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

Peter T. Hraber[1], Bette T. Korber[1,2], Steven Wolinsky[3], Henry A. Erlich[4], Elizabeth A. Trachtenberg[5], and Thomas B. Kepler[6]

***Short Abstract* — Rissanen's minimum description length principle states that the best model to account for some data minimizes the bits required for digital transmission of both the model and the data as encoded via the model. Thus, description lengths are parameter-selection criteria. HLA glycoproteins regulate cellular immune responses to HIV, and are associated with infection progression. HLA polymorphism makes associating alleles with infection outcomes challenging. Comparing description lengths yields allele associations with HIV setpoints, which predict infection progression. Allele associations with viral setpoints support and extend previous studies. Individuals without *B58S* supertype alleles average setpoints 3.6-fold greater than *B58S* carriers.**

***Keywords* — Human Leukocyte Antigens, Human Immunodeficiency Virus, Major Histocompatibility Complex, Immunology, Stochastic Complexity, Information Theory.**

I. Background

THE cell-mediated immune response identifies and eliminates infected cells from an individual. This response is regulated by gene products from the major histocompatibility complex, also known as human leukocyte antigens (HLA). HLA loci are the most polymorphic genes known. They encode ubiquitously expressed cell-surface glycoproteins, which present processed peptide to circulating T cells, and thereby discriminate between self and non-self [1]. This diversity provides a repertoire to recognize evolving antigens. HLA polymorphism is a challenge for statistical genetics, due to compounded error rates from repeated hypotheses tests to identify allele associations with variation in phenotypic traits, such as infection progression.

Acknowledgements: This work was supported by funds from the Elizabeth Glazer Pediatric AIDS Foundation, National Cancer Institute, National Institute for Allergy and Infectious Disease, National Science Foundation award #0077503, and the US Department of Energy.

[1]Theoretical Biology & Biophysics, T-10 MS K710, PO Box 1663, Los Alamos NM 87545 USA. E-mail: phraber@lanl.gov

[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501 USA. E-mail: btk@santafe.edu

[3]Feinberg School of Medicine, Northwestern University, Chicago IL 60611 USA. E-mail: s-wolinsky@northwestern.edu

[4]Roche Molecular Systems, 1145 Atlantic Avenue, Alameda CA 94501 USA. E-mail: henry.erlich@roche.com

[5]Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr. Way, Oakland CA 94609 USA. E-mail: etrachtenberg@chori.org

[6]Department of Biostatistics and Bioinformatics, Box 90090, Duke University, Durham NC 2770 USA. E-mail: kepler@duke.edu

The minimum description length principle is a model-selection criterion that balances the needs for parsimony and fidelity, by penalizing equally for the information required to encode both the statistical model and the residual error. The shortest description length indicates the optimal model among those evaluated, including the number of parameters, and hence, the optimal partition of observations into some number of groups. The penalty for evaluating many alternative hypotheses is an additive constant [2-4].

HLA alleles and HIV setpoints from 479 individuals enrolled in the Chicago Multicenter AIDS Cohort Study provided an opportunity to classify alleles into groups with similar HIV plasma RNA levels, or "setpoints", minimizing description lengths to find the best classifier.

II. Results and Discussion

The best model resulted from clustering supertypes of HLA-B alleles into two groups, one with low and one with high setpoints. Supertypes group HLA alleles having similar peptide-binding anchor motifs [5]. Mapping HLA-B alleles to supertypes was possible for 352 individuals. Individuals without *B58S* alleles averaged HIV setpoints 3.6 times greater than carriers of *B58S* supertype alleles. Due to known associations between low setpoint and favorable infection outcome [6], carriers of *B58S* alleles progress more slowly to AIDS than individuals without them. Mechanisms for these associations include variation in epitope specificity and (more likely) selection that favors rare alleles [7,8].

References

[1] Williams A, Au Peh C, Elliott T (2002) The cell biology of MHC class I antigen presentation. *Tissue Antigens* **59**, 3-17.

[2] Rissanen J (1989) *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.

[3] Rissanen J (1999) Hypothesis selection and testing by the MDL principle. *Comput J* **42**, 260-269.

[4] Burnham KP, Anderson DR (2002) *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd ed. Springer, New York.

[5] Sette A, Sidney J (1999) Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* **50**, 201-212.

[6] Mellors JW, et al. (1996) Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* **272,** 1167-1170.

[7] Trachtenberg EA, et al. (2003) Advantage of rare HLA supertype in HIV disease progression. *Nat Med* **9**, 928-935.

[8] Hraber PT, et al. (2006) HLA and HIV infection progression: application of MDL principle to statistical genetics. *LNBI* **4345,** 1-12.