Published in final edited form as:

Methods Mol Biol. 2012 ; 882: 197–213. doi:10.1007/978-1-61779-842-9_12.

Standard Methods for the Management of Immunogenetic Data

Pierre-Antoine Gourraud, Jill A. Hollenbach, Thomas Barnette, Richard M. Single, and Steven J. Mack

Abstract

In this chapter, we outline some basic principles for the consistent management of immunogenetic data. These include the preparation of a single master data file that can serve as the basis for all subsequent analyses, a focus on the quality and homogeneity of the data to be analyzed, the documentation of the coding systems used to represent the data, and the application of nomenclature standards specific for each immunogenetic system being evaluated. The data management principles discussed here are intended to provide a foundation for the data analysis methods detailed in Chaps. 13 and 14. The relationship between the data management and analysis methods covered in these three chapters is illustrated in Figure 1.

The application of these data management principles is a first step toward consistent and reproducible data analyses. While it may take extra time and effort to apply them, we feel that it is better to take this approach than to assume that low data quality can be compensated for by large sample sizes.

In addition to their relevance for analytical reproducibility, it is important to consider these data management principles from an ethical perspective. The reliability of the data collected and generated as part of a research study should be as important a component of the ethical review of a research application as the security of those data. Finally, in addition to ensuring the integrity of the data from collection to publication, the application of these data management principles will provide a means to foster research integrity and to improve the potential for collaborative data sharing.

Keywords

Data management; Data standards; High polymorphism; HLA; Immunogenetics; KIR

1. Introduction

1.1. Avoiding the “Garbage-In Garbage Out” Predicament

In recent years, large amounts of genetic data have become available to the research community, increasing the potential for new findings in ways that have not previously been possible. The need for standardized data management and statistical analysis increases as more data become available in order to mitigate the variability between individual studies. The difficulty of dealing with heterogeneity between studies is not really new, but it becomes more significant as the amount of accessible data increases.

Sir Arthur Conan Doyle summarized the problem in 1890 in his second novel, “The Sign of Four”:

while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to.

Individuals vary, but percentages remain constant. So says the statistician.

Throughout its history, the field of immunogenetics has provided a continuously changing perspective of individual immunological characteristics. Different levels of information were involved: the source materials level (sera, DNA, RNA, etc.), the experimental level (with the switch from pure serological techniques to a mix of serological and molecular techniques, and the advent of parallel data acquisition instruments), and the analytical level (with development of computer-based data storage and statistical modeling). While immunogenetic data production capacity has clearly advanced significantly, it is not clear that the biostatistical capacity to analyze such quantities of data has developed in a similar manner.

In addition, immunogenetics is distinctive in that it is relevant to both very basic and applied fields of research. Transplantation is the seminal example of gene-based personalized medicine; basic knowledge in molecular anthropology, the history and genetic structure of populations, the genetic components of disease, forensic medicine, and genome evolution all benefit from the progress in immunogenetics. With rivers of data now flowing from high-throughput typing systems, a rigorous statistical approach is needed more than ever to navigate this new immunogenetic sea. In this chapter, we propose some foundational guidelines for the analysis of immunogenetic datasets, focusing on ensuring quality and homogeneity of data in order to escape from the “Garbage-in Garbage-out” predicament. Supplementary materials for this chapter can be found online at <http://methods.immunogenomics.org>.

2. Data Organization and Storage

2.1. Organization of a Master Data File

The “master data file” is the foundation of a quality data analysis. Ideally, a master data file will contain all the data needed for the analysis in a single electronic document, which can be validated, transferred, and exported. While individual software tools may require data to be organized in different ways (i.e., different data input formats), the master data file should be the source of all data analyzed regardless of the organization of those data. This first step crystallizes all the thought that has been put in the study design and the development of the hypothesis(es) under investigation. Study design is beyond the focus of this chapter but sample selection, phenotype assessment, and required sample size computations clearly warrant close attention. As shown in Figure 2, the master data file gathers demographic, phenotypic, and genotypic information; see supplementary Table S1 for a sample master data file.

2.2. Characteristics of a Master Data File

A typical master data file consists of (1) a series of rows dedicated to each analytical unit of the study and (2) a series of columns each consisting of either a “primary” or “derived” variable.

