Open access • Journal Article • DOI:10.1038/NG.3964

Graphtyper enables population-scale genotyping using pangenome graphs

Hannes P. Eggertsson, Hannes P. Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir ...+18 more authors

Institutions: University of Iceland, Amgen, Reykjavík University, RMIT University

Published on: 25 Sep 2017 - Nature Genetics (Nature Publishing Group)

Topics: Genotyping, Sequence analysis, Sequence (medicine) and Population

Title

Graphtyper: Population-scale genotyping using pangenome graphs

Authors

Hannes P. Eggertsson[1,2], Hakon Jonsson[1], Snaedis Kristmundsdottir[1,3], Eirikur Hjartarson[1], Birte Kehr[1], Gisli Masson[1], Florian Zink[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Ingileif Jonsdottir[1,4], Daniel F. Gudbjartsson[1,2], Pall Melsted[1,2], Kari Stefansson[1,4], Bjarni V. Halldorsson[1,3]

[1]deCODE genetics / Amgen Inc., Sturlugata 8, Reykjavik, Iceland

[2]School of Engineering and Natural Sciences, University of Iceland, Reykjavík, Iceland

[3]School of Science and Engineering, Reykjavik University, Reykjavík, Iceland

[4]Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland

Corresponding authors: Hannes P. Eggertsson (hannese@decode.is), Bjarni V. Halldorsson (bjarnih@decode.is)

Abstract

A fundamental requisite for genetic studies is an accurate determination of sequence variation. While human genome sequence diversity is increasingly well characterized, there is a need for efficient ways to utilize this knowledge in sequence analysis. Here we present Graphtyper, a publicly available novel algorithm and software for discovering and genotyping sequence variants. Graphtyper realigns short-read sequence data to a pangenome, a variation-aware graph structure that encodes sequence variation within a population by representing possible haplotypes as graph paths. Our results show that Graphtyper is fast, highly scalable, and provides sensitive and accurate genotype calls. Graphtyper genotyped 89.4 million sequence variants in whole-genomes of 28,075 Icelanders using less than 100,000 CPU days, including detailed genotyping of six human leukocyte antigen (HLA) genes. We show that Graphtyper is a valuable tool in characterizing sequence variation in population-scale sequencing studies.

Introduction

Advances in DNA sequencing technology have improved characterization of sequence diversity in the human genome and have resulted in refinements of the reference sequence[1–4]. The human reference sequence is extremely useful, but it represents a consensus of genomes and therefore it does not capture sequence variation within or between populations[5,6].

In the latest version of the human reference genome (GRCh38), there are several alternate loci where the sequence variation is too complex to be represented with a single sequence. These loci are generally highly polymorphic, and many are known to co-segregate with disease and are therefore of great interest in population genetics. The most prominent example, the human leukocyte antigen (HLA) region, is known to associate with a number of immune mediated human diseases[7]. Given the importance of this region, it has been further characterized in the IPD-IMGT/HLA database[8], which contains a large collection of known HLA allele sequences. Such variation should be included in genome diversity analyzes.

Short-read sequencing is the standard in genome-wide sequence analysis. Most common approaches for discovering sequence variants involve aligning sequence reads to a reference genome[9] and searching for variants as alternative sequences in read alignments (Figure 1a i). However, some reads cannot be aligned to a reference genome, particularly those originating from highly polymorphic regions and regions absent from the reference genome. Reference genome alignments are also generally done without awareness of variation, causing mapping bias towards the reference allele and misalignments around indels[10,11].

Richer data structures that utilize the large amount of available sequence variation data promise to alleviate some of the limitations of previous methods[12–15]. Although approaches that find polymorphisms in reference-free assemblies have been developed to avoid these limitations[16,17], *de novo* assembly algorithms remain computationally expensive, have less sensitivity[17], and use data structures that have a complex coordinate system.

Pangenomes[12,18,19] have recently been proposed to counter weaknesses of both reference alignments and *de novo* assemblies by extending the linear reference alignments with variation-aware alignments[20]. Pangenomes incorporate prior information about variation, allowing read aligners to better distinguish between sequencing errors in reads and true sequence variation. Unlike *de novo* assembly algorithms, pangenomes represent sequence variation with respect to the reference genome, enabling a direct access to its annotated biological features. Variation-aware data structures, such as pangenomes, also allow read mapping and genotype calling to be performed in a single step[12].

Graph-like data structures with directed edges have commonly been used to represent pangenomes[19,21–24]. In an idealized pangenome graph, nodes represent sequences and the sequence of every genotyped individual genome is a path in the graph, but not necessarily vice versa. A number of algorithms have recently been developed that tackle the problems of graph construction, indexing and alignment of sequence reads to graphs[19,21,25–27], Paten *et al*.[24] provide a recent survey of current efforts. However, there is no method that combines these operations and uses the resulting alignments to update the graph with novel variation for the purpose of variant calling[12].

Here we present Graphtyper, a method and software for discovering and genotyping sequence variants in large populations using pangenome graphs. Graphtyper realigns all

sequence reads of a genomic region, including unaligned and clipped sequences, to a variation-aware graph (Figure 1a ii). Concomitantly, it aligns sequence reads and genotypes sequence variants present in its graph. Furthermore, Graphtyper discovers novel single nucleotide polymorphisms (SNPs) and short sequence insertion or deletion variants (indels), which can be used to update the pangenome graph (Methods).

An important benefit of Graphtyper's realignment step is to improve read alignments near indels. Figure 2a shows how Graphtyper represents three common sequence variants, a 40-bp deletion and two SNPs. Using variation-aware realignment, Graphtyper is capable of better characterization of the region's variation than previous methods, with no Mendelian errors (Figure 2b) and no falsely reported additional sequence variants around the indel (Supplementary Table 1) due to misaligned sequence reads (Supplementary Figure 1).

Results

Data structures and genotyping pipeline Graphtyper uses a reference sequence and optionally all known sequence variants as input to construct pangenome graphs. Sequence reads mapped to a genomic region of the reference sequence, including unaligned and trimmed reads, are realigned to the pangenome graph. Using these graph alignments, Graphtyper discovers variants within the genomic region. This process is iterated several times (Supplementary Note 4), i.e., a pangenome graph is constructed, indexed and aligned with sequence reads, from which novel variants are discovered and previously discovered variants are genotyped (Figure 1b).

