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. Across 256 challenging medically relevant loci, Locityper achieves median quality value (QV) above 35 from both long and short read data, outperforming state-of-the-art Illumina and PacBio HiFi variant calling pipelines by 10.9 and 1.7 points, respectively. Furthermore, Locityper provides access to hyperpolymorphic HLA genes and other gene families, including KIR, MUC and FCGR. With its low running time of 1h35m 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, KIR, and CYP2 gene families, have been developed[19–24].

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 is a general-purpose genotyper that 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[25] program and the UK Biobank[26], 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 or identified through stochastic optimization (Methods). 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.

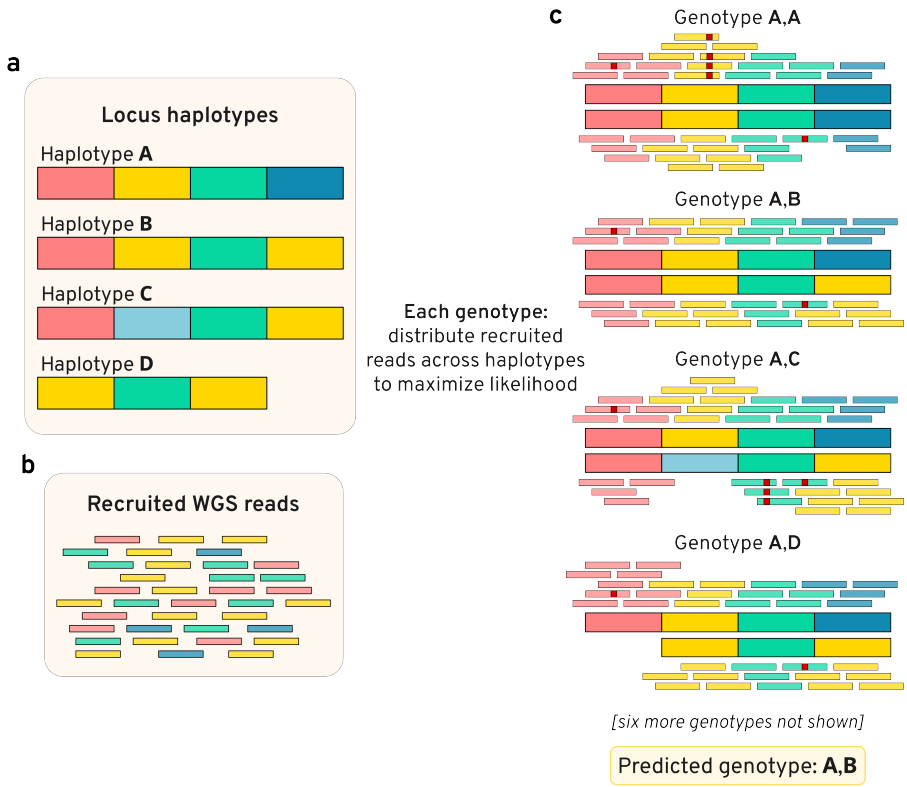

Figure 1. Illustration of the locus genotyping approach. **a**, Reference panel of four locus haplotypes *(A..D)*. **b**, Whole genome sequencing (WGS) reads, recruited to any of the haplotypes. For illustrative purposes, haplotypes and reads are colored by homologous blocks (information, unavailable 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]. Note that the actual number of haplotypes for any given locus may be smaller, as some haplotypes are not unique, while others are only partially assembled. After removing genes that overlap pangenomic variants longer than 300 kb and merging

