Published in final edited form as:

Immunogenetics. 2015 September ; 67(9): 479–485. doi:10.1007/s00251-015-0857-y.

Very long haplotype tracts characterized at high resolution from *HLA* homozygous cell lines

Paul J. Norman¹, Steve Norberg², Neda Nemat-Gorgani¹, Thomas Royce², Jill A. Hollenbach³, Melissa Shults Won², Lisbeth A. Guethlein¹, Kevin L. Gunderson², Mostafa Ronaghi², and Peter Parham^{1,*}

¹Departments of Structural Biology and Microbiology & Immunology, Stanford University School of Medicine, Stanford, CA94305, USA

²Illumina Inc, 5200 Illumina Way, San Diego, CA92122, USA

³Department of Neurology, University of California, San Francisco School of Medicine, San Francisco, CA94158, USA

Abstract

The *HLA* region of chromosome 6 contains the most polymorphic genes in humans. Spanning ~5Mbp the densely packed region encompasses approximately 175 expressed genes including the highly polymorphic *HLA class I* and *II* loci. Most of the other genes and functional elements are also polymorphic, and many of them are directly implicated in immune function or immune-related disease. For these reasons this complex genomic region is subject to intense scrutiny by researchers with the common goal of aiding further understanding and diagnoses of multiple immune-related diseases and syndromes. To aid assay development and characterization of the classical loci, a panel of cell lines partially or fully homozygous for *HLA class I* and *II* was assembled over time by the International Histocompatibility Working Group (IHWG). Containing a minimum of 88 unique *HLA* haplotypes, we show this panel represents a significant proportion of European *HLA* allelic and haplotype diversity (60–95%). Using a high-density whole genome array that includes 13,331 *HLA* region SNPs, we analyzed 99 IHWG cells to map the coordinates of the homozygous tracts at a fine scale. The mean homozygous tract length within chromosome 6 from these individuals is 21Mbp. Within *HLA* the mean haplotype length is 4.3Mbp, and 65% of the cell lines were shown to be homozygous throughout the entire region. In addition, four cell lines are homozygous throughout the complex *KIR* region of chromosome 19 (~250kbp). The data we describe will provide a valuable resource for characterizing haplotypes, designing and refining imputation algorithms and developing assay controls.

Keywords

HLA class I; HLA class II; HLA region genes; Homozygous; HLA imputation; KIR

*Corresponding author: paul.norman@stanford.edu. Telephone (1) 650-723-7456.

Introduction

The *HLA* region of chromosome 6 harbours the most polymorphic human genes (Horton et al. 2004; Tennesen et al. 2012; Robinson et al. 2015), has vital roles in immune defence and reproduction, and exerts a dominant impact on immune-mediated diseases and syndromes (Parham et al. 2013; Trowsdale et al. 2013). The most direct application for clinical *HLA* genotyping is in transplantation, where millions of individuals are genotyped worldwide to enable tissue matching of donors with patients (Terasaki 1969; Petersdorf 2013). *HLA* genotyping is also critical for avoiding specific drug sensitivity (Illing et al. 2013) and assists accurate diagnoses or analyses of many autoimmune diseases (de Bakker et al. 2012). To aid both the characterization of *HLA* variation and development of techniques to detect it, a panel of immortalized *HLA* homozygous cell lines was established almost 30 years ago by the International Histocompatibility Working Group (IHWG) (Yang et al. 1987). The cell panel and its genetic characterization continue to be updated and remain valuable resources for discovering and mapping *HLA* region variation (Marsh et al. 1996; Horton et al. 2004; Dorak et al. 2006; Mickelson E et al. 2006; Hosomichi et al. 2013; Larsen et al. 2014). As many of the cell lines were generated from consanguineous individuals, they also present an opportunity to study *HLA* haplotypes that are identical by descent (IBD). Thus, to reduce complexity during long-range haplotype assembly and ultimately provide unambiguously-phased reference sequences for the entire *HLA* region we genotyped these cell lines to define the precise coordinates of their homozygous tracts. Using a high-density whole genome SNP array, we analyzed 13,331 *HLA* region SNPs from 99 *HLA* homozygous cells. The mean IBD haplotype length was 4.3Mbp, and 65% were homozygous throughout the entire 5.0Mbp *HLA* region of chromosome 6. In addition, four of the cell lines were homozygous throughout the *KIR* region of chromosome 19.

