Predicting *KIR* Structural Haplotypes with Novel Sequence Signatures from Short-read Whole Genome Sequencing

David Roe[1,2*], Rui Kuang[1,3]

[1] Bioinformatics and Computational Biology, University of Minnesota, 4-192 Keller Hall, 200 Union Street SE, Minneapolis, MN 55455

[2] Center for International Blood and Marrow Transplant Research, 500N 5th St, Minneapolis, MN 55401-1206, USA

[3] Department of Computer Science and Engineering, University of Minnesota, 4-192 Keller Hall, 200 Union Street SE, Minneapolis, MN 55455

* Corresponding author roex0050@umn.edu

Abstract

Keywords: KIR Netherlands WGS immunology

The killer cell immunoglobulin-like receptor (KIR) proteins evolve to fight pathogens and mediate the body's reaction to pregnancy. They also play roles in autoimmune diseases, cancer, transplantation, and immunotherapy. *KIR* genes are under selection pressure for variation at both the structural/haplotype and allele/base levels. Consequently, current reference alleles number in the thousands of exon-level variants, hundreds of full-genes, and dozens of full haplotypes. These multi-faceted variations make it impossible to interpret *KIR* haplotypes from abundant short-read genome sequencing data using existing methods. Here, we developed an efficient computational approach to accurately interpret *KIR* haplotypes with short reads from whole genome sequencing (WGS). We designed synthetic candidate sequence probes of size 25 base pairs (25mers) by analyzing known allele sequences. We then queried whole genome sequences of The Genome of the Netherlands (GoNL) with the designed probes to infer genotypes at the gene level, and imputed the haplotype structures using family relationships and expectation-maximization to resolve the ambiguities in haplotype inference. From the results, we report individual haplotype pair predictions, haplotype frequencies, and novel markers for 19 genes, 24 intergene regions, 18 haplotypes, and 8 half-haplotypes. 92% of GoNL cohort can be explained in 18 haplotypes and the frequencies are similar to expectations from previous studies. We applied the combination of gene, intergene, and haplotype associations to leverage new information from short-read sequences. These methods can also be used to predict *KIR* haplotypes and discover *KIR* markers in other populations, and the interpretation algorithm is deployable as in an easy-to-use container.

Introduction

Human chromosome 19q13.4 contains a ~150-250 kilobase region encoding approximately 16 genes – depending on one's definition – of the natural killer cell immunoglobulin-like receptor (*KIR*) family. These genes are ~4-16 kilobases long and evolved via tandem duplication during primate evolution[1][2]. The *KIR* receptors recognize Human Leukocyte Antigen (*HLA*) class I molecules and contribute to Natural Killer (NK) cell functions via activating or inhibiting signals. These receptor-ligand pairs coevolved under selection pressures from reproduction and pathogenic defense[3], and it is believed that *KIR* genes have undergone a balancing selection via 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[4][5][6]. Mostly European-ancestry *KIR* sequences have been publicly deposited via 34 full-haplotype sequences and approximately 2500 full- or inter-gene sequences[7][5][8]. Haplotype structures are divided into two classes. Class 'A' contains one haplotype and its deleted forms. Class '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 of the haplotypes are variants of the B haplotypes. Except for some 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 regions separated by the relatively large and recombinant *KIR3DP1-KIR2DL4* intergene region.

To date, most large-scale *KIR* genotyping has been based on polymerase chain reaction (PCR) via sequence-specific-oligonucleotide (SSO) hybridization, sequence-specific-primer typing (SSP), or

quantitative (qPCR)[9][10][11][4][12][13][14]. Smaller-scale, but higher-resolution short-read sequencing has been demonstrated via amplicon, long-range PCR, or exon capture[15][16]. And even smaller-scale, but full-haplotype sequencing has been demonstrated with long-read sequencing via fosmid preparation[5]. One group has imputed KIR haplotype structures via microarray-genotyped SNPs[17].

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 largely ignored, mis-interpreted, or under-interpreted in current whole genome sequencing studies. Although one method has been demonstrated to predict partial haplotypes based on exonic sequences[16], no method has thus far demonstrated prediction of full structural haplotypes from WGS at population scale.

This study focuses on the prediction of full structural haplotypes from WGS at population scale. In particular, we propose to systematically evaluate small markers for *KIR* regions, as other researchers have noted that single nucleotide polymorphisms (SNPs) can mark *KIR* regions[18][19][17], and we had also observed potential candidate SNPs in visual analysis of haplotype multiple sequence alignments. SNPs have a long and distinguishable history for marking stretches of haplotypes. Recently, 'kmer'-based approaches have also been widely adopted for association analysis, although in many ways (e.g., design, information content) they are more similar to sequence-specific oligonucleotides (SSO). Popular for graph-building during sequence

assembly and population analysis in metagenomics, sets of sequences of length k ('kmers') can also be used for haplotype-marking like SNPs. Several new bioinformatics tools make such comparisons accurate and computationally tractable for populations of genomes[20][21].

We developed and evaluated a new algorithm to predict *KIR* structural haplotypes from short-read whole genome sequences at population scale. A broad overview of the workflow as depicted in Figure 1 shows how we used The Genome of the Netherlands (GoNL) WGS and previously-reported *KIR* sequences to make individual GoNL haplotype predictions and discover new *KIR* markers. In the experiments, we predicted haplotype frequencies for 244 Dutch mother/father/child families. Also, we systematically associated kmer sequences with gene, intergene, and haplotype regions to discover and quantify their respective markers. New genome-searching algorithms and family-based WGS empowered these new *KIR* discoveries. GoNL researchers gain insight into their population via haplotype frequencies, and future association studies may be empowered with individual haplotype-pair predictions.

Subjects and methods

Figure 1 depicts the key steps in our analytic and experimental methods. In brief, we used reference alleles to discover *KIR*-region-specific 25mers (Figure 1a) for genotyping *KIR* regions (Figure 1b and 1c), from which we made individual haplotype-pair predictions (Figure 1d) and subsequently associated sequence markers to those haplotypes and their constituent regions (Figure 1e).