The underlying pangenome data structure is a directed acyclic graph (DAG) where edges connect nodes that contain a DNA sequence (Supplementary Note 1). Graphtyper takes as input a reference genome and a list of known variants. Each known variant is a record of a chromosomal position, a reference allele, and one or more alternative alleles. First, variant records with overlapping reference alleles are merged into a single record (Figure 3a). Second, *allele nodes* are constructed, containing the sequence and start position of each allele of the variant records. Third, *reference nodes* are constructed between two adjacent variant records, storing the corresponding reference sequence and its start position. Finally, nodes at adjacent positions are connected. Paths in the graph alternate between *reference* and *allele* nodes and nodes that share a start position are parallel to each other. Each character in an allele node sequence is given a position equal to the first position of the node plus the character's offset from that position (Figure 3b). Allele node positions longer than the reference allele are assigned new unique positions (z_1 and z_2 in Figure 3b) to avoid

conflicts with the following positions. The final graph represents the reference sequence and all haplotypes in the population as paths.

Aligning sequence reads by traversing the graph is time consuming. To expedite graph alignments, the graph structure is preprocessed by creating an index that maps k-mers to their start and end positions in the reference genome and to overlapping allele nodes (if any) (Figure 3c, Methods). Read alignment then follows the seed-and-extend paradigm (Figure 3d-3h, Methods, and Supplementary Note 2).

The output of each iteration is a file in variant-call format (VCF) including both newly and previously discovered variants, which Graphtyper uses to update the graph in the next iteration (Methods).

Population-scale genotyping We compared Graphtyper to seven widely used genotyping pipelines on human chromosome 21 in a set of 691 whole-genome sequenced Icelanders (Table 1). Of these, 404 individuals were contained in 230 trios (parent-offspring trio families). The genotypers used were Genome Analysis ToolKit UnifiedGenotyper (GATK UG)[28], GATK-Lite UnifiedGenotyper (UGLite), GATK HaplotypeCaller (HC), GATK HC GVCF joint genotyping (HC joint), Samtools[29], Platypus[17], and FreeBayes[30] (Supplementary Note 4). Known sequence variants were not given to Graphtyper as input, all pipelines were given the same BAM files and reference sequence (GRCh38).

Our results show that GATK UG, Graphtyper and Samtools all had comparable compute times and completed the genotyping in between 576 and 594 hours (Table 1). The other five genotypers required considerably greater compute times (1,030-12,964 hours).

We assessed the raw output of all eight genotyping pipelines to compare them independent of filtering technique and to include analysis of all germline variation, somatic variation, and

wrongfully reported variation due to sequencing or alignment errors. Compared to other genotypers, Graphtyper called a large number of SNPs (406,087) with a reasonably high ratio of transitions (Ti) to transversions (Tv) (1.49). We observed that all eight genotypers had a large excess of alternative alleles with a transmission rate below 50% (Supplementary Figure 2). We also observed higher Ti/Tv ratios among alleles with higher transmission rates (Supplementary Figure 3). Motivated by these realizations, we estimated the number of germline alternative alleles based on the transmission rate of the alternative alleles in the 230 trios (Methods). Graphtyper detected the largest number of estimated germline alternative alleles in the trios (267,057), followed by GATK UGLite (264,753) and GATK UG (264,447) (Table 1).

We found 105,302 SNPs and 7,694 indels that were called by all eight genotypers and have been reported as common (minor allele frequency > 1% in any population) in dbSNP build 149. In the 230 trios, Graphtyper called these sequence variants with a mean transmission rate of 49.98%, very close to the expected 50%. Graphtyper had the highest Mendelian accuracy (99.52%) and the lowest number of missing genotype calls (0.201%) (Table 1). We also compared SNP calls to our in-house microarray genotypes (Methods), all genotyping pipelines were highly concordant (>99%).

From our comparison of genotypers, we concluded that Graphtyper and GATK UG were the two best genotypers for population-scale genotyping in terms of performance, accuracy and sensitivity. We assessed a call set of highly confident Graphtyper sequence variants using our own filtering criteria and filtered the GATK call sets (UG, HC and HC joint) using their available ‘best practices’ filtering criteria (Supplementary Note 4). Graphtyper achieved substantially lower estimate of false discovery rate (FDR) (2.19%) than the other call sets

(10.26-31.22%), but also had lower estimated number of germline alternative alleles (200,984) than the other call sets (214,801-240,020) (Supplementary Table 2).

We measured their scalability by genotyping chromosome 21 on a dataset of 15,220 Icelanders, in which there are 1,729 trios (3,863 unique individuals). Our results show that Graphtyper scales much better than GATK UG (Figure 4), with GATK UG using approximately 2.5x more time for computations than Graphtyper (Table 2). The compute time used by Graphtyper per sample did not increase substantially when the sample size increased from 691 to 15,220 (changed from 0.842 hr/sample to 0.867 hr/sample), while GATK UG used 2.65x more compute time per sample (changed from 0.834 hr/sample to 2.206 hr/sample).

Based on the transmission of alternative alleles the 1,729 trios, we observed that the FDR increased for Graphtyper and GATK UG compared to the 230 trio dataset in both raw and filtered call sets. We estimated that Graphtyper detected more germline alternative alleles (308,204) with a significantly lower FDR (8.89%) than GATK UG (305,404 and 22.62%, respectively) in the filtered call sets (Table 2).

Single sample genotyping We assessed the single sample genotyping performance of Graphtyper on a well-studied parent-offspring trio (NA12878, NA12891 and NA12892). Whole-genome sequence data (50x 101-bp paired-end Illumina HiSeq 2000) of these samples are publicly available through the Platinum Genome project[31]. We genotyped each sample independently using the same genotyping pipelines as in our population-scale experiment. We ran Graphtyper with and without initializing its graph structure with publicly available common (minor allele frequency > 1% in any population) sequence variants (dbSNP build 150).

We assessed sequence variant call sets of the offspring (NA12878) by comparing it to the set of publicly available high-confidence variant calls[31] to measure variant recall rate and precision. Based on the genotyping of the parents (NA12891 and NA12892), we estimated FDR and the number of transmitted germline alternative alleles in the trio (Methods).

Our results show that even without the knowledge of known variation, Graphtyper has a considerably better recall rate (98.14%) than the other genotypers (90.24-95.91%), high precision (99.774%), and overall the highest number of validated calls (4,081,193) (Table 3). As expected, the knowledge of common dbSNP variants increased Graphtyper's recall rate (to 98.46%), in particular at non-SNP sites where it increased from 91.23% to 93.38%. Consistent with its measured high recall rate, we also estimated that Graphtyper called the highest number of germline alternative alleles in the trio (5,991,012 and 5,874,556 with and without dbSNPs, respectively), substantially more than the other genotypers (5,190,838-5,562,776). However, Graphtyper had the longest compute time (154.1 hours), as the time of constructing and indexing a graph is relatively long for only a single sample.

