Title

In-depth genetic analysis of 6p21.3 reveals insights into associations between HLA types and complex traits and disease

Authors

Matteo D'Antonio[1,†], Joaquin Reyna[2†], Agnieszka D'Antonio-Chronowska[1], Marc-Jan Bonder[3], David Jakubosky[4], Hiroko Matsui[1], Erin N. Smith[2], Oliver Stegle[3], Naoki Nariai[2], and Kelly A. Frazer[1,2,*]

[1] Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA, 92093, USA

[2] Department of Pediatrics and Rady Children's Hospital, University of California, San Diego, La Jolla, CA, 92093, USA

[3] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

[4] Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, CA, 92093, USA

* To whom correspondence should be addressed. Tel: +1 (858) 246-0208; Email: kafrazer@ucsd.edu.

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Present Address: Naoki Nariai, Illumina, Inc., 5200 Illumina Way, San Diego, CA 92122, USA

Abstract

The highly polymorphic major histocompatibility (MHC) region encodes the human leucocyte antigen (HLA) gene complex and is associated with many autoimmune and infectious diseases. Despite the importance of this interval, comprehensive genetic studies interrogating associations between HLA types, expression of non-HLA genes and disease, have not yet been conducted. To address this issue, we collected high-coverage whole genome sequence from 419 individuals and performed HLA typing at the highest resolution. Using RNA-seq from matched iPSC lines, we conducted an in-depth eQTL analysis using "personalized" transcripts, which significantly improved estimated expression levels of HLA genes, and showed HLA types have genetic associations independent from SNPs. We leveraged the eQTL results to examine associations between expression levels of non-HLA genes and disease. As a proof-of-principle, we investigated RNF5, whose protein product is a novel drug target in cystic fibrosis. We observed that decreased expression of *RNF5* was associated with the 8.1 ancestral haplotype, which was previously found associated with protection against infection in cystic fibrosis. Overall, our study shows that genetically dissecting the MHC region provides novel insights into mechanisms underlying associations of this interval with disease.

Introduction

The 4 Mb major histocompatibility (MHC) region on chromosome 6p21.3 is highly polymorphic, encodes the human leucocyte antigen (HLA) gene complex, and has been associated through genome-wide association studies (GWAS) with many autoimmune and infectious diseases (Gough and Simmonds 2007; Blackwell et al. 2009; Matzaraki et al. 2017). The extreme polymorphism rate (up to more than ten times higher than the genome average (Norman et al. 2017)), high gene density, and complex linkage disequilibrium (Miretti et al. 2005; Jensen et al. 2017) at this locus makes determining HLA types (the alleles of HLA genes) at high resolution and performing expression quantitative trait loci (eQTL) analyses challenging. The inability to genetically interrogate the MHC region has made it difficult to understand the mechanisms underlying the hundreds of GWAS associations mapping into the interval.

HLA types are determined at varying levels of resolution and have a highly standardized naming system (Marsh et al. 2010). While HLA types have historically been described at two-digit resolution (e.g. HLA-A*01) based on serological reactions, the development of array-based technologies (Levine and Yang 1994; Okada et al. 2015) provided methods to estimate HLA types at four-digit resolution (e.g. HLA-A*01:01), which distinguishes between variants resulting in non-synonymous amino acid sequence differences. More recently, sequencing-based methods (Boegel et al. 2012; Warren et al. 2012; Hosomichi et al. 2013; Kim and Pourmand 2013; Bai et al. 2014; Szolek et al. 2014; Huang et al. 2015; Nariai et al. 2015; Dilthey et al. 2016; Ka et al. 2017; Xie et al. 2017; Lee and Kingsford 2018) improved HLA typing resolution, incorporating synonymous variants (six-digit resolution, e.g. HLA-A*01:01:01) and non-coding regulatory variants (eight-digit resolution, e.g., HLA-A*01:01:01:01). Currently, there are more than 21,000 alleles for all 30 MHC region genes that have more than one allele at eight-digit resolution in the IPD-IMGT/HLA database (Robinson et al. 2016), and more than 14,000 alleles for the six classical HLA genes (HLA-A, -B, -C, -DQA, -DQB, and -DRB) that are routinely used for HLA typing in clinical settings. Given that known alleles can share highly similar sequences, with many alleles differing by a base-pair substitution, it is challenging to correctly assign an individual's HLA types using whole genome sequencing (WGS) data. Several computational approaches (Boegel et al. 2012; Warren et al. 2012; Hosomichi et al. 2013; Kim and Pourmand 2013; Bai et al. 2014; Szolek et al. 2014; Huang et al. 2015; Nariai et al. 2015; Dilthey et al. 2016; Ka et al. 2017; Xie et al. 2017; Lee and Kingsford 2018) have been developed to predict HLA types at eight-digit resolution by using WGS data and reference HLA sequences in the IPD-IMGT/HLA database (Robinson et al. 2016). These computational approaches have been evaluated using small numbers of samples, shallow (average 7X coverage) WGS data and/or limited to the six classical HLA genes (Bai et al. 2014; Huang et al. 2015; Bauer et al. 2016; Ka et al. 2017; Xie et al. 2017; Lee and Kingsford 2018). Hence, accuracy and effectiveness of HLA typing using deep WGS data for hundreds of individuals across all 30 MHC region genes in the IPD-IMGT/HLA database at eight-digit resolution has not been fully investigated yet. Furthermore, how genetic regulatory variation included in eight-digit resolution naming influences the expression of different HLA types has not been fully explored.

