Expanded methylome and quantitative trait loci detection by long-read profiling of personal DNA

Cristian Groza, Bing Ge, Warren A. Cheung, et al.

Genome Res. 2025 35: 644-652 originally published online March 20, 2025
Access the most recent version at doi:10.1101/gr.279240.124

Expanded methylome and quantitative trait loci detection by long-read profiling of personal DNA

Cristian Groza,[1] Bing Ge,[2] Warren A. Cheung,[3] Tomi Pastinen,[3] and Guillaume Bourque[4,5,6]

[1]*Université de Montréal, Montréal Heart Institute, Montréal, Québec H1T 1C8, Canada;* [2]*McGill University, McGill University and Genome Quebec Innovation Centre, Montréal, Québec H3A 2T8, Canada;* [3]*Children's Mercy Hospital and Research Institute, Genomic Medicine Center, Kansas City, Missouri 64108, USA;* [4]*McGill University, Human Genetics, Montréal, Québec H3A 0C7, Canada;* [5]*Canadian Center for Computational Genomics, McGill University, Montréal, Québec H3A 2R7, Canada;* [6]*Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University, Montréal, Québec H3A 0G1, Canada*

Structural variants (SVs) are omnipresent in human DNA, yet their genotype and methylation statuses are rarely characterized due to previous limitations in genome assembly and detection of modified nucleotides. Also, the extent to which SVs act as methylation quantitative trait loci (SV-mQTLs) is largely unknown. Here, we generated a pangenome graph summarizing SVs in 782 de novo assemblies obtained from Genomic Answers for Kids, capturing 14.6 million CpG dinucleotides that are absent from the CHM13v2 reference (SV-CpGs), thus expanding their number by 43.6%. Using 435 methylomes, we genotyped 4.06 million SV-CpGs, of which 3.93 million (96.8%) are methylated at least once. Nonrepeat sequences contribute 1.59×10^6 novel SV-CpGs, followed by centromeric satellites (6.57×10^5), simple repeats (5.40×10^5), *Alu* elements (5.07×10^5), satellites (2.17×10^5), LINE-1s (1.83×10^5), and SVA (SINE-VNTR-*Alu*) elements (1.50×10^5). Centromeric satellites, simple repeats, and SVAs are overrepresented in SV-CpGs versus reference CpGs. Similarly, methylation levels in SV-CpGs are more variable than in reference CpGs. To explore if SVs are potentially causal for functional variation, we measured SV-mQTLs. This revealed over 230,464 methylation bins where the methylation is associated with common SVs within 100 kbp. Finally, we identified 65,659 methylation bins (28.5%) where the leading QTL variant is an SV. In conclusion, we demonstrate that graph pangenomes provide full SV structures, the associated methylation variation, and reveal tens of thousands of SV-mQTLs, underscoring the importance of assembly based analyses of human traits.

[Supplemental material is available for this article.]

The completion of the first telomere-to-telomere genome (Nurk et al. 2022) also enabled the first epigenomic characterization of a complete human genome (Gershman et al. 2022). This milestone epigenome characterized histone modifications and DNA methylation in previously unsolved and structurally polymorphic regions of the human genome, including centromeres, transposable elements, and tandem repeats. More generally, the DNA sequences omitted from current reference genomes are likely a source of substantial epigenetic activity. Expanding the nonreference results to a larger number of human genomes and epigenomes can expose population variation with potential new insights on trait variation and disease. The ability to survey epigenomes was recently augmented by long-read technologies that simultaneously characterize the sequence of personal genomes, resolving polymorphic structural variants (SVs), together with their epigenomic status (Yue et al. 2022; Cheung et al. 2023; Sigurpalsdottir et al. 2024).

Moreover, computational methods that can compile personalized genomes into pangenome graphs, can capture megabases of nonreference sequences and integrate SVs from a cohort of genomes (Li et al. 2020; Hickey et al. 2023; Garrison et al. 2024). The publication of the draft human pangenome reference also facilitates the study of SVs and their features at scale in a range of data sets (Liao et al. 2023; Groza et al. 2024). Indeed, such developments allow mapping epigenomic data directly to SVs and exploring the epigenetic status of regions that were not included in the reference genome (Groza et al. 2020, 2023). For example, it would be interesting to explore the link between SVs and 5mC base modifications, given the well-known connection between DNA methylation and gene expression (Razin and Cedar 1991; Breiling and Lyko 2015; Dhar et al. 2021).

A pangenome can support genotyping SVs across the same samples or a wider set of samples, which fits well within the assumptions of most methylation quantitative trait locus (QTL) studies. Thus, mapping methylation data to pangenomes to correct reference bias and recover more signals (Wulfridge et al. 2019) and then correlating the resulting methylation features with SVs is a promising approach that is enabled by pangenomes. Therefore, the tools necessary to answer long-standing questions regarding the epigenetic status of SVs (Daron and Slotkin 2017; Groza et al. 2023; Sun et al. 2023) and their associations with other quantitative traits are increasingly accessible.

Here, we use a pangenome comprising 782 haplotype-resolved de novo assemblies from the Genomic Answers for Kids (GA4K) Consortium (Cohen et al. 2022; Kane et al. 2023) and the 94 Human Pangenome Reference Consortium (HPRC) assemblies (Liao et al. 2023) to survey 435 5mC methylomes derived

Corresponding authors: tpastinen@cmh.edu, guil.bourque@mcgill.ca
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279240.124.

35:644–652 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/25; www.genome.org

from whole genome sequencing of blood using HiFi long-reads. With this pangenomic approach, we identify nonreference CpGs within SVs, characterize their population frequency and methylation status, and associate SV-QTLs with methylation variation over the entire genome.

Results

Pangenomes characterize the methylation status of CpGs in SVs

We expected each of the 782 GA4K de novo assemblies to contain a number of structurally variant and nonreference CpGs. These assemblies have a mean N50 of 19.1 Mbp (Supplemental Fig. S1), which is sufficient to resolve most SVs. To recover these sequences, we constructed a genome graph using minigraph (Li et al. 2020) starting with CHM13v2 (Nurk et al. 2022) as a backbone, followed by the 94 HPRC assemblies (Liao et al. 2023), before adding the 782 GA4K samples. In total, this added 14.6 million CpGs in alternative nodes (SV-CpGs, see Fig. 1A, for illustration) for GA4K, on average 16,600 SV-CpGs per sample, on top of the 33.5 million CpGs that exist in CHM13v2 (a gain of 43.6%). At the same time, the pangenome grew by 713 Mb (Supplemental Fig. S2, a gain of 23%), yielding 2.05×10^4 CpGs per megabase of nonreference sequence, compared to only 1.08×10^4 CpGs per megabase of reference sequence.

Next, we obtained and aligned 435 GA4K blood methylomes matching a subset of the samples in this pangenome, which were sequenced at a mean coverage of 26× (Supplemental Fig. S3) and show a mean read N50 of 13.6 kbp (Supplemental Fig. S4). These samples include some trios, where the parents were also sequenced with long-reads, albeit at a lower coverage. Then, we annotated each CpG in the pangenome with the methylation level found in these samples (Methods).

