Defining *KIR* and *HLA class I* Genotypes at Highest Resolution Using High-Throughput Sequencing

Paul J. Norman*,[1] Jill A. Hollenbach,[2] Neda Nemat-Gorgani,[1] Wesley M. Marin,[2] Steven J. Norberg,[3] Elham Ashouri,[1] Jyothi Jayaraman,[4] Emily E. Wroblewski,[1] John Trowsdale,[4] Raja Rajalingam,[5] Jorge R. Oksenberg,[2] Jacques Chiaroni,[6] Lisbeth A. Guethlein,[1] James A. Traherne,[4] Mostafa Ronaghi,[3] and Peter Parham.[1]

1. Departments of Structural Biology and Microbiology & Immunology, Stanford University School of Medicine, Stanford, CA94305, USA.
2. Department of Neurology, University of California San Francisco School of Medicine, San Francisco, CA94158, USA.
3. Illumina Inc., 5200 Illumina Way, San Diego, CA92122, USA.
4. Division of Immunology, Department of Pathology and Cambridge Institute for Medical Research, University of Cambridge, CB2 1QP, UK.
5. UCSF Immunogenetics and Transplantation Laboratory, San Francisco, CA 94143, USA.
6. UMR 7268 ADÉS, Aix-Marseille Université-EFS-CNRS, Marseille, France.

* Correspondence paul.norman@stanford.edu

Keywords: KIR, HLA class I, highly polymorphic, gene content diversity, immunity

Abstract

The physiological functions of natural killer (NK) cells in human immunity and reproduction depend upon diverse interactions between killer cell immunoglobulin-like receptors (KIR) and their HLA class I ligands: HLA-A, -B and -C. The genomic regions containing the *KIR* and *HLA class I* genes are unlinked, structurally complex, and highly polymorphic. They are also strongly associated with a wide spectrum of disease, including infections, autoimmunities, cancers, and pregnancy disorders, as well as the efficacy of transplantation and other immunotherapies. To facilitate study of these extraordinary genes, we developed a method that captures, sequences and analyzes the 13 *KIR* genes and the *HLA-A*, *-B* and *-C* genes from genomic DNA. We also devised a bioinformatics pipeline that attributes sequencing reads to specific *KIR* genes, determines copy number by read depth, and calls high-resolution genotypes for each of the *KIR* genes. The method was validated using DNA from well-characterized cell lines, by comparison with established methods of *HLA* and *KIR* genotyping and by determining *KIR* genotypes from the 1000 Genomes Project sequence data. This identified 116 novel *KIR* alleles, which were all demonstrated authentic by sequencing them from source DNA using standard methods. Analysis of just two *KIR*, showed that 22% of the 1000 Genomes individuals have a novel allele or a structural variant. The method we describe is suited to the large-scale analyses that are needed for characterizing human populations and defining the precise *HLA* and *KIR* factors associated with disease. The methods are applicable to other highly polymorphic genes.

Introduction

The human leukocyte antigen (*HLA*) complex of chromosome 6 is the most polymorphic region of the human genome.[1] This variation is driven by pressure to resist diverse pathogens, but also underlies susceptibility to autoimmunity and other inflammatory diseases of major importance to human health.[2] HLA class I molecules are expressed on the surface of most tissue cells, where they interact with receptors on the surface of lymphocytes, effector cells of the immune system.[3] Natural killer (NK) cells are innate and adaptive lymphocytes that destroy infected or tumor cells having aberrant expression of HLA class I; they also regulate trophoblast invasion during early pregnancy.[4] NK cell activity is genetically modulated through differential expression of polymorphic killer cell immunoglobulin-like receptors (KIR) that recognize HLA class I molecules.[5] Only recently has the *KIR* genomic region been characterized to high resolution.[6] Consequently, re-examination of diseases having long established associations with specific *HLA* polymorphisms is revealing a strong and collective influence from *KIR* polymorphism.[7-10]

The *KIR* locus of chromosome 19q13.4 is characterized by an unusually high diversity, both in the numbers of genes and their alleles.[11] The region varies in size from 100kbp to 350kbp due to structurally diverse haplotypes with duplicated segments, large deletions and gene fusions.[12; 13] As a consequence of this plasticity, the 13 distinct *KIR* genes (MIM: 604936-7, 604945-7, 604952-6, 605305, 610095, 610604) are combined in numerous ways. Haplotypes have between four and 20 *KIR* genes, with the most common *KIR* region haplotype having seven genes.[14] To varying degrees each *KIR* gene is polymorphic, with more than 600 *KIR* alleles currently defined.[15] *KIR* and *HLA class I* (*HLA-A*, *-B* and *-C:* MIM: 142800, 142830 and 142840) polymorphism are actively co-evolving,[16] suggesting that many more *KIR* alleles and haplotypes await discovery. During the last three decades over 10,000 *HLA class I* alleles have been characterized in specialized clinical HLA laboratories[15] and similar intensive study will be needed to characterize *KIR* diversity.

The function of an HLA class I molecule is to bind a peptide, usually a nonamer, inside a cell and take it to the cell surface, where the complex of peptide and HLA class I is engaged by KIR and other lymphocyte receptors.[17; 18] On healthy cells, the peptides bound by HLA class I molecules derive from normal human proteins and do not stimulate an immune response. On infected or transformed cell surfaces, pathogen-specific or tumor-specific peptides are bound to HLA class I, and gross changes in the surface level of HLA class I can be induced. All such differences activate lymphocytes and the immune response.[4; 19] For the interactions of KIR with HLA class I to be effective they have to respond to a wide diversity of tumors and pathogens, many of which are rapidly evolving.[20] This has been achieved by having a diversity of interactions within each individual, and differences in those interactions from one individual to another. The latter provides barriers that can impede the spread of infection within families, communities and populations.

Crucial features that distinguish *KIR* and *HLA* alleles from those of most other genes are the depth, breadth and functional importance of their sequence divergence. Thus, alleles can differ by multiple nucleotide substitutions, and three or four alternative nucleotides are present at functionally critical positions. *KIR* and *HLA* alleles segregate as constituents of distinct lineages, which are further diversified by intra-genic as well as inter-genic recombination.[13; 21] In turn, these lineages are maintained in all human populations and both genomic regions exhibit clear evidence for the impact of balancing selection.[22; 23] Moreover, the strong, highly reproducible signals of natural selection observed for the *HLA class I* and *KIR* regions point to their genomic variation being critical for human survival.[24; 25]

The development of methods to assess the nature and extent of *KIR* genomic diversity has been limited by the complexity of the region. The widely used methods that exist for typing *KIR* focus principally on gene content.[12; 26-28] In contrast, the methods being used to determine allelic variation are costly, time-consuming[6; 16; 29] and unsuitable for high-throughput studies. The results of the few allele-level population studies of *KIR*[16; 29-32] however, show that such investigation is likely to be informative. For example, some *KIR* are restricted to population groups of specific geographic ancestry.[30; 31] Other *KIR* have lost expression, but appear common and widely distributed.[29; 32] To extend such studies to other populations, as well as disease cohorts, we have developed a sequencing and bioinformatics method that determines complete *KIR* and *HLA class I* genomic diversity.

Material and Methods

Overview

To target *KIR* and *HLA class I* genes for next generation nucleotide sequencing (NGS), we designed sets of specific oligonucleotide probes to capture the *KIR* region (140-240kbp) and the individual *HLA-A*, *-B* and *-C* genes (each ~3kbp) from libraries prepared from sheared genomic DNA. We then developed a bioinformatics pipeline (PING: Pushing Immunogenetics to the Next Generation) specifically to convert sequence data obtained from the highly polymorphic *KIR* genes into high-resolution genotypes. A summary of the pipeline is shown in Figure 1A. PING first sorts the sequence reads to isolate those that represent fragments from the *KIR* genomic region from those that do not (a process termed filtering). The final *KIR* genotypes are obtained from these filtered reads, using a composite of two core modules that describe the gene and allele content for each individual and also return information on newly-identified SNPs and recombinant alleles. The first module (PING_gc), which determines the *KIR* gene copy-number, is used to inform the second module (PING_allele) that generates allele data (Figure 1A and Figure S1). Each module is split into two sub modules. KIR Filter Fish (KFF), which is used in both main modules, probes the *KIR* sequence data with specific sequence search strings and determines which genes (KFFgc) or alleles (KFFallele) are present. The function served by KFF is equivalent to genotyping with sequence-specific oligonucleotide probes (SSOP).[33] To complement KFF, MIRAgc (based around the program MIRA)[34] and Son of Samtools (SOS; based around Samtools)[35] create alignments to reference sequences in order to determine the gene and allele content, respectively. The output is designed to comply with the genotype list (GL string) format that is used for reporting *HLA* and *KIR* data by clinical transplantation laboratories.[36] We validated the typing obtained from the complete capture/NGS and bioinformatics method by using standard molecular techniques, and further tested the bioinformatics component using existing data sets from whole-genome sequencing experiments. A summary of the data generated or otherwise obtained is shown in Figure 1B. *KIR* and *HLA class I* allele sequences used for probe design and as reference data were obtained from the ImmunoPolymorphism Database (IPD; see Web Resources).[15] Throughout this paper any unique DNA sequence that spans a coding region (coding DNA sequence: CDS) is considered a distinct allele. An explanation of KIR and HLA nomenclature is given in Appendix A.

Human Subjects and Data

Ethical approval for this study was obtained from the Stanford University Administrative Panels on Laboratory Care and Human Subjects in Medical Research and The Committee on Human Research at the University of California, San Francisco. Written informed consent was obtained from all individuals.

