Locityper: targeted genotyping of complex polymorphic genes

Timofey Prodanov[1,2,✉], Elizabeth G. Plender[3,4], Guiscard Seebohm[5], Sven G. Meuth[6], Evan E. Eichler[3,7] & Tobias Marschall[1,2,✉]

[1] Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University, 40225 Düsseldorf, Germany. [2] Center for Digital Medicine, Heinrich Heine University, 40225 Düsseldorf, Germany. [3] Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA. [4] Basic Sciences Division and Computational Biology Program, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA. [5] Institute for Genetics of Heart Diseases, Department of Cardiovascular Medicine, University Hospital Münster, 48149 Münster, Germany. [6] Department of Neurology, Medical Faculty, Heinrich Heine University, 40225 Düsseldorf, Germany. [7] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.

✉Corresponding authors: `timofey.prodanov@hhu.de` and `tobias.marschall@hhu.de`

The human genome contains numerous structurally-variable polymorphic loci, including several hundred disease-associated genes, almost inaccessible for accurate variant calling. Here we present Locityper, a tool capable of genotyping such challenging genes using short and long-read whole genome sequencing. For each target, Locityper recruits and aligns reads to locus haplotypes, for instance extracted from a pangenome, and finds the likeliest haplotype pair by optimizing read alignment, insert size and read depth profiles. Locityper accurately genotypes up to 194 of 256 challenging medically relevant loci (95% haplotypes at QV33), an 8.8-fold gain compared to 22 genes achieved with standard variant calling pipelines. Furthermore, Locityper provides access to hyperpolymorphic HLA genes and other gene families, including KIR, MUC and FCGR. With its low running time of 1h10m per sample at 8 threads, Locityper is scalable to biobank-sized cohorts, enabling association studies for previously intractable disease-relevant genes.

Introduction

Single-nucleotide variants (SNVs) are the most abundant class of genetic variants segregating in the human population and are at the same time easy to access using microarray or short-read sequencing platforms. Unsurprisingly, virtually all genome-wide association (GWAS) studies seeking to map genotypes to phenotypes have therefore been focusing on SNVs. In contrast, structural variants (SVs), which are 50bp in size or longer, are much more challenging to characterize and more than half of all SVs per sample are missed by short-read based variant discovery[1–3], despite their biomedical relevance[4,5]. This difficulty to analyze SVs from short reads is largely driven by their common formation through homology-associated mechanisms, leading to repetitive and complex sequence contexts[2]. Almost 750 genes contain "dark" protein-coding exons, where read mapping and variant calling cannot be adequately performed[6] and around 400 medically relevant genes are almost inaccessible due to their repetitive nature and high polymorphic complexity[7]. Of them, 273 genes are widely used for variant calling and assembly benchmarking[8,9]. Long read technologies are needed to address this problem[10–12] and recent long-read based genome assembly strategies indeed lead to haplotype-resolved genome assemblies of diploid samples that routinely resolve many previously intractable complex genetic loci[13,14]. Nevertheless, long read sequencing of large cohorts remains prohibitively expensive, signifying the need for accurate short read based genotyping.

In the meantime, high quality assemblies are available for hundreds of human haplotypes and give rise to a pangenome reference[2,8,15]. The genetic variation encoded therein can serve as a basis for genotyping workflows by mapping reads to a pangenome graph[16,17] or through k-mer based genome inference[18]. While genome inference with Pangenie[18] has expanded the set of accessible structural variants considerably[8], it exhibits limitations at complex loci with few unique k-mers. As an alternative strategy, methods for targeted genotyping of genes of special interest, such as the HLA and KIR gene families, have been developed[19–21]. Even though these methods can achieve high accuracy, they typically rely on specific gene structure and cannot be easily scaled to include more targets.

Here, we propose a new tool, called Locityper, to leverage genome assemblies in a pangenome reference or custom collections of locus alleles for fast targeted genotyping of complex loci. Locityper can efficiently process both short and long read data, and it integrates a range of different signals based on read depth, alignment identity, and paired-end distance in a statistical model to infer genotype likelihoods. This provides an opportunity to genotype and analyze a diverse set of previously understudied genes for already available large sequencing datasets such as the 1000 Genomes Project cohort and large biobanks like the All-of-Us[22] program and the UK Biobank[23], where disease association studies can be performed.

Results

Overview of the method

Locityper is a targeted genotyping tool designed for structurally-variable polymorphic loci. For every target region, Locityper finds a pair of haplotypes (locus genotype) that explain input whole genome sequencing (WGS) dataset in a most probable way. Locus genotyping depends solely on the reference panel of haplotypes, which can be automatically extracted from a variant call set representing a pangenome (VCF format), or provided as an input set of sequences (FASTA format). Before genotyping, Locityper efficiently preprocesses the WGS dataset and probabilistically describes read depth, insert size, and sequencing error profiles. Next, Locityper uses haplotype minimizers to quickly recruit reads to all target loci simultaneously.

At each locus, Locityper estimates a likelihood for every possible locus genotype by distributing recruited reads across possible alignment locations at the corresponding haplotypes (Figure 1). The likelihood function is defined in such a way to prioritize read assignments with smaller number of sequencing errors; optimal insert sizes across the read pairs; and stable read depth without excessive dips or rises. We show that finding a maximum likelihood read assignment can be formulated as an integer linear programming (ILP) problem (Methods), for which Locityper employs existing ILP solvers and stochastic optimization. Finally, Locityper identifies genotypes with the highest joint likelihood, calculates its quality score, and outputs the most probable read alignments to the two corresponding haplotypes.

Locityper performs well on various sequencing technologies, including short-read Illumina sequencing, high-accuracy PacBio HiFi long reads, and error-prone PacBio CLR and Oxford Nanopore long reads. It is able to efficiently process both mapped and unmapped input reads. Across all steps, Locityper efficiently utilizes multiple computing cores to perform the analysis as fast as possible.

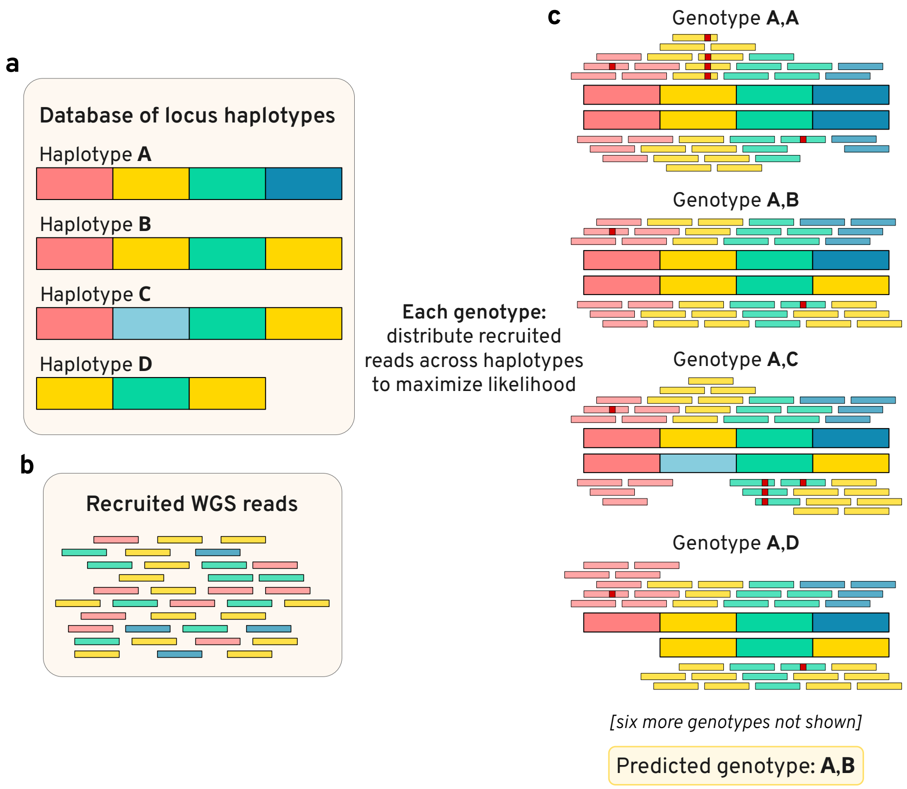
a
Database of locus haplotypes
Haplotype A
Haplotype B
Haplotype C
Haplotype D
b
Recruited WGS reads
c
Genotype A,A
Genotype A,B
Genotype A,C
Genotype A,D
Each genotype: distribute recruited reads across haplotypes to maximize likelihood
[six more genotypes not shown]
Predicted genotype: A,B

Figure 1. Illustration of the locus genotyping approach. a, Database of four locus haplotypes *(A..D)*. **b**, Whole genome sequencing (WGS) reads, recruited to any of the haplotypes. For illustrative purposes, haplotypes and reads are colored by homologous blocks (information, unavailable to Locityper). **c**, Optimal assignments of reads to various genotypes, where small red squares show read alignment mismatches or indels. Genotype *A,B* has the highest joint likelihood due to a small number of alignment errors and no lack or excess of read depth.

Locityper accurately genotypes a wide range of challenging loci

In order to evaluate Locityper's targeted genotyping accuracy, we utilized a list of 273 challenging medically relevant (CMR) genes, collected by Wagner et al.[7] For each target locus we used a reference panel of up to 90 haplotypes extracted from the recent phased whole genome assemblies[8] released by the Human Pangenome Reference Consortium (HPRC)[15]. After removing genes that overlap pangenomic variants longer than 300 kb and merging nearby genes (see Methods), we retained 256 loci covering 13.9 Mb and fully encompassing 265 CMR genes and 23 other protein coding genes (see Supp. Table 1). Then, we used Locityper to genotype 40 Illumina, 40 simulated short-reads, 20 PacBio HiFi and 20 Oxford Nanopore (ONT) HPRC WGS datasets. Each of the datasets was processed twice: first, with the full reference panel of 90 HPRC haplotypes; and then in a leave-one-out (LOO) setting, where the two relevant sample haplotypes were removed from the database beforehand.