Initially, we found that these methylomes cover 7.99 million of the 14.6 million SV-CpGs that exist in the pangenome. However, a portion of these CpGs are supported by few reads due to low depth in some methylomes (Supplemental Fig. S3) and poor mappability of SV sequences. Therefore, we chose to focus on CpGs supported by at least 5 reads in a given individual (Sigurpalsdottir et al. 2024). After filtering, we counted 4.06 million SV-CpGs with methylation data, of which 3.93 million (96.8%) show a methylation level equal or above 50% in at least one of the 435 methylomes (Fig. 1B). For comparison, 30.5 of the 32.1 million covered reference CpGs (95.2%) (Fig. 1B) show a methylation level above 50% in the same samples (reference 5mCpGs). Indeed, a similar proportion of SV and reference CpGs are methylated at least once across the frequency spectrum (Supplemental Fig. S5). To further validate our ability to measure methylation levels in the pangenome, we successfully recapitulated a majority of the methylation differences between haplotypes (Supplemental Fig. S6) that are known to exist at human imprinting control regions (Jima et al. 2022).

Saturation analysis over these methylomes shows that additional new methylomes are expected to contribute 2050 SV-5mCpGs and 2750 SV-CpGs (Fig. 1C). Similarly, we expect to discover an additional 594 reference CpG and 1210 reference 5mCpGs in further additional genomes. We continue to discover reference CpGs at a low rate because some reference alleles are rare and will only be observed after sampling many methylomes (Supplemental Fig. S7). Statistical testing suggests that SV-CpGs

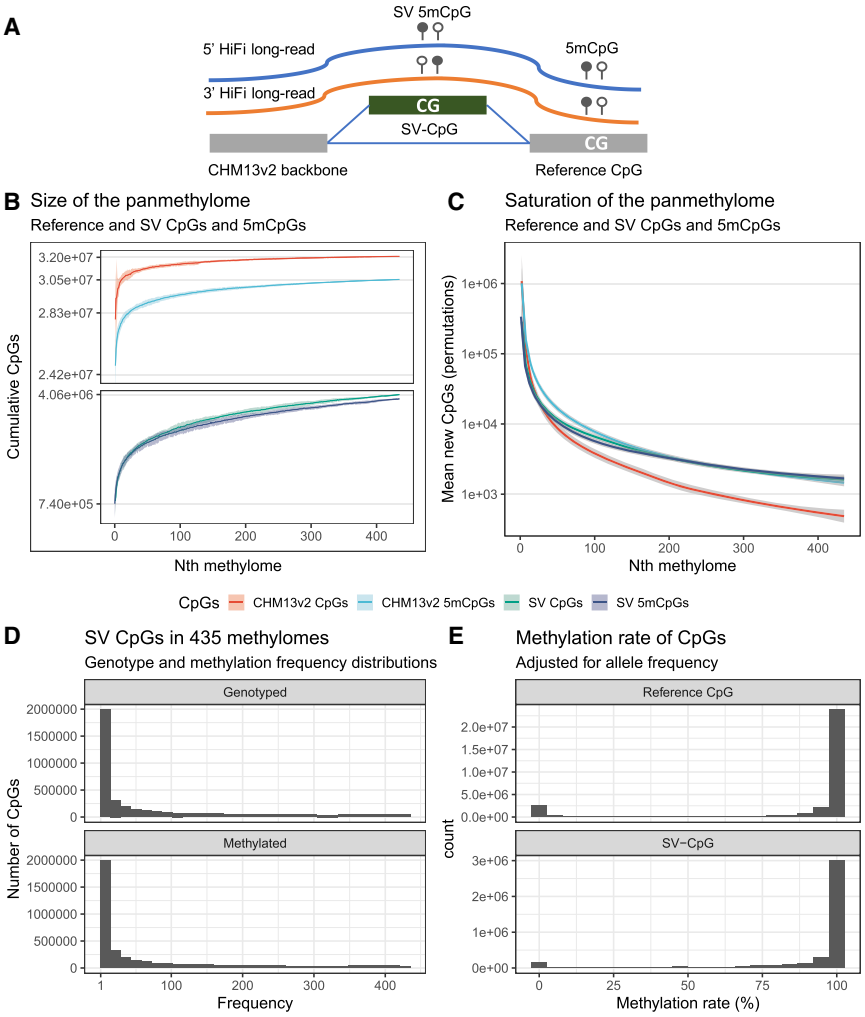

Figure 1. Illustration, number, frequency, and methylation of nonreference CpGs. (*A*) Illustration depicting the CHM13v2 backbone (gray) of the GA4K pangenome with a reference CpG, and an insertion (green) containing an SV-CpG that is not present in the reference. Also illustrated are HiFi long-reads featuring CpGs in a methylated state that align in this region to the positive and negative strand. (*B*) The cumulative number of reference CpGs, 5mCpGs, SV-CpGs, and SV-5mCpGs in the 435 methylomes. (*C*) The rate of change in the saturation of CpGs is shown in *B*. (*D*) Frequency distribution of SV-CpGs (red) and SV-5mCpGs (blue). (*E*) Observed methylation rates across SV- and reference CpGs, adjusted for allele frequency by counting only samples that carry any given CpG.

are discovered at a higher rate than reference CpGs and that SV-5mCpGs are discovered at a similar rate to SV-CpGs (Supplemental Table S1). This result is in line with our expectation from population genetics, since SV-CpGs are a subset of all CpGs in a population and most mammalian CpGs are highly methylated (Hattori et al. 2004).

A large fraction of nonreference CpGs are methylated

Knowing the size of the above panmethylome, we asked what was the frequency distribution of SV-CpGs and SV-5mCpGs. We genotype 7.03×10^5 singleton SV-CpGs, 2.51×10^6 with a genotype frequency below 10% and 5.72×10^4 with a genotype frequency above 90% (Fig. 1D). Some of these were methylated and we counted 7.02×10^5 singleton SV-5mCpGs, 2.64×10^6 with a methylation frequency below 10% and 4.00×10^4 with a frequency above 90% (Fig. 1D). However, many 5mCpGs were rare because they lie on rare alleles. Therefore, we calculated the methylation rate, where we adjusted for allele frequency and only considered samples that carry the CpG (Fig. 1E; Methods). After we calculated methylation rates, we observed 1.31×10^5 SV-CpGs (3.21%) and 1.53×10^6 reference CpGs (4.78%) that were never methylated in any methylome and have a methylation rate of 0% (Fig. 1E). Also, the average methylation rate was 90.4% for SV-CpGs and 86.5% for reference CpGs (Supplemental Fig. S8). The higher methylation rate in SV-CpGs may be due to enrichment in repeat elements that usually contribute to structural variation and are targets of methylation. Lastly, we counted dynamic CpGs that have a methylation rate between 15% and 85%, yielding 4.26×10^5 (10.5%) SV-CpGs and 2.20×10^6 (6.88%) reference CpGs (Fig. 1E).

Using the graph approach, we were able to view these methylation patterns in many haplotypes, including SVs, across hundreds of samples in polymorphic regions like the *HLA* (Fig. 2) and the *KIR* locus (Supplemental Fig. S9). In these representations, we clearly see patterns of methylated and unmethylated CpGs within nonreference sequences in the *HLA* and *KIR* loci, a task that was not possible with reference genomes that describe only one haplotype. For instance, multiple insertions, including a full LINE-1 insertion, of various sizes lie in the *HLA-DQA1* and *HLA-DQB1* loci and harbor CpGs in highly methylated states, as we traverse the graph from the 5′ to the 3′ end (Fig. 2).

