nature genetics

Graphtyper enables population-scale genotyping using pangenome graphs

Hannes P Eggertsson[1,2], Hakon Jonsson[1], Snaedis Kristmundsdottir[1,3], Eirikur Hjartarson[1], Birte Kehr[1,4], Gisli Masson[1], Florian Zink[1], Kristjan E Hjorleifsson[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Ingileif Jonsdottir[1,5], Daniel F Gudbjartsson[1,2], Pall Melsted[1,2], Kari Stefansson[1,5] & Bjarni V Halldorsson[1,3]

A fundamental requirement 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 use 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 the 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 both small and population-scale sequencing studies.

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 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 cosegregate with disease and are therefore of great interest in population genetics. The most prominent example, the HLA region, is known to associate with a number of 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 analyses[9].

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[10] and searching for variants as alternative sequences in read alignments (**Fig. 1a**). 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 performed without awareness of variation, causing mapping bias toward the reference allele and misalignments around indels[11,12].

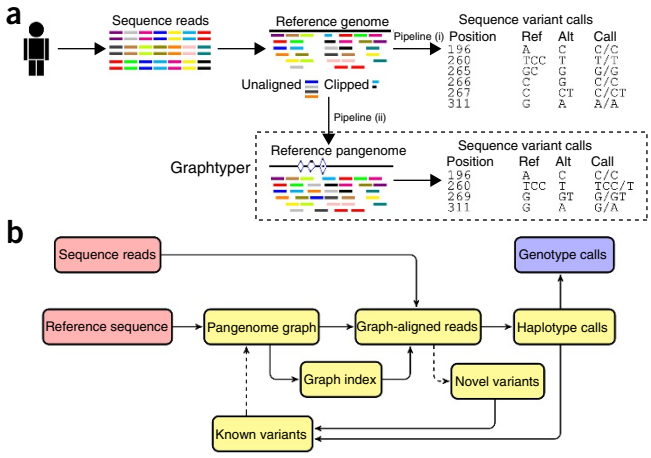

Figure 1 Genotyping pipeline designs. (**a**) Overview of two genotyping pipeline designs. Pipeline (i): a commonly used genotyping pipeline, where sequence reads are aligned to a reference genome sequence and sequence variants are called from discordances between the reads and the reference. Pipeline (ii): Graphtyper's genotyping pipeline. Sequence reads are realigned to a variants-aware pangenome graph and variants are called on the basis of 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.

[1]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland. [2]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [3]School of Science and Engineering, Reykjavik University, Reykjavik, Iceland. [4]Berlin Institute of Health (BIH), Berlin, Germany. [5]Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland. Correspondence should be address to H.P.E. (hannese@decode.is) or B.V.H. (bjarnih@decode.is).

Received 20 June; accepted 1 September; published online 25 September 2017; doi:10.1038/ng.3964

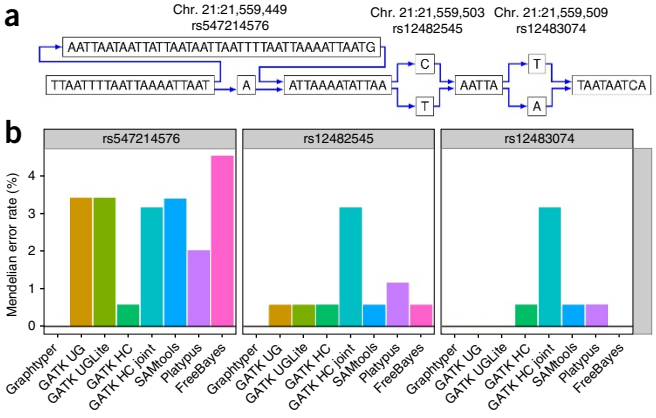

Figure 2 Importance of variation-aware alignment. (**a**) The genomic region chr. 21: 21,559,430–21,559,518 (GRCh38) and three previously reported sequence variants represented with a pangenome graph. (**b**) Mendelian error rates of the three previously reported sequence variants called by eight genotypers.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

Richer data structures that use the large amount of available sequence variation data promise to alleviate some of the limitations of previous methods[13–16]. Although approaches that find polymorphisms in reference-free assemblies have been developed to avoid these limitations[17,18], *de novo* assembly algorithms remain computationally expensive, have less sensitivity[18], and use data structures that have a complex coordinate system.

