Methylation-aware long-read phasing significantly improves genome-wide haplotype reconstruction

Aaron Pfennig[1] and Joshua M. Akey[1,*]

[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

[*] Corresponding author: jakey@princeton.edu

Abstract

Haplotypes are linear sequences of co-inherited alleles along individual chromosomes and are central to genetic mapping, clinical variant interpretation, and inference of population history. However, accurate genome-wide haplotype reconstruction remains challenging. Long-read sequencing has the potential to dramatically improve haplotype inference, but existing methods do not directly leverage all the information embedded in these data. Here, we present LongHap, a read-based phasing method that integrates sequence and 5-methylcytosine (5mC) information in a unified probabilistic framework. By leveraging differentially methylated sites, LongHap resolves phase relationships between variants that are inaccessible to sequence-based approaches alone. Across multiple datasets and sequencing platforms, LongHap increases phase block lengths by up to 30% while substantially reducing switch error rates. LongHap rigorously embeds complex structural variants into the broader haplotype context using loopy belief propagation, enabling improved phasing of INDELs and other variant classes that are inherently difficult to resolve. Methylation-aware phasing also improves the accuracy and contiguity of haplotypes spanning rare variants and structurally complex, medically relevant genes across diverse ancestries, facilitating the interpretation of compound heterozygosity and haplotype-specific regulatory architectures. These results establish methylation-aware phasing as a general framework for improving genome-wide haplotype reconstruction, with broad applications across genetics and genomics.

Introduction

Genetic variation is inherently organized into haplotypes—combinations of alleles that are inherited together on individual chromosomes—and many fundamental biological processes operate at this level rather than at individual variant sites[1]. Haplotype structure shapes patterns of linkage disequilibrium, governs the joint effects of multiple variants on gene regulation and protein function, and is critical for diverse applications ranging from association mapping and clinical variant interpretation to inference of population history and natural selection[2,3].

To facilitate powerful haplotype-based interpretation of genetic variation, phasing approaches have been developed that fall into three main categories: statistical, pedigree (trio), and read-based phasing[4]. Using large reference panels and parental sequencing data, statistical and trio phasing can infer haplotypes on a chromosome-scale[5,6], but rare variant phasing remains challenging, particularly in complex genomic regions, since statistical inference is underpowered and short reads are difficult to map in these regions[7,8]. Read-based phasing, in contrast, directly analyzes raw sequencing data to identify alleles that co-occur across heterozygous sites within individual reads[9–13], avoiding reference panel biases and enabling rare variant phasing without requiring parental sequencing data[7]. However, the ability of read-based approaches to phase variants over long distances has historically been limited by short read lengths.

Improved accuracy and declining costs have made long-read sequencing with PacBio and Oxford Nanopore Technologies (ONT) viable alternatives to short-read sequencing [14,15]. The longer read lengths of these technologies dramatically improve the ability of read-based phasing approaches, such as WhatsHap, HapCUT2, and HaploMaker, to infer accurate haplotypes over longer genomic segments[9–12]. A recent method, called LongPhase, achieves chromosome-scale phasing with ultra-long ONT (UL-ONT) data, while improving phasing contiguity and

computational efficiency over WhatsHap and HapCUT2[13]. However, LongPhase relies on simplified modeling of read evidence, which can limit its ability to robustly phase complex variants and regions affected by alignment uncertainty. In addition to DNA sequence variation, modern long-read sequencing technologies capture additional layers of information, including base modifications such as 5-methylcytosine (5mC), which can provide orthogonal evidence of haplotype identity. Despite the availability of epigenetic information in long-read sequencing data, only one method, MethPhaser[16], has been developed to harness 5mC data in long reads to *post hoc* refine the haplotypes inferred by another read-based phasing tool. Thus, no read-based phasing method exists that integrates sequence and epigenetic information in a unified framework, leaving orthogonal signals in long-read sequencing unexploited.

To address these limitations, we developed LongHap, a read-based phasing method that seamlessly integrates sequence and 5mC methylation information from PacBio HiFi and ONT sequencing data in a unified framework (Figure 1). By integrating methylation signals, LongHap resolves phase relationships that cannot be inferred from sequence data alone. This enables the extension of phase blocks across gaps lacking read overlap, reducing switch errors by up to 5% and increasing phase block contiguity by 30%. Crucially, LongHap's novel embedding of structural variants (SVs) into the broader haplotype context through loopy belief propagation extends the benefits of methylation-aware phasing to all variant classes. In addition to these genome-wide performance gains, we demonstrate that integrating methylation information enables phasing of challenging, medically relevant genes and rare variants that are difficult to phase otherwise, with direct implications for interpreting compound heterozygosity in clinical settings[5,17,18]. These results highlight the power of exploiting the full information content of long-read sequencing data rather than simply extending short-read paradigms to long-read data. The

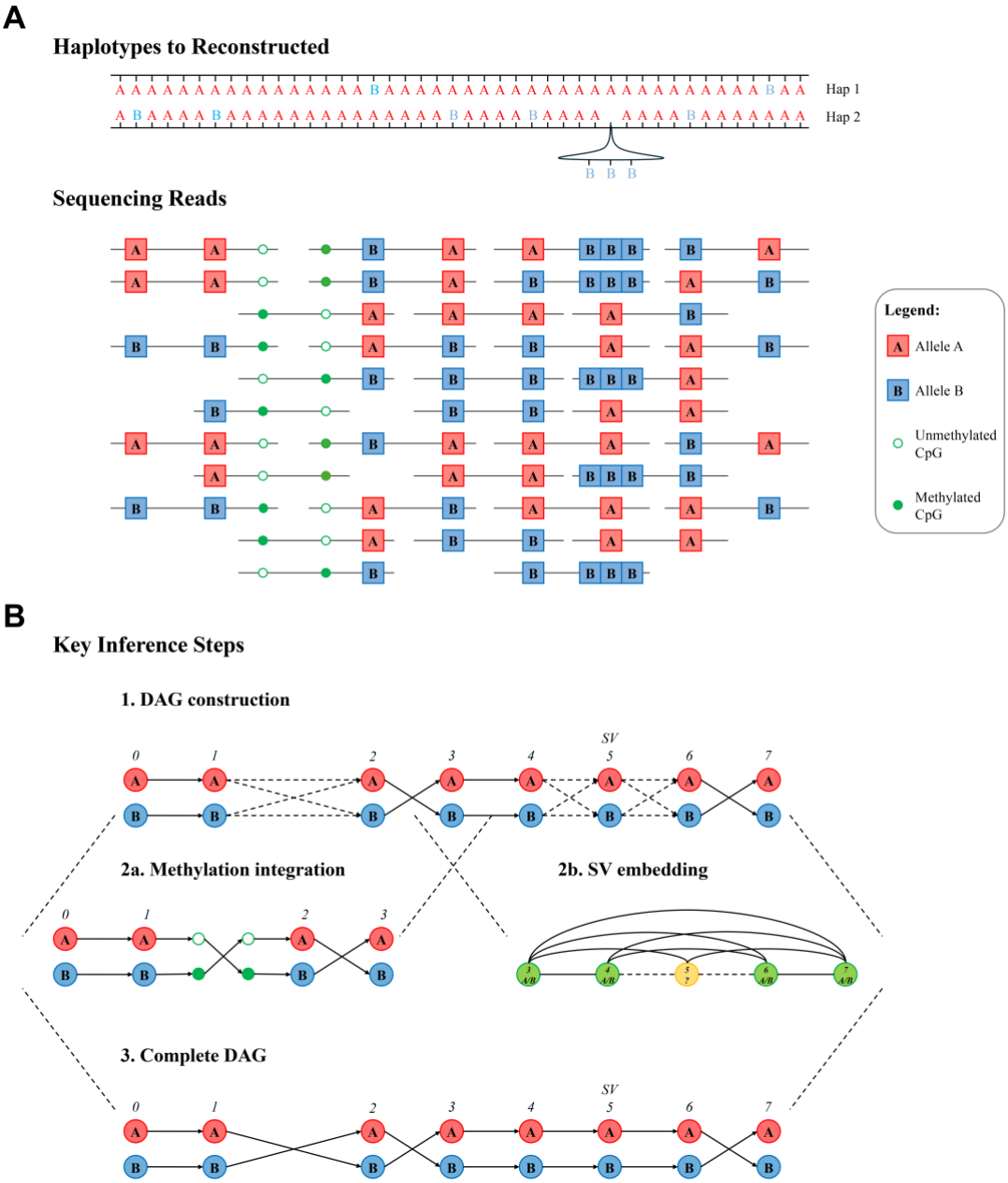

Figure 1. Schematic overview of LongHap. A) The haplotype structure at a locus that we want to reconstruct from long-read sequencing data. A and B denote alleles, and the bubble represents a three-base pair insertion on haplotype 2. Open and filled green circles indicate unmethylated and methylated CpG sites, respectively. Sequencing reads across the region are shown below the haplotypes. **B)** Schematic of the three key steps LongHap uses to infer haplotypes from read-based sequence and methylation variation. LongHap constructs a directed acyclic graph (DAG) in which each layer represents a heterozygous variant and edge weights reflect the co-occurrence of alleles across individual sequencing reads. Where adjacent phase blocks cannot be connected by sequence evidence alone, LongHap iteratively identifies differentially methylated sites on-the-fly and uses them to assign reads to maternal and paternal haplotypes, resolving transitions between two previously ambiguous layers. Finally, for complex or low-support variants, such as the structural variant at position 5 in the example shown, LongHap extracts a local subgraph and applies loopy belief propagation to incorporate long-range haplotype context, re-estimating transition probabilities into and out of the target variant. The complete DAG is then decoded using a Viterbi-like algorithm to identify the most likely haplotype configuration.

substantial improvements in phasing accuracy and contiguity afforded by LongHap will facilitate inferences into the biology and evolution of complex eukaryotic genomes.

