AMERICAN SOCIETY FOR HISTOCOMPATIBILITY & IMMUNOGENETICS

15000 Commerce Parkway • Suite C • Mt. Laurel, NJ 08054
Phone: 856.638.0428 • Fax: 856.439.0525 • Web site: www.ashi-hla.org • E-mail: info@ashi-hla.org

November 20, 2013

Genomic Data Sharing Policy Team
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

To the Genomic Data Sharing Policy Team,

We are clinical scientists and researchers in the fields of transplantation, histocompatibility, immunogenetics and immunogenomics, responding to the Request for Public Comments on the Draft NIH Genomic Data Sharing (GDS) Policy issued on September 20th 2013. Collectively, we have extensive clinical and research experience with genes in the Major Histocompatibility Complex (MHC) and Leucocyte Receptor Complex (LRC) regions of the human genome. These regions (on chromosomes 6 and 19, respectively) are central to the study of disease etiology, diagnosis and therapy, but are poorly represented in current genome assemblies due to high levels of polymorphism and structural variation. Our comments, while specific to these immunogenomic regions, apply to all regions of the genome that display high levels of polymorphism and structural variation.

We would like to address the human genomic data submission expectations outlined in section IV.C.1 of the GDS Policy. Appendix A of the GDS Policy states that submission of Level 1 processed data (aka, initial sequence reads) is not expected for human data if those reads are included in a Level 2 aligned sequence file (e.g. BAM format), and states that Level 2 processed data (i.e., DNA sequence aligned to a reference sequence or *de novo* assembly) are expected for submission. However, Level 1 processed data that have not been aligned to a reference sequence or that have been excluded from *de novo* assembly do not appear to be expected for submission. As we detail below, the continual discovery of new HLA and KIR sequence polymorphisms can rapidly invalidate the interpretation of aligned and *de novo* assembled sequences for these genes, reducing their utility for future studies. Furthermore, the current state of the primary and alternative alignments for the MHC and LRC regions of the genome is insufficient to permit reliable and accurate alignment of initial sequence reads for these regions. The sharing of unmapped or unaligned initial sequence reads is critical for the investigation of these regions.

The MHC region on human chromosome 6p21.3 is the most medically relevant region of the human genome. More than 100 infectious, autoimmune and pharmacological disease phenotypes and cancers are associated with genetic variation in the MHC[1-9], and in particular with the Human Leucocyte Antigen (HLA) genes. HLA molecules are critical components of the adaptive immune system, mediating the specific destruction of infected cells and production of antibodies. In addition, HLA molecules interact functionally with Killer-cell Immunoglobulin-like Receptor (KIR) molecules, key components of the innate immune system that also play critical roles in transplantation and disease[10-20]. The HLA and KIR (chromosome 19q13.4) regions are the most polymorphic in the human genome[2, 21, 22, 23, 24]; both are polygenic and highly dense with homologous genes[2, 21, 25, 26], and both display extensive structural variation[27, 28]. Due to extensive genetic variation observed for these genes among

human populations, study of the HLA and KIR genes is also a model for health disparities research[29].

The investigation of HLA and KIR polymorphism is an active and ongoing pursuit; as of October 2013, 678 unique KIR gene sequences and 9,945 unique HLA gene sequences have been identified[23, 24]. These sequences are housed in the IPD-KIR[24] and IMGT/HLA[23] Databases, which are updated on a regular basis. Both were most recently updated in October of 2013, with the addition of 79 new KIR sequences, and 439 new and extended HLA sequences. The IMGT/HLA and IPD-KIR Databases are the primary resources for the alignment of HLA and KIR initial sequence reads and for the validation of *de novo* sequence assemblies of these reads, and are the only resources available to relate gene sequences to the HLA or KIR allele nomenclature[30, 31], complex naming systems that are key for investigations of these genes[32, 33].

