Improved genome inference in the MHC using a population reference graph

Alexander Dilthey1, Charles Cox2, Zamin Iqbal1, Matthew R Nelson3 & Gil McVean1

Although much is known about human genetic variation, such information is typically ignored in assembling new genomes. Instead, reads are mapped to a single reference, which can lead to poor characterization of regions of high sequence or structural diversity. We introduce a population reference graph, which combines multiple reference sequences and catalogs of variation. The genomes of new samples are reconstructed as paths through the graph using an efficient hidden Markov model, allowing for recombination between different haplotypes and additional variants. By applying the method to the 4.5-Mb extended MHC region on human chromosome 6, combining 8 assembled haplotypes, the sequences of known classical HLA alleles and 87,640 SNP variants from the 1000 Genomes Project, we demonstrate using simulations, SNP genotyping, and short-read and long-read data how the method improves the accuracy of genome inference and identified regions where the current set of reference sequences is substantially incomplete.

The current paradigm for analyzing human genomes using high-throughput sequence (HTS) data is to map to a single haploid reference sequence in which there is no representation of variation1–3. Across much of the genome, such exclusion has little effect on the accuracy of genome inference because of the relatively low genetic diversity of humans4. However, for some regions, such as the major histocompatibility complex (MHC) region on chromosome 6, which contains the human leukocyte antigen (HLA) genes, there is very substantial sequence and structural variation5. Such diversity can result in poor genomic characterization in individuals who carry sequence that is either missing or highly divergent from the single reference. Other locations of high diversity include the KIR6 region, olfactory gene clusters7, ancient inversions such as that at 17q21.31 (refs. 8–10) and regions of recurrent genomic rearrangement11, many of which have substantial influence on phenotype and disease risk. In many of these cases, multiple alternative haplotypes have been characterized and are available. For example, there are seven alternative plus one primary (PGF) MHC haplotypes in the human reference (GRCh37). More generally, sequencing projects have greatly advanced understanding of human genetic variation12–14, and using such information to help characterize human genomes represents an important and unsolved problem.

The problem of the single-reference approach and the potential for using known MHC variation are demonstrated in Figure 1. Mapping to the standard reference (the PGF MHC haplotype) results in large fluctuations in coverage and many poorly mapped reads (Fig. 1a). However, when the reference is augmented with an additional haplotype, identified by comparing the classical HLA genotypes of the sample with those of the eight reference haplotypes and noting that one of the eight haplotypes is a close match, read coverage and alignment are greatly improved (Fig. 1b,c).

Using prior information about variation raises five main challenges. First, a data structure for representing genomic variation must be defined, which can accommodate multiple sources of information, from assembled reference sequence (such as the ALT paths in GRCh37) to catalogs of small variants such as the 1000 Genomes Project12,14. Second, algorithms must exist for matching HTS data to the variation-aware reference structure. Third and potentially simultaneously with step two, additional variation not yet represented in the reference data structure must be detected. Fourth, because most functional information (such as gene location and structure) uses the coordinates of a single linear reference, information from a variation-aware reference must be projected onto a primary sequence. Finally, benchmarks must be established to validate and compare the output from a variation-aware reference tool chain to that provided by existing approaches.

Thus far, these challenges have only been partially addressed. Traditional multiple-sequence alignments (MSAs), representing inter- and intraspecies genetic variation, have been generalized to partial-order alignment (POA) graphs15, to represent shared sequence and mosaic sequences arising from recombination, and then further to A-Bruijn16 and cactus17 graphs to provide support for rearrangements and duplications. However, these graphs have not been used in the assembly of individual genomes from HTS data. Conversely, multiple approaches of mapping individual reads to variation-aware data structures have been proposed18–22. However, none of these are practical for representing a heterogeneous catalog of population variation with large and small events and the additional mutation- and recombination-driven differences found between reference material and the sample being studied.
Here we present a solution to these challenges. We describe a structure for representing known variation called a population reference graph (PRG) and a series of algorithms that enable characterization of the genomes present in an individual from HTS data. We build on previous work for using colored de Bruijn graphs to analyze sequence variation but also take advantage of the existing tool chain for read mapping and variant calling. To demonstrate the value of the method, we develop a PRG for the MHC region and combine simulation with analysis of empirical data on SNP genotypes, classical HLA types, and short-read and synthetic long-read Moleculo data.