2.2.1. Units of Analysis—The master data file is generally structured using a single row dedicated to each analytical unit of the study. The analytical unit is the entity for which the required items of data have been collected and compiled. In general, analytical units are individual subjects, although in some cases the major entity being analyzed in a study is not an individual; for example, a transplantation study may focus on donor/recipient pairs, or an epidemiological study may focus on cases and controls. When nuclear families are analyzed, it may also be more appropriate to consider the nuclear family itself as the analytical unit of the study rather than its constituent members. In these cases, data on more than one individual may be included on a single row of the master data file. Conversely, when the study design involves repeated measures (over time or under different circumstances), observations of a single subject are usually split over several rows. This may be the case for example, when investigating the consistency of a measure between different techniques or between different operators.

2.2.2. Variables—Each column (or field) in the master data file should be dedicated to a variable. A variable is an observed or measured characteristic that can have more than one value and to which a numerical measure or a category from a classification can be assigned. It describes a single item in the data set. Where possible, each column should be numerically coded, and missing values must be given a conventional code as well. It is often useful to give each variable a short (but informative) name, which is stored in the first row of each column (also known as the column header). These variables are sometime called “raw” or “primary” variables, because they directly encode information; “analysis” or “derived” variables encode information that can be derived from the raw variables, and can be included in the master data file as well. For example, body mass index (BMI) can be calculated from weight and height variables, but it may be useful to include a calculated BMI in the master data file as well. The master data file may include other variable fields that will contain variable data generated as part of the analyses.

In general, the master data file should be designed to exclude redundancies between fields. At this early stage of the analysis, it is also important to identify the type of the variables being included in the master data file; the two most important variable types are “quantitative” and “qualitative” variables, and the statistical method that will be applied will depend on the type of variable being analyzed.

1. Quantitative variables: Quantitative variables provide numerical measures of a given characteristic being observed, and can be either “continuous” or “discrete.” Continuous quantitative variables have an infinite number of possible values that are not countable, whereas discrete quantitative variables have either a finite number of possible values or a countable number of possible values. For example, in hematopoietic stem cell allogeneic transplantation, the dose of CD34+ cells (e.g., $4.61 \times 10^6 \text{ kg}^{-1}$ of body weight) is a continuous quantitative variable. When these values are rounded (e.g., to $5 \times 10^6 \text{ kg}^{-1}$ of body weight), the variable may be considered as a discrete quantitative variable. In general, mathematical and statistical operations can be carried out on quantitative variables.

2. Qualitative variables: Qualitative (or categorical) variables allow for classification of individuals based on a characteristic, and can be either “nominal” or “ordinal.” Nominal

qualitative variables define unordered or non-hierarchical categories, whereas ordinal qualitative variables can be used to order or rank the observations. For example, leukemia subtypes can be defined with a nominal qualitative variable; 1 = acute lymphoblastic leukemia (ALL); 2 = acute myelogenous leukemia (AML), 3 = chronic lymphocytic leukemia (CLL); 4 = chronic myelogenous leukemia (CML), 5=unclassified leukemia, 999 = data missing, whereas stages of progression to type-1-diabetes (T1D) can be defined with an ordinal qualitative variable; 0 = normal glucose regulation, 1 = impaired fasting glucose, 2 = impaired glucose tolerance, 3 = diabetes mellitus, 999 = data missing. In general, qualitative variables are descriptors and not numbers, and mathematical operations are not appropriate. Instead, qualitative variables are useful for stratifying and dissecting patterns derived from the analysis of quantitative variables.

2.2.3. Data Dictionary—In addition to the numerically encoded variables, a master data file should also contain (or be associated with) a dictionary defining each variable and its associated coded values. The minimal items for this data dictionary are the short name of the variable, a detailed description of its content, a specification of the meaning of each associated numerical code, and the code used for missing values. Any additional information or comment that can be useful for a better characterization of the variable has to be mentioned. This effort will ensure sustainability of the data.

Figure 3 shows an example of the data structure in a master data file: numerically encoded data and data dictionary; see supplementary Table S2 for a sample data dictionary for Table S1. Figure 1 presents a “roadmap” to the landscape of statistical techniques; it summarizes the analytical approach: it goes from raw data storage, through analysis of missingness of the data, the use of classical descriptive statistical techniques (both numerical and graphical) to final statistical modeling fueling the interpretation of the data. Figure 1 is also available online as supplementary Figure S1.

