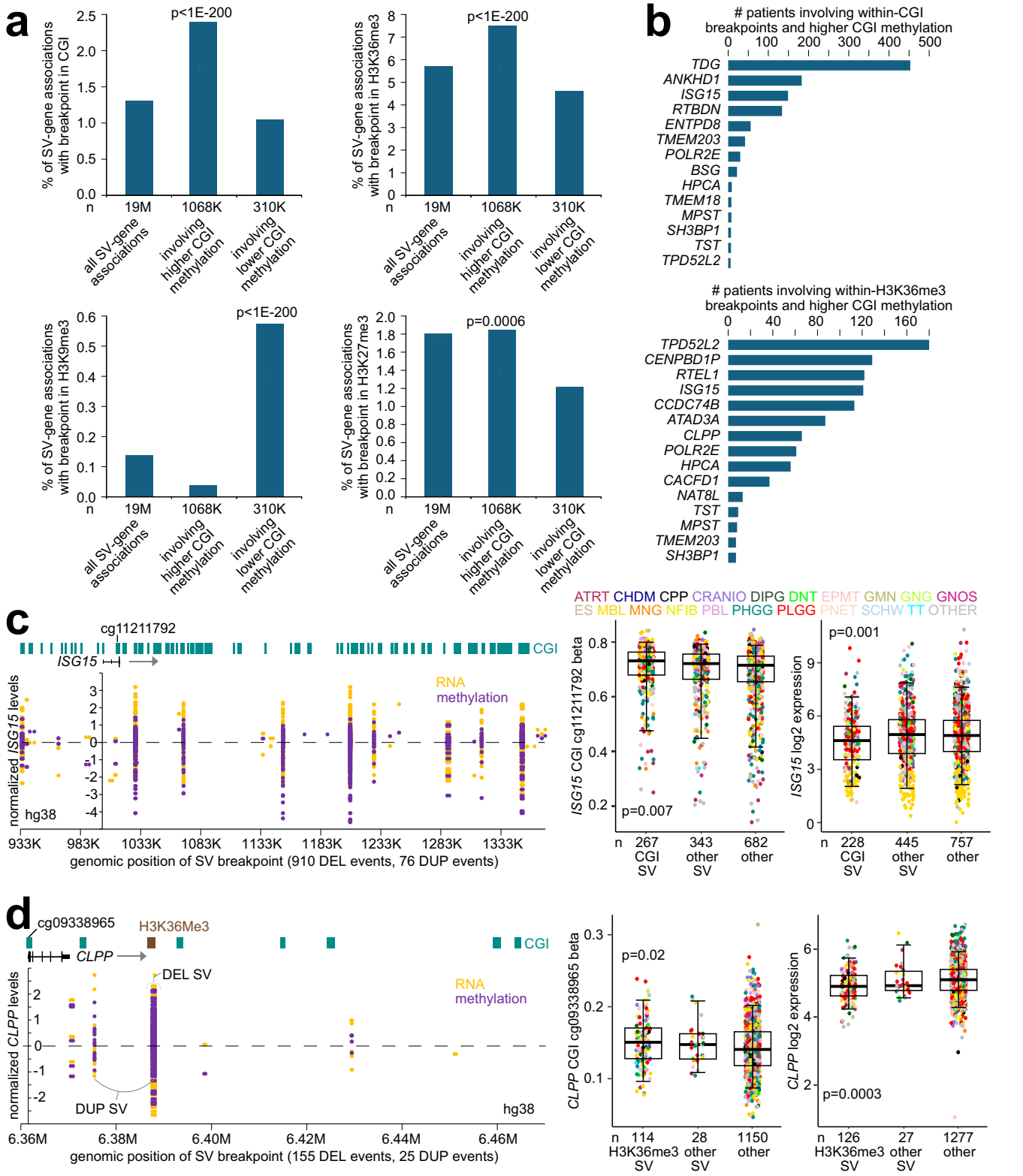
Regarding the within-CGI and within-H3K36me3 germline SV breakpoint associations, we focused on the top genes for which over half of patients harboring within-element breakpoints had CGI methylation greater than the median and for which there was a corresponding SV-expression association ($p < 0.05$, linear model) in the opposite direction from that of the CGI methylation (Fig. 4b and Supplementary Data 5). Regarding the within H3K9me3 associations, very few involved genes with positive CGI methylation association (Supplementary Data 5). Top genes involving within-CGI germline SV breakpoint associations included *ISG15* (Fig. 4c), while top genes involving within-H3K36Me3 germline SV breakpoint associations included *ISG15* (Fig. S5b,c) and *CCLP* (Fig. 4d). Regarding *ISG15*, germline SVs with breakpoints falling within any of the CGIs in the genomic region spanning 68 kb upstream of the gene to 355 kb downstream were associated with both higher methylation of a CGI within *ISG15* (involving 149 patients with methylation greater than the median, Supplementary Data 5) and lower *ISG15* expression (Fig. 4c). Regarding *CCLP*, recurrent duplication and deletion SVs with breakpoints falling within a H3K36me3 element ~17.7 kb downstream of the gene were associated with higher *CCLP* CGI methylation (involving 66 patients) and decreased *CCLP* expression (Fig. 4d). In some instances, the observed SV-associated differences for methylation or expression may appear subtle, consistent with normal phenotypic variation, where such small molecular changes over the patient's life may conceivably have an effect.

Differential DNA methylation near enhancers involving nearby germline SV breakpoints

We also looked beyond CGIs to examine DNA methylation changes involving gene enhancers and nearby germline SV breakpoints, as DNA methylation at distal enhancer regions has been implicated in gene regulation[24–27]. For each of 30292 enhancers[53], we mapped the nearest DNA methylation probe and determined its methylation association with nearby SV breakpoints within 100 kb of the enhancer (Fig. 5a). We found 254 enhancers with an SV-methylation association significant at FDR < 10% (linear modeling with covariates, Fig. 5b and Supplementary Data 6). For 159 of the 254 significant enhancers, SV breakpoints associated with lower methylation, paralleling our previous observations involving somatic SVs in CBTN cohort[33], where most of the significant enhancers there also had a negative SV-methylation association. Significant fractions of the germline SV-enhancer methylation associations present in the CBTN cohort also involved SV-expression associations for genes within 500 kb of the enhancer (Fig. 5c). Of the 2484 genes with a nearby enhancer (within 500 kb) having SV-associated lower methylation ($p < 0.01$), 69 and 78 had positive and negative SV-expression associations ($p < 0.01$, linear model, any genomic region window), respectively, representing significant gene set overlaps ($p = 0.02$ and $p = 0.0004$, respectively, one-sided Fisher's exact tests). Similarly, of the 1768 genes with a nearby enhancer having SV-associated higher methylation, 57 and 50 had positive and negative SV-expression associations, respectively ($p = 0.001$ and $p = 0.05$, respectively, one-sided Fisher's exact tests).

NBPF1 and *NUDT2* are two example genes with associated enhancers showing differential methylation in conjunction with nearby germline SV breakpoints (Fig. 5d,e). Regarding *NBPF1*, a recurrent inversion SV of size ~333 kb upstream of the gene had breakpoints within 6 kb of the gene start, involving 77 patients of 1292 with tumor DNA methylation data. *NBPF1* has an enhancer element within the gene boundaries, for which DNA methylation (by probe cg12434908, closest to the enhancer) was markedly higher in tumors from patients with the inversion SV, corresponding to lower expression levels (Fig. 5d). *NBPF1* also had a CGI methylation array probe (cg01599189) ~ 537 bp upstream of the gene that also showed higher methylation in association with the inversion SV, though of borderline

https://doi.org/10.1038/s41467-025-60110-y



statistical significance ($p = 0.035$ compared to $p < 1E\text{-}8$ for the enhancer methylation, Supplementary Data 3). Of all genes within 500 kb of the *NBPF1*-associated enhancer, only *NBPF1* had a significant SV-expression association (Supplementary Data 6). Regarding *NUDT2*, a common and recurrent duplication SV of size 118 bp, with DGV entry nsv1079004, occurred upstream of the gene and involved 662 of the 1292 patients. An enhancer element further upstream of nsv1079004 showed lower DNA methylation in patients with the SV, while *NUDT2* showed correspondingly higher expression (Fig. 5e). Of all genes within 500 kb of this enhancer, only *NUDT2* had a significant SV-expression association. No CGI array probes for *NUDT2* had significant SV-methylation associations.