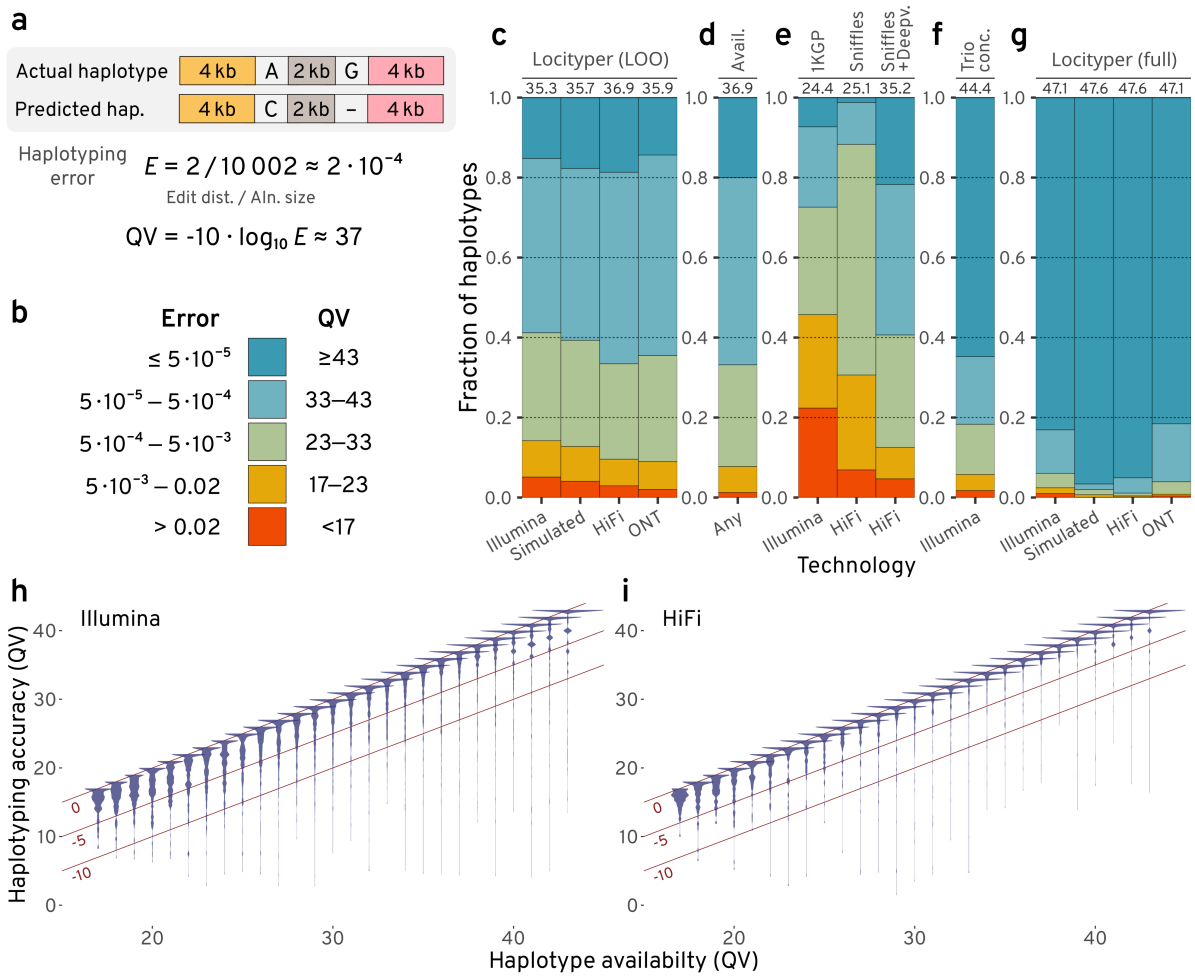

Figure 2. Haplotype accuracy definition and analysis at 256 challenging medically relevant loci. a, Haplotyping error is calculated as sequence divergence between actual and predicted haplotypes. Quality value (QV) is a Phred-like transformation of the haplotyping error. **b**, Approximate correspondence between haplotyping error and QV bins. **c**–**g**, Fraction of haplotypes across 256 loci and multiple samples, distributed into five QV bins (also see Supp. Table 2). Median QV is printed on top of each bar. **c**, Locityper accuracy in the leave-one-out (LOO) configuration. **d**, Haplotype availability (QV between actual and closest available leave-one-out haplotypes). **e**, Haplotyping accuracy of the 1KGP call set; as well as Sniffles and Sniffles+Deepvariant variant calling. **f**, Concordance of Locityper predictions across 563 unrelated trios. **g**, Locityper accuracy using the full reference panel. **h**–**i**, Correspondence between haplotype availability and haplotyping accuracy based on Illumina (**h**) and PacBio HiFi (**i**) WGS datasets. Availability <17 and >43 is not shown, but follows the same trend. Red lines mark 0, 5 and 10 point QV loss.

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).

To measure haplotyping error, we calculated sequence divergence between actual and predicted haplotypes (Figure 2a) and corresponding Phred-like[27] quality values (QV), widely used for genome assembly evaluation[28,29]. Then, we distributed haplotype predictions into five bins based on their QV (<17, 17–23, 23–33, 33-43 and ≥43), where a haplotype from the last two bins (QV≥33) differs from an actual haplotype by no more than 5 bp per 10 kb (Figure 2b), which is competitive with long read genome assemblies from ONT data[30]. 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.

First, we genotyped 40 Illumina WGS datasets from the HPRC cohort. Each dataset was processed using the leave-one-out (LOO) approach, where the two relevant sample haplotypes were excluded from the reference panel. Across 20,350 cases where locus haplotypes were fully assembled, Locityper achieved median QV=35.27, with 58.8% haplotypes having QV≥33 (15.2% for QV≥43). On the other hand, 9.1% haplotypes had QV between 17 and 23 and 5.1% haplotypes had QV≤17 (see Figures 2c, 3).