2.3. Storage of Data

The best way to appropriately organize the data usually requires building a database. Because many research groups do not have the information technology (IT) resources to do so, a spreadsheet application (e.g., Microsoft Excel or its equivalent) is often used instead. Although very easy to use, accidental sorting, deletions, and unintended changes to the data can occur with such spreadsheet applications. For example, when HLA allele names are stored in Excel, it is common for allele names that should be treated as text (e.g., “0101” or “01:01”) to be treated as numbers or times (e.g., 101 or 1:01). Such data rearrangements may be difficult to detect and will certainly affect the results of the study. If spreadsheet applications are used to manage allele name data, individual datasets should be stored and transmitted as text-formatted “flat files,” and measures should be taken to ensure the validity of allele names prior to analysis.

3. Nomenclature Management

The high degree of polymorphism observed for immunogenetic genes and loci, coupled with the structural variation of immune genes, has resulted in regularly updated allele name nomenclatures that keep pace with the growing complexity of these systems, but each

nomenclature revision brings with it the potential for confusion when data generated using different nomenclature versions are analyzed together (the so-called immunogenetic tower of Babel) (1). This is a common problem for HLA and KIR genes, alleles, and haplotypes; MHC microsatellites, single nucleotide polymorphism (SNP), copy number variants (CNVs), and small insertion-deletion changes (indels). It is essential that the data in an analysis share a common nomenclature, and that the specific nomenclature version under which a dataset was generated (or validated if the dataset is compiled from multiple sources) is identified in the master data file.

3.1. HLA Nomenclature Variation

The HLA nomenclature for allele names has changed considerably over the last several decades. Each allele name contains a string of numbers arranged in “domains” that correspond to polymorphisms at the serological, protein, synonymous nucleotide, noncoding nucleotide, and expression levels. The nomenclature has been expanded as new alleles, and new forms of polymorphism have been identified, and the features of three major versions of this nomenclature are outlined in Table 1.

With each improvement to the nomenclature, some allele names have been deleted (2) or changed (e.g., B*1522 changed to B*3543 in 2002 after DNA sequences outside of exons 2 and 3 became available). With the adoption of the current nomenclature version, many allele names have changed in nonobvious ways (e.g., the DPB1*0502 allele changed to *104:01 in order to maintain the sequential order of DPB1 protein sequences), and the locus identifier for HLA-C locus alleles changed from Cw* to C* (3). If the nomenclature version pertinent to a given dataset has not been explicitly identified, it is important to be able to make an educated guess about the likely version under which the data were generated.

3.1.1. Antigen Recognition Sequence Nomenclature—Rather than treating each allele as distinct, it is often useful to analyze alleles that share the peptide sequences that constitute the antigen recognition sequence (ARS) or that share the nucleotide sequences that encode the ARS as a combined category. The ARS is alternatively referred to as the peptide-binding domain or peptide-binding region (respectively abbreviated as PBD or PBR). The ARS of class II HLA molecules is encoded by exon 2, and the class I ARS is encoded by exons 2 and 3. While the version 2 HLA nomenclature was in effect, allele-codes identifying alleles that share the same ARS encoding exon sequences were developed. For example, the “A*010101g” and “A*02 G1” codes have been used to represent all HLA-A alleles that shared exon 2 and 3 nucleotide sequences with the A*010101 allele (4, 5).

Version 3 of the HLA nomenclature includes official codes for identifying alleles that share an ARS or ARS-encoding exon sequence. Alleles that share an identical ARS are included in a P group (e.g., A*01:01P includes all HLA-A alleles that share the ARS with the A*01:01:01:01 alleles), and alleles that share identical exon 2 or exons 2 and 3 nucleotide sequences are included in a G group (e.g., A*01:01:01G includes all HLA-A alleles that share their exon 2 and exon 3 sequences with A*01:01:01:01) (3).

3.1.2. Nomenclature Conversion Resources—Several resources are available for the conversion of allele names between nomenclature versions. The ImMunoGeneTics

(IMGT)/HLA web site and database provides tables for the conversion of version 2 allele names to version 3 names, and houses a web-based conversion tool for individual allele names at http://www.ebi.ac.uk/imgt/hla/convert_name.html (6).

