KIR gene content imputation from single-nucleotide polymorphisms in the Finnish population

Jarmo Ritari[1]*, Kati Hyvärinen[1], Jukka Partanen[1], Satu Koskela[1]*

[1]Finnish Red Cross Blood Service, Helsinki, Finland

*Correspondence: jarmo.ritari@veripalvelu.fi; satu.koskela@veripalvelu.fi

Abstract

The killer cell immunoglobulin-like receptor (KIR) gene cluster on chromosome 19 encodes cell surface glycoproteins that bind class I human leukocyte antigen (HLA) molecules as well as some other ligands. Through regulation of natural killer (NK) cell activity KIRs participate in tumour surveillance and clearing viral infections. KIR gene gene copy number variation associates with the outcome of transplantations and susceptibility to immune-mediated diseases. Inferring KIR gene content from genetic variant data is therefore desirable for immunogenetic analysis, particularly in the context of growing biobank genome data collections that rely on genotyping by microarray. Here we describe a stand-alone and freely available gene content imputation for 12 KIR genes. The models were trained using 818 Finnish biobank samples genotyped for 5774 KIR-region SNPs and analysed for KIR gene content with targeted sequencing. Cross-validation results demonstrate a high mean overall accuracy of 99.2% (95% CI: 97.8-99.7%) which compares favourably with previous methods including short-read sequencing based approaches.

Introduction

Killer cell immunoglobulin-like receptors (KIRs) regulate the activity of natural killer (NK) cells and a subset of T cells via inhibitory and activating signals. Through their KIR molecules NK cells detect phenotypic change in a target cell. KIRs recognise human leukocyte antigen (HLA) class I molecules as cognate ligands, limiting to particular HLA allotypes within the serological HLA-C1 and C2 allele groups (Wroblewski, Parham & Guethlein, 2019) HLA-Bw4 motif, HLA-A3/11, HLA-G and HLA-F (Garcia-Beltran et al., 2016). The functional difference between inhibitory and activating KIRs is determined by the presence or absence of a cytoplasmic immunoreceptor tyrosine-based inhibitory (ITIM) protein motif, respectively. In the absence of constitutive signaling conveyed by an inhibitory KIR binding to its class I ligand, NK cell cytotoxic activity and cytokine production are triggered (Lanier, 2008).

According to the missing-self hypothesis, NK cells recognise tumour or virally infected cells that attempt to evade T cell mediated immunity by downregulating their cell surface HLA-molecules that present intracellular antigens to T cells.

Activating KIRs, in contrast, are thought to recognise surface molecules indicative of aberrant host cell activity such as an exceptionally high surface density of HLA class I molecules even though in some cases the ligand remains unknown (Ivarsson, Michaëlsson & Fauriat, 2014). Activating KIRs have lower affinity to their ligands than inhibitory KIRs (Stewart et al., 2005) most likely owing to NK cell education to maintain self-tolerance. However, upon receiving a sufficient positive stimulus, they are able induce NK cell activation and target cell lysis. A vast majority of genetic associations of KIRs with cancer, autoimmunity and infectious diseases are attributed to variation in activating KIRs (Parham & Guethlein, 2018).

The KIR gene cluster on the human chromosome 19q13.4 encodes fifteen relatively homologous KIR genes and two pseudogenes, constituting two main haplotypes: A and B (https://www.ebi.ac.uk/ipd/kir/sequenced_haplotypes.html). The group A haplotype consists of functional *KIR3DL3, KIR2DL3, KIR2DL1, KIR2DL4, KIR3DL1, KIR2DS4* and *KIR3DL2* genes of which all except *KIR2DS4* are inhibitory. The group B haplotype, on the other hand, is more diverse being characterised by the presence of at least one of *KIR2DS2, KIR2DL2, KIR2DL5, KIR2DS5, KIR3DS1, KIR2DS3* or *KIR2DS1* genes (Bashirova et al., 2006). Thus, the group B haplotype harbours several activating KIR genes, whereas the only activating receptor, *KIR2DS4,* of the group A is in a significant proportion of Caucasians a non-functional truncated variant (Maxwell et al., 2002; Bontadini et al., 2006), rendering about 40% of group A homozygotes solely inhibitory. Approximately 55% of haplotypes are mixtures between group A and B (Middleton & Gonzelez, 2010), making the haplotype strucure highly variable in the population. Allelic diversity within KIRs is equally high with at least a few hundred known polymorphisms (https://www.ebi.ac.uk/ipd/kir/stats.html), which can affect class I ligand affinity (Carr, Pando & Parham, 2005; Frazier et al., 2013).