We also filtered the Graphtyper call sets (Supplementary Note 4) and compared it with GATK's call sets filtered according their 'best practices' guidelines. After filtering, Graphtyper's recall rate was reduced to 96.47% and its estimated FDR reduced from 6.06% to 4.69% (Table 3).

28,075 Icelandic whole-genome samples We used Graphtyper to genotype the autosomes and chromosome X of 28,075 whole-genome sequenced Icelandic samples. The samples have a mean sequencing depth of 35.3x (s.d. 7.9x; range 2-200x) stored in a total of 2.12 PB of BAM files. The overall compute time for genotyping was 97,917 CPU days or 83.7 CPU

hours per sample on average. Graphtyper genotyped 89.4 million sequence variants: 1.1 million complex variants, 6.4 million indels, and 81.9 million SNPs with a Ti/Tv ratio of 1.04.

The compute time of genotyping chromosome 21 in 28,075 Icelandic samples was 27,853 CPU hours or 0.99 CPU hours per sample on average. Compared to Graphtyper's chromosome 21 genotyping of 691 samples, the sample size 40-folded, the number of sequence variants increased by 220%, but the compute time per sample only increased by 17.6%.

HLA typing The IPD-IMGT/HLA database[8] contains known HLA allele sequences identified with a field (usually two digits) hierarchical colon separated identifier. The first field denotes the HLA allele family, the second field denotes the subtype within the family, the third field denotes groups with synonymous substitutions within the subtype, and the fourth field denotes allele differences in non-coding regions.

Based on known HLA allele sequences, we created graphs for six important HLA genes: HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQA1, and HLA-DQB1 (Methods). Using these graphs, we were able to HLA type the same dataset of 28,075 Icelanders in a single genotyping-only iteration. Our results show high diversity of HLA allele families in the Icelandic population (Supplementary Table 3).

The total compute time of the HLA genotyping of the six genes was 2,609 hours, or 5.6 minutes per sample. The compute time of Graphtyper for the HLA region was orders of magnitudes lower than other genotypers[32,13] (Supplementary Note 6). Previously, deCODE genetics laboratory performed HLA typing of the six genes with a PCR based method at 2-digit ($n = 647$) and 4-digit ($n = 368$) resolutions. These previous typings are in good concordance (95.1-100% 2-digit; 91.6-100% 4-digit) with Graphtyper's HLA genotype calls

(Table 4). Upon manual inspection, we concluded that a large fraction of the discrepancy between the two methods are most likely explained by sample mix-up (Supplementary Note 6).

Discussion

Previous genotypers use read alignments to linear reference genomes, which limits their performance in polymorphic regions. To better characterize sequence diversity we implemented a novel variation-aware data structure and developed efficient algorithms in a software called Graphtyper. Graphtyper locally realigns sequence reads from a genomic region to a pangenome graph, and concomitantly genotypes sequence variants in all individuals. We show that combining these two steps is not only practical, but improves sensitivity and is more scalable than other genotyping methods. Our results show that Graphtyper has the highest Mendelian accuracy at previously reported variant sites among the genotypers in our comparison.

Graphtyper can use known variants as input, further improving sensitivity. When using dbSNP as part of the input, Graphtyper fails to recall only 0.73% of SNP variants in the Platinum genome dataset, a rate 5 times lower than the 3.61% missed by the best competitor. Additionally, the graph representation allows us to construct graphs with known sequence variation in the HLA region and accurately genotype known alleles of six HLA genes. Our HLA types are in good concordance to previously PCR verified HLA types. Graphtyper's ability to determine genotype calls for more sequence variants, including those that have complex representation, such as the HLA region may help geneticists in characterizing genomes and their impact. Despite these successes, additional work is required, for example, currently Graphtyper cannot call structural variants.

The computational requirements of many genotypers are so large that it is infeasible to effectively apply them to population-sized data sets. For large datasets, the computational requirements of Graphtyper are significantly lower than previous methods, requiring full

utilization of a 10,000 core computer cluster for 10 days, compared to an estimated minimum of 25 days for GATK UG.

It is important to note that our current pipeline still relies on the linear reference sequence and BWA for global read alignments in order to assign reads to a region. To completely remove bias towards the reference genome and fully utilize the promise of pangenome analysis requires developing robust methods for graph alignment, some of which are on the horizon[24,25,27]; one such notable project is vg (https://github.com/vgteam/vg). Our results further show the importance of replacing the linear reference with richer data structures to improve our understanding of how sequence diversity impacts diseases and other phenotypes.

References

1. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).
2. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–444 (2015).
3. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
4. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
5. Seo, J.-S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature* **538,** 243–247 (2016).
6. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
7. Tiwari, J. L. & Terasaki, P. I. *HLA and Disease Associations*. (Springer New York, 1985).
8. Robinson, J. *et al.* The IPD and IMGT/HLA database: Allele variant databases. *Nucleic Acids Res.* **43,** D423–D431 (2015).
9. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
10. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
11. Shao, H. *et al.* A population model for genotyping indels from next-generation sequence data. *Nucleic Acids Res.* **41,** e46–e46 (2013).
12. Computational Pan-Genomics Consortium, T. C. P.-G. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* bbw089 (2016). doi:10.1093/bib/bbw089
13. Dilthey, A. T. *et al.* High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. *PLOS Comput. Biol.* **12,** e1005151 (2016).
14. Paten, B., Novak, A. & Haussler, D. Mapping to a Reference Genome Structure. (2014).
15. Huang, L., Popic, V. & Batzoglou, S. Short read alignment with populations of genomes. *Bioinformatics* **29,** i361–i370 (2013).
16. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44,** 226–232 (2012).
17. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46,** 1–9 (2014).
18. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28,** 57–63 (2010).
19. Sirén, J., Välimäki, N. & Mäkinen, V. Indexing graphs for path queries with applications in genome research. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **11,** 375–388 (2014).
20. Schneeberger, K. *et al.* Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10,** R98 (2009).

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

22. Sirén, J. Indexing Variation Graphs. (2016). doi:10.1137/1.9781611974768.2

23. Church, D. M. *et al.* Extending reference assembly models. *Genome Biol.* **16,** 13 (2015).

24. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27,** 665–676 (2017).

25. Sirén, J. Indexing Variation Graphs. in *2017 Proceedings of the Ninteenth Workshop on Algorithm Engineering and Experiments (ALENEX)* 13–27 (Society for Industrial and Applied Mathematics, 2017). doi:10.1137/1.9781611974768.2

26. Kehr, B., Trappe, K., Holtgrewe, M. & Reinert, K. Genome alignment with graph data structures: a comparison. *BMC Bioinformatics* **15,** 99 (2014).