RESULTS

The population reference graph (PRG)

A PRG is a directed, acyclic graphical model for genetic variation generated by using known allelic relationships between sequences (Fig. 2a,b and Supplementary Fig. 1). The graph is constructed in three steps (see the Supplementary Note for details). First, reference sequences are aligned using standard MSA methods. Second, a graph structure is generated from the MSA by collapsing aligned regions with sequence identity over a defined k-mer size. This structure is related to the POA graph, although it differs in preserving more information about local haplotype structure, which is important for read alignment in regions of high sequence diversity. Third, small variants, defined (as in VCF) by a reference position and alternative alleles, are added to all valid paths (i.e., a SNP cannot be added to a path with a deletion). Here we use the primary assembly and seven MHC ALT sequences from GRCh37, along with SNPs from Phase 1 of the 1000 Genomes Project and classical HLA allele sequences from the International Immunogenetics Information System (IMGT) at key HLA class I and class II loci (Supplementary Table 1). The resulting graph structure can be thought of as a generative model for genomes. From a limited set of input sequences, many different paths through the graph are possible, capturing the effect of recombination.

Using the PRG to infer individual genomes

The use of HTS in humans largely relies on genome(s) being closely related to the reference, thus enabling reads to be mapped accurately and with appropriate certainty. We extend this idea by inferring the (diploid) path through the PRG that most closely resembles the two haplotypes of the sample. Specifically, by comparing the HTS data from a sample to the PRG, we construct a diploid personalized reference genome, here referred to as a chromotype (which could be generalized to higher ploidies or mixtures). To infer new variation, we map reads to the chromotype and use existing variant calling software. A chromotype for a diploid is best understood as a bifurcating-merging subgraph of the PRG, analogous to paired homologous chromosomes with bubbles at regions of divergence.

To infer chromotypes, we exploit the computational efficiency of hidden Markov models (HMMs). Briefly, HTS data are summarized (using Cortex) by the counts of each string of length k (k-mer). Similarly, the set of k-mers present in the PRG is enumerated, eliminating those that occur at more than one level within the PRG (those that are paralogous) and are hence uninformative (Fig. 2c). Finally, by using a probabilistic model for the emission of k-mers (Online Methods), the Viterbi algorithm infers the maximum-likelihood chromosome (Fig. 2d). Note that this approach does not preserve long-range haplotype phase information and cannot detect variants absent from the PRG. In addition, the focus on diagnostic k-mers limits our ability to analyze low-complexity regions, such as segregating segmental duplications, where read depth information is required for genotyping.

To detect new variation, the inferred chromotype is decomposed into two haplotypes (with arbitrary phasing between adjacent bubbles), which then replace the homologous region in the primary reference. Reads are mapped to the two resulting reference genomes and placed at their best position across the two reference genomes, as measured by mapping quality, or uniformly, if mapping qualities are identical (Fig. 2e). A standard variant caller is used to discover new alleles independently in the two mappings, and a heuristic algorithm modifies the chromotype, incorporating new variants (Fig. 2f). We have also developed an algorithm for mapping reads directly to the chromotype; however, this method is currently too slow for analyzing millions of reads and hence was only used for Moleculo validation.

Validation and comparison to other methods

To assess the value of the PRG approach in characterizing variation within samples, we used simulations and empirical data analysis. We compared four approaches to characterizing variation:

1. As a baseline, we used a single reference (the PGF haplotype within the MHC region from GRCh37) and looked at the effect of calling a sample as everywhere homozygous reference (PGF reference).
2. We used a read mapping approach (Stampy followed by Platypus) in which the components were designed explicitly
Figure 2  Schematic showing the construction and application of a PRG. (a) Multiple sources of information about genetic variation, including alternative reference haplotypes (lines), classical HLA alleles (rectangles), and SNPs or short indels (triangles), are aligned. Colors indicate divergent sequence; dashes indicate gaps. (b) A PRG is constructed from the alignment, resulting in a generative model for variation within the region. SNPs, indicated by diamonds, are added as alternative paths to all valid backgrounds (excluding sequence with gaps or a third allele at the position). (c) The PRG is compared to the de Bruijn graph constructed from reads obtained from a sample. Informative k-mers (those that are found at only one level in the PRG) are identified (dark blue). k-mers found elsewhere in the genome (yellow) are ignored. (d) A HMM is used to infer the most likely pair of paths through the PRG, allowing for read errors, resulting in an individualized reference chromotype for the sample. (e) Two haploid genomes are constructed from the reference chromotype, with arbitrary phasing between adjacent bubbles, and reads (light blue lines) from the sample are aligned and assigned (on the basis of mapping quality) to a reference, thus identifying places where the sample contains new variation (red circles; only one path through the chromotype is shown). (f) Newly discovered variants modify the reference chromotype, resulting in the inferred chromotype for the sample.