While the mechanisms underlying the majority of associations between the MHC region and disease are unknown, there are two competing hypotheses that explain why this interval is associated with hundreds of different traits. One hypothesis proposes that HLA gene associations are due to immune response towards self or foreign antigens ("altered self-antigen")

(Oldstone 1998; Yin et al. 2013; Klein et al. 2014). In support of this hypothesis, several studies have found that autoimmune diseases such as rheumatoid arthritis are associated with the presence of epitopes present in specific HLA types (Gregersen et al. 1987; Bodis et al. 2018; Okada et al. 2018). A second hypothesis proposes that the HLA gene associations with complex traits are proxies for a non-HLA gene ("mistaken identity") (Holoshitz 2013). Previous studies have shown that the expression of both HLA and non-HLA genes are associated with the genotypes of disease-associated variants present in the MHC region (Fehrmann et al. 2011; Dendrou et al. 2018), suggesting that the "mistaken identity" hypothesis could underlie some of the genetic associations. Altogether, these studies suggest that there are complex interactions between genetic variation, gene function of both HLA and non-HLA genes, and disease, indicating that a combination of both the "mistaken identity" and the "altered self-antigen" hypotheses may explain the hundreds of associations between the MHC region and human GWAS traits and disease.

At present, a comprehensive genetic study of the MHC region using HLA types at eight-digit resolution has not yet been conducted. To address this gap, we used deep WGS from 419 individuals to call eight-digit HLA types and RNA-seq data from 361 matched induced pluripotent stem cells (iPSCs) to investigate associations between SNPs, HLA types, expression of non-HLA genes and disease. We observed that common two-digit HLA types frequently were subdivided into multiple HLA types at eight-digit resolution and that using "personalized" transcript sequences corresponding to the HLA types detected in the individual resulted in improved accuracy of estimated HLA gene expression levels. We found that HLA types have genetic associations independent of SNPs, genes in the MHC region are significantly enriched for having their expression levels modulated by regulatory variants, and that the MHC region contains four groups of genes (both HLA and non-HLA) that have alleles with highly correlated expression. Notably, we observed that the expression levels of 65 genes (14 HLA genes and 51 non-HLA genes) were associated with genetic variants involved in complex traits and disease. Finally, we leveraged the results of our in-depth eQTL analysis to identify putative mechanisms of non-HLA genes underlying disease associations in the MHC region. Using this approach, we showed that the 8.1 ancestral haplotype (8.1AH), which spans the whole MHC region and is known to be associated with protection against infection in Cystic Fibrosis (CF) patients (Laki et al. 2006), is associated with decreased expression of *RNF5*. The protein encoded by *RNF5* is a drug target for patients with the F508del mutation in the CFTR gene, because its downregulation stabilizes the CFTR protein, thereby rescuing its function (Tomati et al. 2015; Sondo et al. 2018). Taken together, our study shows the importance of genetically dissecting associations in the 6p21.3 interval to gain insights into molecular mechanisms underlying complex traits and diseases and provides a resource for further genetic investigation of the MHC region and the potential identification of novel therapeutic targets.

Results

We analyzed a total of 419 individuals, of which 273 (152 females and 121 males) were from the iPSCORE resource (Panopoulos et al. 2017) and 146 (85 females and 61 males) were from the HipSci resource (Kilpinen et al. 2017) (Table S1). Of the 419 individuals, 80% (336) were of European ancestry, while the remaining were of diverse ethnic backgrounds (Asian (30), Multiple ethnicities (20), Hispanic (18), African American (7), Indian (5), and Middle Eastern

(3)) and thus harbored HLA alleles from multiple human populations (Table S1). While the HipSci individuals were all genetically unrelated to one another, the iPSCORE individuals comprised 56 families containing between 2 and 14 members, and included 25 monozygotic twin pairs and 17 non-overlapping trios. In total, there were 311 genetically unrelated individuals. Whole-genome sequence (WGS) data generated from fibroblasts were available for all 419 individuals; additionally, for 59 of the HipSci individuals WGS was available from one iPSC clone, and 84 HipSci individuals had WGS from two iPSC clones (504 clones in total). The family structure and the presence of multiple WGS data from the same individual, enabled us to benchmark HLA typing performance based on concordance in twin pairs, concordance between fibroblast-iPSC pairs, and Mendelian inheritance in trios.

High quality WGS spanning the MHC region

Prior to conducting HLA typing, we assessed the quality of the WGS in the HLA region (chr6:29640168-33115544) by determining SNP density in consecutive 10-kb windows across the region. The SNP density in the WGS data varied widely: 1) for the 273 iPSCORE genomes from 1 SNP/10kb to 900 SNPs/10kb (mean = 149.2 SNPs/10kb) compared with the genome average of 79.6 SNPs/10kb Figure 1A); and 2) for the 377 HipSci genomes (146 fibroblast samples and 231 iPSCs) from 1 SNP/10kb to 995 SNPs/10kb (mean = 155.4 SNPs/10kb) compared with a genome-wide average of 58.1 SNPs/10kb. The region with the highest SNP density contained the *HLA-DQA1* and *HLA-DQB1* genes, which is consistent with previous studies(Norman et al. 2017). To determine whether the high SNP density interfered with read alignment, we analyzed the coverage depth across the region. We observed that the coverage (mean 49.8 ± 12.1 for iPSCORE and 37.9 ± 7.1 for HipSci), overall, was comparable with the genome average (52X and 38X, respectively, Figure 1B). We found, on average, 98.1% and 96.5% of the MHC region had a read depth >20X, and for each iPSCORE and HipSci sample only 284 ± 122 bp and 252 ± 115 respectively were not covered by any reads. These findings showed that all 650 WGS samples in the iPSCORE and HipSci collections were of high quality.