Repeats account for more than half of the nonreference methylome

To determine the source of SV-CpGs and SV-5mCpGs, we ran RepeatMasker on the pangenome and tallied the number of CpGs that overlap repeats. We found that 1.59×10^6 SV-CpGs did not overlap any repeats (39.0%), of which 1.51×10^6 (95.2%) were methylated at least once (Fig. 3A). Second, centromeric satellites accounted for 6.57×10^5 of SV-CpGs (16.2%), of which 6.44×10^5 (97.9%) were methylated. Mobile elements also contributed, with 5.07×10^5 SV-CpGs (12.4%) in *Alu* elements (5.02×10^5 methylated, 99.2%), 1.83×10^5 (4.50%) in LINE-1s (1.80×10^5 methylated, 98.2%), and 1.50×10^5 (3.70%) in SVA (SINE-VNTR-*Alu*) elements (nearly all methylated, 99.7%) (Fig. 3A).

Then, we asked if any particular repeats contribute disproportionately to SV-5mCpGs relative to reference 5mCpGs. Here, we found that sequences that were not repeats were depleted in SV-5mCpGs, accounting for 44.2% of reference 5mCpGs but only for 38.4% of SV-5mCpGs. Similarly, *Alu*s contribute 23.8% of reference 5mCpGs but only 12.7% of SV-5mCpGs. LINE-1s, endogenous retrovirus elements (ERVs), and other nonreference repeats were also depleted (Fig. 3B). On the other hand, nonreference multiallelic sequences such as satellites (0.768% vs. 5.31%), centromeric satellites (2.63% vs. 16.3%), simple repeats (2.54% vs. 13.3%), and SVAs (0.514% vs. 3.81%, known to contain tandem repeats) were overrepresented in SV-5mCpGs (Fig. 3B).

Indeed, when we computed the frequency distribution of SV-5mCpGs and stratified by repeat category (Fig. 3C), we noted that SV-5mCpGs in overrepresented repeats like satellites, simple repeats, and SVAs were rarer than those in underrepresented repeats, which was consistent with multiallelic SVs contributing a large number of rare alleles to the pangenome.

Finally, we wanted to know if the methylation rate of SV-CpGs varies across repeat categories and if it differs from reference CpGs. For this purpose, we calculated the methylation rate of every CpG in the pangenome and plotted the resulting distributions stratified by repeats, separating SV-CpGs from reference CpGs (Fig. 3D; Supplemental Fig. S10; Methods). In these analyses, we observed the expected high methylation rates across all families, with the median methylation rate always exceeding 95% in both reference and SV-CpGs. We note that SV-CpGs that are not derived from repeats show higher mean methylation rates than similarly annotated reference CpGs (87.5% vs. 78.1%) (Fig. 3D). Meanwhile, SV-CpGs derived from repeats such as *Alu*s, LINE-1s, and SVAs show lower mean methylation rates than reference CpGs. Mann–Whitney *U* tests indicate that these differences are statistically significant at a false discovery rate (FDR) < 0.05, except for the smallest families with small sample sizes (Supplemental Table S2).

Pangenomes enable the mapping of SV-QTLs

We were interested to see if any of the SVs contained in the GA4K pangenome were QTLs for DNA methylation in our methylomes. To this end, we aligned and genotyped against the pangenome 470 haplotype-resolved de novo assemblies that featured 235 matching methylomes from the same probands. These assemblies with matching methylomes are a subset of the 782 assemblies that compose the pangenome. In these 470 assemblies, we found 373,138 nonreference SV alleles, many exceeding the lengths that could be captured within short reads (Supplemental Fig. S11). For QTL mapping, we selected SV alleles that had a minimum frequency of 5% and a maximum frequency of 95% and resulted in a total of 160,064 SV alleles. The SV alleles were distributed in 97,746 loci, highlighting the ability of pangenomes to characterize multiallelic regions. We also partitioned the CHM13v2 backbone reference into nonoverlapping 200 bp methylation bins and calculated the average methylation level in these bins (Methods). Then, we identified SV alleles and methylation bins within 100 kbp flanking sequence and performed 124.9×10^6 SVs act as methylation quantitative trait loci (SV-mQTL) tests (Supplemental Fig. S12). We detected 230,464 methylation bins that were in QTL with 76,677 SV alleles in 59,872 loci at FDR < 0.05 (Fig. 4A; Methods). For example, we identified a 2216 bp deletion of a proximal enhancer that increases methylation in its vicinity (Fig. 4B,C).

Next, we queried the distance distribution between the methylation bins and the associated SV-mQTLs to determine the ranges of interaction between SVs and methylation bins (Fig. 4D). We tallied 4144 methylation bins that fully overlap with their SV-mQTL, meaning the bin was within a structurally variant region of the backbone reference genome, and another 53,460 bins within 10 kbp of their SV-mQTL. Moreover, we found 14,970 methylation bins that were more than 90 kbp away from their SV-mQTL, at the limit of the allowed flanking distance. Overall, we observed a

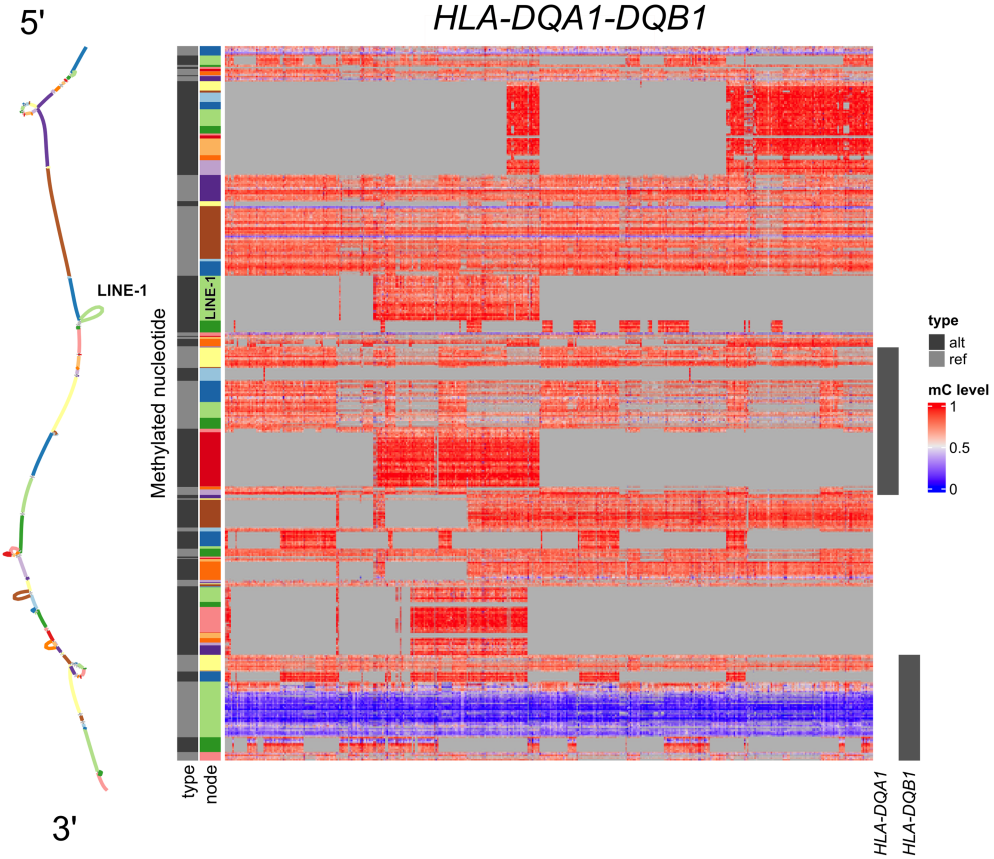

Figure 2. Heatmap visualization of methylation patterns of CpGs (rows) in the *HLA-DQA1-DQB1* locus across 435 methylomes (columns) in the GA4K pangenome. CpGs are ordered *top* to *bottom*, in the 5′ to 3′ direction as they appear in a haplotype and on the corresponding subgraph on the *left*. CpGs are also annotated by the node in the graph (the node row annotation) and whether it is a reference or nonreference CpG (the type row annotation, alt or ref). The *right* annotation shows the genes that overlap the bubbles in which the CpGs lie. Light gray cells in the heatmap are CpGs that are not genotyped in that methylome. Also indicated is the location of the full LINE-1 insertion in the graph and on the heatmap.