To measure genotyping error, we calculated sequence divergence between actual and predicted haplotypes (Figure 2a) and corresponding Phred-like[24] quality values (QV), widely used for genome assembly evaluation[25,26]. Then, we distributed haplotype predictions into four bins based on their QV (<17, 17–23, 23–33 and ≥33), where a haplotype from the last bin (QV ≥ 33) differs from the actual haplotype by no more than 5 bp per 10 kb (see Figure 2b), which is competitive with long read genome assemblies from ONT data[27]. After genotyping, we marked potentially incorrect genotype predictions

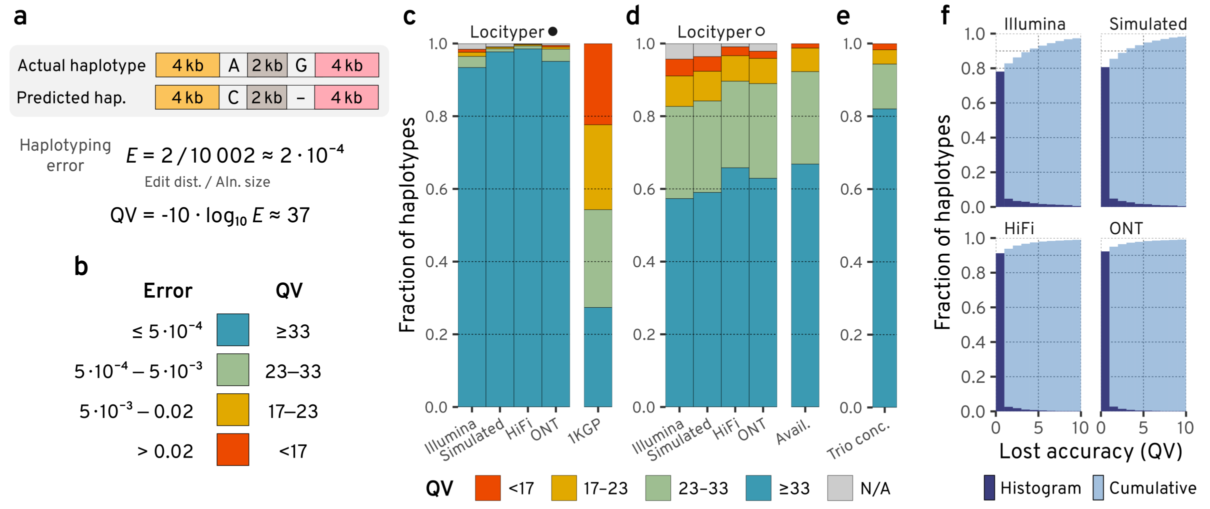

Figure 2. Haplotype accuracy definition and analysis at 256 challenging medically relevant loci. a, Haplotyping error is calculated as sequence divergence between actual and predicted haplotypes. Quality value (QV) is a Phred-like transformation of the haplotyping error. **b**, Approximate correspondence between sequence divergence (haplotyping error) and QV bins. **c**, Haplotyping accuracy for full-database Locityper (marked by a filled circle) and the 1KGP call set for up to 40 HPRC samples, respectively. Haplotypes that failed filtering are shown with gray. **d**, Locityper accuracy in the leave-one-out setting (LOO; shown with a white circle) and the corresponding haplotype availability (QV between actual and closest available haplotypes) across up to 40 HPRC samples. **e**, Locityper concordance at 602 Illumina WGS trios from the 1KGP. **f**, Accuracy, lost by Locityper in the LOO setting across HPRC samples—histogram of differences between best available and predicted QVs. Cumulative fraction is shown with light blue.

based on the number of unexplained reads and the similarities between top genotype predictions; consequently, we will split genotypes into those that *failed* and those that *passed* filtering. Note that the haplotypes were compared across the whole locus, including both coding and non-coding regions, which avoids the need for gene annotations on highly variable haplotypes.

Using the full reference panel, 20,028 Locityper haplotypes (98.4%) passed filtering out of a total of 20,350 fully assembled sample-locus haplotypes across the 256 CMR loci and 40 Illumina WGS samples. Among passed predictions, 94.9% haplotypes had QV ≥ 33, and additional 3.1% haplotypes had QV between 23 and 33 (see Figures 2c and 3). Although some genes remain challenging for accurate genotyping, at 194 (234) loci ≥ 95% haplotypes had QV ≥ 33 (≥ 23).

Even though HPRC assemblies are very accurate, they may include assembly or phasing errors, especially at challenging loci. To remove this factor from the performance analysis, we used ART Illumina[28] to simulate 40 short-read datasets and processed them with Locityper. Unsurprisingly, the tool showed even higher accuracy on simulated datasets, producing 99.1% genotypes, which passed filtering; of them 98.6% with QV ≥ 33 (Figure 2c and Supp. Figure 1a).

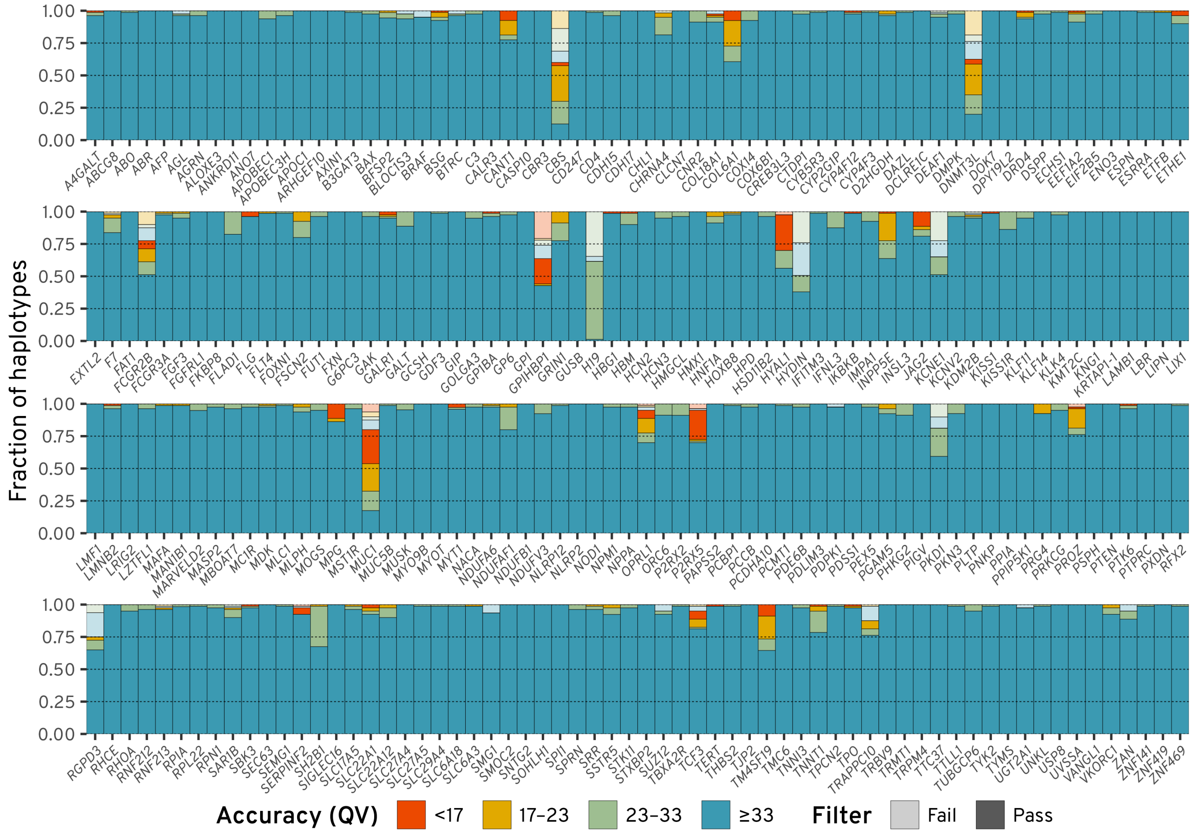

Figure 3. Locityper haplotyping accuracy for 40 Illumina WGS datasets. Predicted haplotypes across 256 challenging medically relevant loci are binned into four groups according to their haplotyping quality value (QV). Semi-transparent colors show predictions that failed post-genotyping filtering.

Locityper significantly outperforms state-of-the-art short-read variant calling pipelines

Recently, New York Genome Center (NYGC) researchers presented an extensive variant calling pipeline and used to it to call phased single nucleotide variants, indels and structural variants across high coverage Illumina WGS datasets for 3,202 samples from the expanded 1KGP (1000 Genomes Project) cohort[3]. As the 1KGP call set is phased, we were able to reconstruct local haplotypes at the 256 CMR loci, as well as calculate genotyping accuracy for 39 HPRC samples present in the 1KGP sample set. Even though the NYGC pipeline utilizes state-of-the-art variant callers, 1KGP haplotypes had significant divergence from the actual sample haplotypes, with only 27.4% haplotypes showing QV ≥ 33 and another 22.3% haplotypes with QV < 17 (see Figure 2c). In total, 1KGP haplotypes were overwhelmingly accurate (≥ 95% haplotypes at QV ≥ 33) at only 22 loci (at 83 loci for QV ≥ 23; see Supp. Figure 1b).

Locityper produces high accuracy genotypes based on long reads

Locityper is not limited to short reads: it is able to process various long read WGS datasets, including accurate PacBio HiFi and error-prone Oxford Nanopore (ONT) data (see Figure 2c). For these two technologies, 99.8% and 99.5% Locityper haplotypes passed filtering, respectively; of them, 98.7% and 95.6% were very accurate with QV ≥ 33, while only 0.1% and 0.5% haplotypes had QV < 17, respectively (see Supp. Figure 1c–d).

Locityper achieves near optimal accuracy in the leave-one-out setting

In the LOO setting, Locityper produced 95.7% Illumina-based haplotypes that passed filtering; among them 86.4% haplotypes had QV ≥ 23, including 59.9% haplotypes with QV over 33 (Figure 2d and Supp. Figure 2a). Even though Locityper achieves smaller accuracy in the LOO setting, it still reconstructs 2.2 times as many highly accurate haplotypes (QV ≥ 33) compared to the 1KGP call set and 4.6 times smaller number of inaccurate haplotypes (QV < 17).

By design, Locityper always associates an input WGS sample with two existing locus haplotypes, and it is not able to predict haplotypes missing from the database. Therefore, Locityper LOO accuracy is limited to haplotype *availability*—similarity between the actual haplotypes and the closest haplotype remaining in the LOO database. 66.8% haplotypes across 40 samples and 256 CMR loci had high availability (closest haplotype with QV ≥ 33); this percentage rises to 92.3% when considering closest haplotypes with QV ≥ 23 (Figure 2d and Supp. Figure 3). In general, Locityper was able to predict haplotypes, close to the best available: *lost accuracy* (difference between best possible and predicted QVs) was under 5 and 10 for 91.4% and 97.3% haplotypes, respectively (see Figure 2f).