Discovery and interpretation of KIR gene and haplotype associations in large biobank genome data collections can be facilitated by imputation of KIR content from single-nucleotide polymorphisms (SNPs) genotyped by microarray. Furthermore, in organ or stem cell transplantation setting the KIR locus offers additional genetic information for donor selection and prediction of clinical outcome (Cooley et al., 2010; Impola et al., 2014; Littera et al., 2017), and for many of these clinical genome datasets SNP microarray provides the most cost-effective genotyping platform as well. To date, several KIR copy number or gene

content analysis methods have been implemented for sequencing data (Norman et al., 2016; Maniangou et al., 2017; Wagner et al., 2018; Chen et al., 2020; Roe & Kuang, 2020), but to our knowledge only one SNP-based approach exists so far (Vukcevic et al., 2015). These approaches reach a high accuracy which makes KIR inference reliable enough for research and even practical clinical applications. However, regarding biobank data, a stand-alone application that does not require submitting individual genotype data to external servers would be essential. To this end, we have implemented a random forest (RF) based KIR gene content prediction in the R environment exploiting SNP data. The reference data used for model fitting comprises KIR genotypes determined by targeted sequencing and 5774 genotyped SNPs in the KIR chromosomal region. Based on prediction of an independent subset of data, our results demonstrate a mean overall accuracy of 99.2% which is comparable to previously published methods.

Materials and Methods

Subjects

Blood donor genomic DNA and genotype data were obtained from the Blood Service Biobank, Helsinki, Finland. The samples were collected from Finnish blood donors who had given an informed consent for biobank research according to the Finnish Biobank Act (688/2012), supervised by the Finnish Medicines Agency (Fimea). The use of the biobank data in this study was approved by the Blood Service Biobank (002-2018).

Genotyping

Genotyping of samples was originally performed on a customized ThermoFisher Axiom array at the Thermo Fisher genotyping service facility (San Diego, USA) as a part of the FinnGen project. After the embargo period, the genotypes were returned to the Blood Service Biobank.

Genotype calling and quality control steps are described in finngen.gitbook.io/documentation/methods/genotype-imputation. The array markes files can be downloaded from www.finngen.fi/en/researchers/genotyping. The protocol for genotype data liftover to hg38/GRCh38 is described in detail in

www.protocols.io/view/genotyping-chip-data-lift-over-to-reference-genome-xbhfij6?version_warning=no, and genotype imputation protocol is described in www.protocols.io/view/genotype-imputation-workflow-v3-0-xbgfijw. KIR genotyping at absence-presence level of 818 samples was purchased from Histogenetics LLC (NY, USA).

Imputation models

The outline of the modelling set up is depicted in Figure 1. The random forest model for predicting KIR gene content was implemented with R v4.0.4 (R Core Team, 2021) using the library ranger v0.12.1 (Wright & Ziegler, 2017). The model error was estimated by dividing the data randomly into two equal subsets to one of which a RF model for each KIR gene was fitted while the other subset was used for prediction with the fitted model. SNP dosage values in the KIR region on chr19 were used as predictor variables, and the KIR gene content (1 for presence, 0 for absence) as determined by targeted sequencing served as the target phenotype variable. Feature selection within the model fitting was implemented using the permutated importance metric. Variants achieving an importance >1 x 10^{-5} were accepted into the model. The final model was fitted with the full dataset (Figure 1).

