MiDAS - Meaningful Immunogenetic Data at Scale

Maciej Migdal[1], Dan Fu Ruan[2], William F. Forrest[3], Amir Horowitz[2], Christian Hammer[4,5]

Human immunogenetic variation in the form of HLA and KIR types has been shown to be strongly associated with a multitude of immune-related phenotypes. We present MiDAS, an R package enabling statistical association analysis and using immunogenetic data transformation functions for HLA amino acid fine mapping, analysis of HLA evolutionary divergence as well as HLA-KIR interactions. MiDAS closes the gap between inference of immunogenetic variation and its efficient utilization to make meaningful discoveries.

The major histocompatibility complex (MHC) is the region in the genome with the highest density of statistical associations with disease phenotypes. The majority of these associations are related to the central role of classical Human Leukocyte Antigen (HLA) proteins in immune responses in the context of autoimmunity, infectious disease, and also cancer.[1] Another complex genomic locus relevant for immune responses is the leukocyte receptor complex (LRC) on chromosome 19, which, among other genes, harbors the killer cell immunoglobulin like receptors (KIR). KIR predominantly mediate function and education of Natural Killer (NK) cells, but can also be found on subsets of T cells.[2] They display a high degree of copy number as well as allelic variation. Many KIR are receptors for HLA class I ligands, but these interactions are highly specific and depend on individuals' HLA and KIR genotypes, segregating on different chromosomes.[2]

The extreme amount of genetic variation has made it challenging to accurately characterize individuals' HLA and KIR genotypes, but besides dedicated typing methods, there are now multiple tools available for inference from next generation sequencing or single nucleotide polymorphism (SNP) array genotyping data at scale.[3–5] However, the availability of immunogenetic variation data is only the first necessary step in uncovering and understanding its role in immune-related traits, and statistical considerations are more complex when compared to the millions of common single nucleotide polymorphisms (SNPs) or copy number variants (CNVs) in our genomes that predominantly have two allelic states.

Statistical association analyses of immunogenetic variants often focus on the presence vs. absence of single HLA alleles. They are most often analyzed on 2-field level (formerly '4-digit', e.g. *HLA-DRB1*15:01*), which defines the protein structure of the HLA protein, as well as the composition of its peptide binding groove and thus the repertoire of antigens it can present. HLA alleles can also be grouped on 1-field level, which often corresponds to the serological antigen carried by an allotype,[6] or on the level of supertypes, which present overlapping peptide repertoires based on their main anchor specificities.[7] In addition, typing data and resulting association statistics can be available on the level of G groups, which contain alleles that have identical nucleotide sequences across the exons encoding the peptide binding domains (exons 2 and 3 for HLA class I and exon 2 for HLA class II alleles).[8]

MiDAS accepts HLA genetic data in up to 4-field (8-digit) resolution, checks it for consistency with official HLA nomenclature,[6] and can reduce its resolution or transform it into supertypes or G groups, to allow consistent results reporting and cross-study comparability (**Table 1**). MiDAS includes a function to test for deviations from Hardy-Weinberg equilibrium (HWE) and provides the option list HWE P values or directly filter out significant alleles, and it is also possible to quickly compare allele frequencies in input data sets with published frequencies across different populations based on data from a comprehensive online database.[9]

In spite of the vast number of statistical associations in the MHC locus, the complex linkage disequilibrium in the region combined with the proximity of genes with different immune-related or non-immune functions can make it difficult to pinpoint causal genes and variants.[10] However, due to the availability of protein sequences for most known HLA alleles, it is possible to use HLA allele data to generate new variables for each amino acid position in a protein that differs across individuals. Such an approach was used to demonstrate that five variable amino acids across three HLA proteins explained most of the MHC association with seropositive rheumatoid arthritis.[11]

MiDAS facilitates this process by inferring variable amino acid residues for all imported individuals with HLA allele data (**Figure 1**), based on