for high-sensitivity detection and genotyping of short indels and clustered variants (‘Platypus’). The resulting VCF was converted into a chromotype (Online Methods) for comparison.

3. From the PRG, we assessed the Viterbi chromotype, representing a ‘best-guess’ diploid path through the PRG (‘PRG-Viterbi’). These were also reported as VCFs.

4. From the PRG, we assessed the mapping-modified Viterbi chromotypes, containing variants not represented in the PRG (‘PRG-Mapped’). These were also reported as VCFs.

Simulations

To verify that the method and implementation could work, before validation based on empirical data, we simulated high-coverage HTS data (101-bp paired-end reads from a genome with 30× coverage with an empirical error distribution) for 20 individuals. Each simulated diploid genome consisted of two random paths through the PRG for the extended MHC region. The simulated genomes carried a mixture of recombination events between the original eight MHC haplotypes, SNPs and structural variants of varying size (insertions and deletions from 1–125,000 bp). From the simulated data, we inferred, for each sample, the pair of paths through the PRG using the HMM and measured allele concordance with the simulated paths at each level within the PRG (Supplementary Table 2). Across all levels (broadly corresponding to positions in the sequence), 99.89% of alleles were correctly recovered. The accuracy at heterozygous SNP positions was similar (99.83%) and dropped slightly for indel positions (ranging from 95.8 to 100%; Fig. 3a, b).

Experiment 1: comparison to SNP array data

To assess the ability of the PRG approach to genotype variation at sites of high uniqueness within the genome, we measured allele concordance at SNP positions within the extended MHC region independently interrogated through array genotyping and HTS, including for one sample (NA12878) at 60× coverage with 100-bp paired-end reads and Illumina Omni 2.5M array data and five clinical samples (CS2–CS6) at 30× coverage with 90-bp paired-end reads and Illumina 1M array data (Online Methods).

The accuracy of all approaches was high (Fig. 3a): there was ≥97.38% allele concordance with the Illumina Omni 2.5M array (NA12878) and ≥99.53% allele concordance with the Illumina 1M array (CS2–CS6). Comparing the array genotype concordance of Platypus-generated genotypes and PRG-generated genotypes (PRG-Viterbi and PRG-Mapped), we found that both approaches yielded comparable accuracies (97.75% versus 97.45% for the 2.5M array and 99.57% versus 99.66% for the 1M array; Fig. 3c and Supplementary Table 3).

Figure 3  Simulation study and empirical validation. (a) Allele concordance between simulated data (20 simulated diploid individuals; 101-bp reads at 30X diploid coverage with empirical error distribution) and the Viterbi path through the PRG, stratified by simulated variant type (SNP or structural variant; SV) and genotype. (b) Allele concordance in simulations at sites heterozygous for structural variants of different lengths. (c) Allele concordance between SNP array genotypes and chromotypes from each method for NA12878 (squares; Illumina Omni 2.5M array) and the CS2–CS6 samples (stars; Illumina 1M array), stratified by whether the array specifies the genotype as homoygous (hom) or heterozygous (het). Results are shown for the mapping-based approach (Platypus; red), the Viterbi path through the PRG (PRG-Viterbi; pink) and the Viterbi path after mapping to the reference chromotype (PRG-Mapped; blue). (d) Allele concordance between classical HLA genotypes and HLA-DRB1 (measured at a per-base level) and chromotypes from each method for NA12878 and the CS2–CS6 samples (the range of accuracy across CS2–CS6 is displayed as vertical bars). Classical HLA genotypes were inferred from sequence-based HLA typing (Online Methods).
Of the 285 sites at which the array genotypes for NA12878 disagreed with the Viterbi chromotype, in 55 cases this difference was driven by the Viterbi chromotype specifying a gap character suggesting the presence of an indel that could interfere with array genotyping. We manually inspected the alignment\(^1\) of NA12878 reads for these sites and found clear evidence for the presence of a deletion in 33 of the 55 cases (visualizations of read mapping at all positions are provided as the Supplementary Data Set). These findings suggest that a substantial fraction of the discrepancy between array and PRG approaches results from array errors at polymorphic indels. The cause of the remaining discrepancies is not understood.