To develop and validate the complete capture/NGS and bioinformatics method we generated data from three sources of human genomic DNA:

1. A panel of IHWG lymphoblastoid B cell lines.

Genomic DNA was extracted from 97 International Histocompatibility Working Group (IHWG) cell lines. These cells have been used extensively in developing methods for genotyping polymorphic loci including *KIR* and *HLA.*[37-41] Most of the cell lines (93%) are homozygous for the *HLA-A, -B,* and *-C* genes.[37-41] A substantial majority of the IHWG cells (80%) are derived from donors of European origin and represent many of the common *HLA* alleles.[42; 43] Also studied was genomic DNA from a chimpanzee B cell line, derived from Clint[44] (Yerkes pedigree number C0471), a chimpanzee of the *Pan troglodytes verus* (western chimpanzee) subspecies, and subject of the Chimpanzee Genome sequence study.[44]

2. West African Trios

Genomic DNA samples from 30 family trios (both of the parents and one child) from Mali in West Africa were analyzed.[45]

3. European control samples

De-identified DNA samples from 188 unrelated healthy individuals of European origin, with no history of chronic disease, were studied. These samples were selected at random from a larger data set (n=500) developed as controls for genome wide association studies of multiple sclerosis (MS [MIM: 126200]).[46] These samples were used because their high-resolution *HLA-A*, *-B*, and *-C* genotypes had been determined using Sanger sequencing.[46]

4-5. Analysis of existing NGS data

To validate the PING pipeline, we used existing sequence read data from an additional two sources, described below under 4 and 5. To extract *KIR* specific sequences from these data sets, read-pairs that mapped within the *KIR* region (hg19 coordinates: 19:55,228,188-55,383,188) or an unallocated region of chromosome 19 (GL000209.1) that corresponds to an alternative *KIR* haplotype were identified using SAMtools 0.1.18.[35]

4. Fifteen KhoeSan individuals,[47] who had also been *KIR* genotyped using standard lower throughput methods of pyrosequencing and Sanger sequencing.[16]

5. The 1000 genomes project data.[48] All 2,532 of the whole-exome sequenced individuals described in the May 2013 release were targeted.[48] To ensure sufficient quantity of sequence reads for the analysis, samples were excluded if there were fewer than 25 reads that map to exon 3 of *KIR3DL2* or *KIR3DL1/2v* (See Appendix A: KIR and HLA nomenclature). There were 420 of the 1000 genomes samples excluded on this basis and the remaining 2,112 were genotyped. When novel *KIR* alleles were identified in the 1000 genomes data set, genomic DNA from the source samples was purchased from the Coriell repository for confirmation of their sequence using standard molecular methods.

Laboratory Methods

Design of capture oligonucleotide probes

To account for gene content variation of the *KIR* region, we targeted a panel of independently generated reference *KIR* haplotypes,[6; 14] that together represent all of the 13 recognized *KIR* genes. First we designed probes against the two complete *KIR* haplotypes (FP089703 and FP089704) that were generated from the PGF cell line, which was the source of the human reference sequence for the *KIR* region.[6] We used end-to-end tiling, with strand-swapping, to design non-overlapping 80-mer probes to match these reference sequences. We then designed a similar set of probes using a further 27 complete *KIR* haplotype sequences[6; 14] and all *KIR* sequences included in the January 2013 release of the IPD/KIR database.[15] In this second stage, probes that differed by more than three nucleotides from the corresponding segment of the initial reference haplotypes were selected for use. We did not mask any repetitive elements in the target haplotypes. The *KIR* genomic region targeted by the probes is equivalent to that covered by (hg19) chr19:55,228,188-55,383,188 and chr19: unmapped GL000209.1, which are the two *KIR* haplotypes present in the hg19 reference genome. In a similar manner to the *KIR* probes, we designed probes against the alleles of the classical *HLA class I* genes present in PGF, which was also the source of the human reference sequence for the *HLA* region.[1; 14] These probes were supplemented with probes designed against the 6,795 *HLA class I* sequences reported in the January 2013 release of the IPD/HLA database.[15] A total of 10,456 capture probes were used.

Preparation of biotinylated capture probes.

The set of capture oligonucleotides, each one comprising a unique sequence flanked by the common sequences, GGTGATTGCGTATCT (PTL3) and CATGTCGTGGGAATT (PTR3), was synthesized by CustomArray (Bothell, WA). This set of oligonucleotides was pooled and amplified in a single PCR, using primers with sequences corresponding to PTL3 and PTR3. The PCR comprised 1x Titanium Taq buffer (Clontech, Mountain View, CA), 1uM each of biotin-PTL3 and PTR3 primers (Integrated DNA Technologies (IDT), Coralville, IA), 0.2 uM dNTPs with 12.5% dUTP (Roche, Indianapolis, IN), 1ul (1 unit) Uracil-DNA Glycosylase (UDG; New England Biolabs), 1M betaine, 3ul (15 units) Amplitherm polymerase (Epicentre, Madison, WI), 0.2 ng of the pool of capture oligonucleotides and H_2O added to a final volume of 100ul. PCR cycling conditions were as follows: 37^{o}C (10 min), 95 ^{o}C (3 min), (95^{o}C (30 s), 55^{o}C (30 s), 72^{o}C (30 s) x 28), 72^{o}C (10 min.), and 10^{o}C (hold).

The biotinylated PCR product (100ul aliquot) was then bound to streptavidin-coated magnetic beads (Illumina, Inc) that had been pre-washed with 100ul 6x Hybridization Buffer (HB:1 M NaCl, 0.5M phosphate buffer, 0.05% Tween-20) and suspended in 100ul 12x HB. The incubation was carried out for 30 min at room temperature in HB and with agitation. The beads, now coated with biotinylated oligonucleotides, were then washed: once with 100ul 6x HB, twice with 100ul 0.2x HB, once with 100ul 0.1nM NaOH and 100ul 10 mM EDTA, and, lastly, once with 100ul 0.2x HB. Biotinylated oligonucleotides were eluted from the beads using 0.1 mM EDTA, then concentrated using speed vacuum to a final concentration of 2.5nM for each capture probe.

Library preparation, enrichment and sequencing

The protocol was based on the Truseq Nano method for library preparation (Illumina Inc., San Diego, CA). The DNA samples we used are described below. For each sample, 300ng genomic DNA (as determined by Qubit instrument: ThermoFisher, Sunnyvale, CA) were sheared into 800bp fragments using a Covaris S220 instrument (Covaris, Woburn, MA). The library preparation was then performed according to the manufacturer's instructions, with 96 unique 'dual index' combinations used individually to label the library obtained from each DNA sample, and the following modifications: 1. To clean/size-select the samples following end-repair, 70.2ul sample purification beads plus 89.8ul H_20 were added to 100ul sample. 2. In the final PCR the 72°C extension time was changed from 30s to 90s to account for the 800bp fragment length.

Enrichment for *HLA/KIR* region sequences was performed using a modified version of the Nextera Rapid Capture Exome enrichment protocol (Illumina Inc., San Diego, CA), a solution-based target capture assay. The libraries of genomic DNA, indexed uniquely for each sample as described above, were pooled prior to their hybridization with the capture probes. Thus, each hybridization mix (100ul) contained 96 uniquely indexed sequence libraries (62.5ng for each library: 6000ng in total), 50 pM of each biotinylated capture probe and hybridization buffer (CT3, and all subsequent buffers from Illumina Inc., San Diego, CA). The hybridization mix was incubated at 95^{o}C for 10 mins, gradually cooled by 2^{o}C per minute to 58^{o}C then maintained at 58^{o}C for 90 mins. In this reaction, fragments of genomic DNA that contained targeted *KIR* and *HLA* sequences became specifically hybridized to biotinylated capture probes.

In the next reaction, 100ul of streptavidin-coated magnetic beads were used to separate the specific hybridized genomic DNA away from the non-specific un-hybridized genomic DNA. The biotin present in hybrid DNA molecules was bound to streptavidin on the beads, leaving the non-specific DNA in solution. The DNA preparation enriched for the targeted *KIR* and *HLA* genes was then eluted from the beads. Binding of the hybridization product to the beads was achieved by 30 min incubation with agitation at 1000rpm on a plate shaker at room temperature. To clean the product, the streptavidin beads were removed from solution using a magnetic separator, mixed with 200ul Enrichment Wash Solution (Illumina Inc., San Diego, CA), and incubated at 50^{o}C for 30 minutes. This wash step was repeated. To elute the enriched DNA from the beads we added 23ul of elution mix (made from 1.5ul 2M NaOH plus 28ul Elution Buffer 1), incubated for 5 minutes at room temperature and neutralized with 4ul Elute Target Buffer 2. The eluted material was then subject to a second round of enrichment from the hybridization step onwards. After the gradual cooling step, the hybridization mix was maintained at 58^{o}C for 14-18 hours.

An aliquot of 10ul of the enriched DNA preparation was subject to PCR amplification in a 50ul reaction mix containing 5ul of PCR Primer Cocktail and 20ul of Nextera Enrichment Amplification Mix PCR cycling was performed as follows: 98°C for 30s, 17 cycles of 98°C for 10s, 60°C for 30s, and 72°C for 30s with a final elongation step at 72°C for 5 min. Amplified material was purified with 40ul of Sample Purification Beads and eluted in 30ul of RSB (Illumina Inc. San Diego, CA).

NGS strategies

Set 1. (IHWG cell lines) The enriched libraries were sequenced using a HiSeq2000 instrument and sequencing chemistry (Illumina Inc., San Diego, CA). Samples were clustered and paired-end sequencing performed with the TruSeq SBSv3-HS kit (Illumina Inc., San Diego, CA). The sequencing read length was 2 x 101bp.

Set 2. (Trios and chimpanzee). The enriched libraries obtained from these samples were sequenced using a HiSeq2500 instrument and sequencing chemistry (Illumina Inc., San Diego, CA). The sequencing read length was 2 x 250bp. These samples were also genotyped for *HLA-A*, *-B* and *-C* using SSOP, and *Lineage II KIR* using pyrosequencing[23] (See Appendix A: KIR and HLA nomenclature).

Set 3. (European controls) These samples were analyzed using a MiSeq instrument (Illumina Inc., San Diego, CA) with V3 chemistry and the sequencing read length was 2 x 300bp.

Enrichment efficiency

Enrichment efficiency was estimated by mapping unprocessed sequence reads to the human reference (hg19) using BWA-MEM[49] (with k=16) and counting the number of reads that map within the target coordinates of hg19.

