bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title: Fast and Accurate Genomic Analyses using Genome Graphs

Authors: Goran Rakocevic,[1,2,‡] Vladimir Semenyuk,[1,2,‡] James Spencer,[1,2] John Browning,[1,2] Ivan Johnson,[1,2] Vladan Arsenijevic,[1,2] Jelena Nadj,[1,2] Kaushik Ghose,[1,2] Maria C. Suciu,[1,2] Sun-Gou Ji,[1,2] Gülfem Demir,[1,2] Lizao Li,[1,2] Berke Ç. Toptaş,[1,2] Alexey Dolgoborodov,[1,2] Björn Pollex,[1,2] Irina Glotova,[1,2] Péter Kómár,[1,2] Yilong Li,[1,2] Milos Popovic,[1,2] Wan-Ping Lee,[1] Morten Källberg,[1] Amit Jain,[1,2] Deniz Kural[1,2,*]

Affiliations:

[1]Seven Bridges Genomics, Inc, Cambridge, MA 02140

[2]SBGD, Inc, Cambridge, MA 02140

* Corresponding author. Email: deniz.kural@sbgenomics.com

‡ These authors contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract:

The human reference genome serves as the foundation for genomics by providing a scaffold for sequencing read alignment, but currently only reflects a single consensus haplotype, impairing read alignment and downstream analysis accuracy. Reference genome structures incorporating known genetic variation have been shown to improve the accuracy of genomic analyses, but have so far remained computationally prohibitive for routine large-scale use. Here we present a graph genome implementation that enables read alignment across 2,800 diploid genomes encompassing 12.6 million SNPs and 4.0 million indels. Our graph genome aligner and variant calling pipeline consume around 5.5 and 2 hours per high coverage whole-genome-sequenced sample, respectively, comparable to those of state-of-the-art linear reference genome-based methods. Using orthogonal benchmarks based on real and simulated data, we show that using a graph genome reference improves read mapping sensitivity and produces a 0.5 percentage point increase in variant calling recall, which extrapolates into 20,000 additional variants being detected per sample, while variant calling specificity is unaffected. Structural variations (SVs) incorporated into a graph genome can be directly genotyped from read alignments in a rapid and accurate fashion. Finally, we show that iterative augmentation of graph genomes yields incremental gains in variant calling accuracy. Our implementation is the first practical step towards fulfilling the promise of graph genomes to radically enhance the scalability and precision of genomic analysis by incorporating prior knowledge of population characteristics.

One Sentence Summary: Genome graphs incorporating common genetic variation enable efficient variant identification at population scale.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Main Text:

Introduction

Completion of the human reference genome was a landmark achievement in human genetics (*1*), establishing a standardized coordinate system for annotating genomic elements and comparing individual human genomes. The human reference genome has since become the cornerstone of genomics and plays an indispensable role in research on population genetics, disease genetics, cancer genomics, epigenetics, transcriptomics and personalized medicine (*2–7*). For most applications, a pivotal function of the human reference genome lies in providing a scaffold for mapping and assembling short DNA sequences (reads) into longer consensus contigs. This crucial initial step significantly impacts the quality of all ensuing analyses and ultimately the ability to draw conclusions of clinical significance from a sample cohort being studied.

The current human reference genome is represented as a linear haploid DNA sequence (*8*). This structure poses theoretical limitations due to the prevalence of genetic diversity in human populations: any given human genome has on average 3.5-4.0 million single nucleotide polymorphisms (SNPs) or small insertions and deletions (INDELs) and around 2,500 large structural variations (SVs) compared with the reference genome (*2*, *9*), including sizable genomic segments missing from the reference (*10*). This genetic divergence may cause sequencing reads to map incorrectly or fail to map altogether (*11*, *12*), particularly when they span SV breakpoints. Read mapping accuracy thus varies significantly across genomic regions in a given sample, and across genetically diverged samples. Misplaced reads may in turn result in both missed true variants (false negatives) and incorrectly reported false variants (false positives), as well as hamper other applications that rely on accurate read placement, including RNA-sequencing, chromatin immunoprecipitation-sequencing (ChIP-seq) and copy number

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variation detection (*11*, *12*). Identifying SVs is particularly challenging: despite the large number of SVs already characterized (*9*), current methods for genotyping SVs do not make use of these data, but instead rely on detecting complex combinations of abnormal read alignment patterns to detect SVs (*13*, *14*). The genomic representation of SVs presents additional challenges. Since existing methods identify SVs independently between samples, the detected SVs cannot easily be matched using their inferred genomic coordinates, but further ad hoc procedures are needed to merge individual SV calls into common SV events (*9*). We refer to the insufficient representativeness of the linear reference genome and the consequent adverse impact to genomic analyses as 'reference bias'.

Reference bias is caused by the fact that existing genetic variant information is not considered in read mapping. Among the 1000G populations studied, 9-19% of the variants have a within-population allele frequency of >50%, of which 0.3-2.9% are fully fixed on a population-level (Figs. S1 to S4), raising the possibility of addressing reference bias by generating population-specific linear reference genomes using the major alleles of each population (*15*–*18*). However, such major allele references would still fail to include 81-91% of the variants in a given population (Figs. S3 and S4), and thus do not present a global solution for reference bias. Moreover, having distinct population-specific reference genomes with different coordinate systems would complicate any genomic analyses comprising multiple populations. An alternative approach to capture the genetic diversity in the human reference genome is to augment it with alternate haplotype sequences at genetically diverged loci (*19*). This approach has been applied since the previous human reference version, GRCh37 (*19*), but introduces inefficiency through sequence duplication between the main linear genome and the alternate haplotypes.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Recent large-scale resequencing efforts have comprehensively cataloged common genetic variants (*2*, *20*, *21*), prompting suggestions to make use of this information through multi-genome references (*22*), which have been suggested to alleviate reference bias by facilitating read mapping (*23*, *24*). Despite these promising observations, currently available implementations of multi-genome graph references are either orders of magnitude slower than conventional linear reference genome-based methods on human WGS data (one example is BWBBLE (*25*)) or are intended for use with small genomes (*22*) and small regions within large genomes (*23*, *24*, *26*, *27*). Therefore, the genome-wide impact of using multi-genome references for human genomic analyses has not been assessed yet, although whole-genome workflows using graph genomes are under active development (*28*, *29*).

Here we present a graph genome toolkit for building, augmenting, storing and querying graph genomes composed of a population of genome sequences. Our algorithms are, to our knowledge, the first graph genome implementation that achieve comparable reference indexing and read alignment speed to BWA-MEM (*30*), a widely-used linear reference genome aligner. We show that graph genomes improve the mapping accuracy of next-generation sequencing (NGS) reads on the genome-wide level. Our NGS read alignment and variant calling pipeline (Graph Genome Pipeline) leveraging our graph genome data structure outperforms the current gold-standard linear reference-genome pipeline (*31*) comprising BWA-MEM (*30*) and GATK HaplotypeCaller (*31*), as measured by multiple orthogonal benchmarks. By including breakpoint-resolved SV polymorphisms into the graph genome, we demonstrate that SVs can be genotyped rapidly and accurately with a simple read counting approach. As novel genetic variation data are accumulated in graph genomes, incremental improvements in read mapping and variant calling

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

accuracy can be achieved. This will allow our approach to scale and improve with the expanding genomic variant catalogs.

A Computationally Efficient Graph Genome Implementation

We implemented a graph genome data structure that represents genomic sequences on the edges of the graph (Fig. 1A, Materials and Methods). A graph genome is constructed from a population genome sequences, such that each haploid genome in this population is represented by a sequence path through the graph. The genome sequences are provided using VCF files indicating genetic variants with respect to a standard linear reference genome, which is provided using a FASTA file. Thus, the standard linear reference genome is just one of the paths through the graph genome. For representational purposes, the linear reference genome path is labeled as the initial edge, and all coordinates of genetic variants are reported with respect to it. This ensures backward compatibility of graph coordinates to linear reference genome coordinates. In practice, a graph genome is built by iteratively adding edges corresponding to a non-reference allele, terminating at nodes corresponding to genomic loci on the initial edge. Insertions are represented as cyclic edges starting and terminating at the same node, but our mapping algorithm enforces acyclic traversal of the graph (Fig. 1A). Genomic features such as tandem repeats expansions and inversions are represented as insertions or sequence replacements in the graph. Variants are inserted into existing variant branches in a backward compatible manner, enabling variant discovery and representation in genomic regions absent from the linear reference genome (*10*) (Fig. 1A). For querying a graph genome, we use a hash table that associates short sequences of length *k* (*k*-mers) along all valid paths in the graph with their graph coordinates (Fig. 1B). Uninformative *k*-mers that occur exceptionally frequently are omitted (Materials and Methods).

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We implemented a suite of accompanying algorithms optimized for computational resource consumption. A graph genome from the complete 1000G variant set can be built and indexed in less than ten minutes in total (Fig. 1C). Such a graph reference can be stored in less than 30 gigabytes (GB) of memory or just over 1 GB of disk space (Fig. 1C). Loading a stored graph reference into memory takes less than two minutes. Over tenfold range in the number of variants included, time, memory and disk storage consumption only grew around fourfold, twofold and 50%, respectively (Fig. 1C).

Improved Read Mapping Accuracy using Graph Genomes

To support genomic analyses on our graph genome implementation, we developed a graph aligner for short reads that uses the *k*-mer index for seeding followed by local read alignment against the graph (Materials and Methods). The read alignments against a path in the graph are projected to the standard reference genome and output to a standard BAM file, with the alignment path along the graph reported using custom annotation tags. Thus, the output format of our graph aligner maintains full compatibility with existing genomics data processing tools. When an unambiguous projection is not possible, for example for reads fully mapped within a long insertion variant, the reads are placed to the closest reference position, so that downstream analysis tools can access these reads conveniently.

