Computational Methods for Sequencing and Interpreting Killer-cell Immunoglobulin-like Receptors (KIRs) at Multiple Resolutions

A Dissertation

SUBMITTED TO THE FACULTY OF THE UNIVERSITY OF MINNESOTA BY

David Roe

IN PARTIAL FULFILLMENT OF THE REQUIERMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Dr. Rui Kuang, Advisor

January 2021

© David Roe, 2021

Abstract

It is now routine to sequence an individual's DNA, either whole genome sequences (WGS) or targeted, as part of a patient's medical plan. An exception is an ~200 kilobase region in chromosome 19 containing the killer-cell immunoglobulin-like receptors (KIR). These genes encode proteins that influence the actions of natural killer (NK) cells based on whether or not they bind with peptide-bound human leukocyte antigen receptors. This is evolutionary important to fight pathogens and mediate pregnancy. Modern medicine has correlated low-resolution KIR with many diseases and treatments, although the findings are often relatively vague and sometimes contradictory due to low resolution interpretation and/or small cohort sizes. The current best practices for KIR genotyping are to determine the presence/absence (PA) or copy number variation (CNV) of each gene using oligo- or primer-based polymerase chain reaction.

The goal of our research is to advance DNA sequencing and interpretation of human KIR haplotypes. To that end, we have created algorithms and workflows to enhance interpretations of KIR at resolutions from PA genotyping via short-read WGS to full-haplotype assembly from long-read targeted or whole-genome sequences. First, we developed the first workflow to efficiently and accurately capture, sequence, assemble, and annotate full KIR haplotype sequences. As part of this workflow, we designed small sequences to capture the DNA fragments. Next, we use the alignment pattern of those short sequences across finished KIR haplotypes to define and annotate haplotype structures. The results show, for the first time, that the KIR region is composed of 9 genes in 14 loci. Next, we annotated all 68 reported human haplotypes, aligned them at the structural level, and then refined the alignment down the base level, providing the first KIR haplotype multiple sequence alignment. These efforts have led to this region being the best annotated and most diverse in the human genome reference. We next leveraged the MSA to discover PA markers and leveraged them in the first KIR WGS genotyping application. It was evaluated independently and reported to be at least 97% accurate.

These discoveries and inventions are the culmination of several computational methods we have developed that interpret KIR under different typing resolutions. This multi-resolution aspect is crucial to overall understanding; it improves resolution at any given level by leveraging references and/or markers from other resolutions. From low-resolution genotyping from any kind of DNA sequence to the first efficient full-haplotype assembly method, these results advance interpretation of this important genetic region to the personal and population levels.

Table of Contents

List of tables

List of figures

1. Introduction

1.1. Functional roles and medical relevance

The protein coding killer-cell immunoglobulin-like receptor (KIR) genes span ~10-16 kb each, with pseudogenes that are ~5 and ~13 kb. Alleles from any two genes are over 85-98% identical. Frequent recombination has created several functional fusion genes. The function of the KIR proteins is to provide transmembrane signaling in natural killer (NK) cells that, along with other receptors, can lead to the release of cytokines or to the death of the target cell (infected, cancerous, foreign, etc.). Some of these ligand-receptor interactions stimulate the NK cell to react, while some inhibit the NK cell from reacting until the ligand is missing. Although some KIR binding partners are unknown or putative, generally KIR recognize human leukocyte antigen (HLA) and its peptide.

These KIR functions are important to human health. Whether or not these transmembrane receptors bind with peptide-bound HLA class I molecules determines how they help educate, activate, and inhibit NK cell functionality, including cytotoxicity and cytokine release. The accumulated evidence seems to indicate that KIR evolved with HLA to balance its pathogen defense effects with its effects on reproduction via the embryo-uterus interface(1)(2)(3). KIR effects are generally tissue-specific and tissue-variable. Within a given NK cell, KIR expression is stochastic, depending on the epigenetic profiles, inter- and intra-genic content of the haplotypes, binding alleles or lack thereof, methylation status, alternative splicing, and randomness(4)(5)(6). Besides investigating KIR in the context of viruses and pregnancy, medicine is also studying its effects in cancer, hematopoietic stem cells transplants, various autoimmune diseases, and immunotherapy(7)(8)(9).

1.2. Genomic context

KIR gene names reflect their protein structures(10). First the prefix “KIR”, followed by the number of extracellular binding domains (“2D” or “3D”), followed by a short or long intracellular signaling domain (“S”, L”), followed by an index. *KIR2DS1,* for example, has two extracellular domains (‘2D’) and a short intracellular domain (‘S’); it is the first gene named with that structure (‘1’), and *KIR2DS2* is the second. The Immuno Polymorphism Database for KIR (IPD-KIR) names KIR gene alleles, records their DNA-RNA-protein relationships, and annotates each gene’s alleles in a multiple sequence alignment (MSA)(11). It contains over 300 full-length DNA and almost 1000 protein reference alleles in the latest release, 2.9.0.

Human chromosome 19q13.4 contains a ~150-250 kilobase region encoding the 14 loci of the natural killer-cell immunoglobulin-like receptor (KIR) family. The genes are ~4-16 kilobases long and evolved via tandem duplication during primate evolution(12)(13). A few dozen KIR full sequence haplotypes and approximately 2500 full- or inter-gene sequences have been publicly deposited(14)(15)(11). It is believed that KIR genes have undergone a balancing selection via mutations, duplications and deletions into two broad categories of haplotypes, in which one category tends to vary more at the allelic level and the other tends to vary more at the structural (gene content and order) level(16)(15)(17). The two categories are labeled ‘A’ and ‘B’ (Figure 1)(18). Category ‘A’ contains one haplotype and its deleted forms. Category ‘B’ haplotypes are more structurally diverse and contain a variety of insertions and deletions. Generally, the A haplotype occurs with 50-60% frequency, haplotypes that are half-A and half-B occur with 30-40%, and the rest (<=20%) of the haplotypes are variants of the B haplotypes. Except for some rarer deleted forms, KIR haplotypes are structurally variable around 4 ‘framework’ genes (*KIR3DL3, KIR3DP1, KIR2DL4, KIR3DL2*), with *KIR3DL3* through *KIR3DP1* defining the proximal (or ‘centromeric’) region and *KIR2DL4* through *KIR3DL2* defining the distal (or ‘telomeric’) region, with the two gene-rich regions separated by the relatively large *KIR3DP1-KIR2DL4* intergene region, which contains a particularly strong recombination hotspot.

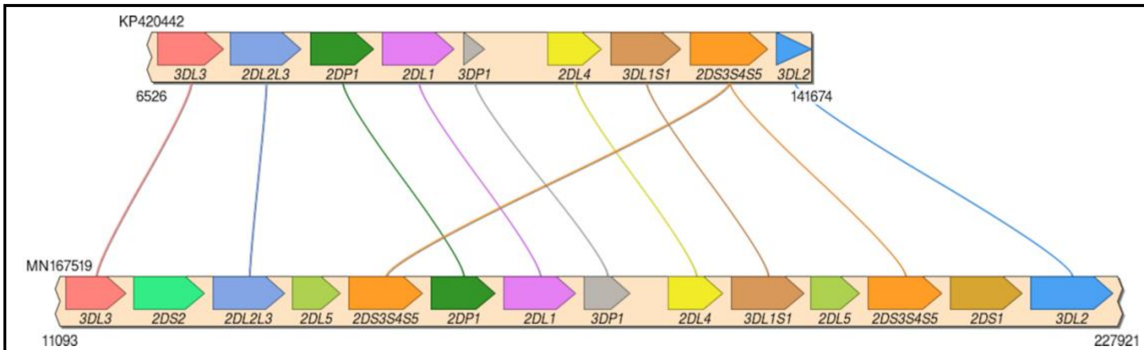

Figure 1. A canonical ‘A’ haplotype and ‘B’ haplotype. An ‘A’ haplotype (top) and a ‘B’ haplotype (bottom). Gene names have been abbreviated for space. Lines indicate the same gene on different haplotypes. The *KIR3DP1-KIR2DL4* intergene region splits the haplotypes into a ‘centromeric’ region to the 5’ (proximal) and a ‘telomeric’ region to the 3’ (distal). The A haplotype is called ‘cA01~cA01’ and the B haplotype is called ‘cB01~tB01’.

Like much of chromosome 19, the KIR region is dense with repetitive elements, which have provided the mechanisms for its recent evolution by tandem duplication and homologous recombination events. Dozens of distinct gene-content haplotypes are seen in Europeans alone(15)(16)(19)(20). Previous reports have documented over ten distinct common haplotype structures(14). KIR haplotypes are named in two halves: 'c' for centromeric (i.e., proximal) and 't' for telomeric (i.e., distal) separated by a recombination hotspot contained in the ~10 kb intergenic region between *KIR3DP1* and *KIR2DL4*. Each half is also labelled 'A' or 'B', designating one of two families of haplotypes, based on the gene content(18). The A family haplotype denotes haplotypes with one main gene content structure and relatively large allelic variation. The B family of haplotypes denotes a class of haplotypes with relatively more structural variation and less allelic variation. The haplotype named 'cA01~tB01', for example, means the first (01) centromeric A region in cis ('~') with the first telomeric B region.

1.3. Current state of DNA interpretation

Almost all medical studies that include KIR genotype at the PA or copy number variation (CNV) via PCR primers or oligos specific to a gene. It is difficult to interpret the KIR region with high-throughput sequencing reads for an individual human genome when the structural arrangements are unknown; indeed, it is difficult even when the structural haplotypes are known, since the read length is too short to map unambiguously to the repetitive and homologous KIR genes. As a consequence, the reads from KIR region are ignored, as to the best of our knowledge, there are currently no algorithms to interpret KIR from WGS. Single nucleotide polymorphism (SNP)-based KIR interpretation is more commonly applied. For example, KIR*IMP is a web-application to predict genes and haplotypes from microarray SNP genotypes(21). As an algorithm whose raw data is microarray calls, KIR*IMP can interpret KIR from genome wide SNP arrays, but it is not applicable to interpret KIR from raw sequences.

Genetic interpretation of exonic-or-lower resolutions from next generation sequencing (NGS) is often ambiguous and unphased, and therefore limits precise understanding of how KIR sequences affect phenotypes. The importance of high-throughput high-resolution typing is exemplified by the fact that the genes contain extensive exonic SNP and short insertion/deletion (indel) variations which rivals that of its binding partner: HLA class I(1)(22). Over 300 full-length DNA and almost 1000 protein reference alleles have been reported in IPD-KIR(23). All resolutions except haplotyping are ambiguous or require statistical phasing from few references. A cost-effective high-throughput method that could characterize all the sequences within the KIR haplotypes in cis could advance that understanding and clarify previously ambiguous and/or contradictory evidence. To date, the only approach for full haplotyping was to physically separate and amplify maternal and paternal haplotypes via fosmids for subsequent sequencing(24)(15)(19)(14)(25), a process whose expense has generally prohibited its use in large-scale association studies. While high-resolution haplotyping by fosmid clones or full gene by PCR is costly and inefficient, low resolution genotyping of gene presence/absence or copy number provide limited information for functional analysis and association tests.

1.4. Novel contributions

The overall goal of our research is to create computational methods and reference libraries to interpret high-throughput DNA KIR sequences. We generate interpretations and/or evaluate others' interpretations at multiple DNA resolutions: haplotype, gene, exon, gene copy number variation, gene presence/absence, and single nucleotide polymorphism, as shown in Figure 2. Haplotypes can be interpreted at multiple resolutions. We use the term 'sequence haplotype' to indicate an ordered DNA sequence of one parental chromosome. We use the term 'structural haplotype' to indicate the ordered gene content of one parental chromosome. Similarly, we use the term 'exon haplotype' to indicate the ordered exon sequences of one parental chromosome; KIR exon genotyping usually sequences a subset of exons (usually exons 4-6) responsible for the extracellular antigen recognition domain. Table 1 shows our publications and algorithms in the multi-resolution context.

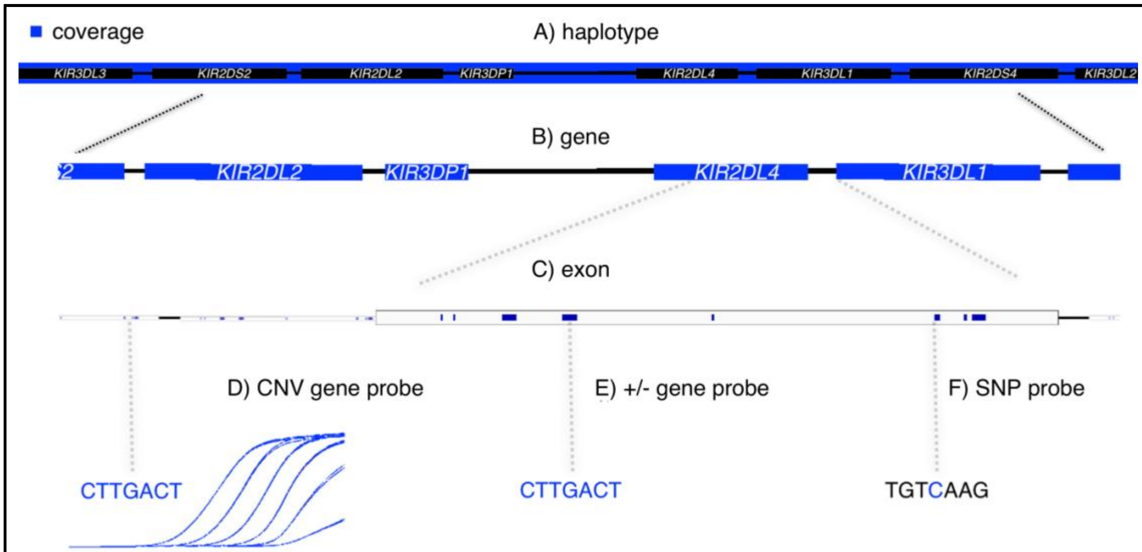

Figure 2. Multi-resolution interpretation. Blue bars indicate DNA coverage at each resolution. A) Full sequence haplotype (~100-250kb). B) Full gene (~10kb). C) Exons (dozens to hundreds of bases). D) Short probes (1-2 dozen bases) via qPCR indicating the copy number of the genes. E) Short probes (1-2 dozen bases), via qPCR indicating presence/absence of the genes. F) Traditional SNP.

pub details	algorithm	input	output
PLoS ONE (2012)	exonic haplotype-pair predictions with EM	exonic genotypes from targeted PCR	exonic haplotype-pair predictions & exonic haplotype frequencies
PLoS ONE (2016)	structural haplotype-pair predictions with EM	PA/CNV genotypes from targeted PCR	structural haplotype-pair predictions & structural haplotype frequencies
Genes & Immunity (2017)	sequence haplotype assembly and reference library	fosmid-generated long-read sequences	sequence haplotigs and new reference sequence haplotypes
Frontiers in Immunology (2020)	sequence haplotype assembly and reference library	targeted & WGS long-read sequences	sequence haplotigs and new reference sequence haplotypes
Frontiers in Immunology (2020)	motif and sequence haplotype MSA	sequence haplotype sequences	alignment and annotation of haplotypes/genes/exons/introns
Frontiers in Immunology (2020)	genotyping with kmer associations	WGS	PA genotype and structural haplotype-pair predictions

Table *1*. Publications. Column 1 contains the journal name and year of publication. Column 2 contains the lead authors of the publication. Column 3 is a short description of the main results. Column 4 describes the resolution of the input. Column 5 describes the resolution of the output.

1.4.1. Haplotype resolution

Our primary contributions enable efficient sequence haplotype assembly (Figure 2A, Table 1 row 4), and our primary findings stem from these assemblies (Table 1, rows 5-6). For the assembly workflow, we developed an efficient set of probes to capture long DNA fragments for sequence haplotype assembly; a software workflow error corrects the reads, bins the reads per gene, assembles the corrected reads, and annotates the haplotigs for gene and exon boundaries. Virtual probes can be used *in vitro* to capture KIR DNA. *in silico*, they can be used to bin the reads, and their alignment order across a haplo-tig/type annotates the locations of the genes. We used these structural patterns to align all human haplotypes at a high level, and then used traditional alignment to refine to the base level. This contribution is the first multiple sequence alignment of all public human haplotypes (Table 1, line 5).

1.4.2. Gene and exon resolution

The first human haplotype MSA empowered the first algorithm and software programs (Table 1, lines 4-5) to annotate sequence haplo-types/tigs with the locations of the genes and exons (Figure

2B-C). The order of the haplotype capture probes enable structural (gene location) annotation of the assembled or reference haplotypes, and our creation of the MSA enabled discovery of each gene's exon/intron boundaries. This is the first general KIR annotation system, and we applied it to annotate all haplotypes with respect to reference exonic and genic allele names.

Given exon-resolution sequences from short, targeted PCR (Figure 2C), we predicted individual exon haplotype-pair predictions and exon haplotype frequency estimates (Table 1, line 1) on a cohort of over 500 individuals. The results revealed patterns of exonic linkage disequilibrium, the first reports of its kind.

1.4.3. CNV and PA resolution

Most KIR genotyping data, both for personal medicine and population studies, is reported as the presence/absence or copy number variation (Figure 2D-E) of each gene. Giving these genotypes, we developed algorithms to predict structural haplotype pair structures at the personal level and haplotype frequencies at the population level (Table 1, lines 1 and 2) on a large and diverse transplantation cohort of over 10,000 individuals in 5 broad populations. The results revealed patterns of structural linkage disequilibrium, some of which were population-specific and some of which were common to all populations.

We also developed the first algorithm to genotype KIR from any WGS (Table 1, line 6), and we applied it to the entire cohort of 748 individuals from the Genomes of the Netherlands project. These results, along with others from an independent group, suggest the algorithm is >= 97% accurate.

1.4.4. SNP resolution

Our multiple sequence alignment of KIR haplotypes allowed us to discover markers (or lack thereof) for structural haplotypes (Table 1, line 6). In addition, we evaluated the only open software for predicting genes and structural haplotypes from SNPs (Figure 2F). The analysis shows some weak spots of the SNP genotyping algorithm.

1.5. Overview

The following sections present the research in a multi-resolution context. Sequence haplotyping is described in section 2. This is the most important single contribution, since it is the first to provide an efficient approach to obtain, assemble, and annotate full KIR haplotypes. It also provides the information base for other contributions: the first description of the structural nature of KIR genes, loci, and haplotypes, as well as the first MSA of all reference haplotype sequences. In section 3, the topic switches to interpretating CNV and PA resolutions from non-high-throughput sequencing and describes how these genotypes are used to infer haplotype structures in the context of five populations. Section 4 describes the first genotyping algorithm for high-throughput- and/or WGS-sequences. It also makes some important observation about the only SNP-based KIR genotyping assay. Finally, section 5 draws some conclusions from the overall work and speculates on possible future contributions.

2. Long-read sequencing, assembly, and application of full haplotypes

2.1. Current state of sequence haplotyping

Interpretation of genic-or-lower resolutions from high-throughput sequencing is often ambiguous and unphased, and therefore limits precise understanding of how KIR sequences affect phenotypes. The importance of high-throughput high-resolution typing is exemplified by the fact that the genes contain extensive exonic SNP and short insertion/deletion (indel) variations which rivals that of its binding partner: HLA class I(1)(22). Over 300 full-length DNA and almost 1000 protein reference alleles have been reported in IPD-KIR(23). All resolutions except haplotyping are ambiguous or require statistical phasing from few references. A cost-effective high-throughput method that could characterize all the sequences within the KIR haplotypes in cis could advance that understanding and clarify previously ambiguous and/or contradictory evidence. To date, the only approach for full haplotyping was to physically separate and amplify maternal and paternal haplotypes via fosmids for subsequent sequencing(24)(15)(19)(14)(25), a process whose expense has generally prohibited its use in large-scale association studies. While high-resolution haplotyping by fosmid clones or full gene by PCR is costly and inefficient, low resolution genotyping of gene presence/absence or copy number provide limited information for functional analysis and association tests.