Validation of the results from capture/NGS using established methods

Previously described protocols were used for pyrosequencing[16] and real-time PCR.[12] Standard Sanger DNA sequencing reactions were performed in forward and reverse directions using BigDye Terminator v3.1 and analyzed using an ABI-3730 sequencer (ABI, Foster City CA). To verify the sequences of novel alleles, PCR products were cloned using the pCR2.1-TOPO vector (Invitrogen, Carlsbad, CA) and sequenced using M13 and internal primers. For each individual in whom an allele was identified for the first time, five or more clones corresponding to the allele were sequenced. HLA class I genotyping was performed using Luminex bead-based SSOP hybridization as described previously[16] except where indicated otherwise.

Bioinformatics methods: **Pushing Immunogenetics to the Next Generation (PING) pipeline**

Harvesting KIR specific reads

For sequence data obtained by the capture/NGS sequencing method, sequence reads specific to the *KIR* region were identified and harvested using Bowtie 2.[50] The 29 sequenced *KIR* haplotypes and all *KIR* alleles from the IPD/KIR database[6; 14; 15] were concatenated to create a single reference file for this purpose. The equivalent of 70,000 reads (at 2 x 300 bp) that passed this filter stage were taken per sample for *KIR* genotyping.

KIR gene-content module (PING_gc)

This module is divided into two complementary components, the first (KFFgc) based on string searches of raw data and the second (MIRAgc) on read-depth following alignment to reference sequences.

1. *KIR gene-content determination by virtual probes (KFFgc)*

The sequence read files specific to the *KIR* region were first processed using the ‘fastq quality trimmer’ function of Fastx Toolkit 0.0.13 (see Web Resources). From an alignment of all human *KIR* sequences (IPD/KIR Release 2.6.1, 17 February 2015),[15] all possible 32bp sequences specific to just one of the *KIR* genes were generated, and those sequences covering polymorphic positions removed. For each *KIR* gene, this process produced a set of probes that is gene-specific but not allele-specific. For each gene, ten such probes were selected at random and used for determining *KIR* gene-content. We then counted the total number of exact matches to each probe sequence, and its reverse complement, present in the forward and reverse (read1.fastq and read2.fastq) sequence read files for each sample. As every *KIR* haplotype has one copy of *KIR3DL3*,[6; 14; 16] the presence/absence of other *KIR* genes was determined from the mean number of probe-hits per target *KIR* divided by the mean number of *KIR3DL3* probe-hits on *KIR3DL3*. We used a ratio of 0.2 as the threshold, and values above this were considered positive.

2. *KIR gene copy number determination by analysis of read depth (MIRAgc)*

As shown previously, the depth of reads aligned to a reference can be used to estimate structural variation on a genomic scale.[51; 52] We incorporated this concept into the *KIR* genotyping pipeline. Using MIRA 4.0.2[34], the *KIR* region specific sequence read pairs were aligned simultaneously against one reference sequence for each *KIR* gene. Using values extracted from the ‘contigstats’ results table from the MIRA output, the number of reads that align to each *KIR* gene were divided by the number of reads aligning to *KIR3DL3.* This comparison produced a clustering of read ratios that correlates with differences in *KIR* gene content (Figure 2 and Figure S2). For example, for *KIR2DL4*, which is absent from some haplotypes and duplicated on others,[13; 53] we observed three clusters of read ratios, corresponding to one, two or three copies of *KIR2DL4* (Figure 2). For *KIR2DL5*, which can be present at centromeric and telomeric locations in the *KIR* haplotype,[6; 54] we observed five clusters of read ratios, corresponding to individuals having zero, one, two, three or four copies of *KIR2DL5* (Figure 2). Representative examples of clustering results derived for all the *KIR* are given in Figure S2.

In very rare cases there may be copy number variation of *KIR3DL3*. For example, an individual with duplicated *KIR3DL3* on one haplotype was observed in a study of >3,500 subjects.[12] Such individuals would then show an unusual copy number for all of the *KIR* genes, and hence would be identified for manual inspection. The individual identified with a duplicated *KIR3DL3* had a normal genotype of two *KIR3DL2* copies,[12] so in this case *KIR3DL2* could be used as a substitute standard.

KIR allele genotyping module (PING_allele).

This module is also divided into two complementary components, the first (KFFallele) based on string searches of raw data and the second (SOS) on read-depth following alignment to reference sequences. A summary of the KFFallele workflow is shown in Figure S1A, and SOS in Figure S1B and Figure S1C.

1. High-resolution KIR genotyping with virtual probes (KFFallele)

For each *KIR* gene, an alignment of coding region alleles was generated. From these, every possible unique 32-mer sequence that did not overlap an exon boundary was generated and the resulting pool screened to remove all 32-mers that are present in another *KIR* gene. A 'hit table' was generated using the original allele alignment and bespoke R scripting (allgenos_hit_table_2.R). The hit table was passed on to a random forest algorithm using the 'RandomForest' package of the R computer program suite.[55] This algorithm ranks the probes according to the amount of information they contribute to the final answer. The purpose of the random-forest step is to obtain the most efficient subset of probes, thereby increasing the computational speed of genotype assignment. The output consists of a series of probe subsets that have an increasing proportion of the total set (5%, 10%, 15% …etc.). Empirical testing found the most efficient subset is usually the 40% most informative probes. The final probe subset was applied to the original allele alignment, to produce a hit table of the expected probe pattern for all possible genotypes (again using allgenos_hit_table_2.R). The number of exact matches to each probe sequence, or its reverse complement present in the forward and reverse files (read1.fastq and read2.fastq) for each sample, was then counted. Aggregate counts above a threshold of 10 were considered positive, and the results compared with the genotype 'hit table' using a bespoke R script (KFFsums.R) to assign the *KIR* allele genotype.

2. KIR sequence alignment genotype by SOS

Data used to generate filters and reference sequences were obtained from the IPD/KIR database (Release 2.6.1, 17 February 2015) and the set of 29 complete *KIR* haplotype sequences.[6; 14; 15] For each given *KIR* gene, all available allele sequences were selected for use as a positive filter and all allele sequences of all other *KIR* genes were used to form a negative filter. Sequence reads specific to the given *KIR* were selected by mapping them to the positive filter, using Bowtie 2.[50] This mapping was performed non-stringently (>=97% nucleotides match) to allow for detection of unknown SNPs. All reads that aligned to the positive filter were retained, and those that did not were excluded. The retained reads were mapped stringently (>=99% match) to the negative filter, and in this case those that aligned were excluded. Finally, the selected reads were aligned to a single reference sequence, chosen for each *KIR* (Figure S3), and the SNP variants ascertained using SAMtools/bcf version 1.2.[35; 50]

The resulting variant call files (vcf) were analyzed using a custom R script (jSOS) to generate genotypes based on the known combinations of SNPs (i.e. the unique *KIR* alleles available from IPD). A post-filtered and aligned read depth of 20 was used as the minimum for calling the genotype at any given nucleotide position. The jSOS algorithm determines all of the possible allele combinations based on the genotypes of those SNPs that achieve the threshold read depth. Thus, if a specific SNP fails to reach the threshold, the ambiguity of the final allele call will increase, but an incorrect allele-level genotype will not be returned. When the combination of SNPs does not correspond to a known pair of alleles, jSOS identifies that a novel combination of known SNPs is present. In these situations, the known allele most likely present is identified as well as the new combination (we view this as the most parsimonious genotype, and the least parsimonious could be two novel combinations present). The jSOS algorithm also identifies novel SNPs. When manual confirmation was sought, such as these cases of novel polymorphism, alignments for visual inspection were created using MIRA 4.0.2, and examined using Gap4 of the Staden package.[56]

HLA class I genotypes

The *HLA class I* allele compositions were determined using NGSengine 1.7.0 (GenDX software, Utrecht, NL), kindly provided by Wietse Mulder and Erik Rozemuller, with the 'IMGT 3.18.0 combined' reference set. There was no pre-filtering for *HLA* genes and the data was analyzed directly with the software. The one exception was the removal of reads mapping to *HLA-Y*, an *HLA class I* pseudogene present on a subset of *HLA* haplotypes.[57] Presence or absence of *HLA-Y* was determined using sequence specific string searches of the fastq data. For the IHWG cells these assignments agreed with those previously determined using PCR.[43] The assignment of *HLA class I* alleles was further validated by analyzing the sequence data with Assign MPS 1.0 (Conexio Genomics, Fremantle Australia), kindly provided by Damian Goodridge. Any discrepancies between the two methods, or with previously obtained results, were resolved by designing virtual probes that distinguish the two possibilities in question and searching the unprocessed sequence read data (as described for KFF). Then sequence reads specific to the locus under question were extracted and aligned to reference sequences (as described for SOS) and inspected manually. Here, the reference sequences used were chosen based on the *HLA class I* genotype of the respective sample. The majority of *HLA class I* alleles are not fully characterized through all exons and introns (>95%)[15]. Thus for some of the validations, the second or third fields of resolution were compared.

Haplotype assignments

Haplotypes derived from homozygous cell lines were unambiguous. Haplotypes were assigned for the family trios by segregation. For all other individuals, centromeric and telomeric allele-level

KIR haplotypes were assigned for each individual using the expectation–maximization (EM) algorithm of haplo.stats implemented in the R programming language (see Web Resources).

Results

Validation of the *KIR* capture method

We applied our capture/NGS method to DNA extracted from 97 publically available cell lines (sample Set 1; Material and Methods), originally collected by the IHWG to facilitate study of *HLA* genes.[40] PGF is one of these cell lines having both *KIR* haplotypes characterized previously by standard methods.[6; 14] Data we obtained from PGF were therefore used to assess the quality of *KIR* sequences produced. The PGF *KIR* sequence reads were mapped back onto the two conventionally determined haplotype sequences. For PGF *KIR* haplotype 1 (FP089703: 137,813bp), our method gave 100% coverage, a mean read depth of 49.4x and a read depth of >10x for 99.5% of the haplotype (Figure 3A-B). For PGF *KIR* haplotype 2 (FP089704: 142,732bp), we obtained 99.99% coverage, a mean read depth of 49.8x and a read depth >10x for 99.5% of the haplotype. There are two short gaps in our haplotype 2 sequence, comprising 18 nucleotides in total (Figure 3C). The missing sequences are not predicted to be of functional importance (Figure 3C). Thus our method gives full coverage of the *KIR* region when targeting haplotypes that were among those included in the probe design. However, for population studies it is important that our probe sets can cope with a wider (and unknown) range of diversity, with no 'allelic dropout'. As a test, we applied our method to genomic DNA from the subject of the chimpanzee genome project.[44] We mapped the reads obtained to the two full *KIR* haplotype sequences (BX842589 and AC155174) that were previously characterized from this individual.[25; 58] These sequences represent both haplotypes of the *KIR* region and include ten of the fourteen chimpanzee *KIR* genes. We obtained 98.8% coverage for both of the haplotypes, with mean read depth of 130x for haplotype 1 and 110x haplotype 2. This success in capturing the chimpanzee *KIR* region strongly indicates that our method captures the full range of human *KIR* haplotypes.