At other sequencing datasets, predicted haplotypes were even closer to optimal: lost accuracy was under 5 QV for 93.3%, 97.5% and 97.8% haplotypes based on Simulated, HiFi and ONT reads, respectively; and was under 10 for 98.2%, 99.0% and 99.1% haplotypes (Figure 2f and Supp. Figure 2b–d). In particular, 99.1% HiFi-based LOO haplotypes passed post-genotyping filters and of them 66.4% and 90.5% had QV ≥ 33 and ≥ 23 to the actual haplotypes, respectively.

This analysis shows that Locityper performs extremely well when required haplotypes are present in the reference panel, and achieves near-optimal accuracy with only limited haplotype sets. Growing number of haplotypes in pangenomes[15] are likely to increase Locityper accuracy even further.

Locityper produces concordant trio predictions

In addition to the HPRC dataset, we genotyped the full 1KGP cohort of 3202 Illumina WGS samples, including 602 trios. At each of the 256 CMR loci and for each trio we calculated concordance—similarity between child and parent haplotypes (see Methods). As Figure 2e shows, the vast majority of trio haplotypes were concordant (82.0% and 94.4% with QV ≥ 33 and ≥ 23, respectively). Moreover, average concordance QV surpassed 42 and did not drop below 31 at any locus (see Supp. Figure 4).

Locityper accurately genotypes HLA and KIR genes

In order to evaluate Locityper ability to genotype hyperpolymorphic genes, we examined genes from two medically relevant genomic regions: Major histocompatibility locus (MHC) covering over 4 Mb and over 200 genes[29]; and killer-cell immunoglobulin-like receptor (KIR) gene cluster spanning 150 kb and 17 genes[30]. The two regions contain extremely polymorphic HLA and KIR genes that play an essential role in adaptive and innate immune systems[31,32]. As Locityper genotypes target loci based solely on the sequences of available haplotypes, it is not limited to gene bodies and can utilize intergenic sequence, gene order and presence/absence of copy-number-variable genes. On the other hand, Locityper may require a large collection of assembled MHC or KIR haplotypes to accurately genotype out-of-sample individuals.

Multiple specialized tools have been developed for genotyping the MHC locus[19,20,33]; the newest of them being T1K[21], a state-of-the-art[34] genotyper for HLA and KIR genes that is capable of processing whole genome and whole exome short read sequencing data. To compare T1K and Locityper accuracy, we genotyped 40 Illumina HPRC WGS datasets at 23 genes and 13 pseudo genes from the MHC locus as well as 6 genes and 3 pseudogenes from the KIR locus, all combined into 25 target loci with sum length slightly over 1 Mb. Similarly to CMR loci benchmarking, we run Locityper twice: using the full database and in the LOO configuration.

Across the 40 HPRC samples and 36 genes from the MHC locus, Locityper achieved full match with baseline annotation (correctly predicted all fields in the HLA nomenclature[35,36]) in 99.5% cases with the full database and 87.9% in the LOO setting, compared to T1K's 65.0% (Figure 4a,c). When not requiring full matches and evaluating copy-number accurate protein product prediction (second

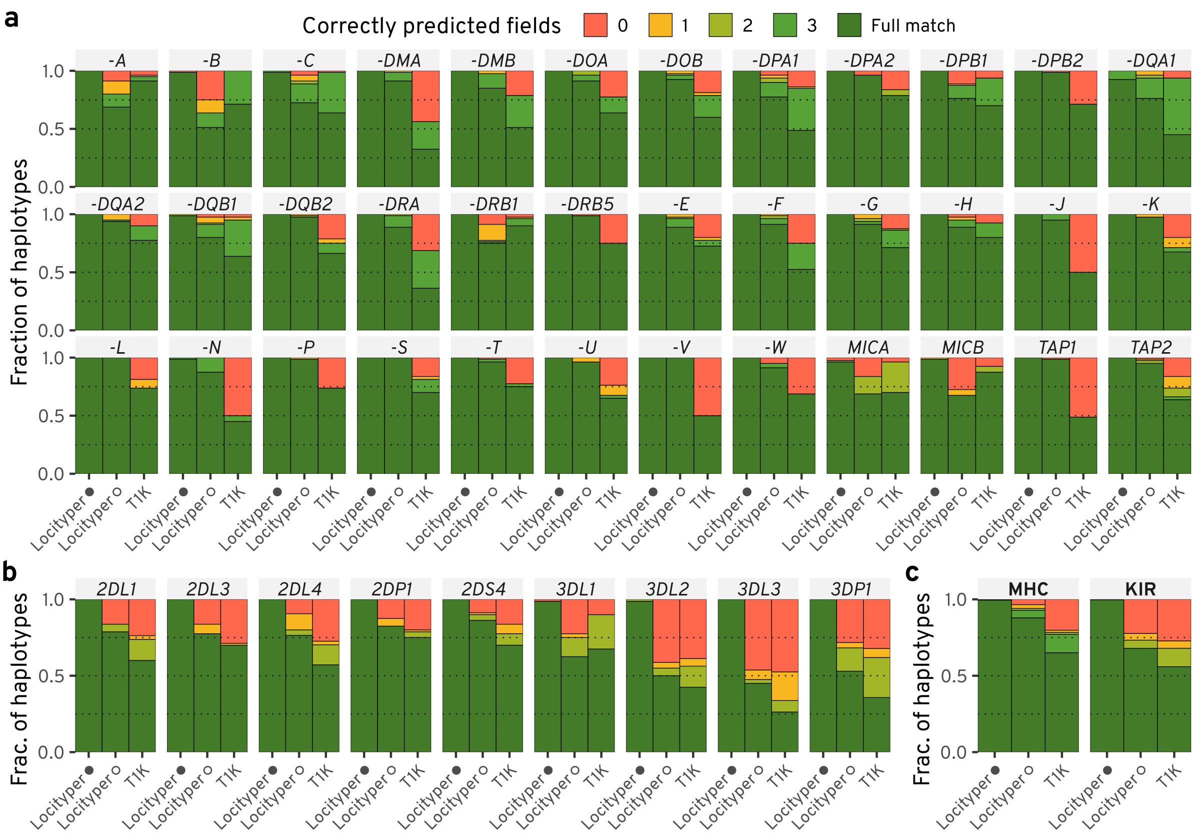

Figure 4. Haplotyping accuracy for 40 HPRC samples at the MHC and KIR loci. Accuracy is shown for Locityper with the full database (denoted by black circle); Locityper in the leave-one-out setting (white circle); and T1K. Fully predicted alleles, as well as correctly identified missing copies, are colored with dark green *(Full match)*, due to the varying number of allele fields in the HLA/KIR gene nomenclature[35,37]. Otherwise, haplotypes are colored according to the number of correctly predicted fields. **a**–**b**, Haplotyping accuracy at 36 (pseudo)genes from the MHC locus (**a**) and 9 (pseudo)genes from the KIR gene cluster (**b**). **c**, Fraction of haplotypes of various accuracy, aggregated across all MHC and KIR genes/pseudogenes.

nomenclature field), T1K accuracy rose to 78.3%, while Locityper accuracy grew to 99.5% and 94.0% using full and LOO databases, respectively. Similarly, at the 9 KIR genes, Locityper correctly predicted protein products in 99.9% (full database) and 73.3% (LOO) and achieved full match in 99.7% and 67.9% cases, respectively. Using the same metrics, T1K accuracy reached 68.0% (protein product) and 55.9% (full match) (see Figure 4b–c).

Some protein products were present in only one HPRC sample, consequently, such samples cannot be correctly annotated by Locityper in the LOO setting. Cases like that appeared in the most polymorphic HLA and KIR genes and explained 41.3% and 19.0% Locityper LOO errors, respectively (see Supp. Figure 5). At the same time, in many cases T1K predicted smaller copy number than required, explaining 85.7% and 42.5% of all errors at the MHC and KIR loci, respectively. When ignoring these two error types (missing allele predictions and unavailable protein groups), Locityper and T1K achieve similar accuracy when predicting protein products—96.4% and 96.2% at the MHC locus, as well as 77.5% and

78.7% at the KIR genes (Supp. Figure 5). Overall, the general purpose tool Locityper performs in a competitive manner even when compared to T1K, which is specifically designed for HLA/KIR genes.

Locityper accurately genotypes disease associated gene families

Although the set of CMR genes includes a wide variety of genetically diverse genes, several important polymorphic gene families are underrepresented in it. One such highly heterogeneous gene family is the mucin gene cluster (*MUC1–MUC24*)[40]. Mucin genes encode large glycoproteins that are essential to barrier maintenance and the defense of epithelial tissues. All canonical mucins harbor a large exon that contains Variable Number Tandem Repeats (VNTRs). These VNTR sequences vary per mucin, yet each extensively encode serine and threonine residues for glycosylation[41]. The gene family can be broken up into two subgroups—that of the cell surface, i.e. tethered mucins and the secreted mucins. In the tethered mucins, single VNTR domains contain variation in total motif copy number and motif usage (Figure 5a); however, the secreted, gel-forming mucins harbor potential variation in VNTR domain copy number, VNTR motif copy number, VNTR motif usage, and cys domain copy number[42,43] (Figure 5b).