Results

Overview of LongHap

LongHap is a read-based variant phasing method that integrates methylation signals native to long-read sequencing data, such as PacBio Revio HiFi and ONT sequencing, into a unified framework. Conceptually, LongHap reframes haplotype phasing as a problem of joint probabilistic inference from multiple, partially independent signals embedded within sequencing reads. As input, LongHap requires variant calls in VCF format, aligned sequencing reads in BAM format, and unphased methylation calls (Figure 1A). LongHap then uses a stepwise approach to phase SNVs, INDELs, and SVs (Figure 1B). First, LongHap phases pairs of variants based on sequence information. In a second step, LongHap dynamically identifies differentially methylated sites and leverages them as additional, phase-informative markers to resolve variant pairs that could not be confidently phased based on sequence information alone, thereby extending initially inferred phase blocks. LongHap then embeds complex and low-support variants into the broader haplotype context using graph theory, that is, loopy belief propagation, for more robust phasing (Figure 1B). Finally, LongHap outputs a phased VCF and, optionally, haplotagged read alignments, phase block coordinates, and the set of differentially methylated sites used for phasing.

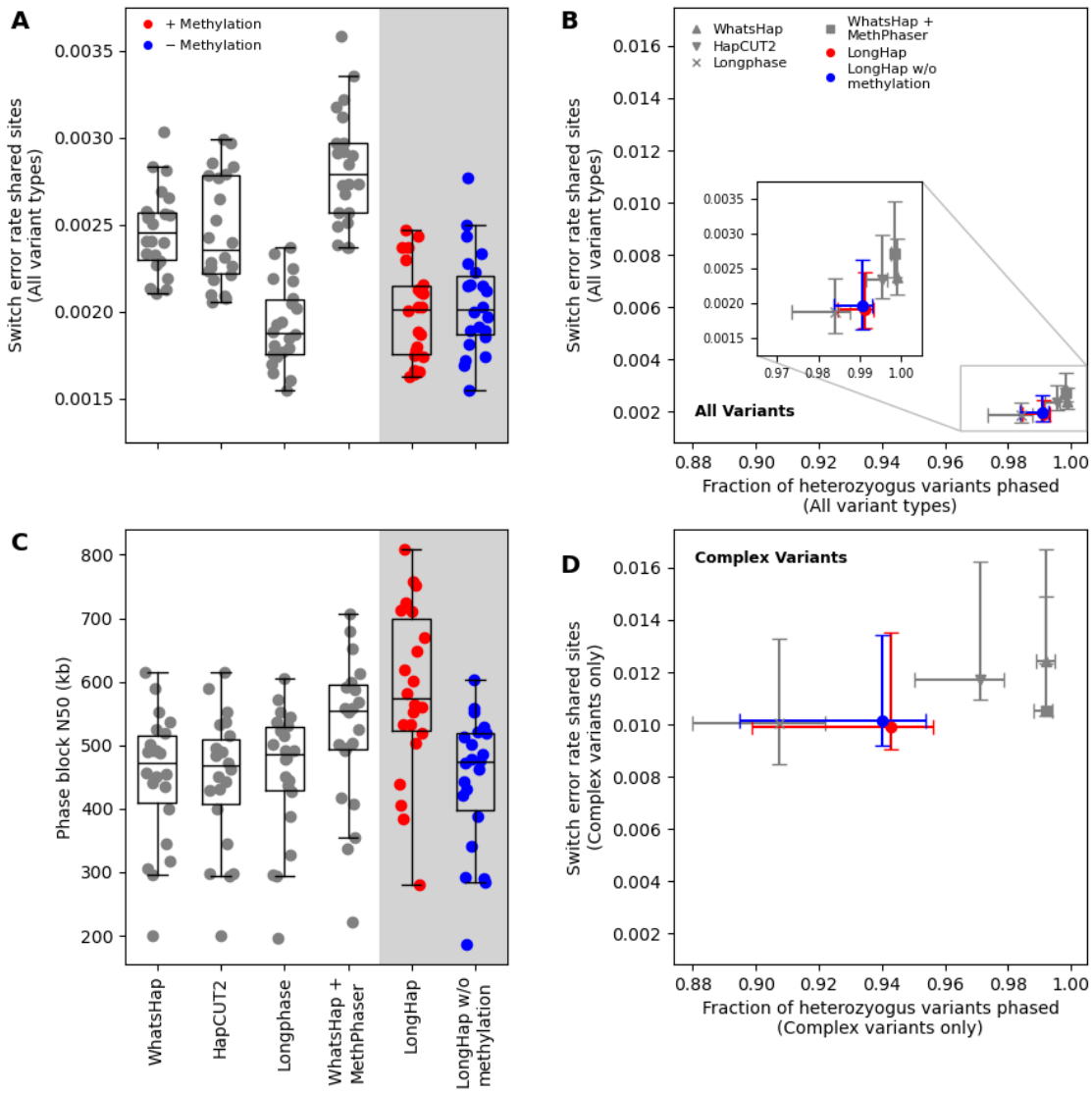

Figure 2. LongHap optimizes the trade-off between minimizing the switch error rate and maximizing the fraction of phased variants. A) Switch error rates for different read-based variant phasing tools using 38x PacBio Revio HiFi data for individual chromosomes of the well-characterized genome of HG002. LongHap makes 164 fewer errors when considering methylation information (4,860 vs. 5,024), and 2,117 fewer errors than WhatsHap + MethPhaser (6,977), another tool that integrates methylation information. LongPhase makes the fewest errors (4,754) but phases the fewest sites of all tools. Switch error rates were calculated based on the intersection of sites phased by all tools (96.13% of all heterozygous sites). **B)** By leveraging methylation information, LongHap phases more variants than LongPhase, while maintaining a lower error rate than all other tools. Switch error rates were calculated based on the intersection of heterozygous sites phased by all tools (96.13% of all heterozygous sites). **C)** Simultaneously, LongHap's integration of methylation information significantly increases the phase block N50 (mean N50: 584 kb compared to 443 kb), a greater gain than achieved by WhatsHap + MethPhaser (mean N50: 523 kb compared to 450 kb for WhatsHap alone). **D)** Similar to **(B)** LongHap phases more INDELs and SVs than LongPhase, while maintaining a lower error rate than all other tools, through the integration of methylation information and its rigorous embedding of INDELs and SVs into the broader haplotype context. Switch error rates were calculated based on the intersection of INDEL and SV sites phased by all tools (84.80% of all heterozygous sites with complex variants).

92 **LongHap outperforms all existing tools**

93 We evaluated LongHap's performance using PacBio Revio HiFi and regular and ultra-long
94 ONT long-read sequencing datasets for HG002 that were made available by the T2T consortium
95 and Human Pangenome Reference Consortium (HPRC)[19,20]. For all autosomes, we calculated the
96 switch error rates, the percentage of phased heterozygous sites, and phase block length N50s for
97 LongHap, WhatsHap, HapCUT2, LongPhase, and WhatsHap + MethPhaser using WhatsHap
98 compare and stats[9–11,13,16]. To increase comparability and highlight the benefits of integrating
99 methylation information, we compare LongHap's performance to the other tools without and with
100 methylation information. No tool phases all heterozygous sites, and we thus limit the comparisons
101 to the intersection of biallelic sites that were phased by all tools.

102 A key challenge in read-based phasing is simultaneously minimizing switch errors, while
103 maximizing the fraction of variants phased. LongHap advances this trade-off, outperforming
104 existing tools on this combined criterion across long-read sequencing technologies. Using PacBio
105 Revio HiFi data, LongHap achieves a switch error rate of 0.196% with methylation information,
106 while phasing 99.13% of heterozygous sites (Figures 2A & B). In contrast, LongPhase achieves a
107 marginally lower switch error rate (0.192%) but phases 18,394 fewer sites (98.42%). Conversely,
108 WhatsHap and HapCUT2 phase more sites but with considerably higher error rates (~0.246%)
109 (Figure 2B). Notably, LongHap's integration of methylation information results in 164 fewer
110 errors than sequence-only phasing. In contrast, MethPhaser, the only other tool that incorporates
111 methylation, increases the switch error rate (0.281%) (Figure 2A & B), underscoring the advantage
112 of jointly modeling sequence and epigenetic signals over *post hoc* refinement. As discussed below,
113 the importance of balancing the switch error rate with the fraction of phased heterozygous sites is
114 exacerbated for noisier ONT sequencing data.

