Tissue Antigens ISSN 0001-2815

14th International HLA and Immunogenetics Workshop: Report of progress in methodology, data collection, and analyses

R. M. Single[1], D. Meyer[2], S. J. Mack[3,4], A. Lancaster[5], H. A. Erlich[3,4] & G. Thomson[5]

1 Department of Mathematics and Statistics, University of Vermont, Burlington, VT, USA
2 Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, Brazil
3 Roche Molecular Systems, Alameda, CA, USA
4 Children's Hospital Oakland, Research Institute, Oakland, CA, USA
5 Department of Integrative Biology, University of California, Berkeley, CA, USA

Key words
anthropology; biostatistics; International Workshop; Major Histocompatibility Complex database; population genetics

Correspondence
Richard Single
Department of Mathematics and Statistics
University of Vermont
Burlington
VT
USA
Tel: 1 802 656 8631
Fax: 1 802 656 2552
e-mail: Richard.Single@uvm.edu

doi: 10.1111/j.1399-0039.2006.00767.x

Abstract

The Biostatistics Component of the 13th International Histocompatibility Workshop (IHWS) developed the *PyPop* (Python for Population Genomics) software framework for high-throughput analysis and quality control (QC) assessments of highly polymorphic genotype data. Since its initial release, the software has had several new analysis modules added to it. These additions, combined with improved data filtering and QC modules, facilitate analyses of data at different levels (allele, haplotype, amino acid sequence, and nucleotide sequence). Since the 13th IHWS, much of the human leukocyte antigen (HLA) data from the workshop, QCed via PyPop and other methods, have been made publicly available through the Major Histocompatibility Complex database web site at the National Center for Biotechnology Information (http://ncbi.nih.gov/mhc/). The Anthropology/Human Genetic Diversity component (AHGDC) data have been used in a variety of studies. Prugnolle et al. used this data to corroborate a model of pathogen-driven selection as a factor related to high levels of diversity at HLA loci. Using a comparative genomics approach contrasting results for HLA and non-HLA markers, Meyer et al. analyzed a subset of the 13th IHWS AHGDC data and showed that HLA loci show detectable signs of both natural selection and the demographic history of populations.

Introduction

The overarching goals of the work carried out by the Biostatistics and Anthropology/Human Genetic Diversity components (AHGDC) are to improve our understanding of the global distribution of human leukocyte antigen (HLA) alleles and haplotypes, the forces (selective and demographic) that have shaped these distributions, and the implications of these distributions for anthropological, disease association, and transplantation studies. Recent work that has been done in pursuit of these goals is summarized in the chapters described below.

Chapter overviews

Population samples and genotyping technology

The population sampling and genotype data generation efforts of the 13th International Histocompatibility Workshop (IHWS) AHGDC continued in the 14th Workshop, with new investigators continuing to be recruited to the project, and new high-resolution class I and class II genotype data being generated for 112 population samples, representing ~12,000 sampled individuals, from around the world. The genotypes and allele and haplotype frequencies

© 2007 The Authors
Journal compilation **69** Suppl. 1 (185–187) © 2007 Blackwell Munksgaard

for the 95 population samples genotyped as part of the 13th IHWS are available to the public through the National Center for Biotechnology Information's Major Histocompatibility Complex database (dbMHC) resource (http://ncbi.nih.gov/mhc/). The goals for this resource include aiding in the study of: HLA allelic diversity; the evolution of HLA polymorphism; the evolution and migration of human populations; best practices for maintaining and searching bone marrow donor registries; and HLA-associated disease susceptibility. A detailed description of recent and future updates to the AHGDC data at dbMHC along with the genotyping technology used is given in Mack et al. (3).

Python for Population Genomics update: a software pipeline for large-scale multilocus population genomics

Although initially developed for HLA data, the PyPop (Python for Population Genomics) software (http://www.pypop.org) is applicable to any genotype-level and allele frequency data (1, 2). This software was used for analyses by the AHGDC, hematopoietic stem cell transplantation, and disease components of the 13th IHWS to facilitate analyses for the IHWS (4–7) and has been used in several subsequent studies by these and other groups (listed in reference 3).

The various data filters available in PyPop facilitate across population analyses, comparisons of allele-level and amino acid-level population genetic results, and data quality control. The allele-to-sequence filter has made it possible to address questions about the level at which selection may be acting in the MHC. It is hypothesized that single amino acid sites may be a target of selection, because changes at specific amino acids of an HLA molecule can confer the ability to bind a different repertoire of peptides. The goal is to detect specific amino acid residues that may be subject to selection in all populations or in populations from specific geographic regions. The amino acid sequence translation capabilities of PyPop, along with the AHGDC data, have made it possible to expand upon work by Salamon et al. (8) and Valdes et al. (9), which identified selection at specific class I and II amino acid sites, in a large set of populations with high-resolution typing. A detailed description of these and other enhancements to the PyPop software framework since the 13th IHWS is given in Lancaster et al. (10).