Germline SVs and associated CGIs involving patient survival

While most CGIs with germline SV-associated differential methylation in the CBTN cohort would presumably have no role in cancer, a subset

Fig. 4 | Germline SV breakpoints falling within CGIs and histone epigenetic marks are associated with differential CGI methylation. a For each of four epigenetic elements—CGIs, H3K36me3, H3K9me3, and H3K27me3—the percentages of patient-level germline SV-gene associations with an SV breakpoint falling within the element are tabulated for the entire set of SV-gene associations and the subsets of associations involving higher or lower CGI methylation, respectively ($p < 0.01$ by linear modeling across all patients, with methylation beta levels in the patient tumor greater or less than the median across tumors in a direction consistent with the global association). Enrichment p-values by chi-squared test. n, numbers of SV-gene associations. **b** For within-CGI germline SV breakpoint associations (top) and within-H3K36me3 SV breakpoint associations (bottom), the numbers of patients with higher CGI methylation (greater than the tumor median) combined with SV breakpoint falling within the epigenetic element, for each the top genes shown. For the top genes shown, over half of patients harboring within-element SV breakpoints had CGI methylation greater than the median, and there was a corresponding SV-expression association in the opposite direction from that of the CGI methylation. **c** Left: CGI methylation (purple) and mRNA (orange) levels of *ISG15* corresponding to SV breakpoints in the genomic region surrounding the gene. Each point represents a patient. Methylation and mRNA values are respectively normalized to standard deviations from the median (using logit- and log2-transformed values, respectively). Right: Boxplots of *ISG15* CGI methylation (beta) and log2 expression by germline SV class. *P*-values comparing the first group (SV breakpoints within CGI) versus all other patients by *t*-test on logit-transformed (methylation) or log2-transformed values (RNA). Boxplots represent 5%, 25%, 50%, 75%, and 95%, with n representing the numbers of patients. **d** Similar to part c, but for *CCLP*-associated germline SVs (including SVs involving H3K36me3 marks) and *CCLP* differential CGI methylation and expression.

of these CGIs and related genes may have potential roles. Germline SVs associated with patient survival in pediatric brain tumors may involve genes having biological roles in more aggressive disease and may represent candidate cancer risk variants for genetic testing. Overall survival data were available for 1317 CBTN patients with germline WGS data (Supplementary Data 1). For each of the 14148 genes represented in the CBTN SV breakpoint and CGI methylation probe datasets, we associated the germline SV breakpoint pattern with survival (by Cox, accounting for tumor histologic type) for each of the SV breakpoint matrices respectively involving the four genomic region windows examined for SV-CGI methylation associations (Supplementary Data 7). Using a strict FDR cutoff of 10%, 226, 82, 375, and 356 top CGI probes had the genes respectively associated with patient survival for gene-centric regions 1 Mb, within gene, 100 kb upstream, and 100 kb downstream (Supplementary Data 7), and using a relaxed p-value cutoff of 0.05, about 50% more significant genes were found over chance expected due to multiple testing, indicative of information in the patient germline representing a factor in patient outcomes. In conducting integrative analyses to identify top CGIs involving SV-associated survival patterns, we used a more relaxed p-value cutoff for each individual analysis to limit false negatives[13,18,54,55]. In particular, for each of the genomic region windows 100 kb upstream of the gene, 100 kb downstream of the gene, and within the gene body, we observed highly significant overlap between CGI probes with SV-methylation association ($p < 0.01$ by linear model with covariates) and CGI probes for genes with germline SV breakpoint patterns positively associated with pediatric brain tumor patient survival (one-sided Cox $p < 0.05$, correcting for tumor histologic type, Fig. 6a). The observed enrichment patterns suggested potential roles involving some of the observed SV-CGI methylation associations with more aggressive pediatric brain tumors.

Across all four genomic region windows examined, we identified 1155 CGI probes for which germline SV breakpoints near the annotated gene had both positive associations with worse overall survival and either positive or negative associations with differential CGI methylation (Fig. 6b and Supplementary Data 7, 1006 CGI probes having positive SV-methylation association). As we had not used CGI methylation associations with survival to derive the 1155-CGI signature, we tested whether the signature could collectively predict patient outcome based on tumor DNA methylation data, whereby the signature significantly stratified patients into high-, low-, and intermediate-risk groups (Fig. 6c, Log rank $p < 1E\text{-}5$, accounting for tumor histologic type). Several factors, not limited to SVs, may underlie the survival-associated methylation patterns in tumors observed across the CBTN cohort.

Genes represented in the 1155-CGI signature included *POLD4*, which encodes a component of the DNA polymerase delta complex. Germline deletion SV breakpoints 412 kb downstream of *POLD4* were associated with higher CGI methylation and lower expression (Fig. 6d and e). *POLD4*-associated germline SV breakpoint patterns, higher *POLD4* CGI methylation, and lower *POLD4* expression were all associated with worse patient survival (Fig. 6f). Only a fraction of the germline SV-CGI methylation associations involved coordinated SV-expression associations (Fig. 2c), which was also the case with the above 1155-CGI signature (Supplementary Data 7). Of all the genes involving the 1155-CGI signature, *POLD4* was the most notable in having coordinated SV-mRNA associations along with robust germline SV and mRNA survival associations. Other CGI probes involved survival associations with both methylation and germline SVs but not for the annotated gene mRNA (Fig. S6a and b). Of the other genes individually highlighted above in our study, only *NBPF1* and *RPSA* had mRNA associations with patient outcomes, but with worse rather than better outcomes, as similarly observed for other tumor suppressor genes such as *TP53* (Fig. S6c).

Discussion

Our present study systematically catalogs the specific genes and nearby CGIs with germline SV-associated differential expression and methylation, respectively, across pediatric brain and CNS tumors. Most of these associations would represent normal phenotypical variations across individuals. A subset of genes highlighted in our study as having established cancer roles or patient survival associations could warrant further investigation in the context of pediatric brain tumors. Genes of interest may involve both common and rare germline SVs. The respective sets of germline SV-methylation and SV-expression associations for pediatric brain tumors differed substantially from analogous associations for adult cancers, reflecting the need to explore SV-methylation associations across different cellular and disease contexts. Our catalog of associations represents a resource for future studies involving normal germline variation across brain tissues or disease-associated variants. Knowledge on genetic predisposition for brain and CNS tumors remains sparse[56]. Specifically, GWAS or case-control studies associating non-coding germline SVs, by and large, remain to be carried out, analogous to similar studies focusing on SNPs[56,57]. Any disease-associated non-coding germline SVs arising in future studies would need context to help interpret the association in terms of what nearby genes appear affected. Our study results can provide this context, revealing any expression and methylation associations involving SVs of interest. Our study measured the effect of germline SVs on methylation in the tumor environment and not a normal environment, and there could potentially be higher order interactions between the tumor environment and the cell that manifest as differential methylation not observed in normal cells.