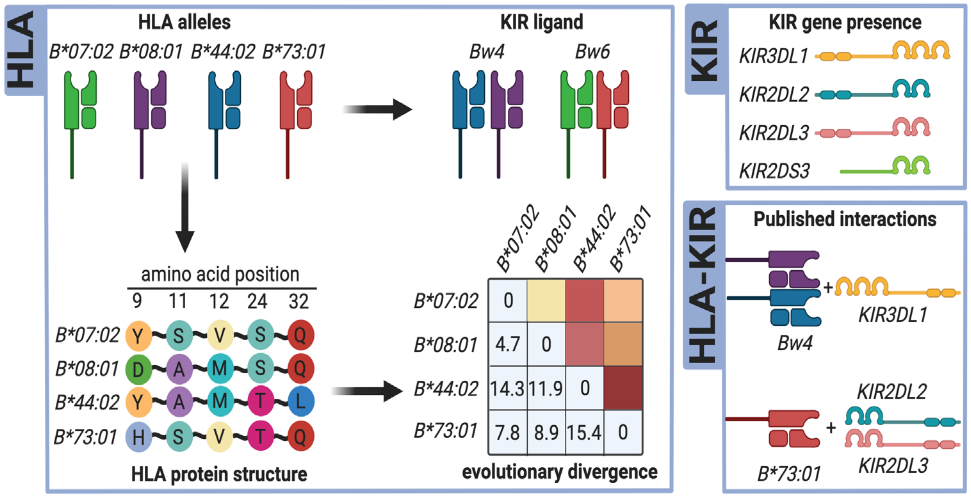


Figure 1. MiDAS data transformation functions.
MiDAS can transform HLA and KIR input data to test association hypotheses beyond single allele or KIR gene approaches. HLA alleles can be grouped according to their interactions with KIR, and sequence information is used to infer variable amino acid positions for statistical fine-mapping. Amino acid level information is also used to calculate evolutionary divergence of HLA allele pairs for a given gene. If both HLA and KIR data are available, biologically validated receptor-ligand interactions can be coded according to the definitions summarized by Pende et al.[2]
Figure created with BioRender.com

[1]Roche Global IT Solution Centre (RGITSC), Warsaw, Poland. [2]Department of Oncological Sciences, Precision Immunology Institute, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Department of OMNI Bioinformatics, Genentech, South San Francisco, CA 94080, USA. [4]Department of Cancer Immunology, Genentech, South San Francisco, CA 94080, USA. [5]Department of Human Genetics, Genentech, South San Francisco, CA 94080, USA.
Correspondence: hammer.christian@gene.com

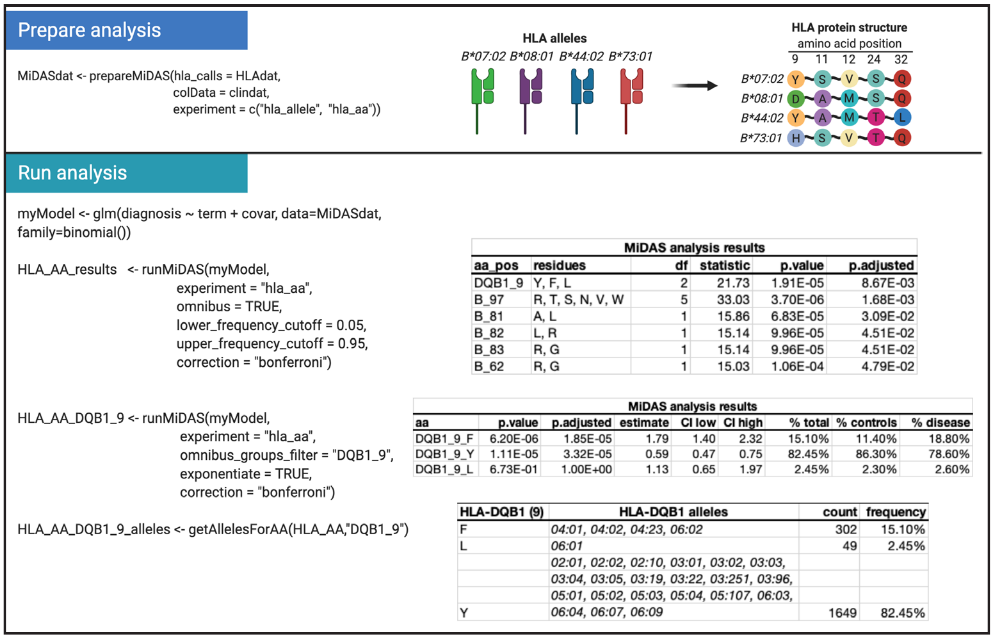
Variable type	MiDAS experiment name	Definition	Reference	Example use case
HLA alleles	hla_alleles	HLA allele status at 1- to 4-field resolution	12	1
	hla_supertypes	HLA class I alleles grouped into supertypes	7	28
	hla_g_groups	HLA alleles grouped according to identical nucleotide sequence in peptide binding domains	8	29
HLA amino acids	hla_aa	Variable amino acid positions and residues based on HLA allele sequence alignments	12	11
HLA intra-individual diversity	hla_het	Heterozygosity vs. homozygosity of each classical HLA gene		13
	hla_divergence	HLA class I evolutionary divergence as measured by Grantham's distance	30,31	15
HLA NK ligand status	hla_NK_ligands	Bw4 / Bw6, C1 / C2 allele group inference based on HLA allele matching table	2,32	17,18
KIR gene presence	kir_genes	Presence or absence of specific KIR genes (binary variable)	33	34
HLA-KIR interactions	hla_kir_interactions	Experimentally verified ligand-receptor interactions between HLA class I and KIR	2	19-21
Custom	hla_custom, kir_custom	User-provided dictionaries for custom analyses		