Pangenomes[13,19,20] 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[21]. Pangenomes incorporate prior information about variation, allowing read aligners to distinguish better 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 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[13].

Graph-like data structures with directed edges have commonly been used to represent pangenomes[20,22–25]. 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[20,22,26–28]; Paten *et al.*[25] 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[13].

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 (**Fig. 1a**). Concomitantly, it aligns sequence reads and genotypes sequence variants present in its graph. Furthermore, Graphtyper discovers novel SNPs and short sequence insertion or deletion variants (indels), which can be used to update the pangenome graph (**Fig. 1b** and Online 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

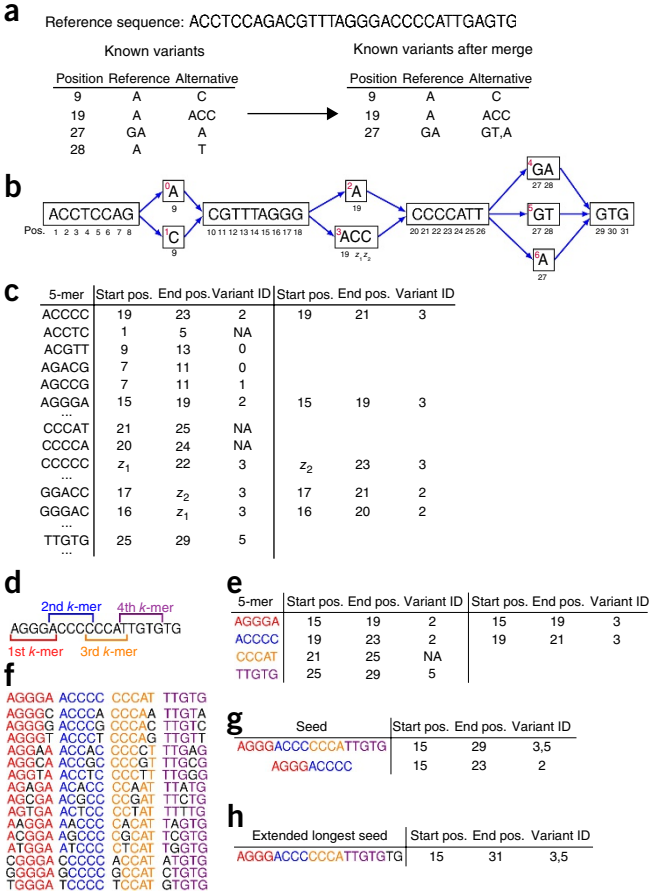

Figure 3 Graphtyper's sequence alignment algorithm. (**a**) An example reference sequence and its known variation. All overlapping variants are merged. (**b**) Constructed pangenome reference graph. We draw the path of the reference sequence as the topmost path. (**c**) The index data structure with $k = 5$. 5-mers in the graph are mapped to a list of its start position, end position, and a variant ID that 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 lookup 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 lookup. (**h**) Final graph alignment after extending the longest seed.

SNPs. Using variation-aware realignment, Graphtyper is capable of better characterization of the region's variation than previous methods, with no Mendelian errors (**Fig. 2b**) and no falsely reported additional sequence variants around the indel (**Supplementary Table 1**) due to misaligned sequence reads (**Supplementary Fig. 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**); that is, a pangenome graph is constructed, indexed, and aligned with sequence reads, from which novel variants are discovered and previously discovered variants are genotyped (**Fig. 1b**).