Because almost all variant sites reported in NA12878 were present within the PRG, we also assessed the accuracy of variant detection and genotyping for sites by comparing calls to an independent call set on the same data generated through \textit{de novo} assembly with Cortex\(^2\). At sites within the graph, we found that all methods performed well (allele concordances: Platypus, 96.7%; PRG-Viterbi, 96.7%; PRG-Mapped, 97.2%). At sites not in the PRG, all methods showed poorer performance, although the mapped step substantially improved accuracy (Platypus, 65.9%; PRG-Viterbi, 40.2%; PRG-Mapped, 55.1%).

**Experiment 2: comparison to classical HLA data**

In regions of high sequence diversity, such as the classical HLA alleles, single-reference mapping and variant calling methods may perform poorly because of the density of mismatches to the reference. To assess the accuracy of different methods at the classical HLA loci, we compared the per-base diploid genotypes inferred by mapping and PRG approaches to those expected from the results of sequence-based typing of the highly polymorphic exons of class I (HLA-A, HLA-B and HLA-C) and class II (HLA-DQA1, HLA-DQB1 and HLA-DRB1) genes in NA12878 and CS2–CS6. We analyzed agreement with the reference sequence for the reported allele (in HLA nomenclature, this means XX:XX:01 or XX:XX:01:01 at six- or eight-digit resolution, respectively, although we note that typing was not carried out at this resolution). This analysis is distinct from classical HLA typing, where the presence of a particular set of haplotypes is inferred.

For class I loci (HLA-A, HLA-B and HLA-C), we found comparable and high (typically ≥99%) accuracy for all methods (no comparison had a difference with significance \(P < 0.01\) by paired \(t\) test; Fig. 3d and Supplementary Table 4). In contrast, for class II loci, PRG methods were significantly more accurate at HLA-DQA1 \((P = 0.002)\) and HLA-DRB1 \((P = 0.001)\) than Platypus, with no difference between the PRG-Viterbi and PRG-Mapped methods. For example, at HLA-DRB1, we found 97.19% allele concordance with the PRG-Mapped genotypes versus 89.85% concordance with the mapping-based genotypes in the CS2–CS6 samples. The main difference between class I and class II loci is the existence of polymorphic paralogs and pseudogenes within

![Figure 4](image_url)

**Figure 4** Recovery of chromotype \(k\)-mers from HTS data. (a) Number of recovered (blue) and non-recovered (red) \(k\)-mers present in chromotypes inferred by the four methods (as for Fig. 3c with the addition of a single reference represented by the PGF MHC haplotype). A \(k\)-mer is counted as recovered if it appears in HTS data from NA12878 (~60x coverage with 100-bp paired-end reads represented by an uncleaned Cortex graph; \(k = 31\)). Chromotypes within regions of clustered variants are disentangled using a greedy algorithm before evaluation, optimizing so that the disentangled haplotypes contain as many \(k\)-mers recovered in the sample as possible (Supplementary Note). (b) Spatial pattern of \(k\)-mer recovery along the extended MHC region for each of the four chromotypes showing the location of classical HLA loci. Recovery fraction was averaged over 1-kb windows.