Materials and Methods

Target region coordinates—The classical *HLA* region encompasses 4.7Mbp of chromosome 6p21, and includes all of the *HLA Class I* and *II* genes (Horton et al. 2004). The region is extended herein to 5.0Mbp to account for non-overlapping segments of the two haplotypes that have been fully sequenced, COX and PGF (Stewart et al. 2004). Thus we define the *HLA* region as spanning chr6:28477897-33500000 bp, and flanked by SNP numbers rs2108925-rs449242. The closest genes within these markers are *GPX5* (28493788-28502728) and *ZBTB9* (33422355-33377699). The ‘extended *HLA* region’ (Horton et al. 2004) begins approximately at SNP rs9358871 (chr6:25726675): encompassing *HIST1H2AA* (25726290-25726790) as the gene at the 5’ extreme. The *KIR* region of chromosome 19 is flanked by SNP numbers rs1325158-rs17771967 (chr19:55226402-55380214 bp). All coordinates are shown according to Human Genome build 37 (hg19).

DNA samples and conventional *HLA* genotypes

DNA was extracted from 99 EBV-transformed B cell lines. These were chosen from knowledge of their *HLA Class I* and *II* genotypes (<http://www.ihwg.org/>), because they have the lowest possible heterozygosity within the 5.0Mbp *HLA* complex (Supplementary Fig. 1).

Low passage cells were grown from archive material available at Stanford, or purchased from the IHWG repository. All the cells are homozygous for *HLA-B*, 97% are homozygous for *HLA-A*, *-C* or *-DRB1*, 95% for *-DQB1* and 78% for *-DPB1*. The panel consisted of 87 cells from the 10th IHCW (Yang et al. 1987) and 12 cells from the 12th and 13th IHCW (Marsh et al. 1996; Mickelson E et al. 2006), and included all eight cells targeted by the MHC sequencing consortium (Horton et al. 2004). At least fifty of the cells derive from consanguineous donors (Supplementary Fig. 1), so their haplotype pairs are likely identical by descent (IBD). All of these cells have been genotyped for *HLA class I* and *II* by multiple investigators and clinical transplant centres worldwide (Mickelson E et al. 2006; Hosomichi et al. 2013). DNA samples were genotyped for confirmation to '4-digit' resolution of *HLA-A*, *-B* and *-C* using LABType SSO reagents (One Lambda, Canoga Park, CA) and a Luminex 100 reader (Luminex, Austin TX). In addition, the cells were genotyped for presence of the *HLA-Y* pseudogene, using primers and conditions described previously (Gleimer et al. 2011). *HLA-Y* marks the presence of a ~70kb structural variant detected on some *HLA* haplotypes (Watanabe et al. 1997). We also used data obtained from 90 Europeans, 44 Chinese Han, 45 Japanese and 90 Yorubans studied in the HapMap phase II project (Myers et al. 2005) and who were genotyped using the same whole-genome 1.1M SNP array as the IHWG panel. From this set, we were able to assemble a control sample set of 86 individuals, sex and ethnically matched with 86 of the *HLA*-homozygous cell line donors.