The constant increase in sequence polymorphism knowledge at these loci means that any *de novo* and aligned sequence assemblies for these genes will rapidly become obsolete, as the initial sequence reads need to be realigned and *de novo* assemblies revalidated in the context of each database update. The GDS Policy's expectation of the sharing of aligned or assembled Level 2 processed data alone for these genes means that those data can never be reevaluated in the context of future IMGT/HLA or IPD-KIR database updates. The cultural, medical and scientific ramifications for this information loss are unacceptable; this loss alone should be sufficient reason to reexamine the GDS Policy.

Because of sequencing and assembly challenges posed by the high level of polymorphism and structural variation within the MHC and LRC regions, these regions are poorly represented in Genome Reference Consortium (GRC) assembly GRCh37.p13; seven alternate locus assemblies are available for the MHC region, and eight alternative haplotypes are available as novel assemblies for the LRC region. All of these reference and alternative assemblies describe haplotypes common only in European populations[28, 34, 35, 36, 37, 38, 39] and do little to represent the extensive divergence and polymorphism observed in the USA and worldwide[40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50]. Moreover, these alternative alignments are largely incomplete and insufficiently reflect even basic levels of established structural variation and polymorphism for the HLA and KIR genes. We anticipate that a large number of complete reference assembly sequences will be needed to enable reliable genomic investigations and personalized medical applications for the MHC and LRC regions.

For example, of the five major structural variants (the DR1, DR51, DR52, DR53 and DR8 haplotypes) recognized for the HLA-DRB genes[27], only sections of the DR52 and DR53 haplotypes are represented in the alternative MHC alignments, while a section of DR51 is represented in the primary assembly. The DR1 and DR8 haplotypes, which constitute 38% of known HLA-DRB polymorphisms[23], are not represented in GRCh37.p13.

The primary assembly for the LRC includes a complete KIR haplotype representing a single genomic structure common in European populations [28]. Although many KIR haplotypes of differing gene content are known, of the eight novel assemblies for the LRC, only one (RefSeq NW_003571055.1) includes a KIR haplotype of alternative genomic structure. These LRC haplotype structures are medically important; their variation is associated with reproductive disorders, as well as decreased relapse after

bone marrow transplant, the bone marrow graft versus leukemia reaction and bone marrow graft versus host disease[13, 14, 16, 20].

In addition, large gaps in the alignments for these regions omit characteristic genes. Four of the seven alternative locus assemblies for the MHC omit the HLA-DRB1 gene and three omit the HLA-B gene, both of which are present in all individuals. None of these alternative locus assemblies contain the ~70kb segment that includes HLA-Y and associated genes. As noted above, many KIR genes are not represented in GRCh37.p13. However, these missing genes are represented in UCSC's hg19 assembly, and are present in one of four chromosome 19 unlocalized genomic contigs included in GRCh37.p13. This discrepancy between the h37 and hg19 assemblies illustrates the limitations of current alignment methods for this important genomic region.

The high levels of similarity among HLA genes and pseudo genes and among KIR genes and pseudogenes pose challenges for their *de novo* assembly and complicate the use of the alignments in assembly GRCh37.p13. The sequences of many HLA pseudogenes and gene fragments can be erroneously mapped to HLA genes[51], and it is likely that this has occurred in the primary and alternative locus assemblies. This degree of error makes it impossible for meaningful information to be obtained from Level 2 processed data.

Overall, the sharing of aligned or assembled Level 2 processed data alone will severely limit the research community's capacity for novel investigations and meta-analyses of immunogenomic data. For example, the potential introgression of HLA polymorphisms from archaic human species in the modern human population[52] could not have been detected through the analysis of Level 2 processed data; only the availability of all initial sequence reads for the Neanderthal and Denisovan genomes made this work possible.

Given the ongoing detection of new HLA and KIR polymorphisms and the current state of the genome assembly, it is our opinion that acceptance of aligned or assembled Level 2 processed data for the MHC and LRC regions is insufficient to meet the GDS Policy's stated goal of ensuring the responsible sharing of research data. We recommend that the GDS Policy require that any shared Level 2 processed data permit the complete regeneration of Level 1 processed data, making initial sequence reads that have not been aligned to a reference sequence or that have been excluded from *de novo* assembly available for future studies. This will allow shared genomic data to be reevaluated in the context of future improvements in the genomic assembly and alignment methodologies, and future expansions of relevant reference polymorphism databases. Ultimately, this approach to data sharing will represent an investment in the future of genomic investigation, stimulating novel research efforts and fostering improved clinical outcomes.