Table 1 Raw sequence variant call comparison of 691 whole-genome-sequenced Icelanders on chromosome 21

	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	1.5	1.38	0.7
Indels	47,866	53,310	53,525	46,779	73,934	75,363	110,347	33,086
MNPs[a]	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 build 149)	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 alt. transmission rate (%)	49.98	50.08	50.08	50.01	50.01	50.11	49.47	50.17
Mean missing call rate in trios (%)	**0.20**	0.29	0.29	0.33	0.25	0.38	0.45	0.26
Mendelian accuracy (%)	**99.52**	99.48	99.48	99.37	99.41	99.38	99.11	99.44
Microarray SNP comparison								
Correctly inferred genotypes	3,267,641	3,273,959	3,273,959	3,270,243	3,270,590	**3,274,628**	3,177,098	2,967,527
Correctly inferred alleles	6,547,170	6,555,670	6,555,670	6,550,301	6,550,875	**6,557,029**	6,426,358	6,127,836
Site recall rate (%)	97.06	97.33	97.33	97.09	97.22	**97.43**	93.02	80.38
Precision at recalled sites (%)	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 (h)	582	**576**	1,640	12,964	1,216 (87)[b]	594	3,173	1,030
Time per sample (h)	0.842	**0.834**	2.373	18.761	1.76 (0.13)[b]	0.86	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)	45.4	52.72	45.86	307.47	53.58	**2.69**	50.15	196.03

Where meaningful, the best performing algorithm on any metric is indicated in bold. [a]Multinucleotide polymorphisms. [b]CPU time of the joint calling step.

The underlying pangenome data structure is a directed acyclic graph (DAG) where edges connect nodes that contain a DNA sequence (**Supplementary Note**). 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 (**Fig. 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 (**Fig. 3b**). Allele node positions longer than the reference allele are assigned new unique positions (z_1 and z_2 in **Fig. 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) (**Fig. 3c** and Online Methods). Read alignment then follows the seed-and-extend paradigm (**Fig. 3d–h**, Online Methods, and **Supplementary Note**).

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 (Online 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)[29], GATK-Lite UnifiedGenotyper (UGLite), GATK HaplotypeCaller (HC), GATK HC GVCF joint genotyping (HC joint), SAMtools[30], Platypus[18], and FreeBayes[31] (**Supplementary Note**). To ensure a fair comparison between genotyping pipelines, no known sequence variants were given to Graphtyper as input and 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 h (**Table 1**). The other five genotypers required considerably greater compute times (1,030–12,964 h).

We assessed the raw output of all eight genotyping pipelines to compare them independently of filtering technique and to include analysis of all germline variation, somatic variation, and erroneously reported variation due to sequencing or alignment errors. In comparison to other genotypers, Graphtyper called a large number of SNPs (406,087)

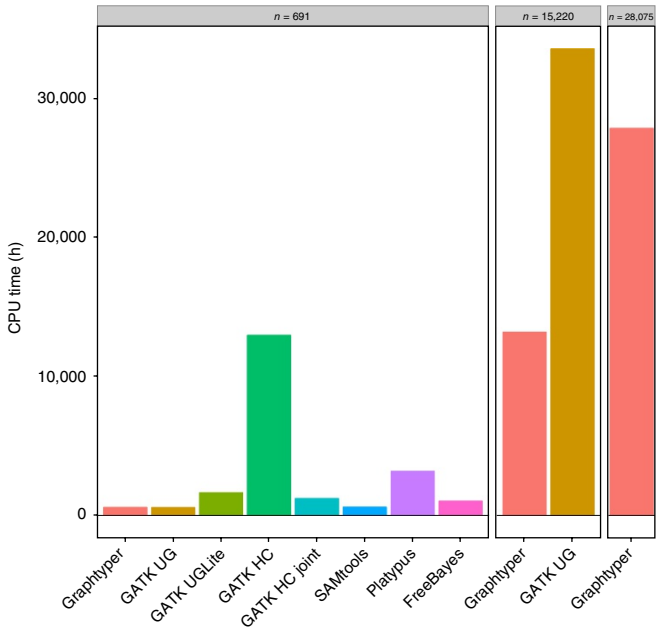

Figure 4 Genotyping time summary. Compute times required to genotype chromosome 21 on three whole-genome sequence data sets. All genotyping pipelines were run once.