115 Beyond switch error rates, phase block contiguity determines how much of the genome can be
116 interpreted in a haplotype-resolved manner. While all methods achieve similar contiguity without
117 methylation data (~443 kb), incorporating methylation signals increases LongHap's phase block
118 N50 by 31.9% to 584 kb, outperforming WhatsHap + MethPhaser (523 kb; +16.2%) (Figure 2C).
119 In summary, these results demonstrate that integrating methylation information enables LongHap
120 to extend phase blocks and resolve phase relationships that cannot be inferred from sequence data
121 alone. This leads to a 3.3% reduction in genome-wide switch errors and a >30% increase in phase
122 block contiguity.

123 **LongHap accurately co-phases complex structural variants**

124 Complex variants such as INDELs and SVs are inherently more difficult to phase than SNVs
125 due to sequencing errors and alignment ambiguities around breakpoints. LongHap addresses these
126 challenges by co-phasing INDELs and SVs along with SNVs, combining local realignment to
127 resolve alignment uncertainty with loopy belief propagation to efficiently embed complex variants
128 into the broader haplotype context. In contrast, WhatsHap and HapCUT2 are primarily designed
129 for SNV phasing[9–11], while LongPhase treats INDELs and SVs similarly to SNVs, without
130 explicitly modeling alignment uncertainty or complex variant structure[13].

131 To assess the benefits of this principled approach, we separately evaluated the accuracy of each
132 tool for INDELs and SVs. As with genome-wide benchmarks, LongHap advances the Pareto
133 frontier for complex variant phasing. Without methylation information, LongHap achieves a
134 switch error rate comparable to LongPhase (1.01% vs 1.00%), while phasing substantially more
135 complex variants than LongPhase (94.29% vs 90.73%). Crucially, with methylation information,
136 LongHap attains the lowest switch error rate of all tools (0.99%) and phases 12,929 (3.56%) more
137 INDELs and SVs than LongPhase (Figure 2D). This demonstrates that methylation information is

138 particularly valuable for complex variant phasing, where sequence evidence alone is most likely
139 to be ambiguous or insufficient around variant breakpoints.

140 **LongHap capitalizes on longer read lengths of ONT sequencing data**

141 Integrating methylation information yields consistent improvements across all ONT
142 sequencing platforms, demonstrating that LongHap's methylation-aware framework generalizes
143 across modern long-read sequencing technologies (Figures S1 & S2). With ONT data, LongHap
144 achieves a lower switch error rate than WhatsHap, HapCUT2, and WhatsHap + MethPhaser
145 (0.110% without and with methylation information compared to 0.129%, 0.135%, and 0.141%,
146 respectively) (Figure S1A), while phasing substantially more heterozygous sites than LongPhase
147 (98.37% - 99.43% for LongHap compared to 93.81% - 96.42% for LongPhase) at a slightly higher
148 switch error rate (0.110% vs. 0.081%) (Figure S1B). The difference between the tools becomes
149 most acute for INDELs and SVs, with LongPhase phasing only 74.93% and 70.74% of
150 heterozygous INDELs and SVs compared to 94.85% and 95.04% for LongHap with ONT and UL-
151 ONT data, respectively (Figures S1D, S2D).

152 LongHap also successfully capitalizes on the greater read length in ONT and UL-ONT
153 sequencing data, achieving chromosome-scale phasing (average phase block N50s: 1.97 Mb and
154 80.7 Mb, respectively). In contrast, WhatsHap fails to translate the longer read lengths into
155 meaningfully longer phase blocks (average phase block N50s: 1.47 Mb and 5.82 Mb), and
156 MethPhaser cannot compensate for WhatsHap's poor phase block N50 (average phase block N50s:
157 1.69 Mb and 7.22 Mb) (Figures S1C & S2C).

158 Together, these data highlight the fundamental trade-off between accuracy and completeness
159 in haplotype inference, whereby methods that minimize switch error rates often do so by excluding
160 difficult-to-phase variants. LongHap optimizes this trade-off, phasing more variants than

161 LongPhase but fewer than WhatsHap or HapCUT2, while maintaining an intermediate switch error
162 rate with ONT and UL-ONT sequencing data.

163 **Gap-bridging properties of phase-informative methylation signals**

164 Leveraging methylation information, LongHap connects 1,636 out of 9,005 phase blocks
165 inferred based on sequence information (18.2%) that could not be resolved by sequence evidence
166 alone using PacBio Revio HiFi data. The average gap between two connected blocks is 32 kb
167 (median: 29 kb; maximum: 233 kb; standard deviation: 12 kb) (Figure 3A). On average, LongHap
168 considers 20.2 differentially methylated sites (median: 15; SD: 19.9) to connect two phase blocks
169 (Figure 3B). These gaps bridged between two phase blocks scale with read length. For ONT and
170 UL-ONT sequencing data, the average distances between two phase blocks connected with
171 differentially methylated sites are 56 kb (median: 51 kb; maximum: 325 kb; standard deviation:
172 24 kb) and 415 kb (median: 239 kb; maximum: 2,579 kb; standard deviation: 626 kb), respectively
173 (Figures S3 & S4).

174 Phase-informative methylation sites are characterized by strongly asymmetric methylation
175 patterns between haplotypes. To confirm that the differentially methylated sites leveraged by
176 LongHap are informative for phasing, we compared the fraction of reads with evidence of
177 methylation, $P(meth) > 0.5$, that support each haplotype. Most sites used by LongHap show
178 differential methylation patterns with no or only an intermediate fraction of reads assigned to one
179 haplotype being methylated, while most reads assigned to the other haplotype are unmethylated
180 (Figure 3C).

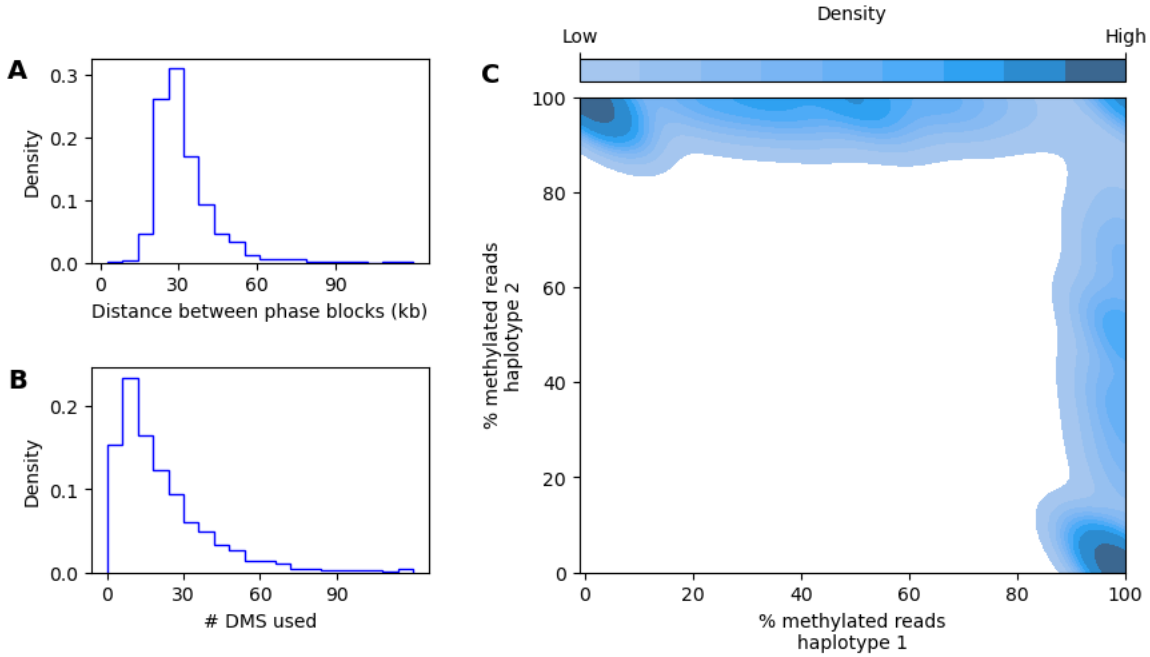

Figure 3. Characterization of differentially methylated sites (DMS) used by LongHap to bridge previously discontinuous phasing blocks. A) Distribution of distances between phase blocks connected using methylation information, using 38x PacBio Revio HiFi data. Most bridged gaps are between 20 and 40 kb, with a small number of gaps exceeding 120 kb. **B)** Distribution of the number of DMS sites used to bridge each gap. Most gaps are resolved using 20 DMS sites, with diminishing frequency at higher counts. **C)** Contour plot the fraction of reads with methylation marks assigned to either haplotype at sites used for phasing by LongHap. Most sites show clear signals of differential methylation, that is, they are truly phase-informative, as most reads are methylated on one haplotype while most reads are not methylated on the other haplotype (top left and bottom right corners). However, LongHap also identifies some false-positive differentially methylated sites (top right corner), injecting noise into the methylation-based phasing. A read is methylated at a given site if $P(meth) > 0.5$.

