

HLA and HIV Infection Progression: Application of the Minimum Description Length Principle to Statistical Genetics

Peter T. Hraber^{*,†}, Bette T. Korber^{*,†}, Steven Wolinsky[‡],
Henry Erlich[§], Elizabeth Trachtenberg[¶], and Thomas B. Kepler^{*,||}

^{*}Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501

[†]Los Alamos National Laboratory, Los Alamos NM 87545

[‡]Feinberg School of Medicine, Northwestern University,
676 North St. Claire, Suite 200, Chicago IL 60611

[§]Roche Molecular Systems, 1145 Atlantic Avenue, Alameda CA 94501

[¶]Children's Hospital Oakland Research Institute,
5700 Martin Luther King Jr. Way, Oakland CA 94609

^{||}Department of Biostatistics and Bioinformatics, and
Center for Bioinformatics & Computational Biology,
Box 90090, Duke University, Durham NC 27708

Classification

Biological Science/Immunology & Physical Science/Applied Mathematics

Corresponding author

Peter T. Hraber

address: Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501

phone: +1 (505) 984-8800

fax: +1 (505) 982-0565

email: pth@santafe.edu

Manuscript information

Text pages: 14

Figures: 1

Tables: 2

Words in abstract: 245 (< 250)

Character count: 45536 (< 47000)

Nonstandard abbreviations

MACS: multicenter AIDS cohort study

MDL: minimum description length

Abstract

The minimum description length (MDL) principle was developed in the context of computational complexity and coding theory. It states that the best model to account for some data minimizes the sum of the lengths, in bits, of the descriptions of the model and the data as encoded via the model. The MDL principle gives a criterion for parameter selection, by using the description length as a test statistic. Class I HLA genes play a major role in the immune response to HIV, and are known to be associated with rates of progression to AIDS. However, these genes are highly polymorphic, making it difficult to associate alleles with disease outcome, given statistical issues of multiple testing. Application of the MDL principle to immunogenetic data from a longitudinal cohort study (Chicago MACS) enables classification of alleles associated with plasma HIV RNA abundance, an indicator of infection progression. We recently reported that MDL analysis of the relationship of HLA supertypes (a classification of alleles by epitope-binding anchor motifs) with HIV RNA levels identifies associations between human genotype and viral RNA. Details of the MDL approach and more extended analyses of HLA and viral RNA are described here. Variation in progression is strongly associated with HLA-B. Allele associations with viral levels support and extend previous studies. In particular, individuals without *B58s* supertype alleles average viral RNA levels 3.6-fold greater than individuals with them. Mechanisms for these associations include variation in epitope specificity and selection that favors rare alleles.

Progression of HIV infection is characterized by three phases: acute, or early, chronic, and AIDS, the final phase of infection preceding death [1]. The chronic phase is variable in duration, lasting ten years on average, but varying from two to twenty years. A good predictor of the duration of the chronic phase is the viral RNA level during chronic infection, with higher levels consistently associated with more rapid progression than lower levels [2]. A major challenge for treating HIV and developing effective vaccination strategies is to understand what contributes to variation in plasma viral RNA levels, and hence to infection progression.

The cell-mediated immune response identifies and eliminates infected cells from an individual. A central role in this response is played by the major histocompatibility complex (MHC), in humans, also known as human leukocyte antigens (HLA). Two classes of HLA genes code for co-dominately expressed cell-surface glycoproteins, and present processed peptide to circulating T-cells, which discriminate between self and non-self [3, 4].

Class I HLA molecules are expressed on all nucleated cells except germ cells. In infected cells, they bind and present antigenic peptide fragments to T-cell receptors on $CD8^+$ T-lymphocytes, which are usually cytotoxic and cause lysis of the infected cell. Class II HLA molecules are expressed on immunogenetically reactive cells, such as dendritic cells, B-cells, macrophages, and activated T-cells. They present antigen peptide fragments to T-cell receptors on $CD4^+$ T-lymphocytes and the interaction results in release of cytokines that stimulate the immune response.

