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. 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. 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.

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. 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, or on the level of supertypes, which present overlapping peptide repertoires based on their main anchor specificities.

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). MiDAS accepts HLA genetic data in up to 4-field (8-digit) resolution, checks it for consistency with official HLA nomenclature, 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.

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. 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.

MiDAS facilitates this process by inferring variable amino acid residues for all imported individuals with HLA allele data (Figure 1), based on
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, which is likely at least in part due to an increased diversity of presented peptides. Further, cancer patients treated with immune-checkpoint inhibitors responded better to the therapy if they had an increased evolutionary divergence in their HLA class I proteins. MiDAS can recode HLA alleles into new variables indicating the presence of all experimentally validated 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. Investigating the role of such HLA-KIR interactions has previously helped to better understand differential risk for pregnancy complications, pathogen immunity, or NK cell activity in recipients of hematopoietic cell transplants.

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). 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. 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. Investigating the role of such HLA-KIR interactions has previously helped to better understand differential risk for pregnancy complications, pathogen immunity, or NK cell activity in recipients of hematopoietic cell transplants.

### Variable type | MiDAS experiment name | Definition | Reference | Example use case
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA alleles</td>
<td>hla_alleles</td>
<td>HLA allele status at 1- to 4-field resolution</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>hla_supertypes</td>
<td>HLA class I alleles grouped into superfamilies</td>
<td>17,18</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>hla_g_groups</td>
<td>HLA alleles grouped according to identical nucleotide sequence in peptide binding domains</td>
<td>17,18</td>
<td>8</td>
</tr>
<tr>
<td>HLA amino acids</td>
<td>hla_aa</td>
<td>Variable amino acid positions and residues based on HLA allele sequence alignments</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>HLA intra-individual diversity</td>
<td>hla_het</td>
<td>Heterozygosity vs. homozygosity of each classical HLA gene</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hla_divergence</td>
<td>HLA class I evolutionary divergence as measured by Grantham’s distance</td>
<td>30,31</td>
<td>15</td>
</tr>
<tr>
<td>HLA NK ligand status</td>
<td>hla_NK_ligands</td>
<td>Bw4 / Bw6, C1 / C2 allele group inference based on HLA allele matching table</td>
<td>2,3,12,15</td>
<td>33, 34</td>
</tr>
<tr>
<td>KIR gene presence</td>
<td>kir_generation</td>
<td>Presence or absence of specific KIR genes (binary variable)</td>
<td>33</td>
<td>34</td>
</tr>
<tr>
<td>HLA-KIR interactions</td>
<td>hla_kir_interactions</td>
<td>Experimentally verified ligand-receptor interactions between HLA class I and KIR</td>
<td>2</td>
<td>19-21</td>
</tr>
</tbody>
</table>

---

**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 (FHL) 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,23 which can be modeled by subsetting HLA-Bw4 further according to amino acid level information. Data transformation can then 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.24 Another use case for custom analyses is the transformation of HLA allele data into quantitative variables such as allele-specific expression levels.25–27 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 immunogenetic 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. The former 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 ‘tidiers’ as introduced in the ‘broom’ package (https://broom.tidyverse.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 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

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).