Comparative analyses and geographic region-specific HLA variation

The AHGDC data have been used to explore the relationship between HLA diversity and pathogen richness (11). The authors also compared results relating genetic diversity and distance from Africa observed in these HLA data with patterns observed in non-HLA data. Meyer et al. (12) contrasted results from HLA data (a subset of the AHGDC populations) and non-HLA data in a large set of populations, representing the major geographic regions of the world. They found that HLA-A, B, C, and DQB1 showed patterns of variation indicative of balancing selection when allele frequencies are compared with their neutral expectations. While high levels of heterozygosity and linkage disequilibrium (LD) were found among HLA loci, a pattern expected for loci under natural selection, similar geographic regional trends were seen in both data sets, indicative of signatures of the demographic history of populations on LD and allele frequency variation. Although large numbers of geographic region-specific HLA alleles were identified, with some at relatively high frequency, the HLA data presented similar levels of population differentiation as that seen in published non-HLA data sets. This result suggests either that selection has a weak effect on population differentiation or that selective regimes are sufficiently complex that no simple genetic signature is evident based on these analyses of population differentiation. A summary of recent results that expand on previous studies of geographic region-specific HLA variation, pointing to similarities and differences with results from other non-HLA studies, is forthcoming.

Summary and continuing efforts

The analysis of inter- and intrapopulation genetic variation at HLA loci has provided valuable information for historical inferences and offered important examples of the effect of natural selection upon variation in human populations. The increasing number of identified alleles raised the question of whether results obtained with serological and low-resolution approaches would be confirmed with higher-resolution typing methods. In addition, specific questions regarding population history and microevolutionary processes (natural selection and genetic drift) can be more powerfully addressed by using larger data sets, typed at the same level of resolution. A key question of interest involves determining the extent to which natural selection has shaped genetic variation at HLA loci, and how the effects of selection at these loci can be discerned from that of demographic history. In this context, the data generated by the 13th IHWG AHGDC have allowed several important questions about HLA diversity to be approached in novel ways (11, 12).

PyPop was used to carry out analyses of selective forces and LD at the amino acid level in the 13th IHWS AHGDC data set. Inferred patterns of selection were distinct between loci, but consistent across populations at each locus. Significant levels of heterozygosity (consistent with balancing selection) were detected at amino acid positions contributing to class I binding pockets A–F, as well as at positions thought to interact with the T-cell receptor. Many amino acid positions that are polymorphic in the total set of

© 2007 The Authors

alleles recognized by the World Health Organization Nomenclature Committee for HLA Factors were invariant in this data set, or displayed significant values of homozygosity, consistent with the action of purifying selection at the population level. Hitchhiking was detected for some positions that are not thought to contribute to the so-called Antigen Recognition Sequence. High values of LD were observed between pairs of variable positions at all loci, and in some cases between positions in different exons. Strong LD was observed in regions that correspond to serological epitopes (e.g. the Bw4/Bw6 epitope). The amino acid-level analyses will continue to be refined, and these and other methods will be applied to new data received as part of the 14th Workshop.

Acknowledgments

This work was supported by National Institutes of Health grants AI49213 and GM35326, FAPESP (Brazil) grant 03/01583-8, and U.S. Department of Energy grant DE-FG02-00ER45828. We wish to thank the participants of the IHWS AHGDC for their contributions of data.

Conflict of Interest Statement

All authors have declared no conflicts of interests.

References

1. Lancaster A, Nelson MP, Meyer D, Single RM, Thomson G. PyPop: a software framework for population genomics: analyzing large-scale multi-locus genotype data. *Pac Symp Biocomput* 2003: **8**: 514–25.
2. Lancaster A, Nelson MP, Single RM, Meyer D, Thomson G. Software framework for the biostatistics core. In: Hansen JA, ed. *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference, Volume I. Seattle: IHWG Press, 2007*.
3. Mack SJ, Sanchez-Mazas A, Single RM et al. Population samples and genotyping technology. *Tissue Antigens* 2007: **69** (Suppl. 1): 188–91.
4. Meyer D, Single RM, Mack SJ, Lancaster AK, Nelson MP, Thomson G. Haplotype frequencies and linkage disequilibrium among classical HLA genes. In: Hansen JA, ed. *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference, Volume I. Seattle: IHWG Press, 2007*.
5. Single RM, Meyer D, Mack SJ et al. Single locus polymorphism of classical HLA genes. In: Hansen JA, ed. *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference, Volume I. Seattle: IHWG Press, 2007*.
6. Single RM, Malkki M, Thomson G, Mather KA, Carrington M, Petersdorf E. Linkage disequilibrium and HLA-A: B: DRB1 haplotype probabilities for class I, II, III microsatellite markers in unrelated donor hematopoietic stem cell transplantation. In: Hansen JA, ed. *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference, Volume I. Seattle: IHWG Press, 2007*.
7. Thomson G, Hongzhe L, Dorman J et al. Statistical approaches for analyses of HLA-associated and other complex diseases. In: Hansen JA, ed. *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference, Volume I. Seattle: IHWG Press, 2007*.
8. Salamon H, Klitz W, Easteal S et al. Evolution of HLA class II molecules: allelic and amino acid site variability across populations. *Genetics* 1999: **152**: 393–400.
9. Valdes AM, Meyer D, McWeeney SK, Nelson MP, Thomson G. Locus and population specific evolution in HLA class II genes. *Ann Hum Genet* 1999: **63**: 27–43.
10. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update – a software pipeline for large-scale multilocus population genomics. *Tissue Antigens* 2007: **69** (Suppl. 1): 192–7.
11. Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol* 2005: **15**: 1022–7.
12. Meyer D, Single RM, Mack SJ, Erlich H, Thomson G. Signatures of demographic history and natural selection in the human MHC loci. *Genetics* 2006: **173**: 2121–42.

© 2007 The Authors