Human HLA loci are among the most diverse known [5, 6]. This diversity provides a repertoire to recognize evolving antigens [6, 7]. Previous studies of associations between HLA alleles and variation in progression of HIV-1 infection have established that within-host HLA diversity helps to inhibit viral infection, by associating degrees of heterozygosity with rates of HIV disease progression [8]. Thus, homozygous individuals, particularly at the HLA-B locus, suffer a greater rate of progression than do heterozygotes [8, 9]. Identifying which alleles are associated with variation in rates of infection progression has been difficult, due in part to the compounding of error rates incurred when testing many alternative hypotheses, and published results do not always agree [10, 11].

This study demonstrates the use of an information-based criterion for statistical inference. Its approach to multiple testing differs from that of standard analytic techniques, and provides the ability to resolve associations between variation in HIV RNA abundance and variation in HLA alleles.

As an application of computational complexity and optimal coding theory to statistical inference, the minimum description length (MDL) principle states that the best statistical model, or hypothesis, to account for some observed data is the model that minimizes the sum of the number of bits required to describe both the model and the data encoded via the model [12, 13, 14]. It is a model-selection criterion that balances the need for parsimony and fidelity, by penalizing equally for the information required to specify the model and the information required to encode the residual error.

The analyses detailed below apply the MDL principle to the problem of

partitioning individuals into groups having similar HIV RNA levels, based on HLA alleles present in each case.

Chicago MACS HLA & HIV Data

The Chicago Multicenter AIDS Cohort Study (MACS) provided an opportunity to analyze a detailed, long-term, longitudinal set of clinical HIV/HLA data [10]. Each participant provided informed consent in writing. Of 564 HIV-positive cases sampled in the Chicago MACS, 479 provided information about both the rate of disease progression and HLA genetic background. Progression was indicated by the quasi-stationary “set-point” viral RNA level during chronic infection. Immunogenetic background was obtained by determining which HLA alleles from class I (HLA-A, -B, and -C) and class II (HLA-DRB1, -DQB1, and -DPB1) loci were present in each individual.

Viral RNA set-point levels were determined after acute infection and prior to any therapeutic intervention or the onset of AIDS, as defined by the presence of an opportunistic infection or CD4⁺ T-cell count below 200 per ml of plasma. Because the assay has a detection threshold of 300 copies of virus per ml [10], maximum-likelihood estimators were adjusted to avoid biased estimates of population parameters from a truncated, or censored, sample distribution [15]. Viral RNA levels were log-transformed so as better to approximate a normal distribution.

High-resolution class I and II HLA genotyping [10] provided four-digit allele designations, though analyses were generally performed using two-digit allele designations because of the resulting reduction of allelic diversity and increased number of samples per allele. Because of the potential for results to be confounded by an effect associated with an individual’s ethnicity or revised sampling protocol, two separate analyses were performed, one using data from the entire cohort, and another using only data from Caucasian individuals. Sample numbers were too small to study other subgroups independently.

HLA supertypes group class I alleles by their peptide-binding anchor motifs [16]. Assignment of four-digit allele designations to functionally related groups of supertypes at HLA-A and -B loci facilitated further analysis. Where they could be determined, HLA-A and HLA-B supertypes were assigned from four-digit allele designations [10]. As with two-digit allele designations for each locus, HLA-A and -B supertypes were assessed for association with viral RNA levels. Cases having other alleles were withheld from classification and subsequent analysis of supertypes.

A description length analysis determined whether HIV RNA levels were non-trivially associated with alleles at any HLA locus.

Description Lengths

The challenge of data classification is to find the best partition, such that observations within a group are well-described as independent draws from a single population, but differences in population distributions exist between groups.

Whether the data are better represented as two groups, or more, than as one depends on the description lengths that result.

We use the family of Gaussian distributions to model viral RNA levels. While the MDL strategy can be applied using any probabilistic model, a log-normal distribution is a good choice for the observed plasma viral RNA values. First, the description length of the model and of the data given the model is calculated as described below, grouping all of the observations into one normal distribution, L_1 . Next, the data are broken into two partitions, L_2 , and the log-RNA values associated with HLA alleles are partitioned to minimize the description length given the constraint that two Gaussian distributions, each having their own mean and variance, are used to model the data.