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 Fig. 2**). We also observed higher Ti/Tv ratios among alleles with higher transmission rates (**Supplementary Fig. 3**). Motivated by these realizations, we estimated the number of germline alternative alleles on the basis of the transmission rate of the alternative alleles in the 230 trios (Online 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 (MAF) > 1% in any population) in dbSNP build 149. In the 230 trios, Graphtyper called these sequence variants with a mean alternative allele 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 3,284,976 in-house microarray genotypes (Online Methods). For each call set, we measured array site recall rate and precision at recalled sites, and counted how many genotypes and alleles were concordantly inferred over all array sites. If an array site was not recalled, we interpreted it as a homozygous-reference call ("0/0").

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 high-confidence 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**). Graphtyper achieved a substantially lower estimate of false discovery rate (FDR) (2.19%) than the other call sets (10.26–31.22%) but also had a lower estimated number of germline alternative alleles (200,984) than the other call sets (214,801–240,020) (**Supplementary Table 2**).

We measured scalability by genotyping chromosome 21 on a data set of 15,220 Icelanders[32,33], in which there are 1,729 trios (3,863 unique individuals). Our results show that Graphtyper scales much better than GATK UG (**Fig. 4**), with GATK UG using approximately 2.5× 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 h/sample to 0.867 h/sample), while GATK UG used 2.65× more compute time per sample (changed from 0.834 h/sample to 2.206 h/sample).

On the basis of the transmission of alternative alleles in the 1,729 trios, we observed that the FDR increased for Graphtyper and GATK UG in comparison to the 230-trio data set in both raw and filtered call sets. We estimated that Graphtyper detected more germline alternative alleles (308,204) with a considerably lower FDR (8.89%) than GATK UG (305,404 and 22.62%, respectively) in the filtered call sets (**Table 2**).

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

	Raw		Filtered	
	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
Germline (estimated)	383,998	**397,283**	**308,204**	305,404
FDR (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
Germline (estimated)	340,313	**349,878**	**281,972**	264,441
FDR (estimated) (%)	**58.55**	58.87	**7.51**	10.06
Non-SNPs	158,353	182,078	33,385	100,675
Germline (estimated)	43,685	**47,405**	26,232	**40,963**
FDR (estimated) (%)	**72.41**	73.96	**21.43**	59.31
CPU time (h)	**13,192**	33,573	–	–
Time per sample (h)	**0.867**	2.206	–	–

Where meaningful, the best performing algorithm on any metric is indicated in bold.