We measured the graph aligner runtimes on 30 randomly selected high coverage whole-genome sequencing datasets from the Coriell repository (http://www.ebi.ac.uk/ena/data/view/PRJEB20654). Read alignment against a "global graph genome" containing around 20 million variants (Materials and Methods) required around 5.5 hours per sample when using 36 threads (Fig. 1D). This runtime is comparable to that of BWA-MEM (*30*) (Fig. 1D).

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In order to test the read mapping accuracy of the graph aligner, we simulated diploid individuals with variants observed naturally in the 1000G cohort. While reads without any variants align equally accurately to the reference genome by the graph aligner and BWA-MEM, the graph aligner maintains a high mapping rate and accuracy even in reads containing long INDELs relative to the standard linear reference genome (Fig. 2). Less than 1% of the reads with >10bp insertions and deletions are mismapped using the 1000G graph, whereas this number is two and three times as large with BWA-MEM, respectively (Fig. 2). Even against a linear reference without variants, our graph aligner is able to align more reads containing INDELs than BWA-MEM (Fig. 2). In conclusion, our graph genome aligner significantly improves read alignment fidelity over BWA-MEM while maintaining comparable runtimes.

Graph Genome Pipeline Improves Recall in Variant Detection

We developed a complete graph genome-based variant calling pipeline (Graph Genome Pipeline) that includes a reassembly variant caller and variant call filters as suggested previously (*31*) (Materials and Methods). Generating variant calls from raw FASTQs in the Coriell cohort (29-42x coverage) took on average 7h 10min (σ = 30min) using Graph Genome Pipeline on 36 CPU cores and 20GB of memory. In comparison, the best practices GATK pipeline using GATK HaplotypeCaller (https://software.broadinstitute.org/gatk/best-practices/, hereafter referred to as BWA-GATK) run on same hardware required 50GB of memory and an average of 11h 30min (σ = 3h 16min).

We devised three independent and complementary experiments to compare the variant calling accuracy of our Graph Genome Pipeline against that of BWA-GATK (Supplementary Materials and Fig. S12). The first benchmarking experiment is based on sequencing data simulation, which provides a known ground truth for all variants throughout the genome, but likely incompletely

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reproduces the error modalities of real sequencing data. The second benchmarking experiment uses truth data established by the Genome in a Bottle Consortium (GIAB) (*32*) for five high coverage whole-genome sequenced samples (50x coverage). These truth data cover only about 70% of the genome considered as "high-confidence" regions. This likely excludes ~30% of the genome that is hardest to align and call variants against. Consistent with this view, most contemporary pipelines already achieve greater than 99% variant calling precision and recall with the GIAB samples (https://precision.fda.gov/challenges/truth/results), limiting the utility of these truth sets in detecting further improvements in variant calling accuracy. Our third variant calling benchmark is based on measuring Mendelian consistency in family trios, which is an indirect proxy for variant calling accuracy, but can be conducted on real data throughout the genome. To support this approach, we developed a statistical method to estimate the unknown precision and recall rates of a variant caller using variant calls derived from each member of a family trio (Materials and Methods).

A consistent pattern emerges from the three benchmarking experiments. In most samples across each benchmark, Graph Genome Pipeline improves SNP recall by around 0.5% over BWA-GATK, while their SNP calling precisions are virtually the same (Fig. 3). Both the simulation and the GIAB benchmarks show that Graph Genome Pipeline improves INDEL recall, as well as INDEL calling precision in all but one sample (Fig. 3, A to B). In terms of SNP, INDEL and overall F1 score, Graph Genome Pipeline outperforms all pipelines that participated in the Precision FDA Truth Challenge (Figs. S9 and S10, https://precision.fda.gov/challenges/truth). With around 3.5-4.0 million SNPs and INDELs in an average individual (*33*), the improved recall of Graph Genome Pipeline translates into around 20,000 additional variants being discovered compared with BWA-GATK, with precision remaining comparable.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The GIAB variant call sets provide an estimate for the practical upper limit in achievable accuracy using the standard linear reference genome, since they are carefully curated from an extensive amount of high quality data generated from a combination of several different sequencing platforms, such as 300x PCR-free Illumina sequencing and high coverage 10x Genomics data, and meta-analyzed across a suite of state-of-the-art bioinformatics tools (*32*). Interestingly, among the variants detect by Graph Genome Pipeline but asserted as homozygous reference by GIAB, a substantial proportion (24-47% across four samples, Supplementary Materials) exhibited strong support from alternative sequencing technologies as well as in terms of Mendelian concordance (Supplementary Materials). These variants are often located in variant-dense regions, half (52%) of them are part of the global graph, and most of the remaining variants (46%) are phased with one or multiple nearby variants present in the graph (Fig. S7, Supplementary Materials). Contrary to the linear reference genome, Graph Genome Pipeline is by design able to map reads across known variations without reference bias, which allows it to mitigate the impact of reference bias in repetitive or variant-dense genomic regions (Materials and Methods). We therefore hypothesized that these variants could be real but missed by all other linear reference genome-based pipelines used by GIAB due to reference bias. We successfully carried out Sanger sequencing at 241 and 457 of these “false FP” variants in two GIAB samples, HG001 and HG002, validating 62% and 58% of these variants as real variants missed by GIAB, respectively. Notably, the variants missed by GIAB only derive from the 70% of the genome considered “high-confidence” by GIAB, whilst the remainder of the genome likely contains more variant-dense loci refractory to accurate genomic analyses using linear reference genomes. Thus, our graph genome implementation is able to overcome the practical accuracy limitation of linear reference genomes.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Direct Genotyping of Common SVs using Graph Genomes

Sequence information of SV breakpoints can be incorporated into a graph genome, allowing reads to be mapped across SV breakpoints. Indeed, we found that Graph Genome Pipeline is able to align reads across SVs while BWA-MEM fails to do so with both short Illumina and long PacBio reads, even when the PacBio reads are aligned using parameters tuned for PacBio data (Fig. 4, A to C).

To demonstrate that reads spanning SV breakpoints can be used to directly genotype SVs, we manually curated a dataset of 230 high quality, breakpoint-resolved deletion-type SVs (Table S6) and genotyped them across 49 individuals from the Coriell cohort for which the true SV genotypes are available from the 1000 Genomes Project (*9*) (Materials and Methods). The fractions of reads spanning SV breakpoints segregates cleanly into three clusters based on SV genotype (Fig. 4D) suggesting that graph genome-based SV genotyping could be accomplished even with a simple read counting-based method. Indeed, a simple alternate allele fraction thresholding method correctly identifies 10,836/11,270 (96.1%) of the totality of SV genotypes across all 49 samples. Moreover, SV genotyping across all 230 sites takes less than three seconds per sample. Since SVs genotyped using a graph genome have an automatically defined genomic position with respect to the linear reference genome, they are directly comparable to each other without requiring further coordinate matching.

To compare the SV genotyping performance of the graph aligner to competing technologies, we focused on the GIAB sample HG002, for which both Illumina read and PacBio long read data are publicly available. We manually examined the alignment results from each technology in the 72/230 SVs detected in this sample by the graph aligner. BWA-MEM is unable to align Illumina reads across any of these SVs (Fig. 4A). Similarly, PacBio long read alignment fails in all these

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SVs, even when aligned with BWA-MEM using parameters tuned for PacBio data (Fig. 4C and Fig. S11). Thus, graph genomes improve SV genotyping for short read and long read sequencing technologies alike.

Among the 230 curated SVs are two events, esv3642033 and esv3638126, that are in strong linkage disequilibrium with SNPs significantly associated (P-value $< 5\times10^{-8}$) with breast cancer (rs1436904) and obesity class II risk (rs11639988), respectively (*34*, *35*) (Fig. 4C). We were able to correctly genotype the presence of these two SVs in 48/49 and 49/49 samples, respectively (Fig. 4C). Thus, graph genomes enable rapid and straightforward genotyping of common SVs, including those with clinical and biological relevance.

Graph Genomes Prevent Erroneous Variant Calls Around SVs

Structural variations mediated by certain DNA repair mechanisms can exhibit microhomology, including imperfect microhomology, around their breakpoints (*36*). If an aligner is not aware of an SV, sequencing reads spanning the SV could become erroneously aligned over a region of imperfect microhomology instead, causing mismatches over the region to be spuriously reported as SNPs and INDELs. To quantify this effect in 1000G, we compared the rate of 1000G SNPs around SV breakpoints to their background levels. Within the deleted portions of the 230 curated deletion SVs combined, the total rate of 1000G SNPs is 1.9/base pair (bp), and these SNPs have a transition-transversion ratio (Ti/Tv) of 2.1, which is expected from real biological SNPs (*37*) (Fig. 4E). In contrast, in the ten first base pairs immediately after an SV breakpoint, the total 1000G SNP rate is increased 1.9-fold to 5.5/bp, suggesting that 65% of 1000G SNPs called within 10bp of an SV breakpoint are in fact false (Fig. 4, E to F). The Ti/Tv of these SNPs is 0.72, deviating significantly from the expected ratio of 2.1. Assuming that spurious SNPs have an expected Ti/Tv of 0.5, the FP SNP rate over this region estimated using Ti/Tv is 86%,

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reaching a similar value to that estimated using total SNP counts (Supplementary Materials). In addition to false positive variant calls, we also encountered examples where variants overlapping with an SV erroneously appear homozygous, because BWA-MEM fails to align reads across the SV and thus fails to detect the corresponding SV haplotype (Fig. 4A). Thus, using population variation information in graph genomes mitigates variant calling and genotyping errors.