For fixed $n \times n$ covariance matrix Σ , the description length is $L_\Sigma = \frac{1}{2} \log |\Sigma| + \frac{1}{2} Y' \Sigma^{-1} Y + C$, where Y is the n -component vector of observations and C is the quantity of information required to specify the partition. Logarithms are computed in base two, with fractional values rounded upwards, so that the resulting units are bits. The description length of interest results from integrating L over all covariance matrices with the appropriate structure. In practice, we use Laplace's approximation for the integral [12, 17] which gives, asymptotically, $L = \frac{1}{2} \log |\hat{\Sigma}| + \frac{1}{2} Y' \hat{\Sigma}^{-1} Y + \frac{k}{2} \log n + C$, where k is the number of free parameters in the covariance model, and $\hat{\Sigma}$ is the specific covariance matrix of the appropriate structure that minimizes L_Σ . A more detailed account appears in the Appendix.

The analog of a null hypothesis is the assumption that one group of alleles is sufficient to account for the variation in viral RNA. The description length for one group is: $L_1 = \frac{1}{2} (n + (n - 1) \log s^2 + \log n \bar{x}^2 + 2 \log n)$, where n is the total number of observations, s^2 is the maximum-likelihood estimate of the population variance and \bar{x} is the sample mean, computed as the Winsorized mean [15] because of truncation below the sensitivity limit of the RNA assay.

It follows that the description length for two groups can be computed as:

$$L_2 = \frac{1}{2} \sum_{i=1}^2 (n_i + (n_i - 1) \log s_i^2 + \log n_i \bar{x}_i^2 + 2 \log n_i) + C,$$

where C is an adjustment for performing multiple comparisons. Because additional information is required to specify the optimum partition, the description length is increased by a quantity related to the number of partitions evaluated, such that $C = N \log k$ bits, where N is the number of alleles observed at the partitioned locus. For $k = 2$, $C = N$.

Further partitions of alleles into more than two groups might yield a shorter description length, computed as a summation over terms in the equation for L_2 for each of the k distinct groups.

The shortest description length for any value of k indicates the best choice of model parameters, including the number of parameters, and hence, the optimum partition of N alleles into k groups. We denote this as L^* .

Algorithm

The minimum description length is found by iteratively computing the description length for each possible partition of alleles into groups and taking the minimum as optimal. Iteration consists first of determining the number of alleles, N , at a particular locus, and then incrementing through each of the $k^{(N-1)}$ possible partitions of alleles into k groups, computing the associated description length, and reporting the best results. Each iteration evaluates one possible mapping of alleles to groups. Searching through all possible partitions using the description length as an optimality criterion ensures selection of the best partition as a result of the search.

In this mapping, the ordering of groups is informative, because the ordering gives the relative dominance of alleles for diploid loci. An individual having an allele assigned to the first-order group is assigned to that group. Otherwise, the individual is assigned to the next appropriate group. Two individuals sharing one allele might be placed in either the same group or different groups, depending on the mapping of alleles to groups in a particular iterate. For example, consider how one might group two individuals, one with alleles $A1$ and $A2$ at some locus, and another with alleles $A2$ and $A3$. Whether or not they are grouped together depends on the assignment of alleles to groups, and can be done several different ways. The algorithm enumerates each possible assignment of alleles to groups.

The extent of the search scales as k^N . In practice, the most diverse locus was HLA-B, with 30 alleles when analyzed using two-digit allele designations. For two groups, this gives $2^{30} \approx 10^8$ possible partitions. Serial iteration on an UltraSPARC-IIi 440MHz CPU (Sun Microsystems) requires roughly 36 hours for completion. A parallel implementation requires no message passing, so computing time scales inversely with an increasing number of CPUs, or doubling available processors halves the time for iteration. With many CPUs, the search space of 2^{30} partitions can be exhaustively evaluated in an hour or less. Unfortunately, exhaustively evaluating all three-way partitions is prohibitive, as $3^{30} \approx 2 \times 10^{14}$, over a million-fold increase in computational effort! Supertype classification reduced the diversity of possible partitions and enabled partitioning of the data into more than two groups.

The algorithm was implemented in C and will be distributed on request.

Class I & II HLA Results

The description length for the entire cohort as one group is $L_1 = 934$ bits; for the Caucasian subsample, it is $L_1 = 721$ bits. In general, $L_1 < L_2$ at most loci (Table 1), so the MDL criterion does not support partitioning alleles into groups that are predictive of high or low RNA levels, except at HLA-B, where $L_2 < L_1$. In the subsample, partitioning HLA-C or HLA-DQB1 alleles can also provide preferred two-way splits, though not as well as HLA-B. Further partitioning was intractable because of great allelic diversity, as previously mentioned. Partitions of HLA-B alleles provide the best groupings among all loci. Because $L_2^* < L_1$,

two groups, partitioned by HLA-B alleles, provide a better description than one (Fig. 1a and 1b).