Sincerely Yours,

The ASHI Scientific Affairs Committee

This letter of comment on the Draft NIH Genomic Data Sharing Policy has been authored by the undersigned on behalf of the Scientific Affairs Committee of the American Society of Histocompatibility and Immunogenetics, the Immunogenomics Data Analysis Working Group, and the Immunogenomic Next Generation Sequencing Data Consortium.

Attila Berces, PhD
Chairman, Omixon

Nezih Cereb, MD
CEO, Histogenetics

Professor Frans H.J. Claas, PhD
Director, Eurotransplant Reference Laboratory
Professor Transplantation Immunogenetics, Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, The Netherlands

Lisa Creary, PhD
Post Doctoral Research Scientist, Histocompatibility and Immunogenetics Department, NHS Blood and Transplant, United Kingdom

Sue Davey, MSc, DipRCPath
HLA NGS project leader, Histocompatibility and Immunogenetics Department, NHS Blood and Transplant, United Kingdom

Professor Peter Donnelly, PhD, FRS
Director, Wellcome Trust Centre for Human Genetics
Professor of Statistical Science, Department of Statistics, University of Oxford, United States

Nadim El-Awar, PhD, MBA
Research Scientist
Transplant Diagnostics, Thermo Fisher Scientific

Henry A Erlich, PhD
Senior Scientist, Center for Genetics, Children's Hospital & Research Center Oakland, United States

Professor Marcelo A Fernández-Viña, PhD
Chairman, 17th International HLA and Immunogenetics Workshop
Organizer, Immunogenomic Next Generation Sequencing Data Consortium
Professor, Department of Pathology, Stanford School of Medicine
Director, Histocompatibility, Immunogenetics, and Disease Profiling Laboratory, Stanford Medical School Blood Center, United States

Professor Gottfried Fischer, MD
President, European Federation for Immunogenetics
Associate Professor, Department for Blood Group Serology and Transfusion Medicine, Medical University of Vienna, Austria

Professor Howard M. Gebel, PhD
Vice chair, ASHI Scientific and Clinical Affairs Committee
Professor of Pathology, Emory University Hospital, United States

Daniel Geraghty, PhD

Clinical Research Division, Fred Hutchinson Cancer Research Center, United States

Benjamin Gifford
Research and Development, Life Technologies

Professor Pierre-Antoine Gourraud, PhD MPH
Dept. of Neurology, School of Medicine, University of California San Francisco, United States

Professor Wolfgang Helmberg, MD
Professor of Medicine, Department of Serology and Transfusion Medicine, Medical University Graz, Austria

Jill A Hollenbach, PhD MPH
Chair, ASHI Scientific and Clinical Affairs Committee
Co-Chair, Immunogenomic Data Analysis Working Group
Organizer, Immunogenomic Next Generation Sequencing Data Consortium
Staff Scientist, Center for Genetics, Children's Hospital & Research Center Oakland, United States

Susan H Hsu, PhD
Director, Histocompatibility / Molecular Genetics, American Red Cross

Uma Kanga, PhD
Senior Research Scientist, Department of Transplant Immunology and Immunogenetics, All India, Institute of Medical Sciences, India

Steven J Mack, PhD
Editor In Chief, Human Immunology
Co-Chair, Immunogenomic Data Analysis Working Group
Organizer, Immunogenomic Next Generation Sequencing Data Consortium
Staff Scientist, Center for Genetics, Children's Hospital & Research Center Oakland, United States

Martin Maiers
Organizer, Immunogenomic Next Generation Sequencing Data Consortium
Director of Bioinformatics, National Marrow Donor Program, United States