Incremental Improvement in Variant Calling Recall through Iterative Graph Augmentation

As common genetic variants continue to be cataloged across populations, newly discovered variants can be incrementally added to existing graph genomes to increase the comprehensiveness of the graph while maintaining backward compatibility to samples analyzed using earlier versions of the graph (Fig. 1A). To test whether incremental graph augmentation would improve variant calling, we augmented the global graph with variants detected in ten samples from three super-populations of the Coriell cohort as well as a Qatar genome project cohort (*16*) (Fig. 5, A to B), and compared the variant calls obtained using the global and the four augmented global graphs. Augmenting the global graph genome does not change the number of known variants discovered (0.7% and 0.0% median increase in the number of known SNPs and INDELs called, respectively; Fig. 5, C to D). However, the augmented graphs result in 16.6% and 2.4% additional previously unreported SNPs and INDELs being discovered, respectively (Fig. 5, C to D).

We measured the quality of the detected variants indirectly using Ti/Tv and heterozygous-to-homozygous alternate allele ratio (het/hom) for SNPs, and INDELs, respectively (*37*). For each tested pipeline, known SNPs and INDELs have a Ti/Tv ratio of 2.06 and het/hom ratio of 1.3-2.0, which is within the expected range for common variants (*37*). These indicators deviate from

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the expected values for novel SNPs and INDELs (although to a lesser extent with Graph Genome Pipeline compared with BWA-GATK), implying that there are false positives among the novel variants called by Graph Genome Pipeline (Fig. 5, C to D). However, while the augmented graphs help detect more known and novel variants, Ti/Tv and het/hom of either, with the exception of novel indels in the Qatari samples, are unaffected.

Read alignment and variant calling benchmarking experiments yielded similar observations. Read alignment recall reaches almost 100% if a target sample is aligned against a graph genome that contains all its actual variants (Fig. 2). Likewise, trio concordance and variant calling recall is further improved variant calling in a child is performed using a graph genome augmented with variants detected in the respective parents (Fig. 3, C to E). Thus, incremental augmentation of graph genome references yields cumulative improvements in variant calling recall without an accompanying decrease in precision.

Discussion

Here we presented, to our knowledge, the first computational infrastructure for building, augmenting, storing and querying multi-genome references at comparable computational requirements to prevailing linear reference-based methods. Our graph genomes are constructed from the linear reference genome and genetic variants provided in standard FASTA and VCF formats, and our graph genome aligner outputs read alignments in the standard BAM format against the coordinate system of the linear reference genome. As a result of the backward compatibility and computational efficiency of our graph genome infrastructure, any existing genomics pipeline can swap in our graph genome algorithms to capitalize on the accompanying improvements in accuracy without requiring any other pipeline or hardware upgrade.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Our benchmarking experiments demonstrate that using a graph genome reference improves read mapping and variant calling recall, including that of SVs, without a concomitant loss in precision (Figs. 2 to 5). Our graph aligner is able to readily align reads across breakpoint-resolved SVs included in the graph, unlike linear reference genome-based methods even with long reads (Figs. 4 and S11). These alignments can then be used to directly genotype SVs (Fig. 4). In contrast, existing methods for genotyping INDELs or SVs require specifically designed multi-step algorithms (*13*). Importantly, graph genome-based SV genotyping provides a unified coordinate system for common SVs, solving the problem of ambiguous coordinate representation that complicates SV analyses (*9*). The simplicity, speed and accuracy of graph genome-based SV genotyping raises the prospect for population-scale SV genetics and association studies.

Further improvements to variant calling could be achieved through leveraging haplotype information of SVs and small variants onto the graph paths. Information such as allele frequencies of each variant and linkage disequilibrium between them could be incorporated into the graph, providing additional statistical information for read alignment and variant calling. This approach would provide a computationally efficient alternative to "joint variant calling" (*38*) in leveraging previously accumulated population genetics information when analysing a newly sequence sample. Future efforts from us and others (*39*) will be required to assess the benefits of using graph genomes with encoded allele frequency and linkage disequilibrium information.

The potential benefits of an unbiased multi-genome graph reference are not limited to variant calling, but cover the full range of genomics research. Graph genomes provide an unbiased, representative scaffold for read alignment, which is critical to sequencing alignment quantification applications such as RNA-sequencing, ChIP-seq and CNV calling. Our graph genome implementation can also be used to encode information other than whole human

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genomes. Individual gene families or related microbial strains could be compressed and efficiently searched using our graph genome algorithms. Similarly, the transcriptome could be represented as genomic deletions, allowing RNA-seq reads to be directly aligned across exon-exon junctions (*40*). A personalized graph genome could be constructed from a sequenced germline genome, in order to provide an optimized scaffold for somatic variant detection in matched cancer genomes.

The ability to genotype variants including SVs efficiently and accurately opens the door for myriad clinical and research applications. A graph built from rearranged B or T cell receptors could allow quantitative tracking of mature B or T cells populations. A somatic graph incorporating somatic mutations and SVs of a treated primary tumor could enable sensitive monitoring of tumor relapse at the primary site or in cell-free DNA. A "precision medicine graph" could be designed and continually updated to include all actionable genetic variants in order to support clinical decision making.

As more genetic variants are accumulated on a reference graph genome, cumulative gains in genomics analysis accuracy can be achieved (Figs. 2, 3 and 5). This is consistent with recent efforts to establish population-specific reference panels, which have been shown to contribute to increased accuracy in imputation (*41–43*) and genetic risk prediction (*44*). The current wave of national sequencing projects will extend the catalogs of population-specific genetic variants, which will incrementally improve the prospects for graph genome reference approaches (*24*, *28*, *29*), including ours. Completion of the linear human reference genome marked the beginning of human genomics. Our computationally efficient and flexible graph genome implementation supports the community for a gradual transition towards a graph-based reference system.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References and Notes:

1. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature*. **431**, 931–945 (2004).
2. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).
3. ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. **306**, 636–640 (2004).
4. Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
5. F. S. Collins, H. Varmus, A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).
6. GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
7. Deciphering Developmental Disorders Study, Large-scale discovery of novel genetic causes of developmental disorders. *Nature*. **519**, 223–228 (2015).
8. V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, R. S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K. M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J. T. Simpson, G. Threadgold, J. Torrance, J. M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, H. Li, C.-S. Chin, A. M. Phillippy, R. Durbin, R. K. Wilson, P. Flicek, E. E. Eichler, D. M. Church, Evaluation of GRCh38 and de novo

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* (2017), doi:10.1101/gr.213611.116.

9. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, 1000 Genomes Project Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel, An integrated map of structural variation in 2,504 human genomes. *Nature*. **526**, 75–81 (2015).
10. B. Kehr, A. Helgadottir, P. Melsted, H. Jonsson, H. Helgason, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, A. Gylfason, G. H. Halldorsson, S. Kristmundsdottir, G. Thorgeirsson, I. Olafsson, H. Holm, U. Thorsteinsdottir, P. Sulem, A. Helgason, D. F. Gudbjartsson, B. V. Halldorsson, K. Stefansson, Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588–593 (2017).
11. J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, J. K. Pritchard, Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*. **25**, 3207–3212 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

12. D. Y. C. Brandt, V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, D. Meyer, Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3* . **5**, 931–941 (2015).

13. C. Alkan, B. P. Coe, E. E. Eichler, Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).

14. D. Antaki, W. M. Brandler, J. Sebat, SV2: Accurate Structural Variation Genotyping and De Novo Mutation Detection. *bioRxiv* (2017), p. 113498.

15. N. D. Thanh, P. T. M. Trang, D. T. Hai, N. H. A. Tuan, L. S. Quang, B. Q. Minh, D. Q. Minh, P. B. Son, L. S. Vinh, AB050. Building population-specific reference genomes: a case study of Vietnamese reference genome. *Annals of Translational Medicine*. **3** (2015), doi:10.3978/j.issn.2305-5839.2015.AB050.

16. K. A. Fakhro, M. R. Staudt, M. D. Ramstetter, A. Robay, J. A. Malek, R. Badii, A. A.-N. Al-Marri, C. Abi Khalil, A. Al-Shakaki, O. Chidiac, D. Stadler, M. Zirie, A. Jayyousi, J. Salit, J. G. Mezey, R. G. Crystal, J. L. Rodriguez-Flores, The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Hum Genome Var*. **3**, 16016 (2016).

17. L. Maretty, J. M. Jensen, B. Petersen, J. A. Sibbesen, S. Liu, P. Villesen, L. Skov, K. Belling, C. Theil Have, J. M. G. Izarzugaza, M. Grosjean, J. Bork-Jensen, J. Grove, T. D. Als, S. Huang, Y. Chang, R. Xu, W. Ye, J. Rao, X. Guo, J. Sun, H. Cao, C. Ye, J. van Beusekom, T. Espeseth, E. Flindt, R. M. Friborg, A. E. Halager, S. Le Hellard, C. M. Hultman, F. Lescai, S. Li, O. Lund, P. Løngren, T. Mailund, M. L. Matey-Hernandez, O. Mors, C. N. S. Pedersen, T. Sicheritz-Pontén, P. Sullivan, A. Syed, D. Westergaard, R. Yadav, N. Li, X. Xu, T. Hansen, A. Krogh, L. Bolund, T. I. A. Sørensen, O. Pedersen, R. Gupta, S. Rasmussen, S. Besenbacher, A. D. Børglum, J. Wang, H. Eiberg, K. Kristiansen,

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S. Brunak, M. H. Schierup, Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*. **548**, 87–91 (2017).