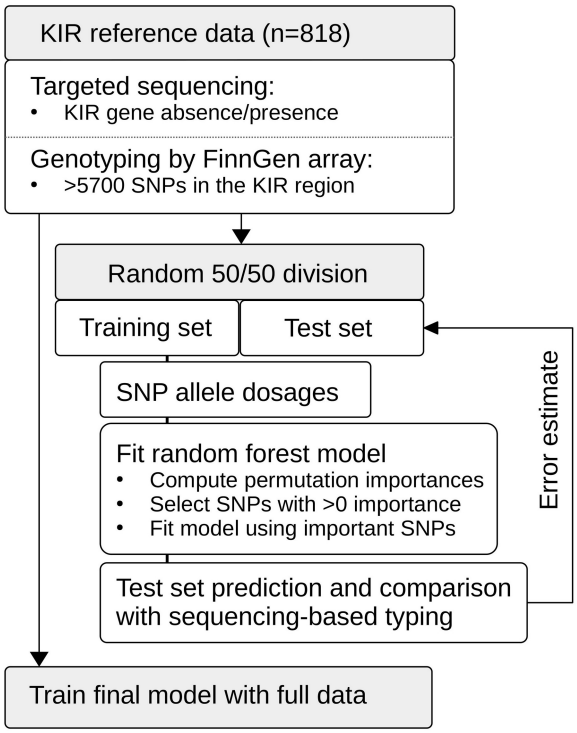

Figure 1. Schematic presentation of the study setup. The reference data set of 818 individuals was genotyped on the FinnGen SNP array, and KIR gene content was determined by targeted sequencing. We used random forests to fit models to the training data set comprising randomly selected 409 individuals. Feature selection was based on the importance metric computed through permutation. The model was re.fitted on SNPs achieving an importance of >0. Based on the test set comprising the other half of the samples, we calculated prediction error estimates for the modeling approach. Finally, we used the whole data set to train complete models.

Accuracy metrics were calculated using the R library caret v6.0-86 (Kuhn, Max, 2020). Positive predictive value (PPV) was defined as (sensitivity * prevalence) / ((sensitivity*prevalence) + ((1-specificity)*(1-prevalence))). Negative predictive value (NPV) was defined as (specificity * (1-prevalence)) / (((1-sensitivity)*prevalence) + ((specificity)*(1-prevalence))). Balanced accuray was calculated as (sensitivity+specificity)/2. Overall accuracy was calculated as the proportion of correct calls from all calls with 95% confidence intervals determined by binomial distribution. The data were managed with the tidyverse v1.3.0 (Wickham et al., 2019) package system.

To compare our method with KIR*IMP, KIR*IMP v1.2.0 (http://imp.science.unimelb.edu.au/kir/) (Vukcevic et al., 2015) was applied to the 818 samples constituting our reference panel. Prior to submitting the data to KIR*IMP, the genotypes were transferred to hg19 coordinates with UCSC LiftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) and phased with shapeit v2.r904 (O'Connell et al., 2014) with default parameters except for burn=10, prune=10, main=50 and window=0.5. The orientations of the SNPs in the phased dataset were harmonised according to the KIR*IMP SNP information file (http://imp.science.unimelb.edu.au/kir/static/kirimp.uk1.snp.info.csv) using a custom R script. In analysing the results, we used KIR2DS4TOTAL and KIR3DL1ex9 CNV imputation results of KIR*IMP to compare with our KIR2DS4 and KIR3DL1 absence-presence imputations, respectively.

Code and data availability

The analysis code and the models are available at https://github.com/FRCBS/KIR-imputation. SNP genotype data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001005457. Further information about EGA can be found on https://ega-archive.org "The European Genome-phenome Archive of human data consented for biomedical research" (http://www.nature.com/ng/journal/v47/n7/full/ng.3312.html).

Results

Accuracy estimates for each KIR gene for prediction of an independent test set are listed in Table 1. In summary, the mean overall accuracy of prediction was 0.992 (95% CI 0.997–0.978). The lowest accuracy of 0.968 (95% CI 0.983-0.945) was obtained for KIR2DS3 while KIR2DL3, KIR2DS2, KIR2DS4 and KIR3DL1 all achieved an overall accuracy of 1. Accuracy estimates for the test data and the RF out-of-bag (training set and full data) are plotted in Figure 2a. SNPs used by the models are listed in Supplementary Table 1.