We identified SV-methylation associations involving known cancer susceptibility genes, including *MSH2*, *RPSA*, and *PALB2*. Any disease-associated germline SVs that remain to be established would likely include many having low penetrance but which could still inform on disease biology. Analogous to results of genomic studies defining the somatic landscape of pediatric brain tumors[19,33,58], germline variants involving disease may not have immediate clinical impact—e.g., as

https://doi.org/10.1038/s41467-025-60110-y

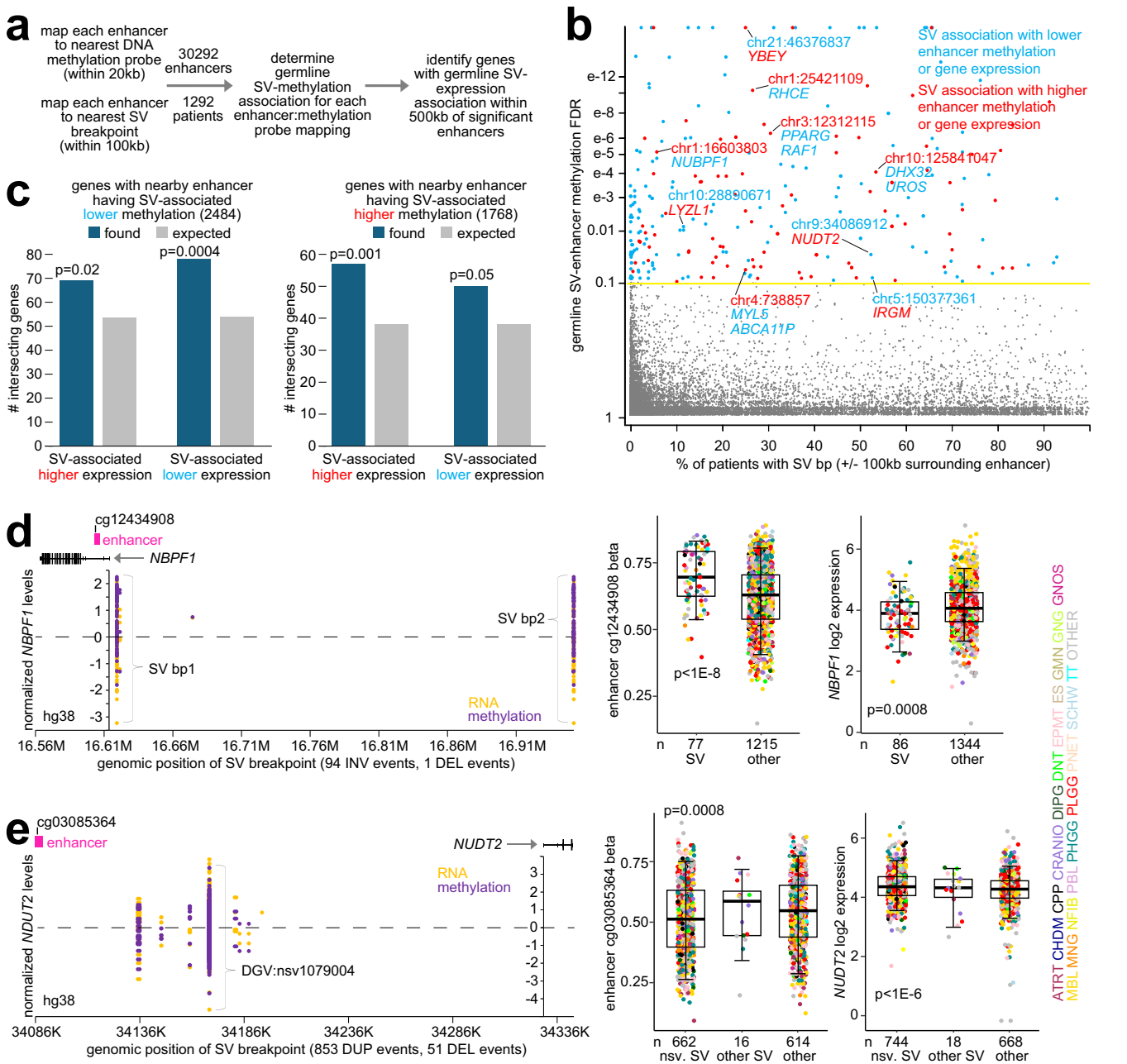
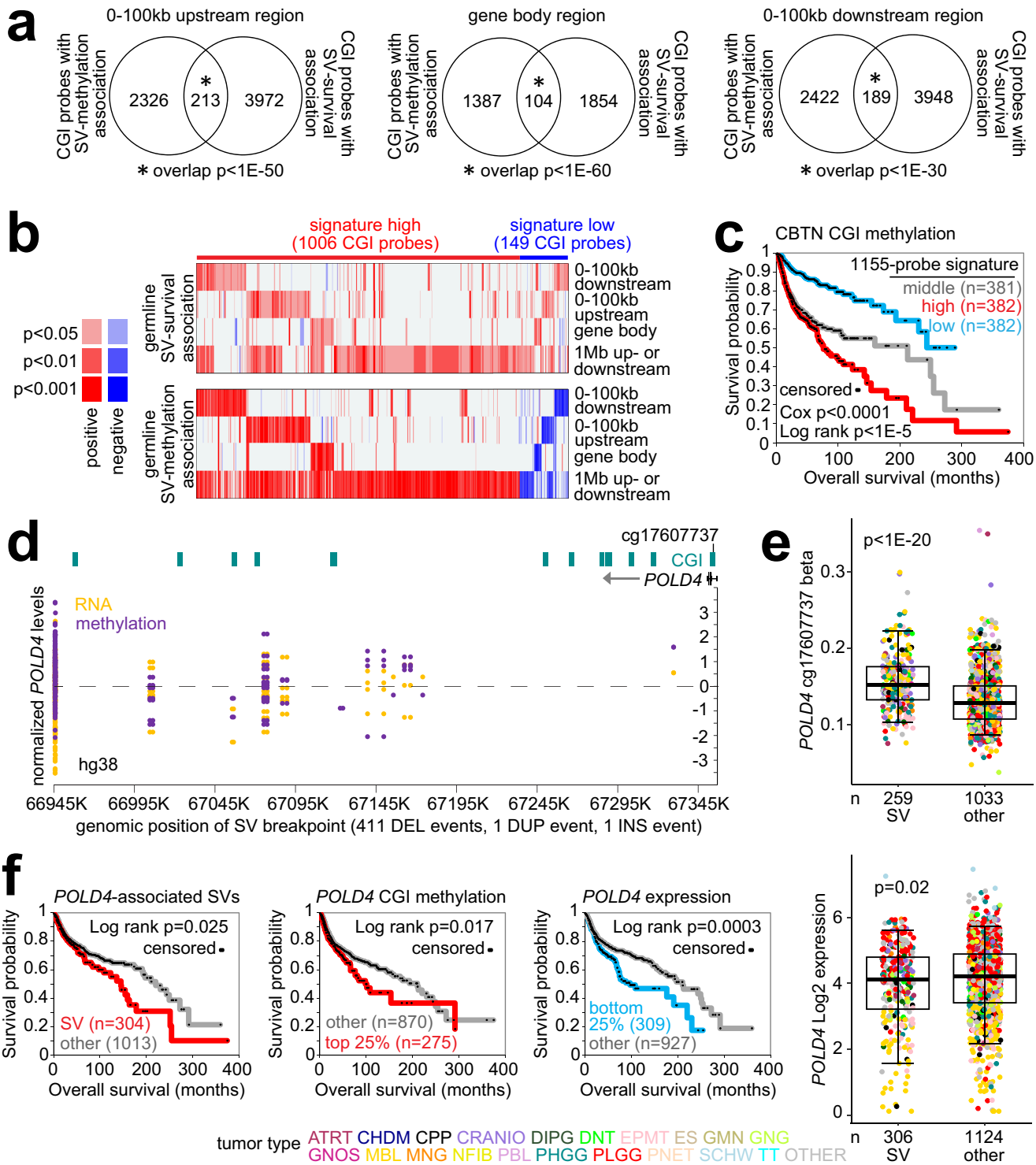