Methylation information enables the phasing of biomedically important loci

To evaluate LongHap's ability to phase a recently curated list of 273 challenging, medically relevant genes (CMRGs)[21], we intersected this gene list with phase block coordinates for each assessed tool. Integrating methylation information substantially improves LongHap's phasing of CMRGs across all metrics. While LongHap and WhatsHap + MethPhaser both fully phase the most CMRGs (211), LongHap achieves the most contiguous phasing of the 247 CMRGs covered by all tools, with the fewest overlapping phase blocks (633 compared to 756 for HapCUT2) and the greatest coverage of phased CMRG base pairs (128.9 Mb compared to 127.3 Mb for WhatsHap) (Table 1). Strikingly, methylation integration alone reduces the number of phase

Table 1. Number of phase blocks inferred by each tool using 38x PacBio Revio HiFi data that overlap challenging, medically relevant genes (CMRGs). 247 CMRGs are covered by all tools. LongHap infers the most contiguous phase blocks overlapping CMRGs, as indicated by the lowest number of phase blocks and the second most overlapping base pairs.

	CMRGs covered	**Phase blocks overlapping CMRGs**	**CMRGs fully phased**	**Common phase blocks overlapping CMRGs**	**Overlapping base pairs of common CMRGs (Mb)**
LongHap	250	635	211	633	128.9
LongHap w/o methylation information	248	714	206	714	125.2
LongPhase	254	741	209	764	127.2
WhatsHap	254	774	210	781	127.3
HapCUT2	253	759	209	756	126.6
WhatsHap + MethPhaser	254	873	211	883	126.7

190 blocks overlapping CMRGs from 714 to 633 (11.3%), representing a direct improvement in the
191 interpretability of these clinically important regions. Without methylation information, LongHap's
192 CMRG phasing is comparable to or slightly worse than other tools, underscoring that methylation
193 is the critical driver of these improvements. Similar patterns are observed with ONT and UL-ONT
194 sequencing data (Tables S2 & S3).

195 To illustrate how adding methylation information improves phasing of CMRGs, we analyzed
196 the *LIX1* gene in more detail (CHM13 coordinates: chr5:97,592,630-97,643,516; Figure 4). *LIX1*
197 is expressed in the central nervous system and thyroid in humans and is involved in cell regulation
198 and differentiation, as well as mitochondrial metabolism[22–25]. *LIX1* has recently been associated
199 with worse prognoses for individuals with gastrointestinal stromal tumors[26]. Without methylation
200 information, LongHap cannot phase the regions spanning the *LIX1* gene, as coverage dropout
201 causes the phase block to break, leaving the gene split across two disconnected blocks. Integrating

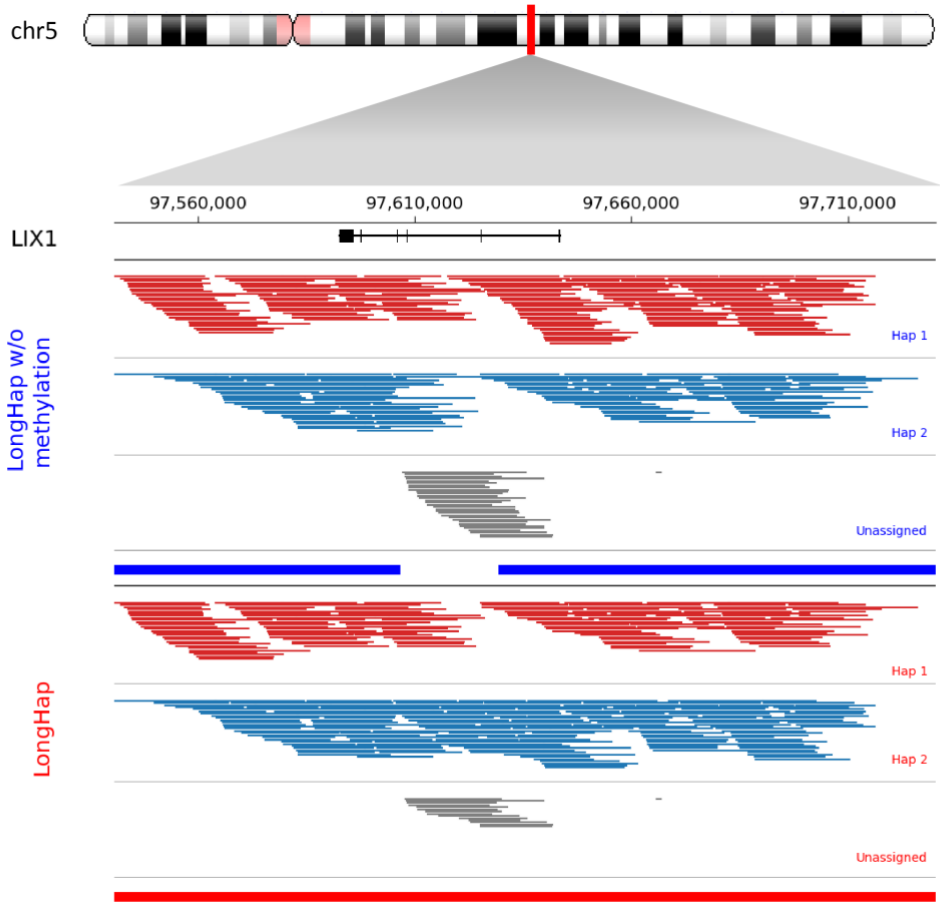

Figure 4. Methylation information enables continuous phasing across the LIX1 locus. Haplotagged read alignments at the LIX1 locus (chr5:97,592,630–97,643,516; CHM13v2.0) for LongHap run without (top) and with (middle) methylation information. Reads assigned to haplotype A are shown in red, haplotype B in blue, and unassigned reads in grey. Without methylation information, a coverage dropout causes the phase block to break into two. By leveraging differentially methylated sites to assign additional reads to distinct haplotypes, LongHap resolves a continuous phase block spanning the entire locus. Inferred phase blocks in the bottom tracks for LongHap without methylation (blue) and with methylation information (red). Genomic coordinates are shown relative to CHM13v2.0.

methylation information allows connecting the two adjacent phase blocks by leveraging phase-informative differentially methylated sites to assign additional reads to distinct haplotypes, producing a single continuous phase block spanning the entire locus (Figure 4).

207 **Genome-wide inference of haplotypes in globally diverse individuals**

208 We investigated LongHap's ability to generate new biological knowledge by analyzing seven
209 samples from diverse genetic ancestries using PacBio Revio HiFi data from the HPRC[20].
210 Specifically, we analyzed two Southern Han Chinese samples (HG00609, HG00658), two
211 admixed American samples from Puerto Rico (HG00738, HG01099), and two sub-Saharan
212 African Mandinka samples (HG02723, HG02615) in addition to HG002, which is of European
213 descent. When comparing LongHap's phasing against the *de novo* genome assemblies from HPRC
214 release 2, LongHap achieves low switch error rates for all samples, ranging from 0.185% to
215 0.486% per chromosome. For the East Asian samples, the switch error rates are higher (0.313% -
216 0.486%) and inferred phase block N50s are shorter (mean: 321 kb) than for African samples
217 (switch error rates: 0.199% - 0.374% per chromosome; mean phase block N50: 1,727 kb). This
218 pattern reflects the higher heterozygosity of African samples, in which the greater density of
219 heterozygous sites enables more contiguous phasing. However, the integration of methylation
220 leads to the same relative improvements in phasing accuracy and contiguity independent of genetic
221 ancestry (Figure S5).