Collection of GoNL family WGS

This study makes use of data generated by the Genome of the Netherlands Project. A full list of the investigators is available from www.nlgenome.nl. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated July 9, 2009 and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). The sequencing was carried out in collaboration with the Beijing Institute for Genomics (BGI). The samples were provided by The LifeLines Cohort Study[22], and generation and management of GWAS genotype data for it, is supported by the Netherlands Organization of Scientific Research (NWO, grant 175.010.2007.006), the Dutch government's Economic Structure Enhancing Fund (FES), the Ministry of Economic Affairs, the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the Northern Netherlands Collaboration of Provinces (SNN), the Province of Groningen, the University Medical Center Groningen, the University of Groningen, the Dutch Kidney Foundation and Dutch Diabetes Research Foundation.

WGS was obtained for 744 (248 mother, father, child trios) from The Genome of the Netherlands(GoNL)[23], a genome sequencing project whose goal is to map the genetic variation within the population of the Netherlands in 250 family trios (750 individuals). Paired-end sequencing of the whole genomes of the population was done on the Illumina HiSeq 2000 platform. Coverage of the *KIR* region were similar to the previously-reported whole-genome average of ~10-15x. Six individuals from two families (220 and 223) were removed due to data quality issues in one individual from each family.

Collection of reference alleles

All full-length, publicly-reported, *KIR* allele sequences were downloaded from IPD-KIR 2.7.1[24] and Roe et al. 2017. The allele sequences were binned into each gene and intergene region ('Reference Alleles' in Figure 1). *KIR2DL5A* and *KIR2DL5B* were considered separately and also collectively as *KIR2DL5*, and similarly with *KIR2DS3, KIR2DS5,* and *KIR2DS3S5*. This resulted in 774 full-length reference sequences in 43 regions, comprised of 19 genes and 24 intergenes (Supplemental Table 1, 'gene intergene' worksheet, columns A and B).

Discovery of 25mer genotypes markers of *KIR* regions

First, we detected markers of each gene/intergenic region by analyzing 25mers in the reference allele sequences. The candidate markers were selected to be region-ubiquitous, i.e. occurring in every allele sequence of the region, and as region-unique as possible by the number of off-locus (region) hits. Then, these 25mer candidate markers were used to genotype *KIR* short read data (Figure 1b & 1c) and be associated with the *KIR* regions based on the haplotype predictions (Figure 1e). KMC 3[21] was used to find all 25mers present in the reference allele sequences of each region (Figure 1a and Supplemental Table 1, column C). These locus-ubiquitous 25mers were then prioritized by the fewest number of off-locus occurrences. In Supplemental Table 1, column D lists the number of 25mers that are both ubiquitous and unique to each region. The top 100 25mers with fewest off-locus hits were kept; if there were multiple markers tied for 100th-best, all 25mers with that number of hits were kept. Given many 25mers with the same number of off-locus hits, the sizes of the gene sets range from 213 (*KIR2DS1*) to 5285 (*KIR2DL4*), and the sizes of the sets of intergenic regions range from 100 (*KIR3DL3-KIR2DS2*) to 9115 (*KIR3DP1-KIR2DL4*) 25mers. The total is 39,254 25mers in 43 regions (Supplemental Table 1, columns E-G). The 25mer (probe)

sets for the 12 non-framework genes and *KIR3DP1-KIR2DL4* exhibit the best-case scenario: all 25mers are region-ubiquitous and region-unique.

The 25mer set of haplotype – as opposed to gene/intergene – probes was generated by selecting all 25mers from the full-gene and full-intergene allele sequences that occur 50 or fewer times, to exclude 25mers likely to cross genes and therefore haplotypes. From this, the gene and intergene candidates were removed as candidates on the theory that genes markers cannot be unique to a haplotype if the genes themselves are not. This resulted in 86 460 candidate haplotype markers.

Genotyping with 25mer probing of WGS short reads

The short reads of each GoNL individual's genome were searched to match the discovered 25mer markers using KMC 3, with workflows implemented in Nextflow[25] and Apache Groovy[26] (Figure 1b). Next, presence/absence (P/A) genotyping for each gene/intergene region was made by taking the peak count of the hits with each gene's markers: if the majority of the markers had a hit count of zero, the gene was considered absent; otherwise, if the majority of makers had a hit count greater than zero, the gene/intergene was considered present (Figure 1c, details in Supplementary Figure 1). Thus, a GoNL genotype call table was generated.

Haplotyping with presence/absence genotypes

The individual haplotype-pair predictions were made by fitting their gene P/A genotypes into all possible pairs of 18 reference haplotypes (Figure 1d, Supplemental Table 2). These potentially-ambiguous predictions were reduced by family patterns and then further by the EM (Expectation-Maximization)-based methods as described and used in Vierra-Green 2012[27]. In each family,

the impossible haplotype-pairs among the trio are removed from the candidate pairs for EM inference. After EM, all individuals were assigned a single most likely haplotype-pair assignment. Haplotype frequencies were calculated from the individual haplotype-pair predictions.

Detecting significant 25mer signatures

The individual haplotype-pair predictions in family-consistent individuals allowed calculation of significance tests between 25mers and regions using Fisher's Exact Test (Figure 1e). The tests were performed for both the gene/intergene regions genotyped in Figure 1c and for the haplotype/half-haplotypes inferred by the EM in Figure 1d.

Refining haplotypes with 25mer signatures

In this final step, the haplotype interpretations were refined, using only the significantly-associated 25mers for the final haplotype-pair predictions. This time, markers for genes, intergenes, and haplotypes were used. First, gene-only genotypes were used to predict potentially-ambiguous haplotype pairs (as in the discovery phase). Then, these haplotype pairs were reduced using intergene markers. Any remaining ambiguity was removed by the EM. Family relationships were used for validation, not prediction.

Results

Allele frequencies

The frequencies of all the *KIR* genes in the genotyping by the 25mer hits in the WGS short reads are within previously reported ranges for Europeans as aggregated and reported via The Allele

Frequency Database[28], except *KIR2DL5B* (45% vs. 23-39%) and *KIR3DL2* (99% vs. 100-100%) (Table 1). *KIR3DL2* is only trivially different, and *KIR2DL5B* was only reported in small cohorts in three populations: Italy, Romania, and Sweden. The frequency of *KIR2DL5* (all alleles of *KIR2DL5A* and *KIR2DL5B*) was within expectation.

