Using UNIFORMAT and GENE[RATE] to Analyze Data with Ambiguities in Population Genetics

José Manuel Nunes

Department of Genetics and Evolution, Anthropology Unit, University of Geneva, Geneva, Switzerland.

Supplementary Issue: Evolutionary Genomics

ABSTRACT: Some genetic systems frequently present ambiguous data that cannot be straightforwardly analyzed with common methods of population genetics. Two possibilities arise to analyze such data: one is the arbitrary simplification of the data and the other is the development of methods adapted to such ambiguous data. In this article, we present an attempt at such a development, the UNIFORMAT grammar and the GENE[RATE] tools, highlighting the specific aspects and the adaptations required to analyze ambiguous nominal data in population genetics.

KEYWORDS: ambiguous genetic data, frequency estimation, EM algorithm, data manipulation, Hardy-Weinberg, linkage disequilibrium

SUPPLEMENT: Evolutionary Genomics

CITATION: Nunes. Using UNIFORMAT and GENE[RATE] to Analyze Data with Ambiguities in Population Genetics. *Evolutionary Bioinformatics* 2015:11(S2) 19–26 doi: 10.4137/EBO.S32415.

TYPE: Methodology

RECEIVED: September 29, 2015. **RESUBMITTED:** December 28, 2015. **ACCEPTED FOR PUBLICATION:** December 31, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1775 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by Swiss FNS grants #31003A_127465 and #31003A_144180 and by COST Action BM0803. The author confirms that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Author discloses no potential conflicts of interest.

CORRESPONDENCE: Jose.deAbreuNunes@unige.ch

Introduction

Diploid systems are characterized by genotypes with two alleles at each locus: one inherited from the mother and the other from the father. The typing technologies used to determine these alleles do not always reveal exactly two alleles; sometimes only one is seen and sometimes more than two seem to be present. An illustration of such data is provided by the typing of the major histocompatibility complex in humans, generally referred to as human leukocyte antigen (HLA), routinely made for clinical purposes and also used for studies of human peopling history and population genetics.[1–9] Such data, usually designated as ambiguous, have the characteristic of not providing a single two-allele genotype (possibly two identical alleles) for every individual typing. Ambiguous data, then, indicate an incertitude in the determination of the two alleles of the genotype of an individual rather than the occurrence of more than two alleles in an individual's genotype.

Essentially, all classical methods of population genetics have been conceived for data without ambiguities and, therefore, are not appropriate to handle ambiguous data. This is even more so for nominal data, where alleles are the distinct forms, potentially infinite as described by Kimura and Crow (see Hedrick[10]), possible for a given locus. A common approach is to preprocess the data in order to eliminate ambiguities and hence obtain single two-allele genotypes.[11] The problem with such approaches is that they always involve arbitrary decisions that, at least in the long run, are inconsistent. For instance, some alleles now well identified with sequencing techniques were initially seen in ambiguous data typing, but they were reported as not present as a result of preprocessing treatments.[12] To tackle this problem, we have applied the alternative approach of adapting population genetics methods to ambiguous nominal data (see references below).

To achieve a generalization of population genetics methods adapted to ambiguous and highly polymorphic diploid data, a number of steps were involved. A fundamental first step was to establish a well-defined format that could clearly and explicitly describe ambiguities, that is, UNIFORMAT, which is described in the following section. A second step was to adapt, prove, and validate algorithms, encode them as programs, and test them empirically. A third point to address was how to handle specific difficulties related to the generalized methods. For instance, when estimating genetic frequencies, it was necessary to take into account the diagnosis of the estimation procedure, such as the number of distinct solutions and its control based on the convergence of the log likelihood. It was also necessary to find strategies to deal with the possible effects of indistinguishable ambiguities (allele blocks), and a further challenge was to find alternatives to essential equilibrium measures, such as Hardy–Weinberg equilibrium (HWE) and linkage disequilibrium (LD), that could take into account the extra information provided by the ambiguities and avoid statistical tests whose assumptions are likely to be broken by ambiguous data.

In this article, we provide an updated description of UNIFORMAT and a unified account of the methods using the GENE[RATE] tools, including references to some applications already running and published studies that use them. As far as we know, the approach presented here is the only one currently available, which is able to handle both highly polymorphic (typical samples include hundreds of alleles per locus and tens of thousands of two-locus haplotypes) and ambiguous data.

The GENE[RATE] framework. The algorithms and programs described in this manuscript have been developed in the context of analyses of ambiguous HLA data.[2,13–15] These tools have been used in a number of studies already published, but here, we provide an integrated description of all the tools and a number of updates and improvements. Furthermore, we want to present this suite of tools in a larger, non-HLA context of ambiguous diploid data for which they are also expected to be useful.