Professor Steven GE Marsh, BSc PhD ARCS
President, International HLA and Immunogenetics Workshop Council
Deputy Director of Research, Anthony Nolan Research Institute
Professor of Immunogenetics, Cancer Institute, University College London, United Kingdom

Professor Diogo Meyer, PhD
Professor, Department of Genetics and Biology, Institute of Biosciences, University of São Paulo, Brazil

Professor Derek Middleton, DSc, PhD, FRCPath
President, International HLA and Immunogenetics Workshop Council
Consultant Clinical Scientist, Transplant Immunology, Royal Liverpool University Hospital Trust
Professor, Institute of Infection and Global Health, University of Liverpool, United Kingdom

Professor Dimitri S Monos, PhD

Director, Immunogenetics Laboratory
Professor of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, United States

Professor Cristina Navarrete, PhD, FRCPath
Head of Histocompatibility and Immunogenetics Research, NHS Blood & Transplant, United Kingdom

Paul J Norman PhD
Senior Research Scientist, Department of Structural Biology, Stanford University School of Medicine, United States

José M Nunez
Laboratory of Anthropology, Genetics and Peopling History, Department of Genetics and Evolution, Anthropology Unit, University of Geneva, Switzerland

John Ord
Research Associate, Medical School, University of Cambridge, United Kingdom

Professor Daniel Ramon, PhD, D[ABHI], HCLD[ABB]
Clinical Assistant Professor
Director, Histocompatibility Laboratory, Department of Pathology, University of Michigan Medical School, United States

Bryan Ray, PhD
Director Research and Development | LIFECODES Products

James Robinson
Senior Bioinformatics Scientist, Anthony Nolan Research Institute, United Kingdom

Professor Alicia Sanchez-Mazas
Director, Laboratory of Anthropology, Genetics and Peopling History, Department of Genetics and Evolution, Anthropology Unit, University of Geneva, Switzerland

Professor David Senitzer, PhD, ABMLI, ABHI
Clinical Professor of Hematology and Hematopoietic Stem Cell Transplantation
Director, Histocompatibility Laboratory, City of Hope National Cancer Center, United States

Professor Richard M. Single, PhD
Assistant Professor, Department of Mathematics and Statistics, University of Vermont, United States

Professor Emeritus Glenys Thomson, PhD
Department of Integrative Biology, University of California Berkeley, United States

Professor Elizabeth A Trachtenberg, MS, PhD, D(ABHI)
Director, HLA, Immunogenetics and Disease Profiling Laboratory, Stanford Blood Center
Professor, Department of Pathology, Stanford University School of Medicine, United States

Christien Voorter, PhD
Transplantation Immunology, Tissue Typing Laboratory, University Hospital Maastricht, The Netherlands

