Method

Phased nanopore assembly with Shasta and modular graph phasing with GFAse

Ryan Lorig-Roach,[1] Melissa Meredith,[1] Jean Monlong,[1] Miten Jain,[2] Hugh E. Olsen,[1] Brandy McNulty,[1] David Porubsky,[3] Tessa G. Montague,[4,5] Julian K. Lucas,[1] Chris Condon,[1] Jordan M. Eizenga,[1] Sissel Juul,[6] Sean K. McKenzie,[6] Sara E. Simmonds,[7] Jimin Park,[1] Mobin Asri,[1] Sergey Koren,[8] Evan E. Eichler,[9,10] Richard Axel,[4,5] Bruce Martin,[7] Paolo Carnevali,[7] Karen H. Miga,[1] and Benedict Paten[1]

[1]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, California 95060, USA; [2]Department of Bioengineering, Department of Physics, Northeastern University, Boston, Massachusetts 02120, USA; [3]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA; [4]The Mortimer B. Zuckerman Mind Brain Behavior Institute, Department of Neuroscience, Columbia University, New York, New York 10027, USA; [5]Howard Hughes Medical Institute, Columbia University, New York, New York 10032, USA; [6]Oxford Nanopore Technologies Incorporated, New York, New York 10013, USA; [7]Chan Zuckerberg Initiative Foundation, Redwood City, California 94063, USA; [8]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20894, USA; [9]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA; [10]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA

Reference-free genome phasing is vital for understanding allele inheritance and the impact of single-molecule DNA variation on phenotypes. To achieve thorough phasing across homozygous or repetitive regions of the genome, long-read sequencing technologies are often used to perform phased de novo assembly. As a step toward reducing the cost and complexity of this type of analysis, we describe new methods for accurately phasing Oxford Nanopore Technologies (ONT) sequence data with the Shasta genome assembler and a modular tool for extending phasing to the chromosome scale called GFAse. We test using new variants of ONT PromethION sequencing, including those using proximity ligation, and show that newer, higher accuracy ONT reads substantially improve assembly quality.

[Supplemental material is available for this article.]

Phased genome assemblies enable a variety of clinically motivated analyses and population studies. For clinical and biological studies, there are many transcriptional and translational outcomes that could result from a given set of mutations in the same gene or regulon, depending on whether they co-occur on the same molecule of DNA (Cordeiro et al. 2006; Walker et al. 2016; Miller and Piccolo 2020). Additionally, understanding how variants are linked enables imputation, and therefore a high-quality set of phased variants can serve as a catalyst for much larger volume experiments, creating greater statistical power for disease association (Marchini and Howie 2010; Peterson et al. 2019). Population genetics also benefits from haplotype information because it provides a means to estimate recombination and gene flow (Song et al. 2017). As a consequence of its many applications, a significant portion of recent efforts in human genomics have become focused on generating a high-quality, genome-wide set of phased variants (Altshuler et al. 2005; Chin et al. 2020; Wang et al. 2022).

Methods for phasing are diverse and can use information from populations or from an individual's sequencing data. At the population level, variants are associated with one another by their co-occurrence across many individuals (Altshuler et al. 2005; Browning and Browning 2007, 2011; Howie et al. 2009; Byrska-Bishop et al. 2022). At the individual level, variants are phased based on co-occurrence in spanning reads. Individual-level methods can be further subdivided into approaches that rely on mapping to an existing assembly and those that use reads directly for creating a phased consensus. When read length or accuracy is limited, mapping-based methods are essential, because mapping is required to find a set of candidate variants that share reads among them. After mapping, reads are usually phased by finding a partition that maximizes the consistency of shared reads among the alleles (Patterson et al. 2015; Edge et al. 2017; Ebler et al. 2019).

In contrast to reference-based methods, de novo methods for read-based phasing generate candidate variants internal to the assembler, following read overlap (Cheng et al. 2021; Rautiainen et al. 2023). The advantage of assembly-based methods is that they do not fall victim to reference bias and can therefore identify variants that would otherwise map poorly, as with repetitive regions or large duplications and inversions (Brandt et al. 2015; Günther and Nettelblad 2019). Reference-based methods can work around this issue by using a draft assembly instead of an

Corresponding authors: rlorigro@ucsc.edu, pacarnev@ucsc.edu, bpaten@ucsc.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278268.123.

34:454–468 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/24; www.genome.org

existing reference. When parental sequencing data are available, the strategy of trio binning simplifies the problem by partitioning the reads using exact subsequences (*k*-mers) from parental data. However, its applications have been limited by the added need for parental information and the inability of *k*-mers to adequately label repetitive or homozygous regions that have few haplotype-specific *k*-mers (Koren et al. 2018).

Despite the variety of phasing methods and the high demand for phased genomes, all the Oxford Nanopore Technologies (ONT)–specific genome assemblers evaluated at the time of Shasta's original publication (three years before this work), including Shasta (Shafin et al. 2020), produced collapsed, haploid assemblies. Some assemblers provided pseudohaplotypes or alternative contigs that were not explicitly phased or evaluated as such. Since then, none of the remaining ONT-based assemblers have adapted their methods to generate true diploid assemblies. Instead, as a result of the emergence of Pacific Biosciences (PacBio) HiFi reads (Wenger et al. 2019), modern diploid assemblers have used shorter, more accurate reads to infer haplotype-specific overlaps or build a large-*k* de Bruijn graph (Cheng et al. 2021; Rautiainen et al. 2023). In the case of Verkko, ONT reads are used to resolve regions of the genome graph that remain tangled after construction with HiFi.

To date, there is only one published example of a nanopore assembler that produces phased haplotypes (Luo et al. 2021), but it does not run to completion on our ultralong ONT data. An earlier paper concluded that nanopore de novo phasing was not practical at the time (Duan et al. 2022). To address this, we present the results of continued development on the Shasta assembler and its new "mode 2" of assembly, capable of using R9 or R10 ultralong nanopore reads to phase variation observed in its sequence graph. Following the overlap stage of assembly, Shasta methods and data structures have been replaced in order to explicitly model sequence "bubbles" or local regions of heterozygous variation and their correlation to one another (see Methods).

Because the basis of assembly phasing lies in the overlapping of reads, highly repetitive or homozygous regions limit phasing in Shasta and other assemblers, such as Hifiasm and Verkko (Cheng et al. 2021; Rautiainen et al. 2023). Although trio-based haplotagging of partially phased assemblies produces accurate and global phasing, parental sequencing is not always feasible, and for a variety of species, it is essentially impossible. To address these shortcomings, proximity ligation data have been used for extending phasing beyond the length of a typical read without the need for parental sequencing (Selvaraj et al. 2013). Sequencing technologies like Hi-C and Pore-C exploit the physical packing of chromatin in the nucleus to ligate proximal regions of DNA molecules, which can be hundreds of millions of base pairs apart along the chromosome (Lieberman-Aiden et al. 2009; Deshpande et al. 2022), exceeding the longest nanopore reads observed (Payne et al. 2019). Pore-C is of particular interest because it does not need a separate Illumina sequencing machine, and as a result of its protocol, it produces many more proximity-based contacts compared with Hi-C at the same coverage, while also not requiring parental data.

The approach described here leverages the assembly graph to phase and extend partial haplotypes. In this scheme, variants are not inferred from the read alignments. Instead, graph topology or sequence homology is used to identify large-scale haplotypic bubbles in the graph, and the information from proximity linkages is used to phase bubbles relative to each other. To find a partition of haplotypes that is consistent with the proximity information, this work uses a new variation of the stochastic optimization methods previously described (Selvaraj et al. 2013; Cheng et al. 2021). Once phases are inferred, chaining can then make use of the information stored in the edges of the graphical fragment assembly format (GFA) representation of the assembly graph to achieve a similar result to scaffolding algorithms (Burton et al. 2013; Putnam et al. 2016). Our proximity-based phasing methods are evaluated on Shasta, Hifiasm, and Verkko graphs, using both Hi-C and Pore-C data, showing flexibility and reusability. In addition, we show results comparing proximity ligation libraries, which are produced as part of our automated workflow. GFAse also provides a means to do parental *k*-mer haplotagging using the succinct variation graph that Shasta produces (see Methods).

Using nanopore sequencing for both long reads and proximity ligated reads (Pore-C), we aim to show a previously undescribed single-sequencer pipeline that is a logistically simpler alternative to hybrid and HiFi-based approaches, while attaining comparable accuracy. We compare the output of this pipeline to a variety of other hybrid methods that use Hi-C, trio Illumina, and PacBio HiFi data types. In addition to this, we aim to evaluate the tradeoffs of quality and cost-efficiency using a series of experiments. For the upper limit of cost-efficiency, we test a single flow cell (FC) of PromethION R10, and for the upper limit of assembly continuity, we evaluate high-coverage, ultralong reads as input. These data points may provide a reference for future projects, which must choose from many possible combinations of sequence inputs and software.

Results

Phased assembly generally has multiple steps for which input type and choice of software or configuration are key to generating a useful product (Fig. 1A). For input data quality, we separately evaluate conventional long reads and data types used for additional phasing (Fig. 1B). For completeness, we compare proximity-ligated reads to the standard of trio Illumina phasing. For evaluating the output of each module, we have primarily used the HG002 benchmark human genome, and we performed supplementary analysis involving HG005 and four other diverse samples from the HPRC Year 1 data release. We also show phasing in diploid non-human species in two different organisms.

Long-read sequencing

To address variability in read length, coverage, and accuracy, these results evaluate phasing in six different combinations of library preparation and chemistry. The effect of read accuracy on phasing is addressed with differing nanopore chemistries: R9 and R10. For the data sets evaluated, R9 has median accuracy of 95.7%–96.1% (Fig. 2A), and R10 has 98.3%–98.9% median accuracy. In ONT R9, three different length distributions and coverages were assembled. The R9 data sets have minimum read lengths of 10 kbp, 35 kbp, and 100 kbp, respectively, so they have been labeled "standard," "ultralong" (UL), and "ultra-ultralong" (UUL) for convenience. Nanopore data sets vary considerably in length characteristics, so cumulative length distributions are plotted for proper comparison (Fig. 2B).