Haplotype frequencies

Per-individual haplotype-pair predictions are included in Supplemental Table 3. Three lists of haplotype pairs are provided: one for the initial fitting of all possible haplotype pairs that could explain the genotype; another that reduces those possibilities by family relationships; and one for the EM-reduced final haplotype pair predictions. Every interpretable individual (under the set of reference haplotypes) had a single mostly-likely pair of haplotypes for the final prediction. The haplotype frequencies are listed in Table 2, column C; they are aggregated from the individual predictions. Columns D-E in Table 2 present the frequencies in the context of our previous study of thousands of a more-generally-European background. They are very similar, but with a few differences. The main difference relative to that study is a higher uninterpretable rate (+7.73%). The next greatest difference is haplotype 4 (cB02~tA01), whose frequency delta is -5.68%.

Detected 25mer markers

The 25mers significantly ($p<10^{-6}$) associated with gene and intergene regions were detected, except the 4 framework genes (*KIR3DL3, KIR3DP1, KIR2DL4, KIR3DL2*) and the *KIR3DP1-KIR2DL4* intergene region. Figure 2 depicts the associations between the sequences (rows) and gene/intergene regions of the cA01~tA01 and cB01~tB01 haplotypes, which also provides the sort order for both axes. The cells are colored by $\log_{10}$ p-value, from 0 (brightest yellow: not significant)

to 10^{-19} (darkest blue: strongly significant). The patchy, blue diagonal from lower-left to upper right is a rough identity line between sequences that originate from a given region and associate to the same region. Although the gene/intergene categories on the x-axis have a consistent width, the region widths on the y-axis vary because some regions are the source of more/fewer significant markers.

Similar to the gene/intergene regions, significantly ($p<10^{-6}$) associated 25mers were discovered for 16 of the 18 reference haplotypes. Of the 8 haplotypes with a frequency over 1%, all but haplotype cA01~tB01 (with centromeric *KIR2DS5*) had significant markers; their distributions are shown in Figure 3. Similarly, 9 of the 14 centromeric or telomeric half-haplotype regions had significant markers, and the distributions of all 4 regions with frequencies >1% are shown in Figure 4.

Discussion

The main findings of this study are: 1) all non-framework genes and non-intervening (i.e., non-*KIR3DP1-KIR2DL4*) intergenes contain many 25mers that uniquely define each region and are ubiquitous to all reported alleles 2) these sequence libraries can be used to genotype individuals from populations with short reads from whole genome sequencing or *KIR*-capture sequencing 3) haplotype pairs can be predicted from the genotypes, and they seem to be accurate based on expectations from previous reports and family relationships 4) 25mers also significantly mark haplotypes and half-haplotypes, at least in this population.

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 (kmer) library, we used KMC 3 to find all 25mers *ubiquitous* to all alleles of each gene and intergene region, and then identify which of them were *unique* (or closest to unique) to each region. This analysis resulted in a library of hundreds-to-thousands of marker 25mers per region (Supplemental Table 1), except for the framework genes (*KIR3DL3, KIR3DP1, KIR2DL4, KIR3DL2*) and the *KIR3DP1-KIR2DL4* intergene region which contained no ubiquitous and unique 25mers. The lack of markers for these regions is not surprising: they are the only regions found in most *KIR* haplotypes, as their 'framework' role indicates. 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 k size other than 25, since the choice gives a reasonable number of significant markers and their lack of off-*KIR* hits as tested the GoNL population WGS confirms the effectiveness as gene/intergenic markers. 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.

The associations between the sequence origin and target-gene/intergene as depicted in Figure 2 provide several interesting interpretations. All the 25mer markers (rows) exhibit significant association with at least one marker. The dark blue along the diagonal from lower-left to upper-right shows that the strongest associations are between 25mers that originate in non-framework regions and mark their own region: no surprise, except possibly that the framework genes had no

associations. The vertical yellow gaps between blue regions show the lack of associations *to* the framework regions, and the gaps between horizontal blue regions show that there are more associations *from* the framework gene and intervening *KIR3DP1-KIR2DL4* region, although their associations are generally less strong.

The interesting observations of the marker rows include how the 25mer markers in and around *KIR2DP1* and *KIR2DL1* mark cA much stronger than cB. Sequences from tA do not mark cA well, but tB markers do associate well with regions in cB. Markers from both *KIR2DS3* (especially) and *KIR2DS5* strongly mark tB, while *KIR2DS5* (but not *KIR2DS3*) marks cB. Also, the mutual-exclusivity of their alleles is evident in the small checkerboard pattern along the diagonal; also, *KIR2DS2* and *KIR2DS2-KIR2DL2* (centromeric) both strongly associate with *KIR2DS5* but not *KIR2DS3*. The large (~13kb) *KIR3DP1-KIR2DL4* intergenic region contains the largest number of significant associations, although they are mostly among the weakest; markers in this region associate with *KIR2DL1*, *KIR2DL1-KIR2DP1*, and *KIR2DP1* best, although some makers from the other regions have stronger associations than 25mers from this intergenic region.

The interesting observations of the gene/intergene-region columns in Figure 2 include the finding that the paralogous gene and intergene regions of *KIR2DL5-KIR2DS3* and *KIR2DL5-KIR2DS5* are marked by most B-only regions; the centromeric *KIR2DS2* and *KIR2DS2-KIR2DL2* mark the paralogous *KIR2DL5-KIR2DS5* stronger than *KIR2DL5-KIR2DS3*, while in the telomere the opposite is true: the telomeric *KIR2DL4-KIR3DS1* and *KIR3DS1* associate with *KIR2DL5-KIR2DS3* stronger than *KIR2DL5-KIR2DS5*. cA regions are marked best by *KIR2DP1*, while conversely, each tA region marks itself best. Regions in cB also widely associate with the markers

in the region, with the strongest associations in *KIR3DL3-KIR2DS2* and *KIR2DS2*; interestingly, tB is best marked by both the centromeric *KIR2DL2-KIR2DL5* as well as the more expected telomeric *KIR2DL4-KIR3DS1* and *KIR3DS1*. The four framework genes and centromere-telomere-intervening region of *KIR3DP1-KIR2DL4* show no significant 25mer associations.