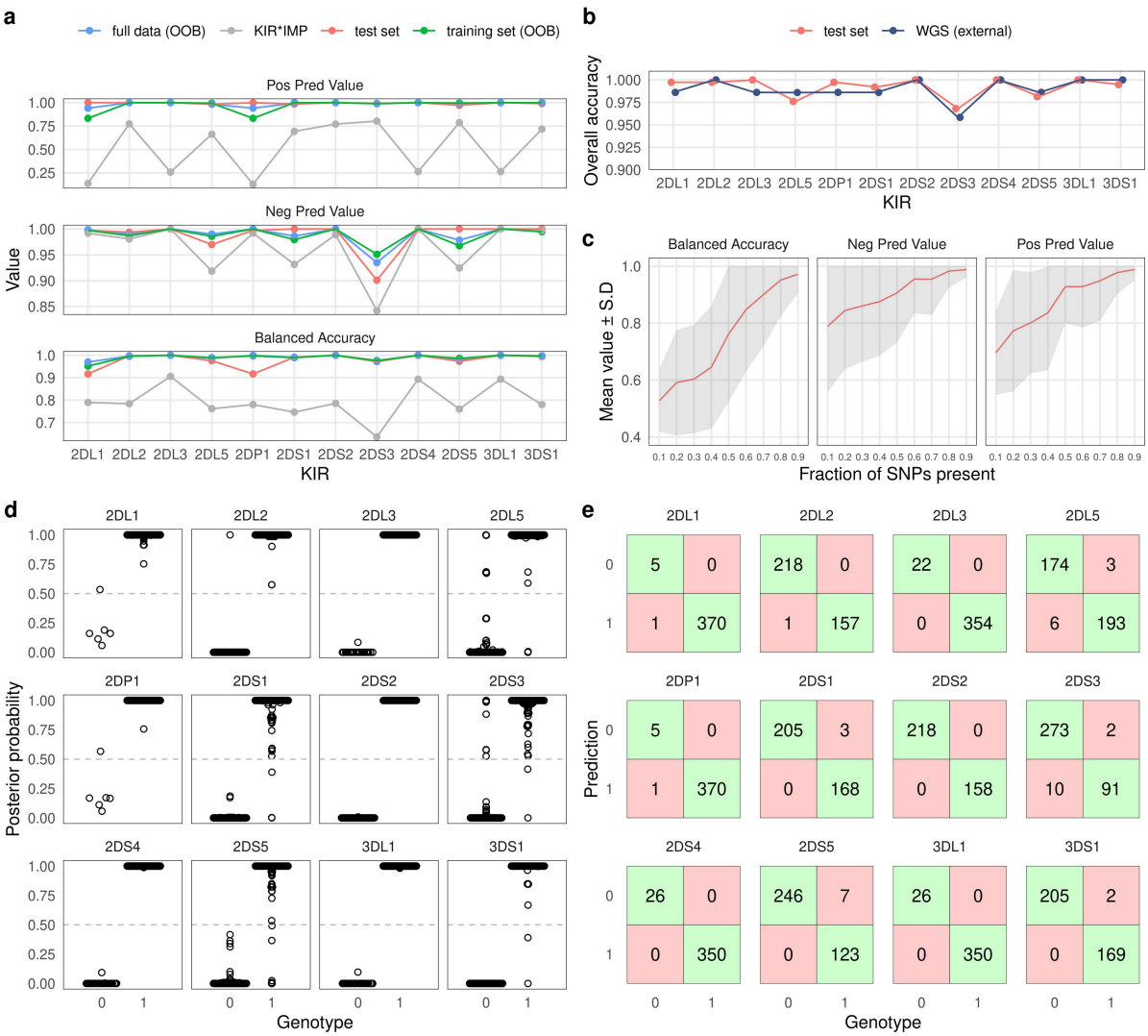

Figure 2. Overview of KIR prediction accuracy. **a)** Prediction performance metrics for the 12 imputed KIR genes. OOB: out-of-bag estimate from random forest models. Test set was predicted by models fitted on the training set. KIR*IMP was applied to the full

dataset. Note the varying scale of the y-axis. **b)** Comparison of overall accuracies between the test set and reported values for the WGS based method kpi extracted from the publication by Chen and co-workers. **c)** Impact of missing SNPs on prediciton performance in the test set. **d)** Posterior probability distributions for test set prediction. Genotypes 0 and 1 denote absence and presence of a KIR gene, respectively. **e)** Confusion tables for the test set predition. Posterior probabilities >0.5 were classified as 'present'.