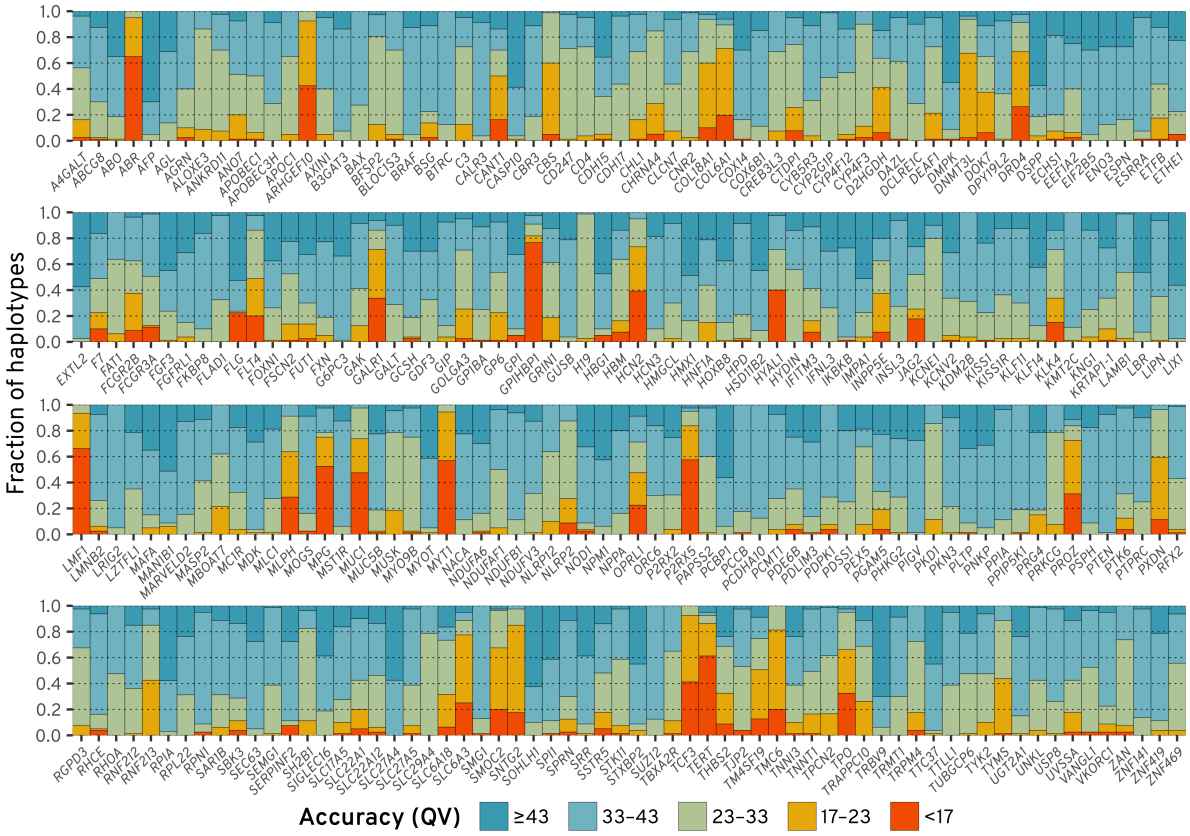

Figure 3. Locityper haplotyping accuracy using leave-one-out reference panel for 40 Illumina WGS datasets. Predicted haplotypes across 256 challenging medically relevant loci are binned into five groups according to their haplotyping quality value (QV). Corresponding haplotype availability can be found in Supp. Figure 2.

Instead of unmapped reads, Locityper can process existing alignments, significantly accelerating the read recruitment stage. This does not lead to lower accuracy, as Locityper predictions for 10 mapped WGS datasets showed virtually identical results (median QV=35.25; see Supp. Table 2).

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[31] to simulate 40 short-read datasets and processed them with Locityper. As expected, the tool showed higher accuracy on simulated datasets, producing median QV=35.65, with 60.7% and 4.0% haplotypes receiving QV≥33 and <17, respectively (see Supp. Figure 1a).

Locityper is not limited to short reads and is able to process various long read WGS datasets, including accurate PacBio HiFi and error-prone Oxford Nanopore (ONT) data. For these technologies, Locityper achieved higher median QVs of 36.90 and 35.95, respectively, and produced 66.6% and 64.5% haplotypes with QV≥33 (18.7% and 14.4% with QV≥43); while only 2.9% and 2.0% haplotypes had QV under 17 (see Supp. Figure 1b–c).

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

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 reference panel. Therefore, Locityper LOO accuracy is limited to haplotype *availability*—similarity between the actual haplotypes and the closest haplotype remaining in the LOO panel. Overall, 66.8% haplotypes had close counterparts in the LOO panel (QV≥33; 20.0% for QV≥43; see Figure 2d and Supp. Figure 2). Inversely, 1.2% and additional 6.5% haplotypes were dissimilar from any unrelated haplotype (QV<17 and 17–23).

An optimal solver, which always finds its closest counterpart from the LOO panel, would achieve median QV=36.93, just 1.66 points higher than Illumina-based Locityper and 0.03 higher than HiFi-based. For Illumina datasets Locityper underperforms on average by just 2.03 QV points compared to the theoretical best, with 95.1% (86.8%) haplotypes trailing by under 10 (5) QV points (see Figure 2h). Even further, across PacBio HiFi datasets Locityper predictions differ from optimal by 0.72 QV points on average; this number drops down to 0.54 when considering well-represented haplotypes (availability ≥33). Overall, 98.7% (96.0%) HiFi-based haplotypes were within the 10 (5) point margin (Figure 2i).

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 significantly outperforms state-of-the-art variant calling pipelines

By identifying two most similar locus haplotypes to a given WGS dataset, Locityper effectively infers diploid sample–locus sequence. Such sequence can be split into structural or short variants; nevertheless,

individual positions may be inaccurate, especially when a variant is missing from the reference panel. To perform a fairer comparison, we examined locus sequences, reconstructed from phased variant call sets.