Fig. 5 | Differential DNA methylation near enhancers associated with nearby germline SV breakpoints. a Schematic of the analysis approach. For each of 30292 enhancers[53], we mapped the nearest DNA methylation probe (within 20 kb)[33] and determined its methylation association with nearby germline SV breakpoints (within 100 kb) by linear modeling correcting for histologic type. **b** Significance of SV-enhancer methylation association (involving breakpoints within the region 100 kb upstream or downstream) is plotted (y-axis) versus the percentage of tumors with breakpoints (x-axis). **c** For top enhancers with significant SV-methylation association ($p < 0.01$ by linear modeling, involving>9 patients), the genes within 500 kb of the enhancer were examined for SV-expression association ($p < 0.01$ by linear modeling with covariates, any region). One-sided Fisher's exact tests compare the number of genes with significant SV-expression association with chance expected. **d** Left: DNA methylation (purple) and mRNA (orange) levels of *NBPF1*, corresponding to SV breakpoints in the genomic region surrounding the gene and its associated enhancer. Each point represents a patient. Methylation and mRNA values are respectively normalized to standard deviations from the median (using logit- and log2-transformed values, respectively). Right: Boxplots of *NBPF1* enhancer methylation (beta) and log2 expression by germline SV class. *P*-values versus other by t-test on logit-transformed (methylation) or log2-transformed values (RNA). Boxplots represent 5%, 25%, 50%, 75%, and 95%, with n representing the numbers of patients. **e** Similar to part d, but for *NUDT2*-associated germline SVs, *NUDT2* differential expression, and differential methylation of nearby enhancer (*p*-values compare DGV:nsv1079004 SV, or "nsv. SV", versus the rest of patients). In the boxplot, "other SV" refers to SVs represented in the *NUDT2* upstream region on the left plot that were not DGV:nsv1079004, where these other SVs were not associated with differential methylation or expression.

part of an assay to assess the risk of developing cancer—but would instead shed light on the specific genes that may contribute to disease. In general, cancer predisposing genes tend to be cancer type specific[59]. Yet, our present study's SV-methylation and SV-expression associations would cut across multiple histologic types rather than being confined to a single type. Cancer risk-associated genes can involve rare germline variants having high penetrance, uncommon variants having moderate penetrance, and common variants having low penetrance[60]. For the low penetrance germline variants, the predictability for disease phenotype may be low and, therefore, difficult to interpret regarding

https://doi.org/10.1038/s41467-025-60110-y



any treatment decisions. Still, it is conceivable that some germline variants may not strongly predispose one to disease and yet have a contributory role in patient survival. For example, it is understood that some alterations present in the germline can determine responses to anticancer therapies and their toxicities[61]. The global associations of germline SV breakpoint patterns with patient survival in CBTN cohort reflects similar findings made in the PCAWG cohort[18], but involving different genes that could warrant further investigation, e.g., as larger multi-omic cohorts become available[20].

The germline SV-methylation associations cataloged here would help explain a fraction of the corresponding SV-expression associations observed. Of the SV-methylation associations, we had focused mainly on genes for which SV-associated differential expression was in the opposite direction from that of methylation, although DNA methylation and expression relationships can also be positive, e.g., when a repressive element is methylated[11,37,62]. At the same time, most germline SV-methylation associations observed here may not have a ready explanation. Most associations involved cis-regulatory regions, with relatively few deletion or duplication SVs spanning CGIs. Our present study made intriguing observations of germline SV breakpoints within epigenetic elements being enriched for differential CGI methylation associations. A significant fraction of SV associations

Fig. 6 | Germline SV breakpoint patterns with differential CGI methylation involving pediatric brain tumor patient survival. a Overlapping CGI probes for genes with germline SV breakpoint patterns associated with worse pediatric brain tumor patient survival and CGI probes with SV-methylation association for each of the genomic region windows represented. Enrichment p-values by one-sided Fisher's exact test. **b** Combining germline SV data with patient survival data and tumor methylation data across the CBTN patients, 1155 CGI probes had both a positive association between germline SV breakpoints near the gene and worse overall survival and a positive or negative germline SV-methylation association, for the same genomic region window. **c** Association of the 1155-CGI probe signature from part b with patient survival in the CBTN cohort, based on scoring of the tumor methylation profiles. The direction of each CGI probe in the 1155-CGI probe signature, as applied here to the CBTN methylation dataset, is based on the direction of the germline SV-methylation association. Survival association *p*-values correct for tumor type. **d** CGI methylation (purple) and mRNA (orange) levels of *POLD4*, corresponding to SV breakpoints in the genomic region surrounding the gene. Each point represents a patient. Methylation and mRNA values are respectively normalized to standard deviations from the median (using logit- and log2-transformed values, respectively). **e** Boxplots of *POLD4* CGI methylation (top) and log2 expression (bottom) by germline SV class. *P*-values versus other by *t*-test on logit-transformed or log2-transformed values. Boxplots represent 5%, 25%, 50%, 75%, and 95%. **f** Associations with worse patient survival for *POLD4*-associated germline SV breakpoint patterns (left), for higher *POLD4* CGI methylation (middle), and for lower *POLD4* expression (right). *POLD4* SV patterns involve breakpoints 412 kb downstream of the gene (as represented in part d). For parts a and b, germline SV breakpoint survival associations by one-sided $p < 0.05$ by Cox accounting for tumor type (minimum of 10 patients with breakpoint in region), and germline SV-expression associations by $p < 0.01$ by linear model correcting for covariates. For parts a and b, Cox *p*-values are one-sided; for part c, Cox *p*-value is two-sided. For parts c-f, n represents numbers of patients.

with higher CGI methylation involved SV breakpoints falling within CGIs (not limited to the differentially methylated CGI) or H3K36me3 histone marks. A significant fraction of SV associations with lower CGI methylation involved SV breakpoints falling within H3K9me3 histone marks. We observed these same overall enrichment patterns in the independent TCGA adult cancer cohort. A strong correlation is known to exist between the genome-wide distribution of DNA and histone methylation, suggesting an intimate relationship between these epigenetic marks that remains to be further elucidated[49]. For most of the SV-CGI methylation associations, the corresponding genes did not appear differential at the expression level, which may point to other roles for DNA methylation that remain to be explored. Along with our present study's results, future studies of combined DNA methylation and non-coding variants by WGS will allow us to map out and understand human phenotypic variation more thoroughly across various tissue-based systems and disease contexts.

Our present study utilized short-read WGS, where future studies utilizing long-read WGS across hundreds of samples would likely uncover additional germline SV-methylation and SV-expression associations. WGS short reads have limitations in finding certain types of SVs, whereas long reads are advantageous for SV calling because they can span problematic regions including long tandem repeats and segmental duplications[36,63]. The methods for detecting SVs from short reads vary in the type of information they exploit[63], and so our present study utilized SV calls made by two different algorithms to help minimize false positive calls, as practiced by others[64,65]. Also, the integration of results between orthogonal platforms (e.g., WGS and DNA methylation) greatly reduces the expected number of false positive associations, as the sets of top associations identified must be significant enough to rise above any noise, e.g., from false positive SV calls or multiple testing of features. Another potential issue regards false negative SV calls, as opposed to false positive calls, as short-read sequencing is likely missing or may have difficulty resolving certain types of SVs (especially in repetitive regions)[66]. In examining fixed genomic region windows near a given gene, our analytical approach would not rely on precise SV mapping or typing. It might be some time before a combined dataset of long-read WGS, DNA methylation, and expression in pediatric brain tumors—comparable to the CBTN—will be made available. However, we anticipate that the overall trends and phenomenon identified here will be further substantiated in future studies, potentially involving additional genes.