Eight-digit HLA typing from WGS data achieves high recall rate and accuracy

We performed HLA typing on 30 genes in the MHC region, including six HLA class I genes (HLA-A, -B, C, -E, -F, -G), eight HLA Class I pseudogenes (HLA-H, -J, -K, -L, -P, -T, -V, -W), twelve HLA class II genes (HLA-DMA, -DMB, -DOA, -DOB, -DPA1, -DPA2, -DPB1, -DBP2, -DQA1, -DQB1, -DRA, -DRB1) and four non-HLA genes (MICA, MICB, TAP1, and TAP2), all which had more than one allele at eight-digit resolution in the IPD-IMGT/HLA database (release 3.30.0) (Robinson et al. 2016). We used HLA-VBSeq(Nariai et al. 2015) to estimate HLA type at eight-digit resolution in the WGS of the 650 samples (Table S2). We also obtained HLA types at lower-digit resolutions (six, four, and two-digit) by removing high digit values from the 8-digit resolution HLA types.

We initially calculated recall rate, measured as the fraction of individuals that could be HLA-typed for each HLA gene. We observed a high recall rate for all 30 genes (mean = 98.5% for both iPSCORE and HipSci samples, Figures 1C,D) and for only three HLA genes in iPSCORE (*HLA-H, HLA-T* and *HLA-K*) and four genes in HipSci (*HLA-H, HLA-T, HLA-K* and *HLA-DRB1*) fewer than 95% of individuals were HLA-typed. The fraction of heterozygous and homozygous HLA

types across the 30 genes was highly similar in the iPSCORE and HipSci individuals (r = 0.92). Across the 30 genes, we observed that the fraction of homozygous HLA types was highly variable and negatively correlated with the number of alleles associated with each gene (r = -0.75 and -0.71, iPSCORE and HipSci samples, respectively). Indeed, the most polymorphic genes (*HLA-A, HLA-B, HLA-C, HLA-DRB1* and *HLA-DBP1*) were homozygous in less than 20% individuals, whereas *HLA-V*, which only has three alleles (V*01:01:01:01, V*01:01:01:02 and V*01:01:01:03, with allele frequency = 0.69, 0.20 and 0.11, respectively, Table S2), was homozygous in more than 45% individuals. These results show that we were able to obtain high recall rates for all 30 genes and that for each gene the fraction of homozygous and heterozygous alleles was consistent with the number of alleles and their frequencies.

To examine the accuracy of the high-resolution HLA types, we determined HLA type concordance across 25 monozygotic twin pairs (defined as twin concordance), Mendelian inheritance concordance from 17 non-overlapping trios and HLA type concordance of 231 fibroblast-iPSC pairs (defined as HipSci concordance). Overall, we found that the median concordance across all genes was very high; however, both the twin pairs (96.7%) and fibroblast-iPSC pairs (95.7%) had slightly higher concordance than the Mendelian inheritance (90.9%) Figures 1E,F,G). These results suggest that HLA-VBSeq produced highly reproducible results when samples with exactly the same two HLA alleles were analyzed; and had slightly lower accuracy calling the same HLA alleles when in different diplotypes. We also analyzed concordance at the derived lower digit resolutions (six, four, and two-digit) and observed that two-digit resolution had the highest concordance for the twin pairs (98.5%), Mendelian inheritance (96.9%) and fibroblast-iPSC pairs (99.2%) (Figure S1). Thus, even when HLA-VBSeq assigned discordant HLA alleles at eight-digit resolution, the vast majority of the time the assigned allele frequently was correct at the two-digit HLA type. These data show that the WGS read depth coverage of both the iPSCORE and HipSci samples were sufficient for HLA-VBSeq(Nariai et al. 2015) to detect eight-digit resolution HLA alleles at a high recall rate and with high accuracy.

Greater number of predicted HLA types at eight-digit resolution

For each HLA gene, we determined the number of unique alleles and their count at two-, four- and eight-digit resolution in the 273 individuals in iPSCORE and 146 HipSci individuals. To determine if the HLA allele frequencies in the iPSCORE and HipSci individuals were representative of what would be observed in a larger population, we initially conducted a comparative analysis with HLA alleles in 3.5 million individuals with diverse genetic backgrounds from The Allele Frequency Net Database (AFND)(Gonzalez-Galarza et al. 2015) (see Methods). We observed high correlation between HLA allele frequencies in AFND and iPSCORE (r = 0.921) and HipSci (r = 0.908) (Figure 2A). These data show that HLA-VBSeq identified a similar number of alleles for each of the 30 genes in the iPSCORE and HipSci individuals, and that both sets of diverse HLA alleles were representative of the human population.