Efficiency of the *KIR* capture method

To estimate the efficiency of the capture process we mapped all sequence reads generated for each individual back to the human genome, and counted those that fell within the target coordinates. Using this measure, the mean enrichment efficiency of the optimized 2 x 300bp sequencing runs (sample Set 3) was 87.01% (sd 5.01). Because the target region represents <0.01% of the human genome, this represents a significant (10,000 x) reduction in sequencing capacity required to analyze the target, compared with whole genome sequencing.

Specificity of sequence read harvesting

To begin the bioinformatics analysis of *KIR* we use a panel reference haplotype sequences as filters to harvest from the main pool of sequenced fragments, any sequence reads that could map to the *KIR* region (described in Material and Methods). As both the capture probes and the reference sequences for harvesting reads were designed using complete *KIR* haplotypes that did not have repetitive elements masked, we performed a further test for specificity on the harvested *KIR* reads. By analyzing 70,000 of these read pairs per individual, we showed a mean of three read pairs (modal value = 0) could map outside the target region of human genome build hg19. Thus, the combination of capture/NGS method and *KIR* sequence read harvesting is highly-specific. We also note that when we generated 2 X 100 bp sequence reads (instead of 2 X 300), up to 12% of the harvested reads potentially originate from repetitive elements outside the *KIR* region. However, 100% of these reads map to a 1.8kb LINE insertion that is located in intron 6 of *KIR3DL2*[13] (Figure S4) and does not overlap with any known control elements. Thus these reads do not affect the subsequent analysis.

Measurement of *KIR* gene copy number (PING_gc)

The PING_gc component of PING specifically determines gene copy number. *KIR3DL3*, a single-copy gene common to all *KIR* haplotypes[6; 14; 16] is used as the standard to which other *KIR* genes are compared (Material and Methods). To assess the correlation between read ratio and *KIR* gene content, we applied PING_gc to the sequence data generated from the 97 IHWG samples. In using the first module of PING_gc (called KFFgc) we observed identical results to those obtained by established methods,[26; 59; 60] producing thirteen distinct *KIR* gene presence/absence genotypes (Figure 4A). We then applied the second module of PING, MIRAgc, and used the observed clustering (Figure 2) to set threshold values for determining *KIR* gene copy numbers (Figure S3). To validate these results, DNA samples from 85 of the same cell lines were studied using an established real-time PCR method for quantifying *KIR* genes.[12] We observed 99.4% concordance between the results obtained by PING_gc and those obtained by the real-time PCR (Figure 4B). Of 10 discordant results, from 1,700 determinations, four involve rare alleles that were not detected by the primers of the real-time PCR assay (*KIR2DL2*009* and *KIR3DL2*076*; the latter allele being discovered during this study). Of the other six discordant results, two were due to false positives of the real-time PCR (as shown independently using a standard PCR method)[27], two were just below the threshold values of the real-time PCR (but clearly positive with PING_gc and standard PCR) and two remain unexplained (Figure 4B). Thus, the discrepancies were likely due to low-frequency errors in sensitivity or specificity of the real-time PCR method. We conclude that analysis of high-throughput sequencing data with the PING_gc module provides precise measurement of *KIR* gene content and copy number, giving almost 100% accuracy. Using PING_gc, we identified 26 distinct *KIR* gene copy number genotypes in the 97 cell lines analyzed (Figure 4C). Two cells have duplicated *KIR3DP1-2DL4-3DL1/S1* segments (Copy number genotypes 10 and 18; Figure 4C) and three cells have haplotypes lacking *KIR2DL4* (genotypes 12, 13 and 15; Figure 4C). In summary, determination of copy number alone increases the resolution of *KIR* genotyping and is an important step towards understanding the role of KIR polymorphism in disease.[61] We next sought to include the allele calling components in validation of the PING pipeline in order to achieve full resolution of *KIR* genotypes.

High resolution genotypes of *KIR* alleles (PING_allele)

PING_allele determines *KIR* allele genotypes according to all known *KIR* coding sequence alleles (Material and Methods). PING_allele was first validated using whole-exome data from a sample of 15 KhoeSan individuals.[47] For these individuals, the *KIR* copy number and allele data produced by PING matched the data obtained previously using the established methods of Sanger sequencing and pyrosequencing-based genotyping of *KIR* genes.[16; 31] Because the *KIR3DL1*, *KIR3DS1* and *KIR3DL2* genes of lineage II KIR (See Appendix A: KIR and HLA nomenclature) exhibit high polymorphism and structural variation,[13; 23; 62] they were chosen as a further test of PING_allele. The pipeline was applied to data obtained, using the capture/NGS method, from 30 family trios from Mali in West Africa (sample Set 2; Material and Methods). In this highly heterozygous population we identified 18 *KIR3DL1/S1*, 15 *KIR3DL2* alleles and three *KIR3DL1/2v* alleles (Figure S5A). These alleles were authenticated using established pyrosequencing and Sanger sequencing methods (Material and Methods), as well as by their segregation in the trios. *KIR3DL1/2v* is a fusion of *KIR3DL1* and *KIR3DL2* that segregates with *KIR3DL1/S1* and expresses a functional protein.[13] Importantly, we correctly identified individuals having distinct combinations of *KIR3DL1/S1* alleles, *KIR3DL1/2v* fusion genes and *KIR3DL1*-deleted haplotypes

(Figure 5). To expand the analysis, we next analyzed 2,112 individuals of the 1000 genomes data set.[48] From their exome sequences, we identified 50 *KIR3DL1/S1*, 46 *KIR3DL2* and five *KIR3DL1/2v* alleles (Figure S5A), as well as 14 *KIR3DL1/S1* duplication and 13 *KIR3DL1/S1* deletion haplotypes (Figure S5B). Such duplicated and deleted *KIR* haplotypes were detected in all 26 populations represented in the 1000 genomes dataset (Figure S5B). These results demonstrate that the capture/NGS method coupled with the copy number and allele components of the PING pipeline correctly identify the extensive and complex variation of lineage II *KIR*. The *KIR3DL1/S1* and *KIR3DL2* genotypes obtained for all individuals analyzed from the 1000 Genomes project are shown in Figure S5C.

Novel Allele Identification by PING

We describe *KIR* variants that were previously undiscovered, but identified and characterized using the methods described here, as 'novel'. The PING pipeline identifies such novel alleles by the presence either of one or more novel SNPs or a novel combination of known SNPs (Material and Methods). To test this 'new allele discovery' component of PING we again used the lineage II *KIR* genes. In the course of analyzing the 1000 Genomes data, we identified 100 novel alleles: 33 *KIR3DL1/S1*, 65 *KIR3DL2,* and two forms of *KIR3DL1/2v*. Defining these alleles are 88 novel SNPs (39 in *KIR3DL1/S1*, 49 in *KIR3DL2*: Figure S6A and Figure S6B) and 17 novel combinations of known SNPs (Figure S6C). Sequences for all the novel alleles were validated by standard methods: PCR amplification from genomic DNA of source material, followed by cloning and/or Sanger sequencing (Material and Methods). Of the 2,112 individuals studied, 229 (10.8%) have at least one novel lineage II *KIR* allele (Figure S5C). A total of 333 different *KIR3DL1/S1*-*KIR3DL2* haplotypes were identified in this analysis (Figure S5C).

High resolution allele and copy number *KIR* genotypes

We applied PING_allele to the *KIR* sequence data obtained from the 97 IHWG cells. This analysis of 13 *KIR* genes identified 144 different *KIR* sequences, 128 corresponding to established alleles and 16 representing novel alleles. The latter were all shown to be authentic using the standard methods described above for *KIR3DL1/S1* and *KIR3DL2* (Figure S6D). By considering all 144 *KIR* variants, we identified a minimum of 104 centromeric and 42 telomeric *KIR* haplotypes in the cell panel (Figure 6 and Figure S7). Consistent with our results are *KIR* genotyping data obtained previously from the IHWG cells, which achieved a limited discrimination of alleles.[6; 59] These analyses demonstrate that PING accurately processes high-throughput sequence data to give accurate high-resolution *KIR* genotypes. The high-resolution *KIR* genotypes of the IHWG cell panel has been compiled (Figure S7), providing a resource for future investigation.

Validation of the high-resolution *HLA class I* capture method

Because KIR and HLA class I glycoproteins are functionally interacting receptors and ligands, the capture/NGS method was designed to capture and analyze the two gene families in the same reaction (Material and Methods). To validate the *HLA class I* sequences obtained by this method, we first analyzed the panel of 97 IHWG cell lines. In previous studies[38-40], including high-density whole genome SNP analysis,[43] 90 of the IHGW cells were judged to be homozygous for *HLA-A*, *B* and *C*, with five of the other seven cells being homozygous for two of the three *HLA class I* genes. Our high-throughput sequencing results are completely concordant with the genotypes previously determined by conventional methods[15] (Figure S8A). In the course of validation, we defined two novel *HLA-C* alleles that encode distinctive proteins (Figure S8B). Because our method gives full-length genomic sequences, including intron and flanking region sequences, all the *HLA -A, -B, -C* alleles of the IHWG cells are now defined at much higher resolution ("four field" see Appendix A: KIR and HLA nomenclature) than previously achieved.