18. F. E. Dewey, R. Chen, S. P. Cordero, K. E. Ormond, C. Caleshu, K. J. Karczewski, M. Whirl-Carrillo, M. T. Wheeler, J. T. Dudley, J. K. Byrnes, O. E. Cornejo, J. W. Knowles, M. Woon, K. Sangkuhl, L. Gong, C. F. Thorn, J. M. Hebert, E. Capriotti, S. P. David, A. Pavlovic, A. West, J. V. Thakuria, M. P. Ball, A. W. Zaranek, H. L. Rehm, G. M. Church, J. S. West, C. D. Bustamante, M. Snyder, R. B. Altman, T. E. Klein, A. J. Butte, E. A. Ashley, Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet.* **7**, e1002280 (2011).
19. D. M. Church, V. A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.-C. Chen, R. Agarwala, W. M. McLaren, G. R. S. Ritchie, D. Albracht, M. Kremitzki, S. Rock, H. Kotkiewicz, C. Kremitzki, A. Wollam, L. Trani, L. Fulton, R. Fulton, L. Matthews, S. Whitehead, W. Chow, J. Torrance, M. Dunn, G. Harden, G. Threadgold, J. Wood, J. Collins, P. Heath, G. Griffiths, S. Pelan, D. Grafham, E. E. Eichler, G. Weinstock, E. R. Mardis, R. K. Wilson, K. Howe, P. Flicek, T. Hubbard, Modernizing Reference Genome Assemblies. *PLoS Biol.* **9**, e1001091 (2011).
20. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. **538**, 201–206 (2016).

21. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536**, 285–291 (2016).

22. K. Schneeberger, J. Hagmann, S. Ossowski, N. Warthmann, S. Gesing, O. Kohlbacher, D. Weigel, Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10**,

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

R98 (2009).

23. B. Paten, A. Novak, D. Haussler, Mapping to a Reference Genome Structure. *arXiv [q-bio.GN]* (2014), (available at http://arxiv.org/abs/1404.5010).

24. A. M. Novak, G. Hickey, E. Garrison, S. Blum, A. Connelly, A. Dilthey, J. Eizenga, M. A. Saleh Elmohamed, S. Guthrie, A. Kahles, S. Keenan, J. Kelleher, D. Kural, H. Li, M. F. Lin, K. Miga, N. Ouyang, G. Rakocevic, M. Smuga-Otto, A. W. Zaranek, R. Durbin, G. McVean, D. Haussler, B. Paten, Genome Graphs. *bioRxiv* (2017), p. 101378.

25. L. Huang, V. Popic, S. Batzoglou, Short read alignment with populations of genomes. *Bioinformatics*. **29**, i361–i370 (2013).

26. A. Dilthey, C. Cox, Z. Iqbal, M. R. Nelson, G. McVean, Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47**, 682–688 (2015).

27. H. P. Eggertsson, H. Jonsson, S. Kristmundsdottir, E. Hjartarson, B. Kehr, G. Masson, F. Zink, K. E. Hjorleifsson, A. Jonasdottir, A. Jonasdottir, I. Jonsdottir, D. F. Gudbjartsson, P. Melsted, K. Stefansson, B. V. Halldorsson, Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* (2017), doi:10.1038/ng.3964.

28. D. Kim, *hisat2* (Github; https://github.com/infphilo/hisat2).

29. *vg* (Github; https://github.com/vgteam/vg).

30. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013), (available at http://arxiv.org/abs/1303.3997).

31. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data. *Nat. Genet.* **43**, 491–498 (2011).

32. J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, E. Henaff, A. B. R. McIntyre, D. Chandramohan, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini, Z. Dzakula, A. Hastie, H. Cao, G. Deikus, E. Schadt, R. Sebra, A. Bashir, R. M. Truty, C. C. Chang, N. Gulbahce, K. Zhao, S. Ghosh, F. Hyland, Y. Fu, M. Chaisson, C. Xiao, J. Trow, S. T. Sherry, A. W. Zaranek, M. Ball, J. Bobe, P. Estep, G. M. Church, P. Marks, S. Kyriazopoulou-Panagiotopoulou, G. X. Y. Zheng, M. Schnall-Levin, H. S. Ordonez, P. A. Mudivarti, K. Giorda, Y. Sheng, K. B. Rypdal, M. Salit, Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. **3**, 160025 (2016).
33. 1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean, An integrated map of genetic variation from 1,092 human genomes. *Nature*. **491**, 56–65 (2012).
34. K. Michailidou, P. Hall, A. Gonzalez-Neira, M. Ghoussaini, J. Dennis, R. L. Milne, M. K. Schmidt, J. Chang-Claude, S. E. Bojesen, M. K. Bolla, Q. Wang, E. Dicks, A. Lee, C. Turnbull, N. Rahman, Breast and Ovarian Cancer Susceptibility Collaboration, O. Fletcher, J. Peto, L. Gibson, I. Dos Santos Silva, H. Nevanlinna, T. A. Muranen, K. Aittomäki, C. Blomqvist, K. Czene, A. Irwanto, J. Liu, Q. Waisfisz, H. Meijers-Heijboer, M. Adank, Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON), R. B. van der Luijt, R. Hein, N. Dahmen, L. Beckman, A. Meindl, R. K. Schmutzler, B. Müller-Myhsok, P. Lichtner, J. L. Hopper, M. C. Southey, E. Makalic, D. F. Schmidt, A. G. Uitterlinden, A. Hofman, D. J. Hunter, S. J. Chanock, D. Vincent, F. Bacot, D. C. Tessier, S. Canisius, L. F. A. Wessels, C. A. Haiman, M. Shah, R. Luben, J. Brown, C. Luccarini, N.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Schoof, K. Humphreys, J. Li, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, F. J. Couch, X. Wang, C. Vachon, K. N. Stevens, D. Lambrechts, M. Moisse, R. Paridaens, M.-R. Christiaens, A. Rudolph, S. Nickels, D. Flesch-Janys, N. Johnson, Z. Aitken, K. Aaltonen, T. Heikkinen, A. Broeks, L. J. V. Veer, C. E. van der Schoot, P. Guénel, T. Truong, P. Laurent-Puig, F. Menegaux, F. Marme, A. Schneeweiss, C. Sohn, B. Burwinkel, M. P. Zamora, J. I. A. Perez, G. Pita, M. R. Alonso, A. Cox, I. W. Brock, S. S. Cross, M. W. R. Reed, E. J. Sawyer, I. Tomlinson, M. J. Kerin, N. Miller, B. E. Henderson, F. Schumacher, L. Le Marchand, I. L. Andrulis, J. A. Knight, G. Glendon, A. M. Mulligan, kConFab Investigators, Australian Ovarian Cancer Study Group, A. Lindblom, S. Margolin, M. J. Hooning, A. Hollestelle, A. M. W. van den Ouweland, A. Jager, Q. M. Bui, J. Stone, G. S. Dite, C. Apicella, H. Tsimiklis, G. G. Giles, G. Severi, L. Baglietto, P. A. Fasching, L. Haeberle, A. B. Ekici, M. W. Beckmann, H. Brenner, H. Müller, V. Arndt, C. Stegmaier, A. Swerdlow, A. Ashworth, N. Orr, M. Jones, J. Figueroa, J. Lissowska, L. Brinton, M. S. Goldberg, F. Labrèche, M. Dumont, R. Winqvist, K. Pylkäs, A. Jukkola-Vuorinen, M. Grip, H. Brauch, U. Hamann, T. Brüning, GENICA (Gene Environment Interaction and Breast Cancer in Germany) Network, P. Radice, P. Peterlongo, S. Manoukian, B. Bonanni, P. Devilee, R. A. E. M. Tollenaar, C. Seynaeve, C. J. van Asperen, A. Jakubowska, J. Lubinski, K. Jaworska, K. Durda, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, N. V. Bogdanova, N. N. Antonenkova, T. Dörk, V. N. Kristensen, H. Anton-Culver, S. Slager, A. E. Toland, S. Edge, F. Fostira, D. Kang, K.-Y. Yoo, D.-Y. Noh, K. Matsuo, H. Ito, H. Iwata, A. Sueta, A. H. Wu, C.-C. Tseng, D. Van Den Berg, D. O. Stram, X.-O. Shu, W. Lu, Y.-T. Gao, H. Cai, S. H. Teo, C. H. Yip, S. Y. Phuah, B. K. Cornes, M. Hartman, H. Miao, W. Y. Lim, J.-H. Sng, K. Muir, A. Lophatananon, S. Stewart-Brown, P.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Siriwanarangsan, C.-Y. Shen, C.-N. Hsiung, P.-E. Wu, S.-L. Ding, S. Sangrajrang, V. Gaborieau, P. Brennan, J. McKay, W. J. Blot, L. B. Signorello, Q. Cai, W. Zheng, S. Deming-Halverson, M. Shrubsole, J. Long, J. Simard, M. Garcia-Closas, P. D. P. Pharoah, G. Chenevix-Trench, A. M. Dunning, J. Benitez, D. F. Easton, Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–61, 361e1–2 (2013).