mean distance of interaction of 37.8 kbp (median 31.9 kbp). When using the much more stringent Bonferroni correction for multiple testing, 21,379 methylation bins pass statistical significance (adjusted *P*-value < 0.05) (Supplemental Fig. S13).

To describe the direction of mQTL effects for SVs, i.e., increase or decrease methylation, we tallied the signs of the strongest effect on each bin across chromosomes and found that SV-mQTLs showed more positive effects than negative effects: 73,702 bins were hypomethylated by SVs and 156,762 were hypermethylated by the strongest SV-mQTL (Fig. 4E). Every chromosome showed more hypermethylating than hypomethylating effects, with some chromosomes contributing slightly more QTLs relative to their size. The abundance of hypomethylating effects could occur in several ways. In some cases, as in Figure 4B, the SV allele disrupts an enhancer sequence (Fig. 4C) that would otherwise potentially upregulate and demethylate the nearby regions. In other cases, the SV allele may be a repeat or a transposon that is targeted for methylation and thus also increases nearby methylation levels, among other mechanisms.

Furthermore, we found that some SV-mQTL alleles are in proximity of dosage-sensitive regions in ClinGen (Rehm et al. 2015). In total, we found 13,460 SV-mQTL alleles with positive or negative effects on nearby methylation overlapping a ClinGen dosage-sensitive region (Supplemental Fig. S14). We show four such SV-mQTLs in Supplemental Figure S15, where the methylation bins are in the *ABR*, *BRAF*, *GPC6*, and *MYT1L* dosage-sensitive genes. These findings suggest that structural variation contributes to hyper- or hypomethylation of DNA near dosage-sensitive genes. Among the overlapped dosage-sensitive genes are 10 genes previously associated with candidate diagnostic SVs in Groza et al. (2024), namely, *WWOX*, *FANCA*, *ACOX1*, *CELF4*, *ALMS1*, *ABCB11*, *A4GALT*, *KMT2E*, *B4GALT1*, and *GALT*. Therefore, mapping SV-mQTLs with pangenomes could improve variant prioritization in clinical diagnosis.

We also wanted to know the methylation state of the SV-QTL alleles since it could be related to their QTL activity. Among the 55,149 SV alleles represented by paths containing at least one node in the pangenome graph, 22,125 alleles did not contain CpGs. In the remaining 33,024 SV alleles with CpGs, the average methylation rate was high, with 28,684 SVs having an average methylation rate above 85% (Supplemental Fig. S16). Moreover, reference and nonreference SV alleles have similar average methylation rate distributions.

Lastly, the pangenome merges SV alleles into multiallelic loci and allows the detection of allelic-specific mQTLs. This enabled us to rank the effect sizes of individual SV alleles in a multiallelic locus on nearby methylation bins (Supplemental Fig. S17). In particular, out of the 230,464 total methylation bins involved in mQTLs, 185,637 bins are associated with only a subset of SV alleles within the same bubble.

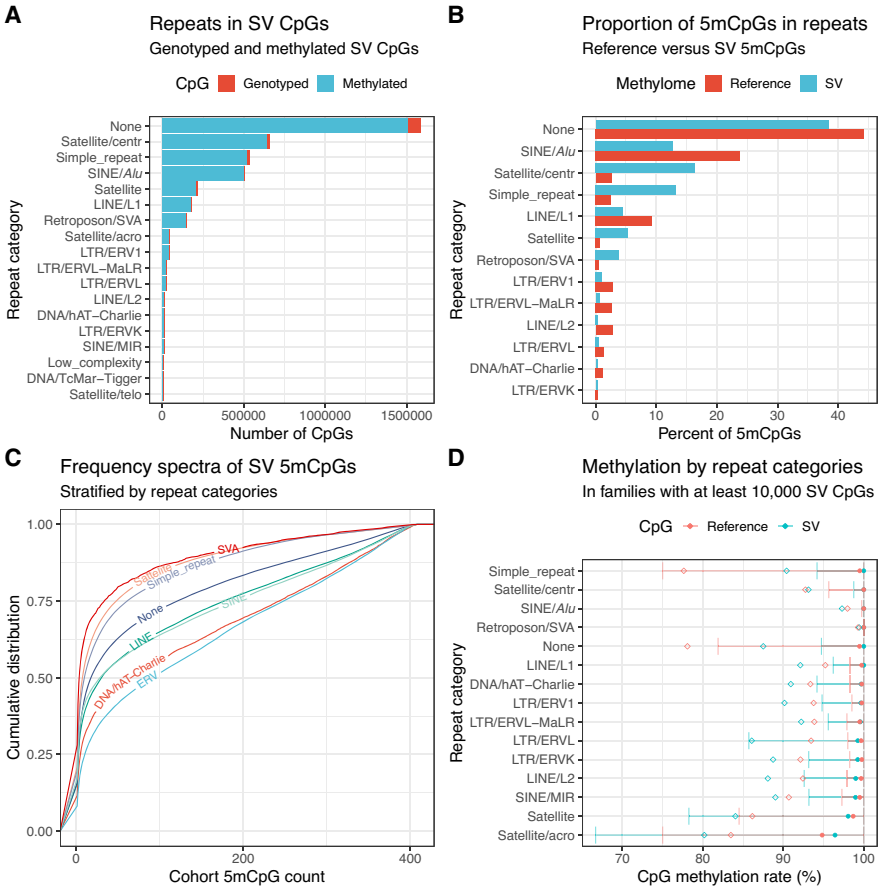

Figure 3. Annotation of sequences that contribute SV-CpGs. (*A*) Number of SV-CpGs and SV-5mCpGs contributed by sequences without repeats (None) and sequences with repeats. (*B*) Proportion of methylated CpGs by repeat category, contrasting reference CpGs and SV-CpGs. (*C*) Frequency distributions of SV-5mCpGs, stratified by repeat category. (*D*) The observed methylation rates of CpGs, as calculated in Figure 1D, are stratified by repeat category. Intervals denote the 25%, 50% (median, dot), and the 75% quantiles. The rhombi denote the means of the distributions.