Entire datasets can be converted from nomenclature version 2 to version 3 using either the Allele Name Translation Tool (ANTT) or Update Nomenclature (UNCL), tools available from the immunogenomics data analysis working group (IDAWG) at <http://tools.immunogenomics.org> (7). UNCL is a web-based tool, and the ANTT is a locally run tool that can be customized to convert between other nomenclatures and name conventions. Both tools accept data in the form of tab-delimited text files, and generate tab-delimited text files of translated data.

The current version of the Helmsberg Sequence COmpilation and Rearrangement Evaluation (SCORE) virtual-DNA analysis database software (8, 9) will convert between version 2 and 3 nomenclatures for allele name data in its database. SCORE will also convert allele names to user-defined nomenclatures.

3.1.3. Allele Name Truncation—Because the polymorphic domains that make up an HLA allele name are arranged in a hierarchical fashion, it is possible to delete domains from the right end of an allele name while retaining important information for analysis. This process of deleting domains is known as truncation (aka right-truncation), and it is important to ensure that truncation has been carried out consistently within a dataset. For example, the full-length A*01:01:01:01 allele name can be truncated to A*01:01:01, A*01:01, and A*01. In some cases, truncation of allele names may be equivalent to applying a P or G group code, but this is not always the case. Depending on the research question being investigated, different levels of truncation may be appropriate (e.g., it may be appropriate to truncate allele names to the peptide level for a study of peptide presentation, but not for a study of allelic diversity and evolution), but all versions of a given allele name must be truncated to the lowest common level for analysis to avoid spurious results.

3.2. Microsatellite Nomenclature (14th Workshop)

In general, microsatellite alleles should be identified using both the number of repeats and the fragment size (10). In order to compare the results of several studies, the correspondence between the repeat number and the various fragment lengths (depending on primer pairs used) should be established. This correspondence list must include specific details (e.g., in the form of UniSTS numbers) of the primer pairs used to genotype each microsatellite (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=unists>). When multiple synonymous names are in use for a given microsatellite allele, the lowest numbered DS6 number should be used as a reference. For example D6S273 should be used to refer to the following synonymous names—142XH6, AFM142xh6, GC378-D6S273. More details can be found on the NCBI's MHC database, "dbMHC" (<http://www.ncbi.nlm.nih.gov/projects/gv/mhc/>).

3.3. Single Nucleotide Polymorphism and Insertion/Deletion Nomenclature

As recently discussed (1), multiple distinct identifiers may exist for a given simple genetic marker (e.g., a SNP or indel), sometimes leading to confusion in the comparison of markers

across studies. This is partly due to the fact that the unique reference sequences applied to structurally variable and highly polymorphic regions like the KIR cluster and the MHC region are not appropriate for these markers. One commonly accepted solution to this problem is to use the NCBI's dbSNP reference SNP (RefSNP) accession ID (rs number or rs#) (11). However, because rs numbers are not necessarily stable identifiers across successive dbSNP releases/builds, all references to an rs# should be accompanied by dbSNP build number. For example, rs16375 is not in use anymore in build 130 of dbSNP; rs1704 should be used instead.

3.4. SNPs: RS# and HGVS Nomenclature

Many other possibilities can be used as complementary indications which directly refer to the sequence changes (e.g., rs2306220:A>G). The other acceptable option is to use the Human Genome Variation Society (HGVS) nomenclature (12, 13). For rs16375/rs1704, an accessioned sequence from EBI, GenBank, or DDBJ can be used as the reference sequence NT 007592.14: 20656832insATTTGTTCATGCCT, but many different accessioned genomic or mRNA sequences can be used.

3.5. KIR Nomenclature

Ideally, the goal in storing and analyzing KIR data is to have some data representation for each chromosome at every locus. The limitations of the typing technology and the variation in KIR haplotype structure dictate that in many cases the second chromosome will be typed as “unknown” (?). Nevertheless, it is important to have data representation for the full genotype in order to facilitate downstream analyses. Each KIR locus typed can have at least one full genotypic record (more in the case of genotypic ambiguity, which is likely for allelic typing results). When a locus is absent from a given haplotype, this absence must be coded in the data (i.e., as an “absent” allele), and should be denoted as “0.”