2.2. A novel sequencing, assembly, and annotation workflow

We developed an approach that leverages PacBio's long-read circular consensus sequence (CCS) reads to span DNA homology, and gene homology to efficiently capture 2-8 kb fragments of DNA. It is a workflow to capture, sequence, assemble, and annotate diploid human KIR haplotypes. And it also has broader implications to other genomic regions with variable or repetitive regions alternating with constant regions. When applied to a cohort of 8 African Americans and a cohort of 8 Europeans, the results demonstrate that every KIR gene and intergene contains constant regions that are targetable by capture probes, and that by targeting the constant regions, the variable regions can be captured and sequenced by standard PacBio workflows. Further, maximizing this paradigm shows that 18 short probe sequences can capture KIR haplotypes and allow their unambiguous assembly. Finally, this is an efficient approach that requires no prior knowledge of the individual or references, only utilizes standard lab workflows, and is available in free and open software.

The goal of the experiment was to create a set of capture probes and a bioinformatics workflow to efficiently assemble full KIR haplotypes from PacBio CCS reads. The experiments are depicted graphically in Figure 3.

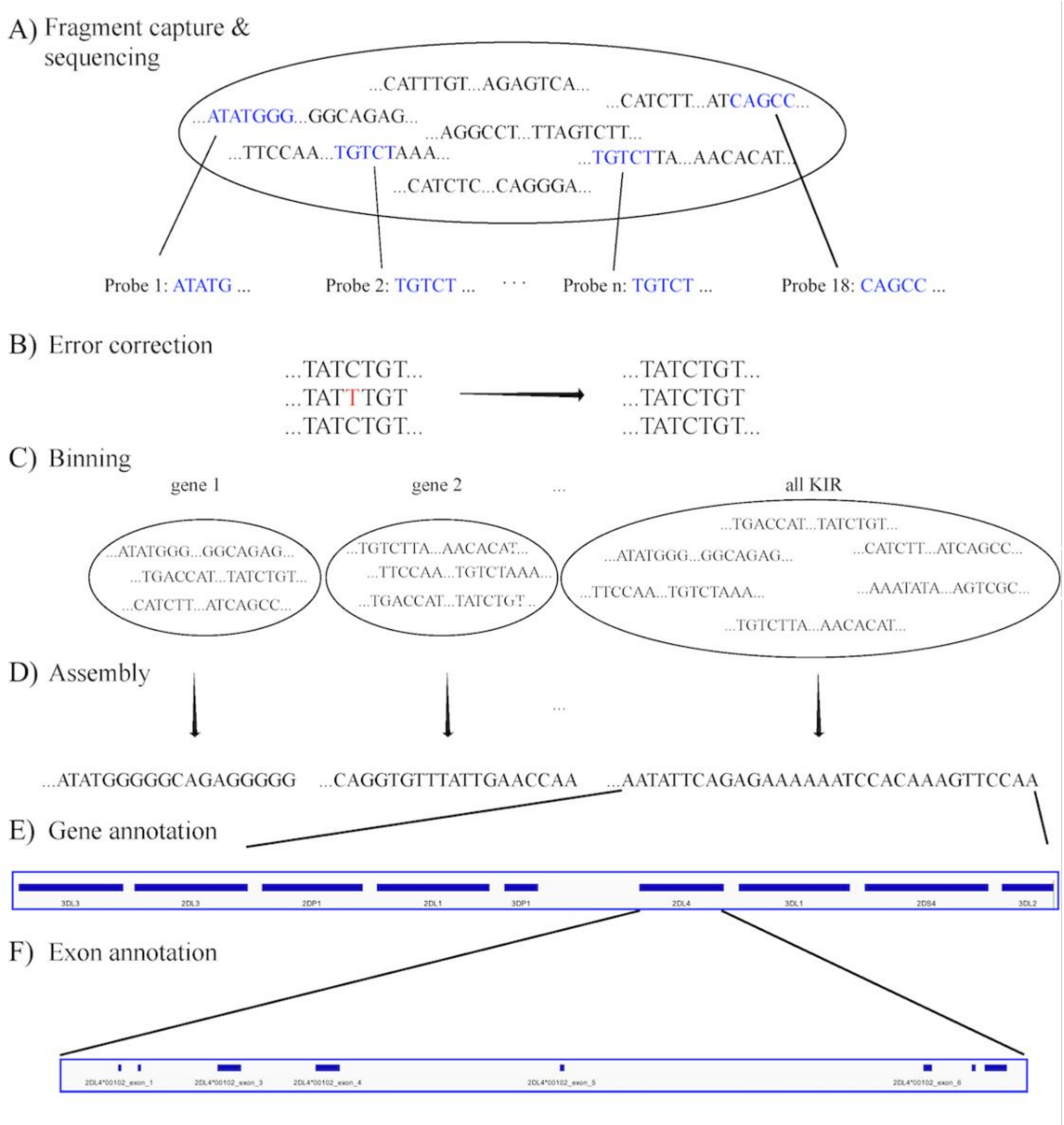

Figure 3. Assembly workflow. The workflow starts with fragment capture *in vitro* or *in silico* and PacBio assembly (A). The sequences are error corrected (B) and binned by KIR region and KIR gene (C) before *de novo* assembly of each bin (D). Finally, the assembled haplotigs/haplotypes are annotated by gene (E) and exon (F).

2.2.1. Workflow

Step 1: Design capture probes

Published KIR haplotypes sequences (36 at the time of this study) as well as all allele sequences from IPD-KIR 2.7.1(11) were used to generate 200 candidate capture probes. The design of the 120-base probes was coordinated with a combination of automated and manual Integrated DNA Technologies (IDT) design tools, including a strain typer alignment tool and a xGen® Lockdown probe design tool. The candidate set of probes was reduced by leveraging sequence homology. First, the haplotype sequences were aligned to two KIR haplotypes (GenBank accessions GU182358 and GU182339). These two reference haplotypes, which together contain all KIR genes, were annotated via RepeatMasker(26)(27). The candidates were prioritized by the highest number of times each aligned to the two reference sequences but not to repetitive elements. The set was chosen by iteratively adding the probe with the highest alignment hit count until both reference haplotypes were covered by less than the expected average DNA fragment length of ~4-5 kb. Ultimately, experiments were conducted on the original set of 200, a minimal set of 15, and a refined set of 18 capture probes.

Steps 2-3: Capture and sequence the DNA fragments per individual

Targeted hybridization probe capture and sequencing was performed (Figure 3A) as described using PacBio unsupported protocol PN101-388-000(39) with the following modifications. Two μg Human Genomic DNA suspended in 200 μL if Elution Buffer was sheered with Covaris G-tubes to 6-8 kb according to manufactures instructions followed by 1:1 PB AMPure bead cleanup. The target fragment size of 6-8 kb was chosen to maximize the ability for capture and to allow proper phasing and the generation of CCS reads, or a consensus sequence of one captured DNA fragment per sequence well. Individual specimens were then library prepped with the KAPA library prep kit (Roche) which consisted of end repair and ligation of uniquely barcoded adapters that also contained the PacBio Universal Primer sequence. After a 0.8X PB AMPure bead clean up samples were enriched with eight PCR cycles in a 200 μL reaction using LA Taq by Takara and PacBio Universal primers followed by a 1:2 PB AMPure bead cleanup. Sample concentrations were measured on the Promega Quantus and 2 μg was size selected for greater than 2 kb fragments on the Blue Pippin System (Sage Sciences).

Eight multiplexed samples for Sequel sequencing were pooled at this point at 0.25 μg per sample and 1.5 μg of pooled size selected DNA, 5 μL of 1 mg/mL Human Cot-1 DNA by Thermo Fisher, and 10 μL of 100 μM PacBio Universal Primer were dried in a Speed Vac with no heat. IDT xGen® lockdown reagents and probes were used to resuspend the genomic/Cot-1/primer mixture according to manufactures instructions and incubated at 70°C for 4 hours followed by IDT xGEN washes®. DNA was then removed from streptavidin beads and further enriched with fifteen PCR cycles using LA Taq. DNA fragments were then library prepped and sequenced eight samples per SMRT cell on the Sequel according to PacBio instructions. Raw sequence data was then demultiplexed and CCS reads generated on SmrtLink 6.0 using 99.9 % subread accuracy filter for generation.

Steps 4-6: Correct, bin, and assemble the sequences

For both targeted and WGS, the fastq sequences were error corrected with LoRMA(28). The sequences were binned for on-KIR and also binned per genic or intergenic region in silico (Figure 3B-C) using the 18 capture probes for on-off KIR detection and 32,230 gene probes. The gene probes are 25mers and are detailed in the KPI manuscript(29). Synthetic probe matching was conducted via bbduk(30) with parameters 'k=25 maskmiddle=f overwrite=t rename=t nzo=t rcomp=t ignorebadquality=t'. This effectively removed any off-KIR sequences and binned the sequences into 15 loci: 12 protein-coding genes, 2 pseudo-genes, and the intergenic region between KIR3DP1 and KIR2DL4. Sequences in each bin were de novo assembled with Canu 2.0(31) (with default parameters except 'genomeSize=200k') for each bin separately and all KIR sequences together (Figure 3D). The assemblies utilized only the captured sequences and were not assisted by any prior information, including individual genotypes or reference libraries.

Step 7: Annotate the assembled sequences

The capture probes were aligned to the haplotype-specific assembled sequences (i.e., haplotigs), and their patterns allowed gene-specific sequences to be extracted from the haplotypes (Figure 3E). The details are presented in section 2.3 and the MSA manuscript (Table 1, line 5). At a high level, the algorithm uses the bowtie2 alignment pattern of the 18 capture probes across the haplotigs/haplotypes to define locus-specific features. Within each feature, the locations of the exons, introns, and untranslated elements were located by searching for inter-element boundaries with 16 base sequences as defined by the full sequence haplotype MSA of 68 human haplotypes (Figure 3F). Each sequence contains 8 bases from one region and 8 from the other. For example, ACACGTGGGTGAGTCC spans the boundary between KIR2DL4's exon two and intron two; the first eight characters are from the second exon, and the last eight bases from the second intron. This matching is flexible up to 3 mismatches if necessary. The locations of these elements allows for the annotation of protein, cDNA, and full-gene alleles with respect to names assigned in IPD-KIR. The haplotigs were ultimately annotated in GenBank's tbl format. BioJava(32) was used for some of the sequence processing. Reports on the assembly (or the raw sequences) were generated from Minimap2(33), Qualimap(34), NanoPack(35), QUAST(36), Simple Synteny(37), and Tablet(38).

2.2.2. Evaluation of the targeted workflow

In our 2017 publication (Table 1, line 3), we have demonstrated the ability to overcome these challenges in the sequencing of the entire KIR region by comprehensively sequencing haplotypes for eight diploid individuals using a combination of fosmid-based library preparation along and PacBio long-read sequencing. The PacBio RS II system was used for sequencing fosmid libraries. These libraries were constructed and screened to obtain complete coverage of both haplotypes for each person. New SMRTbell library preparation approaches were developed for full- length fosmid sequencing to ensure long-read lengths, which provided unambiguous full haplotype coverage. This was the first demonstration of sequence haplotyping on diploids; previous efforts utilized human haploid cell lines. In 2020, we documented (Table 1, line 5) a follow-up project that applied the same fosmid-PacBio approach in 8 African Americans. Along with the 8 Europeans in the first project, this provided 16 diverse and high-quality reference haplotype sequences.

These reference sequences were used as ground truth for the assemblies of those same 16 individuals (32 haplotypes) representing 11 distinct haplotype structures. These reference sequences are shown in Figure 4, which depicts the 11 structures as connections between the same genes in different haplotypes and shows how the structures represent expansion and contraction of the A and B haplotype categories across the *KIR3DP1-KIR2DL4* hotspot. Table 2 shows the results of the assembly compared with the reference sequences. For the 8 Europeans on average, the full set of 200 candidate probes provided 98% coverage, with 99.98% concordance, and it took 1.1 haplotigs to cover 75% of the reference (LG75). When a set of capture probes was reduced to a select 15 and evaluated on both cohorts, the European coverage lowered to 93%, with the same concordance rate (99.98%), and 1.3 LG75. The results for African Americans were very similar: 92% coverage, 99.98% concordance, and 1.6 LG75. When a select 3 more probes were added for a total of 18 capture probes, the assemblies for the 8 African Americans improved to 97% coverage, with 99.97% concordance, and 1.8 LG75. Most of the missing coverage occurred at the 3' end of the haplotypes: in certain *KIR3DL2* alleles and some sequences extending 3' past *KIR3DL2*. Figure 5 shows the alignment of the haplotigs relative to the reference haplotype from the same individual

MN167513 (cA01~tB01, Figure 4) in the 18-probe experiment. It shows a small < 2 kb gap in the assembly in *KIR2DL3*. Otherwise, the haplotigs provide complete and overlapping coverage across the reference haplotype sequence. When all the haplotigs are aligned the reference, the statistics report that it takes 2 haplotigs to assembly 75% of the haplotype and that total coverage is 98.4% with 99.98% concordance with the reference. The haplotigs are colored by base (ACGT) and indicate the haplotigs are concordant. Every gene is spanned by at least one haplotig, and all loci are phased with overlapping haplotigs, which the exception of the gap in *KIR2DL3*.

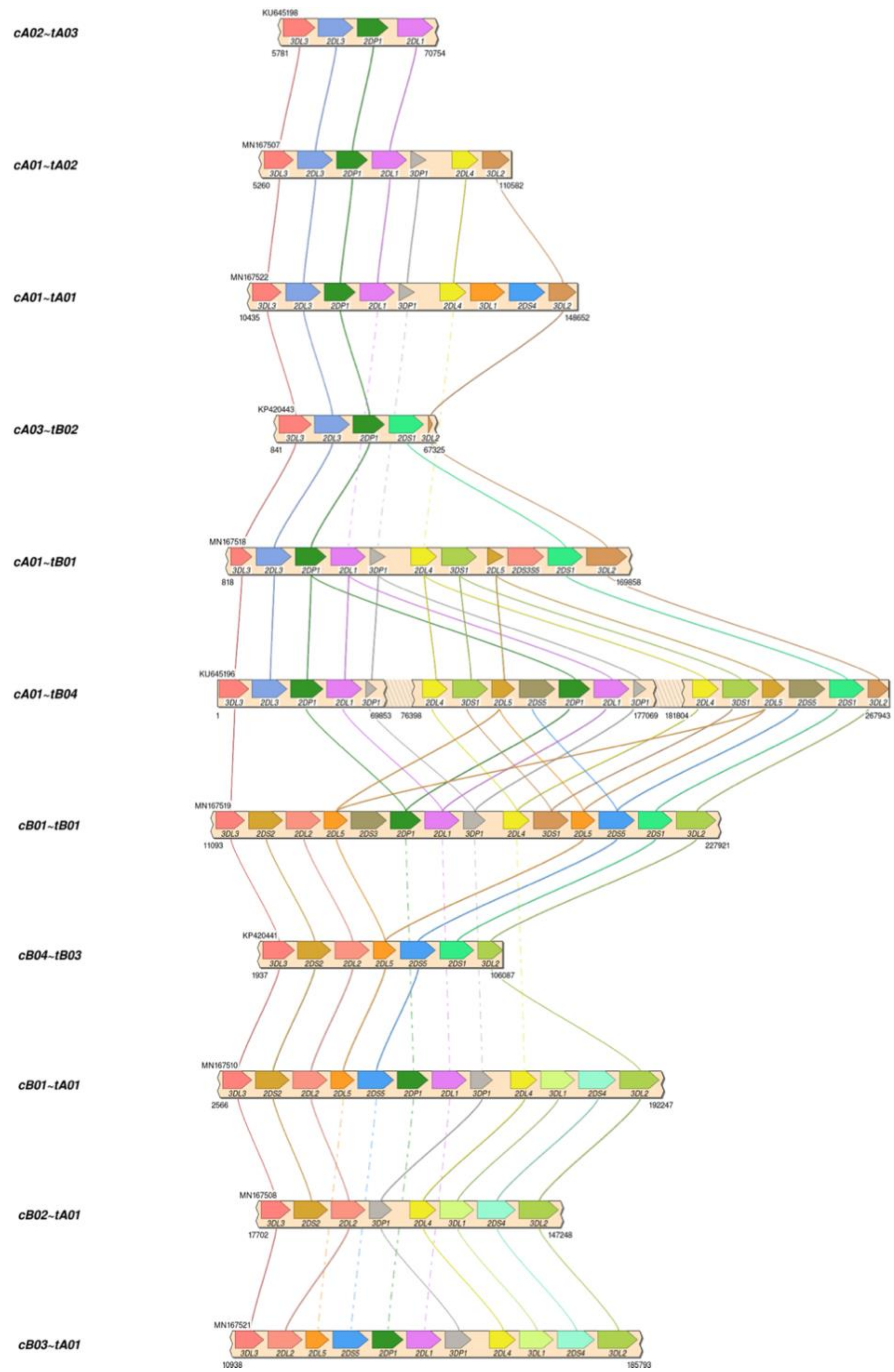

Figure 4. Homologous reference haplotype structures. Each haplotype represents one of the structures previously established via fosmid library preparation and long-read sequencing. The unofficial name of the haplotype is on the left. Lines connect genes with the same name in different structures. Solid lines

connect the same gene in neighboring structures. Dashed lines connect the same gene in non-neighboring structures (i.e., the line goes through one or more neighboring haplotypes).

pop.	probes #	LG75	coverage %	concordance %
EUR	200	1.1	98%	99.98%
EUR	15	1.3	93%	99.98%
AFA	15	1.6	92%	99.98%
AFA	**18**	**1.8**	**97%**	**99.97%**

Table 2. Assembly statistics. For each of experiments in the rows, shown is the population, number of capture probes, the number of haplotigs spanning 75% of the haplotype (LG75), and the coverage of and concordance to the reference.

Figure 5. Assembly haplotigs example. Alignment of assembled haplotigs with reference haplotype sequence MN167513 (cA01~tA01), whose length is 147,345. The gene features are annotated across the top. The haplotigs are stacked below and colored by nucleotide.

2.2.3. Evaluation of WGS workflow

Theoretically, if KIR reads could be removed from WGS, the workflow should be able to assemble haplotypes the same as from targeted sequences. To test this hypothesis, whole genome CCS reads were obtained for an Ashkenazim individual (isolate NA24385) from the Genome In a Bottle (GIAB) consortium, as described in Wenger et al.(39). KIR ground truth was unknown previously. KIR reads were separated from WGS as described above and from there the workflow proceeded as usual from the error correcting step (Figure 3B).

Using the 18 sequences as virtual probes to capture KIR reads from WGS, the KIR assembled into a paternal cA01~tB01 (KIR3DL3*00101~KIR2DL3*00101~KIR2DP1*NEW~KIR2DL1*00302~KIR3DP1*0030202~ KIR2DL4*0050101~KIR3DS1*01301~KIR2DL5A*00101~KIR2DS5*00201~KIR2DS1*00201

~KIR3DL2*00701) haplotype and a maternal cB05~tB01 (KIR3DL3*00301~KIR2DS2*005~KIR2DP1*NEW~KIR2DL1*0040105~KIR3DP1*0030202~KIR2DL4*00501~KIR3DS1*01301~KIR2DL5*00101~KIR2DS5*00201~KIR2DS1*00201~KIR3DL2*NEW). The distal/telomeric halves (*KIR3DP1-KIR2DS1*) are mostly homozygous, with approximately a dozen variants between them. It is possible this region is really deleted in maternal haplotype and should be classified as cB05~tB02 (*KIR3DL3~KIR2DS2/KIR2DS3~KIR2DP1~KIR2DL1/KIR2DS2~KIR3DL2*). Either way, this is the first reference haplotype with a cB05 with tB01 or tB02 in the same haplotype; cB05 has *KIR2DS2/KIR2DS3* fusion (named KIR2DS2*005 in IPD-KIR). KPI(29) confirms cB05~tB01 with cA01~tB01 (or their deleted forms) as the most-likely pair of structural haplotypes. A recently published PacBio full-genome assembly of this individual(39) assembled the cA01~tB01 in paternal haplotig SRHB01000968.1, but the maternal haplotigs do not contain any KIR haplotypes.