Whole genome SNP analysis

Whole genome SNP genotyping (Steemers et al. 2006) was performed using the Infinium HumanOmni1-Quad BeadChip array (Illumina, San Diego CA) according to the manufacturer's instructions, and analysed using PLINK 1.07 (Purcell et al. 2007). Whole-genome SNP data from one sample (WAR) was obtained as part of another study (Moreno-Estrada et al. 2013). There were 123,995 genome-wide markers specific for copy number variants excluded from further analysis as were all (from 14,490) *HLA*-region SNPs with >10% failure rate. Of the SNPs that passed filtering, 402 occurred within *HLA class I* or *II* genes and only 32 of these were in exons. This reflects the high sequence divergence and non-binary nature of *HLA* allelic polymorphism. Following the filtering process 13,331 SNPs in the *HLA* region were analyzed. Any further SNPs with missing data were excluded per individual so that a mean of 13,269 SNPs were used to estimate the homozygous tract lengths.

The size and location of each homozygous tract that is greater than 1Mbp was estimated using PLINK (Purcell et al. 2007), set to tolerate two heterozygous SNPs per 500kbp window. The coordinates of the homozygous tracts overlapping the *HLA* region were adjusted manually to indicate the estimated start point, which likely occurs between the last known heterozygous SNP and the first homozygous SNP within the segment indicated by PLINK. To search for homozygous tracts in the *KIR* region, the analysis was repeated using 200kbp as the minimum homozygous tract size. The coordinates of HapMap II recombination hotspots were obtained from Myers et al. 2005 (Myers et al. 2005) and mapped to hg19 using the UCSC genome browser liftOver tool (Karolchik et al. 2014). The mean resolution of these hotspots within the *HLA* region is +/- 5kb. Principal components

analysis (PCA) plots were generated from genome-wide SNP data in PLINK binary files using the SNPRelate package in R. Only autosomal, non-monomorphic SNPs were considered for analysis, and the first two principal components plotted.

Results and Discussion

To characterize full-length *HLA* haplotypes at high-resolution and with unambiguous phase, 99 individuals were selected from a panel of *HLA* homozygous cell lines (Yang et al. 1987). Of 1,013,490 genome-wide SNPs that were genotyped successfully, 13,331 fall within the 5.0Mbp *HLA* region (median: 1 SNP per 152bp) and a further 3,947 correspond to the additional 3.1Mbp known as the ‘extended *HLA*’ region (median: 1 SNP per 390bp). Comparison of the entire genome data with that from major world populations using PCA confirmed that the majority (80%) of the cell lines were derived from European-origin donors (Fig. 1). Also included in our panel are two Africans, five Amerindians, seven East Asians, one Hispanic and one South Asian individual (Fig. 1). The genome-wide SNP data from the IHWG cell panel are available in human-readable and PLINK binary format at ImmPort under SDY295: EXP13576 (<https://immport.niaid.nih.gov/>).

Present in the *HLA*-homozygous cell line panel are 24 *HLA-A*, 34 *HLA-B*, 17 *HLA-C*, 31 *HLA-DRB1*, 16 *HLA-DQB1* and 15 *HLA-DPB1* alleles (Fig. 2 and Supplementary Figure 1). *HLA-Y* was detected in 12 of the cell lines, and linked to *A*02:05*, *30:01*, *31:01*, *33:01* or *68:02*, as observed previously (Watanabe et al. 1997; Williams et al. 1999; Coquillard et al. 2004). A minimum of 75 distinct *A-B-C* and 88 *A-B-C-DRB1* haplotypes are present in the IHWG sample panel (Fig. 2 and Supplementary Figure 1). Two are common European haplotypes. Each of the two common haplotypes is present in five homozygous cell lines and these include the COX (*A*01:01*, *B*08:01*, *C*07:01*, *DRB1*03:01*) and PGF (*A*03:01*, *B*07:02*, *C*07:02*, *DRB1*15:01*) haplotypes that were determined to completion by the MHC sequencing project (Stewart et al. 2004). We estimate the allele distribution of the cell lines represents 90–95% of the European and 60–80% of the world population by allele frequency (from data available at dbMHC and allelefrequencies.net; Fig. 2). We also estimate the sample panel represents 20–50% of the *HLA* haplotype diversity of Europeans. The high-density SNP genotyping showed that 88.3% SNP markers within the *HLA* region vary in this sample set, which is similar to 86% in a sex and ethnically matched control set of samples that were not selected for *HLA* genotype. Together, these findings show that the sample set represents a significant proportion of European *HLA* diversity and provides a smaller but reasonable representation of worldwide *HLA* allelic diversity, despite having been selected for *HLA* homozygosity.