sequence alignments from the IPD-IMGT/HLA database.[12] It is then possible to perform a likelihood-ratio ('omnibus') test for each variable amino acid position in HLA proteins, determine the effect estimates for all residues at associated positions, and also to map the spectrum of HLA alleles that contain each respective residue (**Table 1**, **Figure 2**).

Intra-individual diversity of HLA alleles, assessed in terms of heterozygosity versus homozygosity or evolutionary divergence, is considered a useful proxy for the diversity of antigens that can be presented by an individual's HLA proteins. For example, HIV-positive patients with full heterozygosity for *HLA-A*, *-B* and *-C* were shown to progress more slowly to AIDS,[13] which is likely at least in part due to an increased diversity of presented peptides.[14] Further, cancer patients treated with immune-checkpoint inhibitors responded better to the therapy if they had an increased evolutionary sequence divergence in their HLA class I proteins.[15] MiDAS can recode HLA alleles into new variables indicating heterozygosity at each locus, as well as Grantham's distance for HLA class I genes (**Table 1**, **Figure 1**). Grantham's distance can be calculated for amino acids in the whole peptide binding region of HLA class I molecules, or restricted to the B- or F- binding pockets individually.[16]

Beyond their central role in antigen presentation, HLA class I molecules also function as ligands for KIR, and thus are able to impact NK cell education and function. Beyond interactions between specific HLA alleles and KIR, HLA alleles can also be grouped by MiDAS according to common epitopes into HLA-Bw4, -Bw6, -C1, and -C2 alleles (**Figure 1**).[2] HLA-Bw4 alleles show experimentally verified interactions with KIR3DL1, whereas HLA-Bw6 alleles have no known interaction with inhibitory KIR. HLA-C1 alleles show strong affinity for KIR2DL3, whereas HLA-C2 alleles show only weak affinity for KIR2DL3, but strong affinity for KIR2DL1. In terms of examples for disease relevance, HLA-Bw4 is a risk factor for ulcerative colitis in Japanese, and homozygosity for HLA-C1 was shown to be associated with reduced risk of relapse in patients with myeloid leukemia after transplantation.[17,18]

Hypotheses including a potential NK cell involvement benefit from the availability of both HLA and KIR typing data. MiDAS can load KIR data indicating the presence or absence of individual KIR genes, and perform association analysis on the level of these genes. But more importantly, if both HLA alleles and KIR data are available, it generates new variables indicating the presence of all experimentally validated interactions as summarized by Pende et al.[2] Investigating the role of such HLA-KIR interactions has previously helped to better understand differential risk for pregnancy complications,[19] pathogen immunity,[20] or NK cell activity in recipients of hematopoietic cell transplants.[21]



```
MiDASdat <- prepareMiDAS(hla_calls = HLAdat,
                         colData = clindat,
                         experiment = c("hla_allele",  "hla_aa"))
```

```
myModel <- glm(diagnosis ~ term + covar, data=MiDASdat,
family=binomial())

HLA_AA_results   <- runMiDAS(myModel,
                             experiment = "hla_aa",
                             omnibus = TRUE,
                             lower_frequency_cutoff = 0.05,
                             upper_frequency_cutoff = 0.95,
                             correction = "bonferroni")
```