The current implementation depends on other software and computer libraries in addition to the key components presented in the following section, all of which constitute a suite of facilities to perform routine tasks in population genetics data analyses. The GENE[RATE] tools are available online at two addresses: the original page: http://geneva.unige.ch/generate and a version tightly integrated with an HLA database, http://hla-net.eu/tools, whose usage is described in a previous work.[16] Examples of UNIFORMAT files and their preparation using common computer programs are provided in the website. These include text documentation, the slides of a tutorial presentation, and a screencast on the preparation and validation of UNIFORMAT files.

The essence of the framework is a suite of computer programs that handle highly polymorphic and ambiguous data organized around a data format that allows for the expression of complex phenotypes. The GENE[RATE] tools currently include utilities to convert files from and into a few data formats and to recode datasets by transliterating some alleles or combinations of alleles into others – an operation required, for instance, in studies involving several population samples, all of which are not typed at the same time or in the same laboratory. The tools also include programs to estimate a number of one- and two-locus parameters relevant for population genetics, whose rationale is presented in the following section.

Methods

UNIFORMAT version 3. To describe ambiguous data, we formally defined the UNIFORMAT grammar.[17] This grammar allows the user to express all kinds of ambiguities that can occur in diploid data and includes abbreviations that may help express some usual cases, such as possible but uncertain homozygous, one "known" allele and several possible second alleles and untyped loci. Examples of data written in UNIFORMAT are given as follows:

```
#  a simple double heterozygous
#  for HLA-A and a locus with alleles k– and k+
id  A*01,A*02:01  k–,k+

#  an untyped case for first locus
id  @  k+,k–
#  a homozygous-or-blank-heterozygous for the second locus
id  A*02,A*11  k+

#  a real homozygous for first locus and a multiple allele for second locus
id  A*01,A*01  B*07,B*14:01&B*14:02&B*14:05

#  a case of multiple allele pairs for a locus
id  B*07,B*14:01|B*07,B*14:02|B*07,B*14:05
```

The formal specification of the current version of UNIFORMAT in a BNF-like form (see Levine[18] for BNF-like descriptions), slightly simplified from the one actually used in the parser implementation by omitting terms related to error control, follows below:

```
sample      : data EOF | EOF

data        : data case | case

case        : IDENT LOCI_SEP full_pheno

full_pheno  : full_pheno LOCI_SEP locus_pheno | locus_pheno
locus_pheno : multi_alp | ALLELE | MULTI_AL
            | AL_UNDET | AL_UNDET_NO_BLANK

multi_alp   : multi_alp ALP_SEP basic_alps | basic_alps

basic_alps  : ALLELE AL_SEP ALLELE
            |ALLELE AL_SEP MULTI_AL
            |MULTI_AL AL_SEP ALLELE
            |MULTI_AL AL_SEP MULTI_AL
```

In easy-to-read language, UNIFORMAT can be described as follows: a sample is a file where each line represents an individual; each line starts with an identifier and is followed by the phenotypes for the locus or loci of the individual; all the parts are separated by white space; each locus phenotype is a string (without spaces inside it) where allele pairs (that we call alps) are separated by a vertical bar, "|", and alleles of a pair by a comma, with the proviso that the left allele is "smaller" than the right one. This last condition prevents an allele pair from appearing under two different forms (such as a_1, a_2 and a_2, a_1), which is likely to introduce confusion. It also enables simplification of the algorithms, allowing for huge speedups during computer-intensive calculations because no comparisons are required. The condition assumes that allele names can be ordered in some way, for instance, alphanumeric sorting, or, for computer implementations, numeric sorting (easily implemented by translating allele names into numbers representing their positions

in a list of allele names). Allele names can also be used to describe haplotypes, by separating the locus allele names with a tilde sign ("~"). All the semantics of single allele names are valid for haplotype names, and currently, these can be mixed, which might lead to strange, but useful expressions. For instance, A*01~B*08,A*02 represents a typing for which one haplotype is known to be exactly A*01~B*08, while the other haplotype is only known to have A*02 at its first locus. Actually, haplotype data can be represented as a special case of (a highly polymorphic and ambiguous) single locus data.