What is the composition of the optimum groupings? For the entire cohort, the following alleles were associated with low viral RNA levels: B^*13 , B^*27 , B^*38 , B^*45 , B^*49 , B^*57 , B^*58 , and B^*81 . The remaining alleles, associated with greater viral RNA than the first group, are: B^*07 , B^*08 , B^*14 , B^*15 , B^*18 , B^*35 , B^*37 , B^*39 , B^*40 , B^*41 , B^*42 , B^*44 , B^*47 , B^*48 , B^*50 , B^*51 , B^*52 , B^*53 , B^*55 , B^*56 , B^*67 , and B^*82 . As described earlier, having any alleles associated with the first group is sufficient for an individual to be assigned to the group having lower viral RNA.

How robust are these assignments of alleles to groups? Four alternative groupings provide description lengths within one bit of the optimum. They do not dramatically rearrange the assignment of individuals to groups, but do provide insight as to which alleles are assigned to either group with less confidence. Among near-optimal partitions, alleles B^*82 and B^*67 were assigned to groups other than in the optimum partition.

In the Caucasian subsample, alleles B^*13 , B^*27 , B^*40 , B^*45 , B^*48 , B^*49 , B^*57 , and B^*58 are associated with lower viral RNA, and the remaining alleles, B^*07 , B^*08 , B^*14 , B^*15 , B^*18 , B^*35 , B^*37 , B^*38 , B^*39 , B^*41 , B^*44 , B^*47 , B^*50 , B^*51 , B^*52 , B^*53 , B^*55 , and B^*56 , or lack of any alleles from the first group, are associated with greater viral RNA levels. Two nearly optimal partitions assigned alleles B^*47 and B^*48 to the second group. Fig. 1 illustrates the distributions of viral RNA levels from this subsample, as one group (Fig. 1c) and as the best partition at HLA-B (Fig. 1d).

To summarize the most robust inferences from the analyses of two-digit allele designations, individuals having HLA-B alleles B^*13 , B^*27 , B^*45 , B^*49 , B^*57 , or B^*58 were associated with lower viral RNA levels than their counterparts lacking these alleles.

Comparison of groupings obtained via the MDL approach with more traditional means for statistical inference, a two-tailed, two-sample, Welch modified t-test, which does not assume equal variances, and its non-parametric variant, the Wilcoxon rank-sum test [18], was very favorable. In each case, the null hypothesis was that of no difference between the group mean log-transformed viral RNA levels, and the alternative hypothesis was that the means differ. Both tests agreed in rejecting the null hypothesis in favor of the alternative ($P < 10^{-10}$).

HLA Supertype Results

Assigning the diploid, co-dominantly expressed HLA-A alleles to four HLA-A superotypes [16], $A1s$, $A2s$, $A3s$, and $A24s$, was possible for 399 individuals. The mapping of HLA-B alleles to five superotypes, $B7s$, $B27s$, $B44s$, $B58s$, and $B62s$, was made for 352 individuals. The resulting decrease in allelic diversity enabled analysis for $k > 2$.

Description lengths of the best k -way partitions of supertype alleles for HLA-A superotypes are: $L_1 = 793$, $L_2 = 782$, $L_3 = 789$, and $L_4 = 794$ bits. The best description length results from a two-way split, though a three-way split also

yields a shorter description length than that obtained from one group. The best partition of HLA-A supertypes assigned individuals having *A1s* alleles to the low RNA group.

For HLA-B supertypes, $L_1 = 704$, $L_2 = 691$, $L_3 = 693$, and $L_4 = 697$ bits (Fig. 1e). The best model results when $k = 2$. Overall, individuals lacking *B58s* alleles averaged viral RNA levels 3.6-times greater than individuals having *B58s* supertype alleles (Fig. 1f). Thus, individuals with *B58s* alleles have significantly lower viral RNA levels than individuals without them.