27. Maciuca, S., Elias, C. D. O., McVean, G. & Iqbal, Z. A natural encoding of genetic variation in a burrows-wheeler transform to enable mapping and genome inference. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9838,** 222–233 (2016).

28. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

29. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

30. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv Prepr. arXiv1207.3907* 9 (2012). doi:arXiv:1207.3907 [q-bio.GN]

31. Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27,** 157–164 (2017).

32. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30,** 3310–6 (2014).

33. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

34. Eggertsson, H. P. Gyper: A graph-based HLA genotyper using aligned DNA sequences. (2015).

Methods

Icelandic DNA data The Icelandic samples were whole-genome sequenced at deCODE genetics[2] using Illumina HiSeq and HiSeqX sequencing machines[33] and aligned to the GRCh38 human reference genome using the BWA MEM algorithm[9]. All sequenced individuals were also SNP chip typed using Illumina Human Hap or Omni chip arrays. DNA was isolated from both blood and buccal samples.

All participating subjects signed informed consent. Personal identities of the participants and biological samples were encrypted by a third party system approved and monitored by the Data Protection Authority. Approvals for these studies were provided by the National Bioethics Committee and the Data Protection Authority in Iceland.

Sequence read alignment In Graphtyper, sequence variation of small genomic regions (we used 50 kbp regions this study) are represented with a pangenome graph structure. Sequence reads are realigned to the graph of a region if BWA reported them to be in the same region. First, Graphtyper extracts a set of k-mers from the sequence read, which overlap by one DNA base in the read (Figure 3d), and determines if they are present in the graph using an index structure (Figure 3e). Seeds are generated from matches in the index look-up. If the alignments of two adjacent k-mers overlap by exactly one base, Graphtyper joins their matches into larger seeds (Figure 3g). The longest seeds are then extended (Figure 3h) by finding a path in the graph with the fewest mismatches using a breadth first search algorithm. If no seeds are extended with 12 mismatches or fewer, Graphtyper again extracts a set of k-mers from the read which overlap by one base in a read, but now also k-mers with one mismatch are included (Figure 3f). The process is applied both to a read and its reverse

complement. If both orientations of a read align to the graph, Graphtyper selects the longer alignment or, if they are equally long, the alignment with fewer mismatches.

Novel variant discovery Graphtyper post-processes graph alignments to discover novel small sequence variants. Novel sequence variants are classified as SNPs, indels (up to approx. 50 bp), and complex variation (e.g. multiple nucleotide polymorphisms and microsatellites). For each read uniquely aligned to the graph, Graphtyper starts by determining the position in the reference genome of its first and last aligned position in the graph and extracts the reference sequence between these two positions. Then on each side of the reference sequence, the read is extended by an additional 50 bases plus the number of soft clipped bases on the given side. The read is then locally aligned to the extracted reference sequence using a banded semi-global version of Gotoh's algorithm (Supplementary Figure 4a). Differences in the local alignments are treated as observations of variants (Supplementary Figure 4b).

Once all reads have been processed, Graphtyper outputs sequence variants where there exists a sample that has at least 5 observations of an alternative allele and its frequency is at least 20% (default values).

Genotyping Graphtyper genotype calls sequence variants in the graph by treating the graph alignments as independent observations of each sample's underlying genotype. It genotypes sequence variants in the graph by considering nearby variants together. Given graph-aligned sequence reads of a population, the likelihood that the reads were sampled from a pair of haplotypes is estimated for each sample and the haplotypes with the highest likelihood are determined. To greatly reduce the number of haplotypes considered, all sequence variants located 5 bp or less from each other are grouped (Supplementary Figure 5a) and each

variant group is genotyped independently. Let $H_i = \{h_{i,1}, h_{i,2}\}$ be a multiset of the unknown haplotypes of sample i in a variant group, v, and let $R_i = \{r_{i,1}, r_{i,2}, \ldots, r_{i,|R_i|}\}$ be the sample's multiset of sequence reads aligned by Graphtyper to the variant group v.

For each pair of possible haplotypes, a relative likelihood of the observed reads given the haplotypes $\mathcal{L}(R_i|H_i)$ is computed. We assume that the reads from one individual are independent of other individuals' reads. Graphtyper computes the relative likelihood iteratively as:

$$\mathcal{L}(R_i|H_i) = \prod_{r_{ij} \in R_i} L(r_{ij}|H_i) \quad (1)$$

where the relative likelihood of observing a read $r_{i,j}$ given the pair of underlying haplotypes is set as:

$$L(r_{ij}|H_i) = \begin{cases} 1 & \text{, if both } h_{i,1} \text{ and } h_{i,2} \text{ support the read.} \\ 1/2 & \text{, if exactly one of } h_{i,1} \text{ and } h_{i,2} \text{ support the read.} \\ \varepsilon_{r_{ij},H_i} & \text{, if neither } h_{i,1} \text{ nor } h_{i,2} \text{ support the read.} \end{cases} \quad (2)$$

where $\varepsilon_{r_{i,j},H_i}$ is the relative likelihood of observing an error, given the underlying haplotypes H_i and the read $r_{i,j}$. These relative likelihoods are chosen from the set $\left\{\frac{1}{2^5}, \frac{1}{2^6}, \ldots, \frac{1}{2^{13}}\right\}$ based on how similar the read is to the haplotypes H_i, the base pair quality, mapping quality of the read, and if the read is soft clipped (Supplementary Note 3). Restricting relative likelihoods to this set allows storing only the integer exponents, minimizing storage requirements and avoiding floating point precision problems.

As sequence variants are genotyped in groups, Graphtyper can identify the haplotypes in the population within each group (Supplementary Figure 5b) and remove unobserved haplotypes from the graph (Supplementary Figure 5c). This greatly reduces the number of haplotypes in complex regions.

Sequence variant quality assessment For each sequence variant we estimated the Mendelian error rate as the fraction of incorrectly inferred offspring in trios with two homozygous parents (Supplementary Figure 6a). We defined Mendelian inaccuracy as the estimated Mendelian error rate plus the fraction of trios with a missing genotype call, which are genotypes reported as "." or "./." in the VCF output.