To compare our approach with KIR*IMP, we converted our dataset of 818 samples to hg19 genome build and harmonised the SNP orientations. 126 SNPs out of 5774 could not be lifted over to hg19, and out of the 301 SNPs used by KIR*IMP 249 were found in our input data. SNP allele frequencies between the KIR*IMP reference panel and the input data had Pearson's correlation coefficient of 0.968 (Supplementary Figure 1a). Mean accuracy based on an estimate from the KIR*IMP reference subsetted for the input SNPs was 96.59% (Supplementary Figure 1b). Accuracy metrics for the imputation of our data by KIR*IMP are indicated by grey colour in Figure 2a. In summary, for all 12 included KIR genes we observed a distinctly lower imputation accuracy for KIR*IMP in comparison with our method.

Table 1. KIR imputation accuracy.

KIR_gene	Sensitivity	Specificity	Pos. Pred. Value	Neg. Pred. Value	Precision	Recall	Balanced accuracy	Overall accuracy
2DL1	0.833	1	1	0.997	1	0.833	0.917	0.997
2DL2	0.995	1	1	0.994	1	0.995	0.998	0.997
2DL3	1	1	1	1	1	1	1	1
2DL5	0.967	0.985	0.983	0.97	0.983	0.967	0.976	0.976
2DP1	0.833	1	1	0.997	1	0.833	0.917	0.997
2DS1	1	0.982	0.986	1	0.986	1	0.991	0.992
2DS2	1	1	1	1	1	1	1	1
2DS3	0.965	0.978	0.993	0.901	0.993	0.965	0.972	0.968
2DS4	1	1	1	1	1	1	1	1
2DS5	1	0.946	0.972	1	0.972	1	0.973	0.981
3DL1	1	1	1	1	1	1	1	1
3DS1	1	0.988	0.99	1	0.99	1	0.994	0.995

To compare the level of overall accuracy of our SNP-based method with an established sequencing-based approach, we extracted the results of the evaluation by Chen and coworkers (Chen et al., 2020) for the KIR imputation method kpi (Roe & Kuang, 2020). Figure 2b shows the accuracy of kpi compared

202 with our test set results. The observed values were highly similar with KIR2DS3
203 being the most diffcult gene to impute correctly.

204 Varying numbers of missing SNPs within the KIR region reduced the imputation
205 accuracy in accordance with the fraction of removed variants. At 80% of the
206 SNPs present the accuracy generally remained at a good level but started to
207 increasingly deteriorate after that (Figure 2c).

208 Posterior probability (PP) values of imputation are informative of imputation
209 uncertainty and can be incorporated into association analyses (Zhou et al.,
210 2020). Figure 2d shows the PP distributions for each imputed KIR gene.
211 Classification performances according to the PP threshold where PPs below 0.5
212 were classified as a missing KIR gene are shown by confusion tables in Figure 2e.
213 KIR genes with a higher error rate typically exhibited a PP distribution indicative
214 of higher uncertainty as more values were closer to 0.5 than 1 or 0.

215 Discussion

216 Genome data generated in a growing number of biobank projects is instrumental
217 to detailed immunogenetic analyses of several clinical phenotypes and diseases.
218 Within the current technological and economical constraints SNP microarrays
219 offer a practical way for genotyping hundreds of thousands of individuals. The
220 KIR gene content, despite being a relatively coarse-scale feature, has been
221 shown to influence many immune-mediated disorders (Bashirova et al., 2006;
222 Parham & Guethlein, 2018) and complications in pregnancy (Colucci, 2017).
223 Imputation of KIR gene content from SNPs in a scalable way is therefore essential
224 to analysing and interpreting large biomedical databases. To this end, in the
225 present study we have built a machine learning model for inferring KIR gene
226 content from SNP dosage data for stand-alone application in biobanks and other
227 clinical data collections. Exploitation of random forest for imputing KIRs from SNP
228 genotypes was first implemented in the KIR*IMP software (Vukcevic et al., 2015),
229 which runs on a remote server (http://imp.science.unimelb.edu.au/kir/). The main
230 difference of our method in comparison with KIR*IMP is that it does not require
231 phased data and the models can be downloaded and run locally. However,
232 KIR*IMP produces a more detailed output that includes A and B haplotypes,
233 framework genes KIR3DP1 and KIR2DL4, variants of KIR2DS4 and KIR3DL1 and
234 gene copy numbers. Otherwise, at the level of gene absence-presence, the

imputation accuracy of our method compares favourably not only with KIR*IMP but also to sequencing-based methods (Chen et al., 2020; Roe & Kuang, 2020).