The IHWG cells represent an unusual, and highly selected, sampling of the human population because they are *HLA* homozygous, and many of them derive from consanguineous individuals. To extend the study to heterozygous individuals, we applied the capture/NGS method to 30 West African family trios from Mali (sample Set 2; Material and Methods) and 188 individuals selected at random from a panel of healthy Europeans (sample Set 3). The capture/NGS method achieved full coverage of the *HLA class I* genes, and an example of the result obtained from one individual is shown in Figure 7A. In the Africans, 22 *HLA-A*, 30 *HLA-B* and 15 *HLA-C* alleles were identified. These allele sequences agreed completely with *HLA class I* genotypes we determined using standard probe-based and Sanger sequencing methods. They were also consistent with the observed segregation of *HLA -A, -B, -C* alleles within the family trios (Figure S8C-D). In total, 170 distinct *HLA class I* alleles were identified in the three validation sets (IHWG cell lines, African trios and the Europeans). These 170 alleles, include 62 of the 69 fundamental allele types defined by the first two digits of the HLA nomenclature (Figure 7B), and thus cover most, if not all, of the breadth of *HLA class I* allelic diversity.[21] Thus it is likely that all *HLA-A*, *-B* and *-C* alleles can be captured and sequenced by our method. In support of this thesis, the capture/NGS method robustly detects and sequences *B*73:01* (Ashouri et al. in prep), a unique allele of archaic origin that has by far the most divergent *HLA class I* gene sequence in the modern human population.[63] Further demonstrating our method's robust capacity to target divergent sequences, we successfully captured and sequenced all alleles of the *Patr-A, -B, -C* genes (Figure S8E), the orthologs of *HLA-*

A -*B* -*C*, respectively, from the chimpanzee that was the subject for the chimpanzee genome project.[44] In summation, the breadth and depth of our results give confidence that our method will be able to capture the full range of *HLA class I* alleles.

Discussion

We developed an integrated capture/NGS and bioinformatics method to characterize completely the structure and sequence of the highly polymorphic *KIR* and *HLA class I* genes. The approach enables a focused and extensive definition of this physiologically important variation, which is not possible with any other single method. Our method is also well-suited for genotyping the large cohorts required for insightful study of population genetics and disease association, as well as donor selection for clinical transplantation. All components of the method were validated using panels of DNA samples that represent the observed range of human variation for these complex genomic regions. Both the method and the information we have obtained during its development should prove valuable resources for future studies.

We used DNA from well-characterized immortalized human B cell lines as reference materials during the design and optimization of the laboratory and bioinformatics methods. These panels of IHWG and 1000 Genomes cell lines are generally available for other researchers (see Web Resources). For validation we focused on *lineage II KIR* genes, because they exhibit some of the most extreme and complex genomic variation within the human *KIR* locus. We identified and distinguished deletions and duplications of *KIR3DL1/S1* and presence of the *KIR3DL1/2v* fusion gene, at the same time as defining the alleles of these genes. This was achieved by the combination of quantitative assessment of read depth and virtual sequence probing, coupled with reference alignment. Such independent verification is critical for characterizing structural *KIR* variants, which are not detected by methods that depend only on the alignment of sequence reads to reference haplotypes.[64; 65] In summary, the validation experiments demonstrate our method to be robust and capable of detecting the full range of *KIR* genomic variation.

With few exceptions, studies of *KIR* in human populations and disease cohorts have analyzed *KIR* gene content, but not allelic diversity.[66; 67] Such studies were seminal for showing how *KIR* genomic diversity can shape the immune response and provide resistance to disease.[68] Gene content studies also uncovered the influence of *KIR* diversity on the success of reproduction and bone marrow transplantation.[9; 11] The few studies that have focused on specific *KIR* genes, their allelic diversity and copy number, have refined these disease associations and implicated specific alleles.[61; 69] In the course of validating our method, we identified and characterized 116 novel *KIR* alleles. This new knowledge of *KIR* polymorphism makes substantial contributions to the *KIR* database.[15] For example, the number of *KIR3DL2* alleles was doubled and is now in excess of 100. We also show that 476 (22.5%) of the 1000 Genomes individuals have at least one example of a structural variant or novel allele of *KIR3DL1/S1* or *KIR3DL2* (Figure S5c). All of these genomic variations have potential to influence NK cell function, but they are not visible to typing at the level of *KIR* gene content. A strong case can therefore be made that high-resolution knowledge of *KIR* diversity, in all its forms, will identify additional disease associations and improve the understanding of those already known.

The study of human populations, their evolutionary dynamics, ancestry and disease, has benefited from methods of genome-wide associations (GWAS) that genotype numerous SNP markers in large cohorts of individuals. Such analysis of the *KIR* region has been impractical, because its extraordinary structural diversity leaves few locations suitable for designing binary SNP markers, and many of the *KIR* region genotyping results fail routine quality control filters. Thus the ‘immunochip’, which focuses on immune-system genes[70] and has refined the role of *HLA*-

associated diseases, includes relatively few informative SNPs in the *KIR* region. These SNPs are located in the *KIR3DL3*, *KIR2DL4*, *KIR3DL1/S1* and *KIR3DL2* genes, which were previously assumed to be present in one copy on every haplotype.[14] Our study demonstrates that this is not the reality. In more than 10% of the 1000 Genomes individuals, one of these four *KIR* genes is either deleted, duplicated or part of a fusion gene. We conclude that SNP genotyping within the *KIR* locus using standard binary measurement is of little practical value.

To compensate for the absence of suitable SNPs within the *KIR* genes, an imputation method was recently described, which should give accurate re-assessment of *KIR* gene content diversity for many of the reported GWAS.[28] Imputation of *HLA class I* alleles from GWAS data has been informative for studies of immune-mediated diseases.[71] *HLA* allele imputation has varying efficiency, particularly in non-European individuals, in part because >10,000 alleles are described, but also because imputation relies on linkage disequilibrium, which can extend for shorter genomic tracts in non-Europeans than Europeans.[15; 71] Many of the *KIR* variants and polymorphisms identified by our method are not evenly distributed across human populations. For example, the *KIR3DL1/2v* gene fusion is restricted to Africans, who exhibit the lowest linkage disequilibrium, worldwide. Thus, it is unlikely that imputation will be able to resolve all the structural and allelic diversity of *KIR*.

We employed short-read technology because of its high fidelity. Pressing this point, all the novel SNPs identified by our method were confirmed by independent and well-established sequencing methods. The capture method we used, will probably soon be adapted to obtain longer fragment sizes and read lengths, which should become increasingly valuable as the sequencing error rates decrease.[72] Because we are able to capture and sequence the chimpanzee *KIR* region, we show the method we described likely captures the extent of human *KIR* diversity. Thus there is limited allelic dropout. Alternative methods that do not suffer allelic dropout are whole genome approaches. Because our method targets large numbers of individuals and with low impact on sequencing instrument and reagent resources, our assay provides an economic and practically viable alternative to whole-genome experiments. We also note our bioinformatics pipeline can obtain accurate *KIR* genotypes from any whole-genome sequencing experiments of sufficient mean depth. The pipeline is also designed for application to any highly polymorphic gene system. Our approach is aimed for genotyping very large numbers of individuals but with low impact on computer resources. In those properties it differs from the population graph method of allele designation that has been applied to *HLA*.[73] However, this method could be a valuable complement to our methods if also applied to *KIR*.

The first *KIR* cDNA sequences were reported in the late 1990s.[74; 75] This led to research that revealed the unanticipated scope of genetic complexity and diversity of the human *KIR* gene family.[27] The method we describe here will facilitate determination of a complete description of *KIR* variation in the human population, its interaction and co-evolution with *HLA class I,* and its influence on physiology, disease and immunotherapy.

Appendix A; *KIR* nomenclature

Throughout this paper any unique DNA sequence that spans a coding region (otherwise known as coding DNA sequence; CDS) is considered a distinct allele. Those alleles that encode a unique protein sequence define an allotype. *KIR* genes and alleles are named by the KIR nomenclature committee, formed from members of the WHO Nomenclature Committee for factors of the HLA system, and members of the HUGO Genome Nomenclature Committee.[15] The *KIR* database is part of the ImmunoPolymorphism database (IPD), which is listed in the Web Resources section.

In the nomenclature for alleles, the digit (2 or 3) and letter D following the '*KIR'* prefix indicate whether two (2D) or three (3D) immunoglobulin-like domains are present in the encoded protein. After the letter D is another letter, either L, S or P. The letters L and S refer to the relative length of the cytoplasmic tail: being either short (S) or long (L). Long-tailed KIR are inhibitory receptors, whereas short-tailed KIR are activating receptors. P denotes a pseudogene. Next is a number that distinguishes KIR encoded by different genes, but having the same domain number and signalling function. There are two instances where a pair of *KIR* gene names was combined to form one name after it became apparent they occupied the same locus –these are *KIR2DL2/3* and *KIR3DL1/S1*. Following the gene name, there are three fields of numbers that distinguish the various types of alleles. The first field, of three digits, distinguishes alleles that encode different allotypes. Thus these alleles encode proteins with different amino-acid sequences. The second field, of two digits, distinguishes alleles that encode the same allotype, but differ by one or more synonymous substitutions in the coding region. The third field, also of two digits, distinguishes alleles with the same coding region sequence but have one or more substitutions in introns, flanking regions and other parts of the gene. For example, KIR3DL2 is a receptor with three Ig-like domains and inhibitory signalling function. *KIR3DL2*001* and *KIR3DL2*002* are alleles that encode allotypes with different amino-acid sequence, whereas the *KIR3DL2*00101* and *KIR3DL2*00102* alleles both encode the KIR3DL2*001 allotype, but differ by a synonymous nucleotide substitution in the coding region sequence. The *KIR3DL2*0010101* and *KIR3DL2*0010102* alleles also encode the KIR3DL2*001 allotype, but they differ by nucleotide substitutions in the introns.