Table 3 Comparison of whole-genome sequence variant calls of NA12878

	No sequence variants given								Common dbSNP variants given	
	Raw					Filtered			Raw	Filtered
	Graphtyper	GATK UG	GATK UGLite	GATK HC	SAMtools	Platypus	FreeBayes	Graphtyper	GATK UG	GATK HC	Graphtyper	Graphtyper
SNPs	4,210,841	3,913,454	3,912,894	3,774,031	3,729,409	3,511,646	3,760,288	3,821,418	3,585,462	3,569,701	4,230,056	3,817,459
Transitions/transversions	1.91	1.97	1.97	1.99	2.02	2.02	1.98	1.99	2.04	2.04	1.9	1.99
Indels	726,382	649,301	649,477	781,960	735,279	823,257	617,530	703,251	646,057	771,134	761,794	730,566
MNPs	1,146	0	0	0	0	176,269	96,809	940	0	0	1,199	974
Complex	7,538	0	0	0	0	0	35,463	6,625	0	0	7,626	6,693
Recalled Platinum variants	**4,090,418**	3,967,739	3,967,654	3,997,455	3,874,091	3,760,978	3,813,506	4,020,670	3,862,484	3,918,216	4,103,693	4,030,504
Recall rate (%)	**98.14**	95.20	95.20	95.91	92.95	90.24	91.50	96.47	92.67	94.01	98.46	96.70
Validated variant calls	**4,081,193**	3,963,186	3,963,134	3,994,476	3,861,985	3,757,577	3,798,996	4,011,769	3,857,999	3,915,296	4,094,264	4,021,641
Precision (%)	99.774	99.885	99.886	**99.925**	99.688	99.910	99.620	99.779	99.884	**99.925**	99.770	99.780
Validated SNP calls	**3,567,543**	3,465,168	3,465,145	3,457,324	3,422,248	3,221,031	3,327,170	**3,502,636**	3,360,971	3,380,200	3,568,374	3,501,379
Recall rate (%)	**99.24**	96.39	96.39	96.17	95.20	89.60	92.55	**97.43**	93.49	94.02	99.27	97.40
Precision (%)	99.990	99.991	99.991	**99.998**	99.993	99.996	**99.998**	99.992	99.993	**99.998**	99.986	99.990
Validated non-SNP calls	513,650	498,018	497,989	**537,152**	439,737	536,546	471,826	509,133	497,028	**535,096**	525,890	520,262
Recall rate (%)	91.23	87.70	87.69	**94.29**	78.85	94.25	84.90	90.40	87.52	**93.93**	93.38	92.33
Precision (%)	98.304	99.153	99.159	**99.464**	97.371	99.393	97.032	98.333	99.154	**99.469**	98.330	98.389
Peak memory usage (GB)	7.68	43.97	40.48	44	**1.35**	3.93	2.23	–	–	–	9.15	–
CPU time (h)	154.1	31.1	41.7	71	35.2	**9.4**	22.3	–	–	–	166.5	–
Alt. alleles called in trio	6,253,839	5,754,093	5,757,400	5,736,575	5,439,047	5,826,828	5,596,394	5,529,778	5,272,137	5,434,920	6,374,281	5,589,820
FDR (estimated) (%)	6.06	3.34	3.38	**3.32**	4.56	4.90	4.67	4.69	**2.62**	2.86	6.01	4.57
Germline (estimated)	**5,874,556**	5,562,132	5,562,776	5,546,352	5,190,838	5,541,586	5,335,096	5,270,514	5,133,770	**5,279,402**	5,991,012	5,334,150
SNP alt. alleles	5,322,813	4,948,488	4,948,129	4,684,879	4,554,216	4,350,270	4,662,174	4,642,251	4,473,460	4,405,919	5,366,101	4,643,158
FDR (estimated) (%)	4.69	2.55	2.55	2.16	**1.61**	2.61	3.63	2.99	1.65	**1.60**	4.68	2.94
Germline (estimated)	**5,073,098**	4,822,380	4,821,792	4,583,794	4,480,936	4,236,524	4,493,068	**4,503,294**	4,399,652	4,335,432	5,115,034	4,506,482
Non-SNP alt. alleles	931,026	805,605	809,271	1,051,696	884,831	1,476,558	934,220	887,527	798,677	1,029,001	1,008,180	946,662
FDR (estimated) (%)	13.92	**8.17**	8.44	8.48	19.77	11.61	9.87	13.56	**8.08**	8.26	13.11	12.57
Germline (estimated)	801,458	739,752	740,984	962,558	709,902	**1,305,062**	842,028	767,220	734,118	**943,970**	875,978	827,668

Graphtyper was run with and without being given the knowledge of common dbSNP variation. Where meaningful, the best performing algorithm on any metric is indicated in bold.

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 (50× 101-bp paired-end Illumina HiSeq 2000 reads) of these samples are publicly available through the Platinum Genome project[34]. 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 (MAF > 1% in any population) sequence variants (dbSNP build 150). Our experiments showed that GATK does not benefit from incorporating known dbSNP variants as part of its input (**Supplementary Note**).

We assessed sequence variant call sets of the offspring (NA12878) by comparing it to the set of publicly available high-confidence variant calls[34] to measure variant recall rate and precision. On the basis of genotyping of the parents (NA12891 and NA12892), we estimated FDR and the number of transmitted germline alternative alleles in the trio (Online Methods).