Finally, allele names (ALLELE) and identifiers (IDENT) can use any of the following characters: alphabetic and numeric characters and also characters from the set {*, +, /, –, _, ', : , ~}. Other characters will produce invalid names, and some do have special meaning; hence, LOCI_SEP, the locus separator, is a horizontal white space; AL_SEP, the allele separator, is a comma; ALP_SEP, the alps separator, is a vertical bar; untyped loci are marked with the "@" sign corresponding to AL_UNDET or AL_UNDET_NO_BLANK; a multiple allele, MULTI_AL, is just a suite of alleles (ALLELE) separated by an ampersand, "&"; and EOF means the end-of-file.

The main changes introduced by UNIFORMAT version 3 are the simplification of the locus separator, which can now be any white space (any number of tabs and spaces) and not necessarily one tabulation mark as before, and the integration of the treatment of the haplotype notation. These simplifications are possible because, in this version, white spaces cannot occur inside single locus phenotypes. Although no other changes were made to the grammar definition, its actual implementation as a parser has been much improved by using hash tables to store allele (and haplotype) names, allowing for a faster encoding of input files and decoding of results into output files. The current implementation also allowed the integration of the previous tools into a single tool for validation and recoding. Overall, this simplifies the creation and manipulation of UNIFORMAT files, while maintaining compatibility with version 2 files. Files in UNIFORMAT version 1 can no longer be used because of the HLA nomenclature changes effective in 2010,[19] but they can still be converted to newer versions using the validation tool.

Tools for using UNIFORMAT effectively. The grammar defined earlier has been implemented in a program, UNIFORMAT, that checks if a file is a valid UNIFORMAT file, performs all abbreviation expansions, and, in case of errors, returns an indication of the lines where the errors occur.

The same program can also perform recoding, or transliteration, of valid UNIFORMAT files. Recoding or transliteration is an operation often required when working with population data samples coming from different laboratories, or not typed with the same techniques, or typed at different times using different allele definitions. The transliteration facility makes this operation much simpler, as a single transliteration file can be used for all data sources that are used in a project. The transliteration file is just a list of old and new names; an example is provided in the web page of the tool.

There are many reasons for which "file conversions" are needed; that is why there are many options for this tool. When the data are compatible (usually this means not ambiguous), both forward and backward conversions are available. The formats that have been included in this tool are the ones that we have most commonly found in our practice, but they may be extended in the future.

Some people develop their own typing kits, and, for those, PHENOTYPE might be the tool of choice for a fast and accurate interpretation of the results. This tool expects two inputs: a probe-definition file and a typing-reactivity results file. The output is a UNIFORMAT file that provides the interpretation of the reactivities in terms of allele pairs required to explain them. The originality of this tool is the use of allele pairs rather than lists of alleles, thus avoiding spurious ambiguities such as those resulting from the use of NMDP codes, see discussions in Buhler et al and Sanchez-Mazas et al.[14,20]

Frequency estimation. All methods based on the expectation–maximization (EM) algorithm[21] are actually generalizations of the gene-counting method initially proposed by Ceppellini et al.[22], which were extended to haplotypes in 1995.[23–25] Implementations of the gene-counting methods were further extended to deal with ambiguities (initially a fixed number, around 200; personal communication by J. Clayton in 1995, personal communication) and further generalized to any number of ambiguities and loci around 2005.[26]

The principles of the GENE[RATE] implementation of frequency estimation have been succinctly described,[17] and the relevant mathematical details appear in the Appendix. Basically, it is an implementation of the gene-counting method that allows the use of ambiguous data by representing them as alternative allele pairs. What was and still is particular to the GENE[RATE] implementation is that the algorithm is controlled by an EM algorithm rather than by the changes in frequency estimates (as in the original gene-counting method and most current EM implementations). The algorithm itself is implemented in such a way as to report quality information about the estimates, ie, the number of iterations until convergence and the number of distinct solutions found. This is indeed essential because the estimation procedure does not necessarily have a unique solution.[27] The assessment of the uniqueness of the solution is made by using multiple random starting points and checking if they all converge to the same maximum, as in Excoffier and Slatkin,[24] but, unlike these authors, we report all distinct solutions found and not only the one with the largest log likelihood. If the solution is unique, it means that we have good estimates; if there are multiple solutions, the results should not be used as frequency estimates. Common ways to tackle the problem of getting multiple solutions include increasing the sample size, reducing the level of

resolution of the typing, or recoding some indistinguishable alleles as a single entity. Sometimes these remedies produce the desired effect (ie, a single solution), but sometimes they do not, meaning that there are no acceptable maximum likelihood frequency estimates. From our experience, this is very rarely the case with one-locus alleles and rare with two-locus haplotypes, whereas it happens more frequently with highly ambiguous data, possibly with several loci, and with not so large sample sizes. In practice, this has rarely been a problem for us. The effect of sampling variation is much more important than variation due to the use of ambiguities in the estimation process, as shown in a previous study.[28] The possibility of dealing with ambiguities requires scrutiny from the user, especially if the relative amount of ambiguities in the sample is large. An account of the questions raised by the use of ambiguities is given in Buhler et al.[14]

A further extension of the gene-counting estimation implemented in GENE[RATE] is the replacement of Hardy–Weinberg with a more general model that includes one parameter to allow for deviation from Hardy–Weinberg proportions (see mathematical details in Appendix). This model is a key element for efficient Hardy–Weinberg testing with our method, as described in the following section.