Literature Cited

1. Johnson, A.D. and C.J. O'Donnell, *An open access database of genome-wide association results.* BMC Med Genet, 2009. **10**: p. 6.
2. Shiina, T., et al., *The HLA genomic loci map: expression, interaction, diversity and disease.* J Hum Genet, 2009. **54**(1): p. 15-39.
3. Watson, C.T., et al., *Estimating the proportion of variation in susceptibility to multiple sclerosis captured by common SNPs.* Sci Rep, 2012. **2**: p. 770.
4. Isobe, N., et al., *Genetic risk variants in African Americans with multiple sclerosis.* Neurology, 2013. **81**(3): p. 219-27.
5. Hildesheim, A., et al., Association of HLA class I and II alleles and extended haplotypes with nasopharyngeal carcinoma in Taiwan. J Natl Cancer Inst. 2002. 94(23): p. 1780-9..
6. Chapman, S.J. and A.V. Hill, *Human genetic susceptibility to infectious disease.* Nat Rev Genet, 2012. **13**(3): p. 175-88.
7. Leish, G.E.N.C., et al., *Common variants in the HLA-DRB1-HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis.* Nat Genet, 2013. **45**(2): p. 208-13.
8. Kirino, Y., et al., *Genome-wide association analysis identifies new susceptibility loci for Behcet's disease and epistasis between HLA-B*51 and ERAP1.* Nat Genet, 2013. **45**(2): p. 202-7.
9. Nomura, T. and T. Matano, *Association of MHC-I genotypes with disease progression in HIV/SIV infections.* Front Microbiol, 2012. **3**: p. 234.
10. Santin, I., et al., *Association of KIR2DL5B gene with celiac disease supports the susceptibility locus on 19q13.4.* Genes Immun, 2007. **8**(2): p. 171-6.
11. Brown, M.A., *Genetics and the pathogenesis of ankylosing spondylitis.* Curr Opin Rheumatol, 2009. **21**(4): p. 318-23.
12. King, A.L., et al., *A genome-wide family-based linkage study of coeliac disease.* Ann Hum Genet, 2000. **64**(Pt 6): p. 479-90.
13. Martin, M.P. and M. Carrington, *KIR Locus Polymorphisms: Genotyping and Disease Association Analysis*, in *Methods in Molecular Biology: Innate Immunity*, J. Ewbank and E. Vivier, Editors. 2007, Humana Press Inc.: Totowa, NJ. p. 49-64.
14. Hollenbach, J.A., et al., *Susceptibility to Crohn's disease is mediated by KIR2DL2/KIR2DL3 heterozygosity and the HLA-C ligand.* Immunogenetics, 2009. **61**(10): p. 663-71.
15. Ahlenstiel, G., et al., *Distinct KIR/HLA compound genotypes affect the kinetics of human antiviral natural killer cell responses.* J Clin Invest, 2008. **118**(3): p. 1017-26.
16. Miller, J.S., et al., *Missing KIR ligands are associated with less relapse and increased graft-versus-host disease (GVHD) following unrelated donor allogeneic HCT.* Blood, 2007. **109**(11): p. 5058-61.
17. Kulkarni, S., M.P. Martin, and M. Carrington, *The Yin and Yang of HLA and KIR in Human Disease.* Seminars in Immunology, 2008. **In Press**.
18. Martin, M.P., et al., *Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS.* Nat Genet, 2002. **31**(4): p. 429-34.
19. Khakoo, S.I. and M. Carrington, *KIR and disease: A model system or system of models?* Immunol Rev, 2006. **214**: p. 186-201.
20. Carrington, M. and M.P. Martin, *The impact of variation at the KIR gene cluster on human disease.* Curr Top Microbiol Immunol, 2006. **298**: p. 225-57.
21. Mungall, A.J., et al., *The DNA sequence and analysis of human chromosome 6.* Nature, 2003. **425**(6960): p. 805-11.