For R10, a similar series of length distributions are used, with one key difference: The lowest coverage assembly used only a single PromethION flow cell (Fig. 2B, labeled 1FC). For that data set,

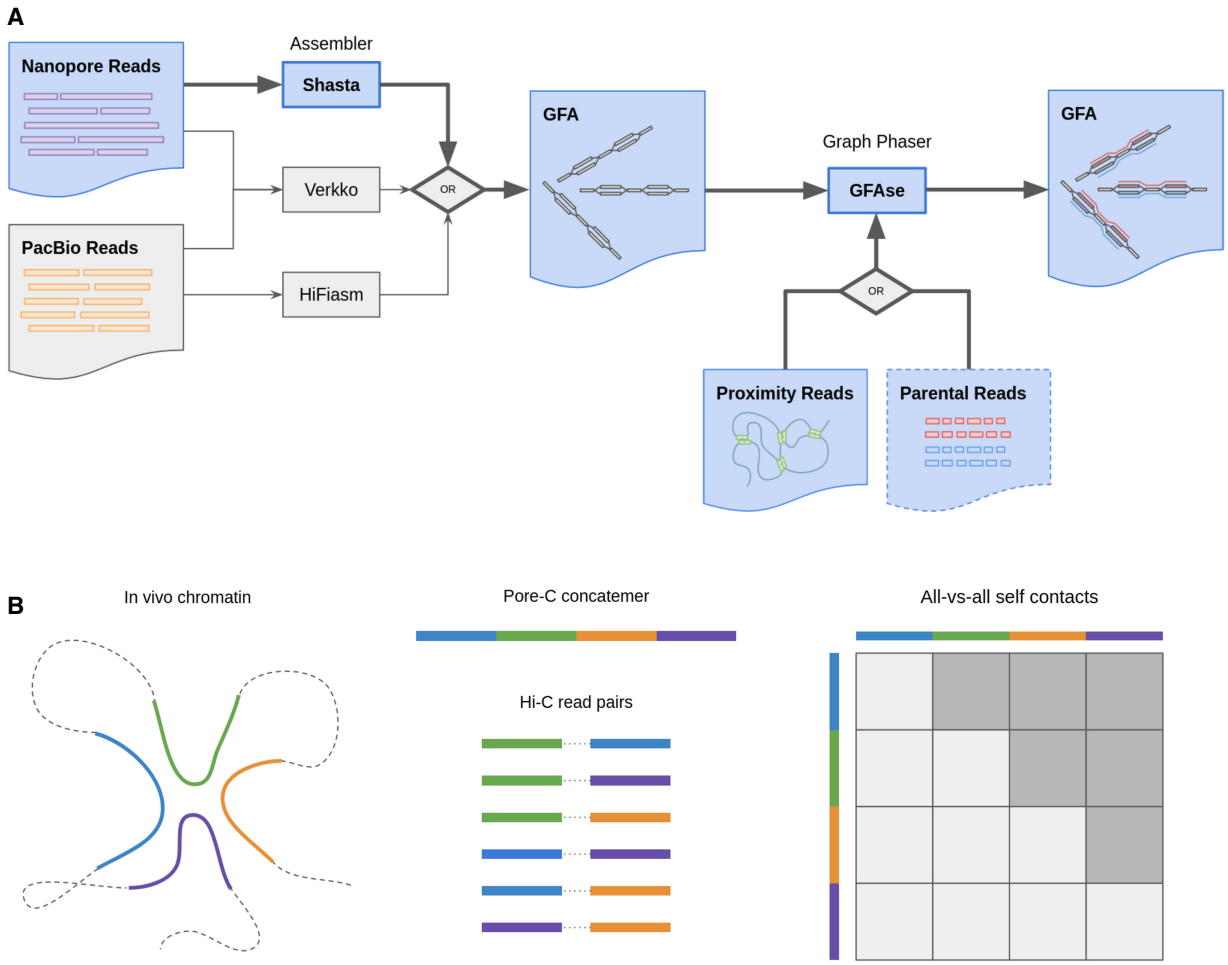

Figure 1. Summary of de novo phasing pipeline using Shasta and GFAse and input proximity ligation data types Pore-C and Hi-C. (*A*) Shasta performs de novo assembly and phases to the extent that is supported by informative variants in the nanopore reads. GFAse then takes a partially phased assembly GFA and extends phasing using orthogonal phasing information. GFAse can perform phasing based on any alignable data type (Hi-C, Pore-C, etc.). For Shasta graphs, GFAse can also use parental sequencing. The pathways with bolded arrows and blue fill are the methods that are previously undescribed. (*B*) Diagram of Pore-C sequencing in comparison to Hi-C. In the all-versus-all contact matrix, shaded squares represent usable contacts, which scale at a rate of $1/2n^2 - n$ for a concatemer of n fragments or subreads.

reads were sheared to maximize throughput. In the R10 UL data set, four flow cells of unsheared DNA were combined to create a data set with 60× coverage, reaching an approximate minimum read length of 50 kbp. Finally, in an attempt to maximize contiguity and find the limits of our methods with these data, we have also assembled a data set from 11 flow cells and down-sampled to 60×, resulting in an effective minimum read length of 165 kbp.

Pore-C and Hi-C sequencing

An early version of the Pore-C protocol was performed and provided by ONT for use in this work and compared with existing Hi-C data sets from the HPRC year 1 data freeze. Notably, these Pore-C libraries do not have a larger modal alignment length than Hi-C (Fig. 2C). The difference between Pore-C's multicontact concatemers and Hi-C's one-to-one paired ligation are shown using alignments. Pore-C concatemers are composed of many smaller reads, or "subreads," analogous to Hi-Cs paired reads. As a result of having many subreads, each Pore-C read has many alignments, and its contacts accumulate in an all-versus-all manner among subreads. The number of subreads (usable for phasing) can be in the dozens (Fig. 2D). In contrast, Hi-C accumulates at most one long-range contact per pair of reads, and in many cases, read pairs contain unmappable reads, which drastically reduces its throughput.

To show the practical differences between Pore-C and Hi-C that are particularly relevant to phasing, these results also show the signal ratio (Fig. 2E) and the total number of contacts (Fig. 2F) for reads that map to a diploid reference. Because mapping quality is used to filter contacts during phasing, the spectrum of signal ratio and the number of contacts are plotted across observed map qualities. Signal ratio is computed using a trio-phased diploid reference to estimate the number of consistent and inconsistent contacts (*cis* or *trans* w.r.t. haplotypes). In summary, contacts from Hi-C consistently have a higher signal ratio, whereas Pore-C produces more contacts.

Assemblers evaluated

For phasing and assembly quality evaluation, we compare to widely used pipelines for diploid assembly that use ONT reads, PacBio CCS,

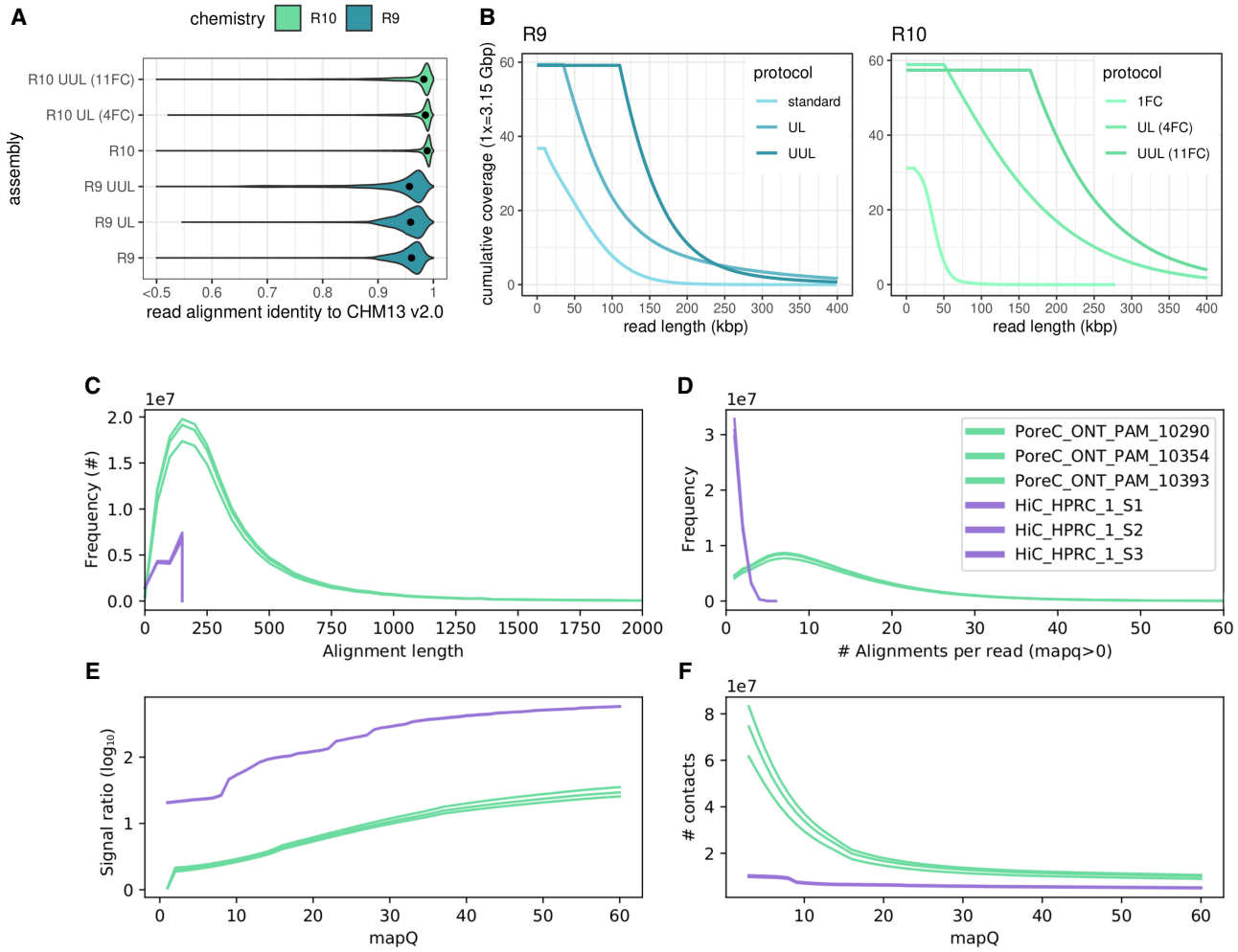

Figure 2. Distributions of read accuracy, coverage, length for reads used in assembly phasing, and phasing signal for proximity-ligated reads. (*A*,*B*) Identity and length metrics for nanopore read sets used in the HG002 evaluation. (*C*–*F*) Pore-C and Hi-C metrics for contacts and signal ratio, measured on a per-library basis. "Alignment length" and "alignments per read" are proxies for subread statistics. Only mappings that are usable for phasing are shown, that is, with mapping quality (mapQ) > 0 in a diploid reference. Signal ratio is computed using a high-quality trio-phased assembly to indicate the number of consistent and inconsistent contacts (see Methods).

or both. For PacBio CCS and hybrid CCS/ONT assemblers, we have included Hifiasm and Verkko (Table 1). As the name implies, Hifiasm is an assembler for PacBio CCS (HiFi) data that has built-in methods for trio and Hi-C phasing. When comparing Shasta to Hifiasm, the relative strengths of PacBio and nanopore become apparent. GFAse is also compared with Hifiasm's native phasing methods to evaluate GFAse's performance. Verkko is used in this comparison as an upper limit, because it uses both high-coverage CCS and ONT reads to generate its assemblies. In this comparison, Hifiasm uses 30× coverage CCS and 30× coverage Hi-C. Verkko "production" assemblies use 42× CCS and 70× ONT >100 kbp, as well as trio Illumina. The "full-coverage" Verkko assembly uses 190× CCS.