Testing HWE. Testing for HWE on HLA data presents a number of difficulties, such as non-observed genotypes, ambiguous data, and the presence of a recessive-like blank allele. To avoid such difficulties, we considered an approach using a likelihood ratio test (LRT). The test consists of comparing the likelihood of frequency estimates under the Hardy–Weinberg model with the likelihood of frequency estimates under a model that generalizes HWE by including an extra parameter (such as an inbreeding coefficient). Thus, we consider two models: one of them being a particular case of the other. Formally, the full model is a function of both the allele frequencies and a parameter, F, which measures deviation from the Hardy–Weinberg model:

$$L_0 = f(p_i, F)$$

The usual Hardy–Weinberg model just sets this extra parameter to zero:

$$L_{HWE} = f(p_i, F = 0) = f(p_i)$$

As usual for LRTs, twice the difference of the log likelihoods follows a chi-square distribution with one degree of freedom. This test presents an advantage over the more usual chi-square, exact, or Monte–Carlo–Markov–Chains (MCMC) tests, in that it is not affected by the problems mentioned at the beginning of this section.

Assessing selective neutrality. Neutrality testing is frequently performed using the revised version of the Ewens–Watterson test proposed by Slatkin, hereafter EWS.[29,30] Using this test with ambiguous data is, however, problematic due to the presence of genotypes that are not determined unequivocally. To tackle this problem, a parametric resampling schema is used in such a way that the EWS is applied to a batch of samples with no ambiguous genotypes, which is randomly drawn from the allele frequencies estimated in the population.[28] The batch of *P*-values obtained in this way is then adjusted to maintain the false discovery rate at its nominal level.[31] This correction for multiple testing improves the Bonferroni correction method originally proposed in Ref. 28. The adjusted values are then used to assess the putative deviations from neutrality.

As an indication of the quality of the assessment, the number of zeros or ones, which are values, respectively, smaller or larger than all others, is reported, but the adjusted *P*-values actually indicated correspond to non-zero or non-one *P*-values. From our experience, getting zeros or ones without also getting small non-zero (typically smaller than 2.5%) or large non-one (typically larger than 97.5%) *P*-values almost always indicates that the number of samples generated is not large enough to capture the full variability of the EWS statistic. Thus, the test should be rerun with a higher number of bootstrapped samples.

Linkage disequilibrium. LD for two bi-allelic loci reduces to a single coefficient; hence, the existence of a haplotype in LD implies that the other three haplotypes are also in disequilibrium. In this case, we can also say that the two loci are in LD. This simple situation does not hold when working with loci having more than two alleles. Therefore, we need to consider separately global LD, and the LD of each individual haplotype. The latter is the same as for bi-allelic loci, ie, the difference between the observed (often estimated) haplotype frequency and the product of the frequencies of the alleles defining the haplotype. The global measure of LD is supposed to provide a summary of the situation, taking into account all possible haplotypes for two given loci. It is possible that two loci are not in global LD while some specific haplotypes for these loci do present strong LD. On the other hand, global LD requires that at least one haplotype exhibits strong LD.

To test the LD, the usual chi-square test (or equivalents such as Fisher's exact test and its MCMC approximations[32]) is in general inapplicable to HLA data, because it cannot handle ambiguous data or the presence of a putative recessive-like blank allele, and it also suffers from the large number of haplotypes that are generally not observed (ie, with estimated frequencies of 0). Instead, our approach consists in using an LRT (as for the Hardy–Weinberg test) for comparing two estimations: the log likelihood of the estimated haplotype frequency and the product of the log likelihoods of the estimated allele frequencies. Under the hypothesis of no LD, the two statistics are expected to be similar. As the model of "no LD" can be seen as a special case of the more general LD model (by setting the disequilibrium coefficients of all haplotypes to zero), the number of degrees of freedom associated with this LRT used is the number of possible haplotypes. Formally, this uses a full model described by the following equation:

$$L_0\left(p_i, q_j, \delta_{ij}\right) \propto \prod_{ij}\left(p_i q_j + \delta_{ij}\right)^{n_{ij}}$$

where n_{ij} is the count of ij haplotypes (observed or estimated). The simple "no LD" model can be described by $L(p_i) \times L(q_j)$, which means that the simple model results from the full model when all δ_{ij} are zero.