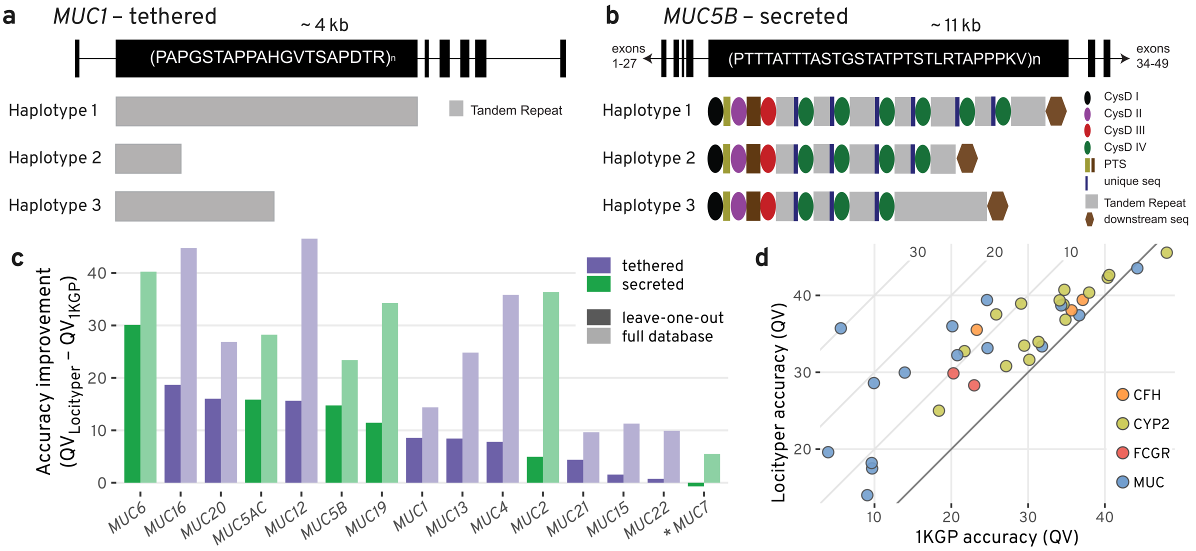

Figure 5. Locityper can accurately genotype mucin and other gene families. a, Gene model of *MUC1*, a mucin tethered to the surface of epithelial cells. *MUC1* harbors a 20 amino acid VNTR repeat sequence and is highly polymorphic in VNTR length, as represented by the example haplotypes 1–3[38]. **b**, Gene model of *MUC5B*, a secreted, gel-forming mucin that is important for homeostasis in the lungs. *MUC5B* encodes an irregular 29 amino-acid VNTR motif that is broken up into separate VNTR domains by cys domains. The number of VNTR domains, cys domains, and VNTR motifs could each contribute to polymorphism among haplotypes at this locus[39]. **c**, Difference in average haplotyping accuracy (QV) between Locityper and 1KGP call set at 15 mucin genes based on 39 Illumina WGS datasets. Improvement for the leave-one-out setting and the full Locityper database are shown with dark and light shades, respectively. Tethered and secreted mucins are shown with purple and green colors; the only non-gel-forming secreted mucin *MUC7* is marked with an asterisk. **d**, Locityper (LOO) and 1KGP call set average genotyping accuracy (QV) across 4 gene families: CFH in orange; CYP2 in light green; FCGR in red; and MUC in blue. Diagonal black line shows zero improvement boundary and diagonal gray lines show 10, 20 and 30 QV improvement.

The presence of these repetitive sequences makes mucins both highly polymorphic and difficult to accurately sequence/genotype in short reads.

Locityper leverages information about both read depth and read alignment for genotyping; therefore, the tool is well suited for the characterization of mucin genetic variation. Based on 39 HPRC Illumina WGS datasets, Locityper (LOO) haplotypes achieved on average 10.5 higher QV compared to the 1KGP call set across 15 examined MUC loci, with the largest improvement observed at *MUC6* and *MUC16* with 30.1 and 18.7 QV, respectively (Figure 5c). The only negative QV difference between Locityper and 1KGP was observed at the non-gel forming *MUC7* gene, where the two haplotype sets showed very high QV values 43.5 and 44.2, respectively.

Further examples of genes that are challenging to address with standard calling techniques are *FCGR2B* and *FCGR3A*, encoding receptors for the Fc region of the immunoglobulin gamma complexes (IgG)[44,45]. IgG binding to FCGR2B induces immune complexes phagocytosis/endocytosis and therewith establishes the basis of antibody production by B-cells. Several transcript variants encoding different isoforms with differing biological function are present (e.g. isoform IIB2 does not trigger phagocytosis unlike other isoforms[46]). Genetic variations in this gene have been reported to cause increased susceptibility to systemic lupus erythematosus (SLE)[47], as well as to malaria, autoimmune hypersensitivity disease, and immune thrombocytopenic purpura[48–50]. The second receptor, FCGR3A, is expressed on natural killer cells (NKc) as an integral membrane glycoprotein[45]. It binds antigen-IgG complexes formed during infection events and thus triggers cytokine production and degranulation by the NK cells. This process is a central effector mechanism limiting viral load and viral propagation in a memory-like manner[51]. Similar to *FCGR2B*, genetic variants in the *FCGR3A* gene have been associated with SLE[47], as well as immunodeficiency, susceptibility to recurrent viral infections, and alloimmune neonatal neutropenia[52–56]. However, genetic analyses of the FCGR genes using high resolution short reads have been proven notoriously difficult due to a recent gene duplication and diversification processes[57]. Nevertheless, at the *FCGR2B* and *FCGR3A* receptor genes Locityper (LOO) improves average QV by 5.3 and 9.6 points compared to the 1KGP call set, respectively (23.0 → 28.3 and 20.3 → 29.9) based on the 39 Illumina WGS datasets (Figure 5d). A larger reference panel would likely improve Locityper's ability to genotype FCGR genes even further, as the tool achieves much higher accuracy (35.9 and 54.0) when using its full database.

Moreover, Locityper (LOO) achieves significant QV improvement (12.2) at the *CFH* gene, associated with age-related vision loss and kidney disorders[58,59]. Finally, Locityper showed on average 4.5 higher QV across 16 protein coding CYP2 genes that play a major role in drug metabolism[60,61]. Out of the CYP2 genes, Locityper achieved the highest improvement at *CYP2U1* (9.8), *CYP2A13* (11.0) and *CYP2W1* (11.7) (see Figure 5d).

Locityper produces more accurate variant calls compared to Pangenie and 1KGP call sets

Pangenie is a pangenome-based short read variant caller, which calls sequence variants and structural variations by counting read k-mers along paths in a pangenomic graph[18]. In order to facilitate comparison to unphased Pangenie call sets, we converted Locityper–predicted locus haplotypes into phased diploid variant calls, and compared Pangenie and Locityper call sets to the ground truth variant calls, extracted from the phased whole genome assemblies for the 40 HPRC samples. Both Locityper and Pangenie were run using their full respective reference panels, which include HPRC samples.

Across the 256 CMR loci, Locityper achieves higher average F_1 score (0.939, median = 0.965) than Pangenie (0.842, median = 0.902), improving both average precision (Pangenie: 0.848, Locityper: 0.945); and recall (Pangenie: 0.843, Locityper: 0.936) by over 9% (see Supp. Figure 6). Although 1KGP call set shows similarly high precision (0.946), it does so at the expense of much lower recall (0.658), achieving a combined F_1 score of 0.746 (median = 0.829). The difference is more pronounced at indels and structural variations, where Locityper, Pangenie and 1KGP call sets attain average F_1 scores of 0.882, 0.728 and 0.550, respectively.

Runtime and memory usage

Locityper WGS preprocessing (executed once per dataset) took on average 16m (minutes) using 8 threads and consumed 15 Gb of RAM for 30× unmapped Illumina WGS input dataset. If a dataset with a similar library preparation was previously preprocessed, read mapping can be skipped, which speeds up WGS preprocessing to under 3m. Next Locityper step, read recruitment, can simultaneously identify reads for multiple target loci. Due to the fact that reading and decompressing input data was the most time-consuming operation, recruitment speed did not depend on the number of loci (1 to 256 tested). Using 8 threads, read recruitment lasted 14m and had negligible memory footprint (<1 Gb). The final Locityper step, genotyping, consumed <10 Gb of RAM and required approximately 9 seconds per locus, depending the target complexity and size: in total, genotyping 256 CMR loci took 38m30s, while genotyping 25 loci covering MHC and KIR loci took 3m55s. Altogether, Locityper analysis of the HLA and KIR genes, including preprocessing and read recruitment, required under 35 minutes using 8 threads.

At the same time, T1K with 8 threads required on average 2h30m and 48m to process the MHC and KIR loci, respectively, and required 2.5 Gb memory. Pangenie calls variants across the whole genome, and, consequently, had heavier runtime and memory footprint: at 24 threads, its pangenome indexing (executed once) and genotyping steps took 34m and 1h40m, respectively, and consumed 60 and 37 Gb of RAM.

In addition to unmapped data, Locityper and T1K can efficiently utilize mapped reads (in BAM/CRAM formats for Locityper and BAM format for T1K) by only recruiting reads aligned to the regions of interest or to alternative contigs, as well as unmapped reads. Additionally, by examining existing alignments, Locityper is able to preprocess WGS datasets almost immediately. Overall, this decreases T1K runtime to 45m and 23m for HLA and KIR loci, respectively, as well as speeds up the full Locityper pipeline for these genes to under 6m.

Discussion

In this study, we present Locityper, a targeted method for genotyping complex polymorphic genes using both short and long-read whole genome sequencing. Locityper implements fast read recruitment to a collection of target loci, and utilizes a carefully balanced probabilistic model to calculate genotype likelihoods based on read alignment, insert size and read depth profiles. Locityper employs integer linear programming and stochastic optimization to find the most likely genotype for each target locus. Locityper departs from the prevalent variant-centric approach, which we argue constitutes a particular limitation for highly polymorphic loci. In contrast, our approach leverages collections of known haplotype sequences, which can be extracted from a pangenome reference or directly provided by the user. By examining larger regions around genes of interest, Locityper inherently makes use of any available information, including intergenic sequence, gene order, structural variants, and copy number of short tandem repeats. Locityper is easy to install via docker, singularity or conda, only requires easy-to-obtain input files, has a small memory footprint and significantly shorter runtime than both T1K and Pangenie.

We demonstrated Locityper's accuracy through excellent agreement to both phased genome assemblies and Mendelian consistency across the 602 family trios included in the 1KGP cohort. When evaluated across a wide range of challenging disease-associated genes, Locityper produces significantly more accurate haplotype predictions compared to a state-of-the art phased variant call set on the 1KGP cohort and outperformed genome-wide pangenome-based genome inference using Pangenie. Locityper's accuracy stays consistently high across various input sequencing technologies, performing well at Illumina, simulated short reads, PacBio HiFi and Oxford Nanopore Technologies (ONT) datasets.

At present, the size of the available collections of reference haplotypes still poses a limitation. To quantify this effect, we performed leave-one-out evaluations, showing that the best available haplotype does not reach QV 33 in more than 30% of cases (Fig. 2d). Therefore, despite Locityper's ability to predict haplotypes close to the best available, the resulting accuracy is not yet ideal for all genes of interest. Significantly larger pangenomes are presently being constructed by the HPRC[15] and we are confident that these future pangenomes will lead to a significant increase in performance on out-of-sample individuals for more complex polymorphic genes. Already now, Locityper outperforms the specialized genotyper T1K across HLA and KIR genes in a LOO setting and shows improved ability to genotype other medically relevant gene families (e.g. MUC and FCGR) using short read WGS.