If either parent is heterozygous we cannot deterministically infer the genotype of the offspring (Supplementary Figure 6b). For those trios we instead calculated the transmission rate of each alternative allele from parent to offspring. The expected transmission rate of germline alternative alleles is 50%. Falsely discovered variation due to sequencing errors and somatic mutations are assumed to transmit at a lower rate. We used the difference of alternative allele transmission rates above and below 50% to estimate the false discovery rate (FDR) using:

$$FDR_{\text{estimated}} = \max\left(\frac{\#(AA_{TMR<50\%}) - \#(AA_{TMR>50\%})}{\#(AA)}, 0\right) \quad (3)$$

Here, $\#(AA)$ is the number of called alternative alleles, and $\#(AA_{TMR>50\%})$ and $\#(AA_{TMR<50\%})$ are the number of alternative alleles with a transmission rate above and below 50%, respectively. We estimated the number of germline alternative alleles using:

$$\#(\text{Germline } AA)_{\text{estimated}} = \#(AA)(1 - FDR_{\text{estimated}}) \quad (4)$$

HLA typing pre-processing We retrieved HLA allele sequences from the IPD-IMGT/HLA database (version 3.23.0, see URLs). We extracted the differences to a VCF file that we used to create the pangenome graphs for HLA typing. A more detailed description of our HLA typing method as well as comparisons to other methods have been published in our previous work[34] and are described in Supplementary Note 7.

Author contributions HPE implemented the Graphtyper software. HPE, PM and BVH designed the Graphtyper algorithm. HPE, DFG, PM, BVH and KS designed the experiments. HPE, EH, GM and FZ ran all evaluated genotypers. HPE and HJ analyzed the call sets. Aslaug Jonasdottir, Adalbjorg Jonasdottir and IJ were responsible for PCR validation. HJ and SK contributed software for the project. HPE wrote the initial version of the manuscript, HJ, SK, BK, PM, BVH and KS contributed to subsequent versions. All authors reviewed and approved the final version of the manuscript.

URLs IPD-IMGT/HLA (http://www.ebi.ac.uk/ipd/imgt/hla/, Github page: https://github.com/ANHIG/IMGTHLA)

Code availability Graphtyper is available at https://github.com/DecodeGenetics/graphtyper (GNU GPLv3 license).

Figures

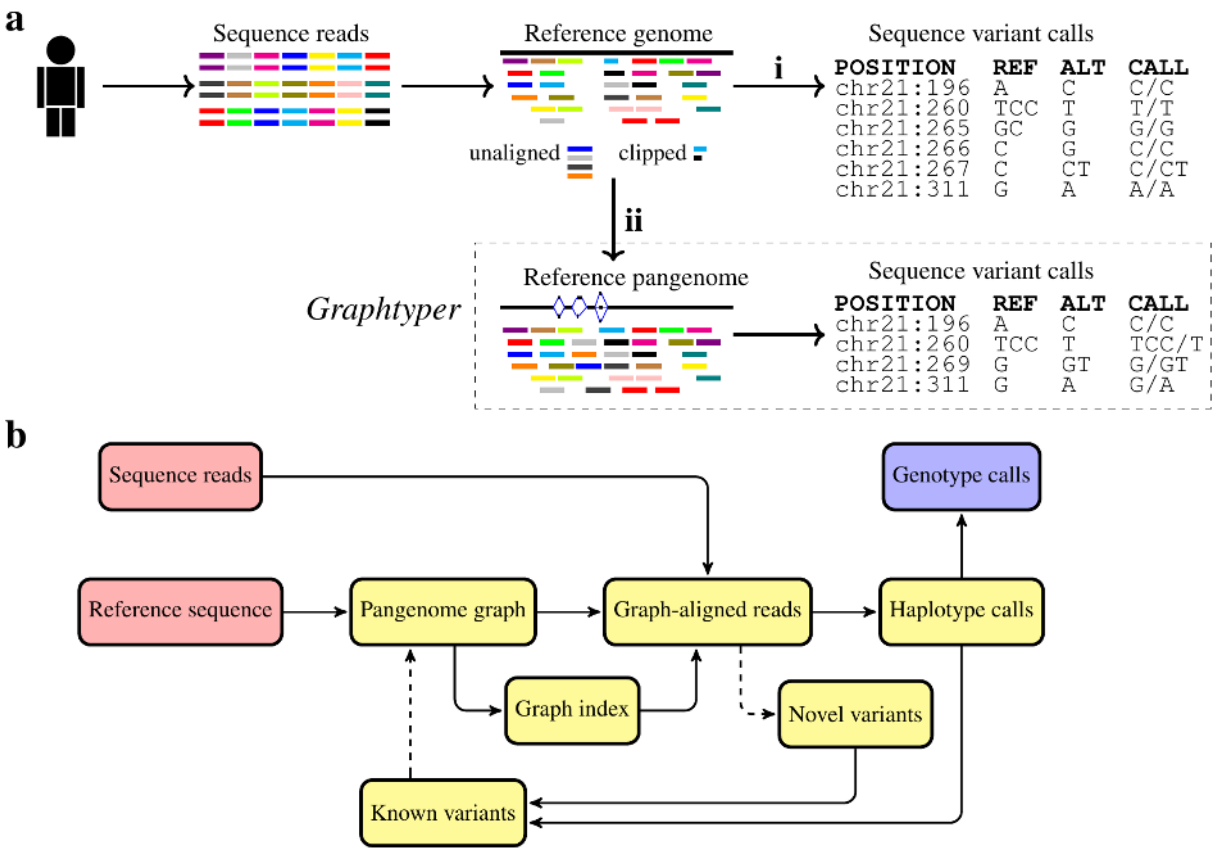


Figure 1: ***(a)*** *Overview of two genotyping pipeline designs.* ***(i)*** *A commonly used genotyping pipeline, where sequence reads are aligned to a reference genome sequence and sequence variant calls are made from sequence discordances between the sequence reads and the reference sequence.* ***(ii)*** *Graphtyper's genotyping pipeline. Sequence reads are realigned to a variant-aware pangenome graph and variants are called based on which path the reads align to.* ***(b)*** *Graphtyper's iterative genotyping process. Dashed paths are optional. As input, Graphtyper requires a reference genome sequence and sequence reads (red) and outputs genotype calls (blue) of variants.*