In all imputation evaluations KIR2DS3 demonstrated the largest error in overall accuracy, followed by KIR2DL5 and KIR2DS5. A common feature shared by these three genes is that their location within the KIR chromosomal region is not fixed but can vary between centromeric and telomeric positions (Hsu et al., 2002; Pyo et al., 2010). Conceivably, this kind of positional variance may confound the identification of predictive SNPs resulting in greater imputation uncertainty. Other challenging genes were KIR2DL1 and KIR2DP1 which both harbour a relatively rare gene absence with a population frequency of about 1.6%, and therefore had few cases in the training data. In this regard, the out-of-bag estimate for the whole dataset might be the most reliable error estimate for these genes, suggesting a balanced accuracy and positive predictive value of about 0.95. Despite some challenges, KIR gene content imputation presents a valuable tool for initial screening and provides a rational basis for further analyses.

Allelic diversity, copy number variation and homologous gene sequences make KIR typing challenging by NGS or microarray probes. Nevertheless, the large number of variants within the genomic region allows extraction of information based on linkage with gene content, even if the causative variants cannot in all cases be directly measured. However, this is also a disadvantage because linkage patterns typically vary between populations and consequently models trained on one population may not be fully transferrable to another. While the informative SNPs used by our method are not specific to the Finnish population as such, but present a set of common genetic variants with relatively similar allele frequencies across European populations, it is not guaranteed that the prediction would achieve as good an accuracy in populations other than Finns. Our method is also limited by the requirement of the availability of informative SNPs in the dataset under analysis. These variants are not genotyped by all microarrays commonly used in genome analysis and therefore selection of a suitable platform is crucial. Another noteworthy limitation is that the method is not capable of identifying alleles. To date, only targeted sequencing based approaches can resolve KIR alleles (Norman et al., 2016; Maniangou et al., 2017; Wagner et al., 2018; Roe & Kuang, 2020). A possible future direction therefore is to extend KIR imputation from SNPs to cover allelic diversity.

Acknowledgements

The study was supported by the Academy of Finland, the Finnish Cancer Association, VTR funding from the Finnish Government, and Business Finland. The funders and biobanks had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank the Finnish Red Cross Blood Service Biobank and the blood donors for providing the samples. FinnGen is acknowledged for providing the SNP genotype data.

Competing interests

The authors declare no competing interests.

References

Bashirova AA, Martin MP, McVicar DW, Carrington M. 2006. The Killer Immunoglobulin-Like Receptor Gene Cluster: Tuning the Genome for Defense. *Annual Review of Genomics and Human Genetics* 7:277–300. DOI: 10.1146/annurev.genom.7.080505.115726.

Bontadini A, Testi M, Cuccia M, Martinetti M, Carcassi C, Chiesa A, Cosentini E, Dametto E, Frison S, Iannone A, Lombardo C, Malagoli A, Mariani M, Mariotti L, Mascaretti L, Mele L, Miotti V, Nesci S, Ozzella G, Piancatelli D, Romeo G, Tagliaferri C, Vatta S, Andreani M, Conte R. 2006. Distribution of killer cell immunoglobulin-like receptors genes in the Italian Caucasian population. *Journal of Translational Medicine* 4:44. DOI: 10.1186/1479-5876-4-44.

Carr WH, Pando MJ, Parham P. 2005. KIR3DL1 Polymorphisms That Affect NK Cell Inhibition by HLA-Bw4 Ligand. *The Journal of Immunology* 175:5222–5229. DOI: 10.4049/jimmunol.175.8.5222.