2.2.4. Assembly conclusions and discussion

The power of this method to assemble repetitive KIR regions without incorporating false non-KIR genomic signals may lie in the strongest recombination hotspot, the 10 kb intervening region between *KIR3DP1* and *KIR2DL4*. Conventionally, KIR haplotype names (e.g., 'cA01~tA01') have been described as two halves ('c' and 't') separated by a recombination hotspot ('~'). The rate of recombination between the two halves is so frequent that any two may be found with each other, despite the relative evolutionary youth of the region. This hotspot stretches over 9 kb between *KIR3DP1* and *KIR2DL4*. Any two alleles of this region are over 99% identical and consist of 13% Alus (SINEs) and 58% LINE1 repeat elements. Figure 6 shows an alignment of the region to itself. The top part of the figure displays the location of the *KIR3DP1-KIR2DL4* intergenic region in the context of the cA01~tA01 primary human genome reference. The bottom half zooms into the intergenic region and shows a dot plot of the alignment. The lines on the dot plot indicate stretches of the haplotype that align with itself, either in the same location or a different location. The red lines indicate matching in the same orientation as the overall haplotype, and the blue indicate matching in the reverse complement. The red and blue horizontal bars at the bottom of the figure detail the location of repetitive elements. The red from 0-2000 simply shows that this region aligns with itself. There is a stretch of 3 kb from ~5200-8200 that reverse complement matches ~400-3400. The yellow boxes highlight Alu repeats: AluSx3 from ~800-1000 is matched in the reverse complement by AluSx1 (~6400-6600) and AluSx4 (~7100-7400) and the same orientation by AluSq2 (~8200-8400). All elements are surrounded by very similar L1 elements for at least 1000 bp on both sides. This stretch of 3000+ bases of reverse complement repeats provides fertile ground for homologous recombination between the two halves of KIR haplotypes. This is the most difficult region to phase and the results demonstrate that the combination of variants and read length is generally high enough to phase full sequence haplotypes with diploid reads. Although this region is an extreme example, the other recombination sites, which are usually internal to genes and result in gene fusions, homologously recombine via the same elements(12)(40)(20)(41)(42), although only the *KIR2DL4-KIR3DP1* intergenic region contains as many elements in both directions(24). Our assay is the only high-throughput method that allows analysis of all of these regions.

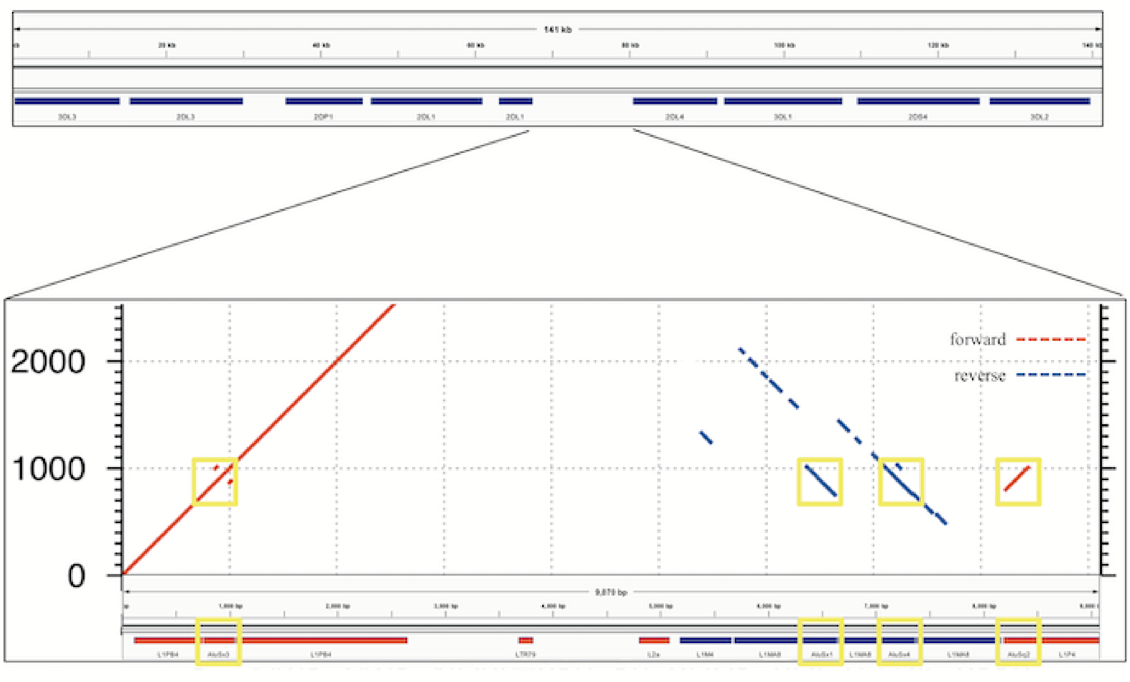

Figure 6. Recombination hotspot in assembly. The top shows the 9 kb haplotype context of the repetitive region of the *KIR3DP1-KIR2DL4* intergene region, whose self-alignment is shown in the bottom half. Red lines indicate alignment of the haplotype with itself in the same orientation, and blue lines indicate reverse complement orientation. The location of the repetitive elements is at the bottom, with red and blue again indicating orientation. The yellow boxes highlight three AluSx and one AluSq elements that align with each other, two in each direction.

These experiments in individuals from diverse populations demonstrate that KIR haplotypes can be efficiently enriched and ultimately assembled using an efficient number of capture probes. The workflow successfully reconstructed both haplotypes from targeted sequencing in 16 individuals and from WGS in 1 individual. Although recent advances in Single-Molecule Real-Time sequencing by PacBio have improved quality and extended its applicability to WGS and highly repetitive regions, these advancements are not sufficient to accurately assemble KIR haplotypes. In our evaluations of self-reported results from other published and pre-published assemblers, we could not find another assembler that correctly assembled diploid KIR haplotypes from PacBio reads alone. The sequences need to be error corrected, separated from off-KIR, and sometimes the reads need to be separated and assembled on a per-gene basis depending on depth and genotypic variation, as more binning helps overcome the challenges of higher multiplexing. We have confirmed successful assembly with multiplexing up to eight individuals, which, we estimate, should lead to costs that rival full-exon short-read sequencing, and an order-of-magnitude more efficient than fosmid-based library preparation due to a more high-throughput library preparation workflow. We report efficiency details as a minimum, because the purpose of these experiments were to demonstrate the capability, not maximum efficiency. Our results suggested that lower bounds efficiency for off/on ratio for KIR capture range between 1.3 to 2.4.

In addition to assembly, our software is the first annotation system for KIR haplotypes. The annotation algorithm leverages the information of the pattern of capture probes across the haplotype sequence to define loci and their exon/intron locations.

In addition to demonstrating an efficient targeted haplotyping strategy, to the best of our knowledge, this the first report of KIR full diploid haplotype assembly from HiFi WGS alone. Our approach was able to assemble both haplotypes from WGS whereas the previously reported whole-genome assembly could not, underlying the necessity of a KIR-specific assembler. Both regions comprising the haplotype it missed (cB05~tB01 or a deleted form) are not in the primary human genome reference, and the two have not been reported together previously. Perhaps this lack of representation in the reference contributed to the missing assembly. Other possibilities include the lack of binning/separation of KIR reads from the rest of the genome before assembling, or differences in the tools used in the workflow. Regardless, this experiment demonstrates the value added by the bioinformatics algorithms, in addition to the targeted capture and assembly.

Our approach leverages sequence similarity across multiple loci that were created by duplication followed by variation. Since this this is a true for many gene families, our approach should be more generally applicable to other regions that have a mix of homologous and variable/repetitive regions relative fragment length and capture characteristics.

The application of KIR genetics in medical research such as immunity, reproduction, and transplantation is encouraging, but limited by the technical difficulties for high-resolution interpretations at large scale and low cost. Here, a KIR haplotyping workflow was presented that can provide full-sequence haplotypes at approximately the same cost as full exon or full gene. For the first time, it allows high-resolution KIR haplotypes in population-sized cohorts, as opposed to lower-resolution genotypes. The analysis pipeline uses domain knowledge to assemble reads generated via well-established sequencing techniques that is accurate enough for personalized precision medicine and scalable to populations. To this point, most KIR association studies focus on variation at only one locus or one functional class to associate, while keeping the rest of the haplotypes static. Future full-haplotype studies will help KIR researchers better study gene combinations, regulatory regions, recombination hotspots, self-regulation, and non-binding factors that influence disease phenotypes. This increased ability will provide completed sets of population-specific reference haplotypes which will, among other things, enhance imputation power of lower resolution data. It allows for new comparisons that will provide insight into evolution and make this region the best annotated in the human genome, despite its complexity. Lastly, this novel approach will provide the capability to discover genetic associations in medically relevant areas such as infections, transplantation, cancer susceptibility, autoimmune diseases, reproductive conditions, and immunotherapy. The open and free software is available at https://github.com/droeatumn/kass and supported by a environment at https://hub.docker.com/repository/docker/droeatumn/kass.

2.3. A unified framework describing and classifying genes and haplotypes

2.3.1. Current state of multiple sequence alignments

The Immuno Polymorphism Database for KIR (IPD-KIR) names KIR gene alleles, records their DNA-RNA-protein relationships, and annotates each gene’s alleles in a multiple sequence alignment(11). It contains over 300 full-length DNA and almost 1000 protein reference alleles in the latest release, 2.9.0.

There is no equivalent of IPD's allele annotation for haplotypes. KIR haplotypes have no official nomenclature, although the majority of contributions to the human genome reference use a convention set by Pyo et al. in 2010(15) and 2013(19). The haplotype names reflect the two-part structure of the region: a proximal centromeric ('c') region is paired with ('~') a distal telomeric ('t') region. Each region is a variant of a 'A' haplotype or 'B' haplotype family, followed by a two-digit index. The haplotype named 'cB02~tA01', for example, is comprised of the second centromeric B region ('cB02') in cis with ('~') the first telomeric A region ('tA01').

Although there are many publications that analyze gene or intergene allele alignments, none report full haplotype multiple sequence alignments (MSAs), and therefore the haplotype nomenclature has not been formalized. Allele gene assignments are evaluated largely by amplification primers or sequence similarly to other alleles, as opposed to location in its haplotype. To help solve these issues, we present a simple bioinformatics approach to represent KIR haplotypes as a string of alignment-based motifs. We applied it by creating a full-haplotype DNA MSA for all 68 human haplotype sequences in the human genome project and a chimp haplotype for an outlier.

2.3.2. A novel method to align all haplotypes

Source sequences and annotation

The source sequences consisted of two newly assembled haplotypes from an Ashkenazim individual (Table 1, line 4), along with all 66 full-sequence alternative haplotypes in the human genome reference(24)(15)(19)(14) and a chimpanzee haplotype (GenBank accession AC155174.2)(25). The two Ashkenazim haplotypes were assembled with PacBio reads obtained from GIAB(39) and have not yet been submitted to genetic databases. All other haplotypes were generated by physically separating chromosomes via fosmid cloning and then sequencing and assembling fragments into full sequence haplotypes. The 68 include 27 new sequences first described in this manuscript, including 24 haplotypes from 12 individuals of African American ancestry that were selected by genotypic diversity and/or lack of representation in the human genome reference. An additional 3 haplotypes (1 Asian, 2 European) of opportunity from two individuals were contributed by Scisco Genetics, who sequenced all non-GIAB haplotypes following previously reported protocols(15)(19). The GenBank entries also provided structural and allelic annotation of all the human haplotypes, except for the two Ashkenazim haplotypes. Broken down by population, the haplotype counts include 31 African or African American, 3 Asian, 2 Ashkenazim, 24 European, 2 Guarani South Amerindian, 2 Romani in Spain, and 4 unknowns. The structures are depicted schematically in Figure 7, excluding cA01~tB04, which contains a large insertion and was excluded for compact visualization.

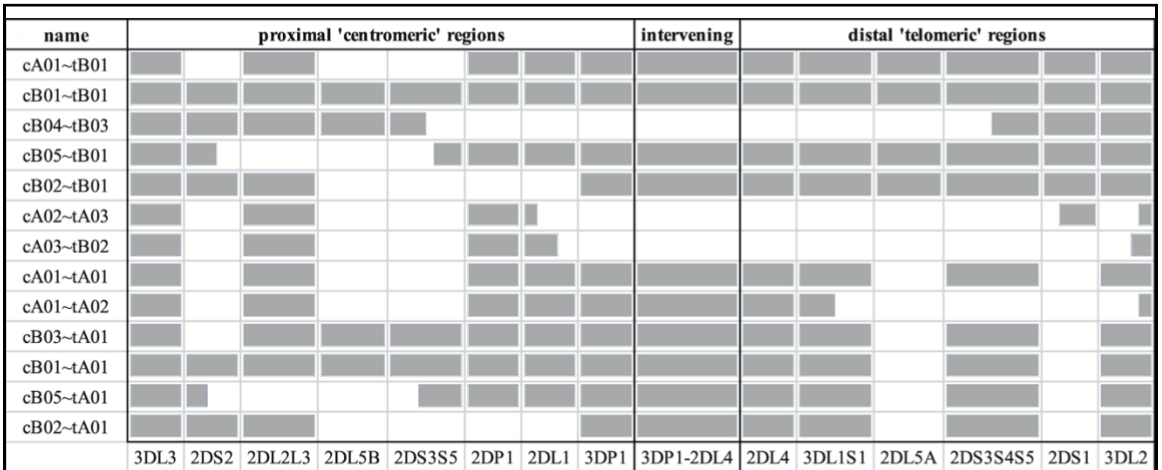

Figure 7. Human-annotated reference haplotype structures. Schematic representation of the informal names and definitions of 13 human haplotype structures. The informal names of the haplotypes are in the first column. The definitions of the centromeric and telomeric regions are in the top row. The abbreviated names of the genes are in the bottom row; the intergenic region between *KIR3DP1* and *KIR2DL4* is represented as *'3DP1-2DL4'*. Grey cells indicates the presence of alleles for that gene; white indicates absence of an allele. Some cells are partially colored to indicate two-gene fusion alleles. The names of the haplotypes, as well as the gene content and locations are taken from the human curated annotations in the human genome reference. Haplotype cA01~tB04 contains a large insertion and was excluded for visualization.

The result of the workflow was a multiple sequence alignment of 68 human and 1 chimp haplotypes. As shown in Figure 8, the first step was to align the 18 capture probes to the haplotype sequences (Figure 8A). Next, the sequence between each pair of probes (e.g., probe 4 to probe 3) was assigned a letter (e.g., C) (Figure 8B). In this way, each haplotype can be represented by a string ('motifs', e.g., '…CIJKL…') whose length is 105 characters or less. These haplotype motifs were aligned to produce a high-level MSA (Figure 8C). The motifs allowed loci to be located in the haplotype sequences. The sequences for all alleles of each locus was aligned separately, and then the MSAs were concatenated to create the full haplotype DNA MSA (Figure 8D). In this two-alignment approach, the motif alignment provides structural annotation, and the locus-specific DNA alignment provides base-level accuracy. The following subsections detail each step.

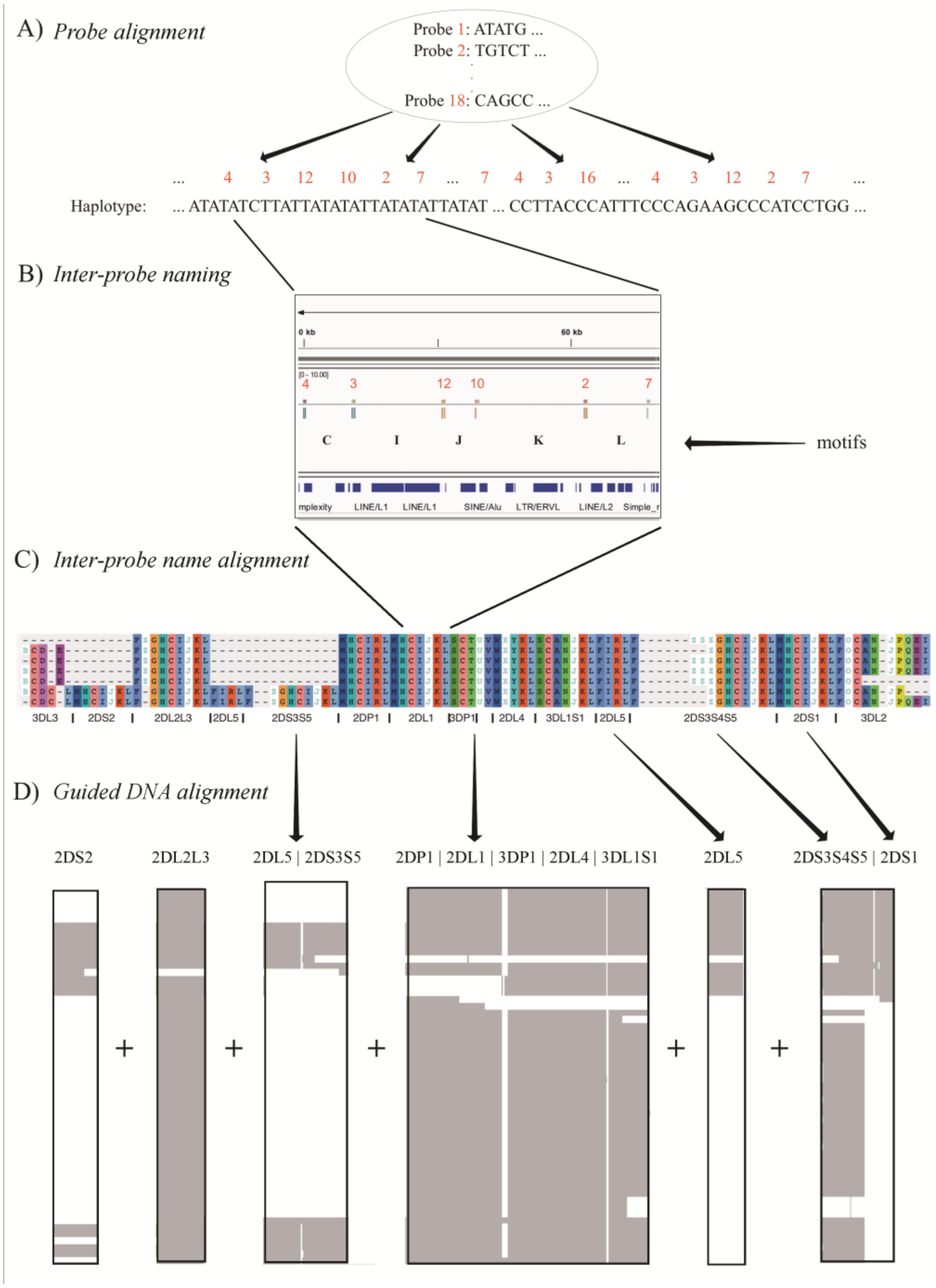

Figure 8. MSA workflow. Step (A) depicts alignment of 18 120 base probes to each haplotype sequence. Step (B) shows how each ordered pair of probes in the alignment is assigned a letter. e.g., the letter 'I' represents the sequence between the alignment of probe 3 and probe 12. Step (C) shows the haplotype motif MSA and how loci are defined in the MSA. Step (D) shows how the DNA in the motif-defined loci were separately aligned and then joined to create the final MSA.

Probe alignment and inter-probe naming

The assembler manuscript (Table 1, line 4) describes a method that uses 18 120 bp probes to capture KIR PacBio long-read sequences and assemble them into haplotypes with an average of 97% coverage and 99.7% concordance compared with reference sequences. When the probes are aligned to the haplotypes (Figure 8A), they align every 2398 bases on average. Probe locations are discovered by alignment with bowtie2 with options '-a –end-to-end –rdg 3,3 –rfg 3,3'. The alignment order of the 18 probes across a haplotype sequence allows that haplotype structure to be succinctly annotated as sequences of probe pairs (Figure 8B). For example, assume the alignment of probe 4 followed by probe 3 is called 'C', probe 3 followed by probe 12 is called 'I', and probe 12 followed by probe 10 is called 'J'. Then the region 'probe 4 to probe 3 to probe 12 to probe 10' can be called 'CIJ'. Probe 1 repeating once in the alignment is 'Z' and twice is 'ZZ', etc. There are 42 such probe pairs in this collection of 69 haplotypes. In this way, haplotypes can be briefly annotated as strings of a 42-character alphabet. See Figure 9 for an example of KP420440, whose full motif is MHCIJKLFGHCIJKLAIRLFZGHCIJKLMHCIRLMHCIJKLSCTUVWXYKLSCNOJKLAIRLFZGHCIJKLMHCIJKLFPCNOJQ. The probe sequences and motif pairs are defined in the Appendix.