For nanopore, HapDup has recently developed a combination of alignment-based and assembly-based methods for phasing long reads (Kolmogorov et al. 2023). HapDup uses a linear unphased Shasta assembly as a starting point, phases aligned reads, and then generates a reference-free consensus for each haplotype. HapDup is specialized for structural variant (SV) detection in low-coverage nanopore data sets, so it is a natural comparison point for phased Shasta assemblies.

Phasing results

To evaluate phasing accuracy, the assemblies presented are aligned to a common reference, and their heterozygous alleles are compared using WhatsHap to an orthogonally phased truth set, produced by NIST's Genome in a Bottle (Zook et al. 2020) consortium. Switch rate indicates how often alleles in the sample switch phase relative to the truth set, and hamming rate indicates the proportion of switched loci. Genotypes with allele sequences that do not both exactly match the reference VCF are not evaluated for phasing, which is accounted for by reporting the number of variants covered.

GFAse trio uses *k*-mers from parental Illumina short reads to phase the heterozygous bubbles in the child Shasta assembly graph. Phasing results show that Shasta + GFAse trio using Illumina reads outperforms Hifiasm trio and Hifiasm Hi-C in terms of median switch rate (~0.0005) and hamming error (~0.0005). Shasta + GFAse trio results are also within range of the Verkko trio "production" assembly that uses both ONT and PacBio input reads. Two chromosomes, Chr 15 and Chr 16, have higher

Table 1. Coverage summaries for HG002 assemblies evaluated in this analysis

		ONT		CCS		
Assembler	Label	Coverage	N50 (kbp)	Coverage	N50 (kbp)	Hi-C/Pore-C Coverage
HapDup v0.4 (Shasta)	HapDup	38	31			
Shasta v0.10.0	R9 standard	37	60			45/30
Shasta v0.10.0	R9 UL	60	80			45/30
Shasta v0.10.0	R9 UUL	60	150			45/30
Shasta v0.10.0	R10 (1FC)	26	32			45/30
Shasta v0.10.0	R10 UL (4FC)	60	130			45/30
Shasta v0.11.1	R10 UUL (11FC)	~60	230			45/30
Hifiasm v0.16 trio	Hifiasm trio			30	17.5	
Hifiasm v0.16 Hi-C	Hifiasm Hi-C			30	17.5	0/30
Verkko v1.1 trio	Production	185.77	81.20	42.66	14.75	
Verkko v1.1 trio	Full coverage	971.71	50.65	169.03	17.22	

hamming rates across all the R10 Shasta assemblies, which may be owing to the difficulty of resolving these chromosomes and identifying inversions relative to the hg38 reference. This method requires short-read sequencing of both parents, which, for various reasons, may not always be feasible or cost-effective. For the more contiguous Shasta assemblies, trio results closely match the phasing results from Hi-C and Pore-C phasing, which do not require any parental sequencing.

Shasta + GFAse assemblies consistently outperform native Hifiasm Hi-C phasing, and in some cases, hamming rates match or beat trio-phased Hifiasm. When comparing the total number of assessed variants, R10 assemblies drastically improve on R9, likely as a result of its lower error rate. This effect can be seen by comparing the standard-length results to UL in R9 and R10. It is also evident from the polished HapDup assembly that polishing has a drastic effect on covered variants, because it also uses a single flow cell R9 Shasta assembly. UL R10 assemblies reach nearly as many variants covered as Verkko while also producing comparable switch and hamming rates. In a direct comparison of Hifiasm's internal Hi-C phasing versus GFAse Hi-C postphasing, the GFAse switch rate is lower, but hamming rate is slightly higher (Supplemental Fig. S1).

Yak trio eval, an orthogonal evaluation method that uses parental short-read *k*-mers to evaluate phasing accuracy, reports an order of magnitude higher switch and hamming error for Shasta assemblies compared with Verkko (Supplemental Table S2). This could be from a combination of greater genome coverage and systematic false positives in the switch analysis. As one possible explanation for the discrepancy between yak and WhatsHap, we observed that some of the yak switch blocks occur in regions of the Shasta assembly that differ from Verkko only in homozygous variants (Supplemental Fig. S4A), and this is particularly evident in Chr X of the male HG002 assembly (Supplemental Fig. S4D). As an additional source of enrichment of errors, we observe that some yak switches occur in repetitive regions that contain homozygous homopolymer errors (Supplemental Fig. S4C). When comparing Shasta Dipcall VCFs directly to Verkko Dipcall VCFs, we observe a variant coverage of 2.48 million at a hamming rate of 0.0003, but the total variants from Dipcall differ significantly in assembly specific insertions and deletions (indels) (Supplemental Table S3).

As evaluated by WhatsHap, GFAse consistently produces a median chromosomal hamming error of <0.001% for Shasta assemblies, which is reduced from an expected chromosomal hamming of ~50% for nonglobally phased assemblies. In the single flow cell, standard-length R10 experiment, R10 produces shorter haplotype bubbles compared with R9, which are then phased together at the chromosome scale with proximity ligation data (Figs. 3, 4D). In the high-coverage UL R10 assembly, a large portion of the variants exist in continuous haplotypes directly from the Shasta assembler (Fig. 4E). Generally, the graphs with high input N50 reach trio-level phasing accuracy when phased with Hi-C. The highly fragmented input graphs such as the Shasta "standard" R9 or the Hifiasm PacBio graphs converge to a less optimal phasing.

In addition to this analysis, a limited number of HG005 R9 and R10 assemblies were evaluated (see Supplemental Fig. S2; Supplemental Table S1), along with four HPRC R9 assemblies (Supplemental Fig. S3). A similar trend in phasing accuracy is observed across all additional individuals, with the caveat that there is no Strand-seq-derived truth set for the HPRC assemblies.

By comparing Pore-C and Hi-C (Supplemental Fig. S1), it is evident that more information is provided by one library of Pore-C than by Hi-C. It takes three pairs (6 × 2 lanes, or ~45×) of Hi-C to reach a phasing that is roughly equivalent to 1FC Pore-C (30×). This is likely because of the higher number of total contacts per Pore-C read. In combination with the single flow cell R10 assembly, we achieve an accurately phased human assembly with a total of 2 PromethION flow cells.

Assembly quality

SVs were evaluated with Truvari and the GIAB HG002 Tier1 SV collection as a truth set. Shasta nanopore assemblies consistently yield SV F1 scores >90%, with a peak score of >95% in the UL 4FC and 11FC R10 assemblies. Most notably, a single flow cell of R10 400-bp nanopore data can reach an F1 score of ~94% (Fig. 4A). HapDup (Kolmogorov et al. 2023), which starts with Shasta assembly, uses a local realignment and polishing to recover short collapses, which is a probable cause for its greater recall. Precisions from R10 assemblies match or exceed HiFi-based methods, but recall values do not, most likely as a result of collapse in the repetitive regions of the genome.

A similar trend is seen in the gene-level analysis performed by asmgene (Fig. 4B), in which the number of full single-copy genes is