Recently, New York Genome Center (NYGC) researchers presented an extensive variant calling pipeline and used 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], of them 39 with HPRC assemblies. Even though the NYGC pipeline utilizes state-of-the-art variant callers, 1KGP haplotypes had significant divergence from the actual sample haplotypes: only 27.4% haplotypes achieved QV≥33 and another 22.3% haplotypes had QV<17, while the median QV was 24.41, almost 11 points smaller than Locityper on Illumina reads (see Figure 2e and Supp. Figure 3).

While short read datasets are notoriously difficult to genotype at complex loci, PacBio HiFi data is arguably the easiest. To put Locityper performance in perspective we examined phased structural variant calls, generated by Sniffles[32] for 20 HiFi datasets. As Figure 2e and Supp. Figure 4a show, Sniffles alone did not achieve high levels of accuracy, producing median QV=25.09. Combining structural variants with short variant calls, produced by Deepvariant[33], raised median QV to 35.19, which is 1.71 points behind Locityper on the same data and 0.08 points behind Illumina-based Locityper. While Sniffles+Deepvariant (Supp. Figure 4b) produced larger fraction of poor haplotypes (4.7% and 7.9% with QV<17 and 17–23 against 2.9% and 6.7% for Locityper), this pipeline also produced a bigger share of extremely accurate haplotypes (21.7% against 18.7%), likely due to Locityper's inability to call novel short variants.

Locityper produces concordant trio predictions

In addition to the HPRC dataset, we genotyped the full 1KGP cohort of 3202 Illumina WGS samples, including 563 trios independent from the HPRC cohort. At each of the 256 CMR loci and for each trio we calculated concordance—similarity between child and parent haplotypes (see Methods). As Figure 2f shows, the vast majority of trio haplotypes were concordant: 64.8% and 81.7% with QV ≥43 and ≥33, respectively. Moreover, median concordance QV surpassed 44.4, and was over 43 at 90% of the loci (see Supp. Figure 5).

Locityper achieves near-perfect accuracy with full reference panel

Finally, we examined Locityper's ability to accurately identify true sample haplotypes using a full reference panel. This experiment should mimic future pangenomes, where almost all haplotypes present in the population would also exist in the reference panel. At each of the sequencing technologies, Locityper achieved extremely high median QV (>47), and produced >93% haplotypes with QV≥33. In this configuration, Illumina and ONT-based haplotypes showed slightly lower accuracy: 83.1% and

81.6% had QV≥43, respectively, while only 1.0% and 0.4% had QV<17. On the other hand, simulated short reads and PacBio HiFi datasets produced almost perfect haplotypes: 96.6% and 95.1% with QV≥43 and ≈0.1% with QV<17 (Figure 2g and Supp. Figure 6). Variant call set, obtained from the Locityper haplotypes using the full reference panel and Illumina data, showed significantly higher F_1 score than the 1KGP call set; as well as higher precision and recall compared to the pangenome-based variant caller Pangenie[18] (see Supp. Information 3.1 and Supp. Figure 7).

Locityper accurately genotypes HLA and KIR genes

In order to evaluate Locityper's 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[34]; and killer-cell immunoglobulin-like receptor (KIR) gene cluster spanning 150 kb and 17 genes[35]. The two regions contain extremely polymorphic HLA and KIR genes that play an essential role in adaptive and innate immune systems[36,37]. 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. As such, Locityper can predict missing genes by selecting padded haplotypes that lack the gene of interest. Nevertheless, 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[19,22,38]; the newest of them being T1K[23], a state-of-the-art[39] 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 26 genes and 14 pseudo genes from the MHC locus as well as 14 genes and 3 pseudogenes from the KIR locus, all combined into 33 target loci with sum length 1.15 Mb.

In the leave-one-out configuration, at the MHC locus Locityper achieved full match with assembly-based allele annotation (correctly predicted all fields in the HLA nomenclature[40,41]) in 88.8% cases, compared to T1K's 64.1% (Figure 4a). At the same time, the two methods correctly predicted protein product (second nomenclature field) in 95.1% and 78.2% cases, respectively. Meanwhile, at the KIR gene cluster Locityper and T1K correctly predicted protein products in 84.9% and 67.1% cases and achieved full match 80.8% and 57.9% cases, respectively (see Figure 4b). When using the full reference panel, containing input samples, Locityper achieved almost perfect accuracy: full match in 99.4% and 99.9% cases at the MHC and KIR loci, respectively.