![Figure 5](image_url)

**Figure 5** Spatial recovery of \(k\)-mers within the HLA class II region. (a) Enlarged view of \(k\)-mer recovery in Figure 4b in the MHC class II region for the chromotypes inferred by the four approaches. (b) Fraction of \(k\)-mers predicted to be present along the region that are also present in the PGF reference haplotype (1-kb windows). (c) Fraction of positions in the chromosome that correspond to gaps in the MSA used to construct the PRG (1-kb windows). Note that the PRG-Mapped chromosome is effectively identical to the PRG-Viterbi path. (d) Fraction of positions in the inferred chromotypes that are heterozygous (lines; note that this includes sites where one allele is a gap character) and the ending points of chromosome bubbles (points). The PGF reference is not shown in b–d.
Figure 6 Alignment of synthetic long-read data to chromotypes. (a) Histogram of the scaled edit distance (the number of non-concordant columns in the alignment between a read and a chromotype divided by the total number of bases in the read) between long-read data (Illumina NA12878 Molecule extended MHC–specific reads) and chromotypes inferred by the four methods. The lower boundary for each interval is omitted for clarity. The inset shows an enlarged view for reads with scaled edit distance >0.01. (b) Left, dot plot between the sequence of a Moleculo contig and the sequence of the non-gap branch of the Viterbi chromotype for NA12878 over the region highlighted in Figure 5a. There is a point (x, y) if and only if the 10-mer beginning at position x in the chromotype segment is identical to the 10-mer beginning (or its reverse complement) at position y in the read. Green indicates the region of the read that, according to the alignment, is matched to the target region (each green point represents a read k-mer between the leftmost and the rightmost read k-mers aligned to the target region). Blue indicates that the match between the k-mer found at positions x in the chromotype and y in the read can be recovered from the alignment. Middle and right, analogous dot plots for the read and the chromotype against themselves, showing that there is no large-scale self-similarity along either sequence.

Experiment 3: k-mer recovery from high-coverage samples

A key notion of the PRG is that it contains the majority of sequence known by the four methods, which are represented in the GRCh37 ALT haplotypes. The very modest gain in accuracy from the mapping step (<1%) likely reflects the very extensive characterization of genetic variation within classical HLA alleles.

Experiment 4: comparison to synthetic long-read Moleculo data

To assess accuracy over longer physical distances than k-mers, we analyzed alignments of synthetic long-read Moleculo data (25× coverage) from NA12878 to chromotypes generated by each approach (Online Methods). We first identified 29,429 reads (median read length of 3,165 bp; for convenience, we refer to the Moleculo sequences as ‘reads’, although they involve an assembly procedure) likely to have arisen from the extended MHC region through the presence of diagnostic k-mers (Online Methods). Read-to-chromotype alignment was performed with a Needleman-Wunsch–like algorithm that aligns gapped graphs instead of sequence, implemented using dynamic programming (Supplementary Note). We measured the scaled edit distance between reads and the chromotype (the number of non-identical characters in read–chromotype global alignment, including gap characters, divided by read length in k-mers) as an indicator of chromotype accuracy.

We found that the mapping approach (Platypus) achieved the highest number of read alignments with zero mismatches (11,338 versus 10,071 for PRG-Mapped). However, both PRG approaches resulted in substantially fewer reads with many mismatches and/or gaps (Fig. 6a and Supplementary Table 6). For example, the total number of alignment columns indicating a deletion in the chromotype decreased from 1,017,231 (Platypus) to 586,852 (PRG-Mapped). Likewise, the number of reads with very poor alignments (more than 150,000 gaps in the aligned read or ≥33% of the aligned chromotype string consisting of new gaps) decreased from 303 to 134. The modified chromotype (PRG-Mapped) had a modest benefit over the Viterbi chromotype (PRG-Viterbi), increasing the number of perfectly matched reads from 8,359 to 10,071. Across the HLA-DRB5 region (determined in the k-mer recovery analysis to be the region most poorly represented by the PRG), we found reads that suggested the presence of an inversion relative to known sequence (Fig. 6b).

DISCUSSION

Within a species (or even within an individual), the effects of mutation, recombination and selection can result in a great diversity of genomes, differing through events ranging from single-nucleotide changes to major rearrangements and gains or losses of sequence. Our hypothesis was that using information about known diversity would aid in the reconstruction of individual genomes from HTS data, particularly
within regions of high sequence and structural variation. To this end, we devised a graph structure for representing such reference variation, a method for using the structure to interrogate short-read HTS data to infer the diploid sequence of an individual and a series of benchmarking tests to evaluate accuracy in comparison to a standard mapping pipeline. By applying the approach to variation within the MHC region, we identified regions where genome inference is improved, sometimes substantially. This work demonstrates the feasibility and potential of using known variation in genome inference from HTS data and represents an important intermediary between mapping to a single reference and full de novo assembly. Our method has immediate application for researchers looking to understand the role of genetic variation within the MHC region for disease risk and drug response and also builds a framework for the analysis of complex variation more generally, not least in those regions with alternative assemblies in humans.