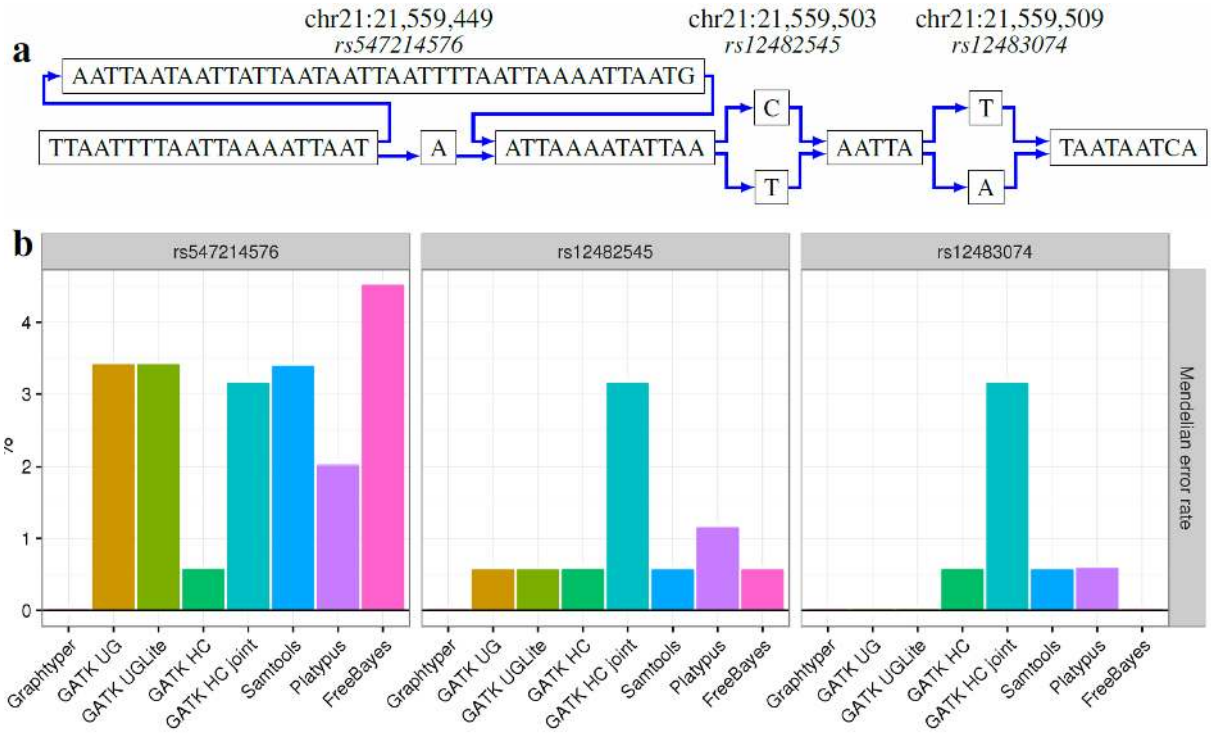


Figure 2: ***(a)*** *The genomic region chr21:21,559,430-21,559,518 (GRCh38) and three previously reported sequence variants represented with a pangenome graph.* ***(b)*** *Mendelian error rates of the three previously reported sequence variants called by eight genotypers. The Mendelian error rate is measured in 230 Icelandic parent-offspring trios.*

a

Reference sequence: ACCTCCAGACGTTTAGGGACCCCATTGAGTG

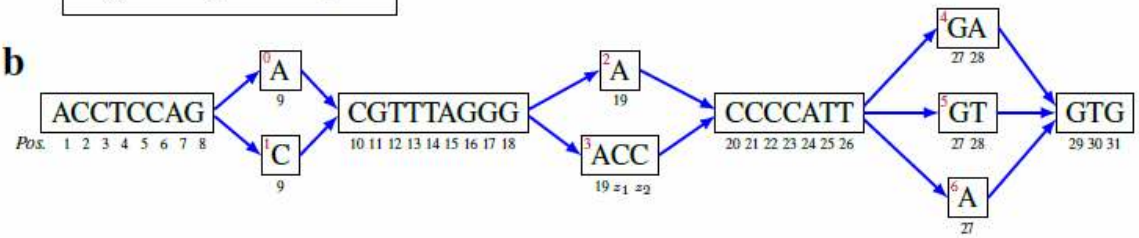
Known variants

Position	Reference	Alternative
9	A	C
19	A	ACC
27	GA	A
28	A	T

Known variants after merge

Position	Reference	Alternative
9	A	C
19	A	ACC
27	GA	GT,A

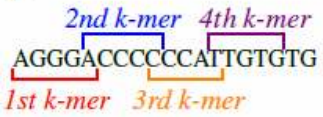
b



c

5-mer	Start pos.	End pos.	Variant ID	Start pos.	End pos.	Variant ID
ACCCC	19	23	2	19	21	3
ACCTC	1	5	NA			
ACGTT	9	13	0			
AGACG	7	11	0			
AGCCG	7	11	1			
AGGGA	15	19	2	15	19	3
...						
CCCAT	21	25	NA			
CCCCA	20	24	NA			
CCCCC	z_1	22	3	z_2	23	3
...						
GGACC	17	z_2	3	17	21	2
GGGAC	16	z_1	3	16	20	2
...						
TTGTG	25	29	5			
...						

d



e

5-mer	Start pos.	End pos.	Variant ID	Start pos.	End pos.	Variant ID
AGGGA	15	19	2	15	19	3
ACCCC	19	23	2	19	21	3
CCCAT	21	25	NA			
TTGTG	25	29	5			

f

AGGGA ACCCC CCCAT TTGTG
AGGGC ACCCA CCCAA TTGTA
AGGGG ACCCG CCCAC TTGTC
AGGGT ACCCT CCCAG TTGTT
AGGAA ACCAC CCCCT TTGAG
AGGCA ACCGC CCCGT TTGCG
AGGTA ACCTC CCCTT TTGGG
AGAGA ACACC CCAAT TTATG
AGCGA ACGCC CCGAT TTCTG
AGTGA ACTCC CCTAT TTTTG
AAGGA AACCC CACAT TAGTG
ACGGA AGCCC CGCAT TCGTG
ATGGA ATCCC CTCAT TGGTG
CGGGA CCCCC ACCAT ATGTG
GGGGA GCCCC GCCAT CTGTG
TGGGA TCCCC TCCAT GTGTG

g

Seed	Start pos.	End pos.	Variant IDs
AGGGACCCCCCATTGTG	15	29	3,5
AGGGACCCC	15	23	2

h

Extended longest seed	Start pos.	End pos.	Variant IDs
AGGGACCCCCCATTGTGTG	15	31	3,5

Figure 3: ***(a)*** *An example reference sequence and its known variation. All overlapping variants are merged.* ***(b)*** *Constructed pangenome reference graph. We draw the path of the reference sequence as the topmost path.* ***(c)*** *The index data structure with k = 5. 5-mers in the graph are mapped to a list of its start position, end position, and a variant ID which it overlaps, if any.* ***(d)*** *Four k-mers are extracted from a sequence read. Each k-mer overlaps its neighbor k-mer by one character.* ***(e)*** *An example look-up of the k-mers from the index data structure from c).* ***(f)*** *All extracted k-mers with a single substitution.* ***(g)*** *Seeds are generated from matches in the index look-up.* ***(h)*** *Final graph alignment after extending the longest seed.*