Chen J, Madireddi S, Nagarkar D, Migdal M, Vander Heiden J, Chang D, Mukhyala K, Selvaraj S, Kadel EE, Brauer MJ, Mariathasan S, Hunkapiller J, Jhunjhunwala S, Albert ML, Hammer C. 2020. In silico tools for accurate HLA and KIR inference from clinical sequencing data empower immunogenetics on individual-patient and population scales. *Briefings in Bioinformatics*:bbaa223. DOI: 10.1093/bib/bbaa223.

Colucci F. 2017. The role of KIR and HLA interactions in pregnancy complications. *Immunogenetics* 69:557–565. DOI: 10.1007/s00251-017-1003-9.

Cooley S, Weisdorf DJ, Guethlein LA, Klein JP, Wang T, Le CT, Marsh SGE, Geraghty D, Spellman S, Haagenson MD, Ladner M, Trachtenberg E, Parham P, Miller JS. 2010. Donor selection for natural killer cell receptor genes leads to superior survival after unrelated transplantation for acute myelogenous leukemia. *Blood* 116:2411–2419. DOI: 10.1182/blood-2010-05-283051.

Frazier WR, Steiner N, Hou L, Dakshanamurthy S, Hurley CK. 2013. Allelic Variation in KIR2DL3 Generates a KIR2DL2-like Receptor with Increased

Binding to its HLA-C Ligand. *The Journal of Immunology* 190:6198–6208. DOI: 10.4049/jimmunol.1300464.
Garcia-Beltran WF, Hölzemer A, Martrus G, Chung AW, Pacheco Y, Simoneau CR, Rucevic M, Lamothe-Molina PA, Pertel T, Kim T-E, Dugan H, Alter G, Dechanet-Merville J, Jost S, Carrington M, Altfeld M. 2016. Open conformers of HLA-F are high-affinity ligands of the activating NK-cell receptor KIR3DS1. *Nature Immunology* 17:1067–1074. DOI: 10.1038/ni.3513.
Hsu KC, Chida S, Geraghty DE, Dupont B. 2002. The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism: Hsu et al · KIR genomic region and KIR haplotypes. *Immunological Reviews* 190:40–52. DOI: 10.1034/j.1600-065X.2002.19004.x.
Impola U, Turpeinen H, Alakulppi N, Linjama T, Volin L, Niittyvuopio R, Partanen J, Koskela S. 2014. Donor Haplotype B of NK KIR Receptor Reduces the Relapse Risk in HLA-Identical Sibling Hematopoietic Stem Cell Transplantation of AML Patients. *Frontiers in Immunology* 5. DOI: 10.3389/fimmu.2014.00405.
Ivarsson MA, Michaëlsson J, Fauriat C. 2014. Activating Killer Cell Ig-Like Receptors in Health and Disease. *Frontiers in Immunology* 5. DOI: 10.3389/fimmu.2014.00184.
Kuhn, Max. 2020. *caret: Classification and Regression Training*.
Lanier LL. 2008. Up on the tightrope: natural killer cell activation and inhibition. *Nature Immunology* 9:495–502. DOI: 10.1038/ni1581.
Littera R, Piredda G, Argiolas D, Lai S, Congeddu E, Ragatzu P, Melis M, Carta E, Michittu MB, Valentini D, Cappai L, Porcella R, Alba F, Serra M, Loi V, Maddi R, Orrù S, La Nasa G, Caocci G, Cusano R, Arras M, Frongia M, Pani A, Carcassi C. 2017. KIR and their HLA Class I ligands: Two more pieces towards completing the puzzle of chronic rejection and graft loss in kidney transplantation. *PLOS ONE* 12:e0180831. DOI: 10.1371/journal.pone.0180831.
Maniangou B, Legrand N, Alizadeh M, Guyet U, Willem C, David G, Charpentier E, Walencik A, Retière C, Gagne K. 2017. Killer Immunoglobulin-Like Receptor Allele Determination Using Next-Generation Sequencing Technology. *Frontiers in Immunology* 8:547. DOI: 10.3389/fimmu.2017.00547.
Maxwell LD, Wallace A, Middleton D, Curran MD. 2002. A common KIR2DS4 deletion variant in the human that predicts a soluble KIR molecule analogous to the KIR1D molecule observed in the rhesus monkey: Maxwell et al : A predicted soluble KIR2DS4 deletion variant. *Tissue Antigens* 60:254–258. DOI: 10.1034/j.1399-0039.2002.600307.x.
Middleton D, Gonzelez F. 2010. The extensive polymorphism of KIR genes. *Immunology* 129:8–19. DOI: 10.1111/j.1365-2567.2009.03208.x.
Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, Jayaraman J, Wroblewski EE, Trowsdale J, Rajalingam R, Oksenberg JR, Chiaroni J, Guethlein LA, Traherne JA, Ronaghi M, Parham P. 2016. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. *The American Journal of Human Genetics* 99:375–391. DOI: 10.1016/j.ajhg.2016.06.023.
O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I, McQuillan R, Fraser RM, Campbell H, Polasek O, Asiki G, Ekoru K, Hayward C, Wright AF, Vitart V, Navarro P, Zagury J-F, Wilson JF, Toniolo D, Gasparini P, Soranzo N, Sandhu MS, Marchini J. 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics* 10:e1004234. DOI: 10.1371/journal.pgen.1004234.