In general, the number of degrees of freedom associated with this likelihood is provided by the number of independent (free) parameters, which is given by considering the restrictions on the sums of the frequencies of all haplotypes carrying a given allele (they must add up to the allele frequency) and on the sum of the allele frequencies at the two loci, that is $\frac{(k_1-1)\times(k_2-1)}{2}$, where k_1 and k_2 denote the number of alleles at each locus. Unfortunately, this number is too large for typical HLA data, and even the largest samples, donor registries with millions of individuals, show only a small part of the total number of haplotypes. To address this issue, we adjust the number of degrees of freedom to the number of possible haplotypes (in the sense of being potentially present in the sample given the two-locus phenotypes). This empirical practice is justified by the close agreement that we have generally observed between the *P*-values provided by it and those of a distribution-free method presented in the following section.

The previous likelihoods provide the first measure of global LD that we report in our outputs (LRT). Given the problems raised earlier about the convergence of the LRT test statistic to its limiting chi-square distribution, we have complemented this LRT test with a resampling procedure, where the null hypothesis of no LD is used to generate a given number of two-locus samples (often 10,000) to which the LRT test is applied. The procedure provides an empirical distribution for the LRT statistic under the hypothesis of no LD, and the reported *P*-value is the position (quantile) of the observed LRT in this empirical distribution. The test statistic is left bounded; therefore, the null hypothesis is only invalidated by extreme values on the right tail of the empirical distribution.

An additional statistic is calculated, not to be used for a global test of LD, but rather to identify individual haplotypes whose frequencies deviate from their expectations more than by chance. (Chance, here, means random sampling, and the deviations are expected to be normally distributed.) This is done by considering the standardized residuals proposed by Agresti[33] for a chi-square test. Residuals of this kind are considered to be more independent from the observed or expected frequencies than other residuals, and unlike other measures of LD such as D and D′ (see Hedrick[10]), they allow for direct comparisons of deviations, even for haplotypes with very different observed or expected frequencies.

Discussion and Conclusion

To our knowledge, currently, the GENE[RATE] tools are the only suite of computer programs that are able to work with highly polymorphic and ambiguous nominal data. These tools are an extension of the classic methods of gene-counting and haplotype estimation published in 1995 (references are mentioned in "Frequency estimation" section) and are shown to be particular instances of the EM algorithm. The framework of nominal alleles is rather distinct from that of sequence data. Nominal alleles refer to longer or shorter chromosomal regions that can span tens to thousands of nucleotides and are just considered as equal or distinct, without taking into account the molecular information. This is often because such information is not simply available given the typing techniques used to produce the data. Such data are clearly not as rich as sequence data, but they present interesting characteristics that make them useful for population genetics analyses. The high polymorphism of the data allows for a high discrimination of populations. The eight loci of the system are considered as segregating independently; they span over a contiguous region of 6 million nucleotides. Finally and most importantly, data for HLA are very abundant, given the clinical relevance of the system. (A detailed discussion of HLA relevance for population genetics has been given by Sanchez-Mazas et al.[13])

As we have stated, not all data ambiguities are of the same nature, and they do not have the same relevance, for nominal and sequence data. It is interesting to see that the approach taken for sequence data is either to make a call or discard the ambiguous position,[34–37] while the approach presented here includes the ambiguities in the calculations. In practice, instead of using two alleles for an individual, what we consider is two probability distributions for each individual (that reduce to a single allele for nonambiguous data).

The spirit of Li's samtools[37] and that of GENE[RATE] are similar, but a number of differences in the nature of the data makes a direct comparison impossible. We do plan, however, to perform such a comparative study using a triple approach: the alleles defined at nominal levels (as usual with HLA data), their translation as sequence data, and the raw reads produced by an NGS typing technique.

The UNIFORMAT grammar and the GENE[RATE] tools described in this article are specially adapted to nominal diploid data with ambiguities. Although primarily developed to solve problems arising with the analysis of the highly polymorphic genetic system HLA, the human MHC, the suite is completely general and applicable to diploid data. It may even accommodate pedigree information by directly specifying known, possibly partial, haplotypes in the UNIFORMAT data file. Actually, these tools have already been used in mixed analyses of HLA, classical and nonclassical genes, and other immunogenetic systems such as KIR and MICA.[38–41]

The web interface is currently the easiest way to use this suite because of the large number of dependencies on external libraries and other software, but the code source will eventually be packaged, including dependency information, in the Debian format[42] or as an R package, and made

publicly available. The web interface is continuously maintained and updated.