As part of this study, we used Locityper to process 3,202 Illumina WGS datasets from the 1KGP and make the obtained genotypes available, which provides a resource for deeper analyses of >300 challenging target loci. Additionally, publicly available Locityper-preprocessed WGS summaries will allow for a faster genotyping of genes that were not a focus of this study across the 1KGP cohort. We envision that Locityper will enable the inclusion of complex loci in GWAS[62] and PheWAS[63] analyses, especially in a larger cohort, such as the All-of-Us program[22] and the UK Biobank[23], which promises to discover many new associations and explain missing heritability. Of note, Locityper's ability to process both short and long reads might prove especially useful for the increasing production of long reads in the context of biobank-scale sequencing efforts.

For a given locus, Locityper aims to find two existing haplotypes that would explain an input WGS dataset in the best way. Consequently, it is not designed to reconstruct a novel haplotype, even if it constitutes a mixture of already known haplotypes. To address this, Locityper outputs read alignments to the top predicted genotypes, which can be used later for visual analysis or variant calling. Combined with assembly polishing[64,65], this could improve genotyping accuracy and allow for reconstruction of previously unobserved alleles—a strategy that we plan to explore in future research.

Different parts of a gene of interest, such as exons, introns, and tandem repeats, might have different relative impact on its biological function, as well as on the allele/protein groups in the corresponding allele classification. Currently, Locityper weights all sequence windows according to their k-mer content and sequence complexity. However, a modification can be made that would either automatically, or with user input, upweight/downweight various parts of the haplotype according to their phenotypic importance. Moreover, two loci with significant homology, e.g. part of a non-tandem segmental duplication, are processed independently, with potentially overlapping sets of recruited reads. Locityper mitigates this problem by tracking the number of off-target k-mers per read/haplotype window. Nevertheless, further method improvements are conceivable, such as using a shared pool of reads for the related loci, similar to the strategy implemented by T1K[21].

In conclusion, Locityper allows for fast and accurate targeted genotyping of challenging polymorphic loci using various sequencing technologies. With the current draft pangenome containing highly accurate phased-genome assemblies, Locityper routinely achieves sequence accuracies above QV 33, which is comparable to genome assemblies from Oxford Nanopore data[27]. As more human haplotypes are represented in pangenomes, we expect the accuracy to improve further, which will facilitate detailed analysis of previously intractable genes, leading to improved diagnostic power and novel disease associations.

Data availability

Locityper-predicted genotypes for 3202 Illumina 1KGP samples, corresponding preprocessed WGS parameters, target loci database and benchmarking results can be found on Zenodo[66] (doi.org/10.5281/zenodo.10977559).

Code availability

Locityper is implemented in the Rust programming language, and can be installed via `conda`, `singularity` and `docker`. Source code is freely available under the terms of the MIT license at github.com/tprodanov/locityper along with installation and usage instructions. Fixed source code version as well as additional benchmarking scripts can be downloaded from Zenodo[67] (doi.org/10.5281/zenodo.10979046).

Methods

In this article, we present a targeted tool *Locityper*, designed for genotyping complex multi-allelic loci. Locityper processes whole genome sequencing (WGS) data produced by various sequencing technologies, including highly accurate short and long reads (such as Illumina and PacBio HiFi data, respectively), as well as error-prone long reads (such as PacBio CLR and Oxford Nanopore data). Locityper can efficiently analyze unmapped reads stored in various formats, as well as mapped reads from sorted and indexed BAM/CRAM files.

Broadly, the method can be split into several steps:

1. Preprocessing target loci,
2. Sample preprocessing (performed once for each WGS dataset),
3. Read recruitment (carried out simultaneously for multiple loci),
4. Locus genotyping and generating BAM files with alignments to the best genotypes.

These steps are described in more detail in the following sections.

Preprocessing target loci

Locityper utilizes solely locus haplotype sequences, and does not require any kind of graph structure on top of them. Locus haplotypes can be provided directly in a FASTA file; alternatively, Locityper can automatically extract locus haplotypes from a pangenome, provided in a VCF format (constructed, for example, by Minigraph-Cactus[68]).

When locus haplotypes are extracted from a VCF file, Locityper tries to extend the locus in such a way that both locus ends do not overlap any pangenomic variation. Additionally, the tool tries to select a position that would produce the largest number of unique canonical k-mers at the edges of the locus (default edge size = 500 bp). In the default configuration, locus extension is limited by 50 kb at each side, but can fail if there is a longer structural variant at the locus boundary. In such cases, the user can either increase the allowed extension size, or set the boundaries manually.

Next, Locityper removes all identical haplotypes and calculates Jaccard distance[69] between multisets of minimizers[70] ($(15, 15)$ by default) for all pairs of haplotypes. These pairwise distances are later used to flag potentially incorrect predicted genotypes. Finally, Locityper finds off-target k-mer multiplicities, calculated as the difference between canonical k-mer counts across the full reference genome (calculated using Jellyfish[71]; recommended $k = 25$) and the corresponding k-mer counts at the reference locus sequence.

WGS dataset preprocessing

Locityper aims to probabilistically describe three features of a given WGS dataset— insert size, error profile and read depth— by examining read alignments to a predefined background region. For human WGS data, we use a 4.5 Mb interval on the chr17q25.1 as the default background region as it contains almost no segmental duplications or other types of structural variations. If input reads are unmapped, Locityper subsamples input reads by a factor of s ($^1/_{10}$ by default) and maps them to the reference genome using Strobealign[72] (short reads) or Minimap2[73] (long reads).

Insert size

Manual examination of several paired-end WGS datasets from the HPRC project[15] indicated that the Negative Binomial (NB) distribution fits insert size distribution the best (see Supp. Figure 7). For a given WGS dataset, we use all fully mapped read pairs (clipping $< 2\%$ of the read length, by default) with high mapping quality (≥ 20). We remove outliers by defining maximum allowed insert size as three times the 99th percentile of the observed insert sizes, and discard violating read pairs. Finally, we obtain the NB distribution parameters using the method of moments. During the next two preprocessing steps we will only use read pairs with insert sizes within the 99.9% confidence interval of the corresponding NB distribution.

Error profile

We use two distributions to describe WGS error profiles. First, we use Beta-Binomial (BB) distribution to evaluate edit distance based on the read length. The distribution is fitted using the maximum likelihood estimation (MLE) based on the remaining read pairs. Obtained BB distribution will be used to distinguish between true and off-target alignments at the genotyping stage.

Second, we calculate match, mismatch, insertion and deletion rates (p_M, p_X, p_I, p_D, respectively), and define alignment likelihood as the product of the corresponding rates to the power of the number of operations. For example, alignment with 100 matches, 1 mismatch and 2 insertions would receive likelihood $p_M^{100} \cdot p_X^1 \cdot p_I^2 \cdot p_D^0$. Note that the probabilities do not sum up to one and are incomparable between reads of different lengths. Nevertheless, this formulation produces fast to calculate probabilities, and provides a way to numerically compare different alignments of the same read.

Read depth

We split the background region into windows of fixed size based on the mean read length, and assign reads to windows based on the middle of the corresponding read alignments. Next, we count the number of primary read alignments assigned to each window (only first mates are counted to preserve window independence)[74].

For each window we calculate GC-content and the fraction of unique k-mers in an area centered around the window. Next, we select windows with many unique k-mers ($\geq 90\%$) and estimate read depth mean and variance across various GC-content values using local polynomial regression[75]. NB parameters are then estimated separately for each GC-content based on the smoothed mean, variance, and subsampling rate (see Supplementary Methods 3.1).

Read recruitment

Following dataset preprocessing, Locityper recruits reads to all target loci. For that, we collect minimizers[70] from each locus and each haplotype (default: (5,15) and (10,15)-minimizers for short and long reads, respectively). Uninformative minimizers, which appear ≥ 5 times off target, are ignored. Locityper compares read and target minimizers in parallel and recruits reads to one or several loci according to one of the following rules: short reads are recruited if a sufficient fraction of minimizers matches the target for all read ends (default: 0.7 and 0.6 for single- and paired-reads).

Only a small part of a long read may overlap a given target locus. Consequently, we recruit a long read if it contains a subregion with sufficiently many minimizer matches. For that, we employ the following heuristic: matching/mismatching informative minimizers are assigned s_+/s_- scores (default: +3/-1), and a read is recruited if it has a continuous subsequence with sum score greater or equal to

$$\left\lceil 2L \cdot \frac{M(s_+ - s_-) + s_-}{m_w + 1} \right\rceil, \tag{1}$$

where L is the subregion length (default: 2000 bp), M is the match fraction (default: 0.5) and $2L/(m_w + 1)$ is the expected number of (m_w, m_k)-minimizers per L bp sequence[76]. This heuristic is useful as it can be quickly evaluated using Kadane's algorithm[77], and as it is not too restrictive: shorter read subregions with a higher match rate may produce a hit, and vice versa.

Genotype likelihood

Read location probabilities

Following the read recruitment, every target locus is genotyped independently from other loci. Reads, recruited to the locus, are aligned to all haplotypes H using either Strobealign[72] or Minimap2[73], depending on the read type. Obtained read alignments are assigned BB p-values according to their edit distances and read lengths. A read pair is retained if both read ends have at least one good alignment ($p \geq 0.01$) to at least one of the haplotypes. All alignments with BB $p < 0.001$ are discarded.

Without loss of generality, we will describe the following steps for paired-end reads and use notation $\mathbf{r} = (r_1, r_2)$ to describe a read pair. Each locus haplotype $h \in H$ is split into non-overlapping windows $W^{(h)}$ of fixed size (same size as in read depth preprocessing); furthermore, we expand $W^{(h)}$ by adding a null window $w_\circ$. Each alignment is connected to a single window w based on the middle point of the alignment, with alignment probability $\mathfrak{p}(r_j, w)$ calculated according to the precomputed error profile. Reads without proper alignment to h are connected to the null window $w_\circ$; we will define $\mathfrak{p}(r_j, w_\circ)$ as $\Lambda \cdot \max_h \mathfrak{p}(r_j, h)$ — the probability of the best r_j alignment to any haplotype, multiplied by a penalty Λ (10^{-5} by default).