Parham P, Guethlein LA. 2018. Genetics of Natural Killer Cells in Human Health, Disease, and Survival. *Annual Review of Immunology* 36:519–548. DOI: 10.1146/annurev-immunol-042617-053149.

Pyo C-W, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, Marsh SGE, Miller JS, Parham P, Geraghty DE. 2010. Different Patterns of Evolution in the Centromeric and Telomeric Regions of Group A and B Haplotypes of the Human Killer Cell Ig-Like Receptor Locus. *PLoS ONE* 5:e15115. DOI: 10.1371/journal.pone.0015115.

R Core Team. 2021. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Roe D, Kuang R. 2020. Accurate and Efficient KIR Gene and Haplotype Inference From Genome Sequencing Reads With Novel K-mer Signatures. *Frontiers in Immunology* 11:583013. DOI: 10.3389/fimmu.2020.583013.

Stewart CA, Laugier-Anfossi F, Vely F, Saulquin X, Riedmuller J, Tisserant A, Gauthier L, Romagne F, Ferracci G, Arosa FA, Moretta A, Sun PD, Ugolini S, Vivier E. 2005. Recognition of peptide-MHC class I complexes by activating killer immunoglobulin-like receptors. *Proceedings of the National Academy of Sciences* 102:13224–13229. DOI: 10.1073/pnas.0503594102.

Vukcevic D, Traherne JA, Næss S, Ellinghaus E, Kamatani Y, Dilthey A, Lathrop M, Karlsen TH, Franke A, Moffatt M, Cookson W, Trowsdale J, McVean G, Sawcer S, Leslie S. 2015. Imputation of KIR Types from SNP Variation Data. *The American Journal of Human Genetics* 97:593–607. DOI: 10.1016/j.ajhg.2015.09.005.

Wagner I, Schefzyk D, Pruschke J, Schöfl G, Schöne B, Gruber N, Lang K, Hofmann J, Gnahm C, Heyn B, Marin WM, Dandekar R, Hollenbach JA, Schetelig J, Pingel J, Norman PJ, Sauter J, Schmidt AH, Lange V. 2018. Allele-Level KIR Genotyping of More Than a Million Samples: Workflow, Algorithm, and Observations. *Frontiers in Immunology* 9:2843. DOI: 10.3389/fimmu.2018.02843.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* 4:1686. DOI: 10.21105/joss.01686.

Wright MN, Ziegler A. 2017. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77. DOI: 10.18637/jss.v077.i01.

Wroblewski EE, Parham P, Guethlein LA. 2019. Two to Tango: Co-evolution of Hominid Natural Killer Cell Receptors and MHC. *Frontiers in Immunology* 10:177. DOI: 10.3389/fimmu.2019.00177.

Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Gagliano Taliun SA, Bi W, Gabrielsen ME, Daly MJ, Neale BM, Hveem K, Abecasis GR, Willer CJ, Lee S. 2020. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nature Genetics* 52:634–639. DOI: 10.1038/s41588-020-0621-6.