Questions, comments, error reports, and suggestions are welcome and can be addressed to the author.

Acknowledgments

This work was supported by Swiss FNS grants #31003A_127465 and #31003A_144180 and by COST Action BM0803. The author thanks Stéphane Buhler and Alicia Sanchez-Mazas for reading a first version of this text and providing valuable comments and suggestions and, more importantly, for their continuous support and advice with all the many details that are an essential part of GENE[RATE].

Author Contributions

JMN is the sole author, responsible for all of the content of this paper. The author has reviewed and approved of the final manuscript.

REFERENCES

1. Sanchez-Mazas A, Fernandez-Viña M, Middleton D, et al. Immunogenetics as a tool in anthropological studies. *Immunology*. 2011;133(2):143–64.
2. Nunes JM, Riccio ME, Buhler S, et al. Analysis of the HLA population data (AHPD) submitted to the 15th International Histocompatibility/Immunogenetics Workshop by using the Gene[rate] computer tools accommodating ambiguous data (AHPD project report). *Tissue Antigens*. 2010;76(1):18–30.
3. Di D, Sanchez-Mazas A. Challenging views on the peopling history of East Asia: the story according to HLA markers. *Am J Phys Anthropol*. 2011;145(1):81–96.
4. Currat M, Poloni ES, Sanchez-Mazas A. Human genetic differentiation across the Strait of Gibraltar. *BMC Evol Biol*. 2010;10:237.
5. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol*. 2005;15(11):1022–7.
6. Sanchez-Mazas A, Lemaître JF, Currat M. Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. *Philos Trans R Soc Lond B Biol Sci*. 2012;367(1590):830–9.
7. Mack SJ, Bugawan TL, Moonsamy PV, et al. Evolution of Pacific/Asian populations inferred from HLA class II allele frequency distributions. *Tissue Antigens*. 2000;55(5):383–400.
8. Cao K, Moormann AM, Lyke KE, et al. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *Tissue Antigens*. 2004;63(4):293–25.
9. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. *Genetics*. 2006;173(4):2121–42.
10. Hedrick P. *Genetics of Populations*. 4th ed. Sudbury, MA: Jones and Bartlett Learning; 2011.
11. Mack S, Tsai Y, Sanchez-Mazas A, et al. 13th International Histocompatibility Workshop Anthropology/Human Genetic Diversity Joint Report. Chapter 3: Anthropology/human genetic diversity population reports. In: Hansen J, ed. Immunobiology of the Human MEG: Proceedings of the 13th International Histocompatibility Workshop and Conference. Vol 1. Seattle: IHWG Press; 2007:580–652.
12. Nunes JM, Buhler S, Sanchez-Mazas A. NO to obsolete definitions: YES to blanks. *Tissue Antigens*. 2014;83(2):119–20.
13. Sanchez-Mazas A, Buhler S, Nunes JM. A new HLA map of Europe: regional genetic variation and its implication for peopling history, disease-association studies and tissue transplantation. *Hum Hered*. 2013;76(3):162–77.
14. Buhler S, Nunes JM, Nicoloso G, Tiercy JM, Sanchez-Mazas A. The heterogeneous HLA genetic makeup of the Swiss population. *PLoS One*. 2012;7(7):e41400.
15. Riccio ME, Buhler S, Nunes JM, et al. 16(th) IHIW: analysis of HLA population data, with updated results for 1996 to 2012 workshop data (AHPD project report). *Int J Immunogenet*. 2013;40(1):21–30.
16. Nunes JM, Buhler S, Roessli D, Sanchez-Mazas A; HLA-net 2013 Collaboration. The HLA-net GENE[RATE] pipeline for effective HLA data analysis and its application to 145 population samples from Europe and neighbouring areas. *Tissue Antigens*. 2014;83(5):307–23.
17. Nunes JM. Tools for analysing ambiguous HLA data. *Tissue Antigens*. 2007;69(suppl 1):203–5.
18. Levine JR, Mason T, Brown D. *Lex & Yacc*. Sebastopol, CA: O'Reilly Media, Inc.; 1992.
19. Marsh SGE, Albert ED, Bodmer WF, et al. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*. 2010;75(4):291–455.
20. Sanchez-Mazas A, Vidan-Jeras B, Nunes JM, et al. Strategies to work with HLA data in human populations for histocompatibility, clinical transplantation, epidemiology and population genetics: HLA-NET methodological recommendations. *Int J Immunogenet*. 2012;39(6):459–72,459–72.
21. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J R Stat Soc Series B Stat Methodol*. 1977;39:1–38.
22. Ceppellini R, Siniscalco M, Smith CA. The estimation of gene frequencies in a random-mating population. *Ann Hum Genet*. 1955;20(2):97–115.
23. Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*. 1995;56(3):799–810.
24. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*. 1995;12(5):921–7.
25. Hawley ME, Kidd KK. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered*. 1995;86(5):409–11.
26. Nunes JM. *Counting Genes* [thesis]. Porto: University of Porto; 2005.
27. McLachlan GJ, Krishnan T. *The EM Algorithm, and Extensions*. Wiley Series in Probability and Statistics. New York, NY: John Wiley & Sons, Inc.; 1997.
28. Nunes JM, Riccio ME, Tiercy JM, Sanchez-Mazas A. Allele frequency estimation from ambiguous data: using resampling schema in validating frequency estimates and in selective neutrality testing. *Human Biol*. 2011;83(3):437–47.
29. Slatkin M. An exact test for neutrality based on the Ewens sampling distribution. *Genet Res*. 1994;64(1):71–4.
30. Slatkin M. A correction to the exact test based on the Ewens sampling distribution. *Genet Res*. 1996;68(3):259–60.
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289–300.
32. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*. 1992;48(2):361–72.
33. Agresti A. *An Introduction to Categorical Data Analysis*. 2nd ed. New York, NY: John Wiley & Sons; 2007.
34. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. *Mol Ecol*. 2013;22(11):3124–40.
35. Lynch M, Xu S, Maruki T, Jiang X, Pfaffelhuber P, Haubold B. Genome-wide linkage-disequilibrium profiles from single individuals. *Genetics*. 2014;198(1):269–81.
36. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a Map-Reduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303.
37. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–93.
38. Wenda S, Faé I, Sanchez-Mazas A, Nunes JM, Mayr WR, Fischer GF. The distribution of MICA alleles in an Austrian population: evidence for increasing polymorphism. *Hum Immunol*. 2013;74(10):1295–9.
39. Di Cristofaro J, Julie DC, Buhler S, et al. Linkage disequilibrium between HLA-G*0104 and HLA-E*0103 alleles in Tswa Pygmies. *Tissue Antigens*. 2011;77(3):193–200.
40. Carlini F, Traore K, Cherouat N, et al. HLA-G UTR haplotype conservation in the Malian population: association with soluble HLA-G. *PLoS One*. 2013;8(12):e82517.
41. Buhler S, Di Cristofaro J, Frassati C, et al. High levels of molecular polymorphism at the KIR2DL4 locus in French and Congolese populations: impact for anthropology and clinical studies. *Hum Immunol*. 2009;70(11):953–9.
42. *The Debian GNU/Linux FAQ – Basics of the Debian Package Management System*. Available at: http://www.debian.org/doc/manuals/debian-faq/ch-pkg_basics.