22. Trowsdale, J. and P. Parham, *Mini-review: defense strategies and immunity-related genes.* Eur J Immunol, 2004. **34**(1): p. 7-17.
23. Robinson, J., et al., *The IMGT/HLA Database.* Nucleic Acids Research, 2013. **41**: p. D1222-7.
24. Robinson, J., et al., *IPD - the Immuno Polymorphism Database.* Nucleic Acids Research, 2013. **41**: p. D1234-40.
25. *Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium.* Nature, 1999. **401**(6756): p. 921-3.
26. Vandiedonck, C. and J.C. Knight, *The human Major Histocompatibility Complex as a paradigm in genomics research.* Brief Funct Genomic Proteomic, 2009. **8**(5): p. 379-94.
27. G. Andersson, *Evolution of the human HLA-DR region.* Front Biosci, 1998. **3**: p. d739-45.
28. Hollenbach, J.A., et al., *Killer cell immunoglobulin-like receptor (KIR) gene content variation in the HGDP-CEPH populations.* Immunogenetics, 2012. **64**(10): p. 719-37
29. Noble, J. A., et al., *HLA class II genotyping of African American type 1 diabetic patients reveals associations unique to African haplotypes.* Diabetes, 2013. **62**(9): p. 3292-9.
30. Marsh, S.G.E., et al., *Nomenclature for factors of the HLA system, 2010.* Tissue Antigens, 2010. **75**(4): p. 291-455.
31. Marsh, S.G.E., et al., *Killer-cell Immunoglobulin-like Receptors (KIR) Nomenclature Report, 2002.* Human Immunology, 2003. **64**: p. 648-654.
32. Hollenbach, J. A., et al., *A community standard for immunogenomic data reporting and analysis: proposal for a STrengthening the REporting of Immunogenomic Studies statement.* Tissue Antigens, 2011. **78**(5): p. 333-44.
33. Mack, S. J., and J. A. Hollenbach. *Allele Name Translation Tool and Update NomenCLature: software tools for the automated translation of HLA allele names between successive nomenclatures.* Tissue Antigens, 2010. **75**(5): p. 457-61.
34. Gragert, L., et al., Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. Hum Immunol, 2013. **74**(10): p. 1313-20.
35. Maiers, M., L. Gragert and W. Klitz. *High-resolution HLA alleles and haplotypes in the United States population.* Hum Immunol, 2007. **68**(9): p. 779-88.
36. Vierra-Green, C., et al., *Allele-level haplotype frequencies and pairwise linkage disequilibrium for 14 KIR loci in 506 European-American individuals.* PLoS One, 2012. **7**(11):e47491.
37. Mack, S. J., et al., *HLA-A, -B, -C, and -DRB1 allele and haplotype frequencies distinguish Eastern European Americans from the general European American population.* Tissue Antigens, 2009. **73**(1): p. 17-32.
38. Mack, S. J., et al., *Human leukocyte antigen-A, -B, -C, -DRB1 allele and haplotype frequencies in Americans originating from southern Europe: contrasting patterns of population differentiation between Italian and Spanish Americans.* Hum Immunol, 2011. **72**(2): p. 144-9.
39. Hollenbach, J. A., et al., *16(th) IHIW: population global distribution of killer immunoglobulin-like receptor (KIR) and ligands.* Int J Immunogenet, 2013. **40**(1): p. 39-45.
40. Norman, P. J., et al., *Unusual selection on the KIR3DL1/S1 natural killer cell receptor in Africans.* Nat Genet, 2007. **39**(9): p. 1092-9.
41. Norman, P. J., et al., *Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, and haplotypes*. Genome Res, 2009. 19(5): p. 757-69.

42. Norman, P. J., et al., *Co-evolution of Human Leukocyte Antigen (HLA) Class I Ligands with Killer-Cell Immunoglobulin-Like Receptors (KIR) in a Genetically Diverse Population of Sub-Saharan Africans*. Plos Genet, 2013. 9(10): e1003938.
43. Tu, B., et al., *HLA-A, -B, -C, -DRB1 allele and haplotype frequencies in an African American population.* Tissue Antigens, 2007. **69**(1): p. 73-85.
44. Ellis, J. M., et al., *Diversity is demonstrated in class I HLA-A and HLA-B alleles in Cameroon, Africa: description of HLA-A*03012, *2612, *3006 and HLA-B*1403, *4016, *4703.* Tissue Antigens, 2000. **56**(4): p. 291-302.
45. Solberg, O. D., et al., *Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies*. Hum Immunol. 2008, **69**(7): p. 443-64.
46. Nunes, J.M., 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): p. 18-30.
47. Buhler, S. and A. Sanchez-Mazas. *HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events.* PLoS One. 2011, **6**(2): p. e14643.
48. Riccio, M.E., 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): p. 21-30.
49. Gonzalez-Galarza, F. F., et al., *16(th) IHIW: extending the number of resources and bioinformatics analysis for the investigation of HLA rare alleles.* Int J Immunogenet, 2013, **40**(1): p. 60-5.
50. Mack, S. J., et al., *Common and well-documented HLA alleles: 2012 update to the CWD catalogue.* Tissue Antigens, 2013. **81**(4): p. 194-203.
51. Major, E., et al., *HLA Typing from 1000 Genomes Whole Genome and Whole Exome Illumina Data*. PLoS ONE, 2013. **8**(11): p. e78410.
52. Abi-Rached, L., et al., *The shaping of modern human immune systems by multiregional admixture with archaic humans.* Science, 2011. **7**;334(6052): p. 89-94.