Our results show that, even without the knowledge of known variation, Graphtyper has a considerably higher 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**). Incorporating 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 dbSNP variants, respectively), substantially more than the other genotypers (5,190,838–5,562,776). However, Graphtyper had the longest compute time (154.1 h) 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**) and compared them with GATK's call sets filtered according to their best-practices guidelines. After filtering, Graphtyper's recall rate was reduced to 96.47% and its estimated FDR was reduced from 6.06% to 4.69% (**Table 3**).

28,075 Icelandic whole-genome samples

We used Graphtyper to genotype the autosomes and X chromosome of 28,075 whole-genome-sequenced Icelandic samples. The samples have a mean sequencing depth of 35.3× (s.d. 7.9×; range 2–200×) stored in a

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	100.00	42	0	0	100.00
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

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. In comparison to Graphtyper's chromosome 21 genotyping of 691 samples, the sample size increased by 40-fold and 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 noncoding regions.

On the basis of 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* (Online Methods). Using these graphs, we were able to HLA type the same data set 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 for the six genes was 2,609 h, or 5.6 min per sample. The compute time of Graphtyper for the HLA region was orders of magnitude less than for the other genotypers[14,35] (**Supplementary Note**). Previously, the deCODE Genetics laboratory performed HLA typing of the six genes with a PCR-based method at two-digit (n = 647) and four-digit (n = 368) resolution. These previous typings are in good concordance (95.1–100%, two-digit alleles; 91.6–100%, four-digit alleles) with Graphtyper's HLA genotype calls (**Table 4**). Upon manual inspection, we concluded that a large fraction of the discrepancy between the two methods is most likely explained by sample mix-up (**Supplementary Note**).

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. We show that combining these two steps is not only practical, but also 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 data set, a rate five 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 with 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, Graphtyper currently cannot call structural variants.

All of the experiments presented here were run on a high-performance computing cluster, but none of the pipelines are limited to such environments. Even with a large computing cluster, the computational requirements of some of the genotypers are so large that it is infeasible to effectively apply them to population-sized data sets. The computational requirements of Graphtyper are considerably lower than with previous methods, requiring full use of a 10,000-core computer cluster for 10 d to genotype the 28,075 whole-genome-sequenced Icelanders, as compared to an estimated minimum of 25 d 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 to assign reads to a region. To completely remove bias toward the reference genome and fully realize the promise of pangenome analysis requires developing robust methods for graph alignment, some of which are on the horizon[25,26,28]; one such notable project is vg (see URLs). Our results further show the importance of replacing the linear reference with richer data structures to improve understanding of how sequence diversity impacts diseases and other phenotypes.

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

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to our colleagues from deCODE Genetics/Amgen for their contributions. We also wish to thank all research participants who provided biological samples to deCODE Genetics.

AUTHOR CONTRIBUTIONS

H.P.E. implemented the Graphtyper software. H.P.E., P.M., and B.V.H. designed the Graphtyper algorithm. H.P.E., D.F.G., P.M., B.V.H., and K.S. designed the