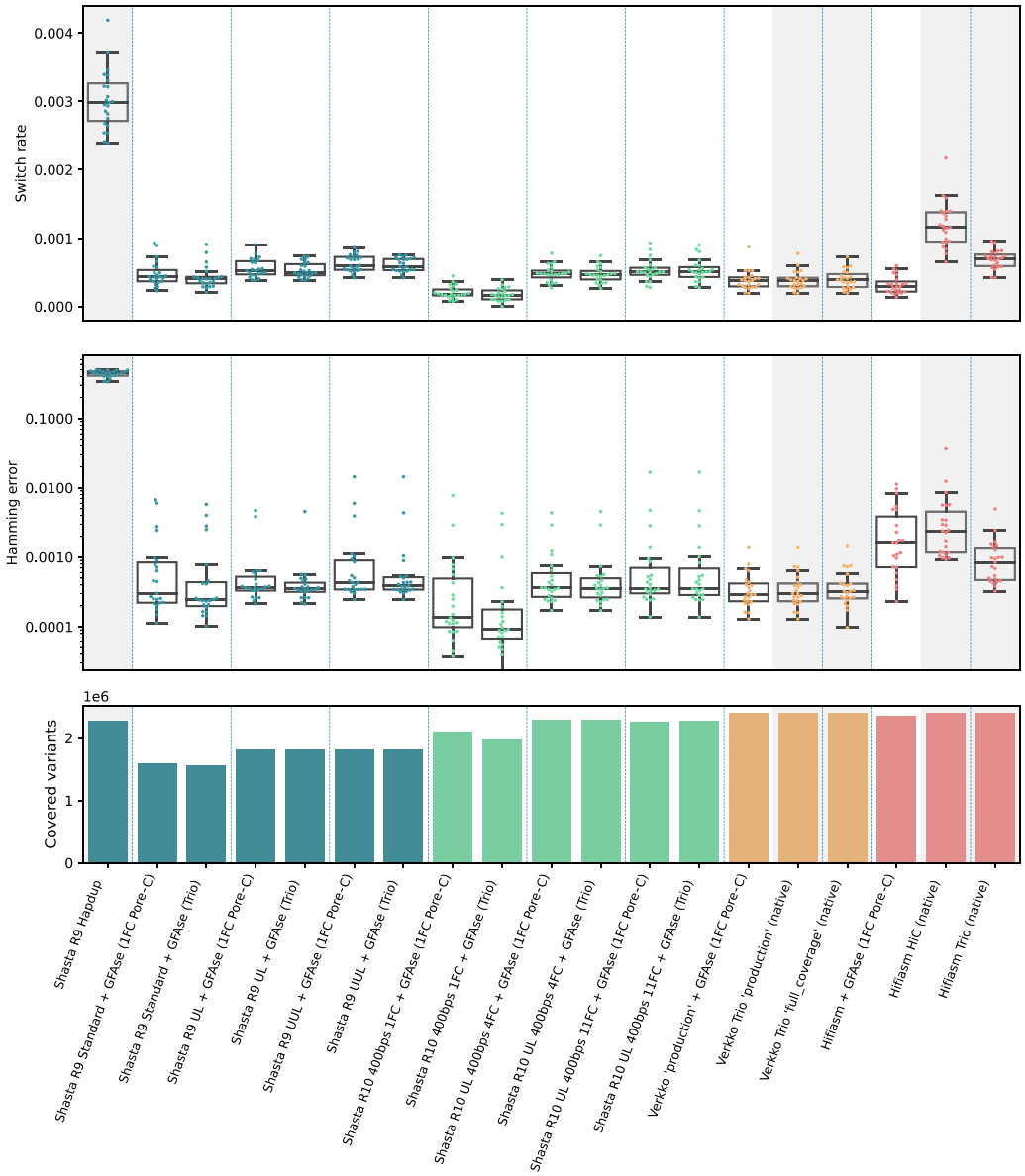

Figure 3. Phasing metrics for HG002 assemblies, as evaluated using the GIAB v4.2.1 benchmark VCF, phased with Strand-seq using WhatsHap (see Methods). All Shasta assemblies are unpolished. Assemblies not phased with GFAse are shaded gray. Each dot represents a chromosome error rate, generated by WhatsHap compare. Native Hifiasm Hi-C uses 30× coverage. Each pair of Hi-C is ~17×. Pore-C flow cells have ~30× yield.

comparable to that of HiFi or hybrid assemblies, but multicopy genes are reduced in comparison. For gene completeness, the highest scoring Shasta assembly reached ~75% multicopy genes at a 99% identity threshold, which is now 10% lower than Hifiasm. In terms of base-level quality, R9 assemblies have base qualities greater than Q30, whereas R10 assemblies are greater than Q40 and PacBio HiFi or hybrid assemblies exceed Q50 (Fig. 4C).

Shasta assemblies are chained and unzipped by GFAse to achieve greater continuity. In theory, only 0.1% of the assembly would need to be assembled as diploid bubbles to produce a fully phased assembly after unzipping, but in practice, mapping and optimization with proximity ligation reads are easier when haplotypes are long. Before chaining, bubble N50s range from <1 Mbp, in the low-coverage experiments, up to 39.7 Mbp, in the highest quality assembly. For the various R9 assemblies, the chained and unzipped assemblies remain consistent in length despite their variable input lengths, which is shown by their largely overlapping post-GFAse NGx distributions (Fig. 4D,E). R10 sequencing protocols are currently limited by throughput, so sheared reads, which do not enable the same level of contiguity as the UL assemblies, were used for the single flow cell experiment. However, for the higher-cost 4FC and 11FC UL experiments, the upper limit on contiguity exceeds our previous most-contiguous R9 assemblies.

Non-human assemblies

To test Shasta's phased assembly methods outside the context of human genomes and base-calling, two species were assembled: the dwarf cuttlefish (*Sepia bandensis*) and the broad bordered

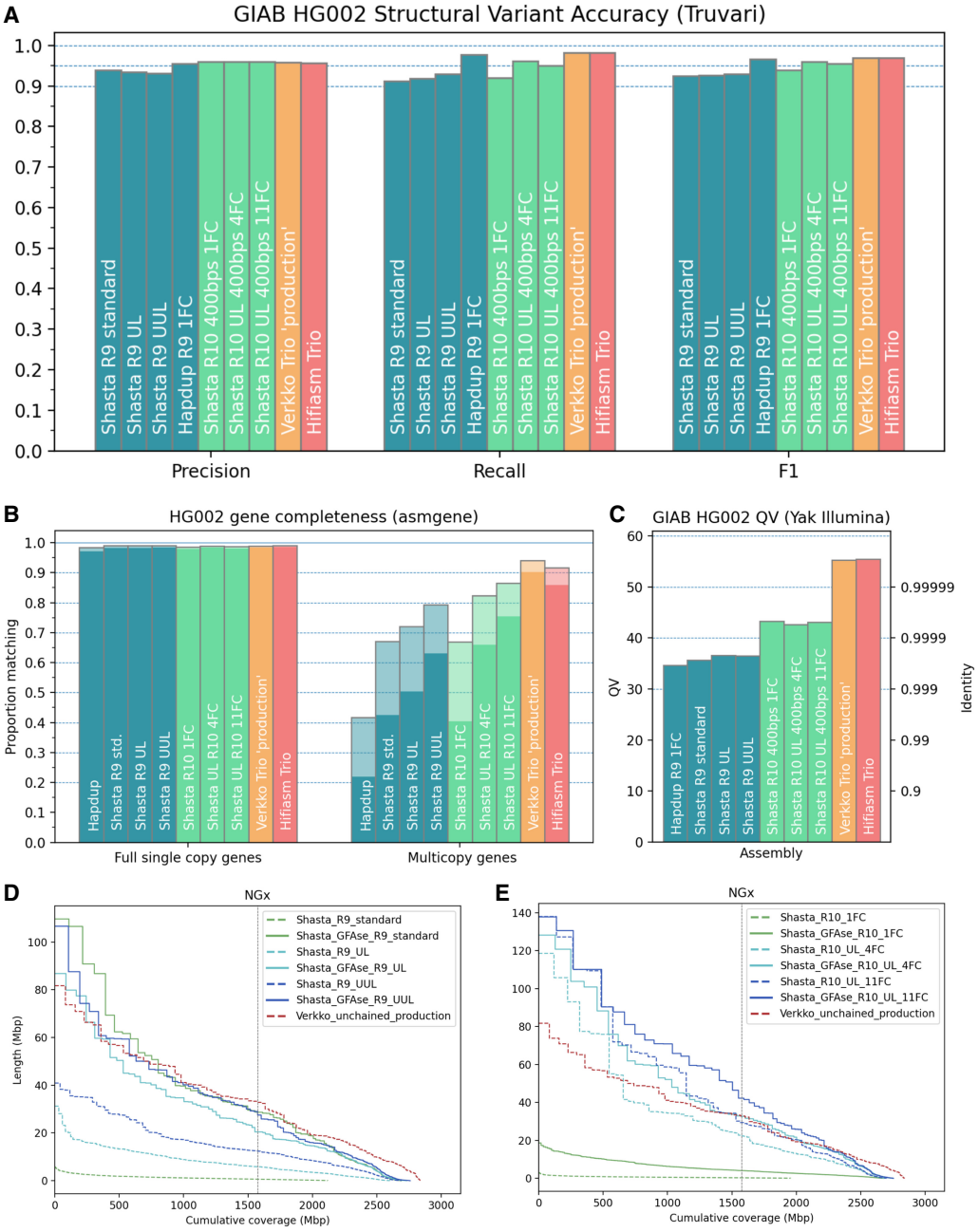

Figure 4. Structural variant, base-level, and gene-level accuracy metrics for HG002 assemblies. (*A*) Base accuracy evaluated using yak with Illumina NovaSeq. (*B*) Gene completeness measured by asmgene using human transcript sequences. "Full single copy" genes only indicate unfragmented, non-duplicated genes, matching transcripts by ≥99% coverage and stratified by >97% (translucent) or >99% identity (opaque). Multicopy genes are similarly stratified. (*C*) SVs evaluated using the GIAB Tier1 benchmark VCF with Truvari. (*D*,*E*) NGx plots for Shasta haplotypes, before and after unzipping bubble chains with GFAse. For comparison, the phased portion of the unchained Verkko "production" assembly is shown. The vertical line indicates the NG50 for each assembly.

yellow underwing moth (*Noctua fimbriata*). *S. bandensis*, previously unsequenced by long reads, was sequenced for this work to a depth of ~105× with 30× at >100-kbp length. *N. fimbriata* was previously sequenced by the Darwin Tree of Life (Holland et al. 2021) to a depth of 87× with an N50 of 28.7 kbp.

In non-human assemblies, truth sets are limited, so this analysis relies on BUSCO to evaluate gene completeness and extent of phasing (Table 2). By comparing the BUSCO score for the full diploid assembly, as well as one haplotype of the diploid assembly, the number of phased and unphased genes can be inferred. In both species presented, complete phased genes are estimated to range from 86% to 89% of the 954 genes in the metazoan data set.

Resource usage

For the slowest assembly evaluated (4FC R10 UL), Shasta runs in 12 h on a 64-thread 1.2-TB AWS instance. This allows for 14 assemblies to be run in the same amount of core hours as a single

Table 2. Non-human Shasta phasing metrics, in terms of BUSCO gene completeness

	Complete and single-copy (S)		Complete and duplicated (D)				
	Both haps	One hap	Both haps	One hap	Total diploid	Total haploid	Diploid N50
Sepia bandensis	5.97%	95.39%	90.15%	0.73%	5,330,371,146	415,993,994	3,794,550
Noctua fimbriata	7.02%	93.29%	91.40%	4.82%	488,722,705	21,756,252	5,094,850

These results use the metazoan data set, which has 954 genes. Percentages are calculated as the proportion of the metazoan genes found, as a single copy or as duplicated (according to BUSCO output), in the haploid or diploid assemblies.

Verkko assembly (Table 3). For unfragmented assemblies, GFAse has variable run time of 2.3 h to 4.6 h using 64 threads, which depends on the number of contacts in the alignments. In the worst-case scenario, with a highly fragmented GFA such as the unphased Hifiasm graph, as well as high-contact data set such as Pore-C, GFAse can take up to 12 h.