Some SVs are stronger methylation QTLs than SNPs

SVs are thought to be enriched in QTLs and have higher effect sizes (Jakubosky et al. 2020). To explore this hypothesis in the GA4K methylomes, we mapped SNP-mQTLs with the same parameters and frequency constraints as SV-mQTLs (Methods). In total, we tested 5,617,307 SNPs for associations with the same methylation bins and found 156,047 SNP-mQTL associated with 178,709 methylation bins at FDR < 0.05 (Supplemental Table S3). Despite SNP-mQTLs being more numerous than SV-mQTLs, individual SVs were an order of magnitude more likely (17.2×) to be associated with methylation: only 2.78% of tested SNPs were mQTLs, in contrast to 47.9% of tested SVs. Then, we checked how often SVs were the leading variants over SNPs in mQTLs. For 65,659 methylation bins, the leading variant was an SV with a larger absolute effect size than any SNP (Fig. 5A; Supplemental Fig. S18, example in Fig. 4B,C). Conversely, 145,453 SNPs were the top variant in 166,317 methylation bins. In terms of variants, 32,947 SV alleles out of 160,064 (20.6%) were the leading variant, compared to only 2.59% for SNPs, or a 7.95-fold enrichment (Supplemental Table S1). Moreover, we found that SV-QTLs interact over longer distances (mean 37.8 kbp) than SNP-QTLs (mean 19.8 kbp) with methylation bins (Supplemental Fig. S19) and that each SV-QTL affects more methylation bins (mean 2.52 bins) than each SNP-QTL (mean 1.15 bins) (Supplemental Fig. S20). For example, the SV-QTL in Figure 4B and C affects methylation over 1 kbp of nearby sequence.

SV-QTLs are enriched in common SVs

When we looked at methylation bins where the leading variant was an SV, we noted that 57,692 bins were QTL with reference alleles having a predominantly positive effect on methylation (Fig. 5B, left). Another 19,329 bins were QTL with nonreference SV alleles, where positive and negative effects were evenly distributed. Next, we looked at methylation bins that were QTL with SVs but the leading variant was a single nucleotide polymorphism (SNP). Here, we found 61,669 methylation bins were QTL with nonreference SV alleles and 54,928 were QTL with reference SV alleles (Fig. 5B, right). Again, reference SV alleles show more positive effects on methylation. Thus, reference SV alleles tend to increase methylation and affect wider regions. More precisely, reference SV alleles were QTL with 2.14 bins on average, while nonreference SVs were QTL with only 1.70 methylation bins.

We plotted the frequency spectra of the SVs that were QTL with the above bins and found that leading reference SV alleles were very common, while leading nonreference SV alleles were the rarest (Fig. 5C). To a lesser extent, the same pattern occurs with reference and nonreference SV alleles that were not leading variants (Fig. 5C). This pattern is likely created by the underlying frequency distribution of reference alleles, which is skewed toward common frequencies and the underlying frequency distribution of nonreference SV alleles, which is skewed toward rare frequencies.

We also explored if the ranges of interaction within QTLs were different between reference and nonreference SV alleles, and between SV alleles that were leading variants and those that were surpassed by SNPs. Here, we found that reference SV alleles were QTL with methylation bins over longer distances than nonreference SV alleles (Fig. 5D). Similarly, leading SV alleles were QTL over longer distances than SVs surpassed by SNPs. Lastly, repeat annotation of SV-QTL alleles again revealed that multiallelic repeats were enriched in nonreference relative to reference SV alleles (Supplemental Fig. S21).

These differences in allele frequency and range of interaction suggest that the more frequent SV alleles interact with methylation more often and over longer distances than younger and less frequent SVs. At the same time, some rare SV alleles showed effects on methylation that were stronger than any nearby SNP.

Leading SV-mQTLs are in proximity to GWAS SNPs

To highlight SVs that may be causal for methylation, we first identified single nucleotide variants (SNVs) that were in high linkage

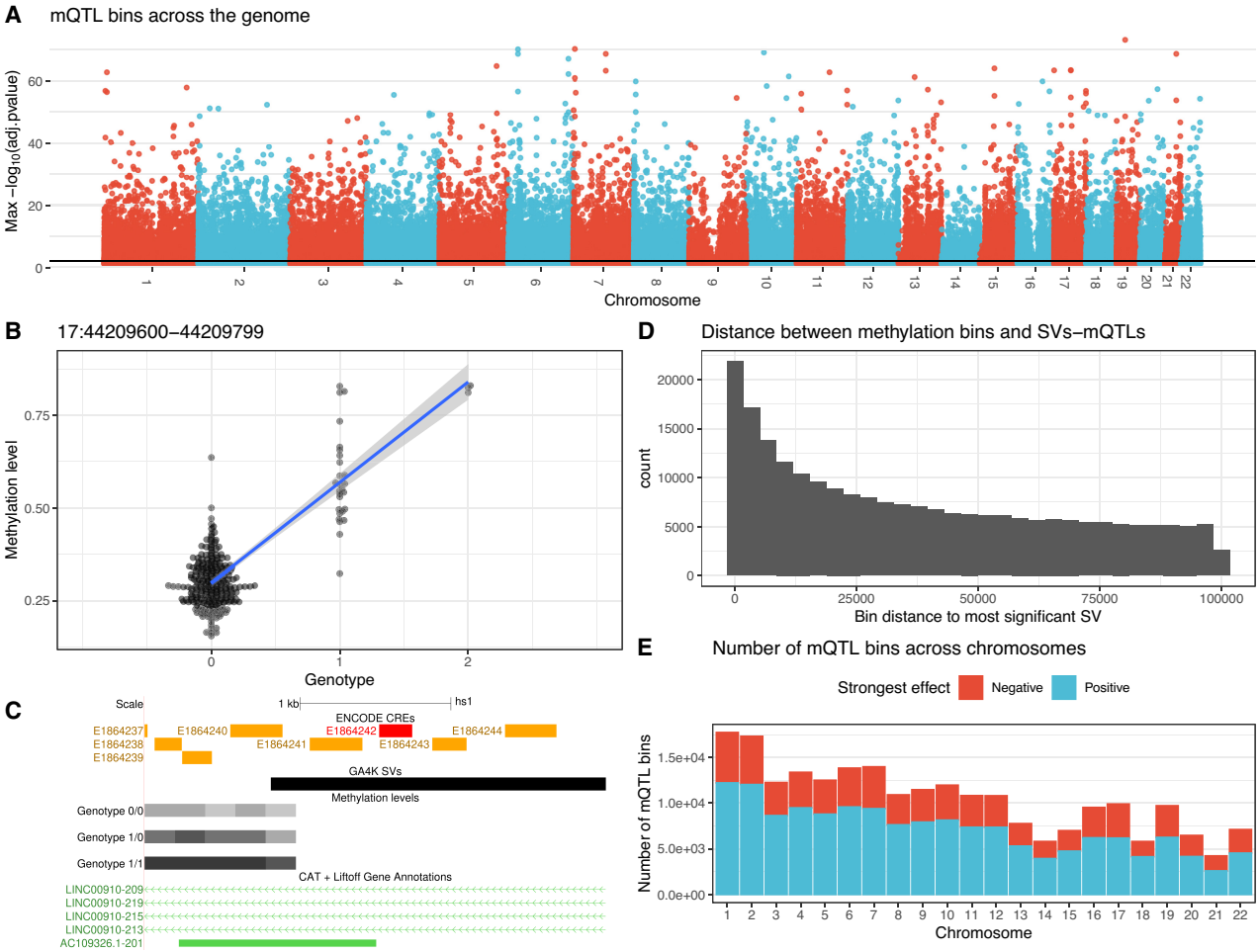

Figure 4. Quantitative trait locus mapping of 5mCpGs averaged in 200 bp methylation bins. (*A*) Manhattan plot of QTL methylation bins associated with an SV at FDR < 0.05 over the entire reference genome backbone. (*B*) Example of a leading SV-QTL interacting with nearby methylation levels. The nearest methylation bin is 34 bp away from the SV. (*C*) The SV allele (GA4K SVs track) is a 2216 bp deletion of CHM13v2.0#Chr17:44,210,434–44,212,650 and overlaps proximal enhancer-like signatures in ENCODE SCREEN (https://screen.encodeproject.org/). Also shown are the mean methylation levels (grayscale) of regions in QTL with the SV, stratified by SV genotype. (*D*) Distribution of distances between SV-mQTLs and their methylation bins. (*E*) Number of methylation bins in QTL with an SV in the GA4K pangenome across chromosomes, stratified by the positive and negative effect of the SV on methylation.