Methods

Patient cohorts

Results are based on data generated by the CBTN. Patients were consented by one of 32 participating sites and enrolled on a local IRB-approved protocol which includes key language to enable prospective collection of, future research on, and sharing of, de-identified surgical specimens, patient demographics, medical history, diagnoses, treatments, and clinical imaging[20]. Tumor molecular profiling data were generated through informed consent as part of CBTN efforts and analyzed here per CBTN's data use guidelines and restrictions. CBTN Member institutions include the following: Akron Children's Hospital, Ann & Robert H. Lurie Children's Hospital of Chicago, Beijing Tiantan Hospital Neurosurgery Center, Children's Healthcare of Atlanta, Children's Hospital of Philadelphia, Children's National Hospital, Children's of Alabama, Dayton Children's Hospital, Doernbecher Children's Hospital, Hassenfeld Children's Hospital at NYU Langone, Hudson Institute of Medical Research, Intermountain Primary Children's Hospital, Johns Hopkins All Children's Hospital, Johns Hopkins Medicine, Joseph M. Sanzari Children's Hospital at Hackensack University Medical Center, Lucile Packard Children's Hospital Stanford, Maria Fareri Children's Hospital at Westchester Medical Center, Meyer Children's Hospital, Michigan Medicine C.S. Mott Children's Hospital, Nicklaus Children's Hospital, Orlando Health Arnold Palmer Hospital for Children, Seattle Children's Hospital, St. Louis Children's Hospital, Sydney Children's Hospital in Randwick, Texas Children's Hospital, UCSF Benioff Children's Hospital, University Children's Hospital Zürich, University of Iowa Stead Family Children's Hospital, UNC Chapel Hill - North Carolina Children's Hospital, UPMC Children's Hospital of Pittsburgh, Wake Forest Baptist Health, and Weill Cornell Medicine.

At the time of this study, 1292 patients had data available for both WGS from the patient blood normal sample (at 30x coverage) and DNA methylation by Illumina array for the tumor sample. Tumor samples in this CBTN cohort spanned at least 33 different tumor histologic types (Supplementary Data 1): APTAD, Adenoma; ATRT, Atypical Teratoid Rhabdoid Tumor; CHDM, Chordoma; CNC, Neurocytoma; CPC, Choroid plexus carcinoma; CPP, Choroid plexus papilloma; CRANIO, Craniopharyngioma; DIPG, Diffuse intrinsic pontine glioma; DNT, Dysembryoplastic neuroepithelial tumor (DNET); EPM, Subependymal Giant Cell Astrocytoma (SEGA); EPMT, Ependymoma; ES, Ewing's Sarcoma; GMN, Germinoma; GNBL, Ganglioneuroblastoma; GNG, Ganglioglioma; GNOS, Glial-neuronal tumor not otherwise specified (NOS); HMBL, Hemangioblastoma; LCH, Langerhans Cell histiocytosis; MBL, Medulloblastoma; MNG, Meningioma; MPNST, Malignant peripheral nerve sheath tumor; NBL, Neuroblastoma; NFIB, Neurofibroma/Plexiform; ODG, Oligodendroglioma; PBL, Pineoblastoma; PCNSL, Primary CNS lymphoma; PHGG, High-grade glioma/astrocytoma (WHO grade III/IV); PLGG, Low-grade glioma/astrocytoma (WHO grade I/II); PNET, Supratentorial or Spinal Cord primitive neuroectodermal; RMS, Rhabdomyosarcoma; SARCNOS, Sarcoma; SCHW, Schwannoma; TT, Teratoma; and Other/unspecified. The CBTN's regulatory for CNS tumors intentionally allowed for the broad collection of abnormal cell growth, not necessarily specific to brain-specific cell types. Likewise, all patients with available CBTN data on combined DNA methylation and germline WGS formed the basis of our study. We also analyzed a CBTN cohort of 1430 patients with combined germline WGS and RNA-sequencing to identify genes with differential expression associated

with germline SVs—1089 of these patients being part of the above 1292 cohort with combined DNA methylation and germline WGS. For some patients, multiple tumors were profiled for DNA methylation or for RNA-seq (e.g., recurrent or progressive and initial tumor); for these patients, one tumor was randomly selected for this present study (Supplementary Data 1). Analysis of patient data was limited to data made publicly available in accordance with the informed consent. Protected CBTN molecular data, including raw sequencing data, are available under restricted access. To gain access to protected data, one can apply to the CBTN and sign a Data Use Agreement. Patient sex was self-reported.

The results here are also based partly on data from The Cancer Genome Atlas (TCGA) Research Network. Previously, we carried out combined germline WGS and RNA-seq analysis for 1218 adult cancer cases from TCGA and the International Cancer Research Consortium (ICGC)[18], with germline SVs being generated as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium efforts[65]. Of the TCGA cases with available DNA methylation data[15], 633 had germline SV calls by WGS as generated by PCAWG. Tumors in this TCGA 633-patient cohort spanned 21 TCGA projects, each representing a specific tumor type. Tumor molecular profiling data were generated through informed consent as part of previously published studies and analyzed per each original study's data use guidelines and restrictions[28,65].

Molecular profiling datasets

Based on WGS, CBTN made Germline SV calls using Manta (v1.4.0)[67] and SVABA (v1.1.0)[68] algorithms. The hg38 reference used for SV calling was limited to canonical chromosome regions. We accessed the germline SV VCF files during January, February, and June of 2024 through the CBTN Cavatica site (https://cbtn.org), using a private link provided by the CBTN for protected patient genomic data. Manta algorithm classified each SV call as one of the following: tandem duplications, insertions, deletions, inversions, and translocations. We used only germline SV calls that passed quality filters in the analyses. Germline SV calls from Manta and SVABA were pairwise joined based on SV position, allowing 200 bp slop at the breakpoints. For this merged germline SV call set, we used the Manta SV genomic coordinates throughout the study and required a minimum SV size of 10 bp (with 99.93% of called SVs in our dataset being of size 50 bp or greater[35]). We obtained data on tumor somatic mutation counts and on somatic mutation in selected genes (by patient) from the PedCBioportal.

Using the combined dataset of 1430 patients with germline SVs combined with RNA expression, we filtered out of the merged germline SV call set any SVs that were over-represented in either the first CBTN cohort or the later CBTN-X01 cohort[58], to minimize any potential batch effects. The combined DNA methylation and germline SV dataset included 228 patients not analyzed previously for germline SVs. With the additional samples, we merged Manta and SVABA SV calls as before and removed SVs with coordinates matching the SVs we removed due to batch effect concerns. A small percentage of germline SV calls represented translocations, which in practice are often false positives caused by misalignments of repeat expansions[39]. Therefore, we removed all germline translocation SV calls from the analyses.

Our previous study analyzed CBTN DNA methylation data by combining them with somatic SVs[33]. For CBTN DNA methylation data generated using the Illumina Infinium MethylationEPIC BeadChip array platform (Illumina, San Diego, CA), we obtained raw IDAT image files from the CBTN Cavatica site under the "methylation" folder. We processed the IDAT files using the minfi package in R Bioconductor[69], with quantile normalization using the Bioconductor preprocessQuantile function to generate the final methylation beta values. Illumina EPIC probe annotation was based on hg19 coordinates. When overlapping methylation array probe coordinates to SV calls based on hg38, we first used the UCSC Genome Browser LiftOver tool to convert the probe coordinates to hg38. Previously[33], we obtained processed RNA-seq data for CBTN tumors from the CBTN Cavatica site from the "CBTN" and "CBTN-X01" folders, from which we generated a batch-corrected dataset to correct for technical differences between the two cohorts.

Regarding *NBPF3*, we additionally obtained germline SV calls with breakpoints involving the region 100 kb of the gene in the CPTAC cohort of combined WGS, methylation, and expression[13]. Germline SV calls were made using both SVABA and Delly algorithms.

Integrative analyses between SVs and DNA methylation

In our present study, we used a similar overall approach for identifying associations between germline SV breakpoints and differential DNA methylation patterns as we had carried out previously between somatic SV breakpoints and DNA methylation[15,33]. Using SVExpress[70], we defined genes with altered DNA methylation involving their associated CGI probe in conjunction with nearby somatic SV breakpoints. For each of four genomic region windows 100 kb upstream of the gene, 100 kb downstream, within the gene body, or 1 Mb upstream or downstream (the latter using the distance metric method), we constructed gene-to-sample SV breakpoint matrices for related gene CGI with germline SV breakpoint patterns. Except for the 1 Mb region, SVExpress constructed the gene-to-sample matrix with entries as 1 if a breakpoint occurs in the specified region for the given gene in the given sample and 0 if otherwise. For the 1 Mb region, we used SVExpress's relative distance metric option[15], whereby breakpoints close to the gene have more numeric weight in identifying SV-expression associations, while breakpoints further away but within 1 Mb can have some influence. The 1 Mb window was considered as genomic elements such as enhancers may impact genes as far as ~1 Mb away[15]. The gene-to-sample SV breakpoint matrices included 21642 unique genes with Entrez identifier, with gene coordinates obtained from Ensembl BioMart (GRCh38.p14)[71].