35. S. I. Berndt, S. Gustafsson, R. Mägi, A. Ganna, E. Wheeler, M. F. Feitosa, A. E. Justice, K. L. Monda, D. C. Croteau-Chonka, F. R. Day, T. Esko, T. Fall, T. Ferreira, D. Gentilini, A. U. Jackson, J. 'an Luan, J. C. Randall, S. Vedantam, C. J. Willer, T. W. Winkler, A. R. Wood, T. Workalemahu, Y.-J. Hu, S. H. Lee, L. Liang, D.-Y. Lin, J. L. Min, B. M. Neale, G. Thorleifsson, J. Yang, E. Albrecht, N. Amin, J. L. Bragg-Gresham, G. Cadby, M. den Heijer, N. Eklund, K. Fischer, A. Goel, J.-J. Hottenga, J. E. Huffman, I. Jarick, Å. Johansson, T. Johnson, S. Kanoni, M. E. Kleber, I. R. König, K. Kristiansson, Z. Kutalik, C. Lamina, C. Lecoeur, G. Li, M. Mangino, W. L. McArdle, C. Medina-Gomez, M. Müller-Nurasyid, J. S. Ngwa, I. M. Nolte, L. Paternoster, S. Pechlivanis, M. Perola, M. J. Peters, M. Preuss, L. M. Rose, J. Shi, D. Shungin, A. V. Smith, R. J. Strawbridge, I. Surakka, A. Teumer, M. D. Trip, J. Tyrer, J. V. Van Vliet-Ostaptchouk, L. Vandenput, L. L. Waite, J. H. Zhao, D. Absher, F. W. Asselbergs, M. Atalay, A. P. Attwood, A. J. Balmforth, H. Basart, J. Beilby, L. L. Bonnycastle, P. Brambilla, M. Bruinenberg, H. Campbell, D. I. Chasman, P. S. Chines, F. S. Collins, J. M. Connell, W. O. Cookson, U. de Faire, F. de Vegt, M. Dei, M. Dimitriou, S. Edkins, K. Estrada, D. M. Evans, M. Farrall, M. M. Ferrario, J. Ferrières, L. Franke, F. Frau, P. V. Gejman, H. Grallert, H. Grönberg, V. Gudnason, A. S. Hall, P. Hall, A.-L. Hartikainen, C. Hayward, N. L. Heard-Costa, A. C. Heath, J. Hebebrand, G. Homuth,

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

F. B. Hu, S. E. Hunt, E. Hyppönen, C. Iribarren, K. B. Jacobs, J.-O. Jansson, A. Jula, M. Kähönen, S. Kathiresan, F. Kee, K.-T. Khaw, M. Kivimäki, W. Koenig, A. T. Kraja, M. Kumari, K. Kuulasmaa, J. Kuusisto, J. H. Laitinen, T. A. Lakka, C. Langenberg, L. J. Launer, L. Lind, J. Lindström, J. Liu, A. Liuzzi, M.-L. Lokki, M. Lorentzon, P. A. Madden, P. K. Magnusson, P. Manunta, D. Marek, W. März, I. Mateo Leach, B. McKnight, S. E. Medland, E. Mihailov, L. Milani, G. W. Montgomery, V. Mooser, T. W. Mühleisen, P. B. Munroe, A. W. Musk, N. Narisu, G. Navis, G. Nicholson, E. A. Nohr, K. K. Ong, B. A. Oostra, C. N. A. Palmer, A. Palotie, J. F. Peden, N. Pedersen, A. Peters, O. Polasek, A. Pouta, P. P. Pramstaller, I. Prokopenko, C. Pütter, A. Radhakrishnan, O. Raitakari, A. Rendon, F. Rivadeneira, I. Rudan, T. E. Saaristo, J. G. Sambrook, A. R. Sanders, S. Sanna, J. Saramies, S. Schipf, S. Schreiber, H. Schunkert, S.-Y. Shin, S. Signorini, J. Sinisalo, B. Skrobek, N. Soranzo, A. Stančáková, K. Stark, J. C. Stephens, K. Stirrups, R. P. Stolk, M. Stumvoll, A. J. Swift, E. V. Theodoraki, B. Thorand, D.-A. Tregouet, E. Tremoli, M. M. Van der Klauw, J. B. J. van Meurs, S. H. Vermeulen, J. Viikari, J. Virtamo, V. Vitart, G. Waeber, Z. Wang, E. Widén, S. H. Wild, G. Willemsen, B. R. Winkelmann, J. C. M. Witteman, B. H. R. Wolffenbuttel, A. Wong, A. F. Wright, M. C. Zillikens, P. Amouyel, B. O. Boehm, E. Boerwinkle, D. I. Boomsma, M. J. Caulfield, S. J. Chanock, L. A. Cupples, D. Cusi, G. V. Dedoussis, J. Erdmann, J. G. Eriksson, P. W. Franks, P. Froguel, C. Gieger, U. Gyllensten, A. Hamsten, T. B. Harris, C. Hengstenberg, A. A. Hicks, A. Hingorani, A. Hinney, A. Hofman, K. G. Hovingh, K. Hveem, T. Illig, M.-R. Jarvelin, K.-H. Jöckel, S. M. Keinanen-Kiukaanniemi, L. A. Kiemeney, D. Kuh, M. Laakso, T. Lehtimäki, D. F. Levinson, N. G. Martin, A. Metspalu, A. D. Morris, M. S. Nieminen, I. Njølstad, C. Ohlsson, A. J. Oldehinkel, W. H. Ouwehand, L. J. Palmer, B. Penninx, C. Power, M. A.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Province, B. M. Psaty, L. Qi, R. Rauramaa, P. M. Ridker, S. Ripatti, V. Salomaa, N. J. Samani, H. Snieder, T. I. A. Sørensen, T. D. Spector, K. Stefansson, A. Tönjes, J. Tuomilehto, A. G. Uitterlinden, M. Uusitupa, P. van der Harst, P. Vollenweider, H. Wallaschofski, N. J. Wareham, H. Watkins, H.-E. Wichmann, J. F. Wilson, G. R. Abecasis, T. L. Assimes, I. Barroso, M. Boehnke, I. B. Borecki, P. Deloukas, C. S. Fox, T. Frayling, L. C. Groop, T. Haritunian, I. M. Heid, D. Hunter, R. C. Kaplan, F. Karpe, M. F. Moffatt, K. L. Mohlke, J. R. O'Connell, Y. Pawitan, E. E. Schadt, D. Schlessinger, V. Steinthorsdottir, D. P. Strachan, U. Thorsteinsdottir, C. M. van Duijn, P. M. Visscher, A. M. Di Blasio, J. N. Hirschhorn, C. M. Lindgren, A. P. Morris, D. Meyre, A. Scherag, M. I. McCarthy, E. K. Speliotes, K. E. North, R. J. F. Loos, E. Ingelsson, Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* **45**, 501–512 (2013).

36. M. McVey, S. E. Lee, MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet.* **24**, 529–538 (2008).
37. J. Wang, L. Raskin, D. C. Samuels, Y. Shyr, Y. Guo, Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*. **31**, 318–323 (2015).
38. K. Nho, J. D. West, H. Li, R. Henschel, A. Bharthur, M. C. Tavares, A. J. Saykin, Comparison of Multi-Sample Variant Calling Methods for Whole Genome Sequencing. *IEEE Int Conf Systems Biol*. **2014**, 59–62 (2014).
39. A. M. Novak, E. Garrison, B. Paten, A graph extension of the positional Burrows-Wheeler transform and its applications. *Algorithms Mol. Biol.* **12**, 18 (2017).
40. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*. **12**, 357–360 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

41. J. Huang, B. Howie, S. McCarthy, Y. Memari, K. Walter, J. L. Min, P. Danecek, G. Malerba, E. Trabetti, H.-F. Zheng, UK10K Consortium, G. Gambaro, J. B. Richards, R. Durbin, N. J. Timpson, J. Marchini, N. Soranzo, Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).

42. E. M. van Leeuwen, L. C. Karssen, J. Deelen, A. Isaacs, C. Medina-Gomez, H. Mbarek, A. Kanterakis, S. Trompet, I. Postmus, N. Verweij, D. J. van Enckevort, J. E. Huffman, C. C. White, M. F. Feitosa, T. M. Bartz, A. Manichaikul, P. K. Joshi, G. M. Peloso, P. Deelen, F. van Dijk, G. Willemsen, E. J. de Geus, Y. Milaneschi, B. W. J. H. Penninx, L. C. Francioli, A. Menelaou, S. L. Pulit, F. Rivadeneira, A. Hofman, B. A. Oostra, O. H. Franco, I. Mateo Leach, M. Beekman, A. J. M. de Craen, H.-W. Uh, H. Trochet, L. J. Hocking, D. J. Porteous, N. Sattar, C. J. Packard, B. M. Buckley, J. A. Brody, J. C. Bis, J. I. Rotter, J. C. Mychaleckyj, H. Campbell, Q. Duan, L. A. Lange, J. F. Wilson, C. Hayward, O. Polasek, V. Vitart, I. Rudan, A. F. Wright, S. S. Rich, B. M. Psaty, I. B. Borecki, P. M. Kearney, D. J. Stott, L. Adrienne Cupples, Genome of The Netherlands Consortium, J. W. Jukema, P. van der Harst, E. J. Sijbrands, J.-J. Hottenga, A. G. Uitterlinden, M. A. Swertz, G.-J. B. van Ommen, P. I. W. de Bakker, P. Eline Slagboom, D. I. Boomsma, C. Wijmenga, C. M. van Duijn, Genome of The Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat. Commun.* **6**, 6065 (2015).

43. M. Nagasaki, J. Yasuda, F. Katsuoka, N. Nariai, K. Kojima, Y. Kawai, Y. Yamaguchi-Kabata, J. Yokozawa, I. Danjoh, S. Saito, Y. Sato, T. Mimori, K. Tsuda, R. Saito, X. Pan, S. Nishikawa, S. Ito, Y. Kuroki, O. Tanabe, N. Fuse, S. Kuriyama, H. Kiyomoto, A. Hozawa, N. Minegishi, J. Douglas Engel, K. Kinoshita, S. Kure, N. Yaegashi, ToMMo Japanese Reference Panel Project, M. Yamamoto, Rare variant discovery by deep whole-genome

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).