experiments. H.P.E., E.H., G.M., and F.Z. ran all evaluated genotypers. H.P.E., H.J., and K.E.H. analyzed the call sets. Aslaug Jonasdottir, Adalbjorg Jonasdottir, and I.J. were responsible for PCR validation. H.J. and S.K. contributed software for the project. H.P.E. wrote the initial version of the manuscript, and H.J., S.K., B.K., P.M., B.V.H., and K.S. contributed to subsequent versions. All authors reviewed and approved the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Genome of the Netherlands Consortium. 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, 1985).
8. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–D431 (2015).
9. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M.R. & McVean, G. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47**, 682–688 (2015).
10. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
11. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
12. Shao, H. *et al.* A population model for genotyping indels from next-generation sequence data. *Nucleic Acids Res.* **41**, e46 (2013).
13. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* http://dx.doi.org/10.1093/bib/bbw089 (2016).
14. 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).
15. Paten, B., Novak, A. & Haussler, D. Mapping to a reference genome structure. Preprint at https://arxiv.org/abs/1404.5010 (2014).
16. Huang, L., Popic, V. & Batzoglou, S. Short read alignment with populations of genomes. *Bioinformatics* **29**, i361–i370 (2013).
17. 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).
18. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
19. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
20. Sirén, J., Välimäki, N. & Mäkinen, V. Indexing graphs for path queries with applications in genome research. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **11**, 375–388 (2014).
21. Schneeberger, K. *et al.* Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10**, R98 (2009).
22. 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**, e82138 (2013).
23. Novak, A.M. *et al.* Genome Graphs. Preprint at *bioRxiv* http://dx.doi.org/10.1101/101378 (2017).
24. Church, D.M. *et al.* Extending reference assembly models. *Genome Biol.* **16**, 13 (2015).
25. Paten, B., Novak, A.M., Eizenga, J.M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).
26. Sirén, J. in *2017 Proceedings of the Nineteenth Workshop on Algorithm Engineering and Experiments (ALENEX)* (eds. Fekete, S. & Ramachandran, V.) 13–27 (Society for Industrial and Applied Mathematics, 2017).
27. Kehr, B., Trappe, K., Holtgrewe, M. & Reinert, K. Genome alignment with graph data structures: a comparison. *BMC Bioinformatics* **15**, 99 (2014).
28. Maciuca, S., Elias, C.D.O., McVean, G. & Iqbal, Z. in *Lecture Notes in Computer Science* **9838**, 222–233 (2016).
29. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
30. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
31. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
32. Jónsson, H. *et al.* Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland. *Nature* http://dx.doi.org/10.1038/nature24018 (2017).
33. Jónsson, H. *et al.* Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* (in press).
34. 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).
35. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).

ONLINE METHODS

Icelandic DNA data. The Icelandic samples were whole-genome sequenced at deCODE Genetics[2,32,33] using Illumina HiSeq and HiSeqX sequencing machines[36], and sequences were aligned to the GRCh38 human reference genome using the BWA-MEM algorithm[10]. 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. The personal identities of the participants and biological samples were encrypted by a third-party system approved and monitored by the Data Protection Authority. The National Bioethics Committee and the Data Protection Authority in Iceland approved these studies.

Sequence read alignment. In Graphtyper, sequence variation of small genomic regions (we used 50-kb regions in this study) is represented with a pangenome graph structure. Sequence reads are realigned to the graph of a region if BWA reports 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 (**Fig. 3d**), and determines whether they are present in the graph using an index structure (**Fig. 3e**). Seeds are generated from matches in the index lookup. If the alignments of two adjacent *k*-mers overlap by exactly one base, Graphtyper joins their matches into larger seeds (**Fig. 3g**). The longest seeds are then extended (**Fig. 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 or fewer mismatches, Graphtyper again extracts a set of *k*-mers from the read that overlap by one base in a read, but now also includes *k*-mers with one mismatch (**Fig. 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 the 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 approximately 50 bp), and complex variation (for example, 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 positions 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 semiglobal version of Gotoh's algorithm (**Supplementary Fig. 4a**). Differences in the local alignments are treated as observations of variants (**Supplementary Fig. 4b**).

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

Genotyping. Graphtyper treats 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 for 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 Fig. 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 variant group v, and let R_i be the sample's multiset of sequence reads aligned by Graphtyper to variant group v.

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

$$\mathrm{L}(R_i|H_i) = \prod_{r_{ij} \in R_i} L(r_{ij} \mid H_i) \tag{1}$$

where the relative likelihood of observing a read r_{ij} given the pair of underlying haplotypes is set as

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

where ε_{r_{ij},H_i} is the relative likelihood of observing an error given the underlying haplotypes H_i and the read r_{ij}. These relative likelihoods are chosen from the set $\{2^{-5}, 2^{-6}, \ldots, 2^{-13}\}$ on the basis of how similar the read is to haplotypes H_i, the base-pair quality, the mapping quality of the read, and whether the read is soft-clipped (**Supplementary Note**). Restricting relative likelihoods to this set allows only the integer exponents to be stored, minimizing storage requirements and avoiding floating-point precision problems.