CGI-level analyses focused on the 133,345 array probes annotated as falling within CGIs. The genes associated with each CGI probe, according to the Illumina platform annotation, were used to construct CGI probe-to-sample breakpoint matrices, using the gene as the reference for the relative breakpoint locations and the gene-to-sample SV breakpoint matrices described above. SVExpress assessed the association between CGI methylation and the presence of an SV breakpoint using linear regression models. We used logit-transformed methylation array beta values (i.e., "M-values"[72]) in the linear regression modeling for DNA methylation data to help the data better align with linear model assumptions[73]. For CGI probes with significant SV-methylation associations considered in downstream analyses, we required significant associations to arise after correction for tumor histologic type. In addition, CGI probes located on the X or Y chromosome needed to remain significant after correction for patient sex in addition to sample histologic type. For the analyses involving within-gene, 100 kb upstream, and 100 kb downstream gene regions, we only considered CGI probes with at least three patients associated with an SV within the given region when estimating FDR by Storey and Tibshirani method[74]. The same germline SV-CGI methylation association analyses were carried out for the cohort of 633 TCGA patients, involving the 114617 CGI probes represented on the Illumina 450 K array platform[15]. Like for the CGI methylation analyses and as previously described for the PCAWG adult cancer dataset[18], we generated germline SV-expression associations for within-gene, 100 kb upstream, 100 kb downstream, and 1 Mb gene centric region windows. We compared the RNA-seq-based SV-expression association results from the 1430-patient cohort with the DNA methylation-based results generated in this present study. When relating CGI methylation results with gene expression results, we relied on the Illumina array annotation for the gene associated with the given CGI probe; in instances where a CGI was annotated as possibly involving multiple genes, we selected one gene (the first one listed) to assign to the CGI probe, to simply the analyses and to minimize multiple testing issues. For about 9.4% of CGI

probes, the probe annotation referred to an additional gene, and in an alternate analysis integrating gene expression and CGI methylation for the alternate gene, just ten CGI probes had expression and methylation inversely associated with SVs ($p < 0.01$ for each) for the same region in the CBTN cohort (Supplementary Data 3).

For each of 30292 enhancers, we mapped the nearest DNA methylation probe (within 20 kb) and determined its methylation association with nearby SV breakpoints (within 100 kb) by linear modeling correcting for histologic type. We utilized the enhancer annotations provided by Kumar et al.[53], using the UCSC Genome Browser LiftOver tool to convert enhancer coordinates from hg19 to hg38. We used SVExpress to construct an enhancer-to-sample matrix with entries as 1 if a breakpoint occurs within 100 kb of the given enhancer in the given sample and 0 if otherwise. For enhancers located on the X or Y chromosome, these needed to remain significant after correction for patient sex in addition to sample histologic type in the linear modeling. When estimating FDR[74], we only considered enhancers with at least three tumors associated with an SV within 100 kb, involving 23241 enhancers and 21246 Illumina methylation array probes.

Integrative analyses of SV breakpoints falling within CGIs and histone epigenetic marks

For each of four epigenetic elements—CGIs, H3K36me3, H3K9me3, and H3K27me3—we examined the percentages of SV-gene associations involving an SV breakpoint falling within the given element, both for all SV-gene associations and for the subsets involving altered CGI methylation. We obtained CGI coordinates from the UCSC Genome Browser (http://genome.ucsc.edu)[75]. We obtained genomic coordinates (hg38) for histone epigenetic marks H3K36me3, H3K9me3, and H3K27me3 by ChIP-seq for K562 cell line from ENCODE consortium datasets (www.encodeproject.org, narrowPeak calls)[8]. For analysis of histone epigenetic marks in TCGA datasets, the hg38 coordinates used in the CBTN analyses were converted to hg19 using the UCSC Genome Browser LiftOver tool. Using SVExpress, we first tabulated germline SV breakpoint-to-gene associations for all genes and samples represented in the combined CGI methylation and germline SV datasets, taking the closest SV within 1 Mb. For each patient-to-gene association, we then determined if that association involved higher or lower CGI methylation. Differential methylation was defined here as $p < 0.01$ by linear modeling (with covariates) across all patients (focusing on the genomic region window for which the gene CGI probe had the lowest SV-CGI methylation association p-value), with the individual patient having both a germline SV breakpoint falling within the most significant genomic region window and methylation beta levels in the tumor greater or less than the median across tumors in a direction consistent with the global association. Where multiple CGI probes referred to the same gene, the probe with the lowest p-value was selected to represent the gene in this analysis, independently of the direction of methylation change to not bias the integration results. For each patient, we also tabulated the set of germline SVs for which one breakpoint was within the boundaries of an epigenetic element of the given class. For all patient-level SV-gene associations represented by the combined datasets, we determined which had the nearest SV involving the gene falling within an element and which involved differential methylation (as described above).

Global patterns involving histone marks may be either cell-type specific or cell-type invariant[8]. For example, of the ~162 K H3K36me3 epigenetic marks identified by ENCODE for K562 leukemia cells, we found about 60% to overlap with H3K36me3 epigenetic marks identified for A549 lung cells. ENCODE data for a brain cell line was unavailable for our study. In addition, the CBTN datasets represent many different brain tumor histologic types, each involving a specific cell type of origin. However, in most instances, we would be uncertain about the tumor cell of origin, which likely would not have any parallel in the ENCODE datasets. Therefore, the overall patterns of histone epigenetic marks in relation to germline SV breakpoints, as reported in our present study, would primarily involve those marks that are cell-type invariant.

Survival analyses

We identified CGI-level correlates of patient survival associated with nearby germline SV breakpoints in the CBTN cohort, with 1317 patients having combined germline SV and overall survival data (Supplementary Data 1). We obtained CBTN patient survival data from the PedCBioportal on May 8, 2024. As different tumor types based on histology or tissue of origin may involve differences in patient survival over time, we utilized statistical models to correct for tumor type, whereby any associations of molecular features with survival would not be explainable by differences involving tumor type representation alone[13,14,76,77]. For associating nearby germline SV breakpoints with patient outcome, we utilized the gene X sample relative distance breakpoint matrix, generated by SVExpress, for a given genomic region in relation to the gene (within the gene boundary, 100 kb upstream of the gene, 100 kb downstream of the gene, or within 1 Mb of the gene start, with relative distances involving the 1 Mb region being weighted using the "relative distance metric" option[15]). For each gene, we used a stratified Cox (accounting for cancer type, using as.factor in R) to associate patient overall survival with the germline SV breakpoint patterns for that gene. Similarly, we identified CGI-level methylation correlates of patient overall survival using Cox correcting for tumor histologic type.

We utilized the gene X sample relative distance breakpoint matrix generated by SVExpress (1 Mb region) to associate nearby SV breakpoints with patient outcomes. For each associated gene, we used a stratified Cox (correcting for histologic type) to associate patient overall survival with the log2-transformed relative distance to the nearest breakpoint for that gene. For each region, FDR calculations were based on the respective set of genes with at least ten patients having a breakpoint in the given region, drawing from the entire set of 14,148 uniquely identified genes represented in the CGI methylation probe dataset (1 Mb region, 14,030 genes; within gene, 2388; 100 kb upstream, 4651; 100 kb downstream, 4678). We also associated mRNA expression of the gene with overall survival using stratified Cox (corrected for histologic type, using as.factor in R). Similarly, we identified CGI-level methylation correlates of patient overall survival using a stratified Cox (correcting for histologic type) on logit-transformed methylation beta values. Combining the germline SV-survival associations with the germline SV-CGI methylation associations, 1155 CGI probes had both a positive association between germline SV breakpoints near the gene and worse overall survival (one-sided $p < 0.05$, Cox analysis incorporating tumor type as a covariate, and at least ten patients with a breakpoint), and a positive or negative germline SV-expression association ($p < 0.01$, linear modeling with covariates), for the same genomic region window, considering any one of the four region windows. For this analysis, we required a positive association between SV breakpoint pattern and survival, as the absence of a germline SV increasing risk of a shorter time to adverse event may be more difficult to interpret.