disequilibrium ($R^2 > 0.95$) with previously known SNPs in the NHGRI-EBI Genome Wide Association Studies (GWAS) Catalog (Sollis et al. 2023). Then, we filtered for SVs showing higher effect sizes than SNVs, that were closer to the methylation bin than the leading SNV and also no further than 10 kbp. In doing so, we obtained a list of 606 SVs that were QTL with 2417 methylation bins (Supplemental Table S4). That is, each putatively causal SV was in QTL on average with 3.99 methylation bins, which was 58% more than the average of SV-QTLs. Moreover, we annotated these SVs with 671 genes that were associated with GWAS SNPs in LD with GA4K SNVs (Supplemental Table S4).

Discussion

On average, we observed 9300 new CpGs in the DNA sequences added by each genome to the GA4K pangenome (0.81 Mbp per genome). Moreover, we were able to characterize the repeat families and other sequences that contribute to new CpGs and showed that at least 96.8% of these new CpGs were methylated in at least one individual. These new CpGs are in SVs that are enriched in short and variable number tandem repeats, which could impact their mappability. Aided by the GA4K pangenome graph, we arranged and sorted this nonreference epigenetic variation in haplotypes that could be compared across many methylomes, allowing for the characterization of complex patterns of methylations within SVs. Moreover, saturation analysis suggests that expanding this pan-methylome with more assemblies and methylomes would continue to add thousands of new SV-CpGs and SV-5mCpGs per sample.

A minority of CpGs are known to be variable, or dynamic, in tissues (Ziller et al. 2013). Similarly, we found that most CpGs in personal reference DNA show predominant hypermethylation, and only 6.88% were variably methylated (15%–85% methylated) in the 435 methylomes. In contrast, 10.5% of nonreference CpGs found in nonreference sequences of the pangenome were variably methylated, showing a larger contribution to epigenetic variation in humans. Moreover, our demonstration of methylation levels in human imprinting control regions (ICRs) indicates that pangenomes could be used to search for imprinted regions in SVs. A future analysis could attempt to find more CpGs with variable methylation in more biological contexts.

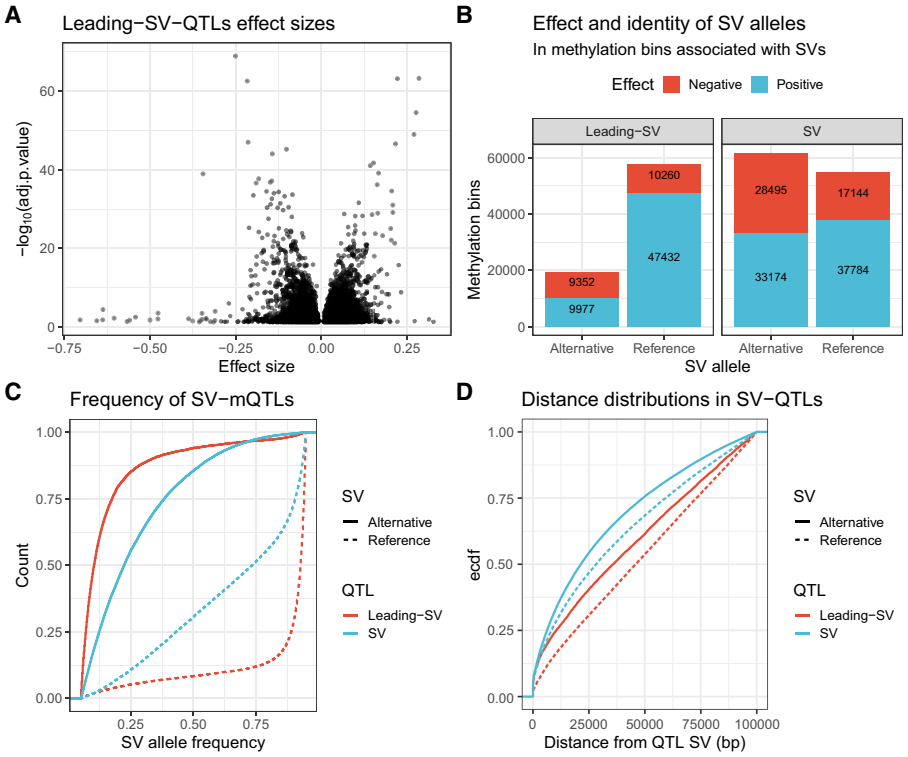

Figure 5. Leading mQTL variants among the pangenome SVs. (*A*) Volcano plots of mQTLs where the leading variant is an SV. (*B*) Number of methylation bins in mQTL with an SV, stratified by reference or nonreference SV allele, and positive or negative effect on methylation. (*C*) Allele frequency distribution of mQTL-SVs, stratified by leading versus nonleading variant, and reference and nonreference SV alleles. (*D*) The distribution of distances between SV-mQTLs and their methylation bins, stratified as in *C*.

We also used the pangenome graph to explore population variation in functional DNA, linking nearly 60,000 SV loci with methylation variation in over 46.1 Mbp of DNA across population assemblies. Parallel analyses of SNV and SV variation in the same samples demonstrated the larger prevalence of methylation among fewer SVs, their impacts extending greater distances, and ability to explain a substantial proportion of mQTLs. Next, we identified two sets of leading SV-QTLs that surpass SNVs in effect size. First, we found a set of leading SV alleles that are common in the population. These tend to be positively associated with DNA methylation and are often included in the reference genome. Second, we found another set of rarer SV alleles that are associated equally with positive and negative effects on DNA methylation. We note that these comparisons did not include methylation bins that lie in new nonreference sequences not mappable by standard mQTL-SNV associations. To include polymorphic methylation bins, statistical models that account for genotype must be developed.

In conclusion, our observations underscore the importance of assessing genomes for the entirety of sequence space not only for structural but also for functional variation (Groza et al. 2023). We also confirm the properties of SVs linking to QTLs impacting at greater distances and at a higher frequency. Hypermethylation in regulatory elements canonically leads to loss of activity, which we previously exploited in the rare variant characterization of long-read sequences (Cheung et al. 2023). The hypermethylating impact we observed for leading SV-mQTLs suggests an important role for studying SVs in gene silencing. Overall, the full-scope structural variation cataloged in pangenome graphs suggests large utility in quantitative trait and disease genetic studies.