Paired end alignment probability of the read pair $\mathbf{r} = (r_1, r_2)$ to windows $\mathbf{w} = (w_1, w_2)$ can be written as $\mathfrak{p}(\mathbf{r}, \mathbf{w}) = \mathfrak{p}(r_1, w_1) \cdot \mathfrak{p}(r_2, w_2) \cdot P_{\text{insert}}(\mathbf{r}, \mathbf{w})$, where the last term is calculated according to the precomputed insert size distribution. For null windows, we define insert size probability as the highest probability achievable under the precomputed insert size distribution. Finally, we will denote the full set of possible read pair locations on haplotype h as $L_{\mathbf{r}}^{(h)} \subset W^{(h)} \times W^{(h)}$ and will define the probability of the read pair $\mathbf{r}$ location to be $\mathbf{w}$ as normalized alignment probability:

$$\mathcal{P}_{\mathbf{rw}} = \frac{\mathfrak{p}(\mathbf{r}, \mathbf{w})}{\sum_{h' \in H} \sum_{\mathbf{u} \in L_{\mathbf{r}}^{(h')}} \mathfrak{p}(\mathbf{r}, \mathbf{u})}. \tag{2}$$

Some parts of the target loci can have high homology to other genomic regions. Consequently, we downgrade the effect of potentially misrecruited reads by setting equal probabilities to all locations for read pairs with less than 5 target-specific k-mers.

Read assignment

Without loss of generality, let us consider a diploid genotype $\mathbf{g} = (h_1, h_2)$. We combine windows across the two haplotypes $W^{(\mathbf{g})} = W^{(h_1)} \cup W^{(h_2)}$. If $h_1 = h_2$, we use two copies of each window, such that $|W^{(\mathbf{g})}|$

is always $|W^{(h_1)}| + |W^{(h_2)}|$. Next, for each read pair $\mathbf{r}$ we concatenate possible locations $L_{\mathbf{r}}^{(h_1)}$ and $L_{\mathbf{r}}^{(h_2)}$ in a similar way to achieve a combined list of locations $L_{\mathbf{r}}^{(\mathbf{g})}$.

We describe read assignment to one of the locations using a boolean vector $T_{\mathbf{r}}$ with exactly one true element $\left(\sum_{\mathbf{w}} T_{\mathbf{rw}} = 1\right)$, where $T_{\mathbf{rw}} = 1$ encodes the statement "true location of the read pair $\mathbf{r}$ is $\mathbf{w}$". Probability of the read assignment T for all read pairs R is a product of all selected location probabilities:

$$P(T \mid R) = \prod_{\mathbf{r} \in R} \sum_{\mathbf{w} \in L_{\mathbf{r}}^{(\mathbf{g})}} T_{\mathbf{rw}} \cdot \mathcal{P}_{\mathbf{rw}}. \tag{3}$$

In order to evaluate genotype concordance with WGS data, we search for a read assignment T that produces maximum joint probability $P(\mathbf{g}, T \mid R) = P(\mathbf{g} \mid T, R) \cdot P(T \mid R) = P(\mathbf{g} \mid T) \cdot P(T \mid R)$.

Read depth likelihood

We define genotype likelihood conditional on the read assignment as the probability for all genotype windows to have copy number (CN) equal to one:

$$P(\mathbf{g} \mid T) = \prod_{w \in W^{(\mathbf{g})}} P\Big(\mathrm{CN}(w) = 1 \,\Big|\, d_w(T)\Big). \tag{4}$$

$d_w(T)$ denotes window w depth according to the read assignment T, which can be written as $\sum_{\mathbf{r} \in R} \sum_{u \in W^{(\mathbf{g})}} \left[T_{\mathbf{r},wu} + T_{\mathbf{r},uw}\right]$. At CN $= 1$, read depth follows the NB distribution with precomputed parameters n and ψ. Bayes' theorem with equal priors produces the following result:

$$\varphi_w(T) = P\Big(\mathrm{CN}(w) = 1 \,\Big|\, d_w(T) = d\Big) = \frac{P_{\mathrm{NB}}\big(d;\ n, \psi\big)}{\sum\limits_{c \in \{1\} \cup C_{\mathrm{alt}}} P_{\mathrm{NB}}\big(d;\ c \cdot n, \psi\big)}, \tag{5}$$

where alternative hypotheses are represented by a set C_{alt}. We found it beneficial to use $C_{\mathrm{alt}} = \{0.5, 1.5\}$, in other words, a half divergence from the expected read depth is considered significant. As unmapped reads are already penalized by low alignment probabilities $\mathfrak{p}(r, w_\circ)$, we define $P\big(\mathrm{CN}(w_\circ) = 1 \,\big|\, d\big) = 1$ for any read depth d.

Window and read weights

Low-complexity regions, as well as short and long repeats evoke various difficulties in read sequencing, recruitment and alignment. In order to assign window weights in a continuous fashion, we define the following two-parametric function $\vartheta : [0,1] \mapsto [0,1]$:

$$\vartheta(x;\eta,q) = \begin{cases} 1 & \text{if } x = 1, \\ 1 - \dfrac{1}{\left(\frac{x}{\eta}\cdot\frac{1-\eta}{1-x}\right)^{2q} + 1} & \text{otherwise.} \end{cases} \tag{6}$$

ϑ exhibits several useful properties: it is a strictly increasing smooth function such that $\vartheta(0) = 0$ and $\vartheta(1) = 1$. Location parameter $\eta \in (0,1)$ defines the break point $\vartheta(\eta;\eta,\cdot) = 1/2$, while the power parameter $q > 0$ controls the "slope" of the function, with larger q producing larger derivative $\vartheta'(\eta;\eta,q)$ (see Supp. Figure 8). Finally, we define window w weight $\zeta_w = \vartheta(x_1;\eta_1,q_1)\cdot\vartheta(x_2;\eta_2,q_2)$ based on the fraction of locus-specific k-mers x_1 and linguistic sequence complexity $x_2 = U_1U_2U_3$, where U_i is the fraction of unique i-mers in the window w out of the maximal possible number of distinct i-mers[78], with default parameters $\eta_1 = 0.2$, $\eta_2 = 0.5$ and $q_1 = q_2 = 2$.

Combined likelihood and likelihood update

Not accounting for window weights, combined likelihood for a genotype $\mathbf{g}$ and read assignment T can be calculated as:

$$P(\mathbf{g}, T \mid R) = \prod_{\mathbf{r}\in R} \sum_{\mathbf{w}\in L_{\mathbf{r}}^{(\mathbf{g})}} T_{\mathbf{rw}}\cdot\mathcal{P}_{\mathbf{rw}} \times \prod_{w\in W^{(\mathbf{g})}} \varphi_w(T), \tag{7}$$

Next, we move the calculations to log-space, add window weights ζ_w, and introduce contribution factors $\Omega_R, \Omega_D \geq 0$, which represent the relative importance of read alignment and read depth likelihoods, respectively. Then, log-likelihood $\mathcal{L}$ can be written in the following way:

$$\mathcal{L}(\mathbf{g}, T \mid R) = \Omega_R \sum_{\mathbf{r}\in R} \sum_{\mathbf{w}\in L_{\mathbf{r}}^{(\mathbf{g})}} T_{\mathbf{rw}} \log\mathcal{P}_{\mathbf{rw}} + \Omega_D \sum_{w\in W^{(\mathbf{g})}} \zeta_w \log\varphi_w(T). \tag{8}$$

Contribution factors Ω_R and Ω_D are necessary as the read alignments can overshadow read depth due to the large number of read pairs and large differences between various read alignments. The two values need to be defined in advance and should sum up to 2: we recommend default values $\Omega_R = 0.15, \Omega_D = 1.85$, as they produced good results across a selection of target loci and sequencing datasets.

Likelihood update

Given log-likelihood $\mathcal{L}(\mathbf{g}, T \mid R)$ for genotype $\mathbf{g}$ and some read assignment T, we can efficiently calculate log-likelihood $\mathcal{L}(\mathbf{g}, T' \mid R)$ for a new read assignment T' if the read assignment has changed for only one read pair; in other words, when $\sum_{\mathbf{r}\in R}\sum_{\mathbf{w}\in L_{\mathbf{r}}^{(\mathbf{g})}} |T_{\mathbf{rw}} - T'_{\mathbf{rw}}| = 2$. Suppose that the read assignment changed for read pair $\mathbf{r}$ from location u, v (in T) to u', v' (in T'). Then, read depth likelihood values $\varphi_w(T')$ will be identical to $\varphi_w(T)$ for all windows except for u, v, u', v', where read depth can quickly be recomputed. This way, log-likelihood can be recalculated in constant time:

$$\begin{aligned} \mathcal{L}(\mathbf{g}, T' \mid R) \;=\; & \mathcal{L}(\mathbf{g}, T \mid R) + \;\Omega_R \cdot \big(\log \mathcal{P}_{\mathbf{r},u'v'} - \log \mathcal{P}_{\mathbf{r},uv}\big) \\ & + \Omega_D \sum_{w \in \{u,v,u',v'\}} \zeta_w \cdot \big(\log \varphi_w(T') - \log \varphi_w(T)\big). \end{aligned} \tag{9}$$

Finding best read assignment

For each genotype $\mathbf{g}$ we aim to find such read assignment T that would maximize joint log-likelihood $\mathcal{L}(\mathbf{g}, T \mid R)$. Locityper implements three approaches for finding such read assignment: Stochastic Greedy approach[79], Simulated Annealing[80] and Integer Linear Programming (ILP)[81]. The first two algorithms start from an arbitrarily generated read assignment T, then iteratively select a random read pair $\mathbf{r}$ and switch its location if it increases the genotype likelihood. In addition to "good" location switches, Simulated Annealing permits "bad" switches (decreasing overall likelihood), gradually restricting frequency of such events.

In an ILP formulation we introduce two sets of unknowns: $x_{\mathbf{rw}} \in \{0, 1\}$ for each read pair $\mathbf{r}$ and each location $\mathbf{w} \in L_{\mathbf{r}}^{(\mathbf{g})}$; and $y_{wd} \in \{0, 1\}$ for each window $w \in W^{(\mathbf{g})}$ and each possible window depth d between zero and maximal possible read depth $D_{\max}$. The problem can be written as follows:

$$\begin{aligned} \text{Maximize} \quad & \sum_{\mathbf{r} \in R} \sum_{\mathbf{w} \in L_{\mathbf{r}}^{(\mathbf{g})}} x_{\mathbf{rw}} \cdot \Omega_R \log \mathcal{P}_{\mathbf{rw}} + \sum_{w \in W^{(\mathbf{g})}} \sum_{d=0}^{D_{\max}} y_{wd} \cdot \Omega_D \zeta_w \varphi_w(d) \\ \text{Subject to} \quad & \sum_{\mathbf{w} \in L_{\mathbf{r}}^{(\mathbf{g})}} x_{\mathbf{rw}} = 1 \quad \forall \mathbf{r} \in R, \\ & \sum_{d=0}^{D_{\max}} y_{wd} = 1 \quad \forall w \in W^{(\mathbf{g})}, \\ \sum_{\mathbf{r} \in R} \sum_{u \in W^{(\mathbf{g})}} \big(x_{\mathbf{r},wu} + x_{\mathbf{r},uw}\big) - & \sum_{d=0}^{D_{\max}} d \cdot y_{wd} = 0 \quad \forall w \in W^{(\mathbf{g})}. \end{aligned} \tag{10}$$

Note, that we can remove variables $x_{\mathbf{r}}$ for trivial read pairs, which map to only one possible location; at the same time, the number of possible read depth variables y_w is exactly one more than the number of non-trivial read pairs mapping to w. Finally, the sum $\sum_{\mathbf{r} \in R} \sum_{u \in W^{(\mathbf{g})}}$ in the third constraint can be limited to windows and read pairs, relevant to the window w. Locityper utilizes two commercial ILP solvers, available under academic licenses: HiGHS[82] and Gurobi[83]. Note that it is possible to state a bigger ILP problem; solution to such formulation would immediately produce the best locus genotype (see Supplementary Methods 3.2). However, we observed that the available ILP solvers are unavailable to quickly and accurately find such solution.