222 Rare variants (MAF < 1%) are particularly challenging to phase using statistical approaches,
223 which are underpowered at low allele frequencies, making read-based phasing essential for their
224 accurate haplotype assignment[8]. By leveraging methylation signals, LongHap reduces switch
225 errors between rare variants and increases the number of rare variants that can be phased.
226 Specifically, using PacBio Revio HiFi data, LongHap makes, on average, 45 fewer errors when
227 phasing rare variants per genome (range: -201 - $+8$; Figure 5A), while phasing substantially more
228 rare variants (range 151 – 822) (Figure 5B). This underscores that the benefit of methylation-aware
229 phasing is the largest in individuals of European and East Asian ancestry, where lower

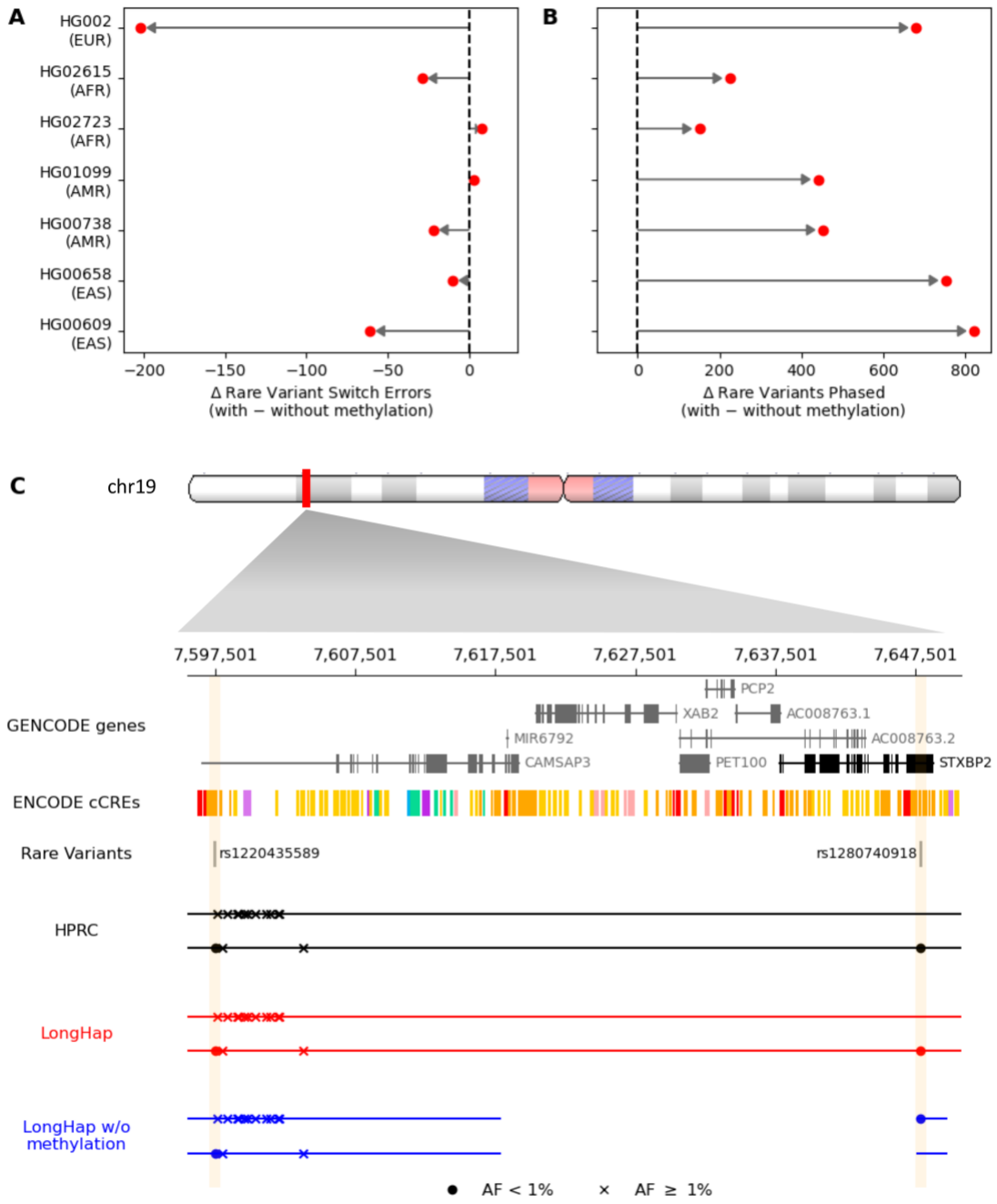

Figure 5. Methylation information improves phasing of rare variants across diverse ancestries and enables haplotype inference at a disease-relevant locus. A) Change in the number of switch errors between rare variants (MAF < 1%) when integrating methylation information across seven individuals of diverse genetic ancestries. Negative values indicate fewer errors with methylation. **B)** Corresponding change in the number of rare variants successfully phased. In all individuals, methylation integration reduces switch errors and increases the number of phased rare variants. **C)** At the STXBP2 locus in HG00609, LongHap without methylation information (bottom) places two rare variants (rs1220435589, rs1280740918) in separate phase blocks, preventing inference of their cis/trans configuration. Integrating methylation information (middle) resolves a continuous phase block, correctly identifying that both variants occur in cis in enhancer elements linked to STXBP2 expression. The HPRC assembly-based phasing (top) is shown as ground truth. Only variants present in the 1000 Genomes Project reference panel are shown. Genomic coordinates are shown relative to CHM13v2.0.

230 heterozygosity creates more frequent gaps in sequence-based phasing that methylation information
231 can bridge.

232 To illustrate this, we examined the *STXBP2* locus in HG00609, a CMRG implicated in
233 Hemophagocytic Lymphohistiocytosis type 5, a fatal autoimmune disease (CHM13 coordinates:
234 chr19: 7,637,764- 7,648,756; Figure 5C)[27–31]. Without methylation information, two rare variants
235 at this locus (rs1220435589, rs1280740918) fall in separate phase blocks, making it impossible to
236 determine their *cis*/*trans* configuration. Consistent with the HPRC assembly, LongHap resolves a
237 continuous phase block spanning both variants and infers that these variants occur in *cis* when
238 leveraging methylation information, which is supported by the HPRC assembly, and is critical for
239 interpreting their combined functional impact, as both variants reside in enhancer elements linked
240 to *STXBP2* expression[32]. This example illustrates how LongHap can resolve clinically critical
241 haplotypes, such as the haplotype-specific regulatory architecture at the STXBP2 locus, that are
242 inaccessible to sequence-only approaches.

243 Discussion

244 LongHap extends read-based phasing by leveraging methylation signals to resolve haplotype
245 relationships that are inaccessible to sequence information alone, enabling more accurate and
246 contiguous phasing, particularly for rare and structurally complex variants. Integrating methylation
247 reduces switch error rates by up to 5% and increases phase block N50 by over 30%, translating
248 into hundreds fewer phasing errors and thousands more phased heterozygous sites genome-wide,
249 with consistent improvements across long-read sequencing platforms (Figures 2, S1, & S2).

250 To our knowledge, LongHap is the first phasing method to selectively apply belief propagation
251 to embed INDELs, SVs, and low-support variants into a broader haplotype context. This targeted
252 embedding – combined with local realignment around complex breakpoints – improves phasing

253 accuracy at challenging loci without incurring the computational overhead associated with globally
254 embedding all variants, as reflected by LongHap's ability to phase a large fraction of INDELs and
255 SVs, while maintaining a low switch error rate and fast runtimes (Figures 2D, S1D, S2D).
256 LongHap phases chromosome 1 of HG002 using 38x PacBio Revio HiFi data with and without
257 methylation information in under ten minutes, substantially faster than MethPhaser, while
258 remaining competitive with existing tools, with runtimes and memory scaling with sequencing
259 depth and read length (Figures S6-S8).

260 By integrating methylation signals directly into the phasing process, LongHap extracts
261 substantially more information than MethPhaser, which leverages methylation signals *post hoc* to
262 refine pre-inferred haplotypes. While MethPhaser slightly increases switch error rates in our
263 benchmarks, LongHap consistently reduces switch errors and extends phase blocks, demonstrating
264 that joint modeling of genetic and epigenetic signals is more effective than *post hoc* refinement
265 (Figures 2, S1, and S2).

266 Importantly, the benefits of methylation-aware phasing extend beyond genome-wide
267 benchmarks to challenging, medically relevant genes and rare variants. Integrating methylation
268 information reduces the number of phase blocks overlapping CMRGs by 11.3% and enables
269 complete phasing of more CMRGs than any other tool, directly improving the interpretability of
270 clinically important genomic regions (Table 1). For rare variants, which are systematically
271 underpowered in statistical phasing approaches[8], methylation integration consistently phases
272 hundreds more rare variants per genome across individuals of diverse ancestries, with the largest
273 gains in individuals of European and East Asian ancestry, where lower heterozygosity creates more
274 frequent sequence-based phasing gaps. Critically, even modest improvement can have outsized
275 clinical consequences since correctly resolving whether two rare variants occur in *cis* or *trans*

276 determines their combined functional impact and is essential for interpreting compound
277 heterozygosity in recessive disorders[5,17,18]. As exemplified by our analyses of two rare variants at
278 the STXBP2 locus, LongHap can facilitate such analyses, aiding the interpretation of genetic
279 variation in clinical settings.

280 By jointly leveraging information from read sequences and methylation patterns, LongHap
281 provides a flexible framework that can be naturally extended to additional epigenetic or read-level
282 signals, such as 5-hydroxymethylcytosine (5hmC), N6-methyladenine (6mA), and other base
283 modifications that become detectable with routine long-read sequencing protocols in the future[33].
284 While orthogonal experimental assays such as chromatin conformation capture can provide long-
285 range phasing information, they are not always practical in population-scale studies due to cost,
286 sample availability, or experimental complexity. As long-read sequencing becomes increasingly
287 prevalent, methods that maximize the information content of these data by integrating
288 complementary signals will be essential.