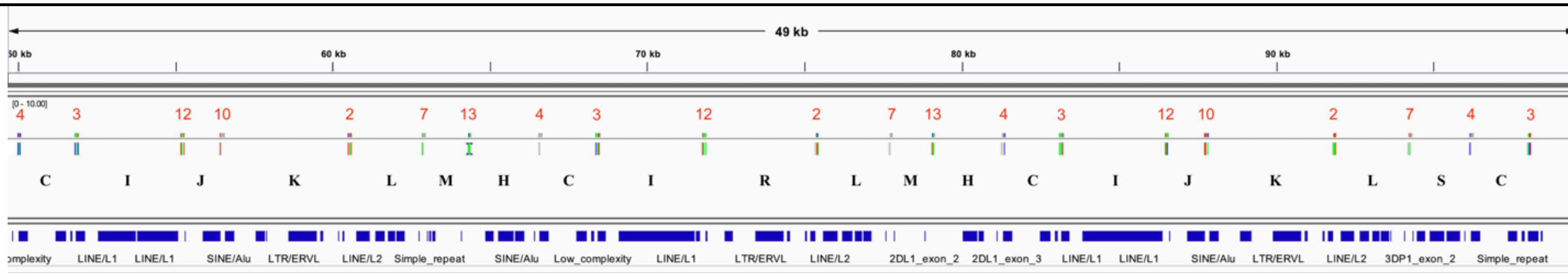

Figure 9. Example of the motif-generating alignment. IGV depiction of the alignment of the 18 set capture probes across a 49 kb region of KP420440 (cB01~tB01). In the middle track, the locations of the probes are displayed by the vertical ticks with the red numbers above them and black letters below them. The red numbers label each probe, and the black letters label each probe pair (i.e., motif). The locations of exons and repeat elements (horizontal blue bars) are on the bottom track. 7 distinct probes align in this window. The motif pattern CIJKL occurs in two locations: 50-63kb and 81-94kb; the variation CIRL occurs 66-79kb.

Inter-probe name alignment

A full-haplotype multiple sequence alignment of the probe motifs was generated for the 68 human haplotypes plus the chimpanzee as an outgroup (Figure 8C). Alignment of the motifs was created with MAFFT(43), but mostly aligned manually with Aliview 1.2.6(44); manual alignment was required to resolve ambiguities, as the motif alphabet is not directly supported by DNA or protein aligners. Ambiguous alignments were resolved by following the human curation of the reference haplotypes.

DNA alignment

MAFFT was used to merge or add these haplotypes in the order cB02~tA01, cA01~tA01, cA02~tA01, cB02~tB01, cA0X~tB0X, cB0X~tA01, and cB0X~tB0X, where 'X' a general number not already used. Then, the full-length DNA sequences were separated into sets as defined by their motif structures (Figure 8D). Each set was aligned separately with MAFFT. The alignment was then edited manually by adding or deleting gaps to conform to locations in the probe motif MSA. Using the loci defined in the motifs, and the motif locations defined in the DNA, the alignment was refined for all alleles in each locus with MUSCLE in Aliview. A high-level depiction of the alignment was created with Jalview 2.11.0's 'Overview Window' functionality and NCBI's Multiple Sequence Alignment Viewer 1.13.1.

The MSA was validated at the structural level by showing that it recapitulates the human curation of the reference haplotypes in GenBank and the human genome reference. The alignment was validated at the allele level by showing each allele is assigned and annotated as expected with respect to IPD-KIR classifications.

The ability of existing software to align the 69 haplotypes was also evaluated with MAFFT stand-alone (--thread 19 –threadtb 10 –threadit 0 –reorder –adjustdirection –anysymbol –leavegappyregion –kimura 1 –auto), PASTA v1.8.5 (docker run –rm -it -v $PWD:/opt/droe smirarab/pasta run_pasta.py), Clustal Omega v1.2.4 (clustalo –threads=29), M-Coffee metaserver 13.41.0.28, MUSCLE v3.8.31 (default parameters), and WebPrank (Updated 8 October, 2017). Implementation was conducted on the web servers when possible or a server running Ubuntu 18.04.4 LTS with 32 AMD OpteronI Processor 6220 and 200 GB RAM and maximum java heap space set to -Xmx100G.

2.4. A multiple sequence alignment of all haplotypes

Validation of the MSA by annotation in human genome reference

Figure 10 shows a multiple sequence alignment of the probe motifs for the human haplotypes, excluding insertion-containing cA01~tB04 for space considerations. It recapitulates the human-annotated structures in Figure 1 with 105 positions. The common ~140 kb cA01~tA01 haplotypes are marked up in ~63 motif characters, and the ~220 kb cB01~tB01 haplotypes are marked up in ~94 motif characters. The average distance between probes is 2398 bases; the maximum distance for non-*KIR3DL3* genes is ~5700 bases and for *KIR3DL3* it is ~7800 bases. The haplotype motifs subdivide into 14 genic and 1 intergenic (*KIR3DP1-KIR2DL4*) loci. From 5' (left) to 3' (right) at the bottom of Figure 5, the 15 abbreviated loci are: 3DL3, 2DS2, 2DL2L3, 2DL5, 2DS3S5, 2DP1, 2DL1, 3DP1, 3DP1-2DL4, 2DL4, 3DL1S1, 2DL5, 2DS3S4S5, 2DS1, and 3DL2. The 14 genic loci consist of 9 distinct motifs: 2DL1, 2DS1, 2DS2 share MHCIJ; 2DL5A and 2DL5B share FIRL; 2DS3, 2DS4, and 2DS5 share FZ+GHCIJKL. The genes can be summarized as a short motif, or in some cases a regular expression. For example, FZ+GHCIJKL means: a F followed by one-or-more Zs followed by GHCIJKL. The chimp haplotype contains 55 motif characters, encoding genes that align to the human *KIR3DL3~KIR2DS2~KIR2DP1~KIR2DL1~KIR3DP1~KIR2DL4~KIR3DL1S1~KIR3DL2*.

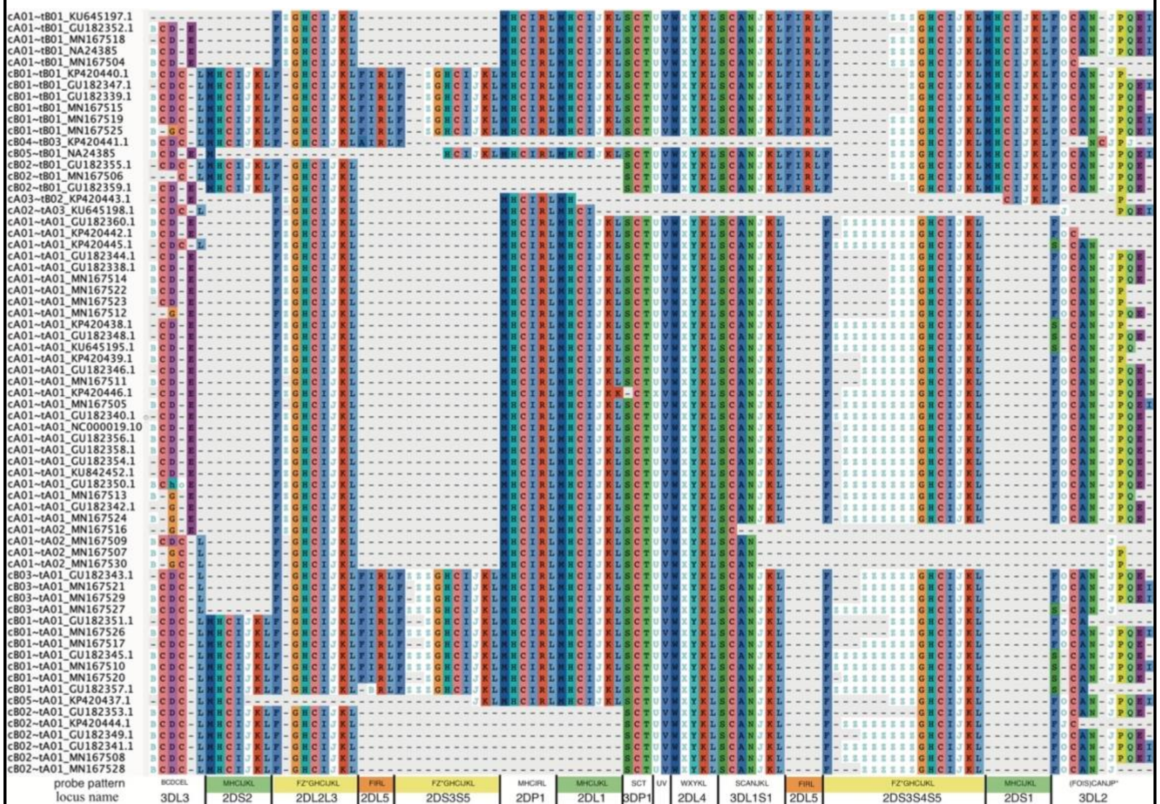

Figure 10. Probe motif multiple sequence alignment of the human haplotypes. The informal haplotype name and GenBank accession number are in the first column. The consensus motif pattern with each abbreviated locus name is described in the bottom row. Some loci share motifs: *KIR2DL1*, *KIR2DS1*, *KIR2DS2* (green); *KIR2DL5A* and *KIR2DL5B* (orange); *KIR2DS3*, *KIR2DS4*, and *KIR2DS5* (yellow). Some gene names are combined when they align at the same locus: *KIR2DL2* and *KIR2DL3* (2DL2L3), *KIR2DS3* and *KIR2DS5* centromeric side (2DS3S5), *KIR2DS3*, and *KIR2DS4*, and *KIR2DS5* telomeric side (2DS3S4S5), and *KIR3DL1* and *KIR3DS1* (3DL1S1). Haplotype cA01~tB04, which contains a 106 kb insertion, was omitted for visualization.

Haplotype sizes by number of loci (not including the *KIR3DP1-KIR2DL4* intergenic region) range from 4 loci (cA04~tA03) to 18 (cA01~tB04). Figure 11 displays a phylogenetic tree from the motifs of the 69 haplotypes, including cA01~tB04 and the chimp haplotype; the clusters recapitulate their human-annotated names shown on the right part of the figure.

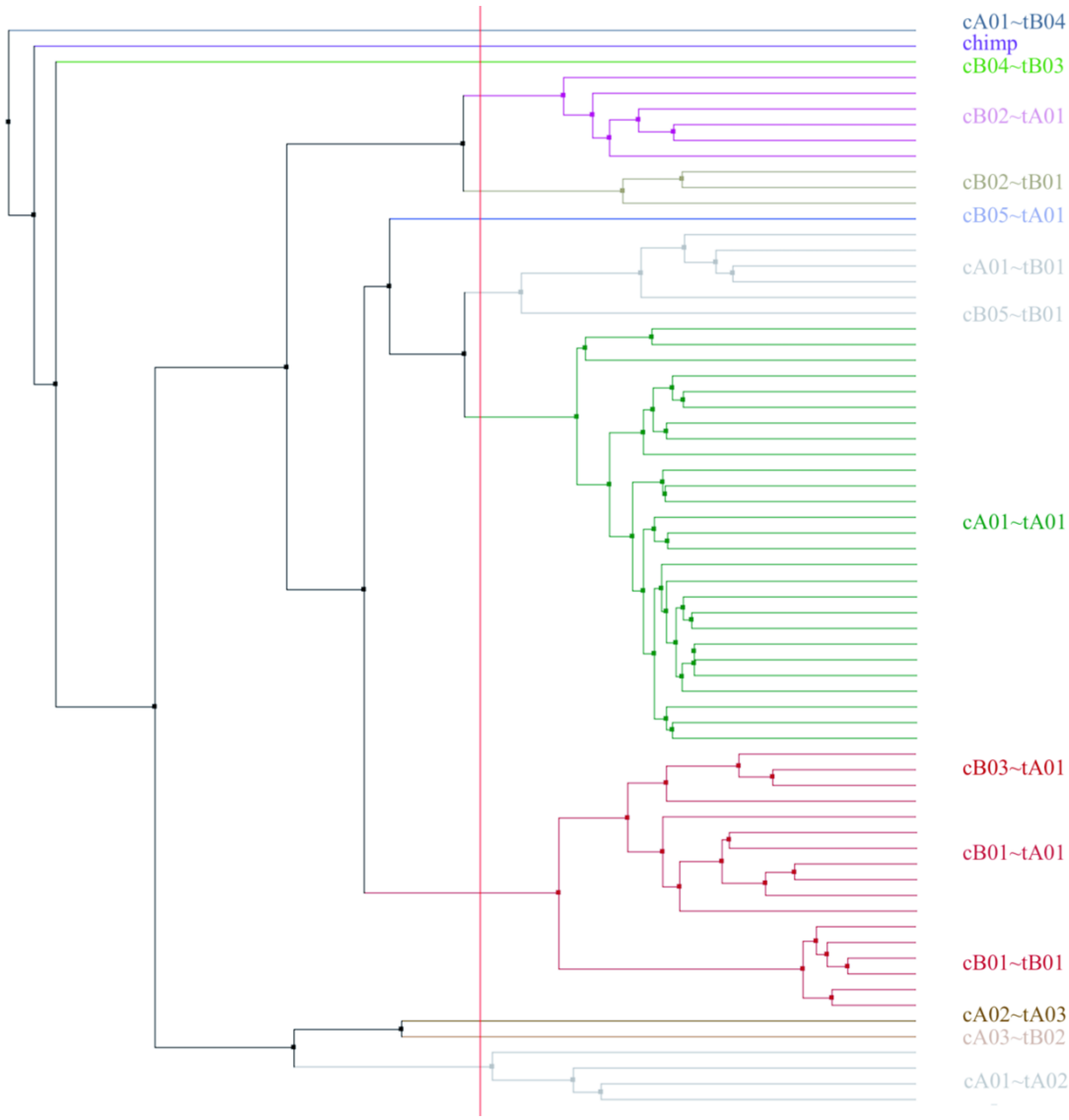

Figure 11. Phylogenetic tree of human haplotype sequences. The tree is made from the MSA from the motifs of all the human haplotypes plus the chimpanzee (n=69). The leaf descriptions on the right include the informal name of the structural haplotype as documented in GenBank. The leaves are colored by their alignment order and the grouping defined by the red vertical line. The tree may not reflect the evolutionary history.

The DNA alignment of the 67 human haplotypes (minus cA01~tB04) has 263556 positions. Figure 12B shows an overview of the DNA haplotype alignment as generated by Jalview's 'Overview Window' function; the names of the haplotypes were added after exporting the overview image from Jalview. The pattern of white (gaps) and grey (aligned sequences) mirror the patterns in the schematic depictions of the human curated haplotypes in Figure 7 (recapitulated in Figure 12A). This evidence shows the DNA alignment matches the human curation at a structural level.

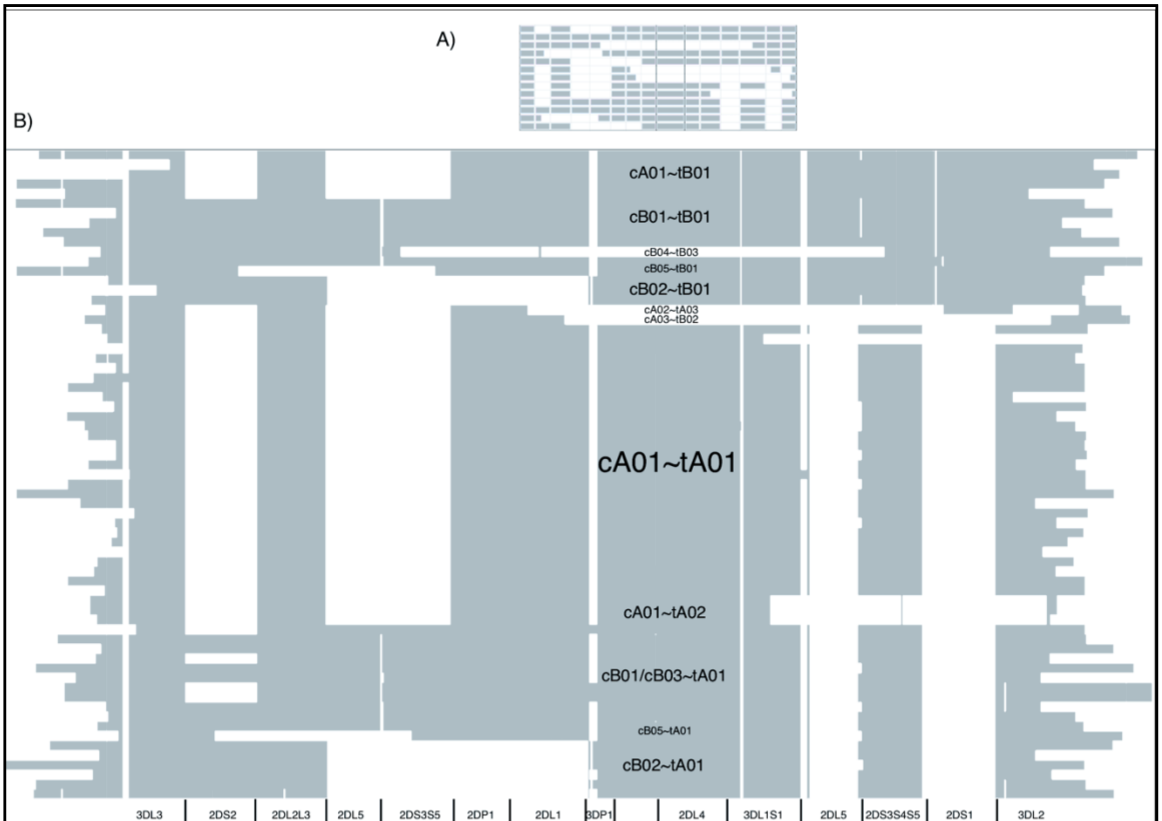

Figure 12. DNA MSA overview. The cartoon of the GenBank annotations from Figure 1 is in (A) and is included to compare with the Jalview overview of the DNA full-haplotype MSA from the 67 haplotypes in (B). The grey denotes alignment of the DNA bases. White denotes gaps between the sequences. cA01~tB04 is excluded for better visualization.

Figure 13 shows a view of the alignment of the 67 haplotypes using NCBI's MSA Viewer. Positions that are in an agreement with consensus are colored in gray, positions that are not in an agreement with consensus are colored in red. In this tool, consensus includes gaps (lack of an allele at that position). For KIR, this causes the regions specific to B haplotypes to show as pure red, and the other regions to show as mixture of red and grey.