Locus genotyping

In order to find the best locus genotype for the input WGS data, Locityper finds the best read assignment and the corresponding genotype likelihood for each possible locus genotype (Figure 1). To speed up the process, we start by calculating log-likelihood in the absence of read depth ($\Omega_D = 0$), which can be efficiently computed by assigning every read to its most probable location. Then, we employ a heuristic filtering by removing all genotypes whose likelihood is 10^{100} smaller than the best likelihood (first 500 genotypes are kept regardless of likelihood). For all remaining genotypes, the best read assignment is found using one of the three approaches, described above. Even though the ILP solvers typically find better read assignments, we use Simulated Annealing as the default solver as it produces decent read assignments in a fraction of ILP solving time.

Splitting locus haplotypes into non-overlapping windows is an intrinsically discrete process. Furthermore, windows can be shifted across different haplotypes due to the presence of indels. Consequently, identical read depth profiles may produce varying read depth likelihoods depending on the window boundaries. To reduce this effect, we perform a procedure, similar to Noise Injection regularization[84], where we randomly move read alignment centers to either direction and reassign reads to windows. In addition, we redefine window GC-content values and weights ζ_w as if the window was randomly moved (actual window boundaries stay fixed). In a default configuration, read and window movement is limited to half window size or 200 bp, whichever is smaller. Repeating noise injection several times (20 by default), together with the stochastic nature of likelihood maximization produces a distribution of log-likelihoods for each genotype.

Finally, Locityper selects a *primary* genotype with the highest average log-likelihood and calculates its Phred quality[24] based on the probability of error: probability that true log-likelihood of any other genotype is higher than true log-likelihood of the primary genotypes, calculated using one-sided Welch's t-test[85].

In addition to quality values, Locityper outputs the number of unexplained reads—reads that map to some, but not to the two predicted haplotypes. Additionally, Locityper iterates over all genotypes and evaluates average Jaccard distance to the primary genotype (see locus preprocessing) weighted by the corresponding genotype probabilities. Such a measure is motivated by the fact that all probable genotypes should be similar to one another. For this study, we marked genotypes as potentially incorrect if weighted distance is over 30 or if there are more than 1000 unexplained reads, which in turn constitute over 20% of all reads for the locus.

Locus selection

Original set of challenging medically relevant (CMR) genes contains 273 protein coding genes[7]. We expanded gene coordinates to a minimum of 10 kb, when needed, and supplied positions as input to Locityper locus preprocessing, allowing an additional coordinate expansion by at most 300 kb to each of the sides (`add -e 300k`). At this stage, eight genes (*ATPAF2*, *CLIP2*, *GTF2I*, *GTF2IRD2*, *IGHV3-21*, *MRC1*, *NCF1* and *SMN1*) were discarded as at least one the gene ends was contained in a 300 kb–long pangenomic bubble. Afterwards, we removed redundant loci (completely contained in another locus), which produced a final set of 256 loci, containing 265 CMR genes.

In a similar fashion, we added 26 loci covering genes from the MHC and KIR gene clusters, as well as 30 loci covering MUC, CFH and CYP2 genes. Full information about the loci can be found in the Supp. Table 1.

Utilized data

Pangenome reference in a variant calling format (VCF) was downloaded from `https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/minigraph-cactus/hprc-v1.1-mc-grch38/hprc-v1.1-mc-grch38.raw.vcf.gz`. Illumina, PacBio HiFi and Oxford Nanopore data for the HPRC samples can be found at `https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working`. NYGC variant calls for the 1KGP samples were downloaded from `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV`. 3202 1KGP Illumina datasets are available on the European Nucleotide Archive under accession codes PRJEB31736 and PRJEB36890.

Simulated Illumina data was constructed using ART Illumina[28] `v2.5.8` with parameters `-ss HS25 -m 500 -s 20 -l 150 -f 15` for all phased haplotype assemblies from the HPRC project, which can be found at doi.org/10.5281/zenodo.5826274.

Benchmarking Locityper

In order to evaluate haplotyping accuracy, we computed full-length alignments between actual and predicted haplotypes using Locityper `align` module. Internally, it finds the longest common subsequence of k-mers using LCSk++[86] and completes the alignment between k-mer matches using Wavefront alignment algorithm[87,88]. Three k-mer sizes are tried (25, 51 and 101), and an alignment with the highest alignment score is returned.

Afterwards, we calculate haplotyping error—sequence divergence between two haplotypes, calculated as the ratio between edit distance Δ and alignment size S (edit distance plus the number of matches). As actual and predicted genotypes consist of two haplotypes, there are two possible actual–predicted haplotype pairings. Out of the two options we select such pairing that produces a smaller ratio between sum edit distance and sum alignment size.

Then, we use Phred-like transformation of haplotyping error $\mathrm{QV} = -10 \cdot \log_{10}(\Delta/S)$ to obtain haplotyping quality values (QV)[25,26]. However, when two haplotypes are completely identical ($\Delta = 0$), QV becomes infinite, which poses problems for average QV calculation. For that reason, we corrected QV definition:

$$\mathrm{QV} = -10 \cdot \log_{10}\left(\frac{\max\{\Delta, 1/2\}}{S}\right). \tag{11}$$

This way, QV difference between edit distances 0 and 1 is the same as between 1 and 2, and equals to $10 \cdot \log_{10} 2 \approx 3$. Constants smaller than $1/2$ were generally even more beneficial for Locityper benchmarking.

In the leave-one-out setting, we calculate lost accuracy as the difference between best possible $\mathrm{QV}_{\mathrm{avail}}$ (QV of the closest remaining haplotype) and $\mathrm{QV}_{\mathrm{pred}}$ of the predicted haplotype. However, as to not penalize well-predicted haplotypes relative to very good possible haplotypes, we modified lost accuracy to be $\min\{\mathrm{QV}_{\mathrm{avail}}, 33\} - \min\{\mathrm{QV}_{\mathrm{pred}}, 33\}$. This way, $\mathrm{QV}_{\mathrm{pred}} = 30$ with $\mathrm{QV}_{\mathrm{avail}} = 50$ will produce lost accuracy = 3 instead of 20.

We considered a trio of locus genotypes concordant if one of the child haplotypes matches well one of the maternal haplotypes, while another child haplotype matches one of the paternal haplotypes. Similarly to haplotyping error calculation, we iterated over 8 possible pairings and selected one with the smallest sum edit distance divided by sum alignment size, as well as calculated a QV score for each of the child haplotypes.

We used Bcftools[89] (`v1.18`) `consensus` command to reconstruct haplotypes from the 1KGP phased variant call set[3]. In the process, we removed contradicting overlapping variant calls, and variants with symbolic alternative alleles (with exception of `<DEL>`), as they cannot be used for haplotype reconstruction.

In order to compare Locityper, 1KGP and Pangenie[18] `v3.02` call sets, we decomposed and normalized variant calls using Vt[90] `v0.57721` commands `decompose_blocksub` and `normalize`, respectively. Before decomposition we removed low quality variants (Locityper genotypes, which failed filtering, and Pangenie variants with quality <10). Then, we used RTG tools[91] `v3.12.1` `vcfeval` module to calculate variant calling precision and recall using existing HPRC Minigraph-cactus[68] representation as a baseline call set.

Finally, we used T1K[21] `v1.0.5` with presets `hla-wgs --alleleDigitUnits 15 --alleleDelimiter :` and `kir-wgs` with all other parameters set to default. Ground-truth HLA and KIR annotation for HPRC assemblies were obtained using Immuannot[92] using allele databases[31,93] IPD-IMGT/HLA `v3.52` and IPD-KIR `v2.12`. If a haplotype contains a novel gene allele, Immuannot may associate it with several existing alleles. In such cases, we evaluated predicted allele by the best-matching existing allele.

In all evaluations, we utilized Locityper `v0.15.1` along with its dependencies Samtools[89] `v1.18`, Jellyfish[71] `v2.2.10`, Minimap2[73] `v2.26-r1175` and Strobealign[72] `v0.13.0`.

Acknowledgements

We thank the MHC working group of the Human Genome Structural Variation Consortium (HGSVC) for valuable feedback on an earlier version of the evaluation. This research was supported in part by funding from the National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI) R01 HG002385 (to E.E.E. and T.M.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author contributions

T.P. and T.M. conceived the project and designed the algorithm. T.P. developed the software and performed the analyses. T.P. and E.G.P. prepared the figures. T.P., E.G.P., G.S., S.G.M., E.E.E. and T.M. wrote the manuscript.

Competing interests

E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc.

References

[1] Chaisson, M. J. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications* **10**, 1784 (2019).

[2] Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).

[3] Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–40 (2022).

[4] Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

[5] Eichler, E. E. Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine* **381**, 64–74 (2019).

[6] Ebbert, M. T. *et al.* Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology* **20**, 1–23 (2019).

[7] Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nature Biotechnology* **40**, 672–80 (2022).

[8] Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

[9] Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).

[10] Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**, 1155–62 (2019).

[11] Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of

long-range sequencing and mapping. *Nature Reviews Genetics* **19**, 329–46 (2018).

[12] Ebler, J., Haukness, M., Pesout, T., Marschall, T. & Paten, B. Haplotype-aware diplotyping from noisy long reads. *Genome Biology* **20**, 116 (2019).

[13] Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–5 (2021).

[14] Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with verkko. *Nature Biotechnology* **41**, 1474–82 (2023).

[15] Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–46 (2022).

[16] Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology* **36**, 875–9 (2018).

[17] Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).