In contrast to T1K, Locityper does not distinguish between exons, introns and intergenic space. This may result in lower accuracy when a haplotype carrying a false gene allele better explains input reads within non-coding sequence. To handle such cases, users may employ a weighted Locityper mode,

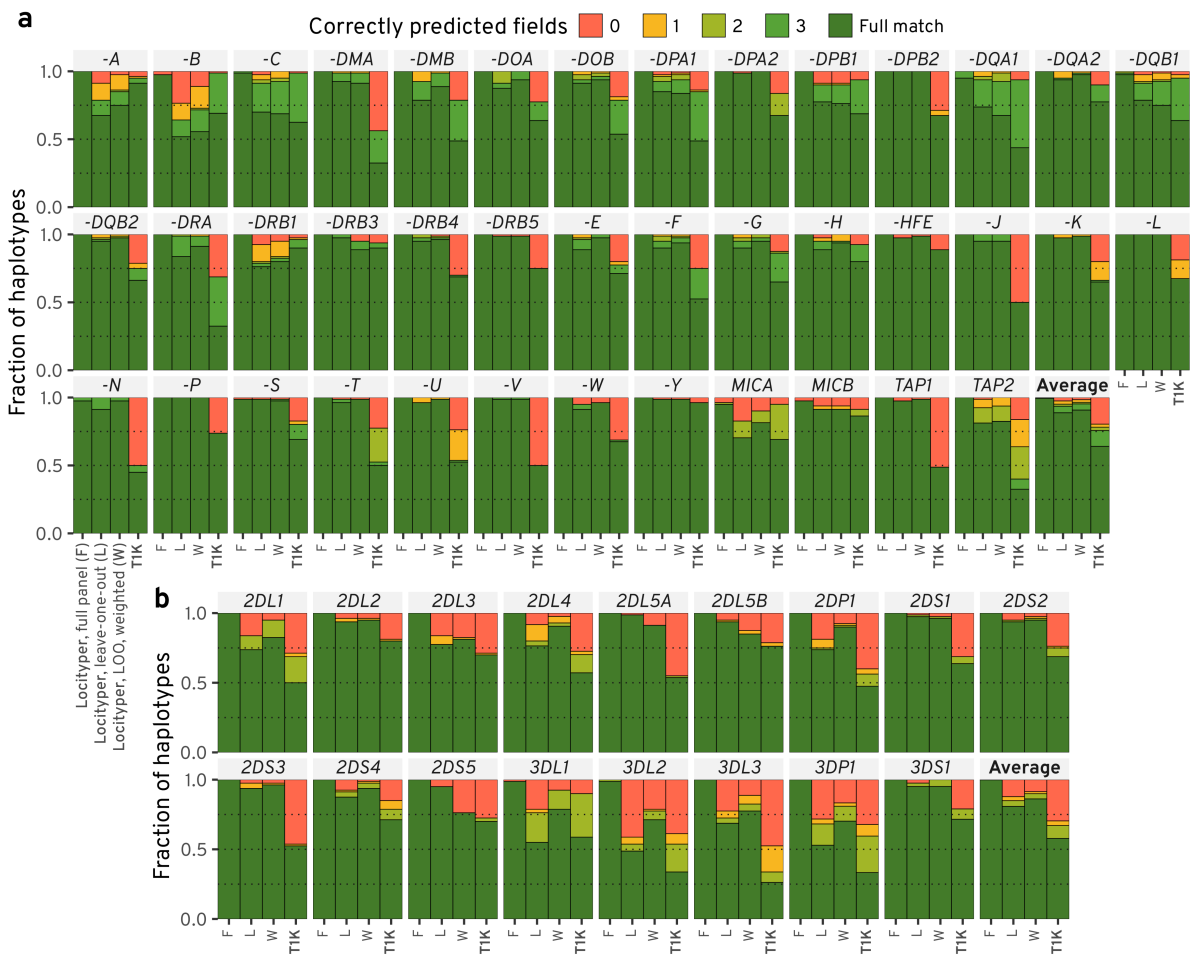

Figure 4. Haplotyping accuracy for 40 HPRC samples at the MHC and KIR loci. Subpanels show fraction of haplotypes, predicted with varying accuracy at 40 (pseudo)genes from the MHC locus (**a**) and 17 (pseudo)genes from the KIR gene cluster (**b**). Fully predicted alleles, as well as correctly identified missing copies, are colored with dark green *(Full match)*, due to the different number of allele fields in the HLA/KIR gene nomenclature[40,42]. Otherwise, haplotypes are colored according to the number of correctly predicted fields. Accuracy is shown for Locityper with the full reference panel (F); Locityper in the leave-one-out setting without and with weights (denoted L and W, respectively); and T1K. Last entry in each panel shows average accuracy across all corresponding genes/pseudogenes.

giving lower value to read depth and read alignments occurring outside of exons (0.1 within introns and 0.005 within intergenic regions). In the weighted mode Locityper accuracy rose to 96.5% (protein product) and 90.8% (full match) at the MHC locus; as well as to 89.9% and 86.2% at the KIR cluster (Figure 4).

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 explained 64.6% and 36.2% of all errors made by Locityper in the weighted mode at the MHC and KIR loci, respectively (see Supp. Figure 8). This is especially noticeable at the hyperpolymorphic *HLA-A*, *-B* and *-DRB1* genes, where protein groups were missing from the LOO panel in 10–22% cases and explains vast majority of Locityper errors. At the same time, T1K often predicted smaller copy number than required, explaining

79.1% and 24.6% of all errors at the MHC and KIR clusters, respectively. When ignoring these two error types (missing copy and unavailable protein groups), Locityper notably outperforms T1K in predicting protein products: 99.0% against 94.5% at the MHC locus, as well as 94.6% against 73.0% at the KIR gene cluster. Overall, the general purpose tool Locityper performs in a competitive manner even when compared to T1K, which is specifically designed for HLA/KIR genes. However, accurate genotyping of the most diverse genes would still likely benefit from larger pangenome sizes.