The most time-consuming step in GFAse is the phase optimization step, which has a complexity that depends on the number of nodes and edges in the contact graph. When running on an unlabeled GFA, the homology detection step incurs an additional cost, but the time consumed by this step is usually limited in comparison to phasing time, as homology detection primarily relies on locality-sensitive hashing (see Methods) to find candidate matches between nodes. Both of these steps are multithreaded, but the phasing step does not benefit from running with more threads than there are independent samples (random seeds) in the stochastic optimizer. The results in Table 3 are shown with 64 threads for simplicity, but the default number of samples is 30, which means that the phasing step would not lose performance until fewer than 30 threads are used, at which point a near linear increase in run time would be expected.

Discussion

In this work, we use the latest advances in nanopore sequencing to simplify the challenge of producing phased de novo assemblies. We show accurate phasing in a two PromethION flow cell pipeline, using one long-read flow cell and one Pore-C flow cell. Phased contiguities yielded by our higher coverage experiments are unmatched in previous nanopore assemblies, and Pore-C as a phasing data type is unexplored in prior publications. The tools presented are efficient and modular, and they rely on sequencing protocols that have a relatively short turn around (Shafin et al. 2020), enabling rapid prototyping.

We focus on nanopore sequencing because, despite its currently lower base-level accuracy relative to its competitors, its ability to sequence native DNA means that its upper read length is essentially limited only by the library preparation and loading procedure, which gives it a unique advantage over methods that rely on synthesis or amplification (Deamer et al. 2016). With recent changes in chemistry and computational methods, this tradeoff in accuracy has reduced, whereas cost and throughput have improved. In three years, since work by Shafin et al. (2020), R9 median read accuracy has increased from 90% to 95%, effectively reducing error by half. In the same time span, protocols for nanopore library preparation have improved average N50s from 42 kbp to >100 kbp, more than doubling. Now, with the R10.4.1 chemistry in production, we see yet another reduction in error, bringing accuracy into the 98%–99% range.

The alternative sequence type to nanopore, PacBio HiFi (CCS), has accuracies of >99.9% (Wenger et al. 2019), but its reads are size-selected, ranging from 10 kbp to 30 kbp. This means that they have the accuracy to distinguish copies of less-diverged repetitive units in the genome but not necessarily the length to span them. Hybrid assemblers have integrated both ONT and HiFi data to accommodate for this, in addition to parental ("trio") Illumina data or Illumina proximity ligation data (Hi-C) to assist with phasing. Hybrid methods have achieved unprecedented accuracy and contiguity by leveraging the strengths of three to four different sequencers; but as a result, they are also costly in terms of resources, logistics, and time.

Using technological advances in ONT sequencing, our results produce a notable improvement over the previous standards for nanopore phasing. Evaluation metrics for phased assembly are approaching that of the hybrid assemblers in gene completeness,

Table 3. Run time performance for various assemblers presented in this paper

	CPU hours	Estimated wall hours (64 vCPU)	Peak RAM (GB)
Verkko	8288.57	129.5	61
Hifiasm	366.9	5.7	150
HapDup	2425.0	24.0	624
Shasta (UUL)	582.4	9.1	1283
Read alignment	307.2	4.8	140
GFAse (Hi-C)	75.0	2.3	84
GFAse (Pore-C)	145.7	4.6	84
GFAse (trio)	0.58	0.58	30

Separate times shown for GFAse Hi-C and Pore-C as a result of its dependence on the number of contacts. For HapDup, CPU hours and wall hours show the sum of multiple steps in the HapDup pipeline, including an initial Shasta assembly, which used 96 cores. GFAse trio runs single-threaded.

contiguity, and phasing accuracy, showing promise for future single-sequencer pipelines. When considering that all Shasta assemblies generated for this paper are unpolished, it is clear that there is room for further improvement. Contiguity and completeness are likely to continue to grow proportionally with additional throughput of long reads. To make full use of the longest reads, further method development is underway to address repetitive regions, which defy the diploid assumption of our current methods. The consistent trajectory of ONT quality and throughput has motivated our work, and we aim to continue to adapt long-read assembly methods to future improvements.

From a software development perspective, the tools presented in this paper are written with the goals of modularity, interpretability, and flexibility of usage. The fully specified graph outputs of Shasta make it an ideal resource for downstream development and analysis, as is shown by the application of GFAse in this context. GFAse uses transparent and reusable data structures and, similar to Shasta, produces comprehensive outputs that describe the homology, proximity linkage, and inferred haplotype chains in the graph. Relying directly on alignments, GFAse is capable of using any data type for phasing that can be aligned to the assembly in BAM format. In theory, long reads, or even conventional linked reads, could be used as phasing information if their alignments span the unphased regions of an assembly. To maximize compatibility, GFAse also accepts a custom contact map as input. This makes it a flexible module for future applications, as long-range linked data types continue to evolve.

Methods

Shasta: de novo phasing with nanopore reads

Shasta is an assembler specialized for nanopore reads, and it uses the overlap layout consensus paradigm of assembly. It starts by reducing reads into vectors of fixed-length subsequences or markers, and then it computes an approximate overlap among reads using a variation of the MinHash algorithm (Broder 1997). Shasta refines its candidate overlaps using alignment in marker space, and reduces the overlap graph by filtering alignments and creating a *k*-nearest neighbor graph. For more details, see the initial 2020 publication or the online documentation (Shafin et al. 2020; https://paoloshasta.github.io/shasta/ComputationalMethods.html).

In this updated version, Shasta performs de novo phasing internally, using only conventional nanopore reads. Shasta uses a data structure referred to as a "phasing graph," built from a marker representation of the reads. The phasing graph is created following the overlap stage of assembly, and it describes the coverage in terms of read IDs covering each branch of a heterozygous diploid bubble found in a graph. This data structure has a strict diploid assumption, not suited for polyploid species. For each pair of bubbles, a Bayesian model computes the probability that they either are uncorrelated or are in one of two possible phased orientations with respect to each other. By iteratively aggregating bubbles with this Bayesian criteria for correlation, groups of phased bubbles are established, whereas uncorrelated error bubbles remain isolated.

Given a set of phased bubbles or a "component," Shasta then identifies local bubble chains within each set, which are bubbles in series, constituting collinearly traversable regions of the graph (https://paoloshasta.github.io/shasta/ComputationalMethods.html). These bubble chains have a strict topology in which elements in the chain can be either homozygous unphased nodes or heterozygous phased pairs of nodes (see Fig. 5). Bubble chains are the basis for subsequent unzipping into haplotypes. For completeness, Shasta generates output GFAs containing the succinct representation of edits, or "detailed" graphs (Fig. 5A), as well as the larger unzipped haplotype representation, or "phased" graphs (Fig. 5B). Both of these representations contain bubble chains with differing length haplotype sequences. The "phased" representation is convenient for downstream phasing because its sequences tend to greatly exceed the length of a read, and multiple variants can be spanned with conventional mapping. On the other hand, the "detailed" representation summarizes the edits between phased haplotypes and represents longer-scale phasing using a path in the GFA. Graphs generated by Shasta contain "blunt" or nonoverlapping nodes, which makes chaining them trivial.

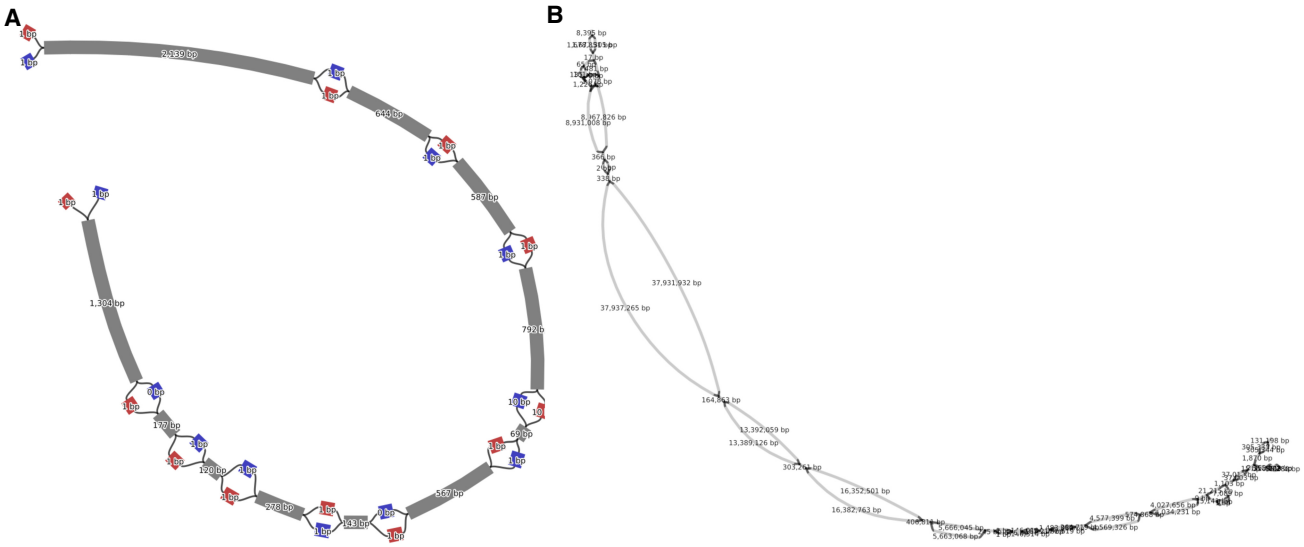

Figure 5. The two types of Shasta output graphs, visualized as a 2D layout in Bandage (Wick et al. 2015) at two different scales. (*A*) A subregion of the "Assembly-Detailed.gfa," showing near-variant scale nodes in a bubble chain and their phasing indicated by colors produced by Shasta. (*B*) A subregion of the "Assembly-Phased.gfa" showing a phased portion of Chr 11 from HG002, which terminates at two tangles, presumably caused by telomeric and centromeric sequences.

GFAse: phasing graphs with proximity ligation data

GFAse can ingest a custom contact map (as a CSV file) or process conventional mappings for phasing information (using BAM format). Hi-C, Pore-C, or other proximity-ligated reads are mapped to the GFA contigs using whichever mapper is most appropriate for the sequence type. The strength of the proximity linkage between any two contigs is updated as a sum, in the form of a weighted edge in a "contact graph," in which the weight is the number of reads linking them. Uninformative mappings that do not cover a heterozygous site are filtered by setting a map quality threshold.