Methods

Creating the pangenome graph

As in Groza et al. (2024), the probands were sequenced with PacBio HiFi, and the genomes were assembled with hifiasm v0.15 in the graph trio binning mode. Parental *k*-mers were counted from parental sequencing using yak count v0.01. We created the GA4K pangenome using `minigraph -xcggs –ggen` (version 0.20-r559) and only keeping structural variation above 50 bp. We started with the CHM13v2 backbone reference, added 94 HPRC genomes (Liao et al. 2023), and finally 782 GA4K assemblies. For the purpose of visualizing methylation on haplotypes in 5′ to 3′ order in a heatmap, we topologically sorted the graph using `vg sort -a topo`.

Genotyping assemblies

To genotype SVs in the GA4K assemblies, we aligned them to the final pangenome and called variants using `minigraph -c –call –vc`. Then, we created a unified genotype matrix across genomes by considering each bubble start, end, source, sink, and path in the pangenome as alleles and then listing the presence of an allele in a genome as 1 and their presence as 0 (see `genotypes.R`). Reference alleles were determined by genotyping CHM13v2 against the pangenome graph.

Annotating CpGs in the pangenome with repeats

We ran RepeatMasker with the Dfam_2.0 (Hubley et al. 2016) database on all nodes in the graph to identify repeats. Then, we overlapped the repeat annotation of nodes with the position of CpGs on these nodes to label each CpG with a repeat annotation. We achieved this by converting the methylation annotation of nodes produced by `nodes_methylation.py` to BED format using `awk (awk -v OFS='\t' '{print $1, $2, $2+1, $3, $4}')` and then intersecting with the repeat annotation using the BEDTools (Quinlan and Hall 2010) command `bedtools intersect -loj`.

Mapping methylation data to pangenome graphs

We wrote `panmethyl` (https://github.com/cgroza/panmethyl) to map methylation data to pangenome graphs. To do so, `panmethyl` takes BAMs processed and annotated with MM and ML tags by PacBio Jasmine. The MM tag describes the position of cytosines in a read and the ML tag describes their methylation probabilities. Panmethyl processes these and extracts HiFi long-reads that are annotated with methylation probabilities at each cytosine. Then, the reads are aligned to the pangenome with `minigraph --vc -c -N 1`, and the methylation probabilities of every cytosine in the read are lifted to their corresponding positions in the pangenome graph. To calculate the per sample methylation level of cytosines in the pangenome, we averaged the

methylation probabilities that map to their respective position. We also indexed every CpG in the pangenome and named CpG that are not in CHM13v2 as SV-CpGs. Using this index, we calculated the population frequency of reference CpGs and SV-CpGs by counting the number of methylomes that cover the CpG in the aligned reads. To calculate the population frequency of reference 5mCpGs and SV-5mCpGs, we counted the number of methylomes where the CpG has an average methylation level across strands of 50% or more (see `merge_cpg_genotypes.py` and `merge_cpgs_methylated.py`) and is covered by at least 5 reads on any strand (Sigurpalsdottir et al. 2024). To calculate the methylation level of a CpG, we average the methylation level of the cytosines on the positive and negative strands. We intersected CpGs with repeats to obtain the population frequency of CpGs and 5mCpGs stratified by repeats.

To check methylation differences at imprinted control regions, we haplotaged the aligned BAMs with WhatsHap (Martin et al. 2023), separated the reads into HP1 and HP2 using the SAMtools (Danecek et al. 2021) commands `samtools view -e [HP]==1` and `samtools view -e [HP]==2`, and ran panmethyl separately for HP1 and HP2 reads in order to obtain phased methylation signals. Then, we retrieved the regions of known human ICRs from https://humanicr.org/ and lifted the coordinates to CHM13v2. In each ICR, for each sample, we calculate the average methylation difference between haplotypes H_1 and H_2 as a percentage relative to the least methylated haplotype using:

$$\text{Methylation fold change (\%)} = \frac{\max(H_1, H_2)}{\min(H_1, H_2)}$$

Calculating the size and saturation of the panmethylome

In GA4K, 435 samples were sequenced with methylation calling enabled. Earlier samples were sequenced without methylation calling and methylation data were not available. We genotyped the presence and absence of CpGs by listing the nodes covered by the aligned methylomes. To obtain saturation curves for the panmethylome, we simulated 10 curves, each with a randomly permuted order of methylomes. In each permutation, we start with the first methylome and progressively add newly discovered SV-CpGs, SV-5mCpGs, and 5mCpGs in subsequent methylomes (see `genotyped_cpg_saturation.py` and `methylated_cpg_saturation.py`). At every iteration, we record the number of accumulated CpGs and the number of new CpGs. Again, we use a minimum coverage threshold of 5 reads to discover a CpG (Sigurpalsdottir et al. 2024). To extrapolate the saturation rate, we fit a logarithmic function on the number of new CpGs across the average of the 10 simulations.

Calculating the methylation rate of CpGs

To know how often each CpG is methylated, we calculated the methylation rate as follows:

$$\text{Methylation rate} = \frac{\text{Number of samples where CpG is methylated}}{\text{Number of samples that carry the CpG}}$$

Again, every CpG must be covered by at least 5 reads (Sigurpalsdottir et al. 2024). We intersected the CpGs with repeat annotations, to obtain methylation rates stratified by repeats. We tested differences in the methylation rate between SV-CpGs and reference CpGs in each family and corrected for multiple testing using the R function `wilcox.test(alternative="two-sided")` and `p.adjust(method="fdr")`.

Mapping methylation QTLs

To map methylation SV-mQTLs and SNP-mQTLs, we binned the CHM13v2 backbone reference genome into nonoverlapping bins that are 200 bp in length. Then, we averaged the methylation levels of every CpG in each bin. CpGs with missing data are not considered in the average methylation level (see `bin_methylation.R`). Bins without CpGs are not considered since they do not contain any CpG methylation. Bins with one or a few CpGs are treated the same as bins with many CpGs. To map QTLs for each methylation bin, we ran linear regression using `lm()` between the methylation level of the bin and the genotype of every SV within 100 kbp (see `run_qtls.R`). Here, the genotype was the number of SV alleles (0, 1, 2) carried by each sample. We did the same for methylation bins and every SNP within 100 kbp. We corrected for multiple testing using `p.adjust(method="fdr")`. To rank and find the leading SV-mQTLs and SNP-mQTLs, we compared the absolute effect sizes of every variant associated with a methylation bin with an FDR < 0.05.