44. A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, E. E. Kenny, Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
45. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
46. R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard, S. E. Devine, An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).
47. Global Alliance for Genomics and Health Benchmarking Workgroup, *Benchmarking Performance Metrics Definitions for SNVs and Small Indels* (https://github.com/ga4gh/benchmarking-tools/blob/master/doc/standards/GA4GHBenchmarkingPerformanceMetricsDefinitions.md).
48. J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, S. W. Scherer, The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–92 (2014).
49. K. Chen, L. Chen, X. Fan, J. Wallis, L. Ding, G. Weinstock, TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res.* **24**, 310–317 (2014).
50. J. G. Cleary, R. Braithwaite, K. Gaastra, B. S. Hilbush, S. Inglis, S. A. Irvine, A. Jackson, R. Littin, M. Rathod, D. Ware, J. M. Zook, L. Trigg, F. M. De La Vega, Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv* (2015), p. 023754.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

51. J. Tarhio, E. Ukkonen, Approximate Boyer–Moore String Matching. *SIAM J. Comput.* **22**, 243–260 (1993).

52. R. S. Boyer, J. S. Moore, A Fast String Searching Algorithm. *Commun. ACM*. **20**, 762–772 (1977).

53. M. Zhao, W.-P. Lee, E. P. Garrison, G. T. Marth, SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS One*. **8**, e82138 (2013).

54. J. Bentley, Programming Pearls: Algorithm Design Techniques. *Commun. ACM*. **27**, 865–873 (1984).

55. P. E. C. Compeau, P. A. Pevzner, G. Tesler, How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* **29**, 987–991 (2011).

56. R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, 1998).

57. G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, M. A. DePristo, From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*. **43**, 11.10.1–33 (2013).

58. 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean, A map of human genome variation from population-scale sequencing. *Nature*. **467**, 1061–1073 (2010).

59. J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, M. Salit, Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

60. D. A. Dmitriev, R. A. Rakitov, Decoding of superimposed traces produced by direct sequencing of heterozygous indels. *PLoS Comput. Biol.* **4**, e1000113 (2008).

61. Ø. A. Haaland, H. J. Skaug, Estimating genotyping error rates from parent–offspring dyads. *Stat. Probab. Lett.* **83**, 812–819 (2013).

62. J. A. Douglas, A. D. Skol, M. Boehnke, Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am. J. Hum. Genet.* **70**, 487–495 (2002).

63. J. Wang, Sibship reconstruction from genetic data with typing errors. *Genetics*. **166**, 1963–1979 (2004).

64. P. C. D. Johnson, D. T. Haydon, Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. *Genetics*. **175**, 827–842 (2007).

65. M. Korostishevsky, I. Malkin, T. Spector, G. Livshits, Parametric model-based statistics for possible genotyping errors and sample stratification in sibling-pair SNP data. *Genet. Epidemiol.* **34**, 26–33 (2010).

66. I. W. Saunders, J. Brohede, G. N. Hannan, Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. *Genomics*. **90**, 291–296 (2007).

67. B. L. Browning, S. R. Browning, Detecting identity by descent and estimating genotype error rates in sequence data. *Am. J. Hum. Genet.* **93**, 840–851 (2013).

68. E. Sobel, J. C. Papp, K. Lange, Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.* **70**, 496–508 (2002).

69. L. Jostins, Inferring genotyping error rates from genotyped trios. *arXiv [q-bio.QM]* (2011),

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(available at http://arxiv.org/abs/1109.1462).

70. I. M. Heid, C. Lamina, H. Küchenhoff, G. Fischer, N. Klopp, M. Kolz, H. Grallert, C. Vollmert, S. Wagner, C. Huth, J. Müller, M. Müller, S. C. Hunt, A. Peters, B. Paulweber, H.-E. Wichmann, F. Kronenberg, T. Illig, Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *Am. J. Epidemiol.* **168**, 878–889 (2008).

71. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

Acknowledgements:

We would also like to thank members of the GA4GH Data Workgroup, Benchmarking, and Reference variation initiatives, in particular Dr. Justin Zook, for insightful discussions and ideas. Dr. Maxime Huvet helped refine the treatment and presentation of ideas behind trio-based benchmarking.

Research reported in this publication was supported in part by UK Department of Health grant SBRI Genomics Competition: Enabling Technologies for Genomic Sequence Data Analysis and Interpretation administered by Genomics England.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

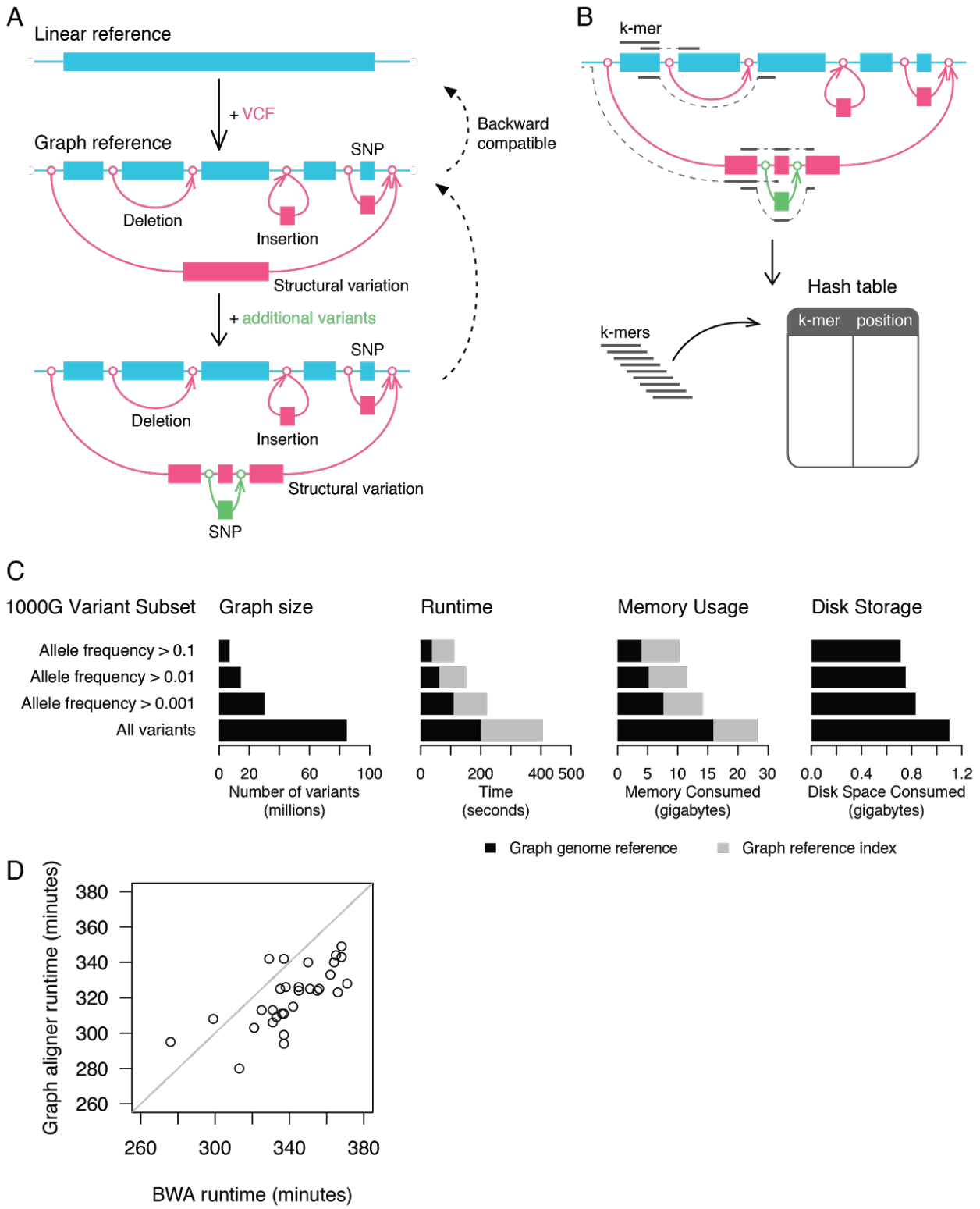

Figure 1. The graph genome architecture and computational resource requirements

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**A**) A graph genome is constructed from a standard linear reference genome in FASTA format augmented by a set of genetic variants provided in VCF format. A graph genome can further be augmented with additional genetic variants in a second VCF file, or in the case of variants within variants using our graph genome toolkit. The coordinate system of each constructed graph genome is backward compatible with that of the linear reference genome.

(**B**) A graph genome is indexed by creating a hash table with *k*-mers along all possible paths of the graph as keys and their corresponding graph genome positions as values. These *k*-mer positions can then be used as seeds for aligning sequencing reads against the graph.

(**C**) Computational resource requirements of building, indexing and storing graph genomes. The resource usage statistics are provided for the standard linear reference genome (GRCh37), which was augmented with different subsets of the 1000G variants. All tests were performed using a single thread on the Amazon AWS instance type *c4.8xlarge*.