To phase the graph, GFAse first identifies diploid, haplotypic bubbles. Two methods are available in GFAse: assembler annotation and sequence similarity search. Efficient similarity search is accomplished with a variation of MinHash (Broder 1997) similar to that used by Shasta (Shafin et al. 2020) and then refined with full-scale alignment with minimap2 (Li 2018). GFAse can use sequence similarity to recognize bubbles or bipartite subgraphs that do not have a strict bubble topology, but the optimizer has a strict diploid assumption because it partitions the nodes of the GFA into two sets.

Phases are optimized using a stochastic method that approximates a solution to the optimization variant of the max-cut problem (Selvaraj et al. 2013; Edge et al. 2017; Cheng et al. 2021). The method depends on an objective function that penalizes inconsistent contacts and rewards consistent contacts. If any two bubbles are compared, there are four possible contacts, and only contacts linking the contigs in matching phases have positive scores. For GFAse, a variation on existing methods was introduced to improve the reproducibility of the stochastic method and to perform better on fragmented graphs in which the state space is much larger. In short, the method takes samples from repeated greedy optimizations of randomly initialized phase states and accumulates a distribution of orientations, which is then used to merge bubbles that are most consistent (Fig. 6). One benefit of sampling many times with few iterations is that samples are independent and can be multithreaded.

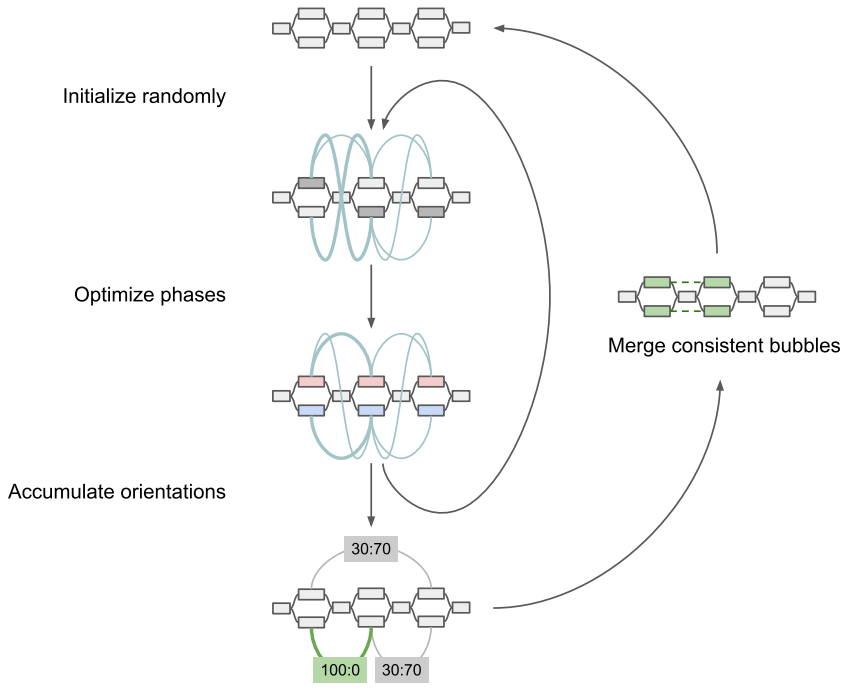

Figure 6. Diagram of sampling method for optimizing proximity linkages in an assembly graph. Edge weights in the contact graph are represented by teal curves. For each inner iteration, a greedily converged phase state is used to update a distribution of orientations among bubbles. Bubbles with the strongest signal at the end of sampling are merged for successive iterations. By the end of each round *r* of merging, the largest possible bubble set is 2 ^ *r* in size.

GFAse: phasing with parental data

Homozygous parental *k*-mers are selected from each parent and used to phase the "detailed" assembly GFA by counting parental *k*-mers in the heterozygous bubbles. To process the parental sequence data, reads are broken into 31-bp *k*-mers using kmc3 (Kokot et al. 2017). Kmc3 subtract was used to identify *k*-mers that are unique to each parent. Finally, unique homozygous *k*-mers are matched to child *k*-mers on heterozygous bubbles assigning a phase to bubble components, using a simple majority vote. Illumina reads for the HG002 Ashkenazi Jewish trio sample were obtained from the publicly available 1000 Genomes Project (fc-4310e737-a388-4a10-8c9e-babe06aaf0cf/working/HPRC_PLUS/HG002/raw_data/Illumina/parents/HG003 and fc-4310e737-a388-4a10-8c9e-babe06aaf0cf/working/HPRC_PLUS/HG002/raw_data/Illumina/parents/HG004).

GFAse: chaining phased graphs

With a phased assembly graph, adjacent bubbles are chained in a manner similar to scaffolding to extend haplotypes. GFAse first loads the GFA using the VG HandleGraph data structure (Eizenga et al. 2020) and identifies tractable regions as anything that follows a strict diploid bubble chain topology. Diploid nodes have exactly one two-hop neighbor and, at most, two direct adjacencies in each direction. Chains are then identified by traversing contiguous subgraphs of labeled nodes. With bubble chains identified, haplotypes are labeled with paths in the GFA formalism. Then they can be "unzipped" trivially by traversing them and duplicating the homozygous nodes into both haplotypes. The strict definition of bubble chains used in this method is intended to maximize fidelity to the input graph by reducing misjoins in the chaining step.

GFAse: input format

To ingest genomic contacts for use in phasing, GFAse accepts either a BAM or a CSV contact map. The BAM can be of any data type, so long as contacts in the genome share a common query name, for example, paired end, long read, Hi-C, Pore-C. The BAM must be grouped or sorted by query name and does not need to be indexed. The alternative CSV is a simple three-column file containing a header and any number of following lines that indicate the weight of an edge in the contact graph (details are in software help output).

GFAse will run to completion on any GFA that meets the v1.0 specification; however, there are some practical considerations:

1. **Overlaps are arbitrarily resolved in the output unzipping step.** Either the left or the right overlap

sequence is arbitrarily chosen as the joining sequence in a phased path.

2. **The cumulative consumed query/target length of a GFA overlap cigar must not exceed the length of the query/target node.** This is not possible in a properly formatted GFA, but it has been observed in both hifiasm and Verkko overlaps. If this input error is encountered, it is recommended to use the --skip_unzip argument.
3. **Edges (L lines) are not required.** For phasing an input that is not in graph form (e.g., in FASTA/Q format), it is possible to simply convert it to a GFA without any L lines, and invoke the --skip_unzip and --use_homology flags to bin the contigs into bin 1 or 2 or unphased. Homology will be inferred entirely independently of the graph edges.
4. **When using an alignment as input, nodes ("segments" or S lines) in the graph should be mappable.** Short nodes that are insufficient in length to map a Hi-C or Pore-C subread (≥150 bp of nonrepetitive sequence) will not accumulate any contacts in the mapping step. Nodes without contacts will not be phased, and GFAse will put them in the "unphased" bin instead of using them to extend a phased chain. In addition to this, when using the homology-based alt detection, nodes must be long enough to confidently map to each other. We recommend starting with an assembly configuration that produces a phase N50 as large as possible without introducing any switch errors. For GFAs in which nodes are not mappable, the user may provide a custom contact map CSV as input.
5. **Paths (P lines) in the GFA are ignored.** Input paths do not influence the phasing step and are not written to any output file. However, GFAse does generate an intermediate GFA that contains additional P lines describing the result of phasing and chaining.

Sequencing and data acquisition

Links to sequences used this work are available at Zenodo (https://doi.org/10.5281/zenodo.10653823).

R9 standard

Sequencing was performed as described by Shafin et al. (2020) and rebasecalled with Guppy v5.0.7.

R9 UL

ONT data from the GIAB consortium (Zook et al. 2016) were rebasecalled with Guppy > v5.0 and combined with the R9 standard data set to provide longer reads.

R9 UUL

DNA extractions from 6 million cells of HG002 were prepared using the Circulomics nanobind CBB kit (PacBio 102-301-900). Libraries were prepared using the ultralong DNA sequencing kit (SQKULK001). The libraries were sequenced on a flow cell R9.4.1 on PromethION for 72 h. Flow cells were washed using the flow cell wash kit (EXP-WSH004) every 24 h. Fresh libraries were loaded after each wash.

R10

DNA extractions from 5 million cells of HG002 were prepared using the PacBio Nanobind CBB kit (102-301-900) according to the UHMW DNA extraction cultured cells protocol (EXT-CLU-001). Standard pipette tips were used to generate a homogenous sample before DNA shearing. DNA was sheared to a target size of 50 kb on Megaruptor 3. Samples were normalized to a 100 µL volume at 50 ng/µL concentration and sheared at speed 27 using the Megaruptor shearing kit (E07010003). DNA size was assessed after shearing on Agilent femto pulse system using the gDNA 165-kb analysis kit (FP-1002-0275). After shearing, DNA size selection was performed using PacBio SRE kit (102-208-300) following the manufacturer's recommendations. Libraries were prepared using the Oxford Nanopore ligation sequencing kit V14 (SQK-LSK114) according to the Oxford Nanopore protocol GDE_9161_v114_revK_29Jun2022. The libraries were sequenced on a flow cell R10.4.1 on PromethION for 96 h. Flow cells were washed using the flow cell wash kit (EXP-WSH004) every 24 h. Fresh libraries were loaded after each wash.

R10 UL

Nanopore sequencing data sets were generated following Oxford Nanopore protocol ULK_9177_v114_revC_27Nov2022. DNA extractions from 6 million cells of HG002 were prepared using Monarch HMW DNA extraction kit for tissue (New England Biolabs T3060). Libraries were prepared using ultralong DNA sequencing kit (SQKULK114). The libraries were sequenced on flow cell R10.4.1 on PromethION for 72 h. Flow cells were washed using the flow cell wash kit (EXP-WSH004) every 24 h. Fresh libraries were loaded after each wash.

Sepia bandensis

Testes tissue (3–5 mg) from an adult male dwarf cuttlefish (*S. bandensis*) was homogenized in PBS using a Dounce homogenizer. This was followed by DNA extraction using the Circulomics nanobind tissue kit (PacBio 102-302-100). Libraries were prepared using the ultralong DNA sequencing kit (SQKULK001). The libraries were sequenced on a flow cell R9.4.1 on PromethION for 72 h. Flow cells were washed using the flow cell wash kit (EXP-WSH004) every 24 h. Fresh libraries were loaded after each wash.