There are several differences between our genotyping approach and the approach in Norman et al.[16], including the way the reference markers were created, the definitions of haplotypes, and the approach to predict the haplotypes. In Norman et al.[16], ten randomly picked 32mer markers were used for genotyping the *KIR* gene content whereas we retained all the hundreds-to-thousands of gene/intergene-specific kmers as candidate markers. We believe that the expanded set of gene/intergene-specific markers is likely to capture the signal from a wider variety of alleles than a random subset, and the kmer tools have since allowed efficient searching of many markers. In addition, we discovered and utilized novel 25mers haplotype markers directly from the WGS, without prior assumptions about the haplotypes they might mark. Bioinformatics tools that allow efficient searching of WGS allow the information to determine the number of relevant markers, as opposed to only using previously-reported alleles. The summary statistics for the gene-specific sequences are given in Supplemental Table 1; their frequencies are in given in Table 1; and the 25mers themselves in Supplemental 4. Although we could also predict the gene copy numbers via the probe hit counts similar as Norman et al., we preferred to use our previously-validated methods to estimate haplotype pairs, using the family-based relationships in GoNL to improve the haplotype (and therefore, gene copy number) predictions and therefore be incorporated prior to EM. We believe this haplotype-first approach is more robust to the variation and quality of large-scale sequencing projects in diverse cohorts.

The haplotype frequencies in Table 2 are presented in the context of our previous study of thousands of a more-generally-European background. The main difference relative to that previous study is a higher uninterpretable rate (+7.73%). The second greatest differences are in haplotype 4 (cB02~tA01) whose frequency is -5.68% and haplotype 3 (cA01~tB01_2DS5), whose frequency delta is -3.43%. The haplotype with the highest positive delta is haplotype 6 (cB01~tA01) at +2.84%. The higher rate of uninterpretable individuals could be explained by several factors. One reason may be the true population differences. Our previous study was a European-American population of donors and recipients of hematopoietic stem cell transplants. This source may enrich the population for more common *HLA*, and therefore more common *KIR* as well. Given that haplotyping from presence/absence genotyping is a parsimonious task relative to a short set of references, it is probably safe to assume that the true uninterpretable rate (relative to this reference set) for GoNL (and Europeans more generally) is closer to the 8% of this study. The data supports the usual 'long tail' of infrequent immune-system haplotypes, as none of the genotypes of 54 uninterpretable individuals occur more than twice.

Our approach also relates to a SNP-based haplotyping method by Vukcevic, et al.[17]. The differences are that the previous study was limited to approximately 1000 potential markers from a SNP array for multiple-marker associations, and their definition of haplotype is by gene copy number, which is less restrictive than our structural haplotypes and potentially confounding/ambiguous because haplotype-pairs as defined by gene copy number are more ambiguous than haplotype-pairs by both gene copy numbers *and order*. In addition, kmer markers derived from WGS are more informative than SNP arrays. SNPs are fixed single nucleotide

polymorphisms surrounded by a certain amount of monomorphism, whereas kmers can incorporate any manner of divergence throughout its sequence. kmer discovery from WGS allows markers not previously reported.

This study coherently integrates high-quality public *KIR* haplotype and allele sequences, algorithms for analyzing high-throughput sequences, and high coverage WGS data from the GoNL project. We have used these resources to contribute *KIR* interpretations to GoNL, to identify contribute informative markers for future studies, and to develop algorithms that can be reapplied to other populations. The software to interpret haplotype-pairs from WGS is managed in GitHub[29] and executable via a Docker container[30].

References

1. 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. Available: http://link.springer.com/article/10.1007/s002510050620

2. Martin AM, Kulski JK, Gaudieri S, Witt CS, Freitas EM, Trowsdale J, et al. 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

3. Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. Nat Rev Immunol. 2013;13: 133–144. Available: http://www.nature.com/nri/journal/v13/n2/abs/nri3370.html

4. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. Genome Res. 2012;22: 1845–1854. doi:10.1101/gr.137976.112

5. Pyo C-W, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, et al. 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. Matsunami H, editor. PLoS ONE. 2010;5: e15115. doi:10.1371/journal.pone.0015115

6. Manser AR, Weinhold S, Uhrberg M. Human KIR repertoires: shaped by genetic diversity and evolution. Immunol Rev. 2015;267: 178–196. Available: http://onlinelibrary.wiley.com/doi/10.1111/imr.12316/full

7. Roe D, Vierra-Green C, Pyo C-W, Eng K, Hall R, Kuang R, et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. Genes Immun. 2017; doi:10.1038/gene.2017.10

8. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2014; doi:10.1093/nar/gku1161

9. Norman P, Cook M, Carey BS, Carrington CF, Verity D, Hameed K, et al. SNP haplotypes and allele frequencies show evidence for disruptive and balancing selection in the human leukocyte receptor complex. Immunogenetics. 2004;56. doi:10.1007/s00251-004-0674-1

10. Houtchens KA, Nichols RJ, Ladner MB, Boal HE, Sollars C, Geraghty DE, et al. High-throughput killer cell immunoglobulin-like receptor genotyping by MALDI-TOF mass spectrometry with discovery of novel alleles. Immunogenetics. 2007;59: 525–537. doi:10.1007/s00251-007-0222-x

11. Vilches C, Castaño J, Gómez-Lozano N, Estefanía E. Facilitation of KIR genotyping by a PCR-SSP method that amplifies short DNA fragments. Tissue Antigens. 2007;70: 415–422. doi:10.1111/j.1399-0039.2007.00923.x

12. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, et al. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. Hum Mol Genet. 2010;19: 737–751. doi:10.1093/hmg/ddp538

13. Jiang W, Johnson C, Simecek N, López-Álvarez MR, Di D, Trowsdale J, et al. qKAT: a high-throughput qPCR method for KIR gene copy number and haplotype determination. Genome Med. 2016;8. doi:10.1186/s13073-016-0358-0

14. Nelson WC, Pyo C-W, Vogan D, Wang R, Pyon Y-S, Hennessey C, et al. An integrated genotyping approach for HLA and other complex genetic systems. Hum Immunol. 2015; doi:10.1016/j.humimm.2015.05.001

15. Maniangou B, Legrand N, Alizadeh M, Guyet U, Willem C, David G, et al. Killer Immunoglobulin-Like Receptor Allele Determination Using Next-Generation Sequencing Technology. Front Immunol. 2017;8. doi:10.3389/fimmu.2017.00547

16. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. Am J Hum Genet. 2016;99: 375–391. doi:10.1016/j.ajhg.2016.06.023

17. Vukcevic D, Traherne JA, Næss S, Ellinghaus E, Kamatani Y, Dilthey A, et al. Imputation of KIR Types from SNP Variation Data. Am J Hum Genet. 2015;97: 593–607. doi:10.1016/j.ajhg.2015.09.005