289 While we showed that harnessing methylation information improves read-based variant
290 phasing in individuals of diverse genetic ancestries, our analyses are based on sequencing data
291 derived from an Epstein–Barr virus–transformed lymphoblastoid cell lines, which may exhibit
292 atypical methylation patterns. However, the consistency of improvements observed across
293 individuals of diverse ancestries suggests that haplotype-specific methylation patterns informative
294 for phasing are a general feature of human genomes rather than a cell line-specific artifact.
295 Nonetheless, additional studies of primary tissues will be useful to characterize the generalizability
296 of methylation-aware phasing.

297 In summary, by jointly leveraging sequence and methylation information embedded in long-
298 read sequencing data within a unified, probabilistic, and computationally efficient framework,

299 LongHap substantially improves haplotype inference across variant classes and genomic contexts,
300 including complex SVs, INDELs, and rare variants in challenging, medically relevant genes. These
301 advances establish methylation-aware phasing as a broadly applicable strategy for studies of
302 genetic variation, disease mechanisms, haplotype-specific epigenetic regulation, and human
303 evolutionary history.

304 Online Methods

305 LongHap

306 *Input for LongHap*

307 As a minimal input, LongHap requires aligned sequencing reads in BAM format and variant
308 calls in VCF format. To incorporate epigenetic information from long-read sequencing data,
309 LongHap can optionally use site-specific methylation calls, generated, for example, with pb-CpG-
310 tools and provided as unphased methylation annotations.

311 *Determining allelic support in individual reads*

312 Using pysam v0.22.1 (https://github.com/pysam-developers/pysam) and cyvcf2 v0.31.1[34],
313 LongHap streams through the BAM file and infers the allele each read supports at individual
314 overlapping heterozygous sites in the VCF. Alignments with a mapping quality less than a
315 specified threshold (by default, 20) are ignored along with secondary alignments, duplicate reads,
316 and reads that are flagged as quality control failures. Unlike other read-based phasing tools,
317 LongHap can include "multiallelic" heterozygous sites, where both haplotypes are different from
318 the reference. On an individual genome level, these sites are still biallelic and simply require a
319 dynamic indexing for accurate phasing. However, for comparability purposes, all described results
320 are based on biallelic sites only.

321 To optimize runtime, LongHap first attempts to infer allelic states based on CIGAR strings and
322 matching subsequences, resorting to local realignments only if the allele supported by a read
323 cannot be confidently identified. For single-nucleotide variants (SNVs), LongHap requires an
324 exact match of bases when using highly accurate PacBio Revio reads, and for noisier ONT reads,
325 a matching base upstream and downstream of the heterozygous SNV is required in addition to a
326 supporting CIGAR string. For insertions of any size, LongHap requires that the CIGAR string
327 supports either no insertion/deletion (INDEL) or structural variant (SV) at all, or an insertion of
328 the exact length as expected based on the VCF record, along with a matching inserted sequence.
329 Similarly, for deletions, the CIGAR string must not support an INDEL, an SV of any size, or a
330 deletion with the exact length specified in the VCF record. If the allelic state of a read at a given
331 heterozygous site cannot be unambiguously identified according to the above criteria, the read-site
332 pair is left for local realignment.

333 If a read-site pair is left for realignment, the read is locally re-aligned to synthetic references
334 representing allele A and B, respectively. Synthetic references are created by embedding both
335 alleles in the surrounding reference sequence. For PacBio, SNVs, INDELs, and SVs are embedded
336 in 33 bps, 100 bps, and 100 bps, respectively. For noisier ONT data, SNVs, INDELs, and SVs are
337 embedded in 66 bps, 200 bps, and 200 bps, respectively. However, the embedding size can also be
338 defined from the command line for specific use cases. Gap open and extension penalties are
339 dynamically set based on sequencing technology, variant type, and sequence context. Using the
340 CIGAR string, we calculate the INDEL rate for each read. That is, we count the number of bases
341 implicated in an insertion or deletion operation and divide it by the total number of CIGAR
342 operations. For PacBio HiFi sequencing, the indel rate is typically less than 0.005, and we set the
343 gap open and extension penalty to seven and two, respectively. For ONT sequencing, the indel rate

can be higher depending on the kit used. For reads with indel rates, r, satisfying $0.005 \leq r < 0.02$, we set the gap open and extension penalties to 5 and 1, respectively. For reads with $0.02 < r \leq 0.05$, we set the gap open penalty to 4, and for reads with $r > 0.05$, we set the gap open penalty to 3, while keeping the gap extension penalty set to 1. Independent of the sequencing technology, we then double the gap open penalty for SNVs. Similarly, because both PacBio and ONT sequencing have higher error rates in homopolymeric sequences, we reduce the gap open penalty by 2 if a heterozygous site falls within a homopolymer, which we define as at least two identical bases surrounding the heterozygous site in both directions. Finally, we realign the read against both synthetic reference sequences, using a semiglobal alignment approach implemented in parasail v1.3.4 (*sg_stats_striped_sat*)[35]. We then assume that a read supports the allele yielding the greater alignment score.

Construction of directed-acyclic graph

After determining the allelic states of reads at overlapping heterozygous sites, we construct a directed acyclic graph (DAG) $\mathcal{G} = (V, E, \omega)$. V, the set of vertices, is defined by the heterozygous sites in the VCF file, including INDELs and SVs. For each heterozygous site, we create a layer with two vertices, representing two alleles (A and B), such that:

$$V = \{(v_A^0, v_B^0), (v_A^1, v_B^1), \dots, (v_A^{L-2}, v_B^{L-2}), (v_A^{L-1}, v_B^{L-1})\},$$

where L is the number of heterozygous sites in the VCF. E, the set of edges, comprises all possible edges between vertices of adjacent layers in $\mathcal{G}$, that is:

$$E = \{(v^i,\ v^{i+1}) \mid v^i \in \{v_A^i, v_B^i\}, v^{i+1} \in \{v_A^{i+1}, v_B^{i+1}\}\ \forall\ i = 0, \dots, L-2\}$$

where (v_A^i, v_B^{i+1}) represents the edge connecting allele A and B at site i and $i+1$, respectively.

365 Edge weights, ω, are calculated by tallying the co-occurrence of alleles at two adjacent
366 heterozygous sites i and $i+1$, calculating the frequency with which allele pairs (A_i, A_{i+1}),
367 (A_i, B_{i+1}) etc., co-occur at sites i and $i+1$. The weights of edges connecting layers i and $i+1$
368 $(\omega(v^i, v^{i+1}))$ are thus defined by:

369
$$\omega(v^i, v^{i+1}) = \begin{bmatrix} \omega(v_A^i, v_A^{i+1}) & \omega(v_A^i, v_B^{i+1}) \\ \omega(v_B^i, v_A^{i+1}) & \omega(v_B^i, v_B^{i+1}) \end{bmatrix}$$

370
$$= \begin{bmatrix} \frac{\sum A_i \to A_{i+1}}{\sum A_i} & \frac{\sum A_i \to B_{i+1}}{\sum A_i} \\ \frac{\sum B_i \to A_{i+1}}{\sum B_i} & \frac{\sum B_i \to B_{i+1}}{\sum B_i} \end{bmatrix}$$

371
$$= \begin{bmatrix} P(A_j|A_i) & P(B_j|A_i) \\ P(A_j|B_i) & P(B_j|B_i) \end{bmatrix}$$

372 Finally, LongHap mirrors the more confident transition, creating a symmetrical $2x2$ matrix for
373 each pair of adjacent heterozygous sites.

374 When analyzing ONT data, we require that each allele is observed on both DNA strands since
375 it was observed that variants supported by reads from only one DNA strand are more frequently
376 false-positive calls[13]. Variants that are only observed on reads from one strand are marked as not
377 phaseable and are excluded from subsequent analyses.

378 *Re-estimating transitions for difficult-to-phase variants using loopy belief propagation*

379 To improve haplotype reconstruction of difficult-to-phase variants, that is, INDELs, SVs, and
380 variants for which at least one allele was observed less than AN times ($AN < 1$ by default),
381 LongHap embeds them into the broader haplotype context by considering long-range edges
382 between non-adjacent heterozygous sites. Note that alleles observed zero times ($AN = 0$) can
383 infrequently occur as the result of alignment and variant calling errors or difficulty in determining

384 the allelic state of individual reads by LongHap. To re-phase m consecutive difficult-to-phase
385 variants $(i, \dots, i+m)$, a subgraph $\mathcal{G}_S = (V_S, E_S, \omega_S)$ is constructed with the vertices $V_S =$
386 $\{i-n, \dots, i, \dots, i+m, \dots, i+m+n\}$, where n is the number of surrounding variants that can be
387 confidently phased, and $E = \{(v^j, v^{j+1}) | j \in i-n, \dots i+m+n\}$. In practice, we choose $n = 2$.
388 Each vertex can take one of two states $x = \{A, B\}$. For each pair of adjacent vertices, we already
389 have a $2x2$ transition matrix $\omega(v^i, v^{i+1})$, indicating probabilities with that vertex i transitions
390 from its current state into either state at vertex $i+1$. In addition to these edges between adjacent
391 pairs, long-range edges are considered, connecting the preceding and succeeding confidently
392 phased vertices with all other vertices in the graph that are more than one step away, that is,
393 $\omega(v^j, v^k) \mid j \in \{i-n, \dots, i-1\} \forall k \in \{j+2, \dots, j+m+n+1\}$ and $\omega(v^j, v^k) \mid j \in$
394 $\{i, \dots, i+m-1\} \forall k \in \{j+2, \dots, j+m+n-1\}$. For example, if $m = 1$ and $n=2$, long-
395 range transitions are inferred between vertex $i-2$ and vertices $i \dots i+2$, vertex $i-1$ and vertices
396 $i+1$ and $i+2$, and vertex i and vertex $i+2$.