There are four phylogenetically distinguished lineages of human *KIR*.[76] Lineage I and III KIR molecules possess two Ig-like domains. Lineage I (KIR2DL4-5) interact with HLA ligands of low polymorphism including HLA-G, and lineage III (KIR2DL1-3, KIR2DS1-5) interact with specific allotypes of HLA-B and -C molecules. Lineage II (KIR3DL1/S1 and KIR3DL2) and V (KIR3DL3) molecules have three Ig-like domains. Lineage II molecules interact with allotypes of HLA-A or -B,[77] and the ligand for lineage V remains unknown. We focused on the lineage II because this lineage is characterized by extensive structural and sequence diversity.[13; 23] Although most human *KIR* haplotypes have two lineage II genes, *KIR3DL1/S1* and *KIR3DL2*, some haplotypes have deleted *KIR3DL1/S1,* some have duplicated *KIR3DL1/S1* and some have an in-frame fusion of *KIR3DL1* and *KIR3DL2*, termed *KIR3DL1/2v*.[13] There are three divergent lineages of KIR3DL1/S1, the inhibitory KIR3DL1*005 and KIR3DL1*015, which are divergent and highly polymorphic, and activating KIR3DS1, which is less polymorphic.[23]

KIR haplotypes form two groups: *A* and *B*. *KIR A* haplotypes have a fixed content of seven genes and two pseudogenes, whereas *KIR B* haplotypes vary in gene content.[6; 14; 27] There are around 20 common *KIR B* haplotypes with different gene content and numerous additional *KIR B* haplotypes that are rare[6; 12; 27]

HLA class I alleles are named using a hierarchical set of fields, separated by colons, and each containing as many different digits as is needed to distinguish all alleles. The first field distinguishes the major alleles, which differ by multiple nucleotide and amino-acid substitutions (e.g. *HLA-A*01* and *HLA-A*02*). The second field distinguishes the subtypes of each major allele (e.g. *HLA-A*02:01* and *HLA-A*02:153*), which encode allotypes that differ by one or more amino-acid substitutions (e.g. HLA-A*02:01 and HLA-A*02:153). The third field distinguishes subtypes that encode proteins of identical amino-acid sequence, but differ by one or more synonymous substitutions within the protein-encoding exons (e.g. *HLA-A*02:01:01* and *HLA-A*02:01:02*). The fourth field distinguishes subtypes that have identical coding-region sequence, but differ by one or more nucleotide substitutions within introns or the transcribed 3' and 5' flanking regions (e.g. *HLA-A*02:01:01:01* and *HLA-A*02:01:01:02*). In addition, suffix letters are used to denote known expression variants (e.g. N denotes a 'Null' allele, for which the encoded protein is not expressed at the cell surface).

Supplemental Data

Supplemental Data include four figures and four Excel spreadsheets.

Acknowledgements

This study was supported by U.S. National Institutes of Health grants U01 AI090905, R01 GM109030, R01 AI17892 and U19 AI119350. Authors Steven Norberg and Mostafa Ronaghi are employees of Illumina Inc.

Web Resources and Accession Numbers

The URLs for data, material and programs used herein are as follows:

Coriell repository, https://catalog.coriell.org/

FastX toolkit, http://hannonlab.cshl.edu/fastx_toolkit/

GenDX NGSengine, http://www.gendx.com/

Haplo.stats, http://www.mayo.edu/research/labs/statistical-genetics-genetic-epidemiology/

ImmunoPolymorphism database (IPD), http://www.ebi.ac.uk/ipd/

International Histocompatibility Working Group (IHWG), www.ihwg.org/

OMIM, http://www.omim.org/

PING scripts (The versions of PING components used to generate the data described here), https://web.stanford.edu/~n0rmski/projectH/

PING executable R files (The latest versions of PING_gc and PINGallele), https://github.com/wesleymarin/

R program, https://www.r-project.org/

Newly-discovered *KIR and HLA class I* allele sequences were submitted to GenBank and the IPD database.[15] Their official names and GenBank accession numbers are given below and in Figure S6 (*KIR* prefix excluded for brevity):

*3DL1*00103* (LN606766), *3DL1*00104* (KP784298), *3DL1*00105* (KP784297), *3DL1*00404* (KP784290), *3DL1*00504* (KP784291), *3DL1*00505* (KP784294), *3DL1*01506* (KP784293), *3DL1*05902* (KP784300), *3DL1*077* (LN606765), *3DL1*088* (LN606767), *3DL1*089* (LN606768), *3DL1*090* (LN606769), *3DL1*091* (LN606770), *3DL1*092* (LN606771), *3DL1*093* (LN606772), *3DL1*094N* (LN606773), *3DL1*095* (KP784289), *3DL1*096*

(KP784285), _3DL1*097_ (KP784286), _3DL1*098_ (KP784301), _3DL1*099_ (KP784288), _3DL1*100_ (KP784299), _3DL1*101_ (KP784292), _3DL1*102_ (KP784295), _3DL1*103_ (KP784296), _3DL1*109_ (KP784287), _3DS1*01304_ (KP784279), _3DS1*01305_ (KP784278), _3DS1*01306_ (KP784276), _3DS1*01307_ (KP784284), _3DS1*104_ (KP784277), _3DS1*105_ (KP784280), _3DS1*106_ (KP784283), _3DS1*107_ (KP784281), _3DS1*108_ (KP784282), _3DL2*00106_ (KJ535483), _3DL2*081_ (KJ535484), _3DL2*073_ (KJ535485), _3DL2*071_ (KJ535486), _3DL2*080_ (KJ535487), _3DL2*00303_ (KJ535488), _3DL2*00703_ (KJ535489), _3DL2*069_ (KJ535490), _3DL2*070_ (KJ535491), _3DL2*072_ (KJ535492), _3DL2*074_ (KJ535493), _3DL2*068_ (KJ535494), _3DL2*00304_ (KJ535495), _3DL2*079_ (KJ535496), _3DL2*077_ (KJ535497), _3DL2*01002_ (KJ535498), _3DL2*076_ (KJ535500), _3DL2*00203_ (KJ535501), _3DL2*082_ (KJ535502), _3DL2*078_ (KJ535503), _3DL2*101_ (LN995832), _3DL2*00708_ (LN995833), _3DL2*100_ (KP784305), _3DL2*102_ (LN995834), _3DL2*00705_ (LN649139), _3DL2*089_ (LN649146), _3DL2*00502_ (LN649136), _3DL2*095_ (LN649155), _3DL2*00204_ (LN649156), _3DL2*00706_ (LN649150), _3DL2*084_ (LN649140), _3DL2*093_ (LN649152), _3DL2*087_ (LN649143), _3DL2*00707_ (LN649153), _3DL2*098_ (KP784302), _3DL2*094_ (LN649154), _3DL2*01003_ (LN649149), _3DL2*00107_ (LN649144), _3DL2*085_ (LN649141), _3DL2*086_ (LN649142), _3DL2*088_ (LN649145), _3DL2*090_ (LN649147), _3DL2*097_ (KP784303), _3DL2*091_ (LN649148), _3DL2*096_ (LN649157), _3DL2*092_ (LN649151), _3DL2*00704_ (LN606764), _3DL2*01004_ (KP784304), _3DP1*01002_ (KP893537), _3DP1*015_ (KP893538), _3DL2*108_ (LN999781), _3DL2*107_ (LN999782), _3DL2*104_ (LN999783), _3DL2*109_ (LN999784), _3DL2*00602_ (LN999785), _3DL2*106_ (LN999786), _3DL2*04302_ (LN999787), _3DL2*00709_ (LN999788), _3DL2*01004_ (LN999790), _3DL2*103_ (LN999791), _3DL2*01902_ (LN999792), _3DL2*021_ (LN999793), _3DL2*07902_ (LN999794), _3DL2*06002_ (LN999795), _3DL2*105_

(LN999796), *3DL2*00109* (LN999797), *3DL2*10002* (LN999798), *2DL1*032N* (KP893536), *2DL3*034* (KP784272), *2DL5A*021* (KP784273), *2DP1*00203* (KP784307), *2DP1*00204* (KP784309), *2DP1*015* (KP784306), *2DP1*016* (KP784275), *2DP1*017* (KP784308), *2DP1*018* (KP784274), *2DP1*019* (KP784310), *2DS3*008* (KP784269), *2DS5*015* (KP784270), *2DS5*016* (KP784271), *HLA-C*01:02:30* (KP893072), *HLA-C*07:18* (KP893073).

References

1. Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, C.C., Jr., Wright, M.W., et al. (2004). Gene map of the extended human MHC. Nat Rev Genet 5, 889-899.
2. Trowsdale, J., and Knight, J.C. (2013). Major histocompatibility complex genomics and human disease. Annual review of genomics and human genetics 14, 301-323.
3. Zinkernagel, R.M., and Doherty, P.C. (1974). Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. Nature 248, 701-702.
4. Vivier, E., Raulet, D.H., Moretta, A., Caligiuri, M.A., Zitvogel, L., Lanier, L.L., Yokoyama, W.M., and Ugolini, S. (2011). Innate or adaptive immunity? The example of natural killer cells. Science 331, 44-49.
5. Campbell, K.S., and Purdy, A.K. (2011). Structure/function of human killer cell immunoglobulin-like receptors: lessons from polymorphisms, evolution, crystal structures and mutations. Immunology 132, 315-325.
6. Pyo, C.W., Guethlein, L.A., Vu, Q., Wang, R., Abi-Rached, L., Norman, P.J., Marsh, S.G., Miller, J.S., Parham, P., and Geraghty, D.E. (2010). Different patterns of evolution in the centromeric and telomeric regions of group A and B haplotypes of the human killer cell Ig-like receptor locus. PloS one 5, e15115.
7. McLaren, P.J., and Carrington, M. (2015). The impact of host genetic variation on infection with HIV-1. Nature immunology 16, 577-583.
8. Ahn, R.S., Moslehi, H., Martin, M.P., Abad-Santos, M., Bowcock, A.M., Carrington, M., and Liao, W. (2015). Inhibitory KIR3DL1 Alleles are Associated with Psoriasis. The British journal of dermatology.
9. Mancusi, A., Ruggeri, L., Urbani, E., Pierini, A., Massei, M.S., Carotti, A., Terenzi, A., Falzetti, F., Tosti, A., Topini, F., et al. (2015). Haploidentical hematopoietic transplantation from KIR ligand-mismatched donors with activating KIRs reduces nonrelapse mortality. Blood 125, 3173-3182.
10. Hollenbach, J.A., Pando, M.J., Caillier, S.J., Gourraud, P.A., and Oksenberg, J.R. (2016). The killer immunoglobulin-like receptor KIR3DL1 in combination with HLA-Bw4 is protective against multiple sclerosis in African Americans. Genes and immunity 17, 199-202.
11. Parham, P., and Moffett, A. (2013). Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. Nature reviews Immunology 13, 133-144.
12. Jiang, W., Johnson, C., Jayaraman, J., Simecek, N., Noble, J., Moffatt, M.F., Cookson, W.O., Trowsdale, J., and Traherne, J.A. (2012). Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. Genome research 22, 1845-1854.
13. Norman, P.J., Abi-Rached, L., Gendzekhadze, K., Hammond, J.A., Moesta, A.K., Sharma, D., Graef, T., McQueen, K.L., Guethlein, L.A., Carrington, C.V., et al. (2009). Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, and haplotypes. Genome research 19, 757-769.
14. Wilson, M.J., Torkar, M., Haude, A., Milne, S., Jones, T., Sheer, D., Beck, S., and Trowsdale, J. (2000). Plasticity in the organization and sequences of human KIR/ILT gene families. Proc Natl Acad Sci U S A 97, 4778-4783.

15. Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P., and Marsh, S.G. (2015). The IPD and IMGT/HLA database: allele variant databases. Nucleic acids research 43, D423-431.

16. Norman, P.J., Hollenbach, J.A., Nemat-Gorgani, N., Guethlein, L.A., Hilton, H.G., Pando, M.J., Koram, K.A., Riley, E.M., Abi-Rached, L., and Parham, P. (2013). Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans. PLoS genetics 9, e1003938.

17. Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennett, W.S., Strominger, J.L., and Wiley, D.C. (1987). The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature 329, 512-518.

18. Boyington, J.C., Motyka, S.A., Schuck, P., Brooks, A.G., and Sun, P.D. (2000). Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand. Nature 405, 537-543.

19. Zhang, N., and Bevan, M.J. (2011). CD8(+) T cells: foot soldiers of the immune system. Immunity 35, 161-168.

20. Das, J., and Khakoo, S.I. (2015). NK cells: tuned by peptide? Immunological reviews 267, 214-227.

21. Parham, P., Lomen, C.E., Lawlor, D.A., Ways, J.P., Holmes, N., Coppin, H.L., Salter, R.D., Wan, A.M., and Ennis, P.D. (1988). Nature of polymorphism in HLA-A, -B, and -C molecules. Proc Natl Acad Sci U S A 85, 4005-4009.

22. DeGiorgio, M., Lohmueller, K.E., and Nielsen, R. (2014). A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS genetics 10, e1004561.

23. Norman, P.J., Abi-Rached, L., Gendzekhadze, K., Korbel, D., Gleimer, M., Rowley, D., Bruno, D., Carrington, C.V., Chandanayingyong, D., Chang, Y.H., et al. (2007). Unusual selection on the KIR3DL1/S1 natural killer cell receptor in Africans. Nature genetics 39, 1092-1099.

24. Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167-170.

25. Abi-Rached, L., Moesta, A.K., Rajalingam, R., Guethlein, L.A., and Parham, P. (2010). Human-specific evolution and adaptation led to major qualitative differences in the variable receptors of human and chimpanzee natural killer cells. PLoS genetics 6, e1001192.

26. Gomez-Lozano, N., and Vilches, C. (2002). Genotyping of human killer-cell immunoglobulin-like receptor genes by polymerase chain reaction with sequence-specific primers: an update. Tissue antigens 59, 184-193.

27. Uhrberg, M., Valiante, N.M., Shum, B.P., Shilling, H.G., Lienert-Weidenbach, K., Corliss, B., Tyan, D., Lanier, L.L., and Parham, P. (1997). Human diversity in killer cell inhibitory receptor genes. Immunity 7, 753-763.

28. Vukcevic, D., Traherne, J.A., Naess, S., Ellinghaus, E., Kamatani, Y., Dilthey, A., Lathrop, M., Karlsen, T.H., Franke, A., Moffatt, M., et al. (2015). Imputation of KIR Types from SNP Variation Data. American journal of human genetics 97, 593-607.

29. Vierra-Green, C., Roe, D., Hou, L., Hurley, C.K., Rajalingam, R., Reed, E., Lebedeva, T., Yu, N., Stewart, M., Noreen, H., et al. (2012). Allele-level haplotype frequencies and pairwise

linkage disequilibrium for 14 KIR loci in 506 European-American individuals. PloS one 7, e47491.

30. Gendzekhadze, K., Norman, P.J., Abi-Rached, L., Graef, T., Moesta, A.K., Layrisse, Z., and Parham, P. (2009). Co-evolution of KIR2DL3 with HLA-C in a human population retaining minimal essential diversity of KIR and HLA class I ligands. Proc Natl Acad Sci U S A 106, 18692-18697.

31. Nemat-Gorgani, N., Edinur, H.A., Hollenbach, J.A., Traherne, J.A., Dunn, P.P., Chambers, G.K., Parham, P., and Norman, P.J. (2014). KIR diversity in Maori and Polynesians: populations in which HLA-B is not a significant KIR ligand. Immunogenetics 66, 597-611.

32. Yawata, M., Yawata, N., Draghi, M., Little, A.M., Partheniou, F., and Parham, P. (2006). Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection and modulation of effector function. The Journal of experimental medicine 203, 633-645.

33. Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1986). Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. Nature 324, 163-166.

34. Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Muller, W.E., Wetter, T., and Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome research 14, 1147-1159.

35. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

36. Milius, R.P., Mack, S.J., Hollenbach, J.A., Pollack, J., Heuer, M.L., Gragert, L., Spellman, S., Guethlein, L.A., Trachtenberg, E.A., Cooley, S., et al. (2013). Genotype List String: a grammar for describing HLA and KIR genotyping results in a text string. Tissue antigens 82, 106-112.

37. Dorak, M.T., Shao, W., Machulla, H.K., Lobashevsky, E.S., Tang, J., Park, M.H., and Kaslow, R.A. (2006). Conserved extended haplotypes of the major histocompatibility complex: further characterization. Genes and immunity 7, 450-467.

38. Marsh, S.G.E., Packer, R., Heyes, J.M., Bolton, B., Fauchet, R., Charron, D., and Bodmer, J.G. (1996). The International Histocompatibility Workshop cell panel. In Genetic diversity of HLA Functional and medical implications, D. Charron, ed. (Paris, EDK), pp 26-28.

39. Mickelson E, Hurley C, Ng J, Tilanus M, Carrington M, Marsh SGE, Rozemuller E, Pei J, Rosielle J, Voorter C, et al. (2006). 13th IHWS Shared Resources Joint Report. IHWG Cell and Gene Bank and reference cell panels. In Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibilty Workshop and Conference, H. JA, ed. (Seattle, IHWG Press), pp 523-553.

40. Yang, S.Y., Milford, E., Hammerling, U., and Dupont, B. (1987). Description of the Reference Panel of B-Lymphoblastoid Cell Lines for Factors of the HLA system: The B-Cell Line Panel Designed for the Tenth International Histocompatibility Wprkshop. In Immunobiology of HLA Histocompatibility Testing, B. Dupont, ed. (New York, Springer-Verlag), pp 11-19.

41. Yang, Y., Chung, E.K., Wu, Y.L., Savelli, S.L., Nagaraja, H.N., Zhou, B., Hebert, M., Jones, K.N., Shu, Y., Kitzmiller, K., et al. (2007). Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor

against SLE susceptibility in European Americans. American journal of human genetics 80, 1037-1054.

42. Mack, S.J., Cano, P., Hollenbach, J.A., He, J., Hurley, C.K., Middleton, D., Moraes, M.E., Pereira, S.E., Kempenich, J.H., Reed, E.F., et al. (2013). Common and well-documented HLA alleles: 2012 update to the CWD catalogue. Tissue antigens 81, 194-203.

43. Norman, P.J., Norberg, S.J., Nemat-Gorgani, N., Royce, T., Hollenbach, J.A., Shults Won, M., Guethlein, L.A., Gunderson, K.L., Ronaghi, M., and Parham, P. (2015). Very long haplotype tracts characterized at high resolution from HLA homozygous cell lines. Immunogenetics 67, 479-485.

44. Consortium, T.C.S.a.A. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69-87.

45. Ba, A., Beley, S., Chiaroni, J., Bailly, P., and Silvy, M. (2015). RH diversity in Mali: characterization of a new haplotype RHD*DIVa/RHCE*ceTI(D2). Transfusion 55, 1423-1431.

46. Isobe, N., Keshevan, A., Gourraud, P.A., Zhu, A.H., Datta, E., Schlaeger, R., Caillier, S.J., Santaniello, A., Lizee, A., Himmelstein, D.S., et al. (2016). Effects of HLA genetic risk burden on MRI disease phenotypes in multiple sclerosis. JAMA Neurology In press.

47. Kidd, J.M., Sharpton, T.J., Bobo, D., Norman, P.J., Martin, A.R., Carpenter, M.L., Sikora, M., Gignoux, C.R., Nemat-Gorgani, N., Adams, A., et al. (2014). Exome capture from saliva produces high quality genomic and metagenomic data. BMC genomics 15, 262.

48. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

49. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

50. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods 9, 357-359.

51. Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome research 19, 1586-1592.

52. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420-426.

53. Martin, M.P., Bashirova, A., Traherne, J., Trowsdale, J., and Carrington, M. (2003). Cutting edge: expansion of the KIR locus by unequal crossing over. J Immunol 171, 2192-2195.

54. Gomez-Lozano, N., Gardiner, C.M., Parham, P., and Vilches, C. (2002). Some human KIR haplotypes contain two KIR2DL5 genes: KIR2DL5A and KIR2DL5B. Immunogenetics 54, 314-319.

55. R Development Core Team. (2008). A language and environment for statistical computing. In R Foundation for Statistical Computing. (Vienna, Austria.

56. Bonfield, J.K., and Whitwham, A. (2010). Gap5--editing the billion fragment sequence assembly. Bioinformatics 26, 1699-1703.

57. Watanabe, Y., Tokunaga, K., Geraghty, D.E., Tadokoro, K., and Juji, T. (1997). Large-scale comparative mapping of the MHC class I region of predominant haplotypes in Japanese. Immunogenetics 46, 135-141.

58. Sambrook, J.G., Bashirova, A., Palmer, S., Sims, S., Trowsdale, J., Abi-Rached, L., Parham, P., Carrington, M., and Beck, S. (2005). Single haplotype analysis demonstrates rapid

evolution of the killer immunoglobulin-like receptor (KIR) loci in primates. Genome research 15, 25-35.