Table 2 summarizes results of assigning HLA-B associations to high or low viral-RNA categories as two-digit allele designations from both the entire cohort and the Caucasian subsample, and as supertypes for those individuals having two alleles that could be assigned to a supertype. Alleles not found in a sample are indicated by a dash. The *B*15* alleles are not shown because their high-resolution genotype designations correspond to four different supertypes.

Overall, the most consistent associations with low viral RNA are among the *B58s*, and with high viral RNA, the *B7s*. Inconsistencies in assignment to a category occur for the *B*13*, *B*27*, *B*45*, and *B*49* alleles, which are in the low viral-RNA group when analyzed as such, but the high viral-RNA group when assigned to supertypes.

When compared with alternative inferential techniques, the difference between group viral RNA levels was highly significant. This and agreement with alleles reported to be associated with variation in viral RNA levels in previously published studies indicate that using the description length as a test statistic can provide reliable inferences.

MDL & Statistical Inference

The traditional statistical solution is to pose a question as follows: suppose that the simpler model (e.g., one homogeneous population) were actually true; call this the null hypothesis. How often would one, in similar experiments, get data that look as different from that expected under the null hypothesis as in the actual experiment?

This technique has limitations when the partition that represents the alternative hypothesis is not given in advance. There are then many potential alternative partitions and the appropriate distribution under the null hypothesis for this ensemble of tests is very difficult to estimate. Furthermore, for proper interpretation, the outcome relies upon the truth of the initial assumption: that the data are distributed as dictated by the null hypothesis.

An alternative is to choose that model that represents the data most efficiently. Here, efficiency is the amount of information, quantified as bits, required to transmit electronically both the model and the data as encoded by the model. This criterion may not seem intuitively clear on first exposure. However, it follows naturally from a profound relationship between probability and coding theory that was discovered, explored, and elaborated by Solomonoff, Kolmogorov, Chaitin, and Rissanen [19, 20, 21, 22, 23].

The idea is quite simple and elegant. It can be illustrated by analogy to the problem of designing an optimal code for the efficient transmission of natural-language messages. Consider the international Morse code. Recall that Morse code assigns letters of the Roman alphabet to codewords comprised of dots (“.”) and dashes (“-”). The codewords do not all have the same number of dots and/or dashes; it is a variable-length code.

Efficient, compact encodings result from the design of a codebook such that the shortest codewords are assigned to the most frequently encoded letters and long codewords are assigned to rare letters. Thus, *e* and *t* are encoded as “.” and “-”, respectively, while *q* and *j* are encoded as “- - . -” and “. - - -”. The theory of optimal coding provides an exact relationship between frequency and code length and thus, probability and description length.

The key departure of MDL from Morse-codelike schemes is that, while Morse code would generally be good for sending messages over an average of many texts, specific texts might be encoded even more efficiently, by encoding not only letters, but letter combinations, common words, or even phrases, perhaps as abbreviations or acronyms. However, if one is to recode for particular texts, one must first transmit the coding scheme. So perhaps one might use Morse code to transmit the details of the new coding scheme and then transmit the text itself with the new scheme. Whether this might yield greater efficiency depends not only on how much compression is achieved in the new encoding, but also on how much overhead is incurred in having to transmit the coding scheme.

The analogy to scientific data analysis is clear. A statistical model is an encoding scheme that encapsulates the regularities in the data to yield a concise representation thereof. The best model effectively compresses regularities in the data, but is not so elaborate that its own description demands a great deal of information to be encoded. The MDL principle provides a model-selection criterion that balances the need for a model that is both appropriate and parsimonious, by penalizing with equal weights the information required to specify the model and the unexplained, or residual error.

Yet another contribution the MDL principle brings to statistical modelling is that the penalty for multiple comparisons is less restrictive than the penalty of compounded error rates incurred with canonical inferential approaches. In order to maintain a desired experiment-wide error rate, the standard adjustment is to make the per-comparison error rate considerably more stringent. With current technology, realistic sample sizes for such studies will generally be less than a thousand and stringent significance levels will be difficult to surpass. Unfortunately, fixing the false-positive error rate does not address the false-negative probability, which may leave researchers powerless to detect effects among many competing hypotheses with limited samples.

Mechanisms

Of HLA supertype alleles, individuals with *B58s* have lower viral RNA levels than those who lack them, even among homozygotic individuals. Naturally,

this leads one to consider mechanisms that underlie patterns found in the data. Elsewhere, we consider two hypotheses to explain the observed associations between HLA alleles and variation in viral RNA [10].