We used the high-density SNP data to measure the coordinates of the homozygous tracts that intersect the *HLA* region. These are shown in Supplementary Figure 2. The mean size of homozygous tract within chromosome 6 of the IHWG cell lines is 21Mbp and the largest is 135.0Mbp (79% of chromosome 6: cell CB6B). Sixty-three of the cells are homozygous through the entire 5.0Mbp *HLA* region and 53 of these are also homozygous through the ‘extended *HLA*’ (Fig. 3). Whilst 49–62% of the IHWG panel is derived from consanguineous donors (Supplementary Figure 1), 66–80% of the cells that are homozygous through the entire *HLA* region are consanguineous (Supplementary Figure 2). Thus the set of

longest haplotypes that we describe is likely to be enriched for ones that are identical by descent (IBD). The IHWG cell lines are homozygous for an average 4.3Mbp (87%) of the 5.0Mbp region and the minimum *HLA* homozygous tract length is 1.0Mbp (cell JY). Four of the homozygous tracts are not continuous through the region, each being broken by a single heterozygous segment of ~200–600kb (BER, EJ32B, HOKKAIDO and SA1: Supplementary Figure 2). One sample (HOM2) had unexpectedly high heterozygosity in the intervals between the classical *HLA* genes that had been genotyped with conventional methods (Fig. 4). As the SNP diversity within these segments was similar to that of the heterozygous samples, this shows that the two *HLA* haplotypes of HOM2 are not IBD, despite the cell line being listed as consanguineous (Supplementary Figure 1). We show similarly that the PGF cell line is heterozygous through the first ~500kb of the targeted region, as previously suspected (Stewart et al. 2004). Four of the cell lines (AWELLS, LKT3, MT14B and SPL) are fully homozygous through the *KIR* region of chromosome 19q13.4, and PF04015 and T7526 have >50% homozygosity of *KIR* (not shown).

The Hapmap II project mapped 29 recombination hotspots to the *HLA* region (Myers et al. 2005). We determined that 52% of the homozygous tract breakpoints in the *HLA* region of the IHWG cells occur within one of these recombination hotspots (Fig. 5). This finding is consistent with a genome-wide average of 50–60% recombination events occurring within hotspots (Myers et al. 2005; Coop et al. 2008). Within the *HLA* region the most frequent haplotype breakpoints of the IHWG cell lines overlap with the previously identified (Jeffreys et al. 2001; Cullen et al. 2002; Stenzel et al. 2004) hotspots, *GABRI* (29571319-29600162), *TAP2* (32790065-32806010), *DMB* (32902747-32908584) and *DOA* (32974614-32977313) (Fig. 5). A further understanding of the characteristics and commonalities of these hotspots is likely to be revealed through complete sequencing of the surrounding regions.

Whole-genome SNP arrays can provide a low resolution scan of genomic structural variation. By virtue of the extended homozygosity in the IHWG panel we were able to identify nine large (>35kb) structural variants that associate with specific *HLA* alleles in this panel (Fig. 6). Surprisingly, the presence of *HLA-Y* is associated with two distinctive patterns of structural variation. Both of these occur near the predicted site of the ~70kb insertion that harbours this gene (telomeric of *HLA-A* and *-G*), but one pattern associates with *HLA-A**31/33 alleles and the other with *A**30:01/68:02 alleles (Fig. 6). This observation suggests presence of two divergent types of *HLA-Y* segment insertion in the *HLA* region. Other major structural variants we detected were associated with *A**03 and *DRB1**15/16 (Fig. 6). Although the method employed here measures structural variation at very low resolution, this knowledge will be beneficial towards full haplotype characterization.