We next investigated the relative number of alleles and their frequencies for each of the 30 genes at the two- and eight-digit resolutions (Table S2, Figure 2B-K, Figure S2). We found that genes with few alleles at two-digit resolution had the largest fold increase in the number of alleles at eight-digit resolution (i.e. the number of alleles at two-digit was negatively correlated with the fold increase at eight-digit: r = -0.32). This observation was due to the fact that common HLA types at

two-digit resolution (>5% allele frequency) were three times more likely to be resolved into a higher number of eight-digit resolution alleles than rare HLA types (5.2 and 1.8 eight-digit HLA types per two-digit HLA type, respectively for common and rare alleles, $p = 1.2 \times 10^{-13}$, Mann-Whitney U test). At two-digit resolution, 13 genes had only one allele, but their number of alleles increased three- (*HLA-V*) to more than ten-fold at eight-digit resolution (*HLA-E*, *F* and *G*, Figure 2B-E). For other genes, such as *MICA*, *MICB* and *HLA-W*, only one or two common two-digit alleles resolved into a higher number of eight-digit alleles (Figure 2F-H), resulting in a doubling of the number of alleles between the two resolutions. The *HLA-A*, *HLA-B* and *HLA-DPB1* genes showed the greatest diversity both at two- and eight-digit resolutions, and we observed for all three genes a 2.7-fold increase in the number of alleles between the two resolutions (Figure 2I-K). Our observations are consistent with the fact that *HLA-A* and *HLA-B* are the most polymorphic genes in humans (~3,900 and ~4,700 described alleles, respectively) (Mungall et al. 2003; Kennedy et al. 2017). These findings show that rare two-digit alleles tend to map directly to rare eight-digit alleles, while common two-digit alleles frequently are subdivided into multiple eight-digit alleles, resulting in greater allele heterogeneity at the higher resolution.

Personalized transcripts improve predicted gene expression levels

Given that the MHC region is the most SNP-dense region in the human genome (Figure 1A), the expression levels of genes in this region measured by RNA-seq may be inaccurate due to large numbers of mismatches between the sequence reads and reference sequence. To overcome this issue, we calculated gene expression levels in the 446 iPSCs (including 215 iPSCORE and 231 HipSci from 146 individuals) using "personalized" transcript sequences (see Methods) corresponding to the HLA types detected in the individual using HLA-VBSeq (Table S3). We found that 19 of the 30 MHC genes were expressed in iPSCs (TPM >2 in at least 10 samples, Table S3). The HLA class I genes (HLA-A, B, C and E, Figure 3A) were highly expressed, consistent with their ubiquitous expression in all cell types, while the HLA class II genes were expressed at lower levels (Figure 3B), as expected due to their primary role in immune cells (Matzaraki et al. 2017). To determine whether using personalized transcripts improved predicted gene expression levels, we tested the differences between reference and personalized TPM and found that their correlation ranged between 0.448 (*TAP2*) and 0.872 (*MICB*, Figure 3C,D, Figure S3), suggesting that a substantial portion of the gene expression heterogeneity captured using personalized transcripts cannot be detected using reference transcripts ($1 - R^2$ ranged between 0.200 and 0.760). The HLA class I genes, which are expressed ubiquitously, tended to be expressed significantly higher using personalized transcripts (Figure 3E, Figure S3), whereas most class II genes, which are expressed mostly in immune cells, were expressed significantly lower in the iPSCs using the personalized sequences (Figure 3F, Figure S3). For instance, we observed that in many samples the expression of *HLA-C* was low (<40 TPM) using reference transcripts, but it increased more than four-fold using personalized transcripts (Figure 3E); while *HLA-DPB2* expression decreased in 98% of samples. These results show that using personalized expression levels of HLA genes improves the accuracy of estimated gene expression levels (Figure 3F).

To examine if HLA types were associated with allele-specific gene expression levels, we compared the expression levels of each allele against all other alleles from the same HLA gene (Figure 3G,H). We quantile-normalized expression of each

gene and compared the TPM distributions between all samples carrying a specific HLA allele and all the other samples. Alleles from all 19 expressed genes showed high variability (Figure 3G,H). We observed high variability in the expression levels of the different alleles, with 118 of the 346 HLA types (34.1%) present in at least two individuals showing a significantly different expression level than the samples not carrying the HLA type (t-test, FDR <5%) (Table S4). (Figure 3G,H). Several genes, such as *HLA-C*, *HLA-DRB1* and *HLA-DQB1* showed more than five-fold differences between the least expressed and the most expressed alleles. Interestingly, we observed that, of the 33 tested eight-digit *HLA-C* types, four were overexpressed all of which map to the same two-digit HLA-C*07 allele that has been previously associated with protection against human cytomegalovirus (Schlott et al. 2018) (Figure 3G). Conversely, we found that of nine eight-digit *HLA-DRB1* types that map to the same two-digit HLA-DRB1*04 allele, four were expressed at significantly lower levels than the other *HLA-DRB1* types (Figure 3H). The HLA-DRB1*04 type has been associated with tuberculosis susceptibility (Souza de Lima et al. 2016) and perhaps the association strength would increase if eight-digit HLA types were considered. Overall, eight-digit HLA-types for nine of the expressed MHC genes, were strongly associated with gene expression differences, suggesting that the associations between HLA types and disease may be driven not only by different binding affinities to specific peptides, but also by different expression levels. Therefore, both non-coding and coding genetic variation in the MHC region may significantly impact gene function and disease susceptibility.