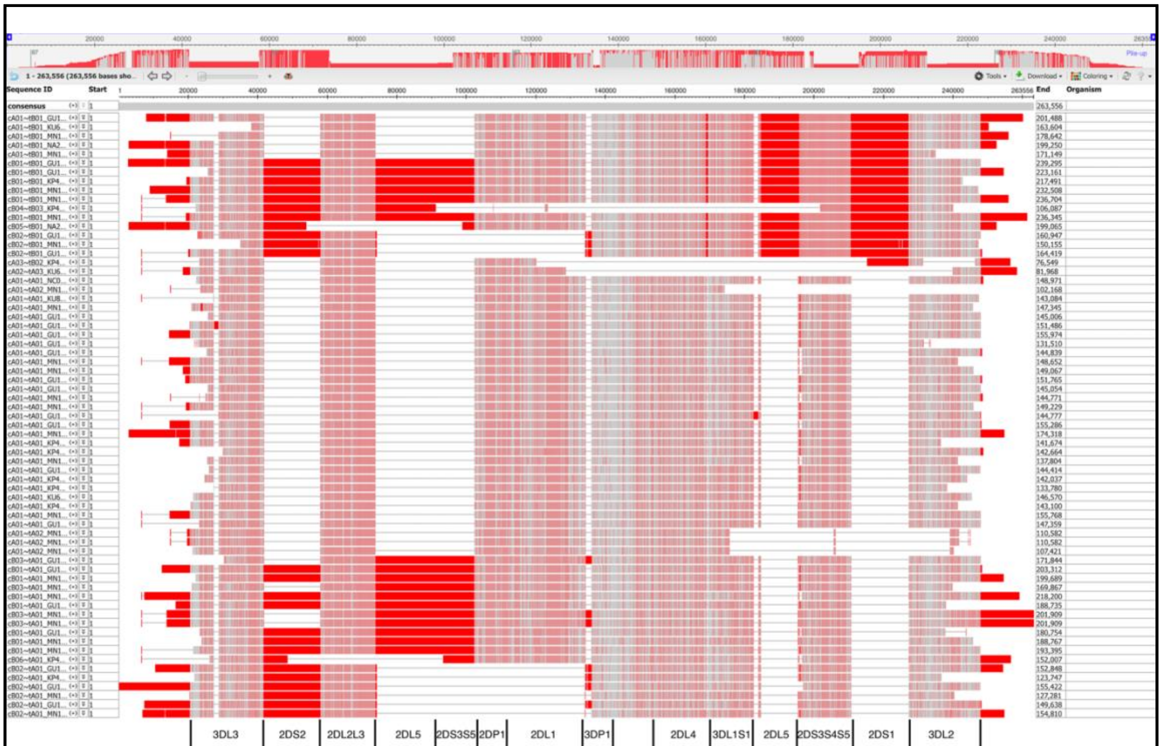

Figure 13. MSA Viewer depiction of the DNA alignment of 67 human haplotypes. Positions that are in an agreement with consensus are colored in gray, positions that are not in an agreement with consensus are colored in red. The location and borders of the genes have been added at the bottom.

Validation of the MSA defined allele assignments by IPD-KIR annotation

The haplotypes were annotated by the names and orders of their alleles. Of the 647 alleles in the 68 haplotypes, 556 (86%) can be assigned names via IPD-KIR 2.9.0, at least at protein resolution. Unnamed alleles occur either because they have not yet been included in the IPD-KIR database, or they are partial alleles in the case of *KIR3DL3* and *KIR3DL2*. 97% of alleles can be named excluding *KIR3DL3*, *KIR2DP1*, and *KIR3DL2*. 39 *KIR2DP1* alleles are unnamed. To check the gene assignment for the 91 alleles that could not be named, the sequences were aligned to a set of 17 full gene alleles, each the first named allele for its assigned gene in IPD-KIR. Each allele being evaluated was considered to be correctly annotated if the allele sequence aligned closet to the IPD-KIR reference to which it was assigned by the motifs. Those results show that every unnamed allele aligned to the reference allele predicted by its motif assignment. The only exception was the GU182360 *KIR3DL2*, which aligns closest to *KIR3DP1*; however, this haplotype sequence is incomplete on the 3' end, and the *KIR3DL2* allele only contains the 2221 5'-end sequences. Alleles that are a fusion of *KIR3DL1* and *KIR3DL2* (e.g., *KIR3DL1* alleles 059-061) are classified as *KIR3DL2* alleles using motifs but are classified as *KIR3DL1* alleles in IPD-KIR.

Comparison with existing MSA software

Of the existing alignment software that was evaluated on the 67 human haplotypes, only stand-alone MAFFT was able to generate an alignment, using the FFT-NS-2 strategy. The input size was too large for Clustal Omega local server, M-Coffee web server, MUSCLE local server, and WebPrank web server, and the other algorithms ran out of memory. Except for large portions of *KIR3DL3*, distal *KIR2DL3/KIR2DS4/2DS3S5*, and *KIR3DL2*, the MAFFT alignment does not recapitulate human-curated KIR haplotypes or gene alleles. It is 75% larger than our motif-guided alignment (960221 vs 263556 positions), and its overview shows the alignment columns are mostly gaps compared with the more expected block shaped column in Figure 12.

2.4.1. MSA conclusions and discussion

Figure 10 suggest 14 human KIR loci and 1 intergenic locus. The genic loci are abbreviated and labeled as 3DL3, 2DS2, 2DL2L3, 2DL5B, 2DS3S5, 2DP1, 2DL1, 3DP1, 2DL4, 3DL1S1, 2DS3S4S5, 2DS1, 3DL2, and the intergenic locus is between 3DP1 and 2DL4 (3DP1-2DL4). There are 9 motif-defined genes in the 14 genic loci (Figure 10). Loci abbreviated 2DL1, 2DS1, and 2DS2 share motif MHCIJKL. Loci 2DL2L3, 2DS3S5, and 2DS3S4S5 share variations of the FZ*GHCIJKL motif. Loci 2DL5A and 2DL5B share variations of FIRL. The motif MSA suggests *KIR2DS4* on the A haplotype shares a locus with *KIR2DS3* and *KIR2DS5* (not *KIR2DS1*) on the B haplotypes. The MSA similarly suggests *KIR2DL2* and *KIR2DL3* share a locus as do *KIR3DL1* and *KIR3DS1*. The Z character is unique to centromeric 2DS3S5 and telomeric 2DS3S4S5 loci and marks the alleles to a certain extent. In both, the *KIR2DS3* alleles have 1 Z in their motifs, *KIR2DS5* has 2 or 3, and *KIR2DS4* alleles are more variable.

Although the motifs correctly annotate the haplotype structures, they do not always do so unambiguously, nor do they always agree with the existing nomenclature as utilized in IPD-KIR. In the current gene nomenclature, there are 16 KIR genes. Some researchers have previously considered *KIR2DL2* and *KIR2DL3* to occupy the same locus as well as *KIR2DS3* and *KIR2DS5*(45)(46). Conversely, *KIR2DL5* was once considered one gene but was split into *KIR2DL5A* for alleles in the centromeric locus and *KIR2DL5B* for alleles in the telomeric locus, creating new 'A' and 'B' designations that are independent from the 'A' and 'B' haplotype classifications(47). The motif MSA suggests the KIR region has 14 loci in 9 gene motifs. Under the existing gene nomenclature, *KIR3DL1* and *KIR3DL2* differ only by an index, since they are both three domain ("3D") long tail ('L') genes; since the extracellular domains of *KIR3DL1/KIR3DL2* fusion alleles are from *KIR3DL1*, those fusion alleles are labelled in IPD-KIR as *KIR3DL1*. They are considered *KIR3DL2* alleles under the motif structure because the proximal portions of *KIR3DL1* and *KIR3DL2* share variations of SCANJ, but the distal portions are different; since the fusions are comprised of proximal *KIR3DL1* and distal *KIR3DL2*, they share a partial proximal motif with *KIR3DL1* and a full motif with *KIR3DL2*. The motif ambiguity between the *KIR3DL1/KIR3DL2* fusion and *KIR3DL2* can be resolved by linkage disequilibrium with *KIR2DS4*, since the fusion lacks *KIR2DS4* and the haplotypes with *KIR2DS4* cannot contain the fusion. Both systems consider the cB05 *KIR2DS2/KIR2DS3* fusion to be *KIR2DS2* (KIR2DS2*005 in IPD-KIR).

The motifs and alignments presented here provide a means to help unify interpretation of the entire KIR region. They can be used to precisely define KIR haplotypes and loci, provide context for assigning alleles (especially fusion alleles) to genes, improve evolutionary inferences, improve imputation, interpret co-expression, and generate markers. The motif probes have been applied to a workflow to capture, assemble, and annotate KIR haplotypes at https://github.com/droeatumn/kass; it includes the ability to annotate KIR contigs/haplotypes as a separate workflow.

2.4.2. Contributions to the human genome project

Structural annotation for the new sequences was confirmed by the workflow of Pyo et al. for annotation of the previously-submitted reference haplotypes. The 27 haplotypes include 8 cA01~tA01, 4 cA01~tA02, 4 cB01~tA01, 3 cB01~tB01, 3 cB03~tA01, 2 cA01~tB01, 2 cB02~tA01, and 1 cB01~tB01. Haplotype MN167506 (cB02~tB01), presumed to be of Asian ancestry, contains a *KIR2DL5B* allele (KIR2DL5B*00804) at the *KIR2DL5A* locus. Four haplotypes (MN167507, MN167509, MN167516, MN167530) with a *KIR3DL1/KIR3DL2* fusion(10) have been deposited in the human genome reference for the first time; they have been labeled as telomeric region 'tA02'. The diversity of structures are depicted visually in Figure 4. In total, the 68 human haplotype structures consist of 27 cA01~tA01, 7 cB01~tA01, 6 cB01~tB01, 6 cB02~tA01, 5 cA01~tB01, 4 cA01~tA02, 4 cB03~tA01, 3 cB02~tB01, and 1 each for cA01~tB04, cA02~tA03, cA03~tB02, cB04~tB03, cB05~tA01, cB05~tB01.

A total of 27 new haplotypes were added to the human genome reference as GenBank accessions MN167504-30, 24 of which are from African Americans. The haplotypes cluster by structure, not population. Haplotypes from Africans or African Americans now constitute 47% of the KIR alternate references in the human genome project, and the human genome project contains more than three times as many alternative references for KIR than any full chromosome. These new haplotypes contain the first deposits of unusual linkages such as KIR2DL5B*00804 in the *KIR2DL5A* locus (MN167506), *KIR3DL1* or *KIR3DL2* (fusion) without *KIR2DS4* (tA02), *KIR2DL2* without *KIR2DS2* (cA03), and a *KIR2DS2* (fusion) without *KIR2DL2* (cB05). MN167526's *KIR2DS2* allele has a 1 base deletion after 285th base in the coding sequence, which leads to a premature stop code at position 435.

3. Diploid CNV/PA haplotype inference with PCR genotyping

3.1. Current state of low-resolution PCR genotyping

Circa 2015, medical association studies suggested it might be valuable to interpret KIR in its structural context and doing so may help resolve ambiguous or contradictory associations found across a wide variety of low-resolution studies.

Within this context, our collective knowledge of structural diversity at the gene level was still coarse, especially for unrelated and non-European populations. Only a couple-dozen fully sequenced haplotypes have been deposited in GenBank for the KIR region(24)(12)(15), although population haplotype frequencies have been reported in many families or CNV-based studies. There have been no unrelated population studies using completely cis-linked haplotypes. Most typings – and therefore studies – have been conducted at gene PA or less often at CNV resolution due to technical and economic considerations of sequencing technologies relative to the characteristics of this large and repetitive region. Haplotypic interpretation can be ambiguous under both PA and CNV resolutions, although the extent of improvement for CNV has not been widely reported.

In order to leverage structural information given low-resolution genotypes, we performed experiments to quantify the effects of population variations, reference haplotypes, and genotyping resolution on population-level haplotype frequency estimations as well as predictions of individual haplotypes (Table 1, lines 1-2).

We generated haplotype frequency estimates for over 10,000 individuals in five populations. We reported on the effects of reference haplotype set and resolution, and we documented the inaccuracies that may occur when using low resolution and large sets of reference haplotypes.

3.2. Novel algorithm for haplotype-pair predictions

3.2.1. Genotypic interpretation under haplotype constraints

Each individual's genotype was inferred as potentially-ambiguous pairs of haplotypes (copy number of 16 genes) from one of the reference sets. Inference was performed using PA resolution for all individuals and again using PA + CNV resolution when available. Some genotypes were interpreted into a pair of haplotypes unambiguously, some ambiguously into multiple pairs, and some genotypes were uninterpretable under the haplotype-pair constraints. When interpretable, haplotype pairs of maximum probability were assigned to each individual.

In Figure 14, the PA genotype data is illustrated in the table 'Genotype data (PA)'. In the table, each row is a genotype, and each column is the presence (Y) or absence (N) of a gene in the sample. The table 'Genotype/Haplotype-pair relationships' shows how likely a genotype is interpreted by a haplotype-pair. For example, a_{ijk} is the probability that genotype g_k is interpreted by the haplotype pair (h_i,h_j). $a_{ijk}=0$ when (h_i,h_j) cannot interpret g_k.

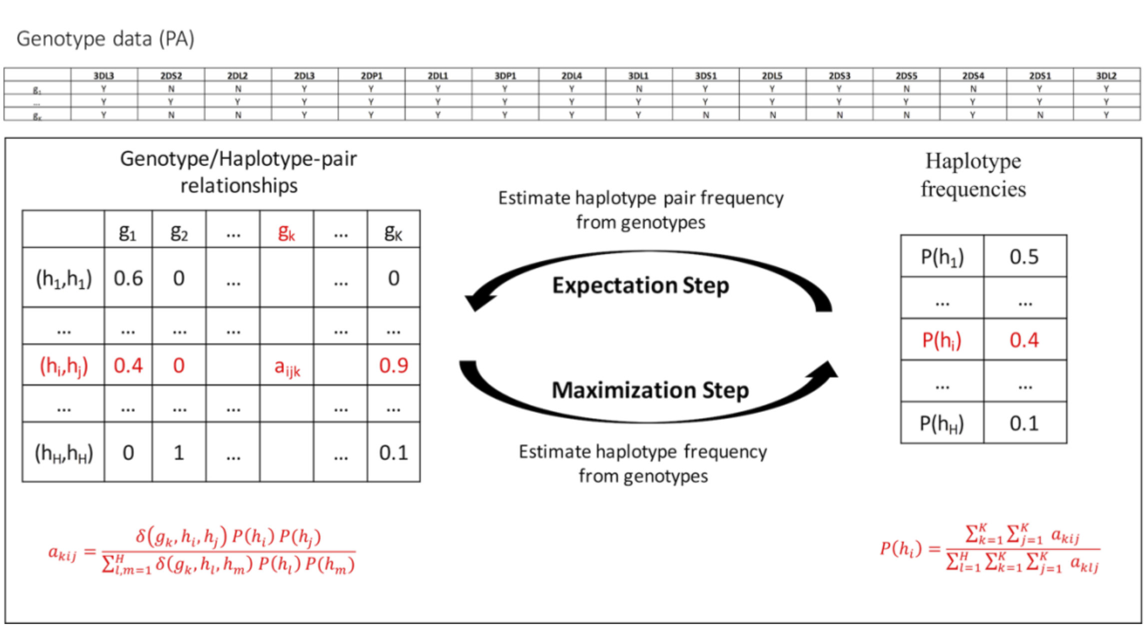

Figure 14. EM workflow. Each individual's genotype g_1 ... g_K) is interpreted as possibly ambiguous pairs of haplotypes (h_1 ... h_H), whose frequencies are initialized evenly. The EM algorithm then alternates between the E and M steps until convergence. The results are haplotype pair assignments for each individual and frequencies estimates for each haplotype.

3.2.2. Haplotype frequency inference

Haplotype frequencies were estimated at PA resolution for all populations (the table 'Haplotype frequencies' in Figure 14). Frequencies were also estimated separately at PA + CNV resolution for the half of AFA with CNV data available. Haplotype pairs and cohort frequencies were estimated using an expectation-maximization (EM) algorithm(48) as illustrated in Figure 14. Given a collection of genotypes and a set of reference haplotypes, the EM algorithm assigns to each genotype the haplotype pair with the maximum probability.

Given haplotypes $\{h_1, h_2, \ldots h_H\}$ and genotypes $\{g_1, \ldots g_K\}$, the likelihood function is

$$L(h_1, h_2, \ldots, h_H: g_1, \ldots, g_K) = \prod_{k=1}^{K} P(g_k|h_1, h_2, \ldots, h_H) = \prod_{k=1}^{K} \prod_{i,j=1}^{H} P\left(g_k \middle| h_i, h_j\right)^{r_{kij}},$$

where r_{kij} is a hidden Bernoulli random variable measuring whether g_k is generated from (h_i,h_j) and

$$LL(h_1, h_2, \ldots, h_H: g_1, \ldots, g_K) = \sum_{k=1}^{K} \sum_{i,j=1}^{H} r_{kij} logP(g_k|h_i, h_j) =$$
$$\sum_{k=1}^{K} \sum_{i,j=1}^{H} r_{kij} (logP(h_i) + logP(h_j)),$$

where $P(h_i)$ is the frequency of haplotype h_i.

3.2.3. EM algorithm

The algorithm iterated between two steps – expectation and maximization -- until convergence. After initializing the haplotypes with equal frequency, the expectation step used the current iteration's estimated haplotype frequencies to calculate haplotype-pair expectations for each genotype, i.e., the entries in the table 'Genotype/Haplotype-pair relationships' in Figure 14:

$$a_{kij} = E\left(r_{kij}\right) = P\left(H_k = \left(h_i, h_j\right)\middle|g_k\right) = \frac{\delta\left(g_k, h_i, h_j\right) P(h_i)\, P(h_j)}{\sum_{l,m=1}^{H} \delta(g_k, h_l, h_m)\, P(h_l)\, P(h_m)}$$

and δ adjusts the statistic for homo/hetero-zygousity:

$$\delta\left(g_k, h_i, h_j\right) = 1 \; if \; h_i = h_j$$
$$\delta\left(g_k, h_i, h_j\right) = 2 \; if \; h_i \neq h_j$$

The maximization step used each genotype's most likely haplotype pair to estimate the frequency of each haplotype illustrated in the table 'Haplotype frequencies' in Figure 14:

$$P(h_i) = \frac{\sum_{k=1}^{K} \sum_{j=1}^{K} a_{kij}}{\sum_{l=1}^{H} \sum_{k=1}^{K} \sum_{j=1}^{K} a_{klj}}$$

An implementation of the algorithm has been deposited in GitHub at https://github.com/droe-nmdp/em1.

3.3. Evaluation in 10,000 individuals from 5 populations

Figure 15 plots the fraction of each population that cannot be explained by each of the three sets of reference haplotypes, i.e., the frequency of individuals not covered by the haplotypes under the PA conditions. In the smallest set of haplotypes HapSet10, these frequencies range from 5.6% in Europeans to almost 20% in African Americans. The largest reference set HapSet63 lowers the uninterpretable frequencies to near 0% for all five populations. African Americans exhibit the highest unexplained rates, followed by Native Americans.

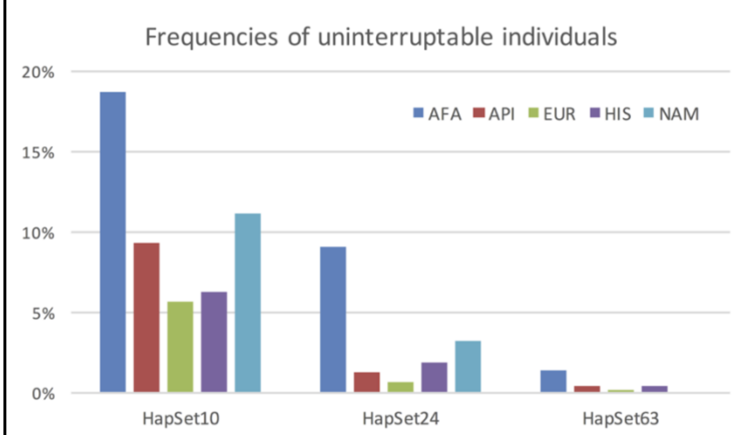

Figure 15. Frequencies of uninterpretable individuals. Each bar represents the percentage of a population that cannot be explained by one of three reference sets of haplotypes. For clarity and distinction, the chart displays a maximum frequency of 20%. Five populations are represented: African American (AFA), Asian Pacific Islander (API), European (EUR), Hispanic (HIS), and Native American (NAM).