59. Garcia, C.A., Robinson, J., Shilling, H.G., Hayhurst, J.D., Flicek, P., Parham, P., Madrigal, J.A., and Marsh, S.G. (2006). KIR gene characterisation of HLA homozygous cell lines. In Immunobiology of the Human MHC Proceedings of the 13th International Histocompatibilty Workshop and Conference, J.A. Hansen, ed. (Seattle, WA, IHWG Press), pp 1203-1207.

60. Cook, M.A., Norman, P.J., Curran, M.D., Maxwell, L.D., Briggs, D.C., Middleton, D., and Vaughan, R.W. (2003). A multi-laboratory characterization of the KIR genotypes of 10th International Histocompatibility Workshop cell lines. Human immunology 64, 567-571.

61. Pelak, K., Need, A.C., Fellay, J., Shianna, K.V., Feng, S., Urban, T.J., Ge, D., De Luca, A., Martinez-Picado, J., Wolinsky, S.M., et al. (2011). Copy number variation of KIR genes influences HIV-1 control. PLoS biology 9, e1001208.

62. Hou, L., Jiang, B., Chen, M., Ng, J., and Hurley, C.K. (2011). The characteristics of allelic polymorphism in killer-immunoglobulin-like receptor framework genes in African Americans. Immunogenetics 63, 549-559.

63. Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., et al. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334, 89-94.

64. Church, D.M., Schneider, V.A., Steinberg, K.M., Schatz, M.C., Quinlan, A.R., Chin, C.S., Kitts, P.A., Aken, B., Marth, G.T., Hoffman, M.M., et al. (2015). Extending reference assembly models. Genome Biol 16, 13.

65. Gargis, A.S., Kalman, L., Bick, D.P., da Silva, C., Dimmock, D.P., Funke, B.H., Gowrisankar, S., Hegde, M.R., Kulkarni, S., Mason, C.E., et al. (2015). Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nature biotechnology 33, 689-693.

66. Gonzalez-Galarza, F.F., Takeshita, L.Y., Santos, E.J., Kempson, F., Maia, M.H., Silva, A.L., Ghattaoraya, G.S., Alfirevic, A., Jones, A.R., and Middleton, D. (2015). Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic acids research 43, D784-788.

67. Hollenbach, J.A., Nocedal, I., Ladner, M.B., Single, R.M., and Trachtenberg, E.A. (2012). Killer cell immunoglobulin-like receptor (KIR) gene content variation in the HGDP-CEPH populations. Immunogenetics.

68. Bashirova, A.A., Martin, M.P., McVicar, D.W., and Carrington, M. (2006). The killer immunoglobulin-like receptor gene cluster: tuning the genome for defense. Annual review of genomics and human genetics 7, 277-300.

69. Martin, M.P., Qi, Y., Gao, X., Yamada, E., Martin, J.N., Pereyra, F., Colombo, S., Brown, E.E., Shupert, W.L., Phair, J., et al. (2007). Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. Nature genetics 39, 733-740.

70. Parkes, M., Cortes, A., van Heel, D.A., and Brown, M.A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet 14, 661-673.

71. de Bakker, P.I., and Raychaudhuri, S. (2012). Interrogating the major histocompatibility complex with high-throughput genomics. Human molecular genetics 21, R29-36.

72. Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M.C., and McCombie, W.R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome research 25, 1750-1756.

73. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M.R., and McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. Nature genetics 47, 682-688.

74. Colonna, M., and Samaridis, J. (1995). Cloning of immunoglobulin-superfamily members associated with HLA-C and HLA-B recognition by human natural killer cells. Science 268, 405-408.

75. Wagtmann, N., Biassoni, R., Cantoni, C., Verdiani, S., Malnati, M.S., Vitale, M., Bottino, C., Moretta, L., Moretta, A., and Long, E.O. (1995). Molecular clones of the p58 NK cell receptor reveal immunoglobulin-related molecules with diversity in both the extra- and intracellular domains. Immunity 2, 439-449.

76. Rajalingam, R., Parham, P., and Abi-Rached, L. (2004). Domain shuffling has been the main mechanism forming new hominoid killer cell Ig-like receptors. J Immunol 172, 356-369.

77. Vivian, J.P., Duncan, R.C., Berry, R., O'Connor, G.M., Reid, H.H., Beddoe, T., Gras, S., Saunders, P.M., Olshina, M.A., Widjaja, J.M., et al. (2011). Killer cell immunoglobulin-like receptor 3DL1-mediated recognition of human leukocyte antigen B. Nature 479, 401-405.

78. Shilling, H.G., Guethlein, L.A., Cheng, N.W., Gardiner, C.M., Rodriguez, R., Tyan, D., and Parham, P. (2002). Allelic polymorphism synergizes with variable gene content to individualize human KIR genotype. J Immunol 168, 2307-2315.

Figure Titles and Legends

Figure 1. Pipeline for analyzing sequence data from highly-polymorphic and structurally divergent haplotypes (*KIR*).

A. The PING (Pushing Immunogenetics to the Next Generation) pipeline has two broad arms and two modules. The first module (PING_gc) determines KIR gene copy numbers, and the second module (PING_allele) determines their alleles. Within each module are two arms. The first arm (KFFgc and KFFallele) is an analysis independent of any alignment or assembly that uses virtual probes to mine the raw data, and the second arm (MIRAgc and SOS) performs filtering and alignment of reads to reference sequences. Thus both copy-number and allele genotype are each derived by two independent methods. The techniques are described fully in the Methods.

B. Lists the data generated using the method described herein (1-3), and those obtained from other sources (4). Shown are the number of individuals, the genotyping results that they are used to validate, and the independent laboratory method used for this purpose is given at the far right.

Figure 2. *KIR* gene copy number genotype determined by read depth.

Shows that the ratio of reads mapping to a specific *KIR* gene relative to those that map to *KIR3DL3* can be used to calculate *KIR* copy number. The results from 97 samples are shown and sorted by the ratio. *KIR2DL4* (left) was present in 1, 2 or 3 copies per individual in the sample set and *KIR2DL5* (right) as 0, 1, 2, 3 or 4 copies.

Figure 3. The *KIR* region is >99.99% covered by the sequence data

A. Shows the target *KIR* region of chromosome 19: the gene locations are shown in orange and pseudogenes in gray. The *KIR* region varies in gene content and shown are two frequent ‘A’ and ‘B’ haplotypes. The ‘KIR’ prefix is omitted from the gene names for clarity (See Appendix A: KIR and HLA nomenclature). The human reference build hg19 has a *KIR A* haplotype. Underneath is a *KIR B* haplotype shown to scale.

B. Shows the read depth following stringent alignment of sequence reads (no base pairs mismatch and duplicates removed) from the PGF cell line to the PGF reference *KIR* haplotypes 1 (light purple) and 2 (dark purple).

C. Coordinates and features of two short gaps in PGF *KIR* haplotype 2. At the right is shown the location of the gaps.

Figure 4. *KIR* gene content and copy number genotypes

A. Gene *content* genotypes derived from all 97 cell lines using the PING pipeline. Black box indicates gene is present and clear box indicates absence. One example from each observed gene content genotype (GC type) is shown, with the number observed shown at the right.

B. Independent validation of the KIR *copy number* genotypes by real-time PCR[12] on 85 samples. There were four discrepancies due to alleles undetected by real-time PCR (allele). There were two false positives (pos) and two false negatives (neg), by real-time PCR. Two discrepancies remain unexplained.

C. Gene *copy number* genotypes derived from all 97 cell lines using PING_gc. Colored rectangle indicates gene present, and the shades represent the copy number as indicated in the key. One example from each observed gene copy number genotype (CN type) is shown, with the number observed at the right.

Figure 5. High resolution allele-level genotyping of *KIR*

Four examples of high-resolution allele and copy-number genotypes of lineage II *KIR* and their segregation in family trios: C - child, F -father, M - mother. Colored boxes show the segregating alleles. All members of family 1 have two alleles each of *KIR3DL1/S1* and *KIR3DL2*. Family 2 shows segregation of the *KIR3DL1/2v* fusion gene (the allele named *KIR3DL1*059*) that consists of exons 1-6 from *KIR3DL1* and 7-9 from *KIR3DL2*.[13; 78] For clarity *KIR3DL1/2v* is shown as an allele of *KIR3DL1*, and so there is no allele of *KIR3DL2* on this haplotype. Family 3 shows segregation of a haplotype that lacks *KIR3DL1/S1* and marked by presence of *KIR3DL2*006*. The gene copy numbers were determined using PING_gc, which indicated one copy of *KIR3DL1* was present in each of individuals 3C and 3M and two copies in 3F. Family 4 shows both the *KIR3DL1/2v* and the *KIR3DL1* negative haplotypes segregating to the child.

Figure 6. High resolution *KIR* allele and copy number genotypes of 97 IHWG cells

Shows four examples of high-resolution allele and copy-number genotypes of *KIR*. Individual 1 (SP0010) is homozygous *KIR-A* haplotype. Individual 2 (CB6B) has two different *B* haplotypes. Individual 3 (E481324) has duplication of three loci (in blue shading: denoted as *3DP1b*, *2DL4b*, *3DL1/S1b*). Individual 4 (LZL) has deletion of the central segment of the *KIR* haplotype (red).

Yellow shading denotes alleles that were identified for the first time in the current study. The full genotypes for each IHWG cell are given in Figure S7.

Figure 7. Capture of *HLA class I* genes for high-resolution allele genotyping

A. Shown is the read depth across each of the *HLA class I* genes from a representative sample (chosen virtue of having closest to the mean number of *HLA*-specific reads). Green lines indicate the coordinates of the exons that were covered. To generate this figure, full gene sequences (~3kb each) were obtained from IPD to represent all of the five *HLA class I* alleles known to be present in this sample (the sample is homozygous for a common allele of *HLA-A)*. Sequence reads were filtered to be specific for *HLA-A, -B -C*, then aligned to these references with high stringency. The read depth was measured using Samtools/bcf.

B. Lists the major *HLA class I* allele types detected in the study.