397 Belief Propagation (BP) provides an efficient means to infer marginal distributions of states at
398 individual vertices by factorizing the graph and propagating previously calculated messages
399 between vertices[36,37]. In acyclic graphs, BP yields an exact solution, while in graphs with cycles,
400 only an approximation can be obtained. Although the above-described directed subgraph has no
401 cycles *a priori*, it does have cycles in its moralized (undirected) form. We thus must approximate
402 marginal distributions at individual vertices, using loopy BP[37,38].

403 Given the above-described moralized subgraph, LongHap determines the neighbors for each
404 vertex based on whether they are connected by an edge. Furthermore, states for each vertex j (ϕ_j)
405 are initialized uniformly, except for the first vertex $(i-n)$, which is initialized to be in state A.
406 The message a node j sends to a neighbor j ($m_{j \to k}(x_k)$) depends on its state ϕ_j, the messages it

receives from its other neighbors, and the local factor connecting j and k ($\psi_{j,k}(x_j, x_k) = \omega(v^j, v^k)$):

$$m_{j\to k}(x_k) = \sum_{x_j} \phi_j(x_j)\psi_{j,k}(x_j, x_k) \prod_{l\in N(j)\setminus\{k\}} m_{l\to j}(x_j)$$

The loopy BP is started at an arbitrary pair of vertices j and k, and an updated $\widehat{m_{j\to k}(x_k)}$ is calculated. Based on $\widehat{m_{j\to k}(x_k)}$, all other messages in the graph are updated, and the algorithm is run iteratively until it converges or reaches a maximum number of iterations (by default, 500). Although convergence is not guaranteed, we found that it virtually always converges within a few iterations. Given the messages received by a node j, its normalized marginal belief ($b_j(x_j)$) is given by:

$$b_j(x_j) = \frac{\phi_j(x_j)\prod_{k\in N(j)} m_{k\to j}(x_j)}{\sum_{x_j}\phi_j(x_j)\prod_{k\in N(j)} m_{k\to j}(x_j)}$$

Similarly, unnormalized pairwise marginal beliefs of vertices j and k are given by:

$$b_{jk}(x_j, x_k) = \phi_j(x_j)\phi_k(x_k)\psi_{j,k}(x_j, x_k)\left(\prod_{l\in N(j)\setminus\{k\}} m_{l\to j}(x_j)\right)\left(\prod_{m\in N(k)\setminus\{j\}} m_{m\to k}(k)\right)$$

$b_{jk}(x_j, x_k)$ can be interpreted as the transition probability between j and k. However, we are only interested in $\omega(\widehat{v^j, v^{j+1}})$ of adjacent heterozygous sites in the graph that is stripped of the long-range connections but implicitly still contains their information, that is, the information of long-range connections is projected in pairwise adjacent connections. $\omega(\widehat{v^j, v^{j+1}})$ is thus obtained by solving:

$$b_{j+1}(x_{j+1}) = \sum_{x_j} b_j(x_j)\,\omega(\widehat{v^j, v^{j+1}})$$

yielding:

426 $$\omega(\widehat{v^j, v^{j+1}}) = \frac{b_j(x_j)+b_{j+1}(x_{j+1})-1}{2b_j(x_j)-1}$$

427 $\omega(\widehat{v^j, v^{j+1}})$ is consistent with the full DAG described above. In the example graph shown in
428 Figure 1B, LongHap aims to calculate $\omega(\widehat{v^4, v^5})$ and $\omega(\widehat{v^5, v^6})$ as described above.

429 *Initial attempt to resolve ambiguous transitions*

430 Although long-read sequencing allows phasing of most pairs of adjacent heterozygous sites,
431 not all transitions can be unambiguously resolved. For some pairs i and $i+1$, all transitions
432 between allelic states may have equal probability, causing a break in the phase block (see below).
433 In such cases, LongHap attempts to extend the phase block by testing whether the transitions from
434 i to $i+2$ can be confidently inferred. If this is the case, LongHap tests if it can use this long-range
435 edge to update the posterior beliefs for the transitions from i to $i+1$ and i to $i+2$, or marks the
436 heterozygous variant at site $i+1$ as not phaseable and connects i to $i+2$.

437 *Leveraging methylation information to resolve remaining ambiguous transitions*

438 After constructing the initial DAG based on sequence evidence, LongHap leverages
439 methylation information inherent to PacBio and ONT long-read sequencing data to resolve
440 remaining ambiguous transitions. These include pairs of adjacent heterozygous sites i and $i+1$
441 that are not overlapped by any read, making it impossible to phase them solely based on sequence
442 evidence.

443 To leverage methylation calls, LongHap first identifies putatively differentially methylated
444 sites, that is, sites at which 20-80% of the reads are methylated. A haplotype is considered
445 methylated if the product of methylation probabilities across all reads assigned to the haplotype is
446 greater than 0.5, and its log10 odds ratio of being methylated versus unmethylated is greater than
447 three, and unmethylated otherwise. Finally, a site is considered differentially methylated and

448 informative for phasing if the inferred methylation states differ between the two haplotypes.
449 LongHap uses these differentially methylated sites as additional markers to connect to phase
450 blacks.

451 LongHap anchors reads based on their allelic states at sites $i-1$ and i and then tests if parental
452 haplotypes are differentially methylated at any candidate sites overlapped by these reads as
453 described above. It then uses the differentially methylated sites to iteratively assign reads to either
454 haplotype and identify additional differentially methylated sites based on the newly assigned reads.
455 This procedure is repeated until either the gap between sites i and $i+1$ is bridged or no more
456 reads can be assigned to either haplotype. If the gap was successfully bridged, LongHap uses all
457 reads assigned to a haplotype to infer the transition probabilities between sites i and $i+1$
458 ($\omega(v^i, v^{i+1})$).

459 *Viterbi-like inference of most likely haplotype*

460 After completing the DAG, LongHap employs a Viterbi-like decoding scheme to identify the
461 most likely haplotypes. That is, there are just transition probabilities and no emission probabilities
462 in our model. At the beginning of the forward pass, we assign equal probabilities of being in state
463 A and B at the first layer. LongHap then calculates the probability of being in either state at each
464 subsequent layer of the DAG, keeping track of the more likely state. Whenever it encounters an
465 ambiguous transition between two layers, that is, all transitions are equally likely, it stops the
466 forward pass, randomly initializes a new phase block, and backtraces the previous phase block.

467

468

469

470 *Output*

471 LongHap writes a phased VCF, providing a PS tag for each phased variant. Optionally,
472 LongHap also provides phase block coordinates, haplotagged read alignments in BAM format (HP
473 tag), and information on the methylation sites used for phasing.

474 **Sequencing data**

475 We retrieved multiple PacBio and ONT sequencing datasets for HG002 for benchmarking
476 purposes from the T2T consortium[19]. Specifically, we downloaded PacBio Revio HiFi reads, ONT
477 Simplex R10.4.1 reads base-called with Dorado (PAW70337), and UL-ONT Duplex R10.4.1 reads
478 base-called with Dorado. We also retrieved PacBio Revio HiFi sequencing data for six additional
479 samples from HPRC: two East Asian samples (HG00609, HG00658), two admixed samples from
480 the Americas (HG00738, HG01099), and two sub-Saharan African samples (HG02723,
481 HG02615)[20]. Table S1 summarizes the characteristics of the sequencing datasets used and provides
482 links to each dataset.

483 **Sequence alignment, variant calling, and methylation calling**

484 Sequences were aligned to CHM13v2.0 using minimap2 v2.28-r1221-dirty with *--map-hifi* and
485 *--map-ont* presets for PacBio Revio HiFi and ONT sequencing data, respectively[39,40]. We then
486 used PacBio Revio HiFi sequence alignments to call SNVs and small INDELs with DeepVariant
487 v1.9.0 and larger SVs with Sniffles2 v2.7.1[41,42]. We removed sites with a genotype quality less
488 than 20 (GQ < 20) and merged DeepVariant and Sniffles2 variant calls. Finally, we left-aligned
489 and merged multiallelic variants using *bcftools norm -m+* [43]. We determined methylation states
490 using *aligned_bam_to_cpg_scores* v3.0.0 from pb-CpG-tools
491 (https://github.com/PacificBiosciences/pb-CpG-tools).