Figure 16 plots estimated PA haplotypes frequencies (including frequencies of uninterpretable individuals) by population for HapSet10 and HapSet24. Under HapSet10 conditions (Fig 6a), the ratio of uninterpretable individuals are EUR(5.6%), HIS(6.2%), API(9.3%), NAM(11.1%), and AFA(18.7%) with a 13% difference in the range. cA01~tA01 (haplotype 1 in the plot) ranges by 8%, from 51%(NAM) to 59%(API). For API and HIS, the second most common haplotype is cA01~tB01 (haplotype 3 in the plot) (9-11%) and it is cB02~tA01 (haplotype 4 in the plot) (9-18%) for AFA, EUR, and NAM. All other haplotypes have estimated frequencies of less than 10%. cB02~tA01 is the only haplotype with a range of over 10%: 6.4% (API) to 18.26% (NAM). Similar results was reported for HapSet24 and HapSet63 (not shown).

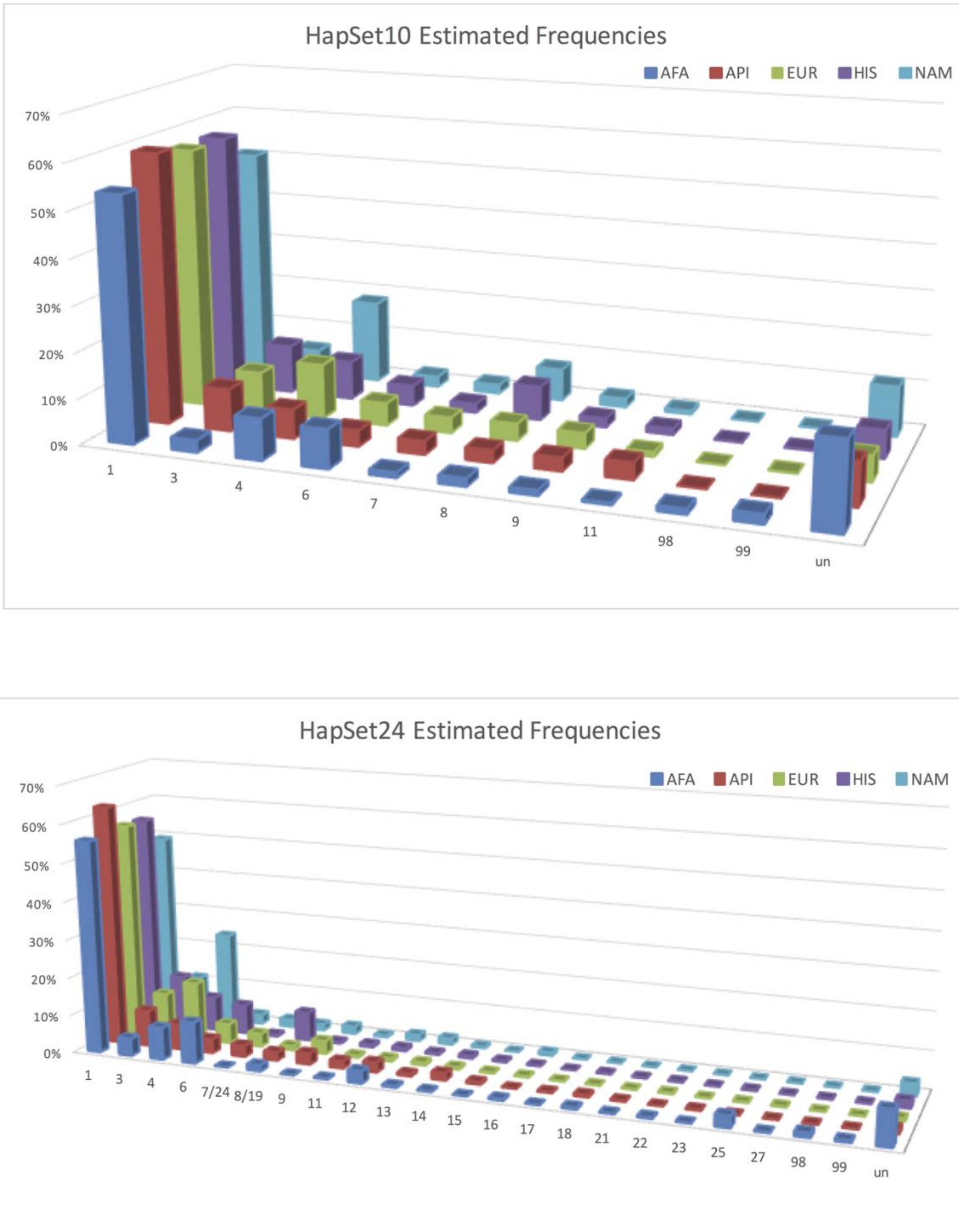

Figure 16. Estimated structural haplotype frequencies for 5 populations for HapSet10 (top) and HapSet24 (bottom) per population and reference haplotype set. Haplotypes in x-axis, frequencies in the y, and populations on the z. 'un' haplotype represents the frequency of individuals that could not be interpreted.

3.4. Conclusions and discussion

The primary purposes of this study were to estimate KIR haplotype frequencies in a relatively large cohort, and to evaluate the variations of the frequencies by population, reference haplotypes, and typing resolution.

Larger sets of reference haplotypes decrease the frequency of unexplainable individuals

All estimates were consistent relative to the previous studies to within 5% except for two haplotypes in the NAM population. Our results showed cB02 associated more frequently with tA01 (+14%) than tB01 (-8.5%). The difference is even stronger when CNV resolution is considered. There are several possible explanations for this difference between U.S. based NAMs relative to various native populations from south of the U.S., including pathogen selection and/or founder effects combined with genetic drift, as similarly noted(49). On the other hand, the differences could be due to the potentially mixed nature of either group, since no further breakdown was available for the 63 individuals and we therefore aggregated the southern group for comparison. The haplotype that exhibits the widest range of frequencies across populations is cB02~tA01, which contains a centromeric B haplotype lacking the *KIR2DL5~KIR2DS3S5~KIR2DL1* region. Comparisons for the three other minority populations are not possible at this time due to our admixed and otherwise biased U.S.-based registry.

The distinguishing features of our computational approach relative to other(50)(51)(52)(53)(54)(55)(56)(57)(58), also mainly EM approaches, are the ability to use multiple or mixed resolutions as well as the size of data sets to which we have applied it.

It can be misleading to interpret PA data with a reference haplotype set that is too large … There are two larger issues that can occur when interpreting low resolution in the context of a large reference haplotype set. Both issues involve the situation when a haplotype is a gene content subset of the other. From the point of view of PA interpretation, one haplotype can be ‘hidden’ behind another and result in misleading estimates.

The first situation can occur when the superset haplotype is relatively frequent. It can lead to the subset haplotype ‘stealing’ frequency from the superset haplotype. For example, haplotype 60 is a subset of haplotype 1: it is a cA01~tA01 haplotype without *KIR3DL3*. Except for the ubiquity of

KIR3DL3, the PA data provides no distinction between 60 and 1 and therefore leads to a higher than expected frequency for 60.

The second situation is the opposite of the first: it can occur when a reference set contains a haplotype that is a gene content superset of a less frequent haplotype. It can lead to the superset haplotype 'stealing' frequency from the subset because adding the relatively frequent gene(s) expands the number of genotypes that can be explained. For example, haplotype 49/51 is a superset of haplotype 4: it is a cB02~tA01 haplotype with *KIR2DP1~KIR2DL1*. Both of those genes have high frequencies such that adding them makes it more accommodating for genotypes with those two usually common genes. This leads to misestimates of haplotype 49/51 with a frequency of 6% and haplotype 4 dropping from 8% to 0.3%. From our CNV analysis, haplotype 4 is exclusively inherited with haplotype 1, and haplotype 49/51 is exclusively inherited with haplotype 16.

Further experiments with mixed PA and CNV resolution show these two situations cannot be remedied by mixed case genotyping wherein some genes (e.g., the four framework genes) are typed at CNV resolution and the rest at PA. In summary, we suggest that reference haplotypes for PA data should not contain haplotypes that are gene-content subsets of other haplotypes. Such a situation can cause hidden haplotypes to erroneously unbalance its frequency with its 'non-hidden' partner: rare hidden haplotypes can be inflated and common hidden haplotypes can be the deflated. This consideration can be valuable for retrospective studies by informing the selection of reference haplotypes given typing resolution(s). It can also be valuable for prospective studies by suggesting which genes should be typed at certain resolutions given a reference set.

The substantial increased clarity that CNV provides is clearly shown when comparing the ambiguity distributions for PA and CNV (details not shown). Higher resolution increased clarity for every reference set. Using the largest set, the higher resolution provided unambiguous haplotype-pair assignments for the majority of individuals.

In summary, we generated haplotype frequency estimates for over 10,000 individuals in five populations, we reported on the effects of reference haplotype set and resolution, and we documented the inaccuracies that may occur when using low resolution and large sets of reference haplotypes. These analytical methods and novel predictions may allow for finer-grained KIR-association analysis, especially when combined with other factors such as expression and HLA ligands.

4. Structural haplotype inference from high-throughput DNA reads

4.1. Current state of high-throughput genotyping

It is difficult to interpret the KIR region with high-throughput sequencing reads for an individual human genome when the structural arrangements are unknown; it is difficult even when the structural haplotypes are known, since the read length is too short to map unambiguously to the repetitive and homologous KIR genes. As a consequence, the reads from KIR region are ignored, as to the best of our knowledge, there are currently no algorithms to interpret KIR from WGS. SNP (single nucleotide polymorphism)-based KIR interpretation is more commonly applied. For example, KIR*IMP is a web-application to predict genes and haplotypes from microarray SNP genotypes(21). As an algorithm whose raw data is microarray calls, KIR*IMP can interpret KIR from genome wide SNP arrays, but it is not applicable to interpret KIR from raw sequences.

Since a general solution for KIR structural interpretation from raw genomic DNA is not currently available, this study implements such an algorithm for the prediction of KIR genes and full structural haplotypes from any type of raw full-region-or-greater genomic sequence at population scale (Table 1, line 6). In particular, we systematically evaluated small markers for KIR genes and then applied those markers to a synthetic KIR probe interpretation (KPI) algorithm for the presence/absence of 16 KIR genes and 16 haplotype structures. Our approach leverages recent bioinformatics innovations for short sequence ('probe') genotyping, along recently published KIR reference haplotypes. The KPI algorithm first efficiently counts the occurrence of each kmer probe in the raw sequences, and then uses multiple probes per gene to call its presence/absence. Those 16 genotypes are then used to generate haplotype-pair predictions. In the experiments, we report 100% accuracy on a test set of synthetic haplotypes for comparisons with known truth. We also report that gene and haplotype-pair predictions for the WGS GoNL cohort are family consistent and compare favorably with reference frequencies in comparison to SNP-based predictions using KIR*IMP.

4.2. The first algorithm for genotyping from WGS

The workflow of KPI consists of three steps,

1. Discover the 25mer gene markers based on a multiple sequence alignment analysis of 68 full-length haplotype sequences.
2. Count the 25mer markers in the reads of genomic DNA per individual to generate the ’individual's 25mer genotype.
3. Predict presence/absence per gene from the marker genotypes for each individual.
4. Predict haplotype pairs from the gene presence/absence calls for each individual.

In the following, we first explain each step and then describe the synthetic data and GoNL data used for the evaluation.

4.2.1. Step 1: Discovering 25mer gene markers

To discover gene marker 25mers, first a multiple sequence alignment was created with 68 publicly deposited full-length haplotypes sequences(59). Briefly, each haplotype was annotated at an average resolution of ~4kbp using a set of 15 120-base markers. This high-level annotation was aligned into a MSA representing a structural alignment of all haplotypes. Then, each subregion was aligned to base pair resolution. This resulted in a full resolution, full haplotype MSA that accurately classifies each allele into a haplotype-defined locus, and it aligns the alleles precisely at each locus. The haplotype and gene annotations of the MSA provided a list of full-length alleles for 16 genes: *KIR2DL1-5, KIR2DS1-5, KIR2DP1, KIR3DL1-3, KIR3DP1*, and *KIR3DS1*. Markers for each gene locus were chosen by selecting all sequences of length 25 (25mers) present in every allele of the gene but not elsewhere in the KIR haplotypes nor the rest of the genome reference GRCh38. The marker sequences are checked in to GitHub at https://github.com/droeatumn/kpi/tree/master/input in text and fasta format.

4.2.2. Step 2: Count 25mer markers

KMC 3, with workflows implemented in Nextflow(60) and Apache Groovy(61) and a software environment implemented as a Docker container, is used to create 25mer databases from sequence or short-read data and match the markers across the datasets. Using KMC 3, we generate the list of all 25mers from the short reads of each individual and then match the 25mers in the marker databases to report the hit counts of each 25mer marker in the individual.

4.2.3. Step 3: Individual genotyping from 25mer markers

KPI calls presence/absence per gene by aggregating the presence/absence genotypes of many small (25mer) markers, each specific to one gene. 25mers with hit counts less than three are considered sequencing errors and set to zero. If the mean hit count of all the markers per gene is zero, then the gene is predicted absent; otherwise, it is called present.

4.2.4. Step 4: Individual haplotyping from genotypes

Haplotype-pair predictions were made by fitting the genotype to all possible pairs of the 16 structural reference haplotypes defined in Figure 17. The numbers and frequencies of the haplotypes are from Jiang et al. 2012(16) (Table 3); some of their haplotypes are combined because Jiang et al. consider certain alleles as separate haplotypes, such as full or deleted alleles of *KIR2DS4*. These 16 haplotypes represent 97% of all haplotypes in the Jiang et al. report.

For the GoNL predictions, haplotype ambiguity was reduced by family trio patterns and then further by the EM (Expectation-Maximization)-based methods(57)(Table 1, lines 1-2). Haplotype frequencies were calculated from the EM-reduced individual haplotype-pair predictions. These haplotype frequency calculations are not possible on the KPI's haplotype-pair predictions because they can be ambiguous.

Jiang et al. 2012 #	3DL3	2DS2	2DL2	2DL3	2DP1	2DL1	3DP1	2DL4	3DL1	3DS1	2DL5	2DS3	2DS5	2DS4	2DS1	3DL2
1/2	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0	1
3	1	0	0	1	1	1	1	1	0	1	1	0	1	0	1	1
11	1	0	0	1	1	1	1	1	0	1	1	1	0	0	1	1
4/5	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1
6/10	1	1	1	0	1	1	1	1	1	0	1	1	0	1	0	1
25	1	1	1	0	1	1	1	1	1	0	1	0	1	1	0	1
7	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1
9	1	1	1	0	1	1	1	1	0	1	1	2	0	0	1	1
8	1	1	1	0	0	0	1	1	0	1	1	0	1	0	1	1
17	1	1	1	0	0	0	1	1	0	1	1	1	0	0	1	1
12	1	1	1	0	0	0	0	0	0	1	1	0	1	0	1	1
18	1	1	1	0	0	0	0	0	0	1	1	1	0	0	1	1
13	1	1	1	0	1	1	2	2	1	1	1	1	0	1	0	1
15	1	1	0	0	1	1	1	1	0	1	1	0	1	0	1	1
16	1	0	0	1	1	1	2	2	1	1	0	0	0	1	0	1
21	1	1	0	0	1	1	1	1	1	0	0	0	0	1	0	1

Figure 17. KPI reference haplotype definitions. Haplotype numeric labels (Jiang et al. 2012) are shown with their definition via gene counts. Following Jiang et al. convention, some haplotypes (e.g., 7, 9) are distinguished by *KIR2DS3/KIR2DS5* alleles instead of structural differences.

Jiang et al. 2012 #	informal names	Jiang et al. 2012 freq.
1/2	cA01~tA01	55.2%
3	cA01~tB01_2DS5	10.9%
11	cA01~tB01_2DS3	1.4%
4/5	cB02~tA01	12.8%
6/10	cB01~tA01_2DS3	6.9%
25	cB01~tA01_2DS5	0.1%
7	cB01~tB01_2DS3_2DS5	2.6%
9	cB01~tB01_2DS3_2DS3	2.1%
8	cB02~tB01_2DS5	2.1%
17	cB02~tB01_2DS3	0.3%
12	cB04~tB03_2DS5	0.8%
18	cB04~tB03_2DS3	0.3%
13	cB01~tB05	0.7%
15	cB05~tB01	0.4%
16	cA01~tB05	0.3%
21	cB05~tA01	0.2%
sum		97.0%

Table 3. KPI reference haplotype names and frequencies. The first column contains the numeric label assigned to haplotypes in Jiang et al. 2012. Column 2 contains the informal names along with a reference frequency in column 3.

4.3. Evaluation with synthetic capture

KPI was evaluated on a synthetic test set. There are six reference haplotype structures with publicly deposited full-length sequences (Figure 17, top six rows). For each of these six structures, one sequence was randomly chosen to represent that structure, and it was paired with a random haplotype sequence from the set of all sequences, with an equal probability for each sequence. dwgsim(62) was used to generate 10,000 2x150 pair reads per haplotype (~20x coverage) with 1% error rate. This provided a simulated six-person validation set of six diploid whole-region short-read sequences, representing all fully sequenced haplotype structures and paired to provide a variety of genotypes.

4.4. Evaluation with GoNL family WGS and Immunochip SNP

KPI was also run on a large real-world example. WGS was obtained from The Genome of the Netherlands(GoNL)(63), a genome sequencing project whose goal is to map the genetic variation within the population of the Netherlands in 250 family trios (750 individuals). The project provided non-paired sequencing of the whole genomes of the population, which was done on the Illumina HiSeq 2000 platform. Coverage of the KIR region were similar to the previously reported(65) whole-genome average of ~10-15x. Two individuals from two different families were removed from the GoNL project for data quality reasons, giving a total of 748 individuals.

KPI's GoNL predictions were compared with results from microarray-based interpretation algorithm KIR*IMP. Illumina Immunochip microarray SNP data was obtained from GoNL. The data was prepared and executed following instructions using KIR*IMP v1.2.0 on 2019-10-05.

To the best of our knowledge, there only exists one method to predict KIR gene content from WGS sequences(64). However, we were unable to obtain results with it for both evaluation data sets. According to the authors, the current version is deprecated and to be replaced soon(65).

4.5. Results

The predictions were evaluated in the small synthetic test set, where truth is known and a large real-world test set, where truth is unknown except for family relationships. Predictions were evaluated by comparing gene and haplotype-pair predictions to: known truth in the synthetic cohort, and family consistency (real-world cohort only), reference frequencies from Jiang et al.'s family copy number study(16), and the Allele Frequencies Database(66) in the real-world cohort. The real-world cohort was also compared with predictions from microarray-based algorithm KIR*IMP, although KIR*IMP was not considered ground truth as it reports accuracies as low as 81% for some genes(21). Haplotype-pair predictions were considered to be family consistent if each parents' two haplotype predictions contained at least one of the child's two predictions and one of the child's haplotypes occurred in one parent and the other haplotype occurred in the other parent.

4.5.1. Synthetic evaluation

Table 4 shows the results of the synthetic tests. The gene present/absent calls were 100% accurate for all genes. Although the haplotype predictions are ambiguous in half of the individuals, all are consistent with the ground truth.

haplotype 1 structure	haplotype 2 structure	haplotype 1 GenBank accession	haplotype 2 GenBank accession	KPI haplotype prediction	haps consist w/truth?	gene pred. accuracy
cA01~tA01	cA01~tA01	GU182344	GU182340	cA01~tA01+cA01~tA01	Y	100%
cA01~tB01	cA01~tA01	KU645197	GU182360	cA01~tA01+cA01~tB01 or cA01~tB01+cA01~tB05	Y	100%
cB01~tA01	cA01~tA01	GU182351	NC000019.10	cA01~tA01+cB01~tA01	Y	100%
cB01~tB01	cB02~tA01	GU182339	GU182353	10 possibilities, including cB01~tB01+cB02~tA01	Y	100%
cB02~tA01	cA01~tA01	GU182341	KP420442	cA01~tA01+cB02~tA01	Y	100%
cB02~tB01	cA01~tA01	GU182359	KP420439	9 possibilities, including cA01~tA01+cB02~tB01	Y	100%

Table 4. Results of KPI synthetic tests. The first four columns detail the sequences from which the tests (n=6 haplotype-pairs) were generated. The fifth column is KPI's haplotype predictions, some of which are summarized for display.