MiDAS analysis results

aa_pos	residues	df	statistic	p.value	p.adjusted
DQB1_9	Y, F, L	2	21.73	1.91E-05	8.67E-03
B_97	R, T, S, N, V, W	5	33.03	3.70E-06	1.68E-03
B_81	A, L	1	15.86	6.83E-05	3.09E-02
B_82	L, R	1	15.14	9.96E-05	4.51E-02
B_83	R, G	1	15.14	9.96E-05	4.51E-02
B_62	R, G	1	15.03	1.06E-04	4.79E-02

```
HLA_AA_DQB1_9 <- runMiDAS(myModel,
                          experiment = "hla_aa",
                          omnibus_groups_filter = "DQB1_9",
                          exponentiate = TRUE,
                          correction = "bonferroni")
```

MiDAS analysis results

aa	p.value	p.adjusted	estimate	CI low	CI high	% total	% controls	% disease
DQB1_9_F	6.20E-06	1.85E-05	1.79	1.40	2.32	15.10%	11.40%	18.80%
DQB1_9_Y	1.11E-05	3.32E-05	0.59	0.47	0.75	82.45%	86.30%	78.60%
DQB1_9_L	6.73E-01	1.00E+00	1.13	0.65	1.97	2.45%	2.30%	2.60%

```
HLA_AA_DQB1_9_alleles <- getAllelesForAA(HLA_AA,"DQB1_9")
```

HLA-DQB1 (9)	HLA-DQB1 alleles	count	frequency
F	*04:01, 04:02, 04:23, 06:02*	302	15.10%
L	*06:01*	49	2.45%
	02:01, 02:02, 02:10, 03:01, 03:02, 03:03,		
	03:04, 03:05, 03:19, 03:22, 03:251, 03:96,		
	05:01, 05:02, 05:03, 05:04, 05:107, 06:03,		
Y	*06:04, 06:07, 06:09*	1649	82.45%

Figure 2. Example of amino acid fine-mapping analysis.
Example analysis flow for HLA amino acid analysis. In the first step, HLA and clinical data were combined in a MiDAS object using the 'prepareMiDAS' function, which also performed HLA data transformation to amino acid level (specified as 'experiment'). Before the association analysis, a statistical model was defined. 'term' is a placeholder that is replaced by each tested amino acid, covariates ('covar') can be categorical or numeric. It is also possible to define interaction terms (e.g. 'term:covar', not shown). 'runMiDAS' was then run twice, first to perform an omnibus test on all variable amino acid positions, and then to calculate effect estimates for all residues (F,Y,L) at the top-associated position (DQB1_9). 'getAllelesforAA' was then used to map all *HLA-DQB1* alleles in the dataset to the three DQB1_9 residues.
Figure created with BioRender.com

Of note, MiDAS also facilitates the testing of specific, more refined hypotheses. For example, amino acid position 80 modulates the interaction of HLA-Bw4 alleles with KIR,[22] which can be modeled by subsetting HLA-Bw4 further according to amino acid level information. Data transformation can also be customized using user-supplied additional data dictionaries. A current shortcoming of MiDAS is that allelic variation of KIR, on top of individual gene presence, is not considered, although it is of demonstrated relevance in modulating interactions between KIR and their respective HLA ligands.[23] Another use case for custom analyses is the transformation of HLA allele data into quantitative variables such as allele-specific expression levels.[24]

MiDAS allows flexible statistical analyses of immunogenetic data with phenotypes on a diverse range of measurement scales, including regression models or time-to-event data. Results are stored as data tables that display nominal and corrected P values, effect estimates, confidence intervals and variant frequencies. It is possible to execute likelihood-ratio ('omnibus') tests, for example to summarize amino acid residues at each position in the protein and identify the most relevant positions as basis for statistical fine-mapping. MiDAS can also perform stepwise conditional analyses to identify multiple statistically independent association signals within and across HLA genes, which is commonly observed.[11] A range of genetic inheritance models can be selected (where applicable: dominant, recessive, additive, overdominant), as well as the preferred method for multiple testing correction and frequency cutoffs for variable inclusion, taking statistical power considerations into account (**Figure 2**).