18. Pontikos N, Smyth DJ, Schuilenburg H, Howson JM, Walker NM, Burren OS, et al. A hybrid qPCR/SNP array approach allows cost efficient assessment of KIR gene copy numbers in large samples. BMC Genomics. 2014;15: 274. doi:10.1186/1471-2164-15-274

19. Vendelbosch S, de Boer M, van Leeuwen K, Pourfarzad F, Geissler J, van den Berg TK, et al. Novel insights in the genomic organization and hotspots of recombination in the human KIR locus through analysis of intergenic regions. Genes Immun. 2014; doi:10.1038/gene.2014.68

20. Pérez N, Gutierrez M, Vera N. Computational Performance Assessment of k-mer Counting Algorithms. J Comput Biol. 2016;23: 248–255. doi:10.1089/cmb.2015.0199

21. Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 2017; doi:10.1093/bioinformatics/btx304

22. Lifelines Biobank [Internet]. Available: http://www.lifelines.nl

23. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. Nat Rev Genet. 2010;11: 559–571. doi:10.1038/nrg2814

24. Robinson J. IPD--the Immuno Polymorphism Database. Nucleic Acids Res. 2004;33: D523–D526. doi:10.1093/nar/gki032

25. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35: 316–319. doi:10.1038/nbt.3820

26. Apache Groovy [Internet]. Available: http://groovy-lang.org/

27. Vierra-Green C, Roe D, Jayaraman J, Trowsdale J, Traherne J, Kuang R, et al. Estimating KIR Haplotype Frequencies on a Cohort of 10,000 Individuals: A Comprehensive Study on Population Variations, Typing Resolutions, and Reference Haplotypes. Toland AE, editor. PLOS ONE. 2016;11: e0163973. doi:10.1371/journal.pone.0163973

28. González-Galarza FF, Takeshita LYC, Santos EJM, Kempson F, Maia MHT, Silva ALS da, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Res. 2015;43: D784–D788. doi:10.1093/nar/gku1166

29. Roe D. KIR Probe Intepretation [Internet]. GitHub; Available: https://hub.docker.com/r/droeatnmdp/kpi/

30. Roe D. KPI Docker container [Internet]. Available: https://hub.docker.com/r/droeatnmdp/kpi/

Table and Figure Legends

Figure 1. Workflow. Boxes a-e label the five steps. a) 25mers were generated (red arrow) that are ubiquitous to all alleles of each gene and intergene (blue arrow) and as unique as possible (minimizing the off-region occurrences) to each region. b) Those candidate 25mers (blue arrow) were counted (red arrow) in the WGS of each GoNL individual (blue arrow). c) Presence/absence was called for each gene/intergene (red arrow) based on the 25mer marker counts (blue arrow), from which (d) (blue arrow) haplotype pairs were predicted (red arrow). e) Then, the 25mers (blue arrow) were associated with genes, intergenes, haplotypes, and half-haplotypes (red arrow).

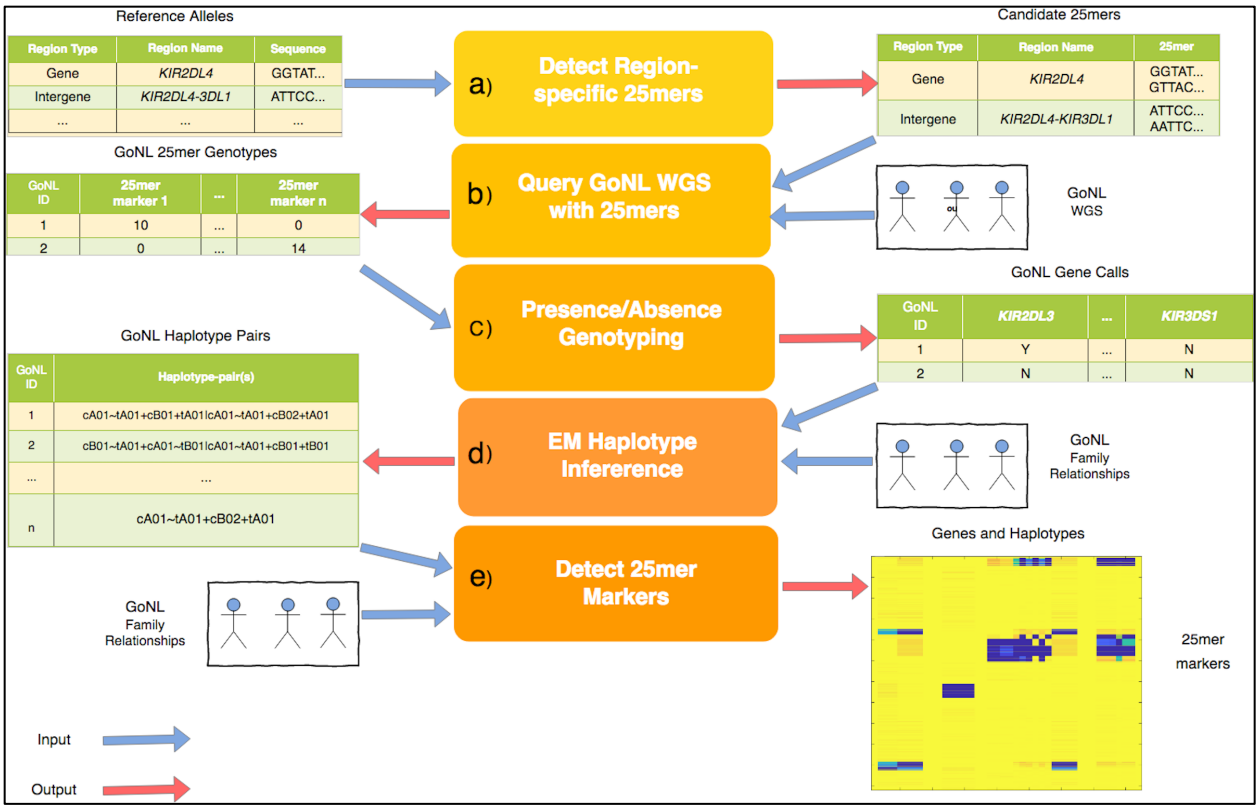

Figure 2. p-value heat map of 25mer-gene/intergene associations. The y-axis represents 56,071 25mer sequences, each with at least one significant ($p<10^{-6}$) association with a gene or intergene region; they are sorted along the vertical axis by reverse order of the cA01~tA01 and cB01~tB01 haplotypes. The x-axis represents the 42 gene and intergene regions ordered by the cA01~tA01 and cB01~tB01 haplotypes. The diagonal from lower-left to upper-right roughly associates origin and destination sequences from the same region. Each cell is colored by significance, from 0 to -19, representing the log p-values. The minimum p-value was limited to 10^{-21}; values lower than this were rounded up.