[18] Ebler, J. *et al.* Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics* **54**, 518–25 (2022).

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

[20] Dilthey, A. T. *et al.* HLA*LA — HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394–6 (2019).

[21] Song, L., Bai, G., Liu, X. S., Li, B. & Li, H. Efficient and accurate KIR and HLA genotyping with massively parallel sequencing data. *Genome Research* **33**, 923–31 (2023).

[22] Mahmoud, M. *et al.* Utility of long-read sequencing for All of Us. *Nature Communications* **15**, 837 (2024).

[23] Sudlow, C. *et al.* Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine* **12**, e1001779 (2015).

[24] Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Research* **8**, 186–194 (1998).

[25] Porubsky, D. *et al.* Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology* **39**, 302–8 (2021).

[26] Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**, 722–36 (2017).

[27] Gustafson, J. A. *et al.* Nanopore sequencing of 1000 genomes project samples to build a comprehensive catalog of human genetic variation. *medRxiv* 2024.03.05.24303792 (2024).

[28] Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–4 (2012).

[29] Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annual Review of Genomics and Human Genetics* **14**, 301–23 (2013).

[30] Biassoni, R. Human natural killer receptors, co-receptors, and their ligands. *Current Protocols in Immunology* **84**, 14.10 (2009).

[31] Barker, D. J. *et al.* The IPD-IMGT/HLA database. *Nucleic Acids Research* **51**, D1053–60 (2023).

[32] Vilches, C. & Parham, P. KIR: diverse, rapidly evolving receptors of innate and adaptive immunity. *Annual Review of Immunology* **20**, 217–51 (2002).

[33] Orenbuch, R. *et al.* arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics* **36**, 33–40 (2020).

[34] Yu, D. *et al.* A rigorous benchmarking of alignment-based HLA callers for RNA-seq data. *bioRxiv* 2023.05.22.541750 (2024).

[35] Marsh, S. G. *et al.* Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* **75**, 291 (2010).

[36] Hurley, C. K. Naming HLA diversity: a review of HLA nomenclature. *Human Immunology* **82**, 457–65 (2021).

[37] Marsh, S. G. *et al.* Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Immunogenetics* **55**, 220–6 (2003).

[38] Okada, E. *et al.* Detecting muc1 variants in patients clinicopathologically diagnosed with having autosomal dominant tubulointerstitial kidney disease. *Kidney International Reports* **7**, 857–66 (2022).

[39] Fowler, J., Vinall, L. & Swallow, D. Polymorphism of the human *MUC* genes. *Frontiers in Bioscience* **6**, 1207–15 (2001).

[40] Cox, K. E. *et al.* The mucin family of proteins: candidates as potential biomarkers for colon cancer. *Cancers* **15**, 1491 (2023).

[41] Guo, X. *et al.* Mucin variable number tandem repeat polymorphisms and severity of cystic fibrosis lung disease: significant association with *MUC5AC*. *PLOS One* **6**, e25452 (2011).

[42] Guo, X. *et al.* Genome reference and sequence variation in the large repetitive central exon of human *MUC5AC*. *American Journal of Respiratory Cell and Molecular Biology* **50**, 223–32 (2014).

[43] Plender, E. G. *et al.* Structural and genetic diversity in the secreted mucins, *MUC5AC* and *MUC5B*. *bioRxiv* 2024.03.18.585560 (2024).

[44] Ye, Q. *et al.* Identification of the common differentially expressed genes and pathogenesis between neuropathic pain and aging. *Frontiers in Neuroscience* **16**, 994575 (2022).

[45] Blázquez-Moreno, A. *et al.* Transmembrane features governing Fc receptor CD16A assembly with CD16A signaling adaptor molecules. *Proceedings of the National Academy of Sciences* **114**, E5645–54 (2017).

[46] Hunter, S. *et al.* Inhibition of Fcγ receptor-mediated phagocytosis by a nonphagocytic Fcγ receptor. *Blood* **91**, 1762–8 (1998).

[47] Kyogoku, C. *et al.* Fcγ receptor gene polymorphisms in Japanese patients with systemic lupus erythematosus: contribution of *FCGR2B* to genetic susceptibility. *Arthritis & Rheumatism* **46**, 1242–54 (2002).

[48] Espéli, M., Smith, K. G. & Clatworthy, M. R. Fcγ RIIB and autoimmunity. *Immunological Reviews* **269**, 194–211 (2016).

[49] Verbeek, J. S., Hirose, S. & Nishimura, H. The complex association of FcγRIIb with autoimmune susceptibility. *Frontiers in Immunology* **10**, 446703 (2019).

[50] Willcocks, L. C. *et al.* A defunctioning polymorphism in *FCGR2B* is associated with protection against malaria but susceptibility to systemic lupus erythematosus. *Proceedings of the National Academy of Sciences* **107**, 7881–5 (2010).

[51] Lee, J. *et al.* Epigenetic modification and antibody-dependent expansion of memory-like NK cells in human cytomegalovirus-infected individuals. *Immunity* **42**, 431–442 (2015).

[52] de Haas, M. *et al.* A triallelic Fcγ receptor type IIIA polymorphism influences the binding of human IgG by NK cell FcγRIIIa. *Journal of Immunology* **156**, 2948–55 (1996).

[53] de Vries, E. *et al.* Identification of an unusual Fcγ receptor IIIa (CD16) on natural killer cells in a patient with recurrent infections. *Blood* 3022–7 (1996).

[54] Grier, J. T. *et al.* Human immunodeficiency-causing mutation defines CD16 in spontaneous NK cell cytotoxicity. *The Journal of Clinical Investigation* **122**, 3769–80 (2012).

[55] Jawahar, S. *et al.* Natural Killer (NK) cell deficiency associated with an epitope-deficient Fc receptor type IIIA (CD16-II). *Clinical & Experimental Immunology* **103**, 408–413 (1996).

[56] Koene, H. *et al.* FcγRIIIa-158V/F polymorphism influences the binding of IgG by natural killer cell FcγRIIIa, independently of the FcγRIIIa-48L/R/H phenotype. *Blood* 1109–14 (1997).

[57] Lejeune, J., Brachet, G. & Watier, H. Evolutionary story of the low/medium-affinity IgG Fc receptor gene cluster. *Frontiers in Immunology* **10**, 1297 (2019).

[58] Donoso, L. A., Vrabec, T. & Kuivaniemi, H. The role of complement Factor H in age-related macular degeneration: a review. *Survey of Ophthalmology* **55**, 227–246 (2010).

[59] Servais, A. *et al.* Acquired and genetic complement abnormalities play a critical role in dense deposit disease and other C3 glomerulopathies. *Kidney International* **82**, 454–464 (2012).

[60] Hoffman, S. M., Nelson, D. R. & Keeney, D. S. Organization, structure and evolution of the *CYP2* gene cluster on human chromosome 19. *Pharmacogenetics and Genomics* **11**, 687–98 (2001).

[61] Zanger, U. M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics* **138**, 103–141 (2013).

[62] Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine* **363**, 166–176 (2010).

[63] Pendergrass, S. *et al.* The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genetic Epidemiology* **35**, 410–422 (2011).

[64] Firtina, C. *et al.* Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm. *Bioinformatics* **36**, 3669–79 (2020).

[65] Warren, R. L. *et al.* ntEdit: scalable genome sequence polishing. *Bioinformatics* **35**, 4430–2 (2019).

[66] Prodanov, T. Locityper loci database, 1KGP genotypes and benchmarking data. https://doi.org/10.5281/zenodo.10977559.

[67] Prodanov, T. Locityper source code. https://doi.org/10.5281/zenodo.10979046.

[68] Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology* 1–11 (2023).

[69] Levandowsky, M. & Winter, D. Distance between sets. *Nature* **234**, 34–35 (1971).

[70] Roberts, M., Hayes, W., Hunt, B. R., Mount, S. M. & Yorke, J. A. Reducing storage requirements for biological sequence comparison. *Bioinformatics* **20**, 3363–9 (2004).

[71] Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–70 (2011).

[72] Sahlin, K. Flexible seed size enables ultra-fast and accurate read alignment. *Genome Biology* **23**, 260 (2022).

[73] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

[74] Prodanov, T. & Bansal, V. Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing. *Nature Communications* **13**, 3221 (2022).

[75] Cleveland, W. S. & Devlin, S. J. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**, 596–610 (1988).

[76] Schleimer, S., Wilkerson, D. S. & Aiken, A. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 76–85 (2003).

[77] Takaoka, T. Efficient algorithms for the maximum subarray problem by distance matrix multiplication. *Electronic Notes in Theoretical Computer Science* **61**, 191–200 (2002).

[78] Trifonov, E. Making sense of the human genome. In *Structure & Methods: Proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics*, 69–77 (1990).

[79] Mirzasoleiman, B., Badanidiyuru, A., Karbasi, A., Vondrák, J. & Krause, A. Lazier than lazy greedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29 (2015).

[80] Pincus, M. A Monte Carlo method for the approximate solution of certain types of constrained optimization problems. *Operations Research* **18**, 1225–8 (1970).

[81] Nemhauser, G. & Wolsey, L. *Linear Programming*, chap. I.2, 27–49 (John Wiley & Sons, Ltd, 1988).

[82] Huangfu, Q. & Hall, J. J. Parallelizing the dual revised simplex method. *Mathematical Programming Computation* **10**, 119–142 (2018).

[83] Gurobi Optimization. Gurobi optimizer reference manual. https://www.gurobi.com (2023).

[84] Grandvalet, Y., Canu, S. & Boucheron, S. Noise injection: Theoretical prospects. *Neural Computation* **9**, 1093–1108 (1997).

[85] Welch, B. L. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).

[86] Pavetić, F., Žužić, G. & Šikić, M. $LCSk$++: Practical similarity metric for long strings. *arXiv* 1407.2407 (2014).

[87] Marco-Sola, S., Moure, J. C., Moreto, M. & Espinosa, A. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**, 456–63 (2021).

[88] Marco-Sola, S. *et al.* Optimal gap-affine alignment in $O(s)$ space. *Bioinformatics* **39**, btad074 (2023).

[89] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

[90] Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–4 (2015).

[91] Cleary, J. G. *et al.* Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *BioRxiv* 023754 (2015).

[92] Zhou, Y., Song, L. & Li, H. Full resolution HLA and KIR genes annotation for human genome assemblies. *bioRxiv* 2024.01.20.576452 (2024).

[93] Robinson, J. *et al.* The IPD-IMGT/HLA database. *Nucleic Acids Research* **48**, D948–55 (2020).