There are, however, many important choices concerning how to represent known variation, the set of variants to be included and how best to use such information in genome inference. These choices must be taken in the light of potential applications, ranging from microbial populations with highly mobile accessory genomes to common rearrangements in cancer. Below, we discuss the key considerations and how the approach described here could be extended or modified.

Choices about the structure and construction of a reference variation graph are intimately linked to its desired functions. Fundamentally, we see two functions of such a structure. First, it should provide a general and intuitive means of referring to variation, in a manner analogous to that of an rsID for SNPs and in a way that closely matches the true distribution of genomes. Our approach was to base the structure on an MSA of known material, allowing for recombination between sequences at aligned regions of identity. As such, the structure makes no attempt to explicitly model events such as duplication or rearrangement that lead to difficulties or ambiguities in alignment. For example, an inversion would be represented by a bubble in the same way that a region of high divergence would be represented. Similarly, the homology within a copy number–variable region would also not be recognized explicitly. To represent such events and the more complex rearrangements and amplifications observed in bacteria and cancer, alternative structures would need to be developed, such as the A-Bruijn or cactus graphs. However, whether such structures are well suited to the problem of inference from HTS data remains to be explored.

In constructing the PRG, we chose to include a wide catalog of information, including short variants, long haplotypes and lists of alleles at classical HLA loci. The comparison with the standard mapping approach suggests that, over much of the extended MHC region, the use of such information under the current implementation leads to little or no gain in accuracy. The choice about what material to include in a graph is a balance between wanting to most fully describe the space of genomic variation and the practical issue of building and using a graph that represents many small and/or rare events, whose inclusion is not necessary and potentially damaging to inference (for example, a duplication seen just once of an otherwise unique region). A pragmatic approach is to say that material should be included if, on average, it leads to better genome inference (here, for example, structural variation in the class II region and class II alleles). However, there are other possible advantages to including more sequence, for example, in reducing the heterogeneity in how complex variants or those in low-complexity regions are reported.

Perhaps the greatest limitation of the approach developed here is in terms of the inference algorithm. By summarizing HTS data as k-mers, we lose the longer-range information within a read and between read pairs. In addition, although the HMM for inferring the underlying reference chromosome is efficient, the approach of mapping reads separately to each of the arbitrarily phased haplotypes is ad hoc. Ideally, reads should be aligned directly to the graph structure, keeping track of the quality of mapping both within and across different levels in the graph. In principle, as demonstrated with the Moleculo data, graph mapping is feasible. However, there is a major challenge in making the process comparable in efficiency to algorithms for mapping to a linear reference. However, if direct mapping of reads to the graph can be achieved, the same HMM structure can be used for genome inference, although we note that the current structure is not well suited to analyzing polymorphic regions with extended identity where the ability to reconstruct the exact underlying sequence (as opposed to some summary, such as copy number) is limited. In theory, it would also be possible to use longer-range information about haplotype structure as a prior on paths through the PRG (such as is used in imputation and the refinement of low-coverage sequencing data).

Finally, we wished to know how unusual the class II MHC region is within the human genome in being poorly served by the paradigm of mapping to a single reference. To assess this, we calculated a genome-wide k-mer recovery map for the Platypus call set on NA12878 (provided as a track for the UCSC Genome Browser; Online Methods). We find that 1% of the human genome has low k-mer recovery (10-kb regions with <90% of predicted k-mers recovered; Supplementary Fig. 3), and these regions affect multiple genes and gene families (Supplementary Table 7). Although some of these regions may reflect large homozygous deletions with respect to the reference, these results suggest that there is an important minority of the genome where the identification and representation of alternative sequences would substantially improve genome inference.

**METHODS**

Methods and any associated references are available in the online version of the paper.