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*)[45]. 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[46]. 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

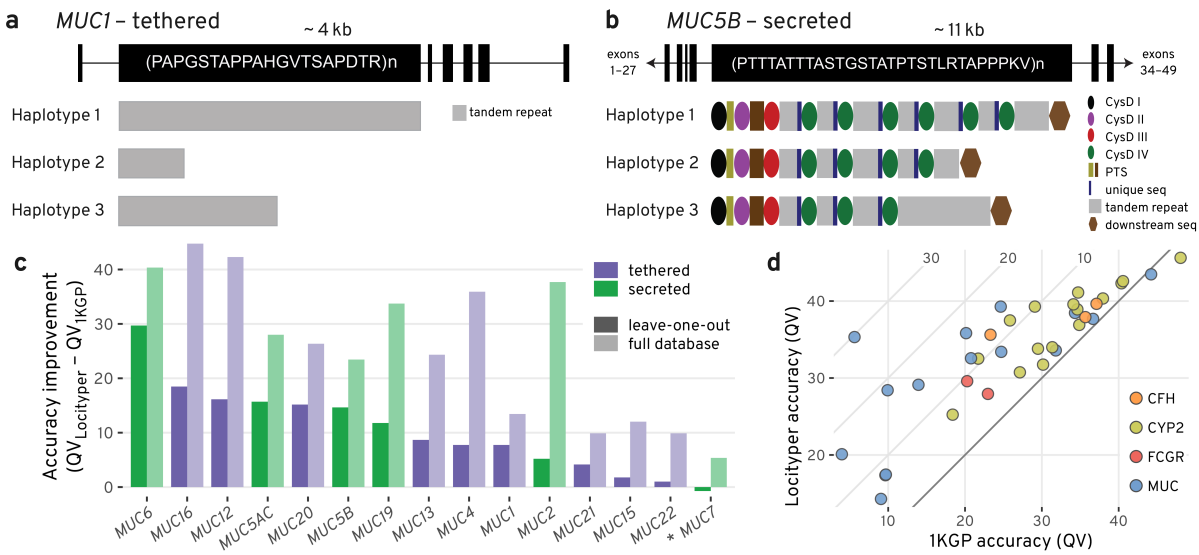

Figure 5. Locityper can accurately genotype mucin and other gene families. **a**, Gene model of *MUC1*, a mucin tethered to the surface of epithelial cells. *MUC1* harbors a 20 amino acid VNTR repeat sequence and is highly polymorphic in VNTR length, as represented by the example haplotypes 1–3[43]. **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[44]. **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.

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[47,48] (Figure 5b). 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 29.7 and 18.5 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)[49,50]. 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[51]). Genetic variations in this gene have been reported to cause increased susceptibility to systemic lupus erythematosus (SLE)[52], as well as to malaria, autoimmune hypersensitivity disease, and immune thrombocytopenic purpura[53–55]. The second receptor, FCGR3A, is expressed on natural killer cells (NKc) as an integral membrane glycoprotein[50]. 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[56]. Similar to *FCGR2B*, genetic variants in the *FCGR3A* gene have been associated with SLE[52], as well as immunodeficiency, susceptibility to recurrent viral infections, and alloimmune neonatal neutropenia[57–61]. 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[62]. Nevertheless, at the *FCGR2B* and *FCGR3A* receptor genes Locityper (LOO) improves average QV by 4.95 and 9.3 points compared to the 1KGP call set, respectively (23.0 → 27.9 and 20.3 → 29.6) 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 (35.6 and 54.0) when using its full reference panel.

Moreover, Locityper (LOO) achieves significant QV improvement (12.3) at the *CFH* gene, associated with age-related vision loss and kidney disorders[63,64]. Finally, Locityper showed on average 4.6 higher QV across 16 protein coding CYP2 genes that play a major role in drug metabolism[65,66]. Out of

the CYP2 genes, Locityper achieved the highest improvement at *CYP2U1* (10.2), *CYP2A13* (10.8) and *CYP2W1* (11.6) (see Figure 5d).

Runtime and memory usage

Locityper WGS preprocessing (executed once per dataset) took on average 16 minutes (m) using 8 threads and consumed 15 Gb of RAM for 30× Illumina WGS dataset. If a dataset with a similar library preparation was previously processed, 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) and lasted under 15m on average. Next, mapping reads to the reference panels across 256 target loci took under 19m using 8 threads, and locus genotyping consumed another 45m. Together, these two steps required approximately 15 seconds per target locus and 7 Gb of RAM. Locityper employs stringent haplotype filtering as the first genotyping step, allowing it to avoid quadratic runtime. Because of that, full analysis based on five times larger reference panels (obtained by artificially mutating existing haplotypes) required only three times as much time (see Supp. Figure 9).