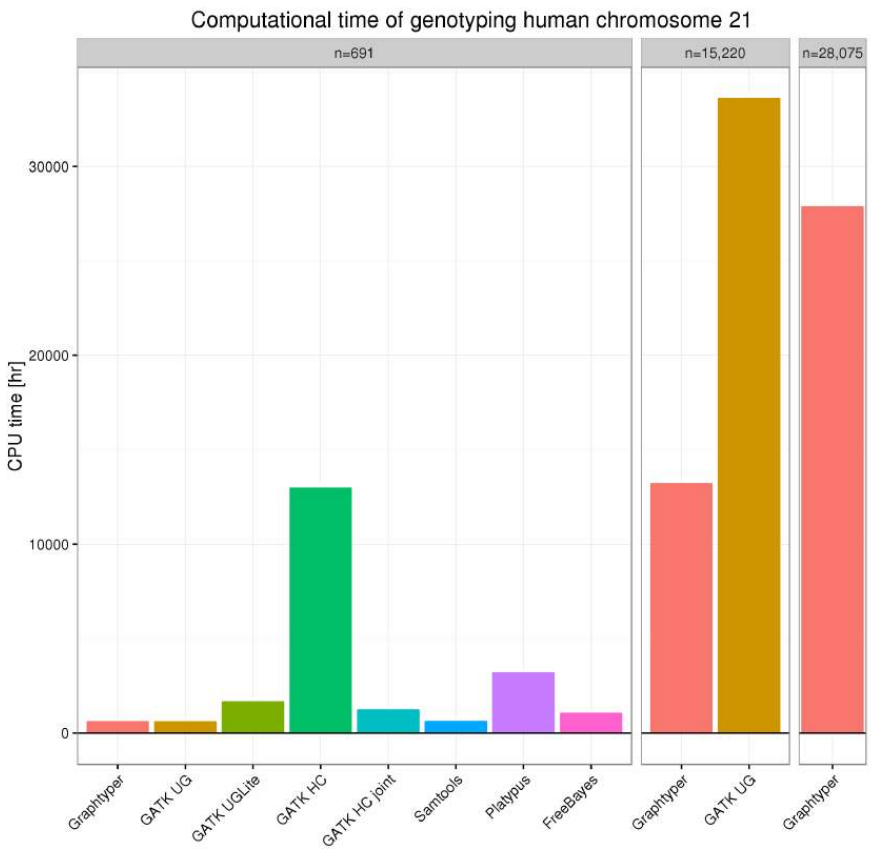


Figure 4: Comparison of compute times required to genotype chromosome 21 on three whole-genome sequence datasets.

Tables

Table 1: Raw sequence variant calls comparison of 691 whole-genome sequenced Icelanders of human chromosome 21.

Genotyping pipeline	Graphtyper	GATK UG	GATK UGLite	GATK HC	GATK HCjoint	Samtools	Platypus	FreeBayes
Sequence variant records	453,288	451,131	451,415	311,731	418,949	411,907	424,000	596,499
SNPs	406,087	397,821	397,890	267,949	352,293	336,544	301,066	562,319
Transitions/Transversions	1.49	1.46	1.46	**1.75**	1.56	1.50	1.38	0.70
Indels	47,866	53,310	53,525	46,779	73,934	75,363	110,347	33,086
MNPs	1,002	0	0	0	0	0	26,086	21,044
Complex	3,682	0	0	0	34,592	0	0	4,532
Common (dbSNP b149)	157,288	**158,700**	158,590	153,543	158,411	157,998	156,280	136,882
SNPs	145,143	145,723	**145,724**	140,533	144,858	145,135	142,417	126,653
Indels	12,145	12,977	12,866	13,010	13,553	12,863	**13,863**	10,229
Alternative alleles called in trios	454,157	447,144	450,241	312,275	435,511	392,960	408,648	448,429
$Germline_{estimated}$	**267,057**	264,447	264,753	237,978	254,427	255,630	228,646	200,776
$FDR_{estimated}$	41.20%	40.86%	41.20%	23.79%	41.58%	34.95%	44.05%	55.23%
SNPs	371,214	366,068	366,019	243,815	307,024	295,707	255,775	364,942
$Germline_{estimated}$	**232,256**	227,858	227,872	206,084	216,448	215,042	183,375	172,226
Non-SNPs	82,943	81,076	84,222	68,460	128,487	97,253	152,873	83,487
$Germline_{estimated}$	34,801	36,589	36,881	31,894	37,979	40,588	**45,271**	28,550
Common dbSNP calls								
Mean transmission rate	49.98%	50.08%	50.08%	50.01%	50.01%	50.11%	49.47%	50.17%
Mean missing call rate in trios	**0.201**%	0.290%	0.289%	0.329%	0.252%	0.375%	0.445%	0.259%
Mendelian accuracy	**99.52%**	99.48%	99.48%	99.37%	99.41%	99.38%	99.11%	99.44%
Recalled microarray calls	3,188,286	3,197,365	3,197,365	3,189,368	3,193,870	**3,200,590**	3,056,002	2,640,485
Concordance	99.79%	99.80%	99.80%	99.78%	99.76%	99.78%	99.20%	**99.90**%
Only ref/ref array calls	99.92%	99.92%	99.92%	99.93%	99.93%	99.93%	99.90%	**99.96**%
Only ref/alt array calls	99.65%	99.63%	99.63%	99.54%	99.52%	99.58%	99.01%	**99.83**%
Only alt/alt array calls	99.71%	99.81%	99.81%	99.80%	99.74%	99.76%	97.94%	**99.85**%
CPU time [hr]	582	**576**	1,640	12,964	1,216 (87*)	594	3,173	1,030
Time per sample [hr]	0.842	**0.834**	2.373	18.761	1.76 (0.13*)	0.860	4.592	1.491
Mean memory [GB]	10.68	50.17	40.55	65.22	51.98	**1.97**	6.31	6.77
Maximum memory [GB]	45.40	52.72	45.86	307.47	53.58	**2.69**	50.15	196.03

*CPU time of the joint calling step.

Table 2: Comparison of Graphtyper and GATK UG genotyping chromosome 21 of 15,220 sequenced Icelanders.