Appendix

Likelihood equations used in the generalized counting methods. The properties of the gene-counting method for frequency estimation have long been known, initially presented as a result of work essentially led by C.A.B. Smith and later framed in the more general expectation–maximization (EM) algorithm of Dempster et al. The general properties of convergence and uniqueness of a solution are verified for some families of distributions, one of which is the general exponential family. Showing that a model used to estimate frequencies leads to likelihood equations that can be seen as resulting from observations of a member of the general exponential family of distributions is then all that is needed to guarantee these properties to such a model of estimation.

We start by considering a generalized description of a sample by means of its genotype descriptions. A uniform representation is given by

$$1 = \sum_{j \geq i} \alpha_{ij} \delta_{ij} p_i p_j \tag{1}$$

where p_i stands for allelic frequencies, δ_{ij} is the zygote indicator, and α_{ij} is the association coefficient for that pair of alleles. The association coefficients are 1 if Hardy–Weinberg holds and would be some constant value (1–F) for an inbreeding-like model. This form provides a unified expression, the free model, whose domain is constrained by

$$\begin{cases} p_i \in [0,1) \\ \alpha_{ij} \in \left[0, \dfrac{1}{\max\{p_i, p_j\}} \right] \\ 1 = \sum_{j \geq i} \alpha_{ij} \delta_{ij} p_i p_j \\ \forall i, j \in \{1, \ldots, k\} : j \geq i \end{cases}$$

The calculation of the degrees of freedom associated with such a model is not mentioned here, but it is easy to see that the constraints on this model are the total sum of allele frequencies and the sums of the frequencies of all the allele pairs carrying a given allele, that is, $\frac{k(k-1)}{2}$, assuming, as until now, that k alleles exist for the locus.