Noctua fimbriata

ONT data were acquired from the Darwin Tree of Life project (Holland et al. 2021).

Generating assemblies

To generate nanopore assemblies, Shasta (v0.10.0 unless otherwise specified) was run with the appropriate configuration for each data type, as follows.

R9 standard

```
--config Nanopore-Phased-May2022
--Reads.minReadLength 10000
--Assembly.mode2.phasing.minLogP 30
```

R9 UL

```
--config Nanopore-UL-Phased-May2022
--Reads.minReadLength 10000
--Reads.desiredCoverage 180000000000
--Assembly.mode2.phasing.minLogP 50
```

R9 UUL

```
--config Nanopore-UL-Phased-May2022
--Reads.minReadLength 110000
--Assembly.mode2.phasing.minLogP 50
```

R10 (1FC)

--config Nanopore-Phased-R10-Fast-Nov2022
--Assembly.mode2.phasing.minLogP 20

R10 UL (4FC)

Configs for *Nanopore-Phased-R10-Fast-Nov2022* and *Nanopore-UL-Phased-May2022* were merged, with *Nanopore-Phased-R10-Fast-Nov2022* taking precedence for any conflicting parameters. The following parameters were then added:

--Assembly.mode2.phasing.minLogP 20
--Reads.minReadLength 50000

R10 UUL (1IFC; Shasta v0.11.1)

--config Nanopore-Phased-R10-Fast-Nov2022
--Kmers.probability 0.05
--MinHash.minBucketSize 20
--MinHash.maxBucketSize 60
--Align.minAlignedMarkerCount 2500
--Reads.minReadLength 170000

Sepia bandensis

--config Nanopore-UL-Phased-May2022
--Reads.desiredCoverage 400G

Noctua fimbriata

--config Nanopore-Phased-May2022

Phasing with GFAse

To phase with GFAse (https://github.com/rlorigro/GFAse), reads are first aligned to contigs using conventional mapping and alignment. Paired Hi-C reads were aligned using BWA-MEM (https://github.com/lh3/bwa), and Pore-C concatemers were aligned using minimap2 (https://github.com/lh3/minimap2). A WDL that combines these steps can be found at GitHub (https://github.com/meredith705/gfase_wdl/tree/main).

Pore-C data were aligned using minimap2 with the following parameters:

```
minimap2 \
-a \
-x map-ont \
-k 17 \
-t 56 \
-K 10g \
-I 8g \
Assembly-Phased.fasta \
porec_reads.fastq.gz \
| samtools view -bh -@ 8 -q 1 - \
> porec_to_assembly.sorted_by_read.bam
```

Hi-C data were aligned using BWA with the following parameters:

```
bwa index Assembly-Phased.fasta \
&& \
bwa mem -t 46 -5 -S -P \
Assembly-Phased.fasta \
HG002.HiC_1_S1_R1_001.fastq \
HG002.HiC_1_S1_R2_001.fastq \
| samtools sort -n -@ 24 - -o hic_to_assembly.sorted_by_read.bam \
```

Shasta assemblies were phased with Hi-C using the following parameters:

```
/home/ubuntu/software/GFAse/build/phase_contacts_with_monte_carlo \
-i hic_to_assembly.sorted_by_read.bam \
-g Assembly-Phased.gfa \
-o /path/to/output/directory/ \
-m 1 \
-t 62
```

Shasta assemblies were phased with Pore-C using the following parameters:

```
/home/ubuntu/software/GFAse/build/phase_contacts_with_monte_carlo \
-i porec_to_assembly.sorted_by_read.bam \
-g Assembly-Phased.gfa \
-o /path/to/output/directory/ \
-m 3 \
-t 62
```

Verkko assemblies need the parameters "--skip_unzip" and "--use_homology" in addition to the above parameters, because bubbles are not labeled, and its homopolymer decompressed GFA has incorrectly specified overlaps that cannot be unzipped and stitched trivially. A map quality minimum of three is used for all Verkko assemblies.

Evaluation

Input data were evaluated using an alignment based QC tool called WamBam, which iterates BAMs and produces read identity and read length stats (source code can be found at GitHub [https://github.com/nanoporegenomics/wambam]). Pore-C statistics were generated via a similar method using the "evaluate_contacts" executable provided in the GFAse repository. Reads were aligned to both haplotypes of the trio phased Verkko "full-coverage" HG002 assembly, and statistics were calculated by accumulating loci and lengths for each mapping of each read ID. For the "signal ratio" calculation, a contact map was constructed by building a graph similar to that used in phasing methods. Any two mappings of the same read ID constitute an edge, and they are binned by the minimum mapping quality of the pair. Edges that cross from one haplotype to another are considered inconsistent with the true phasing and are used to compute a ratio of consistent to inconsistent edges.

As a truth set for phasing, chromosome-length haplotypes were generated using Strand-seq and long reads. To generate haplotypes, we have used a combination of Strand-seq data and PacBio Hifi reads from the same individual (HG005 and HG002). Sparse and chromosome-length haplotypes were generated using Strand-seq data and the R package (R Core Team 2023) StrandPhaseR (version 0.99) as previously described (Porubsky et al. 2017). Next, we detected inverted regions using Strand-seq data and manually curated this list of inversions as previously described (Porubsky et al. 2022). We have used this set of inversions to correct Strand-seq phasing over these regions with the StrandPhaseR function called "correctInvertedRegionPhasing" as previously described (Porubsky et al. 2022). After inversion-phase correction, we generated dense chromosome-length haplotypes using a combination of Strand-seq haplotypes and PacBio long reads as previously described in the integrative phasing framework (Porubsky et al. 2017; Hanlon et al. 2023). Integrative phasing was completed using WhatsHap version 1.0 (Patterson et al. 2015). For integrative phasing, we used a defined set of variant positions (available at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/ChineseTrio/HG005_NA24631_son/NISTv4.2.1/GRCh38/HG005_GRCh38_1_22_v4.2.1_benchmark.vcf.gz and https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NISTv4.2.1/GRCh38/HG002_

GRCh38_1_22_v4.2.1_benchmark.vcf.gz). To include indels into the final callset, we have run WhatsHap with the --indels parameter.

To analyze phasing accuracy, a combination of Dipcall (Li et al. 2018) and WhatsHap was used. Dipcall is a reference-based variant caller. For a set of phased assemblies, it produces a VCF file of single-nucleotide variants (SNVs), as well as small indels. For the male HG002 sample, Dipcall was run using the GRCh38 reference but was set to treat the PAR region as autosomal regions. Phase set tags were manually added to the Dipcall VCF file before being used by WhatsHap. WhatsHap "compare" assesses switch error and hamming distance in the phased assemblies by comparing the phasing of alleles in the NIST's Genome in a Bottle (Zook et al. 2016) truth set to the Dipcall VCF file. WhatsHap "compare" only includes variants in the analysis that have identical alleles in the truth-and-query VCF file, making it robust to SNVs caused by sequencing errors. WhatsHap "stats" were run with a "--chr-lengths" input file to calculate phasing statistics.

Collapses and misassemblies within the genes were evaluated using minimap2 (Li 2018), asmgene, and the publicly available Ensembl genes as input (https://ftp.ensembl.org/pub/release-87/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz). Ensembl cDNA was aligned to the CHM13 v2.0 reference and the HG002 assemblies with minimap2 using "-cx splice:hq" for intra-species cDNA alignment. Asmgene, a part of minimap2, selects the longest isoform from overlapping alignments and counts it as matching the reference if it covers >99% of the transcript length with a mapping identity above an input threshold. To account for ONT's sequencing error profiles, asmgene was run with a mapping identity threshold of 97% instead of the 99% used for HiFi assemblies. Transcripts are counted as single copy if they uniquely align to the reference and as multicopy if they align to multiple loci.

Base quality was estimated using yak (https://github.com/lh3/yak), based on the *k*-mer content of Illumina short reads. Each phased assembly was evaluated separately. *K*-mer in the short reads was counted using "yak count -b 37," and quality values (QVs) were estimated using "yak qv -K 3.2g -l 100k." For HG002, we used the 30× Illumina NovaSeq PCR-free read set publicly available at the Google bucket (gs://deepvariant/benchmarking/fastq/wgs_pcr_free/30x/). For the four samples from the HPRC (HG01993, HG02132, HG02647, and HG03669), we used 30× Illumina short-reads from the high-coverage read set of the 1000 Genomes Project samples (Byrska-Bishop et al. 2022).

Non-human assemblies (*N. fimbriata* and *S. bandensis*) were evaluated using default arguments for BUSCO v5.4.3 and the "metazoan_odb10" data set. To attempt to evaluate the number of phased genes, BUSCO was run twice: once with both haplotypes, and once with one haplotype. For the "both-haplotype" evaluation, the entire diploid assembly was provided to BUSCO. For the "one-haplotype" evaluation, one of each haplotype from the phased regions was removed, and the remaining sequences were evaluated by BUSCO.

Switch error in the phased assemblies was also estimated from Illumina short reads from parents. We used yak to count the *k*-mer in the short reads, as above, and "yak trioeval" to compute the estimated switch error rate. As above, the read sets for HG002's parents were downloaded from the same Google Bucket as for the base-quality evaluation and from the 1000 Genomes Project's data set for the four HPRC samples.

SVs were called from the phased assemblies using dipdiff, a modified version of the SVIM-ASM tool (Heller and Vingron 2020). In HG002 assemblies, the SVs were called against GRCh37 and evaluated with the GIAB SV truth set (Zook et al. 2020) using Truvari (English et al. 2022). Truvari's "bench" command was run with "--no-ref a -r 2000 -C 2000" to ignore missing homozygous calls for the reference allele and to match variants up to 2000 bp away from each other.