	Raw		Filtered	
Genotyping pipeline	Graphtyper	GATK UG	Graphtyper	GATK UG
Sequence variant records	1,101,540	1,160,333	473,813	493,620
SNPs	1,024,677	1,035,206	437,844	423,407
Transitions/Transversions	**1.14**	1.06	2.24	2.27
Indels	81,848	125,127	36,086	70,213
MNPs	3,487	0	133	0
Complex	10,707	0	888	0
Alternative alleles called in trios	979,451	1,032,839	338,266	394,679
$\text{Germline}_{\text{estimated}}$	383,998	**397,283**	**308,204**	305,404
$\text{FDR}_{\text{estimated}}$	**60.79**%	61.53%	**8.89**%	22.62%
SNPs	821,098	850,761	304,881	294,004
Transitions/Transversions	**1.01**	0.92	2.18	2.19
$\text{Germline}_{\text{estimated}}$	340,313	**349,878**	**281,972**	264,441
$\text{FDR}_{\text{estimated}}$	**58.55**%	58.87%	**7.51**%	10.06%
Non-SNPs	158,353	182,078	33,385	100,675
$\text{Germline}_{\text{estimated}}$	43,685	**47,405**	26,232	**40,963**
$\text{FDR}_{\text{estimated}}$	**72.41**%	73.96%	**21.43**%	59.31%
CPU time [hr]	**13,192**	33,573	–	–
Time per sample [hr]	**0.867**	2.206	–	–

Table 3: Comparison of whole-genome sequence variant calls of NA12878. Graphtyper was run with and without the knowledge of common dbSNP variation.

	Raw								Filtered			
Genotyping pipeline	Graphtyper (GT)	GT w/dbSNP	GATK UG	GATK UGLite	GATK HC	Samtools	Platypus	FreeBayes	Graphtyper	GT w/dbSNP	GATK UG	GATK HC
SNPs	4,210,841	4,230,056	3,913,454	3,912,894	3,774,031	3,729,409	3,511,646	3,760,288	3,821,418	3,817,459	3,585,462	3,569,701
Transitions/Transversions	1.91	1.90	1.97	1.97	1.99	2.02	2.02	1.98	1.99	1.99	2.04	2.04
Indels	726,382	761,794	649,301	649,477	781,960	735,279	823,257	617,530	703,251	730,566	646,057	771,134
MNPs	1,146	1,199	0	0	0	0	176,269	96,809	940	974	0	0
Complex	7,538	7,626	0	0	0	0	0	35,463	6,625	6,693	0	0
Recalled platinum variants	4,090,418	**4,103,693**	3,967,739	3,967,654	3,997,455	3,874,091	3,760,978	3,813,506	4,020,670	**4,030,504**	3,862,484	3,918,216
Recall rate	98.14%	**98.46%**	95.20%	95.20%	95.91%	92.95%	90.24%	91.50%	96.47%	**96.70%**	92.67%	94.01%
Validated variant calls	4,081,193	**4,094,264**	3,963,186	3,963,134	3,994,476	3,861,985	3,757,577	3,798,996	4,011,769	**4,021,641**	3,857,999	3,915,296
Precision	99.774%	99.770%	99.885%	99.886%	**99.925%**	99.688%	99.910%	99.620%	99.779%	99.780%	99.884%	**99.925%**
Validated SNP calls	3,567,543	**3,568,374**	3,465,168	3,465,145	3,457,324	3,422,248	3,221,031	3,327,170	**3,502,636**	3,501,379	3,360,971	3,380,200
Recall rate	99.24%	**99.27%**	96.39%	96.39%	96.17%	95.20%	89.60%	92.55%	**97.43%**	97.40%	93.49%	94.02%
Precision	99.990%	99.986%	99.991%	99.991%	**99.998%**	99.993%	99.996%	**99.998%**	99.992%	99.990%	99.993%	**99.998%**
Validated non-SNP calls	513,650	525,890	498,018	497,989	**537,152**	439,737	536,546	471,826	509,133	520,262	497,028	**535,096**
Recall rate	91.23%	93.38%	87.70%	87.69%	**94.29%**	78.85%	94.25%	84.90%	90.40%	92.33%	87.52%	**93.93%**
Precision	98.304%	98.330%	99.153%	99.159%	**99.464%**	97.371%	99.393%	97.032%	98.333%	98.389%	99.154%	**99.469%**
Peak memory usage [GB]	7.68	9.15	43.97	40.48	44.00	**1.35**	3.93	2.23	–	–	–	–
CPU time [hr]	154.1	166.5	31.1	41.7	71.0	35.2	**9.4**	22.3	–	–	–	–
Alt. alleles called in trio	6,253,839	6,374,281	5,754,093	5,757,400	5,736,575	5,439,047	5,826,828	5,596,394	5,529,778	5,589,820	5,272,137	5,434,920
$FDR_{estimated}$	6.06%	6.01%	3.34%	3.38%	**3.32%**	4.56%	4.90%	4.67%	4.69%	4.57%	**2.62%**	2.86%
$Germline_{estimated}$	5,874,556	**5,991,012**	5,562,132	5,562,776	5,546,352	5,190,838	5,541,586	5,335,096	5,270,514	**5,334,150**	5,133,770	5,279,402
SNP alt. alleles	5,322,813	5,366,101	4,948,488	4,948,129	4,684,879	4,554,216	4,350,270	4,662,174	4,642,251	4,643,158	4,473,460	4,405,919
$FDR_{estimated}$	4.69%	4.68%	2.55%	2.55%	2.16%	**1.61%**	2.61%	3.63%	2.99%	2.94%	1.65%	**1.60%**
Non-SNP alt. alleles	931,026	1,008,180	805,605	809,271	1,051,696	884,831	1,476,558	934,220	887,527	946,662	798,677	1,029,001
$FDR_{estimated}$	13.92%	13.11%	**8.17%**	8.44%	8.48%	19.77%	11.61%	9.87%	13.56%	12.57%	**8.08%**	8.26%

Table 4: Comparison of Graphtyper's HLA typings to PCR verified HLA types.

		4 digit resolution				2 digit resolution			
HLA gene	n	Correct	1 error	2 errors	Accuracy	Correct	1 error	2 errors	Accuracy
HLA-A	54	52	2	0	98.15%	52	2	0	98.15%
HLA-B	332	–	–	–	–	314	15	3	96.84%
HLA-C	315	–	–	–	–	290	19	6	95.08%
HLA-DQA1	42	42	0	0	100%	42	0	0	100%
HLA-DQB1	82	80	2	0	98.78%	81	1	0	99.39%
HLA-DRB1	190	163	22	5	91.58%	189	1	0	99.74%