The support, or log likelihood, for such a model is given by

$$\ln L = \sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(\alpha_{ij} \delta_{ij} p_i p_j\right) \tag{2}$$

whose rearrangement gives

$$\ln L = \sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(\alpha_{ij}\right) + \sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(\delta_{ij}\right) + \sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(p_i\right) + \sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(p_j\right)$$

The last terms can be rewritten as

$$\sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(p_i\right) = \sum_{i=1}^{k} \left(\sum_{i=1}^{k} n_{ij} \ln\left(p_i\right) \right) = \sum_{i=1}^{k} \left(\sum_{i=1}^{k} n_{ij} \right) \ln\left(p_i\right)$$

and

$$\sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(p_j\right) = \sum_{i=1}^{k} \left(\sum_{j=1}^{k} n_{ij} \ln\left(p_j\right) \right) = \sum_{j=1}^{k} \left(\sum_{j=1}^{k} n_{ij} \right) \ln\left(p_j\right)$$

Rearranging the indices, this can be further rewritten as

$$\sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(p_j\right) = \sum_{i=1}^{k} \left(\sum_{j=1}^{k} n_{ij} \right) \ln\left(p_i\right)$$

then

$$\sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(p_i\right) + \sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(p_j\right) = \sum_{i=1}^{k} \left(\sum_{j=i}^{k} n_{ij} + \sum_{j=1}^{i} n_{ji} \right) \ln\left(p_i\right)$$

Noting that we restricted our expressions to $j \geq i$ and should therefore rewrite n_{ji} as n_{ij}, we observe that

$$\sum_{j=i}^{k} n_{ij} + \sum_{j=1}^{i} n_{ji} = n_{ii} + \sum_{j=1}^{k} n_{ij} = \sum_{j=1}^{k} \Delta_{ij} n_{ij}$$

and finally rewrite the support Equation (2) as

$$\ln L = \sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(\alpha_{ij}\right) + \sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(\delta_{ij}\right) + \sum_{i=1}^{k} \left(\sum_{j=1}^{k} \Delta_{ij} n_{ij} \right) \ln\left(p_i\right) \tag{3}$$

We proceed now to the identification of the component functions of a member of the exponential family. Their common expression is

$$f(x, \Theta) = a(\Theta) b(x) \exp\left[c^T(\Theta) d(x)\right]$$

We note immediately that

$$\ln b(x) = \sum_{\substack{i=j=1 \\ j \geq i}}^{k} n_{ij} \ln\left(\delta_{ij}\right)$$

and also that

$$a(\Theta) = 1$$

because all the other terms involve parameters of interest (frequencies and association coefficients).

Therefore, we have to show that

$$c^T(\Theta)d(x) = \sum_{\substack{i=j=1 \\ j\geq i}}^{k} n_{ij} \times \ln\left(\alpha_{ij}\right) + \sum_{i=1}^{k}\left(\sum_{j=1}^{k}\Delta_{ij}n_{ij}\right) \times \ln\left(p_i\right)$$

but the right member is just the dot product of two vectors of $k(k+1)/2 + k$ components.

They are

$$d(x) = \left(n_{1,1},\ldots n_{ij},\ldots n_{kk}, \sum_{j=1}^{k}\Delta_{1j}n_{1j},\ldots, \sum_{j=1}^{k}\Delta_{ij}n_{ij},\ldots, \sum_{j=1}^{k}\Delta_{jk}n_{jk}\right)^T$$

and

$$c(\Theta) = \left(\ln\alpha_{1,1},\ldots,\ln\alpha_{ij},\ldots\ln\alpha_{kk},\ln p_1,\ldots,\ln p_i,\ldots,\ln p_k\right)^T$$

These two functions are indeed readily identifiable with the vector of the parameters – or, more precisely, logarithms of the parameters – and a sufficient statistic. Consequently, we have shown that the complete data log likelihood for the free model is a member of the exponential family of distributions.

Calculations made using this model are thus guaranteed to have the same properties as those of the general EM gene-counting method. In the GENE[RATE] framework, this free model is materialized as an HWE model by setting the association coefficients to 1, and as an inbreeding-like model by setting all the association coefficients equal (leading to an estimate of 1–F).