We examined the association of the 1155-CGI probe signature with patient survival in the CBTN tumor methylation profile dataset (as DNA methylation associations with survival were not used to define the 1155-probe signature). The direction of each probe in the 1155-CGI signature, as applied to the CBTN methylation dataset, was based on the direction of the germline SV-methylation association. We scored patient profiles in the CBTN methylation dataset using our previously described "t score" metric[64,78], with the logit-transformed DNA methylation values centered across patients to standard deviations from the median. The t score represents the two-sided t statistic when comparing the average of the signature higher CGIs within each external differential methylation profile with the average of the signature lower CGIs. For example, the t

score for a given sample profile is high when the CGIs high and low in the signature are respectively high and low on average in the external sample profile. We assessed the association of the CGI signature score with patient outcome using Cox and log rank (dividing the cases according to low, high, or intermediate signature scoring, accounting for tumor histologic type).

Statistical analyses

All *p*-values were two-sided unless otherwise specified. Nominal p-values do not involve multiple comparison adjustments, while FDRs involve p-values adjusted for multiple gene feature comparisons. Heat map visualizations were performed using JavaTreeview (version 1.1.6r4)[79]. Figures indicate exact value of n (number of individuals), and the statistical tests used are noted in the Figure legends and next to reported p-values in the Results section. Boxplots represent 5% (lower whisker), 25% (lower box), 50% (median), 75% (upper box), and 95% (upper whisker). Figures represent biological and not technical replicates.

For each feature tested in the linear modeling, corresponding FDRs were computed using the Storey and Tibshirani method[74]: [nominal p-value of the feature] X [total number of features tested] / [number of features in the dataset with p-value less than or equal to the given p-value]. For a top feature set with FDR < 10%, the FDR would suggest that on the order of 10% of these features might have nominal significance due to multiple testing (and, conversely, some 90% of the top features would represent bona fide significant associations). The number of nominally significant features expected at a given *p*-value cutoff (e.g., $p < 0.01$) due to chance alone, as predicted by probability theory ([total number of features] X [nominal p-value cutoff]), was similarly found when shuffling the 1 Mb germline SV breakpoint and methylation datasets and carrying out the same analyses. For the 100 kb upstream, 100 kb downstream, and within-gene region windows, only features with at least three patients with SV breakpoints in the given region relative to the gene were considered, which also limited the numbers of false positives with respect to the FDR calculation.

We relied on a stricter FDR cutoff[74] for defining top methylation features when carrying out probe-level global associations for a single analysis (e.g., CGI-level or enhancer-level SV-methylation associations). When overlapping different top probe-level results sets (e.g., CGI-level or enhancer-level SV-methylation associations involving coordinately expressed genes or germline SV-survival associations involving SV-CGI methylation associations), we used a more relaxed p-value cutoff (e.g., $p < 0.01$) to limit false negatives, helping us identify significant overlap patterns. As reported, the degree of gene set overlap across independent analyses was often highly statistically significant. This practice would be consistent with our previous studies identifying genes with demonstrated functional roles as originally identified using integrative analyses[80,81] and standard analytical methods like GSEA[54,55] that can identify enrichment patterns that would be missed using overly strict FDR cutoffs. For example, when selecting genes based on both SV-methylation and SV-expression associations, using a nominal p-value of 0.01 for each, the number of intersecting features expected for a given region window would be: [total number of features tested] X [nominal *p*-value/2] X [nominal *p*-value/2] X [factor of 2, as we consider both mRNA higher/methylation lower and mRNA lower/methylation higher feature sets]. The number of actual intersecting features found for each region window (considering both mRNA higher/methylation lower and mRNA lower/methylation higher) were 68, 93, and 101 for 100 kb upstream, 100 kb downstream, and within-gene regions, respectively, where the chance expected by multiple testing were, respectively, 1.9, 3.7, and 3.7.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in this study are publicly available. CBTN molecular data (including all raw sequencing and DNA methylation data) are available through the public project on the Kids First Data Resource Portal and Cavatica (https://cbtn.org/) and through the PedCBioPortal (https://pedcbioportal.org/). Protected CBTN molecular data, including raw sequencing data, are available under restricted access. To gain access to protected data, one can apply to the CBTN and sign a Data Use Agreement. The data availability regarding CBTN data, as described above, would apply to our previous studies utilizing CBTN data[13,33,58]. The batch-corrected and harmonized CBTN RNA-seq dataset used in the study (correcting for differences between X01 and initial cohorts) is available via our previous publication[33]. TCGA molecular datasets are available from the Broad Institute Firehose pipeline (http://gdac.broadinstitute.org/) and raw sequencing data are available via the Genome Data Commons (GDC, https://portal.gdc.cancer.gov). The DGV SV catalogs are available via the associated web site (https://dgv.tcag.ca/dgv/app/home). Source data are provided as a Source Data file. Source data are provided with this paper.

References

1. GTEx_Consortium Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
2. Scott, A., Chiang, C. & Hall, I. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res* **31**, 2249–2257 (2021).
3. Zhang, J. & Zhao, H. eQTL studies: from bulk tissues to single cells. *J. Genet Genom*. **50**, 925–933 (2023).
4. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet*. **49**, 692–699 (2017).
5. Conrad, D. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
6. Alkan, C., Coe, B. & Eichler, E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet*. **12**, 363–376 (2011).
7. Flynn, E. & Lappalainen, T. Functional Characterization of Genetic Variant Effects on Expression. *Annu Rev. Biomed. Data Sci*. **5**, 119–139 (2022).
8. The_ENCODE_Project_Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
9. Jakubosky, D. et al. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat. Commun*. **11**, 2927 (2020).
10. Han, L. et al. Functional annotation of rare structural variation in the human brain. *Nat. Commun*. **11**, 2990 (2020).
11. Zhang, Y. et al. High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations. *Nat. Commun*. **11**, 736 (2020).
12. Pcawg_Transcriptome_Core_Group et al. Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).
13. Chen, F., Zhang, Y., Chandrashekar, D., Varambally, S. & Creighton, C. Global impact of somatic structural variation on the cancer proteome. *Nat. Commun*. **14**, 5637 (2023).
14. Zhang, Y. et al. Rearrangement-mediated cis-regulatory alterations in advanced patient tumors reveal interactions with therapy. *Cell Rep*. **37**, 110023 (2021).
15. Zhang, Y. et al. Global impact of somatic structural variation on the DNA methylome of human cancers. *Genome Biol*. **20**, 209 (2019).
16. Zhang, Y. et al. A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Rep*. **24**, 515–527 (2018).
17. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
18. Chen, F., Zhang, Y., Sedlazeck, F. & Creighton, C. Germline structural variation globally impacts the cancer transcriptome including disease-relevant genes. *Cell Rep. Med*. **5**, 101446 (2024).