There may be allele-specific variation in antigen-binding specificity. Some alleles may have greater affinity than others for HIV-specific peptide fragments due to the peptide-binding anchor motifs they present. We were not able to identify any clear association between the frequency of anchor motifs among HIV-1 proteins and viral RNA levels in the Chicago MACS [10], though others have suggested that such a relationship might exist [24].

It may also be the case that frequency-dependent selection has favored rare alleles. Frequent alleles provide the evolving pathogen greater opportunity to explore mutant phenotypes that may escape detection by the host’s immune response. By encountering rare alleles less frequently, the virus has not had the same opportunity to explore mutations that evade the host’s defense response. This hypothesis is corroborated by a significant association between viral RNA and HLA allele frequency in the Chicago MACS sample [10].

Because their predictions differ, these hypotheses could be tested with data from another cohort, where a different viral subtype predominates. That is, if other alleles were associated with low viral RNA than those identified in this study, and an association between rare alleles and low viral RNA levels were observed there, then the second hypothesis would be more viable than the first. Alternatively, if a clear association between antigen peptide-binding anchor motifs and variation in viral RNA levels were found, the first hypothesis would be more viable. Other mechanisms are also possible, and hypotheses by which to evaluate them merit consideration.

Acknowledgments

We thank Bob Funkhouser, Cristina Sollars, and Elizabeth Hayes for sharing their expertise, and researchers of the Santa Fe Institute for insight and inspiration. This research was financed by funds from the Elizabeth Glazer Pediatric AIDS Foundation, the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, National Institutes of Health, National Science Foundation award #0077503, and the U.S. Department of Energy. We have no conflicting interests.

Appendix

In Gaussian Process modeling [25], the population means are treated as random variables and integrated out of the likelihood. The model is then specified entirely by the structure of the covariance matrix Σ , which specifies how each pair of observations is correlated. The covariance is greater for two observations from the same partition than for two observations from different partitions. Any given partition is specified entirely by a corresponding covariance structure.

Partitioning with Gaussian Models. Denote the n observations as the vector Y and the covariance matrix with parameter vector θ by $\Sigma(\theta)$. Let

the number of components of θ (the number of free parameters in the covariance matrix) be k . Then the MDL for the given covariance structure is: $L = \frac{1}{2} \log |\Sigma(\hat{\theta})| + \frac{1}{2} Y' \Sigma(\hat{\theta})^{-1} Y + \frac{k}{2} \log n + C$, where C is the information required to specify the partition or, equivalently, the covariance structure, and $\hat{\theta}$ is the vector of covariance parameters evaluated at maximum likelihood.

One Gaussian Population. The covariance matrix has a component σ_m^2 for the covariance among observations, induced by their sharing an unspecified mean, and an error component σ_ε^2 : $\Sigma = \sigma_\varepsilon^2 I + \sigma_m^2 \mathbf{1}\mathbf{1}'$, with $\mathbf{1}$ the column vector of all ones, $\mathbf{1}\mathbf{1}'$ the matrix of all ones, and I the identity matrix. The inverse is:

$$\Sigma^{-1} = \frac{1}{\sigma_\varepsilon^2} \left(I - \frac{\sigma_m^2}{\sigma_\varepsilon^2 + n\sigma_m^2} \mathbf{1}\mathbf{1}' \right),$$

and the log-determinant: $\log |\Sigma| = (n-1) \log \sigma_\varepsilon^2 + \log(\sigma_\varepsilon^2 + n\sigma_m^2)$.

This gives $L = \frac{1}{2} (n + (n-1) \log \sigma_\varepsilon^2 + \log(\sigma_\varepsilon^2 + n\sigma_m^2) + 2 \log n)$.

We find the maximum likelihood values of the parameters by minimizing over the description lengths. There are two cases.

Case 1: $n^2 \bar{Y}^2 - Y'Y \geq 0$. Here we have $\hat{\sigma}_\varepsilon^2 = (n-1)^{-1} (Y'Y - n\bar{Y}^2)$ and $\hat{\sigma}_m^2 = (n-1)^{-1} (n\bar{Y}^2 - \frac{1}{n} Y'Y)$, so $L = \frac{1}{2} (n + (n-1) \log \hat{\sigma}_\varepsilon^2 + \log n \bar{Y}^2 + 2 \log n)$.