(**D**) Runtimes for BWA and the graph aligner for 30 randomly selected samples from the Coriell cohort. The underlying graph genome contains around 20 million variants and is described in detail in Materials and Methods. Both BWA-MEM and the graph aligner were executed using 36 threads on the Amazon AWS cloud instance type *c4.8xlarge*.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

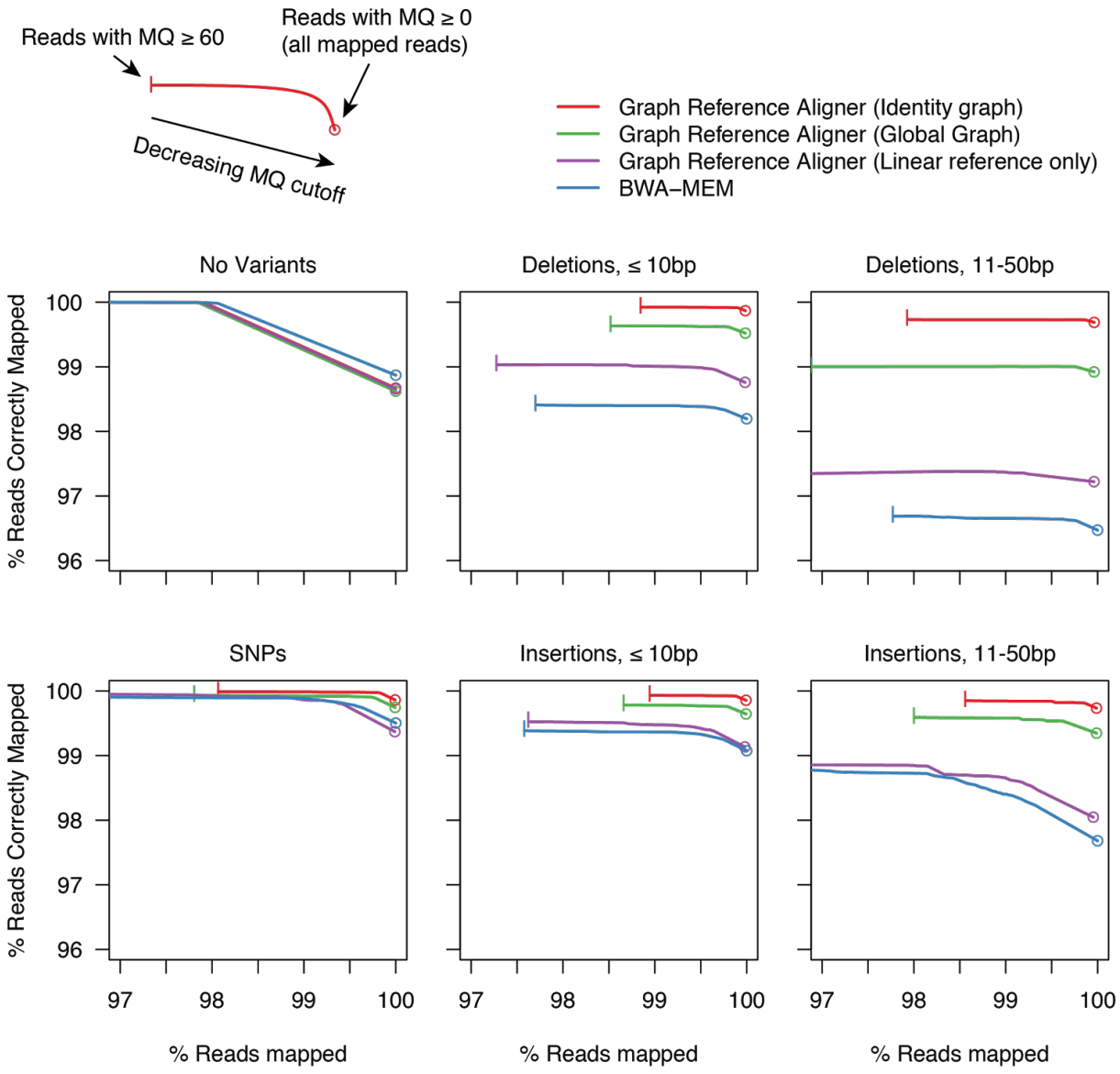

Figure 2. Read mapping accuracy using BWA-MEM and graph genomes

Simulated reads were divided into six categories based on the type of simulated variants they contain relative to the linear reference genome. The percentage of reads mapped correctly was plotted against the percentage of reads mapped for a range of mapping quality (MQ) cutoffs. 'Identity graph' refers to a graph genome containing only the genetic variants present in the respective target sample.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

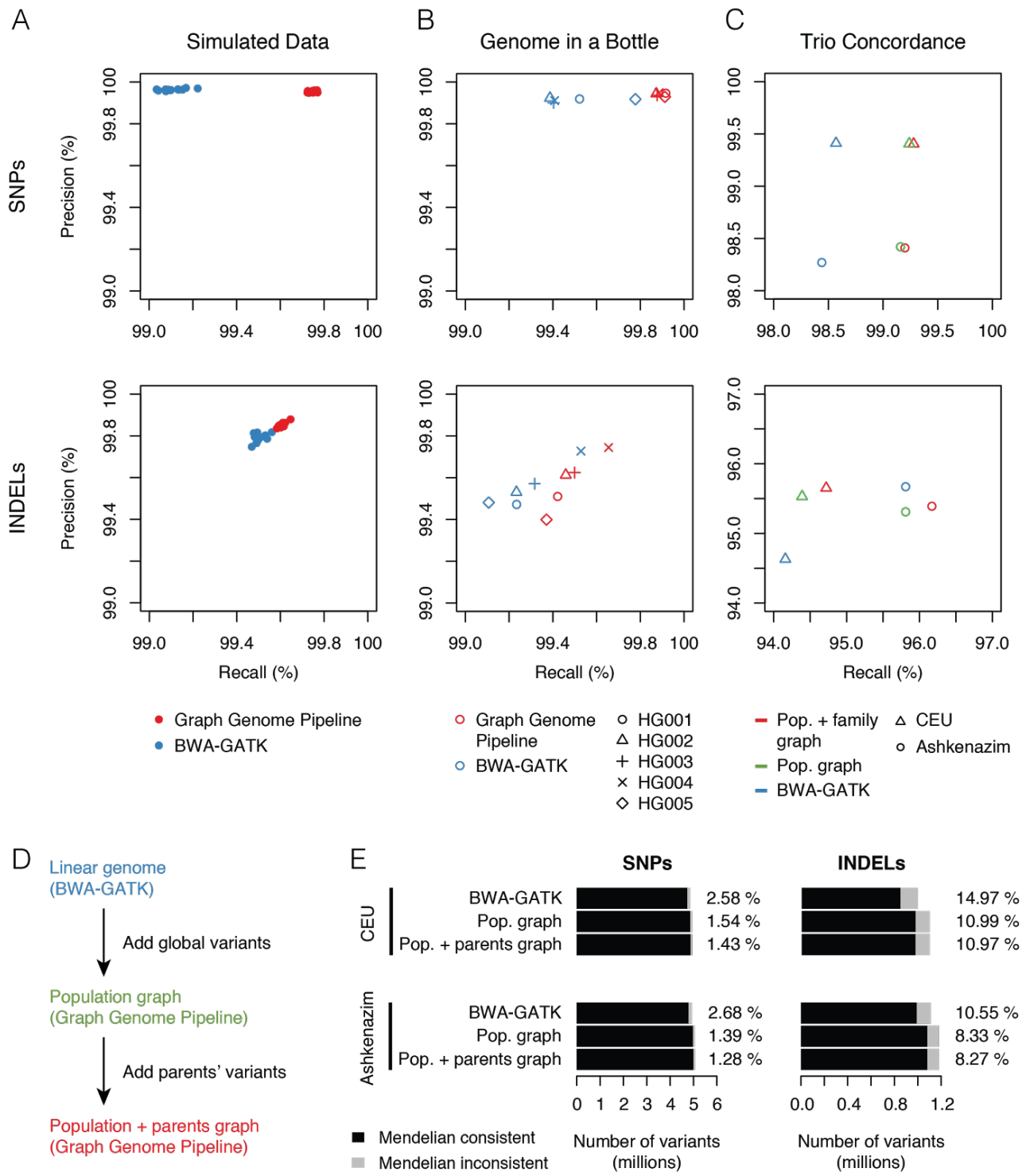

Figure 3. Variant calling benchmarking between Graph Genome Pipeline and BWA-GATK

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**A**) Precision and recall in simulated sequencing data. Every point corresponds to a different simulated individual.

(**B**) Precision and recall benchmarking based on five samples for which the GIAB truth sets are available (we show BWA-GATK results with Hard Filtering as that approach performed better out of the two recommended options; Figure S12 also shows results after VQSR filtering).

(**C**) Statistically estimated precision and recall from the Mendelian consistency data.

(**D**) Schematic representation of the trio concordance analysis.

(**E**) Mendelian concordance of two trios analyzed using BWA-GATK, Graph Genome Pipeline with the global graph and Graph Genome Pipeline using a global graph augmented by the variants detected in either parent of each family.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