492 **Benchmarking of read-based variant phasing methods**

493 We compared the performance of LongHap to WhatsHap v2.3[9,11], HapCUT2 v1.3.4[10],
494 LongPhase[13], and WhatsHap + MethPhaser v0.0.3[16]. For each tool, we calculated phasing statistics
495 using *whatshap stats* and switch error rates relative to the T2T HG002 Q100 variant call set using
496 *whatshap compare*[9,11,19]. Prior to evaluation, variant calls in the T2T HG002 Q100 ground truth
497 call set and phased variant calls were left-aligned, and multiallelic variants were split using *bcftools*
498 *norm -m -any*[43]. As these tools phased variable percentages of heterozygous sites, we calculated
499 switch error rates based on sites phased by all tools to allow a fair comparison.

500 We also evaluated the ability of each tool to contiguously phase a recently curated list of 273
501 challenging, medically relevant genes (CMRGs) made available by the Genome in a Bottle
502 consortium by intersecting their gene bodies with phase blocks using bedtools[21,44]. Although some
503 tools support parallelization, we used single CPU threads for all cores to enable fair comparisons
504 of computational requirements.

505 In addition to HG002, we compared LongHap's variant phasing for six additional individuals
506 from Africa, East Asia, and the Americas to the phasing in the respective assemblies from HPRC
507 release 2 (Table S1)[20], using WhatsHap as described above. To obtain a ground truth set, we called
508 variants for each assembly with dipcall v0.3[45]. We then annotated variant calls for each individual
509 with allele frequencies from the 1000 Genomes Project to specifically evaluated LongHap's ability
510 to phase rare variants (MAF < 1%)[40].

511

512

513

Data availability

No new data has been generated. All used data is publicly available from the website of the T2T consortium and the Human Reference Pangenome Consortium and is summarized in Table S1.

Code availability

LongHap and all code for supplementary analyses are available from: https://github.com/AkeyLab/LongHap.

Acknowledgements

We thank Rob Biermann for providing computational support, helping to improve the computational footprint of LongHap, and all members of the Akey lab for insightful discussions. We also acknowledge that the work reported on in this paper was substantially performed using high-performance computing resources at the Lewis-Sigler Institute for Integrative Genomics and the Princeton Research Computing resources at Princeton University. Princeton Research Computing is a consortium of groups, including the Princeton Institute for Computational Science and Engineering (PICSciE) and Research Computing at Princeton University.

Declaration of Interest

The authors declare no competing interests.

Supplemental information

Figures S1–S8 and Tables S1 and S3

533 **References**

534 1. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase
535 information for human genomics. *Nat. Rev. Genet.* **12**, 215–223 (2011).
536 2. Zhao, H., Pfeiffer, R. & Gail, M. H. Haplotype analysis in population genetics and association
537 studies. *Pharmacogenomics* **4**, 171–178 (2003).
538 3. Abondio, P., Cilli, E. & Luiselli, D. Inferring Signatures of Positive Selection in Whole-Genome
539 Sequencing Data: An Overview of Haplotype-Based Methods. *Genes* **13**, 926 (2022).
540 4. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new
541 developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
542 5. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing
543 of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat. Genet.* **55**, 1243–
544 1249 (2023).
545 6. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale
546 sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
547 7. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies
548 for whole human genomes. *PLOS Genet.* **14**, e1007308 (2018).
549 8. Lalli, J. L., Bortvin, A. N., McCoy, R. C. & Werling, D. M. A T2T-CHM13 recombination map
550 and globally diverse haplotype reference panel improves phasing and imputation.
551 2025.02.24.639687 Preprint at https://doi.org/10.1101/2025.02.24.639687 (2025).
552 9. Patterson, M. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation
553 Sequencing Reads. *J. Comput. Biol.* **22**, 498–509 (2015).
554 10. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for
555 diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).

556 11. Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. 085050 Preprint at
557 https://doi.org/10.1101/085050 (2016).
558 12. Fruzangohar, M., Timmins, W. A., Kravchuk, O. & Taylor, J. HaploMaker: An improved
559 algorithm for rapid haplotype assembly of genomic sequences. *GigaScience* **11**, giac038 (2022).
560 13. Lin, J.-H., Chen, L.-C., Yu, S.-C. & Huang, Y.-T. LongPhase: an ultra-fast chromosome-
561 scale phasing algorithm for small and large variants. *Bioinformatics* **38**, 1816–1822 (2022).
562 14. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-
563 read sequencing. *Nat. Rev. Genet.* **22**, 572–587 (2021).
564 15. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and
565 its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
566 16. Fu, Y. *et al.* MethPhaser: methylation-based long-read haplotype phasing of human
567 genomes. *Nat. Commun.* **15**, 5327 (2024).
568 17. Lassen, F. H. *et al.* Exome-wide evidence of compound heterozygous effects across
569 common phenotypes in the UK Biobank. *medRxiv* 2023.06.29.23291992 (2023)
570 doi:10.1101/2023.06.29.23291992.
571 18. Guo, M. H. *et al.* Inferring compound heterozygosity from large-scale exome sequencing
572 data. *Nat. Genet.* **56**, 152–161 (2024).
573 19. Hansen, N. F. *et al.* A complete diploid human genome benchmark for personalized
574 genomics. *bioRxiv* https://doi.org/10.1101/2025.09.21.677443 (2025)
575 doi:10.1101/2025.09.21.677443.
576 20. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
577 21. Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant
578 autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).

579 22. Le Guen, L., Notarnicola, C. & de Santa Barbara, P. Intermuscular tendons are essential
580 for the development of vertebrate stomach. *Development* **136**, 791–801 (2009).
581 23. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–
582 585 (2013).
583 24. McKey, J., Martire, D., de Santa Barbara, P. & Faure, S. LIX1 regulates YAP1 activity and
584 controls the proliferation and differentiation of stomach mesenchymal progenitors. *BMC Biol.*
585 **14**, 34 (2016).
586 25. Guérin, A. *et al.* LIX1-mediated changes in mitochondrial metabolism control the fate of
587 digestive mesenchyme-derived cells. *Redox Biol.* **56**, 102431 (2022).
588 26. Guérin, A. *et al.* LIX1 regulates YAP activity and controls gastrointestinal cancer cell
589 plasticity. *J. Cell. Mol. Med.* **24**, 9244–9254 (2020).
590 27. Stadt, U. zur *et al.* Familial Hemophagocytic Lymphohistiocytosis Type 5 (FHL-5) Is
591 Caused by Mutations in Munc18-2 and Impaired Binding to Syntaxin 11. *Am. J. Hum. Genet.*
592 **85**, 482–492 (2009).
593 28. Beutel, K. *et al.* Infection of T lymphocytes in Epstein-Barr virus-associated
594 hemophagocytic lymphohistiocytosis in children of non-Asian origin. *Pediatr. Blood Cancer*
595 **53**, 184–190 (2009).
596 29. Cetica, V. *et al.* STXBP2 mutations in children with familial haemophagocytic
597 lymphohistiocytosis type 5. *J. Med. Genet.* **47**, 595–600 (2010).
598 30. Stepensky, P. *et al.* Persistent defective membrane trafficking in epithelial cells of patients
599 with familial hemophagocytic lymphohistiocytosis type 5 due to STXBP2/MUNC18-2
600 mutations. *Pediatr. Blood Cancer* **60**, 1215–1222 (2013).

601 31. Vogel, G. F. *et al.* Disrupted apical exocytosis of cargo vesicles causes enteropathy in FHL5
602 patients with Munc18-2 mutations. *JCI Insight* **2**, (2018).
603 32. Moore, J. E. *et al.* An expanded registry of candidate cis-regulatory elements. *Nature* 1–10
604 (2026) doi:10.1038/s41586-025-09909-9.
605 33. Fu, Y., Timp, W. & Sedlazeck, F. J. Computational analysis of DNA methylation from long-
606 read sequencing. *Nat. Rev. Genet.* **26**, 620–634 (2025).
607 34. Pedersen, B. S. & Quinlan, A. R. cyvcf2: fast, flexible variant analysis with Python.
608 *Bioinformatics* **33**, 1867–1869 (2017).
609 35. Daily, J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence
610 alignments. *BMC Bioinformatics* **17**, 81 (2016).
611 36. Pearl, J. *Probabilistic Reasoning in Intelligent Systems*. (Morgan Kaufmann, 1988).
612 doi:10.1016/B978-0-08-051489-5.50010-2.
613 37. Kschischang, F. R., Frey, B. J. & Loeliger, H.-A. Factor graphs and the sum-product
614 algorithm. *IEEE Trans. Inf. Theory* **47**, 498–519 (2001).
615 38. Yedidia, J. S., Freeman, W. T. & Weiss, Y. Constructing free-energy approximations and
616 generalized belief propagation algorithms. *IEEE Trans. Inf. Theory* **51**, 2282–2312 (2005).
617 39. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
618 3100 (2018).
619 40. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354
620 (2023).
621 41. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks.
622 *Nat. Biotechnol.* **36**, 983–987 (2018).

42. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).

43. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).

44. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

45. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).