MiDAS is freely available as an R package (MIT license), facilitating both hypothesis-driven and exploratory analyses of immunogenetics data (https://github.com/Genentech/MiDAS). Of note, MiDAS is not the first published software for immunogenetics association analysis. However, other tools are significantly more limited in terms of data transformation, analysis functions and statistical model selection.[25–27] A tutorial with example data and analyses is available under https://genentech.github.io/MiDAS/articles/MiDAS_tutorial.html.

Methods

MiDAS data structure

MiDAS accepts HLA types in a format that complies with official HLA nomenclature in up to 4-field resolution (e.g. '*A*02:01:01:01*'),[6] one row per individual, and one column for each allele of each gene (e.g. 'A_1', 'A_2'). KIR data is accepted in a tabular format that indicates presence ('1' or 'Y') or absence ('0' or 'N') of each KIR gene. Example input data tables are provided with the package to help putting users' own data in the right format.

The 'prepareMiDAS' function combines HLA, KIR, and phenotypic data into an object that is a subclass of a MultiAssayExperiment, which we termed 'MiDAS'. HLA input data is transformed into counts tables encoding the copy number of specific alleles, as a basis for statistical analysis. This function also offers the described data transformation options (e.g. NK cell ligands, HLA-KIR interactions).

Compared to the MultiAssayExperiment, 'MiDAS' class makes several assumptions that allow us to use data directly as an input to statistical model functions. The most important assumptions are: the variable names are unique across MiDAS, each experiment has only one assay defined. Further, experiments are defined as matrices or SummarizedExperiment objects. The latter is used in cases where experiment specific metadata are needed for the analysis, including variable groupings used for omnibus tests, or information on the applicability of inheritance models.

Statistical framework

The data analysis framework offered by MiDAS is flexible in terms of choice of the statistical model, often used examples including logistic or linear regression, or cox proportional hazard models for time-to-event analyses. This flexibility is possible due to using 'tidyers' as introduced in the 'broom' package (https://broom.tidymodels.org).

The MiDAS object is passed as a data argument to the function, and the chosen genetic variables are provided using a placeholder variable ('term'). The defined model is passed to the 'runMiDAS' function, where the actual statistical analysis is performed. Here, the placeholder variable is substituted with the actual genetic variables in an iterative manner, allowing to test individual variables for association with the response variable. The use of a placeholder allows the use of more complex statistical models, e.g. gene-environment interactions (e.g. "lm(diagnosis ~ 'term' + sex + 'term':sex)".

'runMiDAS' offers different modes of analysis. By default, the statistical model is iteratively evaluated with each individual genetic variable substituted for the placeholder. Then, test statistics from individual tests are gathered and corrected for multiple testing using a method of choice as implemented in the 'stats' R package. Moreover, 'runMiDAS' offers a conditional mode to test for statistically independent associations of multiple genetic variables, which implements a simple stepwise forward selection method. Here, the iterative comparisons are made in rounds, and for each round the algorithm selects the top associated variable and adds it to the model as a covariate, until no more variables meet the selection criteria. Further, 'runMiDAS' includes an 'omnibus' mode that allows to test the role of multiple grouped variables, using a likelihood ratio test. In particular, this is useful to score amino acid positions according to their omnibus P value, as compared to their individual residues.