Genes in the MHC region are enriched for SNP-eQTLs

We performed a comprehensive eQTL analysis of the MHC region using estimated gene expression levels for the HLA types calculated using personalized transcripts. Of the 383 genes located in the MHC region, we identified 146 (including the 19 genes analyzed above; 15 HLA genes and four non-HLA genes) expressed in iPSCs. We detected associations between the genotype of 83,969 SNPs with MAF >1% located in the MHC region and the expression of each of these 146 genes, using Matrix eQTL (Shabalin 2012). We found 83 genes (eGenes; 15 HLA genes and 68 non-HLA genes) with 81,883 significant eQTLs (32,629 distinct SNPs, referred to as SNP-eQTLs hereafter; Bonferroni-corrected p-value < 0.05, Table S5). Due to the low recombination rate (Trowsdale and Knight 2013; DeGiorgio et al. 2014) and complex LD structure (Miretti et al. 2005; Jensen et al. 2017) in this interval, we initially investigated the distributions of all SNP-eQTLs rather than solely focusing on the top variants. While eGenes on average had 283 SNP-eQTLs, there were 17 eGenes with ten or fewer, and 26 eGenes that had more than 1,000 (Figure 4). In general, SNP-eQTLs were centered around the transcription start site (TSS) of each eGene (mean distance = 33.8 kb, Figure S4), and for 19 eGenes (22.9%), the top SNP-eQTL was localized within 5 kb of their TSS. However, for 12 eGenes (14.5%) the closest SNP-eQTL was more than 500 kb from the TSS and for several eGenes, including *ZNRD1*, *STK19P*, *RNF5* and *HLA-DPB2*, associated SNP-eQTLs spanned the entire MHC region. To identify independent SNP-eQTLs, we repeated the analysis conditioning gene expression on the top SNP-eQTL for each of the 83 eGenes. For 41 eGenes, we observed secondary SNP-eQTLs (conditional on the top SNP-eQTL) and for 15 eGenes we identified a tertiary SNP-eQTL upon conditioning on the top two independent SNP-eQTLs (Figure 4, Table S5). These results show that, within the gene-dense MHC region, a large fraction (83; 56.9%) of the expressed genes are eGenes, many of which (41; 49.4%) are associated with multiple independent regulatory variants. We compared these findings with those of a previous genome-wide eQTL study we

conducted on 215 iPSCORE iPSCs (DeBoever et al. 2017) where we identified significantly fewer genes as eGenes (32.3%; $p = 4.7 \times 10^{-10}$, chi-squared test), of which only a minority (709 out of 5,677 eGene; 12.5%) had multiple regulatory variants. Altogether, our SNP-eQTL analysis shows that both HLA and non-HLA genes in the MHC region are significantly enriched for having their expression levels modulated by regulatory variants.

Genes in MHC region have alleles with highly correlated expression

Given that different HLA types at eight-digit resolution represent a haplotype of many regulatory and coding SNPs rather than a single SNP, we examined if using the 346 HLA types to conduct an eQTL analysis in the MHC region would detect regulatory signals independent from those detected in the SNP-eQTL analysis. To conduct this analysis, we transformed each HLA type into VCF format (see Methods), and calculated associations between each HLA-type at eight-digit resolution and the 146 genes expressed in the HLA region, using Matrix eQTL (Shabalin 2012). After excluding associations between an HLA type and other HLA types of the same gene (Figure 3G, H), we detected 717 associations (referred to as HLA-eQTLs, Bonferroni-corrected p-value < 0.05), between 213 HLA types and 77 genes (eGenes; 15 HLA genes and 62 non-HLA genes, Table S6, Figure 5A). To test if HLA-eQTLs provide independent signals from the SNP-eQTL analysis, we conditioned gene expression on the top SNP-eQTLs for each of the 77 genes. We observed 133 HLA-eQTLs conditional on the top SNP-eQTL, 39 HLA-eQTLs conditional on the top two independent SNP-eQTLs and 31 HLA-eQTLs conditional on the top three independent SNP-eQTLs (10 eGenes; four HLA genes and six non-HLA genes, Figure 5B, Table S6). We observed that associations primarily occurred between HLA-eQTLs and the expression of genes located in their proximity (median distance = 146.2 kb for significant associations and 1.35 Mb for non-significant associations, $p = 2.7 \times 10^{-67}$, Mann-Whitney U test; all self-associations were removed to perform the test). The most significant HLA-eQTLs tended to occur within four distinct genomic regions: 1) chr6: 29640168-30038042, including most of HLA class I genes and HLA class I pseudogenes; 2) chr6:31082526-31629005, including *HLA-B*, *HLA-C*, *MICA* and *MICB*; 3) chr6:32135988-32789609, including most of HLA class II genes, *TAP1* and *TAP2*; and 4) chr6:32916389-33115544, including *HLA-DPA1*, *HLA-DPB1*, *HLA-DPA2* and *HLA-DPB2* (Figure 5A). Two genes (*RNF5* and *C60rf148*) showed strong associations with multiple HLA types in groups 1, 2 and 3. *RNF5* also showed long-range associations with SNP-eQTLs spanning the entire MHC region (Figure 4). These results show that HLA types are associated with the expression of both HLA types of other HLA genes as well as alleles of non-HLA genes in the MHC region independently from SNP-eQTLs, and that the MHC region contains four groups of genes that have alleles with highly correlated expression.