As sequence variants are genotyped in groups, Graphtyper can identify haplotypes in the population within each group (**Supplementary Fig. 5b**) and remove unobserved haplotypes from the graph (**Supplementary Fig. 5c**). In complex regions, this process can greatly reduce the number of haplotype paths in the graph.

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 Fig. 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.

While Mendelian error rate is effective for assessing common alternative alleles, the majority of these alleles are rare and often have no homozygous carriers. When either parent is heterozygous, we cannot deterministically infer the genotype of the offspring (**Supplementary Fig. 6b**). For these trios, we instead calculated the transmission rate of each alternative allele from a parent to his/her offspring. We used the difference between alternative allele transmission rates above and below 50% to estimate the FDR using

$$\mathrm{FDR}_{\mathrm{estimated}} = \max\left(\frac{\#(\mathrm{AA}_{\mathrm{TMR}<50\%}) - \#(\mathrm{AA}_{\mathrm{TMR}>50\%})}{\#(\mathrm{AA})}, 0\right) \tag{3}$$

Here #(AA) is the number of called alternative alleles and $\#(\mathrm{AA}_{\mathrm{TMR}>50\%})$ and $\#(\mathrm{AA}_{\mathrm{TMR}<50\%})$ are the number of alternative alleles with a transmission rate above and below 50%, respectively. The Mendelian laws of inheritance dictate that each allele is equally likely to be transmitted from a parent to his/her offspring. Therefore, in a given variant call set that contains only true germline alternative alleles (FDR = 0%), we would expect $\#(\mathrm{AA}_{\mathrm{TMR}>50\%}) = \#(\mathrm{AA}_{\mathrm{TMR}<50\%})$ and FDR (estimated) = 0%. We also made the assumption that reported non-germline discovered alleles, for example, due to sequencing errors or somatic mutations, are not transmitted. In a call set with no germline alternative alleles (FDR = 100%), we would not expect that alternative alleles are transmitted, $\#(\mathrm{AA}_{\mathrm{TMR}>50\%}) = 0$ and FDR (estimated) = 100%.

On the basis of the above assumptions, we can estimate the number of germline alternative alleles using

$$\#(\mathrm{germline\,AA})_{\mathrm{estimated}} = \#(\mathrm{AA})(1 - \mathrm{FDR}_{\mathrm{estimated}}) \tag{4}$$

HLA typing preprocessing. 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 has been published in our previous work[37] and is provided in the **Supplementary Note**.

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

Data availability. Access to the raw Icelandic sequence data that support the findings of this study is available upon reasonable request from K.S. The data are not publicly available because of Icelandic state law. A **Life Sciences Reporting Summary** is available.

36. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
37. Eggertsson, H.P. *Gyper: A Graph-Based HLA Genotyper Using Aligned DNA Sequences.* MS thesis, Univ. of Iceland, Reykjavík (2015).

doi:10.1038/ng.3964

nature research

Corresponding Author: Hannes Petur Eggertsson

Date: Aug 24, 2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

No sample size calculations were performed. We compared genotyping methods over a wide range of sample sizes (n=1 to n=28,075) to see how they performed on different sample sizes.

2. Data exclusions

Describe any data exclusions.

No data were excluded from the analyses.

3. Replication

Describe whether the experimental findings were reliably reproduced.

NA

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Investigator did not take any part in selecting participants.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Investigator did not take any part in selecting participants.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☒	☐	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☒	☐	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☒	☐	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The source code of our software is available at https://github.com/

Nature Genetics: doi:10.1038/ng.3964

DecodeGenetics/graphtyper. Other software used in the study is described in the Supplementary Note.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	NA
b. Describe the method of cell line authentication used.	NA
c. Report whether the cell lines were tested for mycoplasma contamination.	NA
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	NA

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	There are no covariate-relevant population characteristics of the human research subjects.

Nature Genetics: doi:10.1038/ng.3964