Data access

The 5-base HiFi-GS, HiFi long-read transcript sequencing (Iso-Seq), and WGBS raw and processed data, including assemblies and genotypes, generated in this study have been submitted to NCBI's database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) under accession number phs002206.v5.p1. Raw and processed data are available under restricted access due to IRB regulations and informed consent limiting access to users studying genetic diseases. Data access is provided by dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for certified investigators with local IRB approval in place. The CHM13v2.0 reference genome is available for download at https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/analysis_set/chm13v2.0.fa.gz. Methylation counts are available at Zenodo (https://zenodo.org/doi/10.5281/zenodo.13922419). Panmethyl is available at GitHub (https://github.com/cgroza/panmethyl) and as Supplemental Code. Scripts are available at Zenodo (https:// zenodo.org/doi/10.5281/zenodo.13922419) and as Supplemental Code. A docker container is available at https://hub.docker.com/r/cgroza/panmethyl.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank all families for participating in the Genomic Answers for Kids study. This work was made possible by generous gifts to the Children's Mercy Research Institute and Genomic Answers for Kids program at Children's Mercy Kansas City. We also thank Nick Nolte, Dan Louiselle, and Rebecca Biswell for their work in sample processing, Laura Puckett and Adam Walters for their work in library preparation and sequencing, and the clinical coordination team led by Bradley Belden for their work in clinical coordination. We also thank PacBio for sequencing support for a subset of the samples. T.P. holds the Dee Lyons/Missouri Endowed Chair in Pediatric Genomic Medicine, and B.G. holds the Roberta D. Harding & William F. Bradley, Jr. Endowed Chair in Genomic Research. C.G. is supported by the NSERC PGS D award. G.B. is supported by a Canada Research Chair Tier 1 award and an FRQS Distinguished Research Scholar award. This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada.

Author contributions: G.B. and T.P. conceived and designed the study; C.G. contributed to the study design; C.G., B.G., W.A.C., G.B., and T.P. analyzed the data and interpreted the results of experiments; C.G. prepared figures and drafted the paper; G.B. and T.P. edited and revised the paper; B.G. and W.A.C. provided bioinformatics support; all authors approved the final version of the paper.

References

Breiling A, Lyko F. 2015. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics Chromatin* **8:** 24. doi:10.1186/s13072-015-0016-6

Cheung WA, Johnson AF, Rowell WJ, Farrow E, Hall R, Cohen ASA, Means JC, Zion TN, Portik DM, Saunders CT, et al. 2023. Direct haplotype-resolved 5-base HiFi sequencing for genome-wide profiling of hypermethylation outliers in a rare disease cohort. *Nat Commun* **14:** 3090. doi:10.1038/s41467-023-38782-1

Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, Bansal L, Bartik L, Baybayan P, Belden B, et al. 2022. Genomic answers for children: dynamic analyses of >1000 pediatric rare disease genomes. *Genet Med* **24:** 1336–1348. doi:10.1016/j.gim.2022.02.007

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10:** giab008. doi:10.1093/gigascience/giab008

Daron J, Slotkin RK. 2017. EpiTEome: simultaneous detection of transposable element insertion sites and their DNA methylation levels. *Genome Biol* **18:** 91. doi:10.1186/s13059-017-1232-0

Dhar GA, Saha S, Mitra P, Chaudhuri RN. 2021. DNA methylation and regulation of gene expression: guardian of our health. *Nucleus* **64:** 259–270. doi:10.1007/s13237-021-00367-y

Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, Hagmann J, Vorbrugg S, Marco-Sola S, Kubica C, et al. 2024. Building pangenome graphs. *Nat Methods* **21:** 2008–2012. doi:10.1038/s41592-024-02430-3

Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, et al. 2022. Epigenetic patterns in a complete human genome. *Science* **376:** eabj5089. doi:10.1126/science.abj5089

Groza C, Kwan T, Soranzo N, Pastinen T, Bourque G. 2020. Personalized and graph genomes reveal missing signal in epigenomic data. *Genome Biol* **21:** 124. doi:10.1186/s13059-020-02038-8

Groza C, Chen X, Pacis A, Simon M-M, Pramatarova A, Aracena KA, Pastinen T, Barreiro LB, Bourque G. 2023. Genome graphs detect human polymorphisms in active epigenomic state during influenza infection. *Cell Genomics* **3:** 100294. doi:10.1016/j.xgen.2023.100294

Groza C, Schwendinger-Schreck C, Cheung WA, Farrow EG, Thiffault I, Lake J, Rizzo WB, Evrony G, Curran T, Bourque G, et al. 2024. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat Commun* **15:** 657. doi:10.1038/s41467-024-44980-2

Hattori N, Abe T, Hattori N, Suzuki M, Matsuyama T, Yoshida S, Li E, Shiota K. 2004. Preference of DNA methyltransferases for CpG islands in mouse embryonic stem cells. *Genome Res* **14:** 1733–1740. doi:10.1101/gr.2431504

Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, Abel HJ, Antonacci-Fulton LL, Asri M, Baid G, et al. 2024. Pangenome graph construction from genome alignments with minigraph-cactus. *Nat Biotechnol* **42:** 663–673. doi:10.1038/s41587-023-01793-w

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44:** D81–D89. doi:10.1093/nar/gkv1272

Jakubosky D, D'Antonio M, Bonder MJ, Smail C, Donovan MKR, Young Greenwald WW, Matsui H, Bonder MJ, Cai N, Carcamo-Orive I, et al. 2020. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat Commun* **11:** 2927. doi:10.1038/s41467-020-16482-4

Jima DD, Skaar DA, Planchart A, Motsinger-Reif A, Cevik SE, Park SS, Cowley M, Wright F, House J, Liu A, et al. 2022. Genomic map of candidate human imprint control regions: the imprintome. *Epigenetics* **17:** 1920–1943. doi:10.1080/15592294.2022.2091815

Kane NJ, Cohen ASA, Berrios C, Jones B, Pastinen T, Hoffman MA. 2023. Committing to genomic answers for all kids: evaluating inequity in genomic research enrollment. *Genet Med* **25:** 100895. doi:10.1016/j.gim.2023.100895

Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* **21:** 265. doi:10.1186/s13059-020-02168-z

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617:** 312–324. doi:10.1038/s41586-023-05896-x

Martin M, Ebert P, Marschall T. 2023. Read-based phasing and analysis of phased variants with WhatsHap. In *Haplotyping: methods and protocols* (ed. Peters BA, Drmanac R), pp. 127–138. Springer, New York.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Razin A, Cedar H. 1991. DNA methylation and gene expression. *Microbiol Rev* **55:** 451–458. doi:10.1128/mr.55.3.451-458.1991

Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, et al. 2015. ClinGen—the Clinical Genome Resource. *N Engl J Med* **372:** 2235–2242. doi:10.1056/NEJMsr1406261

Sigurpalsdottir BD, Stefansson OA, Holley G, Beyter D, Zink F, Hardarson MÞ, Sverrisson SÞ, Kristinsdottir N, Magnusdottir DN, Magnusson OÞ, et al. 2024. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biol* **25:** 69. doi:10.1186/s13059-024-03207-9

Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al. 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51:** D977–D985. doi:10.1093/nar/gkac1010

Sun Z, Behati S, Wang P, Bhagwate A, McDonough S, Wang V, Taylor W, Cunningham J, Kisiel J. 2023. Performance comparisons of methylation and structural variants from low-input whole-genome methylation sequencing. *Epigenomics* **15:** 11–19. doi:10.2217/epi-2022-0453

Wulfridge P, Langmead B, Feinberg AP, Hansen KD. 2019. Analyzing whole genome bisulfite sequencing data from highly divergent genotypes. *Nucleic Acids Res* **47:** e117. doi:10.1093/nar/gkz674

Yue X, Xie Z, Li M, Wang K, Li X, Zhang X, Yan J, Yin Y. 2022. Simultaneous profiling of histone modifications and DNA methylation via nanopore sequencing. *Nat Commun* **13:** 7939. doi:10.1038/s41467-022-35650-2

Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, et al. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500:** 477–481. doi:10.1038/nature12433

Received March 17, 2024; accepted in revised form February 11, 2025.