Case 2: $n^2 \bar{Y}^2 - Y'Y < 0$. Here the common mean vanishes, giving $\hat{\sigma}_\varepsilon^2 = \frac{1}{n} Y'Y$, $\hat{\sigma}_m^2 = 0$, so $L = \frac{n}{2} (1 + \log \hat{\sigma}_\varepsilon^2 + \frac{2}{n} \log n)$.

Many Gaussian Populations. Two partitions give two populations. To analyze the HLA/HIV data, we treated these populations as independent. That is, we take the covariance between observations in separate partitions to be zero, and apply the fitting procedure outlined above separately to the two populations. An alternative is to take non-zero covariance between the two populations. This results in a more elaborate estimation procedure, unlikely to yield large efficiency gains because the two degrees of freedom (population means) are essentially mixed into one, with residual error.

The procedure examines each admissible partition and computes the MDL for that partition as the sum of individual description lengths over the two independent populations. The best partition yields the lowest description length over all partitions. This, plus the cost of specifying the partition, is compared with the MDL from the unpartitioned data. If the best partition provides a better representation of the data than the unpartitioned set ($L_k < L_{k-1}$), then the process is repeated in a recursive manner, independently within each of the partitioned populations.

References

- [1] McMichael, A. J. & Rowland-Jones, S. L. (2001) *Nature* **410**, 980-987.
- [2] Mellors, J. W., Rinaldo, C. R., Jr., Gupta, P., White, R. M., Todd, J. A. & Kingsley, L. A. (1996) *Science* **272**, 1167-1170.
- [3] Germain, R. N. (1999) Chapter 9 in *Fundamental Immunology*, fourth edition, ed. Paul, W. E. (Lippincott-Raven, Philadelphia PA), pp. 287-340.
- [4] Williams, A., Au Peh, C. & Elliott, T. (2002) *Tissue Antigens* **59**, 3-17.
- [5] Bodmer, W. F. (1972) *Nature* **237**, 139-145.
- [6] Little, A. M. & Parham, P. (1999) *Rev. Immunogenet.* **1**, 105-123.
- [7] Hill, A. V. S. (1998) *Ann. Rev. Immunol.* **16**, 593-617.
- [8] Roger, M. (1998) *FASEB J.* **12**, 625-632.
- [9] Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., Kaslow, R., Buchbinder, S., Hoots, K. & O'Brien, S. J. (1999) *Science* **283**, 1748-1752.
- [10] Trachtenberg, E. A., Korber, B. T., Sollars, C., Kepler, T. B., Hrabec, P. T., Hayes, E., Funkhouser, R., Fugate, M., Theiler, J., Hsu, M., Kunstman, K., Wu, S., Phair, J., Erlich, H. A. & Wolinsky, S. (2002) *Nat. Med.*, in review.
- [11] Trachtenberg, E. A. & Erlich, H. A. (2001) in *HIV Molecular Immunology 2001*, eds. Korber, B. T., Brander, C., Haynes, B. F., Koup, R., Kuiken, C., Moore, J. P., Walker, B. D. & Watkins, D. (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos NM), pp. I-43-60.
- [12] Rissanen, J. (1989) *Stochastic Complexity in Statistical Inquiry* (World Scientific, Singapore).
- [13] Li, M. & Vitányi, P. (1993) *An Introduction to Kolmogorov Complexity and its Applications* (Springer-Verlag, New York NY).
- [14] Hansen, M. H. & Yu, B. (2001) *J. Am. Stat. Assoc.* **96**, 746-774.
- [15] Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994) *Continuous Univariate Distributions*, volume 1, second edition (Wiley Interscience, New York NY).
- [16] Sette, A. & Sidney, J. (1999) *Immunogenetics* **50**, 201-212.
- [17] Lindley, D. V. (1980) in *Bayesian Statistics*, eds. Bernardo, J. M, DeGroot, M. H., Lindley, D. V. & Smith, A. F. M. (Valencia University Press, Valencia), pp. 223-237.