Where allelic typing is available, a genotype with two alleles represents the simplest case scenario, a heterozygote with the locus present on both haplotypes. However, when we have only one allele detected, e.g., KIR2DL2*002, there are two possible genotypes: “002,002” or “002,0.” At present, most typing systems cannot distinguish these. In these cases, the genotype should be treated as ambiguous, i.e., 002,0/002.

For typing data that is strictly presence/absence, the data can be treated as for a biallelic locus: allele “1” = “present” and allele “0” = “absent.” If the locus is absent, we have the full genotype 0,0; however for locus present, we have an ambiguous genotype, which may be either “1,0” or “1,1.” This can also be represented as an ambiguous genotype : 1,0/1. Because this notation does not consider order, a genotype coding of “0,1” is not used, but would be treated the same as “1,0.”

Some KIR loci have particular additional details to consider prior to analyzing data:

1. KIR2DL2/L3—KIR2DL2 and KIR2DL3 were formerly treated as separate loci. The alleles were not named in a series; therefore, it is important to distinguish between KIR2DL2 and KIR2DL3 when the data is recorded. In the case of this locus, since all typing

systems are able to detect both KIR2DL2 and KIR2DL3 an investigator should always have a full genotype, i.e., no missing data or ambiguity with presence/absence typing.

2. KIR2DL5—KIR2DL5 may be either centromeric or telomeric in the KIR cluster, and possibly on both ends. As such, an individual may have four copies of KIR2DL5; at this time there is no way to distinguish these definitively. Many typing systems will simply type for the presence of KIR2DL5; if typed as “absent,” there is confirmation of genotypes of 0,0 both centromerically and telomerically. However, a typing of ‘present’ gives the genotype 1,0/1 on either or both sides of the KIR complex. Some typing systems currently distinguish “KIR2DL5A” and “KIR2DL5B”; all current data suggest that KIR2DL5A corresponds to the telomeric position and that KIR2DL5B corresponds to the centromeric position.

3. KIR2DS3/S5—As with KIR2DL5, KIR2DS3/S5 may be either centromeric or telomeric in the KIR complex, and possibly on both ends. As such, an individual may have four copies of the gene. Previously, KIR2DS3 and KIR2DS5 were thought to be different loci, and all current typing systems can distinguish between them. However, an individual who is typed, for example, as positive for both KIR2DS3 and KIR2DS5 in a presence/absence typing system may have these two alleles on either or both sides of the KIR cluster. Some data is emerging that suggests that particular alleles of each of these may be either centromeric or telomeric, but this is not yet confirmed. Presently, the only way to ascertain definitively that an individual has both centromeric and telomeric copies of either KIR2DS3 or KIR2DS5 is the case where we have allelic typing with more than two allele calls for the locus.

4. KIR3DL1/S1—As above for KIR2DL2/L3, KIR3DL1 and KIR3DS1 were formerly treated as separate loci. However, it has been recognized for some time that they are allotypes of the same locus, and in this case alleles have been named in a series. As with KIR2DL2/L3, all typing systems minimally detect KIR3DL1 and KIR3DS1, so there should be no missing data/ambiguity in presence/absence typing systems at this locus. If a typing is submitted as, e.g., “3DL1,” we know that the genotype is “3DL1,3DL1.”

Acknowledgments

This work was supported by National Institutes of Health (NIH) grants U01AI067068 (JAH, SJM) and U19 AI067152 (PAG) awarded by the National Institute of Allergy and Infectious Diseases (NIAID) and by NIH/NIAID contract AI40076 (RMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases or the National Institutes of Health.

Glossary

Genetic data

1. **Allele:** Any of the alternative forms (sets of forms) of DNA sequence at locus. These variants may occur for genes and/or genetic markers.

Example: B, HLA-DRB1*01:01:01, HLA-A*01, D6S1666 (184).
2. **Diplotype:** The pair of haplotypes within a given genotype. The chromosomal phase between alleles is always known.

Example: HLA (A*01 B*08 DR*17, A*26 B*27 DR*17).

HLA (A*01 B*27 DR*17, A*26 B*08 DR*17).

When analyzing data for more than one locus, diplotypic data must be distinguished from genotypic data. Because the alleles at different loci in a given genotype can be combined to make many possible haplotype pairs, a genotype must be considered to correspond to multiple diplotypes. Unfortunately, the term “genotype” is sometimes used to refer to a given pair of haplotypes, especially when familial segregation has been studied.