19. Shapiro, J. et al. OpenPBTA: The Open Pediatric Brain Tumor Atlas. *Cell Genom.* **3**, 100340 (2023).
20. Lilly, J. et al. The children's brain tumor network (CBTN) - Accelerating research in pediatric central nervous system tumors through collaboration and open science. *Neoplasia* **35**, 100846 (2023).
21. Gröbner, S. et al. The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327 (2018).
22. Bhootra, S., Jill, N., Shanmugam, G., Rakshit, S. & Sarkar, K. DNA methylation and cancer: transcriptional regulation, prognostic, and therapeutic perspective. *Med. Oncol.* **40**, 71 (2023).
23. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 16–21 (2002). 16.
24. Héberlé, E. & Bardet, A. Sensitivity of transcription factors to DNA methylation. *Essays Biochem.* **63**, 727–741 (2019).
25. Lea, A. et al. Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife* **7**, e37513 (2018).
26. Tong, Y. et al. MICMIC: identification of DNA methylation of distal regulatory regions with causal effects on tumorigenesis. *Genome Biol.* **19**, 73 (2018).
27. Ankill, J. et al. Epigenetic alterations at distal enhancers are linked to proliferation in human breast cancer. *NAR Cancer* **4**, zcac008 (2022).
28. Hoadley, K. et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304 (2018).
29. Zhang, D. et al. Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* **86**, 411–419 (2010).
30. Villicaña, S. & Bell, J. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol.* **22**, 127 (2021).
31. Gunasekara, C. et al. Systemic interindividual epigenetic variation in humans is associated with transposable elements and under strong genetic control. *Genome Biol.* **24**, 2 (2023).
32. Shi, X. et al. Association of CNVs with methylation variation. *NPJ Genom. Med.* **5**, 41 (2020).
33. Chen, F., Zhang, Y., Shen, L. & Creighton, C. The DNA methylome of pediatric brain tumors appears shaped by structural variation and predicts survival. *Nat. Commun.* **15**, 6775 (2024).
34. MacDonald, J., Ziman, R., Yuen, R., Feuk, L. & Scherer, S. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
35. Collins, R. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
36. Dwarshuis, N. et al. The GIAB genomic stratifications resource for human reference genomes. *Nat. Commun.* **15**, 9029 (2024).
37. Lewis, K. & Tollefsbol, T. Regulation of the Telomerase Reverse Transcriptase Subunit through Epigenetic Mechanisms. *Front. Genet.* **7**, 83 (2016).
38. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
39. Sedlazeck, F. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
40. Kleinjan, D. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
41. Oliva, M. et al. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat. Genet.* **55**, 112–122 (2023).
42. Vandepoele, K., Van Roy, N., Staes, K., Speleman, F. & van Roy, F. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Mol. Biol. Evol.* **22**, 2265–2274 (2005).
43. Wassif, C. et al. Mutations in the human sterol delta7-reductase gene at 11q12-13 cause Smith-Lemli-Opitz syndrome. *Am. J. Hum. Genet.* **63**, 55–62 (1998).
44. Forbes, S. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
45. Shi, X., Young, S., Cai, K., Yang, J. & Morahan, G. Cancer susceptibility genes: update and systematic perspectives. *Innov. (Camb.)* **3**, 100277 (2022).
46. DeRycke, M. et al. Targeted sequencing of 36 known or putative colorectal cancer susceptibility genes. *Mol. Genet Genom. Med.* **5**, 553–569 (2017).
47. Zhang, S. et al. RPSA gene mutants associated with risk of colorectal cancer among the chinese population. *Asian Pac. J. Cancer Prev.* **14**, 7127–7131 (2013).
48. Zhen, D. et al. BRCA1, BRCA2, PALB2, and CDKN2A mutations in familial pancreatic cancer: a PACGENE study. *Genet Med.* **17**, 569–577 (2015).
49. Li, Y., Chen, X. & Lu, C. The interplay between DNA and histone methylation: molecular mechanisms and disease implications. *EMBO Rep.* **22**, e51803 (2021).
50. Sun, Z. et al. H3K36me3, message from chromatin to DNA damage repair. *Cell Biosci.* **10**, 9 (2020).
51. Padeken, J., Methot, S. & Gasser, S. Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance. *Nat. Rev. Mol. Cell Biol.* **23**, 623–640 (2022).
52. Ngollo, M. et al. Global analysis of H3K27me3 as an epigenetic marker in prostate cancer progression. *BMC Cancer* **17**, 261 (2017).
53. Kumar, S. et al. Passenger mutations in more than 2500 cancer genomes: Overall molecular functional impact and consequences. *Cell* **180**, 915–927 (2020).
54. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
55. Mootha, V. K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
56. Foss-Skiftesvik, J. et al. Genome-wide association study across pediatric central nervous system tumors implicates shared predisposition and points to 1q25.2 (PAPPA2) and 11p12 (LRRC4C) as novel candidate susceptibility loci. *Childs Nerv. Syst.* **37**, 819–830 (2021).
57. Mack, S. & Northcott, P. Genomic Analysis of Childhood Brain Tumors: Methods for Genome-Wide Discovery and Precision Medicine Become Mainstream. *J. Clin. Oncol.* **35**, 2346–2354 (2017).
58. Zhang, Y., Chen, F., Donehower, L., Scheurer, M. & Creighton, C. A pediatric brain tumor atlas of genes deregulated by somatic genomic rearrangement. *Nat. Commun.* **12**, 937 (2021).
59. Capellini, A., Williams, M., Onel, K. & Huang, K. The Functional Hallmarks of Cancer Predisposition Genes. *Cancer Manag Res.* **13**, 4351–4357 (2021).
60. Sud, A., Kinnersley, B. & Houlston, R. Genome-wide association studies of cancer: current insights and future perspectives. *Nat. Rev. Cancer* **17**, 692–704 (2017).
61. Thavaneswaran, S. et al. Therapeutic implications of germline genetic findings in cancer. *Nat. Rev. Clin. Oncol.* **16**, 386–396 (2019).
62. Peifer, M. et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526**, 700–704 (2015).
63. Mahmoud, M. et al. Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
64. The_Cancer_Genome_Atlas_Research_Network Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
65. The_ICGC-TCGA_Pan-Cancer_Analysis_of_Whole_Genomes_Network Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

66. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
67. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
68. Wala, J. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
69. Aryee, M. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
70. Zhang, Y., Chen, F. & Creighton, C. SVExpress: identifying gene features altered recurrently in expression with nearby structural variant breakpoints. *BMC Bioinforma.* **22**, 135 (2021).
71. Kinsella, R. et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxf.)* **2011**, bar030 (2011).
72. Du, P. et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinforma.* **11**, 587 (2010).
73. Xie, C. et al. Differential methylation values in differential methylation analysis. *Bioinformatics* **35**, 1094–1097 (2019).
74. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
75. Raney, B. et al. The UCSC Genome Browser database: 2024 update. *Nucleic Acids Res.* **52**, D1082–D1088 (2024).
76. Zhang, Y. et al. A Pan-Cancer Proteogenomic Atlas of PI3K/AKT/mTOR Pathway Alterations. *Cancer Cell* **31**, 820–832 (2017).
77. Chen, F., Chandrashekar, D., Scheurer, M., Varambally, S. & Creighton, C. Global molecular alterations involving recurrence or progression of pediatric brain tumors. *Neoplasia* **24**, 22–33 (2022).
78. Cancer_Genome_Atlas_Research_Network Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
79. Saldanha, A. J. Java Treeview-extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
80. Grzeskowiak, C. et al. In vivo screening identifies GATAD2B as a metastasis driver in KRAS-driven lung cancer. *Nat. Commun.* **9**, 273 (2018).
81. Monsivais, D. et al. Mass-spectrometry-based proteomic correlates of grade and stage reveal pathways and kinases associated with aggressive human cancers. *Oncogene* **40**, 2081–2095 (2021).

Acknowledgements

This work was made possible through the resources and datasets made available by the Children's Brain Tumor Network (CBTN). We thank the patients and their families for their participation in the CBTN project. This work was supported by National Institutes of Health (NIH) grant P30CA125123 (C.J.C.).

Author contributions

Conceptualization: C.J.C.; Methodology: C.J.C., F.C., Y.Z.; Investigation: C.J.C., F.C., Y.Z., L.S.; W.L., F.S.; Formal Analysis: C.J.C., F.C., Y.Z.; Data Curation: C.J.C.; Visualization; C.J.C., F.C.; Writing: C.J.C.; Manuscript Review: F.C., Y.Z., L.S., W.L., F.S.; Supervision: C.J.C.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-60110-y.

Correspondence and requests for materials should be addressed to Chad J. Creighton.

Peer review information *Nature Communications* thanks Evelina Miele, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.