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

**ACKNOWLEDGMENTS**

We thank M. Eberle and colleagues at Illumina for early access to the Moleculo data. The study was funded by grants from GlaxoSmithKline and grant...
Online Methods

Algorithms. A full description of the PRG algorithms can be found in the Supplementary Note, including (i) the algorithms used to build PRGs from a set of reference data, (ii) the algorithmic and statistical methods for inferring a best diploid path (chromotype) through the PRG, (iii) the algorithm to discover new variation not present in the PRG and (iv) the graph-mapping algorithm used for the contig analysis.

Data for PRG construction. We defined the extended MHC region as the genomic region spanned by the PGF extended MHC haplotype (identical to the primary human reference in the region: in GRCh37 coordinates, chr6:28,702,185–33,451,429; GenBank ID for GRCh37 chromosome 6, CM00006681). In addition to the PGF extended MHC haplotype, we used seven extended MHC haplotypes from the MHC haplotype project3 (GRCh37, ALT_REF_LOCI_1–ALT_REF_LOCI_7). We created an MSA for the eight haplotypes using the programs FSA25 and MAFFT for refinement24. We used the SNPs identified by the 1000 Genomes Project, Phase 1 release 3, to augment the MHC haplotypes.

We also included all available aligned genomic (i.e., X_gen.txt files) HLA allele sequences from IMGT/HLA25 for the classical HLA alleles at the loci HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1 and HLA-DRB1 as additional scaffold haplotypes. These haplotypes cover all exons and introns of the genes. For many alleles, the genetic sequences are not completely specified over all exons and introns; however, the PRG construction algorithm removes most of the wildcard characters found at the unspecified positions.

The edge probability distributions at each vertex in the PRG are improper; specifically, we assign probability 1 to each edge. This is motivated by the downstream parts of our pipeline, which rely on the Viterbi algorithm for inferring maximum-likelihood personalized haplotypes. With the improper parameterization, each path through the model is equally likely under the Viterbi algorithm, irrespective of how many potential branching points (vertices where there is more than one possible edge to follow) it contains. We used k-mer length k = 31 for creating the k-mer PRG.

In the process of examining the eight extended MHC haplotypes, we discovered an inconsistency in the Ensembl database32. On the SSTO haplotype, HLA-DRB1 and HLA-DRB4 were mapped to the same start coordinate, which was likely caused, according to Ensembl, by a mismapping of exonic sequence of the two transcripts ENST00000549627 and ENST00000548105 (HLA-DRB4 and HLA-DRB1 exon sequence is similar). This issue was fixed in Ensembl release 73.

Whole-genome sequencing data. Subjects CS1 and CS2–CS6 were from four GlaxoSmithKline-sponsored clinical studies: EGF100151, EGF30008, CS1 data were initially aligned to GRCh37 (excluding the alternative loci) on the CLC Genomics Workbench (version 6.5.1), and coverage and intact and broken paired-end read numbers were determined for ~180 kb of sequence surrounding HLA-DRB1. This process was repeated with the addition of the MANN alternative MHC haplotype (identifier ‘ALT_REF_LOCI_4’, GenBank ID GL1000253.1). For all remaining analyses on CS2–CS6, reads were mapped to GRCh37 (excluding the alternative loci) using Stampy3 with BWA1, and variants were called using Platypus 0.1.8 (ref. 19).

Read data for NA12878 from the Illumina Platinum genomes project (HiSeq 2000, ~60× coverage, 100-bp paired-end reads) were obtained from the European Bioinformatics Institute. Reads were aligned to GRCh37 (excluding the alternative loci) using BWA 0.6.2 (ref. 1), and variants were called with Platypus 0.1.8 (ref. 19).

For Platypus 0.1.8-based variant calling in the MHC, we used the command line

```
python $[platypus_executable] callVariants --bamFiles=${bam_path} --output-t=${output_PATH} --bamFiles=${bamPath} --regions=${regions} --nCPU=12 --mergeClusteredVariants=1
```

with the variables substituted with their per-samples values.

For NA12878 mapping with BWA 0.6.2 (and for the remapping step), we used BWA-backtrack (aln/sampe) with parameter --q10 for aln (all other parameters standard values).