The data presented here will be a valuable resource for several reasons. Extreme polymorphism, structural diversity and high repeat content render haplotype assembly from heterozygous individuals prohibitively difficult with most current sequencing technologies. Thus, the principle goal of mapping the homozygous tracts is to reduce complexity during long-range haplotype assembly and ultimately to provide unambiguously-phased reference sequences for the entire *HLA* region. Another goal is to examine true linkage between

potential disease-causing markers, which can be achieved through analysis of these haplotypes that are identical by descent. Further, because the *HLA class I* and *II* genotypes of the IHWG cell lines have been thoroughly characterized, unambiguous knowledge of marker phase can be used to strengthen or test *HLA* genotype imputation algorithms. Finally, although selected for *HLA* homozygosity, the mean tract length genome-wide in the panel is 20% greater than the matched control set (data not shown), and so the *HLA*-homozygous cells will be informative for more general genetic studies. For these reasons we have made all of the data freely available on ImmPort (SDY295: EXP13576). In summary, the IHWG cell line panel represents a significant proportion of European *HLA* allelic and haplotype diversity. The data we describe will provide a valuable resource for characterizing haplotypes, refining imputation algorithms and developing assay controls.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We extend thanks to the members of the John Hansen laboratory at the Fred Hutchinson Cancer Research Center for supplying some of the (IHWG) cell line/DNA samples. This study was supported by U.S. National Institutes of Health grant U01 AI090905.

References

- Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*. 2008; 319(5868): 1395–1398. [PubMed: 18239090]
- Coquillard G, Lau M, Kletzel M, Rodriguez-Marino SG. Identification of two pseudogenes with sequence homology to human and gorilla MHC class IA genes: ancestral haplotype in the Filipino population. *Hum Immunol*. 2004; 65(6):665–673. [PubMed: 15219387]
- Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am J Hum Genet*. 2002; 71(4): 759–776. [PubMed: 12297984]
- de Bakker PI, Raychaudhuri S. Interrogating the major histocompatibility complex with high-throughput genomics. *Hum Mol Genet*. 2012; 21(R1):R29–36. [PubMed: 22976473]
- Dorak MT, Shao W, Machulla HK, Lobashevsky ES, Tang J, Park MH, Kaslow RA. Conserved extended haplotypes of the major histocompatibility complex: further characterization. *Genes Immun*. 2006; 7(6):450–467. [PubMed: 16791278]
- Gleimer M, Wahl AR, Hickman HD, Abi-Rached L, Norman PJ, Guethlein LA, Hammond JA, Draghi M, Adams EJ, Juo S, Jalili R, Gharizadeh B, Ronaghi M, Garcia KC, Hildebrand WH, Parham P. Although divergent in residues of the peptide binding site, conserved chimpanzee Patr-AL and polymorphic human HLA-A*02 have overlapping peptide-binding repertoires. *J Immunol*. 2011; 186(3):1575–1588. [PubMed: 21209280]
- Gonzalez-Galarza FF, Takeshita LY, Santos EJ, Kempson F, Maia MH, Silva AL, Ghattaoraya GS, Alfievic A, Jones AR, Middleton D. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res*. 2015; 43(Database issue):D784–788. [PubMed: 25414323]
- Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S. Gene map of the extended human MHC. *Nat Rev Genet*. 2004; 5(12):889–899. [PubMed: 15573121]
- Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. *BMC Genomics*. 2013; 14:355. [PubMed: 23714642]