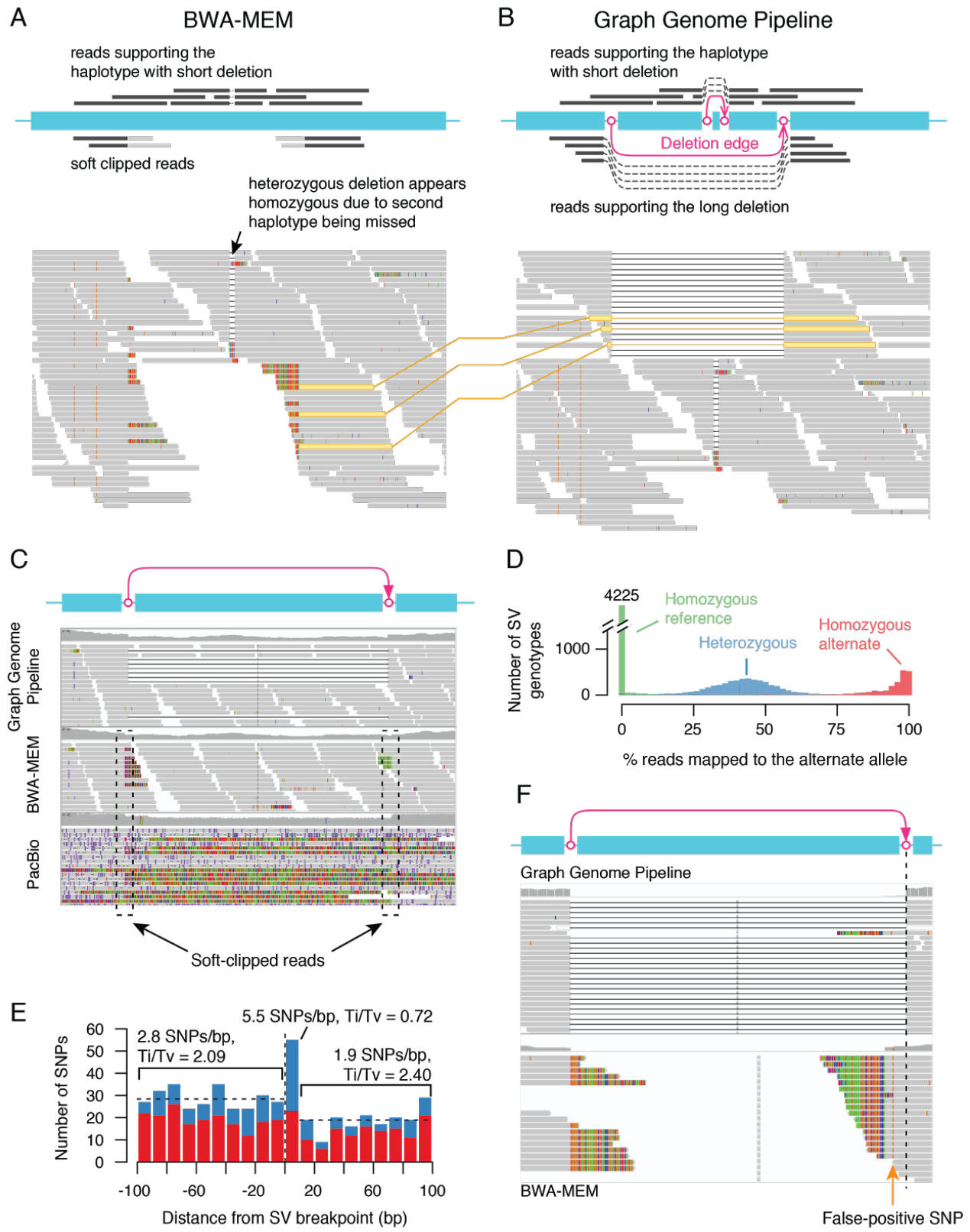

Figure 4. SV genotyping using graph genomes

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**A**) An Integrative Genomics Viewer (*44*) (IGV) screenshot of the BWA alignment over an SV at chromosome 4 position 132,524,277-132,524,548 (GRCh38, Database of Genomic Variants ID: esv3602264), which in this sample (HG01628) overlaps with a small deletion variant. BWA is able to align reads across the small deletion, but not across the SV. Instead, reads on the SV haplotype become soft-clipped (colored vertical bars at the tips of the reads facing the SV). As a consequence of the SV not being called, the small intervening deletion appears homozygous in the read alignment. A schematic of the read alignment pattern is shown above the screenshot.

(**B**) An IGV screenshot of the Graph Genome Pipeline read alignment from the same sample and at the same locus as in panel **A**. A schematic of the local graph structure and read alignment on the graph is depicted above the screenshot. Thanks to the SV being included in the graph genome, reads are correctly aligned across it, revealing the SV haplotype. Reads linked with a yellow line in panels **A** and **B** are the same reads.

(**C**) An IGV screenshot of an SV at chromosome 18 position 26,991,656-26,992,377 (GRCh38, Database of Genomic Variants ID: esv3642033) in sample HG002. This SV is in complete linkage disequilibrium with SNP rs1436904, which is significantly associated with breast cancer risk. Colored vertical bars at the tips of reads indicate soft-clipping. Whereas the graph aligner (top panel) is able to correctly align reads across the SV, BWA fails to do so (middle panel). Similarly, BWA-aligned PacBio reads (bottom panel) using parameters optimized for PacBio (option '-x pacbio' in BWA) also fail to align across the SV.

(**D**) The fraction of reads aligning to an SV branch by 1000G SV genotype (*9*), summarized across 230 SVs and 49 individuals.

(**E**) Total number of transition and transversion SNPs reported by 1000G around the breakpoints of the 230 deletion SVs, grouped by distance from a SV breakpoint in 10bp bins. Positive

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distances (i.e. >0bp) are within a SV, while negative distances (i.e. <0bp) are outside an SV. For example, positions +1 to +10bp correspond to the 10bp closest to an SV breakpoint within a deletion, while positions -1 to -10 bp correspond to the 10bp closest to an SV breakpoint outside a deletion. Red and blue portions of each bar correspond to transition and transversion counts at each bin.

(**F**) Example of an alignment that causes a false positive SNP due to misalignment against the linear reference genome. This sample has a homozygous deletion in this region, and graph aligner (top) aligns reads successfully across it. BWA-MEM (bottom) fails to align reads across the SV, but since the SV has 20bp of imperfect microhomology at the breakpoint, BWA-MEM aligns the reads on the right hand side over the imperfect microhomology region, resulting in a spurious homozygous G>A SNP call.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

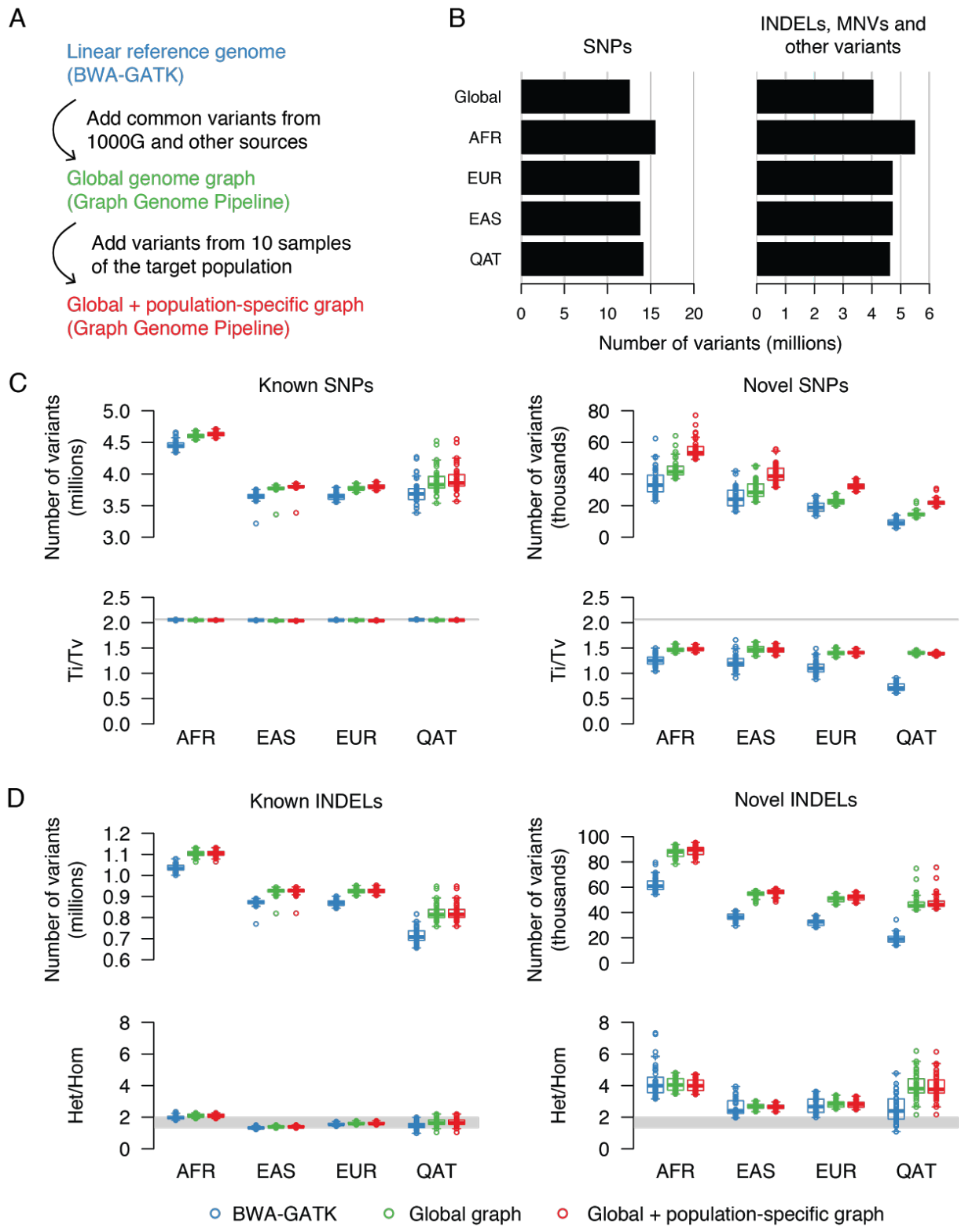

Figure 5. The effect of iteratively augmented graph genomes on variant calling

(**A**) Schematic representation of the graphs generated in the graph augmentation experiment.

bioRxiv preprint doi: https://doi.org/10.1101/194530; this version posted September 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**B**) Numbers of SNPs, INDELs and other variant types in the global and augmented graphs.

(**C**) Counts and transition/transversion ratios of known and novel SNPs called through the BWA-GATK, global graph and augmented graph pipelines. Grey horizontal lines indicate the expected transition/transversion genome-wide ratio (*36*).

(**D**) Counts and het/hom ratios of known and novel INDELs called through the BWA-GATK, global graph and population-augmented graph pipelines. Horizontal bars indicate the average value of each group. Gray rectangles indicate the expected range of population averages for heterozygous/homozygous alternate ratios (*36*).

Supplementary Materials:

Materials and Methods

Figs. S1 to S12

Tables S1 to S3

Captions for Tables S4 to S6

Table S4 (Rakocevic et al. 2017, Graph Genomes.Table-S4-S5.xslx)

Table S5 (Rakocevic et al. 2017, Graph Genomes.Table-S4-S5.xlsx)

Table S6 (Rakocevic et al. 2017, Graph Genomes.Table-S6.xlsx)