Regulatory variants in the MHC region play important roles in complex traits and disease

We examined the role that regulatory variation plays in the large number of genetic associations between the MHC region and complex traits or disease by intersecting 1,611 independent GWAS hits in the MHC region (Buniello et al. 2019) with the 83,969 SNP-eQTLs. We identified 880 SNP-eQTLs (for 65 eGenes; 13 HLA genes and 52 non-HLA genes) that were also GWAS hits for one or more complex traits or disease. Although eGenes were associated on average with 23 SNP-eQTLs (range: 1-423) that were also GWAS hits (Table S5), five genes (*RNF5*, *HLA-DRB1*, *HLA-DQB1*, *HLA-DRB5* and

HLA-DQA1) were associated with significantly more SNP-eQTLs (range: 101-159) corresponding to 185-423 distinct GWAS hits. These results show that the expression of 78.3% (65/83) of all eGenes (identified with SNP-eQTLs) in the MHC region was associated with complex traits and disease. This was significantly greater than what we observed in our previous genome-wide eQTL study conducted on 215 iPSCORE iPSCs (DeBoever et al. 2017), where we identified 1.04% of all eGenes (59/5,677, $p = 3.6 \times 10^{-100}$, Fisher's exact test) had corresponding eQTLs that were also GWAS hits. These results suggest that the differential expression of both HLA and non-HLA genes in the MHC region may contribute to the complex mechanisms linking this genomic locus with hundreds of human traits and disease.

To investigate if the regulatory variation tagged by HLA-types (HLA-eQTLs) contributed to some of the GWAS hits in the MHC region, we intersected the eGenes from the SNP-eQTL analysis with the eGenes from the HLA-eQTL analysis. In total, there were 95 eGenes (15 HLA genes and 80 non-HLA genes) with SNP-eQTLs or HLA-eQTLs, of which 68.4% (65 eGenes; 14 HLA genes and 51 non-HLA genes) had both (Figure 6A, Table S7). For each of these 65 eGenes, the number of HLA-eQTLs was highly correlated with the number of SNP-eQTLs ($r = 0.692$, $p = 1.7 \times 10^{-10}$, Figure 6B). The majority of the eGenes (56, 86.2%) with both SNP-eQTLs and HLA-eQTLs were associated with at least one GWAS hit, of which 35 (59.3%) were non-HLA genes (Table S7). In comparison, only 15 (50%) of eGenes with only SNP-eQTLs were associated with GWAS hits, indicating that eGenes with both SNP-eQTLs and HLA-eQTLs were enriched for being associated with disease ($p = 0.0024$, Fisher's exact test).

Evidence for differential expression of non-HLA genes underlying GWAS signals

Given that we identified GWAS regulatory variants and HLA types as eQTLs for 35 non-HLA genes in the MHC region, we scanned for examples in which associations between complex traits and disease previously attributed to HLA types were likely to be due to the differential expression of a non-HLA gene. Out of the 35 non-HLA eGenes, *RNF5* was the most compelling candidate as it was associated with 3,165 SNP-eQTLs, 279 GWAS hits (150 traits) and the most HLA-eQTLs (30 HLA types from 22 distinct HLA genes, Figure 6B-D, Table S7). Of the 30 HLA-eQTLs, six were HLA types that comprise the 8.1 ancestral haplotype (8.1AH: Figure 6E-I, Figure S5), which spans more than 4 Mb of the MHC region and is known to be associated with many immune system-related diseases (Gambino et al. 2018) and with delayed colonization in cystic fibrosis (CF) (Laki et al. 2006). *RNF5* had two independent SNP-eQTL signals (Figure 4), one in the proximity of the TSS, and the other, which spanned most of the MHC region, and was also conditionally associated with the 22 distinct HLA-eQTLs (Figures 5, 6C). All six HLA types comprising 8.1AH are associated with significantly decreased expression of *RNF5* (Figure 6E-I, Table S6). It has been proposed that the association between 8.1AH and delayed bacterial colonization in CF patients could be due to the impact of the ancestral HLA-types on the microbiota composition in the lungs (Laki et al. 2006). However, our findings combined with the fact that RNF5 is a current drug target in CF patients because its downregulation contributes to stabilizing and rescuing the function of mutant CFTR proteins (Tomati et al. 2015; Sondo et al. 2018), suggests that the association between 8.1AH CF carriers and delayed bacterial colonization could be due to decreased expression of *RNF5* (Pier et al. 1996; Lyczak et al. 2002; Mall and Hartl 2014; McNicholas 2017). In this proposed model, CF patients carrying 8.1AH express *RNF5* at lower levels than non-

carriers, this results in lower protein levels of RNF5 and thereby less degradation of the misfolded mutated CFTR protein, improved Cl^- secretion, lower mucus secretion and delayed colonization by *S. aureus* and *P. aeruginosa* (Figure 6J). While our findings do not preclude that HLA types comprising 8.1AH play a direct role in the protection against colonization in CF, they suggest that reduced *RNF5* expression plays a major factor in the association between 8.1AH and colonization.

Discussion

We used deep whole-genome sequencing data from 419 individuals, a state-of-the-art computational algorithm, HLA-VBSeq (Nariai et al. 2015), and the comprehensive IPD-IMGT/HLA database (Robinson et al. 2016) to construct a large panel of HLA types at eight-digit resolution for 30 genes in the MHC region. By calculating twin, Mendelian inheritance and fibroblast-iPSC concordance, we showed that eight-digit HLA types were accurately predicted using the high-coverage WGS data (96.7%, 90.9% and 95.7%, respectively), which are comparable with other state-of-the-art methods, such as Kourami (94.7% accuracy) (Lee and Kingsford 2018). Furthermore, while several HLA genes, such as *HLA-V*, *E*, *F* and *G* do not have multiple HLA types at two-digit resolution, we were able to detect multiple alleles at eight-digit resolution for all 30 genes and to resolve common two-digit HLA types into multiple eight-digit alleles, showing a greater allele heterogeneity at the higher resolution.