- Illing PT, Vivian JP, Purcell AW, Rossjohn J, McCluskey J. Human leukocyte antigen-associated drug hypersensitivity. *Curr Opin Immunol.* 2013; 25(1):81–89. [PubMed: 23141566]
- Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet.* 2001; 29(2):217–222. [PubMed: 11586303]
- Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* 2014; 42(Database issue):D764–770. [PubMed: 24270787]
- Larsen CE, Alford DR, Trautwein MR, Jalloh YK, Tarnacki JL, Kunnenkeri SK, Fici DA, Yunis EJ, Awdeh ZL, Alper CA. Dominant sequences of human major histocompatibility complex conserved extended haplotypes from HLA-DQA2 to DAXX. *PLoS Genet.* 2014; 10(10):e1004637. [PubMed: 25299700]
- Marsh, SGE.; Packer, R.; Heyes, JM.; Bolton, B.; Fauchet, R.; Charron, D.; Bodmer, JG. The International Histocompatibility Workshop cell panel. In: Charron, D., editor. Genetic diversity of HLA. Functional and medical implications. Vol. 1. Paris: EDK; 1996. p. 26-28.
- Mickelson, E.; Hurley, C.; Ng, J.; Tilanus, M.; Carrington, M.; Marsh, SGE.; Rozemuller, E.; Pei, J.; Rosielle, J.; Voorter, C.; Sayer, D.; Damodaran, A.; Nisperos, C.; HJA. 13th IHWS Shared Resources Joint Report. IHWG Cell and Gene Bank and reference cell panels. In: HJA, editor. Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference. Vol. 1. Seattle: IHWG Press; 2006. p. 523-553.
- Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martinez RJ, Hedges DJ, Morris RW, Eng C, Sandoval K, Acevedo-Acevedo S, Norman PJ, Layrisse Z, Parham P, Martinez-Cruzado JC, Burchard EG, Cuccaro ML, Martin ER, Bustamante CD. Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* 2013; 9(11):e1003925. [PubMed: 24244192]
- Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. *Science.* 2005; 310(5746):321–324. [PubMed: 16224025]
- Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat Rev Immunol.* 2013; 13(2):133–144. [PubMed: 23334245]
- Petersdorf EW. Genetics of graft-versus-host disease: the major histocompatibility complex. *Blood Rev.* 2013; 27(1):1–12. [PubMed: 23182478]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81(3):559–575. [PubMed: 17701901]
- Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* 2015; 43:D423–431. (Database issue). [PubMed: 25414341]
- Sayers EW. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2015; 43(Database issue):D6–D17. [PubMed: 25398906]
- Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. Whole-genome genotyping with the single-base extension assay. *Nat Methods.* 2006; 3(1):31–33. [PubMed: 16369550]
- Stenzel A, Lu T, Koch WA, Hampe J, Guenther SM, De La Vega FM, Krawczak M, Schreiber S. Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Hum Genet.* 2004; 114(4):377–385. [PubMed: 14740295]
- Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coghill P, Dunham I, Forbes S, Halls K, Howson JM, Humphray SJ, Hunt S, Mungall AJ, Osoegawa K, Palmer S, Roberts AN, Rogers J, Sims S, Wang Y, Wilming LG, Elliott JF, de Jong PJ, Sawcer S, Todd JA, Trowsdale J, Beck S. Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res.* 2004; 14(6):1176–1187. [PubMed: 15140828]
- Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM. Evolution and functional

- impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012; 337(6090):64–69. [PubMed: 22604720]
- Terasaki PI. Selection of organ donors. *N Engl J Med*. 1969; 280(23):1304. [PubMed: 4890252]
- Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet*. 2013; 14:301–323. [PubMed: 23875801]
- Watanabe Y, Tokunaga K, Geraghty DE, Tadokoro K, Juji T. Large-scale comparative mapping of the MHC class I region of predominant haplotypes in Japanese. *Immunogenetics*. 1997; 46(2):135–141. [PubMed: 9162100]
- Williams F, Curran MD, Middleton D. Characterisation of a novel HLA-A pseudogene, HLA-BEL, with significant sequence identity with a gorilla MHC class I gene. *Tissue Antigens*. 1999; 54(4): 360–369. [PubMed: 10551419]
- Yang, SY.; Milford, E.; Hammerling, U.; Dupont, B. Description of the Reference Panel of B-Lymphoblastoid Cell Lines for Factors of the HLA system: The B-Cell Line Panel Designed for the Tenth International Histocompatibility Workshop. In: Dupont, B., editor. *Immunobiology of HLA. Histocompatibility Testing*. Vol. 1. New York: Springer-Verlag; 1987. p. 11-19.