3. Gene: The functional and physical unit of heredity. A gene consists of a DNA segment with a specific sequence. It includes information for the synthesis of mRNA molecules that direct the synthesis of proteins.

Example: *ABO Glycosyltransferase* gene, *HLA-DRB1* gene, *KIR-2DS3/S5* gene.

4. Genotype: The genetic makeup at one or more loci of an individual. It refers to a set of alleles carried by an individual (regardless of the expression of those alleles). The chromosomal phase (chromosomal identity of alleles at different loci) between alleles may not be known.

Example: HLA-A (1, 2); HLA-B (8, 44); HLA-DRB1 (03, 04).

KIR: KIR (A, A).

Microsatellites: D6S273 (134, 136) D6S273($\ast(GT)_{19}$, $\ast(GT)_{20}$).

SNP: RS345336443 (G/G).

5. Haplotype: Set of alleles of contiguous loci. They are usually co-transmitted on a parental chromosome.

Example: HLA-A*01-B*08-DRB1*03.

6. Locus: Literally, “place” in Latin, it is the specific usual physical location of genes, an individual genetic marker, or set of genetic markers in a genome.

Example: ABO locus, HLA-DRB locus, KIR-2DS3/S5cen locus, D6S1666 microsatellite locus.

7. Phenotype: The observable expression of alleles as a physical or biochemical trait resulting from the interaction of the genome, the environment, and the experimental settings. In disease studies it may refer to the presence or a manifestation of the disease under study. Disease phenotypes may be reflected in a variety of ways as quantitative or qualitative variables.

**Phenotypic
and
demographic
data**

This term may also refer to a set of alleles (expressed or not) detected by a technique. In codominant or heterozygous situations, phenotypes are noted as pairs of data; each pair is specific to a particular gene and locus.

Example: ABO system: [A]. HLA system: [HLA-A (1, 2); HLA-B (8,44); HLA-DR (3, 4)].

- 1** Admixture: The outcome of interbreeding between members of different populations. An admixed population is generally derived from populations in different geographic regions.
- 2** Collection site: The location where the sample was collected. This can be identified using latitude and longitude coordinates, or by specifying the country or nation, and city/town/village, or other locale where the collection took place.
- 3** Complexity: An ordinal variable that represents an estimate of the degree of admixture and population sub-structure in each population sample.

Example:

Complexity 1: a population sample collected from a single settlement or group of closely related settlements.

Complexity 2: a population sample collected from a group of separate but discrete settlements.

Complexity 3: a population sample collected in a metropolitan area or across an entire nation.

Complexity 4: an admixed population.

- 4** Data management methods: The approaches used in storing and processing the data in preparation for analysis. This can include the formats and programs used to store and edit the data (e.g., a specific spreadsheet program or database system), as well as any modifications that were made to the data between the generation of the data resulting from the typing assay and the inclusion of the data in the master data file. For example, if ambiguities were resolved, the approach used to resolve them should be documented in the data dictionary; if HLA allele data were truncated to a common level, or “binned” into a common sequence category (e.g., treating all alleles that encode the same peptide-binding region as the same allele) this should be documented.
- 5** Ethnicity: A group of individuals (or populations) sharing a common language, culture, or religion, and who are assumed

to share a common ancestry. Ethnicity should be distinguished from geography (e.g., “North American” is not an ethnicity), and though ethnicity is often associated with indigenous nationality (e.g., “Irish,” “Chinese”) qualifiers are often necessary to distinguish ethnicity from nationality (e.g., “Han Chinese”).