References

1. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. Nat Rev Immunol 18, 325–339 (2018).
2. Pende, D. et al. Killer Ig-Like Receptors (KIRs): Their Role in NK Cell Modulation and Developments Leading to Their Clinical Exploitation. Front Immunol 10, 1179 (2019).
3. Chen, J. et al. In silico tools for accurate HLA and KIR inference from clinical sequencing data empower immunogenetics on individual-patient and population scales. Biorxiv 2020.04.24.060459 (2020) doi:10.1101/2020.04.24.060459.
4. Vukcevic, D. et al. Imputation of KIR Types from SNP Variation Data. Am J Hum Genetics 97, 593–607 (2015).
5. Karnes, J. H. et al. Comparison of HLA allelic imputation programs. Plos One 12, e0172444 (2017).
6. Marsh, S. G. E. et al. Nomenclature for factors of the HLA system, 2010. Tissue Antigens 75, 291–455 (2010).
7. Sidney, J., Peters, B., Frahm, N., Brander, C. & Sette, A. HLA class I supertypes: a revised and updated classification. Bmc Immunol 9, 1 (2008).
8. HLA G group definitions. http://hla.alleles.org/alleles/g_groups.html.
9. Gonzalez-Galarza, F. F. et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res 48, D783–D788 (2019).
10. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 54, 15–39 (2009).
11. Raychaudhuri, S. et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet 44, 291–296 (2012).
12. Robinson, J. et al. IPD-IMGT/HLA Database. Nucleic Acids Res 48, D948–D955 (2019).
13. Carrington, M. et al. HLA and HIV-1: Heterozygote Advantage and B*35-Cw*04 Disadvantage. Science 283, 1748–1752 (1999).
14. Arora, J. et al. HLA Heterozygote Advantage against HIV-1 Is Driven by Quantitative and Qualitative Differences in HLA Allele-Specific Peptide Presentation. Mol Biol Evol 37, 639–650 (2019).
15. Chowell, D. et al. Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. Nat Med 25, 1715–1720 (2019).
16. Deutekom, H. W. M. van & Keşmir, C. Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most? Immunogenetics 67, 425–436 (2015).
17. Saito, H. et al. Association between KIR-HLA combination and ulcerative colitis and Crohn's disease in a Japanese population. Plos One 13, e0195778 (2018).
18. Arima, N. et al. Homozygous HLA-C1 is Associated with Reduced Risk of Relapse after HLA-Matched Transplantation in Patients with Myeloid Leukemia. Biol Blood Marrow Tr 24, 717–725 (2018).
19. Colucci, F. The role of KIR and HLA interactions in pregnancy complications. Immunogenetics 69, 557–565 (2017).
20. Jamil, K. M. & Khakoo, S. I. KIR/HLA Interactions and Pathogen Immunity. J Biomed Biotechnol 2011, 298348 (2011).
21. Nowak, J. et al. Role of Donor Activating KIR–HLA Ligand–Mediated NK Cell Education Status in Control of Malignancy in Hematopoietic Cell Transplant Recipients. Biol Blood Marrow Tr 21, 829–839 (2015).
22. Thananchai, H. et al. Cutting Edge: Allele-Specific and Peptide-Dependent Interactions between KIR3DL1 and HLA-A and HLA-B. J Immunol 178, 33–37 (2007).
23. Frazier, W. R., Steiner, N., Hou, L., Dakshanamurthy, S. & Hurley, C. K. Allelic Variation in KIR2DL3 Generates a KIR2DL2-like Receptor with Increased Binding to its HLA-C Ligand. J Immunol 190, 6198–6208 (2013).
24. Ramsuran, V. et al. Epigenetic regulation of differential HLA-A allelic expression levels. Hum Mol Genet 24, 4268–4275 (2015).
25. Pappas, D. J., Marin, W., Hollenbach, J. A. & Mack, S. J. Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline. Hum Immunol 77, 283–287 (2016).

26. Kanterakis, S. et al. SKDM human leukocyte antigen (HLA) tool: A comprehensive HLA and disease associations analysis software. Hum Immunol 69, 522–5 (2008).

27. Choi, W., Luo, Y., Raychaudhuri, S. & Han, B. HATK: HLA Analysis Toolkit. Bioinformatics btaa684- (2020) doi:10.1093/bioinformatics/btaa684.

28. Vejbaesya, S. et al. HLA Class I Supertype Associations With Clinical Outcome of Secondary Dengue Virus Infections in Ethnic Thais. J Infect Dis 212, 939–947 (2015).

29. Rao, P. et al. P135 New frequent HLA-DPB1/DPA1 haplotypes in low resolution typing. Hum Immunol 78, 153 (2017).

30. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. Science 185, 862–864 (1974).

31. Pierini, F. & Lenz, T. L. Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. Mol Biol Evol 35, 2145–2158 (2018).

32. Parham, P. & Guethlein, L. A. Genetics of Natural Killer Cells in Human Health, Disease, and Survival. Annu Rev Immunol 36, 1–30 (2018).

33. Robinson, J., Halliwell, J. A., McWilliam, H., Lopez, R. & Marsh, S. G. E. IPD—the Immuno Polymorphism Database. Nucleic Acids Res 41, D1234–D1240 (2013).

34. Traherne, J. A. et al. KIR haplotypes are associated with late-onset type 1 diabetes in European–American families. Genes Immun 17, 8–12 (2016).