- [18] Venables, W. N. & Ripley, B. D. (1999) *Modern Applied Statistics with S-PLUS*, third edition (Springer, New York NY).
- [19] Kolmogorov, A. N. (1965) *Prob. Inform. Transmission* **1**, 4-7.
- [20] Chaitin, G. J. (1966) *J. Assoc. Comput. Mach.* **13**, 547-569.
- [21] Chaitin, G. J. (1987) *Algorithmic Information Theory* (Cambridge University Press, Cambridge UK).
- [22] Rissanen, J. (1986) *Ann. Statist.* **14**, 1080-1100.
- [23] Rissanen, J. (1999) *Comput. J.* **42**, 260-269.
- [24] Nelson, G. W., Kaslow, R. & Mann, D. L. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 9802-9807.
- [25] Williams, C. K. I. (1997) in *Mathematics of Neural Networks: Models, Algorithms and Applications*, eds. Ellacott, S. W., Mason, J. C. & Anderson, I. J. (Kluwer, Boston MA), pp. 378-382.

Figure Legends

Fig. 1. Description-length comparisons of viral RNA distributions as one (L_1) or two (L_2) groups. Ordinate units are the expected number of observations between two tick marks over the abscissa, or one doubling of viral RNA. Impulses along the abscissa show individual observations, with jitter added to enhance rendering of identical values. (a) Observations (n) from the Chicago MACS cohort lumped into one group, and (b) split into the best partition as two groups, with individuals having alleles B^*13 , B^*27 , B^*38 , B^*45 , B^*49 , B^*57 , B^*58 , or B^*81 assigned to the lower group (n_1), and remaining individuals assigned to the group with greater viral RNA (n_2). (c) Observations from the Caucasian subsample as one group, and (d) as the best split into two groups, where having alleles B^*13 , B^*27 , B^*40 , B^*45 , B^*48 , B^*49 , B^*57 , or B^*58 was the criterion for being assigned to the low viral-RNA group. Observations from individuals having two HLA-B supertype alleles, (e) in one group, and (f) partitioned into two groups, contingent on the presence of B^*58s .

Table 1: Optimum two-way partitions at each locus, with per-locus allelic diversity (N), description lengths without the information cost to specify model parameters ($L_2 - C$), and minimum description lengths (L_2).

	ENTIRE COHORT			CAUCASIAN SUBSAMPLE		
	$n = 479, L_1 = 934$			$n = 379, L_1 = 721$		
LOCUS	N	$L_2 - C$	L_2	N	$L_2 - C$	L_2
CLASS I						
HLA-A	19	916	935	18	703	721
HLA-B	30	887	917*	26	681	707*
HLA-C	14	921	935	13	706	719
CLASS II						
DRB1	13	927	940	13	711	724
DQB1	5	936	941	5	715	720
DPB1	24	927	951	21	710	731

Table 2: HLA-B alleles associated with low (○) or high (●) viral RNA levels.

ALLELE	ENTIRE COHORT <i>n</i> = 479	CAUCASIAN SUBSAMPLE <i>n</i> = 379	SUPERTYPES ONLY <i>n</i> = 352
<i>B7s</i>			
<i>B*07</i>	●	●	●
<i>B*35</i>	●	●	●
<i>B*51</i>	●	●	●
<i>B*53</i>	●	●	●
<i>B*55</i>	●	●	●
<i>B*56</i>	●	●	●
<i>B*67</i>	○/●	–	●
<i>B27s</i>			
<i>B*14</i>	●	●	●
<i>B*27</i>	○	○	●
<i>B*38</i>	○	●	●
<i>B*39</i>	●	●	●
<i>B*48</i>	○/●	○/●	●
<i>B44s</i>			
<i>B*18</i>	●	●	●
<i>B*37</i>	●	●	●
<i>B*40</i>	●	○	●
<i>B*41</i>	●	●	●
<i>B*44</i>	●	●	●
<i>B*45</i>	○	○	●
<i>B*49</i>	○	○	●
<i>B*50</i>	●	●	●
<i>B58s</i>			
<i>B*57</i>	○	○	○
<i>B*58</i>	○	○	○
<i>B62s</i>			
<i>B*13</i>	○	○	●
<i>B*52</i>	●	●	●
OTHER			
<i>B*08</i>	●	●	–
<i>B*15</i>	●	●	–
<i>B*42</i>	●	–	–
<i>B*47</i>	●	○/●	–
<i>B*81</i>	○	–	–
<i>B*82</i>	○/●	–	–