4.5.2. GoNL evaluation

Table 5 shows a summary of the gene prediction results from KPI and KIR*IMP on the GoNL data set. A reference frequency range is included from Allele Frequencies Net, selecting European cohorts >= 500 individuals. Overall agreement between KIR*IMP and KPI for the 16 genes (Table 5, column 6) ranges from 72% to 100%, with a mean of 92%. KIR*IMP differs from the reference haplotype (Table 5, column 7) frequency range by >10% in 4 genes (*KIR2DS2*, *KIR2DL2*, *KIR2DL5*, and *KIR2DS3*) compared with 0 genes for KPI (Table 5, column 8). Both KIR*IMP and KPI differ from the *KIR2DS1* reference by 9-10%, although the two algorithms agree in 98% of individuals for that gene.

gene	Ref. freq.	KIR*IMP freq.	KPI freq.	KIR*IMP - KPI	KIR*IMP & KPI agreement	KIR*IMP - reference	KPI - reference
2DS2	53-54%	39%	52%	-13%	**72%**	**-14%**	**-1%**
2DL2	53-54%	39%	51%	-12%	**72%**	**-14%**	**-2%**
2DL3	90%	96%	92%	4%	92%	6%	2%
2DP1	96%	99%	97%	2%	97%	3%	1%
2DL1	96%	99%	97%	2%	97%	3%	1%
3DP1	100%	100%	100%	0%	100%	0%	0%
2DL4	100%	100%	100%	0%	100%	0%	0%
3DL1	93-94%	96%	96%	0%	100%	2%	2%
3DS1	38-44%	33%	35%	-2%	97%	-5%	-3%
2DL5	53-56%	38%	47%	-9%	**87%**	**-15%**	**-6%**
2DS3	30-31%	10%	29%	-18%	**81%**	**-20%**	**-1%**
2DS5	30-36%	25%	27%	-2%	96%	-5%	-3%
2DS4	92-94%	96%	96%	0%	100%	2%	2%
2DS1	43-44%	33%	34%	-1%	98%	-10%	-9%
average					92%		

Table 5. Summary of KIR*IMP and KPI gene predictions. Frequencies relative to GoNL cohort of 748 individuals. The abbreviated gene name is in column 1. Column 2 lists the reference frequencies from The Allele Frequency Net Database. The predicted frequencies from KIR*IMP and KPI are in columns 3 and 4, respectively. The delta between KIR*IMP and KPI is shown in the column 5. Column 6 shows the agreement between KIR*IMP and KPI. Column 7 shows the delta between KIR*IMP and the reference. Column 8 shows the delta between KPI and the reference. Frequencies with differences >10% are in bold.

Table 6 breaks down the differences between KIR*IMP and KPI in a confusion matrix. In the cases where KIR*IMP calls present ('P') and KPI calls absent ('A') (Table 6, column 2), the largest discrepancies are found in the centromeric genes *KIR2DS2* (8%), *KIR2DL2* (8%), and *KIR2DL3* (6%). In the reverse cases, when KIR*IMP calls absent and KPI calls present (Table 6, column 3), the largest discrepancies are greater and occur with the centromeric *KIR2DS2* (20%), *KIR2DL2* (20%), and the paralogous (centromeric or telomeric) *KIR2DL5* (11%), and *KIR2DS3* (19%).

gene	**KIR*IMP: P KPI: A**	**KIR*IMP: A KPI: P**	**KIR*IMP: P KPI: P**	**KIR*IMP: A KPI: A**
2DS2	8%	**20%**	32%	40%
2DL2	8%	**20%**	31%	41%
2DL3	6%	2%	90%	2%
2DP1	2%	0%	97%	0%
2DL1	2%	1%	97%	0%
3DP1	0%	0%	100%	0%
2DL4	0%	0%	100%	0%
3DL1	0%	0%	96%	4%
3DS1	1%	3%	32%	64%
2DL5	2%	**11%**	36%	51%
2DS3	1%	**19%**	10%	71%
2DS5	1%	3%	24%	72%
2DS4	0%	0%	96%	4%
2DS1	1%	1%	33%	66%

Table 6. Confusion matrix of KIR*IMP and KPI gene predictions. Frequencies relative to GoNL cohort size of 748 individuals. The abbreviated gene names are in column 1. Column 2 lists the cases when KIR*IMP calls present ('P') and KPI calls absent ('A'). Column 3 lists the cases when KIR*IMP calls absent ('A') and KPI calls present ('P'). Column 4 is when they both call present. Column 5 is when they both call absent. Discrepancies >10% are in bold.

Per-individual haplotype-pair predictions were generated for the GoNL. Three lists of haplotype-pairs are made: one for the initial fitting of all possible haplotype-pairs that could explain the genotype (i.e., KPI's output); another that reduces those possibilities by family relationships; and one for the EM-reduced final haplotype-pair predictions.

KIR*IMP makes one most-likely prediction for all individuals. KPI's predictions are sometimes ambiguous, with most predictions (mode) having one haplotype-pair but a mean of 2.3, standard deviation of 2.5, and a maximum of 14 haplotype-pair predictions per individual in the context of the 16 reference haplotypes. The KIR*IMP predictions are family consistent 100% of the time compared with 99.6% for KPI. However, the haplotype pair predictions between the two algorithms are concordant only 58% of the predictions.

Table 7 compares the haplotype predictions between KIR*IMP and the EM-reduced haplotype pair predictions from the KPI output. KIR*IMP fit 100% of its predicted genotypes into 15 of its reference haplotypes. KPI fit 97% of its predicted genotypes into its 16 reference haplotypes. KIR*IMP made predictions for two haplotypes (cA01~tB04 and cB04~tB03, numbered 14, 18, and 12), totaling 0.47%, that are not in KPI's set of reference haplotypes. KPI's haplotype-pair predictions are too ambiguous to summarize in haplotype frequencies.

hap #	reference frequency	KIR*IMP	KPI w/EM		KIR*IMP - reference	KPI w/EM - reference		KIR*IMP - KPI w/EM
1 2	55.20%	71.86%	59.70%		16.70%	4.50%		12.17%
3	10.90%	12.57%	9.60%		1.70%	-1.29%		2.95%
11	1.40%	1.47%	0.60%		0.00%	-0.82%		0.85%
4 5	12.80%	7.49%	15.30%		-5.30%	2.55%		-7.83%
9	2.10%	3.41%	2.90%		1.30%	0.78%		0.52%
7	2.60%	0.33%	3.60%		-2.20%	1.00%		-3.24%
6 10	6.90%	1.80%	5.90%		-5.10%	-1.08%		-4.05%
8	2.10%	0.60%	0.50%		-1.50%	-1.65%		0.12%
17	0.30%	0.00%	0.00%		-0.30%	-0.23%		-0.03%
14*	2.40%	0.40%	0.00%		-2.00%	0.00%		0.00%
18*	0.30%	0.07%	0.00%		-0.20%	0.00%		0.00%
12*	0.80%	0.00%	0.00%		-0.80%	0.00%		0.00%
mean	97.00%	100.00%	98.10%					

Table 7. Comparison of KIR*IMP and EM-reduced KPI haplotype prediction frequencies. Column 2 shows the reference European frequencies, numbered in column 1. KIR*IMP's haplotype frequencies for GoNL haplotypes in column 3, which are EM-reduced in Column 4. Column 5 compares KIR*IMP frequencies with the reference, as column 6 does for EM predictions. Finally, column 7 compares the frequencies of KIR*IMP and the EM-reduced predictions.

4.5.3. Evaluation by an independent third party

KPI was evaluated by Chen et al. as part of a larger effort(67). In a cohort of 72 individuals with ground truth determined by LinkSeq qPCR, they report six mismatches in one sample (possibly swapped), and apart from this 95.8% accuracy for *KIR2DS3* and 100% accuracy for the 15 other genes. As they note, it is now possible to interpret HLA binding alleles and the presence/absence of all their KIR receptors from short-read high-throughput sequencing, and this combination is a valuable advancement for research and medicine. Indeed, they compare KPI favorably with respect to clinical accreditation standards.

4.6. Conclusions and discussion

The findings by Chen et al. are consistent with the results of our synthetic test, whose accuracy was 100% for all genes. One drawback of the design of the synthetic test is that the haplotypes used in the test were also included in the MSA that was used to generate the per-gene probes. However, the main purposes of the synthetic tests were to test the application of the markers to short reads in a variety of genotypes and recover their original geno- and haplo-types; this is value-added compared with simply demonstrating sequences unique to a gene. The almost-perfect results of the Genentech and synthetic experiments, along with a GoNL results that had a 99.6% family-consistency rate and in line with expected frequencies, provide evidence that KPI's gene predictions are very accurate.

The evidence also suggests that KPI's haplotype results are accurate, although often ambiguous: the accuracy in the synthetic test was 100% and the GoNL family-consistency was 99.6, and the predictions allow EM predictions that align with the expected population frequency from the literature.

KIR*IMP's haplotype frequency estimations differ from expectations in some areas. The evidence from comparisons with frequency reports from Jiang et al. 2012 (Table 7, column 5) suggest KIR*IMP overestimated cA01~tA01 (haplotype numbers 1 and 2) and underreported haplotypes containing cB01 or cB02 centromeric regions combined with the tA01 telomeric region (cB01~tA01 and cB02~tA01) in the GoNL cohort. This discrepancy can also be seen in the predicted genotype frequencies, where KIR*IMP relatively under calls the presence of *KIR2DS2*, *KIR2DL2*, and *KIR2DS3* by ~20% and *KIR2DL5* by ~10% compared with KPI and the historical

European frequencies from Allele Frequency Net database (Table 3, column 6); all four of those genes are in cB01, and *KIR2DS2* and *KIR2DL2* are also in cB02. GoNL genotyping was done on the Immunochip, which is the best option according to the KIR*IMP manuscript. With that chip, they report accuracies of 100% for *KIR2DS2*, 98% for *KIR2DL2*, 82% for *KIR2DL5*, 81% for *KIR2DS3*, and 95% for *KIR2DS5* in their Norwegian-German validation cohort. Although the family consistency rate is 100% for KIR*IMP and 96.6% for KPI, their haplotype-pair predictions only agree in 58% of individuals. Without ground truth available, without any reason to expect this cohort to deviate from expectations, and considering KIR*IMP's self-reported accuracy, the evidence suggests that KPI's predictions are more accurate than KIR*IMP's in this cohort and specifically that KIR*IMP under called the presence of genes *KIR2DS2, KIR2DL2, KIR2DS3, KIR2DL5* and haplotypes cB01~tA01 and cB02~tA01. As reviewed recently by Wright et al., this may be particularly relevant in the context of hematopoietic stem cell transplantation, where some case/control studies claim an important role for these regions(68). There are several potential reasons KIR*IMP's predictions may be less accurate than KPI's. The reference haplotypes used for marker discovery for KIR*IMP were defined by copy number genotyping and family relationships; KPI defined its haplotypes using a MSA of full sequence haplotypes. KIR*IMP's input is restricted to a few hundred single nucleotide polymorphisms, whereas KPI can use the entire genomic range of KIR sequences of length 25, which provides the potential for more information per marker and a broader base of markers. KIR*IMP uses a small number of SNPs to call one or more genes, whereas KPI uses dozens-to-thousands of 25mers to call a single gene, One of the steps of KIR*IMP's workflow is to align and phase all the SNPs to one 'A' haplotype, which may be a limitation for genes not on that haplotype; all the gene and haplotypes we found to have lower accuracy rates are not located on the 'A' haplotype. KPI has no alignment or assembly steps. It is also important to note that the primary purpose for the comparison with KIR*IMP was not to evaluate the potential success of predicting KIR genes and haplotypes using SNPs vs sequence reads, but rather to compare the two algorithms. Although both algorithms predict the presence/absence of KIR genes and structural haplotypes, their solution domains are very different: microarray SNP panels vs raw genomic DNA reads. Both algorithms report the lowest accuracy rates for *KIR2DS3* and a ~10% lower frequency rate for *KIR2DS1* in GoNL compared with reference frequencies.

The 85% family consistency rate of the EM-reduced haplotype predictions suggest that KIR haplotype ambiguity cannot be accurately reduced at the individual level via expectation-maximization. However, since the EM-reduced haplotype frequencies are in line with references, it is possible the predictions might aggregate to population-level in a maximum-likelihood manner and therefore perhaps may still be useful for some population genetics purposes.

Traditional lab-based SSO presence/absence genotyping relies on a single short-sequence strategy, an approach that can be applied similarly to synthetic analysis of large amounts of WGS. In this virtual context, primer locations are not needed, and kmer searching is efficient and accurate at populations scales. To develop this synthetic SSO-like (kmer) library, we leveraged the information from a multiple sequence alignment of all full-length haplotypes that are available for this study. We believe this is a more accurate approach than using IPD-KIR reference alleles, because the IPD-KIR reference alleles do not require the haplotype location to be known. In addition, fusion alleles are assigned in IPD-KIR to one of the two parent genes, and therefore large sequences of some alleles are not really from the gene in which they are classified. We used 25 for our 'k' (i.e., sequence size) because BLAST searching indicated this to be a conservative minimum length needed to distinguish a small set of test markers to the KIR region. We did not experiment with any k size other than 25, since the choice gives a reasonable number of significant markers and their lack of off-KIR hits as tested in the GoNL population WGS confirms the effectiveness as gene/intergenic markers. The only fundamental benefit to shorter markers would be in the case when there were no longer markers; however, 25mer markers were found for every gene. The only fundamental benefit to longer markers would be if the markers were not unique to the region; however, all the markers are unique to their region. Many of the 25mers overlap each other, effectively simulating a single longer marker. Similar to the reasoning about probe length, probe mismatches would only need to be relaxed if a locus did not have any markers. Since at least one marker was discovered for each gene, mismatches did not need to be incorporated. Having thus obtained the region markers, we then used the most common ('peak') hit count from each gene/intergene's library of sequences to make the PA genotype calls. This adds a certain amount of allelic flexibility in the algorithm because the ultimate call is an average of all the markers for that gene; if some markers miscall, the overall call for the gene will be unaffected if the majority of the other markers are accurate. Since KPI decomposes the genetic information into 25mers, it works with any collection of DNA reads, as long as the KIR region is included. It works with fasta, fastq, single, paired, short, and long reads. Since the markers are not unique to exons, it will not work with cDNA or exon only reads.

One limitation of the method is that the markers do not mark DNA segments longer than one gene. Perhaps this is primarily due to the frequent recombination between haplotypes. Although recombination has been reported in multiple loci, the hotspot in between the centromeric and telomeric regions is particularly strong, and, in general, any pairing of the two can be expected. We evaluated single markers for haplotypes, but we did not find any. This is particularly relevant in light of the observation that haplotyping from genotypes seems to have limited accuracy under maximum likelihood assumptions (Table 7). It is possible that an algorithm that uses applies multiple markers in a hierarchical or combinational manner may be more successful. For future work, we plan to further evaluate *KIR2DS3* (95.8% accuracy in the Chen et al. evaluation) and evaluate the genotyping in diverse populations. It is possible that population robustness is a weakness of the method, although the fact that almost half of individuals in the discovery cohort are African or African American provides some optimism.

For the WGS data set, KPI averaged less than 1 hour of computing time per individual, with 32 cores of CPU and 32G RAM. The majority of the time is spent using KMC 3 to build the kmer database. kmer counting is an active area of research. Since KPI can easily be altered to use any such application, it has potential for future efficiency improvements.

The markers discovered in this study were enabled by our full-haplotype MSA(59). That manuscript makes observations about the composition and order of sequences within KIR haplotypes, and it reports the implications for our understanding of the relationship between haplotypes, loci, and genes. Here, we have leveraged that basic understanding for practical use by developing a free and open interpretation application, evaluating a SNP interpretation algorithm, and contributing KIR interpretation to an important population genetics resource for the Netherlands and many types of human genetic researchers. We described how the gene markers were discovered from the MSA, and we demonstrated their use to predict genes and haplotype-pairs at high accuracy (97%+) with population scale from any kind of sequence data that includes the full KIR region, including WGS. It was tested on synthetic ground-truth sequences and a large cohort of family WGS. In addition, we compared our algorithm to the leading SNP-based interpretation algorithm. KPI is free software with a GPL3 license and is implemented as a Nextflow workflow backed with an optional Docker environment. It is available at https://github.com/droeatumn/kpi.

5. Conclusions and future work

While high-resolution haplotyping by long-range PCR or fosmid clones is costly and inefficient, low-resolution factors such as gene presence/absence and copy numbers by SNP microarrays and qPCR provide limited information for functional analysis and clinical decisions. Our new algorithms enable researchers to utilize abundant uninterpretable short reads for KIR haplotype analysis and perform associations with diseases at higher resolutions in larger cohorts. The library of the full sequence haplotypes generated in this study will aide KIR researchers to study gene combinations, regulatory regions, recombination hotspots, self-regulation, and non-binding factors that influence disease phenotypes.

These algorithms and bioinformatics workflows allow efficient and affordable complete sequencing and interpretation of KIR. It is the culmination of several computational methods we have developed that interpret KIR under different typing resolutions. This multi-resolution aspect is crucial to overall understanding and it improves resolution at any given level by leveraging references and/or markers from other resolutions. For example, others initially inferred structural haplotypes via low resolution presence/absence genotyping in families. We developed algorithms that use these libraries as references to predict individual structural haplotypes and their frequencies in tens of thousands of individuals with PA or CNV genotypes. This population-scale knowledge allowed us to identify the largest gaps in our reference haplotype libraries. Gaps that we filled by a fosmid and single-molecule sequencing project that was the first to fully-characterize the regions in diploid individuals; our results added 15 alternative haplotypes to the human genome reference project. Then, we used these new references and previously-published alleles to create candidate sequence probes/markers of size 25 (25mers). We then queried all whole genome sequences of GoNL for the probes, used the results to genotype at the gene level, and then imputed the haplotype structures, using family relationships and expectation-maximization to resolve ambiguities. From this, we report individual haplotype pair predictions, haplotype frequencies, and novel markers (with individuals significance scores) for 19 genes, 24 intergene regions, 18 haplotypes, and 8 half-haplotypes. These methods can be also used to predict haplotypes and discover markers in other populations. Finally, we apply these gene, intergene, and haplotype markers in a first-of-its-kind KIR assembler, which uses probes to interpret structural haplotypes from raw medium-length (~5kb) reads, then bins the reads based on their location in the haplotypes, assembles each bin in isolation, and then assembles the consensus of each bin into consensus haplotype sequences. This is the first technique to allow full-haplotype KIR sequencing and interpretation for all, and therefore personalizes a complicated and medically-relevant part of the human genome.

We anticipate these methods will be used to analyze KIR, NK, and immune system effects at larger scales and higher resolution. They should be useful for almost all purposes of genetic analysis: from references libraries, to analysis of populations, to outcomes studies and personal medicine. However, it will require new computational methods to discover the complex associations between the ~250 kb KIR haplotypes and their ~1 mb HLA ligand region. Fortunately, recent advancements in machine learning provide encouragement that we can leverage our expertise in those algorithms to discover these truths.

There is also potential to extend some of our methods to other gene systems. The key to our capture approach is to leverage small, relatively-constant sequences in the midst of larger, more variable or repetitive regions. Since this is a characteristic of many immune system and/or transposable element regions, we may be able to generate efficient capture and possibly annotation for these regions as well.