Figure 1. The IHWG panel comprises mostly European individuals

Shown is a PCA plot using ~1 million genome wide SNP markers from four HapMap populations (Eur, CHB, JPT, Yoruba) and the IHWG panel. There were no Amerindian populations in HapMap II, thus the five Amerindian IHWG samples form a separate cluster (top centre). AMA1 is North African, MANIKA is South Asian (Indian) and MGAR is Hispanic. Also labelled are the two cells, COX and PGF completed by the MHC sequencing consortium (Stewart et al. 2004). PGF is the hg19 reference sequence for the *HLA* region. The genome-wide data set is available at ImmPort under SDY295: EXP13576 (<https://immport.niaid.nih.gov/>).

	HLA	N	% Eur	% World
alleles	A	24	93.8	82.4
	B	34	91.3	60.1
	C	17	95.7	87.3
haplotypes	A-B-C	75	49.5	
	A-B-C-DRB1	88	23.4	

Figure 2. The IHWG panel represents a significant proportion of European HLA diversity

The distribution of *HLA* alleles and haplotypes present in the IHWG panel was compared with their frequencies available at dbMHC and allelefrequencies.net (Gonzalez-Galarza et al. 2015; Sayers 2015) to estimate the proportion of individuals in Europe and worldwide who possess them.

Figure 3. IHWG cell homozygous tracts that intersect the *HLA* region

Shown are homozygous segments determined from high density whole genome SNP analysis. Blue lines denote the interval between the last observed heterozygous SNP and the first homozygous SNP at each end of the homozygous tracts. Red dashed line indicates the 5Mb *HLA* region and the grey dotted line indicates the 'extended' *HLA* region. The precise genomic coordinates of the homozygous tracts for each of the cell lines are given in Supplementary Figure 2.

Figure 4. *HLA* haplotypes of HOM2 are not identical by descent

SNP diversity was calculated in sliding windows of 100bp through the *HLA* region from four cell lines; EN1NOT (heterozygous), COX (homozygous), PGF (partially homozygous; one haplotype is huRef build 19), and HOM2. HOM2 is homozygous for *HLA*-A, -B, -C, -DP, -DQ and -DR by conventional genotyping methods (Fig. 1). The locations of the classical *HLA* genes are shown in the lower panel.

Figure 5. Recombination hotspots and *HLA* haplotype breakpoints

Top; shows the number of heterozygous haplotypes in the IHWG panel (N=99) at each point through the *HLA* region. Inverted triangles show the location of the 29 recombination hotspots predicted to occur within the *HLA* region (Myers et al. 2005). The triangles are filled in red for hotspots overlapping with breakpoints in the IHWG panel, and the size corresponds to the frequency of use in the panel. Lower; shows the number of cells that become heterozygous at each point (53% of breakpoints overlap with the predicted hotspots).

Figure 6. Structural variants detected in *HLA* homozygous regions

Shown are the approximate coordinates of structural variants that track with specific *HLA* alleles (indicated as the Y axis). Green are likely insertions compared to the hg19 reference haplotype as determined by unexpected heterozygosity in homozygous segments. Red is a likely deletion as determined by a tract of failed SNP genotypes. Presence of distinct patterns that were identical amongst individuals sharing *HLA* alleles at a given locus showed these structural variants associate with the specific alleles.