- 6** Family: If individuals in the study belong to discrete familial groups, a family ID is qualitative variable identifying membership in a particular pedigree, as well as the relationship to the index case (proband).
- 7** Geographic region: A specific continental or subcontinental area comprised by multiple nations in which the population is located, or from which the population was derived, if the population is a migrant population. For example, European Americans or European Australians would be assigned to the European region, or to a specific subregion of Europe. Conversely North America would only pertain to Native American/Amerindian/Aleut/Eskimo populations. Populations derived from more than one region (admixed populations) can be assigned to a specific class for the type of admixture (depending on the regions of origin) or included in a single class for all admixed populations. The definitions of each region and admixed class should be defined in the data dictionary.
- 8** Latitude and longitude: Geographic coordinates that specify specific locations on the surface of the Earth. Latitude and longitude values should be recorded in a decimal format, with minutes and seconds indicated as fractions of each degree value. North latitudes and east longitudes should be recorded with positive values, and south latitudes and west longitudes should be recorded with negative values. For example, 35° 20 min south latitude would be recorded as -35.333, and 2° 30 min east longitude should be recorded as 2.5 or +2.5.
- 10** Population: A group of individuals living in a specific geographic area. More specifically, a population is defined such that all pairs of individual members have the opportunity to mate, and are more likely to mate with each other than with members of other populations. A population should be documented in the data dictionary in terms of the pertinent geographic area and the approximate number of included individuals. Population sample: A unique

descriptor for the individuals from a given population that were included in the study. If the study involves multiple sets of individuals (samples) from the same population, each set of individuals should be given a unique name; usually it is sufficient to append the number of individuals to the end of the population name (e.g., antarctica_87, antarctica_207, antarctica_597).

- 11** Population substructure: A barrier to the opportunity of mating between all pairs of individuals in a population.
- 12** Proband: The individual under study, primarily used in family-based disease association studies.
- 13** Status: The status of an individual as affected or unaffected with respect to a disease phenotype, or belonging to a case or a control group.
- 14** Typing assay: The laboratory method(s) and associated protocols used to generate the data included in the analysis. Many of them are described in this volume. Commonly used molecular methods for HLA and KIR genotyping include sequence-specific priming (SSP), sequence-specific oligo probe (SSO or SSOP), sequence/sequencing-based typing (SBT), matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF), and reference strand conformation analysis (RSCA). Serology has been used historically for HLA phenotype data generation. When possible, a description of the assay identifying the assay manufacturer and reagent version/lot employed should be included in the data dictionary. Literature citations or references to specific protocols should also be associated with the methods used, especially if multiple distinct methods have been employed in generating the data.

References

1. Gourraud PA, Feolo M. The Babel Tower revisited: SNPs—Indels—CNVs. Confusion in naming sequence variant always rises from ashes. *Tissue Antigens*. 2010; 75:199–200. [PubMed: 20030785]
2. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Geraghty DE, Hansen JA, Mach B, Mayr WR, Parham P, Petersdorf EW, Sasazuki T, Schreuder GM, Strominger JL, Svejgaard A, Terasaki PI. Nomenclature for factors of the HLA system, 2002. *Tissue Antigens*. 2002; 60:407–464. [PubMed: 12492818]
3. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Fernández-Viña M, Geraghty DE, Holdsworth R, Hurley CK, Lau M, Lee KW, Mach B, Maiers M, Mayr WR, Müller CR, Parham P, Petersdorf EW, Sasazuki T, Strominger JL, Svejgaard A, Terasaki PI, Tiercy JM, Trowsdale J. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*. 2010; 75:291–455. [PubMed: 20356336]