6. Bibliography

1. Wroblewski EE, Parham P, Guethlein LA. Two to Tango: Co-evolution of Hominid Natural Killer Cell Receptors and MHC. *Frontiers in Immunology* (2019) **10**: doi:10.3389/fimmu.2019.00177

2. Bastidas‐Legarda LY, Khakoo SI. Conserved and variable natural killer cell receptors: diverse approaches to viral infections. *Immunology* (2019) **156**:319–328. doi:10.1111/imm.13039

3. Díaz-Peña R, de los Santos MJ, Lucia A, Castro-Santos P. Understanding the role of killer cell immunoglobulin-like receptors in pregnancy complications. *J Assist Reprod Genet* (2019) **36**:827–835. doi:10.1007/s10815-019-01426-9

4. Lau CM, Adams NM, Geary CD, Weizman O-E, Rapp M, Pritykin Y, Leslie CS, Sun JC. Epigenetic control of innate and adaptive immune memory. *Nature Immunology* (2018) **19**:963–972. doi:10.1038/s41590-018-0176-1

5. Parham P, Guethlein LA. Genetics of Natural Killer Cells in Human Health, Disease, and Survival. *Annu Rev Immunol* (2018) **36**:519–548. doi:10.1146/annurev-immunol-042617-053149

6. Bruijnesteijn J, van der Wiel MKH, de Groot N, Otting N, de Vos-Rouweler AJM, Lardy NM, de Groot NG, Bontrop RE. Extensive Alternative Splicing of KIR Transcripts. *Front Immunol* (2018) **9**:2846. doi:10.3389/fimmu.2018.02846

7. Pende D, Falco M, Vitale M, Cantoni C, Vitale C, Munari E, Bertaina A, Moretta F, Del Zotto G, Pietra G, et al. Killer Ig-Like Receptors (KIRs): Their Role in NK Cell Modulation and Developments Leading to Their Clinical Exploitation. *Front Immunol* (2019) **10**:1179. doi:10.3389/fimmu.2019.01179

8. Agrawal S, Prakash S. Significance of KIR like natural killer cell receptors in autoimmune disorders. *Clinical Immunology* (2020) **216**:108449. doi:10.1016/j.clim.2020.108449

9. Chauhan SKS, Koehl U, Kloess S. Harnessing NK Cell Checkpoint-Modulating Immunotherapies. *Cancers* (2020) **12**:1807. doi:10.3390/cancers12071807

10. Marsh SGE, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D, Vilches C, Carrington M, Witt C, Guethlein LA, et al. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Immunogenetics* (2003) **55**:220–226. doi:10.1007/s00251-003-0571-z

11. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research* (2014) doi:10.1093/nar/gku1161

12. Martin AM, Freitas EM, Witt CS, Christiansen FT. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. *Immunogenetics* (2000) **51**:268–280.

13. Martin AM, Kulski JK, Gaudieri S, Witt CS, Freitas EM, Trowsdale J, Christiansen FT. Comparative genomic analysis, diversity and evolution of two KIR haplotypes A and B. *Gene* (2004) **335**:121–131. doi:10.1016/j.gene.2004.03.018

14. Roe D, Vierra-Green C, Pyo C-W, Eng K, Hall R, Kuang R, Spellman S, Ranade S, Geraghty DE, Maiers M. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun* (2017) **18**:127–134. doi:10.1038/gene.2017.10

15. Pyo C-W, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, Marsh SGE, Miller JS, Parham P, Geraghty DE. 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* (2010) **5**:e15115. doi:10.1371/journal.pone.0015115

16. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, Cookson WO, Trowsdale J, Traherne JA. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. *Genome Research* (2012) **22**:1845–1854. doi:10.1101/gr.137976.112

17. Manser AR, Weinhold S, Uhrberg M. Human KIR repertoires: shaped by genetic diversity and evolution. *Immunological reviews* (2015) **267**:178–196.

18. Uhrberg M, Valiante NM, Shum BP, Shilling HG, Lienert-Weidenbach K, Corliss B, Tyan D, Lanier LL, Parham P. Human Diversity in Killer Cell Inhibitory Receptor Genes. *Immunity* (1997) **7**:753–763. doi:10.1016/S1074-7613(00)80394-5

19. Pyo C-W, Wang R, Vu Q, Cereb N, Yang SY, Duh F-M, Wolinsky S, Martin MP, Carrington M, Geraghty DE. Recombinant structures expand and contract inter and intragenic diversification at the KIR locus. *BMC Genomics* (2013) **14**:89. doi:10.1186/1471-2164-14-89

20. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, Middleton D, Carrington M, Trowsdale J. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. *Human Molecular Genetics* (2010) **19**:737–751. doi:10.1093/hmg/ddp538

21. Vukcevic D, Traherne JA, Næss S, Ellinghaus E, Kamatani Y, Dilthey A, Lathrop M, Karlsen TH, Franke A, Moffatt M, et al. Imputation of KIR Types from SNP Variation Data. *The American Journal of Human Genetics* (2015) **97**:593–607. doi:10.1016/j.ajhg.2015.09.005

22. Leaton LA, Shortt J, Kichula KM, Tao S, Nemat-Gorgani N, Mentzer AJ, Oppenheimer SJ, Deng Z, Hollenbach JA, Gignoux CR, et al. Conservation, Extensive Heterozygosity, and Convergence of Signaling Potential All Indicate a Critical Role for KIR3DL3 in Higher Primates. *Frontiers in Immunology* (2019) **10**: doi:10.3389/fimmu.2019.00024

23. IPD-KIR. Available at: https://www.ebi.ac.uk/ipd/kir/

24. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, Beck S, Trowsdale J. Plasticity in the organization and sequences of human KIR/ILT gene families. *Proceedings of the National Academy of Sciences* (2000) **97**:4778–4783.

25. Carrillo-Bustamante P, Keşmir C, de Boer RJ. The evolution of natural killer cell receptors. *Immunogenetics* (2016) **68**:3–18. doi:10.1007/s00251-015-0869-7

26. RepeatMasker. Available at: http://www.repeatmasker.org/

27. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* (2015) **6**:11. doi:10.1186/s13100-015-0041-9

28. Salmela L, Walve R, Rivals E, Ukkonen E. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* (2016)btw321. doi:10.1093/bioinformatics/btw321

29. Roe D, Kuang R. Accurate and Efficient KIR Gene and Haplotype Inference from Genome Sequencing Reads with Novel K-mer Signatures. Bioinformatics (2019). doi:10.1101/541938

30. BBTools. Available at: http://sourceforge.net/projects/bbmap/

31. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. *Genome Res* (2017) **27**:722–736. doi:10.1101/gr.215087.116

32. Lafita A, Bliven S, Prlić A, Guzenko D, Rose PW, Bradley A, Pavan P, Myers-Turnbull D, Valasatava Y, Heuer M, et al. BioJava 5: A community driven open-source bioinformatics library. *PLoS Comput Biol* (2019) **15**:e1006791. doi:10.1371/journal.pcbi.1006791

33. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* (2018) **34**:3094–3100. doi:10.1093/bioinformatics/bty191

34. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* (2015)btv566. doi:10.1093/bioinformatics/btv566

35. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* (2018) **34**:2666–2669. doi:10.1093/bioinformatics/bty149

36. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* (2013) **29**:1072–1075. doi:10.1093/bioinformatics/btt086

37. Veltri D, Wight MM, Crouch JA. SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species. *Nucleic Acids Research* (2016) **44**:W41–W45. doi:10.1093/nar/gkw330

38. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* (2013) **14**:193–202. doi:10.1093/bib/bbs012

39. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* (2019) doi:10.1038/s41587-019-0217-9

40. Martin MP, Bashirova A, Traherne J, Trowsdale J, Carrington M. Cutting Edge: Expansion of the *KIR* Locus by Unequal Crossing Over. *J Immunol* (2003) **171**:2192–2195. doi:10.4049/jimmunol.171.5.2192

41. Sambrook JG. Single haplotype analysis demonstrates rapid evolution of the killer immunoglobulin-like receptor (KIR) loci in primates. *Genome Research* (2005) **15**:25–35. doi:10.1101/gr.2381205

42. Guethlein LA, Older Aguilar AM, Abi-Rached L, Parham P. Evolution of Killer Cell Ig-Like Receptor (*KIR*) Genes: Definition of an Orangutan *KIR* Haplotype Reveals Expansion of Lineage III KIR Associated with the Emergence of MHC-C. *J Immunol* (2007) **179**:491–504. doi:10.4049/jimmunol.179.1.491

43. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* (2017) doi:10.1093/bib/bbx108

44. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* (2014) **30**:3276–3278. doi:10.1093/bioinformatics/btu531

45. Falco M, Moretta L, Moretta A, Bottino C. KIR and KIR ligand polymorphism: a new area for clinical applications?: Relevance of KIR and KIR-L polymorphism. *Tissue Antigens* (2013) **82**:363–373. doi:10.1111/tan.12262

46. Ordóñez D, Meenagh A, Gómez-Lozano N, Castaño J, Middleton D, Vilches C. Duplication, mutation and recombination of the human orphan gene KIR2DS3 contribute to the diversity of KIR haplotypes. *Genes & Immunity* (2008) **9**:431–437. doi:10.1038/gene.2008.34

47. Cisneros E, Moraru M, Gómez-Lozano N, López-Botet M, Vilches C. KIR2DL5: An Orphan Inhibitory Receptor Displaying Complex Patterns of Polymorphism and Expression. *Frontiers in Immunology* (2012) **3**: doi:10.3389/fimmu.2012.00289

48. Do CB, Batzoglou S. What is the expectation maximization algorithm? *Nature biotechnology* (2008) **26**:897–900.

49. Augusto DG, Piovezan BZ, Tsuneto LT, Callegari-Jacques SM, Petzl-Erler ML. KIR Gene Content in Amerindians Indicates Influence of Demographic Factors. *PLoS ONE* (2013) **8**:e56755. doi:10.1371/journal.pone.0056755

50. Yoo YJ, Tang J, Kaslow RA, Zhang K. Haplotype inference for present absent genotype data using previously identified haplotypes and haplotype patterns. *Bioinformatics* (2007) **23**:2399–2406. doi:10.1093/bioinformatics/btm371

51. Erdem E, Erdem O, Türe F. "HAPLO-ASP: Haplotype inference using answer set programming," in *Logic Programming and Nonmonotonic Reasoning* (Springer), 573–578. Available at: http://link.springer.com/chapter/10.1007/978-3-642-04238-6_60 [Accessed June 19, 2014]

52. Erdem E, Türe F. Efficient Haplotype Inference with Answer Set Programming. in *AAAI*, 436–441. Available at: http://www.aaai.org/Papers/AAAI/2008/AAAI08-069.pdf [Accessed June 19, 2014]

53. Nowak RM, Ploski R. NullHap - a versatile application to estimate haplotype frequencies from unphased genotypes in the presence of null alleles. *BMC Bioinformatics* (2008) **9**:330. doi:10.1186/1471-2105-9-330

54. Single RM, Martin MP, Meyer D, Gao X, Carrington M. Methods for assessing gene content diversity of KIR with examples from a global set of populations. *Immunogenetics* (2008) **60**:711–725. doi:10.1007/s00251-008-0331-1

55. Gourraud PA, Gagne K, Bignon JD, Cambon-Thomsen A, Middleton D. Preliminary analysis of a KIR haplotype estimation algorithm: a simulation study. *Tissue Antigens* (2007) **69**:96–100. doi:10.1111/j.1399-0039.2006.762_4.x

56. Augusto DG, Hollenbach JA, Petzl-Erler ML. A deep look at KIR–HLA in Amerindians: Comprehensive meta-analysis reveals limited diversity of KIR haplotypes. *Human Immunology* (2015) **76**:272–280. doi:10.1016/j.humimm.2015.01.025

57. Vierra-Green C, Roe D, Hou L, Hurley CK, Rajalingam R, Reed E, Lebedeva T, Yu N, Stewart M, Noreen H, et al. Allele-Level Haplotype Frequencies and Pairwise Linkage Disequilibrium for 14 KIR Loci in 506 European-American Individuals. *PLoS ONE* (2012) **7**:e47491. doi:10.1371/journal.pone.0047491

58. Wu J, Chen G-B, Zhi D, Liu N, Zhang K. A hidden Markov model for haplotype inference for present-absent data of clustered genes using identified haplotypes and haplotype patterns. *Frontiers in Genetics* (2014) **5**: doi:10.3389/fgene.2014.00267

59. Roe D. A Detailed View of KIR Haplotype Structures and Gene Families as Provided by a New Motif-Based Multiple Sequence Alignment. *Frontiers in Immunology* (2020) **11**:12.

60. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nature Biotechnology* (2017) **35**:316–319. doi:10.1038/nbt.3820

61. Apache Groovy. Available at: http://groovy-lang.org/

62. dwgsim. Available at: https://github.com/nh13/dwgsim

63. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, Ye K, Guryev V, Vermaat M, van Dijk F, et al. The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics* (2014) **22**:221–227. doi:10.1038/ejhg.2013.118

64. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, Jayaraman J, Wroblewski EE, Trowsdale J, Rajalingam R, et al. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. *The American Journal of Human Genetics* (2016) **99**:375–391. doi:10.1016/j.ajhg.2016.06.023

65. PING Github issue 5. Available at: https://github.com/Hollenbach-lab/PING/issues/5

66. Santos EJM dos, McCabe A, Gonzalez-Galarza FF, Jones AR, Middleton D. Allele Frequencies Net Database: Improvements for storage of individual genotypes and analysis of existing data. *Human Immunology* (2016) **77**:238–248. doi:10.1016/j.humimm.2015.11.013

67. Chen J, Madireddi S, Nagarkar D, Migdal M, Heiden JV, Chang D, Mukhyala K, Selvaraj S, Iii EEK, Brauer MJ, et al. In silico tools for accurate HLA and KIR inference from clinical sequencing data empower immunogenetics on individual-patient and population scales.11.

68. Wright PA. Killer‐cell immunoglobulin‐like receptor assessment algorithms in haemopoietic progenitor cell transplantation: current perspectives and future opportunities. *HLA* (2020)tan.13817. doi:10.1111/tan.13817

Appendix

Glossary

term	meaning
AFA	African American
API	Asian or Pacific Islander
CCS	circular consensus reads
CNV	copy number variation
EM	expectation-maximization
EUR	European
exonic haplotype	DNA content of a subset of exons in order of one parental chromosome
GIAB	Genome in a Bottle
genotype	a set of alleles from a single individual, unordered with unknown parental origin
GoNL	Genomes of the Netherlands
GPL	GNU general public license
haplotig	a subset of a sequence haplotype, derived from assembled reads
haplotype	a set of alleles from a single individual, ordered with known parental origin
HLA	human leukocyte antigen
HIS	Hispanic
IPD-KIR	KIR subsection of the Immuno Polymorphism Database
KIR	killer cell immunoglobulin-like receptors
KIR*IMP	KIR imputation (software)
kmer	a short DNA sequence of length k
KPI	KIR probe interpretation (software)
motif	a short, ordered pattern of DNA sequences
MSA	multiple sequence alignment
NAM	Native American
NCBI	National Center for Biotechnology Information
NK	natural killer (cells)
PA	presence/absence
PCR	polymerase chain reaction
sequence haplotype	full DNA content and order of one parental chromosome
SNP	single nucleotide polymorphism
SSO	sequence-specific oligonucleotide
SSP	sequence-specific primer
structural haplotype	gene content and order of one parental chromosome
WGS	whole genome sequences/sequencing

Capture probes

probe	sequence
1	ATATGGGCCTGGAGTGGAGTTATGGGCCTGGAGTGAAGTTATGGGCCTGGAGGTGGAGATACGGGCCTGGAGTGGAGATATGAGCCTGGAGTGGAGATATGGTCCTGGAGTGGAGATATG
2	TGTCTCCTGCCCATGAGCACCACAGTCAGGCCTTGAGGGGATCTTCTAGGGAGACAACAGCCCTGTCTCAAAACTGGGTTGCCAGCTCCAATGTACCAGCAGCTGGAATCTGAAGGCGTG
3	ACCTCCCTGAGGAAACTGCCTCTTCTCCTTCCAGGTCTATATGAGAAACCTTCTCTCTCAGCCCAGCCGGGCCCCAAGGTTCAGGCAGGAGAGAGCGTGACCTTGTCCTGTAGCTCCCGG
4	GATGATAAAGAGAGACACCTTCTAAACTCACAACCTCTCTTCCTAGGAGTCCACAGAAAACCTTCCCTCCTGGCCCACCCAGGTCCCCTGGTGAAATCAGAAGAGACAGTCATCCTGCAA
5	AGAAAAGTGGTGATTGGCCATCTTTAGTCTCAATGTAAACGGTAATACTGATGAGTGTGGAAAAGGCAGGGAAGAGGATTGACAATAAGTGACACTCATTGTTTTCATCTGAGCTTTGAG
6	AACAAACCTGCATGCATACCTTTAATCTACAGTAAAGGTTGAAGTTATTTAAAAATAGGAAGAAGAATTACCCTATACCTAAAGCTAAGATTTTTCCCTTTGAATATTCGTTTCTTCATC
7	CCACTGGGCCTCATGCAAGGTAGAAAGAGCCTGCGTACGTCACCCTCCCATGATGTGGTCAACATGTAAACTGCATGGGCAGGGCGCCAAATAACATCCTGTGCGCTGCTGAGCTGAGCT
8	GTTTAATGTTCTTTTCTGTCTTTAGTATCAGGGTGATGCCAGCCTTATAGAATGAGTAAAGGCCACCCTGGGCAAACAGTGAGACCCATCCCTTTTTAAAAATTATGAGTTTTACAAATT
9	AAAAATGCTGTGTTTGTTCTGCCTGCATTCCTAACTGGGAGGACAAATGCCTGGGGGCTTGAGAAGGGGAAGGAAGGGGAACATTTTTGAGGGTGGTGTGTTTGTAGAGAAGTTCTACTT
10	CACGAACGGTGAACAAGATGCATTTGGCCTCTGCTCTTGGGACACTGATATTGCAGATGGTTAAATGGGAGGGCAGAAAATGAATGCACAAGTGGACCAATAAATGAATGATCCATTGGG
11	ACTGGTGAAACGTAGGGTTGATTTACACACTAAATGAATGAAAGATGGATATAAGCTATGCTTGTGAGGTAGAATCATTTGCAGGGAGGGCTTGCTGGGTTTGATTTTTCCTAGTAGTTT
12	GACTCCCAGGGCCCAATGTTAGATAACAGAGTGTTGGCCATGAACCAAACTCAAAGATTTCCACTGAGTAGAGGACAGACACCCTCATTTCCTCACCTCTCTCCTGTCTCATGTTCTAGG
13	GGCAGAGTCAAGTTCTGTGGGGACCAGGGTTACACTAGGGTGCTCAAAGCTGGGGTGTGTGGTGGGAAAGTGGTAGGAACAGCAGATCCTCTGAGGACAAAGGTGTTACTCACACACTTC
14	ATAAATACTCTCTTCACTGTTGTGTACACAACATTTTAACTAGTTATTCTGGTTTAAATTTAATATTGACTTTATCTACATATCACAATTGATTACTGTGTACAGACTTTCTTTTCTATT
15	TAGACAAATAATGATAGATATATAATGATATCACAAATAATGATAGACATATAGTTGGATAATGACAGATATATAATGATTGATACACAGATAGGGTATTTATATATTGGCTTATGCAAC
16	GTCAGACAAATGTCTTCCCCATTATTTTCTTCTACATGTTTCATAGGTTCAGGCCTTAGACTCATGTTTTTAATCCATTTTCATTTGATTTTTGTGTAAGGTGACAGGTATAGATGCAGT

17	ATCATATGGAAAATATGAAAGCCCCCTGAACCCACCAGCACAGGCCCTGAAATAGGGAAAGTGCTCTGTTCATCACAAGAAACTTTCCCCCTCACCCAAATCCCCCACCTCACCCCTACT
18	CAGCCCTCCGGGAACTAATAGAGGGGGAACTTGCTAACCCCATCATGTGGGGCAGCATTAATCTATTCATGATGGATCCACCTCCATGACTCAAACACCTTCCCATAGGCCCAAACTTCC

Motif assignments

Probe pair	motif character
13-4	H
3-9	X
12-2	R
7-13	M
10-2	K
2-4	k
3-11	D
7-12	b
6-7	a
2-15	i
11-7	E
1-1	Z
6-4	B
2-7	L
3-12	I
1-13	G
3-14	T
12-10	J
14-5	U
7-15	m
5-8	V
11-2	c
5-4	l
7-1	F
4-3	C
5-6	d
8-3	W
5-3	j
9-10	Y

7-4	S
3-16	A
16-12	N
10-17	O
17-18	P
18-2	Q
4-11	e
1-4	f
3-10	g
10-11	h
7-3	o
3-2	p
9-2	q