Altogether, Locityper analysis of the MHC and KIR loci, including preprocessing, required 35m using 8 threads. However, genotyping in the weighted mode was more computationally intensive, raising the total runtime to 1h5m. 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 10m.

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 or 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 563 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 state-of-the art phased variant calling pipelines on Illumina and PacBio HiFi data. 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: overall 33% haplotypes did not have a good representative (QV<33) in the leave-one-out reference panels (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 256 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[67] and PheWAS[68] analyses, especially in a larger cohort, such as the All-of-Us program[25] and the UK Biobank[26], 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[69,70], this could improve genotyping accuracy and allow for reconstruction of previously unobserved alleles—a strategy that we plan to explore in future research.

Currently, two loci with significant homology, e.g. part of a non-tandem segmental duplication, can only be 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[23].

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[30]. 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[71] (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[72] (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[73]).

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.

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[74] with 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 chr17q25.1 as the default background region because it contains almost no segmental duplications or other types of structural variations. Locityper first recruits input reads to the background region (see "Read recruitment"), optionally subsamples them, and then maps them to the reference genome using Strobealign[75] (short reads) or Minimap2[76] (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 10). 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)[77].

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[78]. NB parameters are then estimated separately for each GC-content based on the smoothed mean, variance, and subsampling rate (see Supp. Information 3.2).

Read recruitment

Following dataset preprocessing, Locityper recruits reads to all target loci. For that, we collect minimizers[79] from each locus and each haplotype (default: (10,15)-minimizers). 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 0.5 for single- and paired-end 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[80]. This heuristic is useful as it can be quickly evaluated using Kadane's algorithm[81], and 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[75] or Minimap2[76], 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. This way, insert size between a read end and its unmapped counterpart is assumed to be optimal in order to only penalize unpaired locations once. Finally, we will denote the full set of possible read pair locations on haplotype h as $L^{(h)} \subset W^{(h)} \times W^{(h)}$ and will define the probability of the read pair $\mathbf{r}$ location to be $\mathbf{w}$ as normalized alignment probability:

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

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 $|W^{(h_1)}| + |W^{(h_2)}|$. Similarly, we concatenate possible locations $L^{(h_1)}$ and $L^{(h_2)}$ to achieve a combined list of locations $L^{(\mathbf{g})}$.

We describe read assignment to the genotype $\mathbf{g}$ using a boolean matrix T , where $T_{\mathbf{rw}} = 1$ encodes the statement "true location of the read pair $\mathbf{r}$ is $\mathbf{w}$" and every row contains exactly one true element. Probability of the read assignment T given read pairs R can be described as the total probability of all selected locations:

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

Read depth likelihood

In addition to good alignment probabilities, optimal haplotypes should have stable haploid read depth. Corresponding conditional probability can be written as

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

Here, $d_w(T)$ denotes window w depth according to the read assignment T, defined as $\sum_{\mathbf{r}} \sum_u \left[T_{\mathbf{r},wu} + T_{\mathbf{r},uw}\right]$. At CN = 1, read depth follows the NB distribution with precomputed parameters n and ψ. Bayes'

theorem with equal priors produces the following result:

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

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 $\vartheta : [0, 1] \mapsto [0, 1]$:

$$\vartheta(x; \eta, q) \;=\; \begin{cases} 0 & \text{if } x = 0, \\ \dfrac{1}{\left(\frac{\eta}{x} \cdot \frac{1-x}{1-\eta}\right)^q + 1} & \text{otherwise.} \end{cases} \tag{6}$$

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 controls the "slope" of the function, with larger q producing larger derivative $\vartheta'(\eta; \eta, q)$ (see Supp. Figure 11). 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 $x_2 = U_1 U_2 U_3$, where U_i is the fraction of unique i-mers in the window w out of the maximal possible number of distinct i-mers[82], with default parameters $\eta_1 = 0.2$, $\eta_2 = 0.5$ and $q_1 = q_2 = 4$.

Locityper accepts explicit user-defined weights for each basepair of the input haplotypes, e.g. down-weighting non-coding sequence. In such cases, ζ_w is multiplied by average weight across window w, while each read receives its own weight based on the maximum explicit weight under the primary alignments of both read ends. After that, read weights are used as multipliers to log location probabilities $\log \mathcal{P}$.

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:

$$\begin{aligned} P\big(\mathrm{CN}(\mathbf{g}) = 1, T \mid R\big) &= P\big(T \mid R\big) \times P\big(\mathrm{CN}(\mathbf{g}) = 1 \mid T\big) \\ &= \prod_{\mathbf{r} \in R} \sum_{\mathbf{v} \in L^{(\mathbf{g})}} T_{\mathbf{rv}} \cdot \mathcal{P}_{\mathbf{rv}} \times \prod_{w \in W^{(\mathbf{g})}} \varphi_w(T). \end{aligned} \tag{7}$$

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:

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

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 factors should sum up to 2 in order to generate the same range of likelihoods as in the unweighted case ($\Omega_R = \Omega_D = 1$). We use default values $\Omega_R = 0.15$ and $\Omega_D = 1.85$, as they produced good results across a selection of target loci and sequencing datasets.