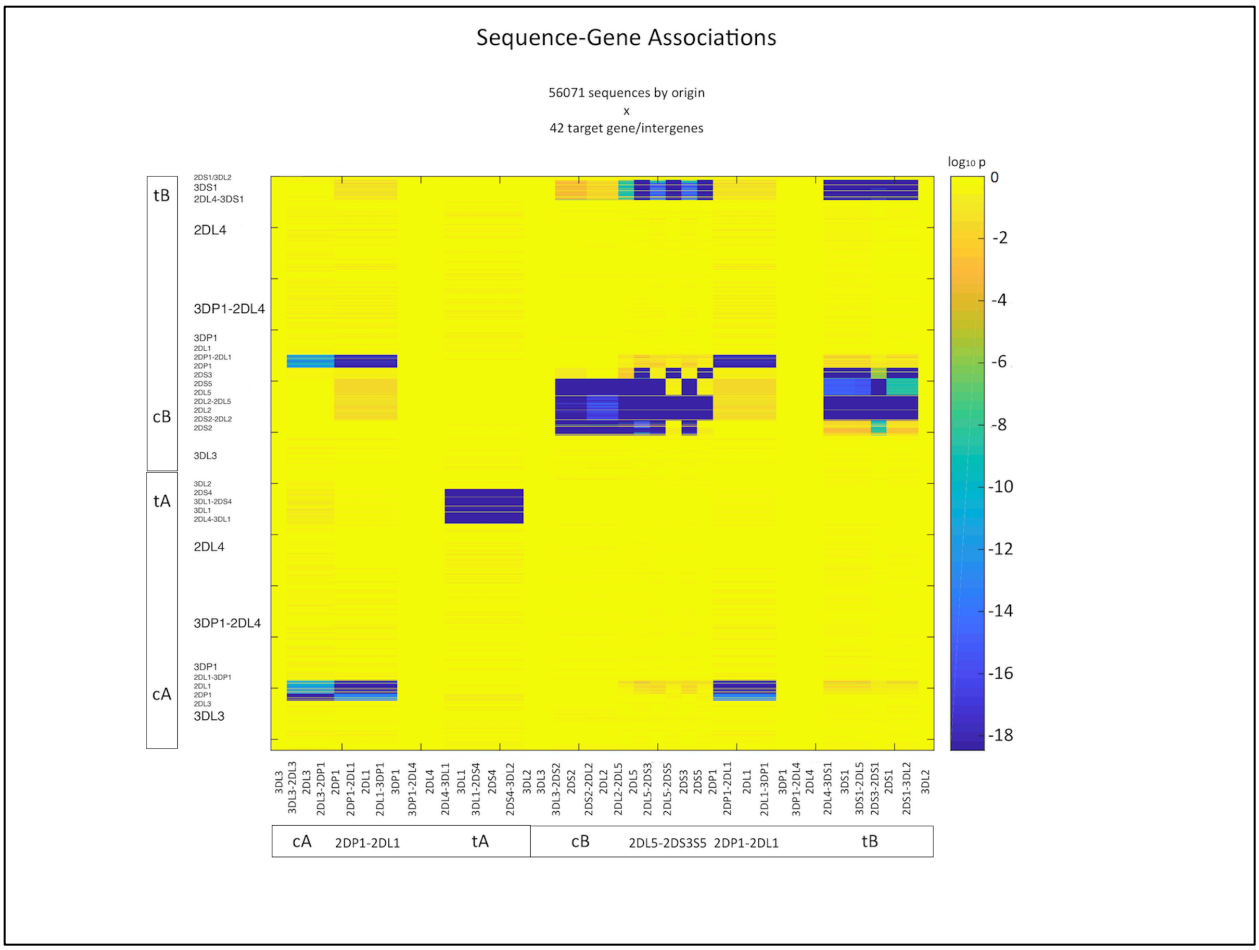

Figure 3. Distribution of p-values of significantly-associated 25mer per haplotype. The range of p-values of significantly-associated ($p<10^{-6}$) sequences are shown with 'Box and Whisker' plot per haplotype with frequency > 1%. The p-value on the y axis ranges from 0 to 10^{-16}, and the number of sequences per region range from 162 for cB01~tB01 (*KIR2DS3*) to 12593 for cB01~tA01 and cB01~tB01 (*KIR2DS5*). No significant associations were found for haplotypes 3, 4, and 6.

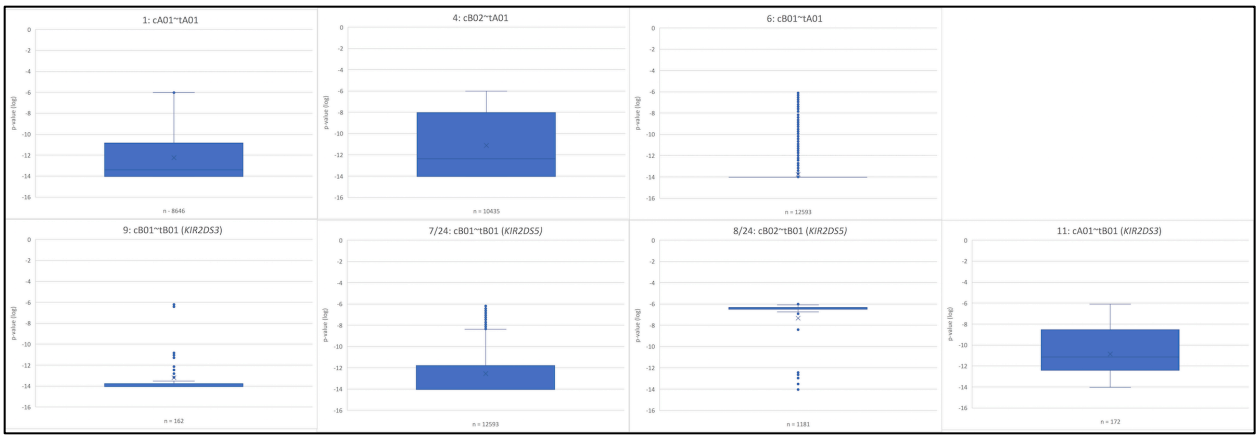

Figure 4. Distribution of p-values of significantly-associated 25mer sequences per half-haplotype. The range of p-values of significantly-associated ($p<10^{-6}$) sequences are shown with 'Box and Whisker' plots per half-haplotype. The p-value on the y axis ranges from 0 to 10^{-16} except, cB02 and tB01 (*KIR2DS3*) which end at 10^{-10} and 10^{-14} respectively. The B regions have low thousands of 25mers for the cB01 and tB02 centromeric regions and low ten-thousands for the both variants of tB01. No significant associations were found for regions cA01, tA01, tB03, tB04, and tB05.