Since the human MHC region on 6p21 has been typically excluded from QTL studies, due to its high SNP frequency and complex LD structure, we investigated whether combining eight-digit HLA types with "personalized" expression levels of HLA genes could enable genetic association studies. We found that using personalized expression levels of HLA genes resulted in improved estimates of allele-specific expression levels and the identification of HLA genes whose alleles have significantly different expression levels. The personalized HLA gene expression levels also enabled an in depth eQTL study of the MHC region which showed that: 1) HLA types have genetic associations independent of SNPs; 2) genes in this interval are significantly enriched for having their expression levels modulated by regulatory variants; and 2) within the interval there are four groups of genes that have alleles with highly correlated expression levels. Our findings suggest that the associations between HLA types and complex traits and disease are likely driven not only by differential binding affinities to specific peptides, but also by differential expression levels of both HLA and non-HLA genes in the MHC region.

As a concrete example of this point, we determined that decreased expression of *RNF5* was associated with six HLA types comprising the 8.1 ancestral haplotype (8.1AH), which previously has been associated with delayed colonization in cystic fibrosis (CF) (Laki et al. 2006). Interestingly, RNF5 was recently described as a potential novel drug target in CF, because, when RNF5 activity is decreased, the mutant CFTR protein becomes stabilized and its function as a cAMP-dependent chloride channel in the lung epithelium is partially rescued (Tomati et al. 2015; Sondo et al. 2018). Impaired ion transport mediated by CFTR is associated with reduced airway hydration and decreased mucociliary clearance, which leads to increased vulnerability to bacterial infection (Munder and Tummler 2015; Dechecchi et al. 2018). Hence, we

postulate that downregulation of *RNF5* expression due to the presence of specific regulatory variants in 8.1AH is the likely mechanism underlying delayed colonization by *S. aureus* and *P. aeruginosa*. Taken together, our study shows that understanding the associations between SNPs, HLA types and proximal phenotypes in the 6p21.3 region can lead to the identification of causal mechanisms underlying disease associations, and potentially novel drug targets.

Methods

Whole-genome sequencing data

iPSCORE Resource: Whole genome sequencing (WGS) was performed for 273 individuals at Human Longevity, Inc. (HLI) as described in DeBoever et al. (DeBoever et al. 2017). Briefly, DNA was isolated from blood or skin fibroblasts (254 and 19 samples, respectively) using DNEasy Blood & Tissue Kit. A Covaris LE220 instrument was used to quantify, normalize and shear the DNA samples, which were then normalized to 1 µg, DNA libraries were made using the Illumina TruSeq Nano DNA HT kit and their size and concentration determined using LabChip DX Touch (Perkin Elmer) and Quant-iT (Life Technologies), respectively. We normalized all concentrations to 2-3.5 nM, prepared combined pools of six samples and sequenced them at 150 bp paired-end on Illumina HiSeqX. WGS data was generated at an average of 1.1 billion paired-end reads (depth of coverage: 52X) per sample (range 720 million to 3.1 billion). WGS reads were aligned to the reference genome (GRCh37/hg19) including decoy sequences (hs37d5) (Auton et al. 2015) using BWA-MEM version 0.7.12(Li and Durbin 2010). Fastq file quality was estimated using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). All BAM files were processed using GATK best practices(Van der Auwera et al. 2013; Zhang et al. 2017), as described in DeBoever et al. (DeBoever et al. 2017) to detect single nucleotide polymorphisms (SNPs) genome-wide. The genome-wide SNP density (79 SNPs per 10-kb) was calculated by dividing the total number of SNPs in the 273 iPSCORE individuals (20,500,225) by the length of the genome (2,900,434,419, excluding undefined "N" nucleotides).

HipSci Resource: WGS was performed for 146 individuals and 231 associated iPSC lines. Fibroblasts obtained from each HipSci individual were reprogrammed and between one to two iPSC lines were derived. We downloaded CRAM files for the 377 samples (146 fibroblast samples and 231 iPSCs) from the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena, study PRJEB15299) and transformed to BAM files using Samtools(Li et al. 2009a). The WGS data was an average of 835 million paired-end reads (depth of coverage: 38X) per sample (range 450 million to 1.8 billion). BAM files were processed GATK to detect SNPs on chromosome 6.

HLA typing from whole-genome sequencing data with HLA-VBSeq

HLA-VBSeq(Nariai et al. 2015) estimates the most probable HLA types from WGS data at eight-digit resolution by simultaneously optimizing read alignments to a database of HLA type sequences and the abundance of reads on the HLA types by variational Bayesian inference. HLA typing was carried out as follows. First, we identified 30 target genes that