Likelihood update

Given log-likelihood $\mathcal{L}_T^{(\mathbf{g})}$ for genotype $\mathbf{g}$ and some read assignment T, we can efficiently calculate log-likelihood $\mathcal{L}_{T'}^{(\mathbf{g})}$ for a new read assignment T' if the read assignment has changed for only one read pair. Suppose that the read assignment changed for read pair $\mathbf{r}$ from location uv (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:

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

Finding best read assignment

For each genotype $\mathbf{g}$ we aim to find such read assignment T that would maximize joint log-likelihood $\mathcal{L}_T^{(\mathbf{g})}$. Locityper implements three approaches for finding such read assignment: Stochastic Greedy approach[83], Simulated Annealing[84] and Integer Linear Programming (ILP)[85]. 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 $\mathbf{w} \in L^{(\mathbf{g})}$; 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:

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

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[86] and Gurobi[87]. Note that it is possible to state a bigger ILP problem removing the need to iterate over all possible genotypes (see Supp. Information 3.3). However, we observed that existing ILP solvers are unable to quickly and accurately find a solution to such problem.

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[88], 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[27] 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[89]. Additionally, we redefine genotype probabilities as the probability of having the highest true likelihood, calculated as the product of inverse t-test p-values for all pairwise genotype comparisons.

Moreover, Locityper outputs the number of unexplained reads—reads that map to some, but not to the two predicted haplotypes. Finally, Locityper outputs weighted Jaccard distance between minimizers of the primary genotype and other probable genotypes. In an unambiguous prediction, this value should be low as all likely genotypes should be similar to each other. Users can employ these values for conservative post-genotyping filtering, for example in the HiFi-based LOO evaluation, discarding 20.2% genotypes with over 50 unexplained reads raises median QV from 36.9 to 38.2.

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 33 loci covering genes from the MHC and KIR gene clusters, as well as 31 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[31] `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++[90] and completes the alignment between k-mer matches using Wavefront alignment algorithm[91,92]. 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)[28,29]. 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.

We consider a trio of locus genotypes concordant if one of the child haplotypes closely matches one of the maternal haplotypes, and another closely matches one of the paternal haplotypes. Similarly to

haplotyping error calculation, we iterate over 8 possible combinations; select one with the smallest sum edit distance divided by sum alignment size; and calculate QV score for each of the child haplotypes.

To compare Locityper with state-of-the-art PacBio HiFi pipelines, we obtained existing[8] unphased Deepvariant[33] `v1.1.0` SNP and indel calls for PacBio HiFi HPRC datasets, which we phased using WhatsHap[93] `phase v2.3`. Next, we used WhatsHap `haplotag` to assign reads to haplotypes and used Sniffles[32] `v2.4` to generate phased SVs. Finally, we used RTG[94] `vcfmerge v3.12.1` to generate merged Sniffles + Deepvariant call set.

We used Bcftools[95] (`v1.21`) `consensus` to reconstruct haplotypes from each of the three phased variant call sets (Sniffles, Sniffles + Deepvariant, and 1KGP[3]). In the process, we removed contradicting overlapping variant calls, as well as variants with symbolic alternative alleles (with exception of `<DEL>`), as they cannot be used for haplotype reconstruction.

Finally, we used T1K[23] `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[96] using allele databases[36,97] IPD-IMGT/HLA `v3.55` and IPD-KIR `v2.13`. 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 `v0.18.0` along with its dependencies Samtools[95] `v1.21`, Jellyfish[74] `v2.2.10`, Strobealign[75] `v0.13.0` and Minimap2[76] `v2.26-r1175`.

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] Numanagić, I. *et al.* Cypiripi: exact genotyping of *CYP2D6* using high-throughput sequencing data. *Bioinformatics* **31**, i27–i34 (2015).

[21] Numanagić, I. *et al.* Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. *Nature communications* **9**, 828 (2018).

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

[23] 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).

[24] Hari, A. *et al.* An efficient genotyper and star-allele caller for pharmacogenomics. *Genome Research* **33**, 61–70 (2023).

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

[26] 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).

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

[28] 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).

[29] 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).

[30] 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).

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

[32] Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology* 1–10 (2024).

[33] Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018).

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

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

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

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

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

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

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

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

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

[43] 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).

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

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

[46] 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).

[47] 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).

[48] Plender, E. G. *et al.* Structural and genetic diversity in the secreted mucins, *MUC5AC* and *MUC5B*. *The American Journal of Human Genetics* **111**, 1700–16 (2024).

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

[50] 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).

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

[52] 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).

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

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

[55] 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).

[56] 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).

[57] 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).

[58] 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).

[59] 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).

[60] 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).

[61] 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).

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

[63] 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).

[64] 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).

[65] 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).

[66] 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).

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

[68] 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).

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

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

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

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

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

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

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

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

[77] 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).

[78] 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).

[79] 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).

[80] 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).

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

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

[83] 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).

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

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

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

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

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

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

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

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

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

[93] Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. *bioRxiv* 085050 (2016).

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

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

[96] Zhou, Y., Song, L. & Li, H. Full resolution HLA and KIR genes annotation for human genome assemblies. *Genome Research* **34**, 1931–41 (2024).

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