For Stampy alignment, we used (in addition to file input and output parameters) the following command line options:

```
--bwoptions=-t 2 -q10 /tmp/hs37d5 --keeprorder -v0 --solexa
```

Simulations. Genomes were simulated from the PRG by independently sampling two paths with uniform choices at junctions. We concatenated the edge labels induced by each path, removed ‘gap’ characters and used the strings thus generated as a sample’s two haplotypes from which to generate reads. The number of starting reads (read length, 101 bp) at each position was Poisson distributed with mean such that the average depth was 30×. Accuracy was assessed by comparing the true underlying genotype at each level of the PRG with the genotype inferred from the Viterbi path. Specifically, we used the scoring system shown in Supplementary Table 8 to measure the number of correct alleles.

Allele concordance was the sum over sites of the score obtained divided by the maximum possible score (i.e., two times the number of sites analyzed). The same table was used to measure ‘accuracy’ in the empirical data analysis. In the ‘reads with error’ case, we used an empirical error model based on the PCR-free data from NA12878, which achieved an average per-base error rate of ~0.1%. Our simulations were limited in that we treated the simulated paths as a sample’s complete genome; that is, we did not include additional variation.

Validation data. SNP arrays. Individual SNP array data were provided by GlaxoSmithKline for samples CS1–CS6 (Illumina 1M array). We used publically available Illumina Omni 2.5M SNP array data from the 1000 Genomes Project for NA12878.

HLA genotypes. Individual HLA genotypes (reported to four-digit accuracy using ‘g’ nomenclature) are given in Supplementary Table 9.

k-mer recovery from short read data. See the Supplementary Note for details of how k-mer recovery was estimated.

Synthetic long-read Moleculo data. For the Molecule-based validation, identified contigs likely to have originated from the extended MHC region were identified using the following strategy:

1. We computed the set of all k-mers (k = 31) occurring in the k-merified extended MHC PRG. We called all k-mers occurring in this set ‘extended MHC k-mers’.

2. We computed the set of all k-mers (k = 31) occurring in the human reference genome, excluding the region covered by the extended MHC PRG. We called all k-mers in this set ‘reference k-mers’. Note that some k-mers were both extended MHC k-mers and reference k-mers. We called k-mers that were extended MHC k-mers but not reference k-mers ‘extended MHC–unique k-mers’.

3. We filtered Moleculo reads according to the following criteria:
   a. Fraction of extended MHC k-mers ≥ 0.8.
   b. ≥2 extended MHC-unique k-mers spanning a stretch of at least 50 bases (in between the two k-mers). For each read, we select the maximum stretch MAXSTRETCH spanned by two such extended MHC–unique k-mers.
   c. Within MAXSTRETCH, fraction of extended MHC–unique k-mers ≥ 0.5.
d. Within MAXSTRETCH, fraction of reference k-mers ≤ 0.3.

e. If a read passed these tests, we truncated the read to MAXSTRETCH and aligned it to the PRG.

Runtime. Most algorithms are multithreaded (openMP); hence, total effective runtime depends on local system configuration. We give example runtimes for generating the VCFs for NA12878 on a multicore machine; VCF generation (PRG-Viterbi) takes 1.8 h of wall time (6.6 h of CPU time), whereas VCF generation (PRG-Mapped) takes ~5 h of wall time (5 h of CPU time). Note that this does not include the actual whole-genome remapping process (2×), which is typically carried out on a cluster.

Genome-wide analysis of k-mer recovery in NA12878. Genome-wide k-mer recovery from the Platypus VCF was measured as for extended MHC–specific k-mer recovery, with the exception that we counted k-mers that contained undefined characters (N’s) as recovered, whereas we counted them as absent for extended MHC validation. We provide a wiggle plot with results from the genome-wide k-mer recovery analysis (200-bp bins) (see URLs) that can be used within the UCSC Genome Browser.

31. Holdsworth, R. et al. The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. Tissue Antigens 73, 95–170 (2009).
32. Flicek, P. et al. Ensembl 2013. Nucleic Acids Res. 41, D48–D55 (2013).
33. Spraggs, C.F., Parham, L.R., Hunt, C.M. & Dollery, C.T. Lapatinib-induced liver injury characterized by class II HLA and Gilbert’s syndrome genotypes. Clin. Pharmacol. Ther. 91, 647–652 (2012).