All forms of analysis and evaluation in this paper have been automated with workflow description language (Voss et al. 2017), and made portable using Dockstore (O'Connor et al. 2017) for reproducibility and convenience, and are available as follows: reference-based proximity linkage evaluation, (https://dockstore.org/workflows/github.com/meredith705/gfase_wdl/evaluate_contacts:main); reference-based phasing evaluation coupled with alignment-based gene completeness evaluation (https://dockstore.org/workflows/github.com/meredith705/gfase_wdl/dipcall_whatshap_asmgene:main); VCF-based SV evaluation (https://dockstore.org/workflows/github.com/meredith705/gfase_wdl/gfase_sv_evaluation:mai); reference-based alignment quality evaluation (https://dockstore.org/workflows/github.com/nanoporegenomics/wambam/wambam:main); and Yak trio k-mer-based phasing evaluation (https://dockstore.org/workflows/github.com/meredith705/gfase_wdl/gfase_ base_qv_trio_evaluation:main).

Data access

The Pore-C data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1066887. The cuttlefish assembly generated in this study has been submitted to the NCBI BioProject database under accession numbers PRJNA1082730 and PRJNA1082729 for primary and alt haplotypes. All data links and scripts required to reproduce this work are available at Zenodo (https://doi.org/10.5281/zenodo.10653823) and as Supplemental Code. Shasta source code is available at GitHub (https://github.com/paoloshasta/shasta). GFAse source code is also available at GitHub (https://github.com/rlorigro/GFAse).

Competing interest statement

S.J. and S.K.M. are employees of Oxford Nanopore Technologies and shareholders and/or share option holders of Oxford Nanopore Technologies. S.K. has received travel funds to speak at events hosted by Oxford Nanopore Technologies. E.E.E. is a scientific advisory board (SAB) member of Variant Bio. P.C. was an employee of the Chan Zuckerberg Initiative during the time most of this work was performed. K.H.M. is a SAB member of Centaura and has received travel funds to speak at events hosted by Oxford Nanopore Technologies.

Acknowledgments

This work was funded in part by the National Institutes of Health under award numbers R01HG010485, U24HG010262, U24HG01 1853, OT3HL142481, U01HG010961, and OT2OD033761. A portion of the R10 HG002 data set labeled "UUL" in this work was supported by startup funds (Miten Jain, Genome Technology Laboratory, Northeastern University). Cuttlefish Nanopore work was supported by Oxford Nanopore Technologies grant SC20130149 (awarded to Mark Akeson, UCSC Nanopore Group). We acknowledge the support of the Oxford Nanopore Technologies staff in generating this data set, in particular the pore-C data.

References

Altshuler D, Donnelly P, The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320. doi:10.1038/nature04226

Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. 2015. Mapping bias overestimates reference allele frequencies at the *HLA* genes in the 1000 Genomes Project phase I data. *G3 (Bethesda)* **5:** 931–941. doi:10.1534/g3.114.015784

Broder AZ. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, Salerno, Italy, pp. 21–29. IEEE.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81:** 1084–1097. doi:10.1086/521987

Browning SR, Browning BL. 2011. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* **12:** 703–714. doi:10.1038/nrg3054

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31:** 1119–1125. doi:10.1038/nbt.2727

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* **185:** 3426–3440.e19. doi:10.1016/j.cell.2022.08.004

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18:** 170–175. doi:10.1038/s41592-020-01056-5

Chin C-S, Wagner J, Zeng Q, Garrison E, Garg S, Fungtammasan A, Rautiainen M, Aganezov S, Kirsche M, Zarate S, et al. 2020. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat Commun* **11:** 4794. doi:10.1038/s41467-020-18564-9

Cordeiro JM, Barajas-Martinez H, Hong K, Burashnikov E, Pfeiffer R, Orsino A-M, Wu YS, Hu D, Brugada J, Brugada P, et al. 2006. Compound heterozygous mutations P336L and I1660V in the human cardiac sodium channel associated with the Brugada syndrome. *Circulation* **114:** 2026–2033. doi:10.1161/CIRCULATIONAHA.106.627489

Deamer D, Akeson M, Branton D. 2016. Three decades of nanopore sequencing. *Nat Biotechnol* **34:** 518–524. doi:10.1038/nbt.3423

Deshpande AS, Ulahannan N, Pendleton M, Dai X, Ly L, Behr JM, Schwenk S, Liao W, Augello MA, Tyer C, et al. 2022. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nat Biotechnol* **40:** 1488–1499. doi:10.1038/s41587-022-01289-z

Duan H, Jones AW, Hewitt T, Mackenzie A, Hu Y, Sharp A, Lewis D, Mago R, Upadhyaya NM, Rathjen JP, et al. 2022. Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in nanopore and HiFi assemblies with Hi-C data. *Genome Biol* **23:** 84. doi:10.1186/s13059-022-02658-2

Ebler J, Haukness M, Pesout T, Marschall T, Paten B. 2019. Haplotype-aware diplotyping from noisy long reads. *Genome Biol* **20:** 116. doi:10.1186/s13059-019-1709-0

Edge P, Bafna V, Bansal V. 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27:** 801–812. doi:10.1101/gr.213462.116

Eizenga JM, Novak AM, Kobayashi E, Villani F, Cisar C, Heumos S, Hickey G, Colonna V, Paten B, Garrison E. 2020. Efficient dynamic variation graphs. *Bioinformatics* **36:** 5139–5144. doi:10.1093/bioinformatics/btaa640

English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. 2022. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol* **23:** 271. doi:10.1186/s13059-022-02840-6

Günther T, Nettelblad C. 2019. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet* **15:** e1008302. doi:10.1371/journal.pgen.1008302

Hanlon VCT, Porubsky D, Lansdorp PM. 2023. Chromosome-length Haplotypes with StrandPhaseR and Strand-seq. *Methods Mol Biol* **2590:** 183–200. doi:10.1007/978-1-0716-2819-5_12

Heller D, Vingron M. 2020. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36:** 5519–5521. doi:10.1093/bioinformatics/btaa1034

Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium. 2021. The genome sequence of the broad-bordered yellow underwing, *Noctua fimbriata* (Schreber, 1759). *Wellcome Open Res* **6:** 345. doi:10.12688/wellcomeopenres.17490.1

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5:** e1000529. doi:10.1371/journal.pgen.1000529

Kokot M, Dlugosz M, Deorowicz S. 2017. KMC_3: counting and manipulating *k*-mer statistics. *Bioinformatics* **33:** 2759–2761. doi:10.1093/bioinformatics/btx304

Kolmogorov M, Billingsley KJ, Mastoras M, Meredith M, Monlong J, Lorig-Roach R, Asri M, Alvarez Jerez P, Malik L, Dewan R, et al. 2023. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat Methods* **20:** 1483–1492. doi:10.1038/s41592-023-01993-x

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* **36:** 1174–1182. doi:10.1038/nbt.4277

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. 2018. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15:** 595–597. doi:10.1038/s41592-018-0054-7

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293. doi:10.1126/science.1181369

Luo X, Kang X, Schönhuth A. 2021. phasebook: haplotype-aware de novo assembly of diploid genomes from long reads. *Genome Biol* **22:** 299. doi:10.1186/s13059-021-02512-x

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11:** 499–511. doi:10.1038/nrg2796

Miller DB, Piccolo SR. 2020. Compound heterozygous variants in pediatric cancers: a systematic review. *Front Genet* **11:** 493. doi:10.3389/fgene.2020.00493

O'Connor BD, Yuen D, Chung V, Duncan AG, Liu XK, Patricia J, Paten B, Stein L, Ferretti V. 2017. The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. *F1000Res* **6:** 52. doi:10.12688/f1000research.10137.1

Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, Schönhuth A. 2015. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J Comput Biol* **22:** 498–509. doi:10.1089/cmb.2014.0157

Payne A, Holmes N, Rakyan V, Loose M. 2019. BulkVis: a graphical viewer for Oxford Nanopore bulk FAST5 files. *Bioinformatics* **35:** 2193–2198. doi:10.1093/bioinformatics/bty841

Peterson RE, Kuchenbaecker K, Walters RK, Chen C-Y, Popejoy AB, Periyasamy S, Lam M, Iyegbe C, Strawbridge RJ, Brick L, et al. 2019. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179:** 589–603. doi:10.1016/j.cell.2019.08.051

Porubsky D, Garg S, Sanders AD, Korbel JO, Guryev V, Lansdorp PM, Marschall T. 2017. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat Commun* **8:** 1293. doi:10.1038/s41467-017-01389-4

Porubsky D, Höps W, Ashraf H, Hsieh P, Rodriguez-Martin B, Yilmaz F, Ebler J, Hallast P, Maria Maggiolini FA, Harvey WT, et al. 2022. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185:** 1986–2005.e26. doi:10.1016/j.cell.2022.04.017

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* **26:** 342–350. doi:10.1101/gr.193474.115

Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, Koren S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41:** 1474–1482. doi:10.1038/s41587-023-01662-6

R Core Team. 2023. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Selvaraj S, Dixon JR, Bansal V, Ren B. 2013. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **31:** 1111–1118. doi:10.1038/nbt.2728

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of 11 human genomes. *Nat Biotechnol* **38:** 1044–1053. doi:10.1038/s41587-020-0503-6

Song S, Sliwerska E, Emery S, Kidd JM. 2017. Modeling human population separation history using physically phased genomes. *Genetics* **205:** 385–395. doi:10.1534/genetics.116.192963

Voss K, Van der Auwera G, Gentry J. 2017. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. *F1000Res* **6:** 1381. doi:10.7490/f1000research.1114634.1

Walker MA, Mohler KP, Hopkins KW, Oakley DH, Sweetser DA, Ibba M, Frosch MP, Thibert RL. 2016. Novel compound heterozygous mutations expand the recognized phenotypes of *FARS2*-linked disease. *J Child Neurol* **31:** 1127–1137. doi:10.1177/0883073816643402

Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, Popejoy AB, Asri M, Carson C, Chaisson MJP, et al. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604:** 437–446. doi:10.1038/s41586-022-04601-8

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37:** 1155–1162. doi:10.1038/s41587-019-0217-9

Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31:** 3350–3352. doi:10.1093/bioinformatics/btv383

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3:** 160025. doi:10.1038/sdata.2016.25

Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al. 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* **38:** 1347–1355. doi:10.1038/s41587-020-0538-8

Received July 19, 2023; accepted in revised form March 19, 2024.

Phased nanopore assembly with Shasta and modular graph phasing with GFAse

Ryan Lorig-Roach, Melissa Meredith, Jean Monlong, et al.

Genome Res. 2024 34: 454-468 originally published online April 16, 2024
Access the most recent version at doi:10.1101/gr.278268.123