4. Cano P, Klitz W, Mack SJ, Maiers M, Marsh SG, Noreen H, Reed EF, Senitzer D, Setterholm M, Smith A, Fernández-Viña M. Common and well-documented HLA alleles: report of the Ad-Hoc committee of the American society for histocompatibility and immunogenetics. *Hum Immunol*. 2007; 68:392–417. [PubMed: 17462507]
5. Robinson, J.; Mistry, K.; Marsh, SGE. Exon identity and ambiguous typing combinations. Anthony Nolan Research Institute; 2010. http://www.ebi.ac.uk/imgt/hla/pdf/ambiguity_v2280.pdf
6. Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SG. The IMGT/HLA database. *Nucleic Acids Res*. 2011; 39(Database Issue):D1171–D1176. [PubMed: 21071412]
7. Mack SJ, Hollenbach JA. Allele Name Translation Tool and Update NomenCLature: software tools for the automated translation of HLA allele names between successive nomenclatures. *Tissue Antigens*. 2010; 75:457–461. [PubMed: 20412076]
8. Helmberg W, Lanzer G, Zahn R, Weinmayr B, Wagner T, Albert E. Virtual DNA analysis—a new tool for combination and standardised evaluation of SSO, SSP and sequencing-based typing results. *Tissue Antigens*. 1998; 51:587–592. [PubMed: 9694350]
9. Helmberg W. Storage and utilization of HLA genomic data—new approaches to HLA typing. *Rev Immunogenet*. 2000; 2:468–476. [PubMed: 12361090]
10. Gourraud PA, Cambon-Thomsen A, Dauber EM, Feolo M, Hansen J, Mickelson E, Single RM, Thomsen M, Mayr WR. Nomenclature for HLA microsatellites. *Tissue Antigens*. 2007; 69(Suppl 1):210–213. [PubMed: 17445203]
11. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001; 29:308–311. [PubMed: 11125122]
12. den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mut*. 2000; 15:7–12. [PubMed: 10612815]
13. den Dunnen, J. Nomenclature for the description of sequence variants. Human Genome Variation Society; 2010. <http://www.hgvs.org/mutnomen/>
14. Bodmer JG, Marsh SGE, Parham P, Erlich HA, Albert E, Bodmer WF, Dupont B, Mach B, Mayr WR, Sasazuki T, Schreuder GMT, Strominger JL, Svejgaard A, Terasaki PI. Nomenclature for factors of the HLA system, 1989. *Tissue Antigens*. 1990; 35(1):1990.
15. Who Nomenclature Committee. Nomenclature for factors of the HLA system, 1987. *Tissue Antigens*. 1988; 32:177–187. [PubMed: 3217934]
16. Bodmer JG, Marsh SG, Albert ED, Bodmer WF, Dupont B, Erlich HA, Mach B, Mayr WR, Parham P, Sasazuki T, et al. Nomenclature for factors of the HLA system, 1990. *Hum Immunol*. 1991; 31(3):186–194. [PubMed: 1890020]
17. Bodmer JG, Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Charron D, Dupont B, Erlich HA, Mach B, Mayr WR. Nomenclature for factors of the HLA system, 1995. *Tissue Antigens*. 1995; 46:1–18. [PubMed: 7482491]
18. Bodmer JG, Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Charron D, Dupont B, Erlich HA, Fauchet R, Mach B, Mayr WR, Parham P, Sasazuki T, Schreuder GM, Strominger JL, Svejgaard A, Terasaki PI. Nomenclature for factors of the HLA system, 1996. *Tissue Antigens*. 1997; 49:297–321. [PubMed: 9098945]

Fig. 1.

A roadmap to the landscape of statistical techniques. This roadmap is a guide summarizing potential approaches to data analysis. Starting with raw data in the center, analyses proceed first through an assessment of missing data, and then the application of classical descriptive statistical techniques (both numerical and graphical), before a set of final statistical tests, specific to the type of data and the hypothesis being tested, and which inform the interpretation of the data, are selected.

Fig. 2.

The benefits of a single master data file. A master data file integrates all data pertinent to a given study in a standard fashion. Genotype and phenotype data that must be compiled for a research project are shown on the *right*. These are all incorporated into a single master data file, which is stored in a database, or as an MS Excel table.

Fig. 3. Characteristics of a Master Data File and associated Data Dictionary. A data dictionary provides context for the numerically encoded observations recorded in the master data file. Each field (*column*) in the master data file corresponds to a row in the data dictionary, where it is described in a standardized fashion.

Table 1

Features of the three nomenclature versions for classical HLA allele names

Nomenclature		Domain size (number of characters)				
Version	Epoch	Serologic ^a	Protein	Synonymous	Noncoding	Expression variants
1	1987–2002	2 ^b	2 ^b	1 ^c	2 ^d	N ^d , L ^e
2	2002–2010	2	2	2 ^f	2	L, N, A ^g , C ^g , S ^g
3	2010+	2	2+ ^h	2+ ^h	2+ ^h	A, C, N, L, S, Q ⁱ

^aThe concept of a serologic specificity does not apply to DPB1 alleles. The first two digits of allele names for unique DPB1 protein sequences have been assigned in numerically (14)

^bImplemented in 1987 (15)

^cImplemented in 1990 (16)

^dImplemented in 1995 (17)

^eImplemented in 1996 (18)

^fThis domain was expanded to two digits in 2002 (2)

^gImplemented in 2002 (2)

^hThese domains were explicitly defined as colon-delimited fields in 2010, and can accommodate any number of variants (3)

ⁱImplemented in 2010 (3)