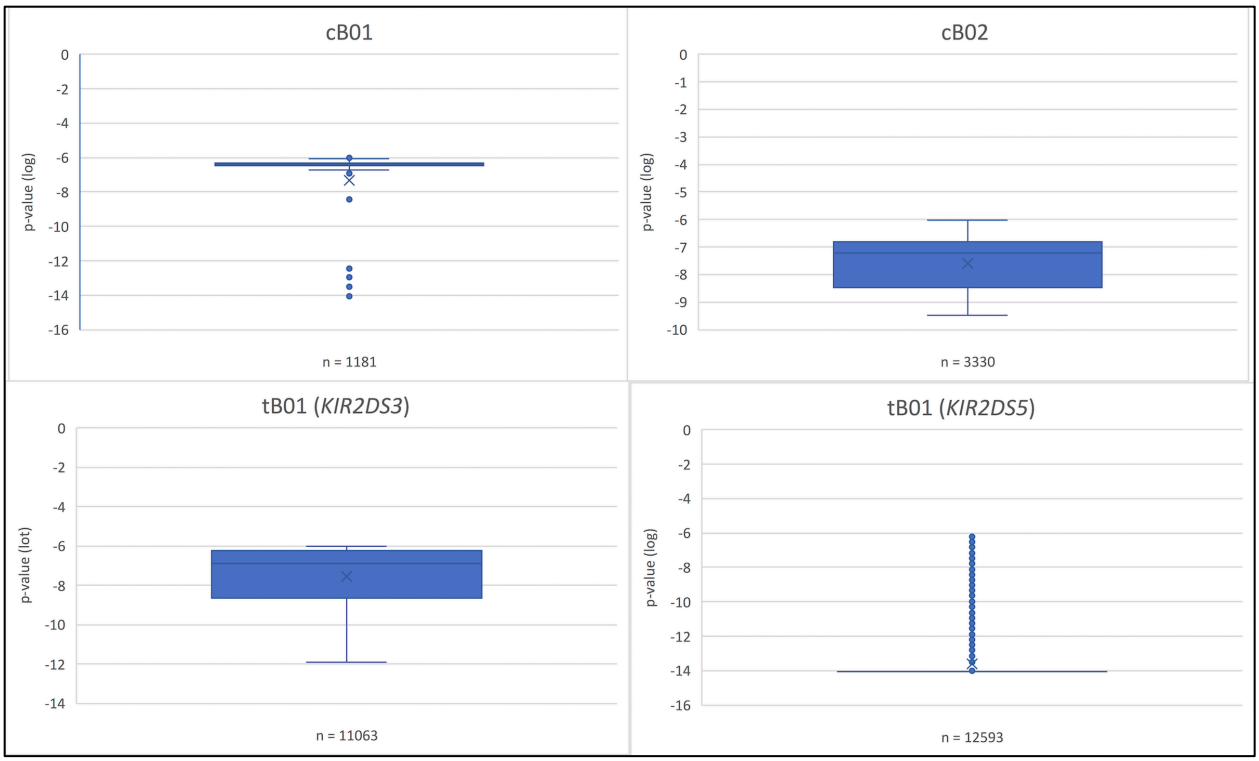

Table 1. Gene and intergene frequencies. Column A contains the number of individuals with at least one predicted copy of the gene and intergene regions in column B. Column C contains the frequency, which is compared with the range of European frequencies in The Allele Frequencies Database in columns D through F.

occurrence count	region	occurrence frequency	allelefrequencies.net range	within range
720	2DL1	97%	89-98%%	TRUE
723	2DL1-3DP1	97%	N/A	N/A
356	2DL2	48%	41-61%	TRUE
250	2DL2-2DL5B	34%	N/A	N/A
326	2DL2-3DP1	44%	N/A	N/A
677	2DL3	91%	57-94%	TRUE
676	2DL3-2DP1	91%	N/A	N/A
741	2DL4	100%	100%	TRUE
738	2DL4-3DL1	99%	N/A	N/A

247	2DL4-3DS1	33%	N/A	N/A
337	2DL5	45%	40-61%	TRUE
280	2DL5A-2DS5	38%	N/A	N/A
337	2DL5B	45%	23-39%	FALSE [Italy, Romania, Sweden reported]
302	2DL5B-2DS5	41%	N/A	N/A
718	2DP1	97%	94-98%	TRUE
707	2DP1-2DL1	95%	N/A	N/A
452	2DP1-2DS1	61%	N/A	N/A
240	2DS1	32%	15-48%	TRUE
239	2DS1-3DL2	32%	N/A	N/A
381	2DS2	51%	42-60%	TRUE
373	2DS2-2DL2	50%	N/A	N/A
191	2DS2-2DP1	26%	N/A	N/A
205	2DS3	28%	23-37%	TRUE
191	2DS3-2DP1	26%	N/A	N/A
246	2DS3-2DS1	33%	N/A	N/A
341	2DS3S5	46%	N/A	N/A
707	2DS4	95%	83-96%	TRUE
740	2DS4-3DL2	99%	N/A	N/A
192	2DS5	26%	21-36%	TRUE
268	2DS5-2DP1	36%	N/A	N/A
315	2DS5-2DS1	42%	N/A	N/A
700	3DL1	94%	76-97%	TRUE
704	3DL1-2DS4	95%	N/A	N/A
737	3DL2	99%	100%	FALSE
741	3DL3	100%	100%	TRUE
744	3DL3-2DL3	100%	N/A	N/A
377	3DL3-2DS2	51%	N/A	N/A
741	3DP1	100%	97-100%	TRUE
741	3DP1-2DL4	100%	N/A	N/A
255	3DS1	34%	26-46%	TRUE

252	3DS1-2DL5A	34%	N/A	N/A

Table 2. Haplotype frequencies. Column A contains the haplotype identifiers for this study. Column B contains some conventional nomenclature for the haplotype names. Column C contains the predicted haplotype frequencies, aggregated from the individual predictions. Column D contains the haplotype frequency predictions from Vierra-Green et al. 2016, a cohort of thousands of Europeans. Column E reports the difference between the frequencies in this study minus the 2016 frequencies.

haplotype	nomenclature	frequency	2016 frequency	delta
1	cA01~tA01	56%	56%	-0.45%
3	cA01~tB01_2DS5	8%	11%	-3.43%
4	cB02~tA01	10%	15%	-5.68%
6	cB01~tA01	8%	5%	2.84%
7/24	cB01~tB01_2DS5	1%	3%	-2.29%
8/19	cB02~tB01_2DS5	0%	1%	-0.76%
9	cB01~tB01_2DS3	1%	3%	-2.89%
11	cA01~tB01_2DS3	2%	1%	1.38%
12	cB04~tB03_2DS5	1%	1%	0.12%
13	cB01~tB05	1%	1%	0.33%
14	cA01~tB04	1%	1%	0.44%
15	cB05~tB01	1%	0%	0.57%
16	cA01~tB05	1%	0%	0.37%
17	cB02~tB01_2DS3	2%	0%	2.42%
18	cB04~tB03_2DS3	0%	0%	-0.11%
21	cB05~tA01	0%	0%	-0.11%
22		0%	0%	-0.23%
23	cA01~tB06	0%	0%	0.04%
uninterp		8%	1%	7.73%
other		0%	0%	-0.30%
		100%	100%	

Supporting Information

Below are the descriptions, but for size reasons, only the file depicting the reference haplotypes has been included in the preprint. Feel free to contact the corresponding author for more information.

Supplemental Figure 1. Algorithm details**.** Command to query the markers per genome and details as to how each region was genotyped.

Supplemental Table 1. Counts of reference alleles and 25mers. For the 'gene intergene' worksheet, the 'locus' column contains the name of the gene or intergene region. 'num full-sequences reference alleles' indicates the size of the pool of reference alleles from which the 25mers were generated. 'num 25mers ubiquitous to region' contains the number of 25mers that are found in every allele of the locus. 'num 25mers ubiquitous and unique to region' indicates the number of 25mers that are in every allele of that locus and not in any reference alleles of any other locus. 'num 25mers in final set' contains the number of 25mers in the final probe set. It is the number of ubiquitous-and-unique probes plus the ubiquitous-only probes up to a count of 100. The number is often greater than 100 because all ties were kept. The 'haplotype' worksheet contains the number of haplotype candidate 25mers broken down by their region of origin.

Supplemental Table 2. Reference haplotypes. Each haplotype is represented by a pseudo-numeric ID in column A and a nomenclature term in column B. Some haplotypes have some structural notes in column C. The other columns indicate the presence/absence of 16 genes.

haplotype	nomenclature	3DL3	2DS2	2DL2	2DL3	2DP1	2DL1	3DP1	2DL4	3DL1	3DS1	2DL5	2DS3	2DS5	2DS4	2DS1	3DL2
1	cA01~tA01	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0	1
3	cA01~tB01_2DS5	1	0	0	1	1	1	1	1	0	1	1	0	1	0	1	1
4	cB02~tA01	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1
6	cB01~tA01	1	1	1	0	1	1	1	1	1	0	1	1	0	1	0	1
7/24	cB01~tB01_2DS5	1	1	1	0	1	1	1	1	0	1	2	1	1	0	1	1
8/19	cB02~tB01_2DS5	1	1	1	0	0	0	1	1	0	1	1	0	1	0	1	1
9	cB01~tB01_2DS3	1	1	1	0	1	1	1	1	0	1	2	2	0	0	1	1
11	cA01~tB01_2DS3	1	0	0	1	1	1	1	1	0	1	1	1	0	0	1	1
12	cB04~tB03_2DS5	1	1	1	0	0	0	0	0	0	0	1	0	1	0	1	1
13	cB01~tB05	1	1	1	0	1	1	2	2	1	1	1	1	0	1	0	1
14	cA01~tB04	1	0	0	1	2	2	2	2	0	2	2	1	1	0	1	1
15	cB05~tB01	1	1	0	0	1	1	1	1	0	1	1	0	1	0	1	1
16	cA01~tB05	1	0	0	1	1	1	2	2	1	1	0	0	0	1	0	1
17	cB02~tB01_2DS3	1	1	1	0	0	0	1	1	0	1	1	1	0	0	1	1
18	cB04~tB03_2DS3	1	1	1	0	0	0	0	0	0	0	1	1	0	0	1	1
21	cB05~tA01	1	1	0	0	1	1	1	1	1	0	0	0	0	1	0	1
22	cA03~tB02 wo 3DL2	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0
23	cA01~tB06	1	0	0	1	2	2	2	2	1	1	1	1	0	1	0	1

Supplemental Table 3. Individual haplotype pair predictions. Columns A and B contain the family name and relationship. Columns C-E contain haplotype list (hl) pairs separated with a '|'. For example, '1+3|16+3' means the haplotype prediction is either haplotypes 1 and 3 or 16 and 3. The hl in column C represents all possible haplotype pairs fitting the P/A genotypes. Column D hl is the C hl reduced by family relationships. E is D hl reduced by EM. Columns G and H count the number of possible haplotype pairs in columns in the first (column C) and final (column E) respectively. Columns I and J show the average of the two.

Supplemental Table 4. Gene/intergene-25mer associations. A tab-separated text table of 56072 25mer sequence markers (column A), their originating gene or intergene region (column B), and p-values of associations with all genes and intergene regions (columns C-end). 14153 of the sequences have at least one significant ($p<10^{-6}$) association.

Supplemental Table 5. Haplotype-25mer associations. A tab-separated text table of 86643 25mer sequence markers (column A), their most significant association (column B), and p-values of associations with all haplotypes and half-haplotypes (columns C-end). 57443 of the sequences have at least one significant ($p<10^{-6}$) association.