had more than one allele at eight-digit resolution in the IMGT/HLA database (release 3.30.0) (Robinson et al. 2016), which included six HLA class I genes (HLA-A, -B, C, -E, -F, -G), eight HLA Class I pseudogenes (HLA-H, -J, -K, -L, -P, -T, -V, -W), twelve HLA class II genes (HLA-DMA, -DMB, -DOA, -DOB, -DPA1, -DPA2, -DPB1, -DBP2, -DQA1, -DQB1, -DRA, -DRB1) and four non-HLA genes (MICA, MICB, TAP1, TAP2). Second, for each of the 650 samples (273 iPSCORE and 377 HipSci), reads which aligned to the target genes were extracted from each BAM file using coordinates from Gencode v19 (https://www.gencodegenes.org/releases/19.html) and SAMtools version 1.2 (Li et al. 2009a). Third, for each sample the extracted reads were re-aligned to the collection of all genomic HLA sequences in the IPD-IMGT/HLA database (release 3.30.0)(Robinson et al. 2016), allowing each read to be aligned to multiple reference sequences using the "-a" option in BWA-MEM. The expected read counts for each HLA type were obtained with HLA-VBSeq software (http://nagasakilab.csml.org/hla/). For each HLA gene, only the alleles with mean coverage ≥ 20% of the average coverage calculated over the whole genome were considered, and a target HLA genotype was determined as follows: 1) If no allele passed the threshold, then there were not enough reads aligned to correctly identify an HLA type, and hence no alleles were called; 2) If there was only one allele that passed the threshold, and the depth of coverage was two or more times greater than that of the threshold, then the HLA locus was considered to be homozygous for that HLA allele; 3) If there was only one allele that passed the threshold, and the depth of coverage was less than twice that of the threshold, then the allele was called heterozygous with the second allele not determined; 4) If there were two or more alleles that passed the threshold, the alleles were sorted based on the depth of coverage (from high to low), if the depth of coverage of the top allele was more than twice as that of the second one, then the HLA locus was called homozygous for the top allele; 5) If there were two or more alleles that passed the threshold, the alleles were sorted based on the depth of coverage (from high to low), if the depth of coverage of the top allele was less than twice that of the second one, then the two alleles with the highest coverage were selected as the HLA diplotype. For HLA types of lower-digit resolutions (two, four, and six) the eight-digit resolution was converted by removing high digit values.

Determining recall, twin concordance, Mendelian inheritance concordance and fibroblast-iPSC concordance of HLA type detection

Accuracy of HLA typing for each gene at eight-digit resolution was assessed by calculating: 1) recall, the fraction of determined HLA types for each gene; 2) HLA type concordance in 25 monozygotic twin pairs in the iPSCORE resource; 3) Mendelian inheritance concordance across 17 trios in the iPSCORE resource; and 4) HLA type concordance in 231 fibroblast-iPSC pairs. The recall was calculated independently for the 273 iPSCORE samples and the 377 HipSci samples as the number of determined HLA types divided by the total number of samples. To further analyze recall for all genes, we calculated the number of HLA types for each gene and determined the correlation between the number of HLA types and fraction of homozygous individuals. For each gene, we calculated concordance as the fraction of HLA types that matched in the 25 monozygotic twin pairs in iPSCORE or the 231 fibroblast-iPSC paired genomes from the same individual in the HipSci resource. We calculated Mendelian inheritance concordance for each gene as the percentage of HLA types segregating in non-overlapping trios (i.e. in families with multiple trios, each individual was only used once

for this analysis). If one or more HLA types were undetermined for a given pair/trio, we excluded the pair/trio during the concordance calculation.

Analyzing HLA type frequency

We initially determined which HLA type resolution to use to determine if the HLA frequencies we observed were representative of those present in diverse human populations. In the Allele Frequency Net Database (AFND) (Gonzalez-Galarza et al. 2015), 3,556,301 people were genotyped for 12 out of the 30 genes we examined in the MHC region at two-digit resolution, 3,469,268 people (17 genes) at four-digit, 124,721 people (14 genes) at six-digit, and 10,212 people (7 genes) at eight-digit, which would result in testing 226 alleles at two-digit resolution (115 iPSCORE and 101 HipSci), 395 alleles at four-digit resolution (222 iPSCORE and 173 HipSci), 321 alleles at six-digit resolution (174 iPSCORE and 147 HipSci), and 42 alleles at eight-digit resolution (21 iPSCORE and 21 HipSci). We conducted the allele frequency comparative analysis using four-digit resolution to maximize the number of individuals in AFND, the number of genes and the number of alleles. The allele frequency of each HLA type in iPSCORE, HipSci and in the AFND was calculated by dividing the total number of individuals containing the given HLA type by the total number of people in each cohort and a correlation value was calculated after fitting a linear model.

Estimating personalized transcript expression levels of HLA genes

We estimated the expression level of each HLA type using iPSC samples from both the iPSCORE and HipSci resources. For 215 of the iPSCORE individuals, fibroblasts were reprogrammed, an iPSC clone obtained and RNA-seq data generated and processed as described in DeBoever et al. (DeBoever et al. 2017). For the 231 iPSC samples in HipSci, RNA-seq data was downloaded from ENA and processed using the same pipeline used for the iPSCORE iPSC RNA-seq data. The IMGT-IPD HLA database contained cDNA sequences corresponding to each allele at eight-digit resolution for 27 of the 30 genes including: six HLA class I genes (HLA-A, -B, C, -E, -F, -G), five HLA Class I pseudogenes (HLA-H, -J, -K, -L, -V,), twelve HLA class II genes (HLA-DMA, -DMB, -DOA, -DOB, -DPA1, -DPA2, -DPB1, -DBP2, -DQA1, -DQB1, -DRA, -DRB1) and four non-HLA genes (MICA, MICB, TAP1, TAP2). Of these 27 genes, 15 HLA genes and four non-HLA were expressed in at least 10 of the 446 iPSCs (215 iPSCORE and 231 HipSci, TPM > 2). For each individual, we estimated allele-specific expression by replacing the canonical cDNA reference sequences of the 19 expressed genes in Gencode v19 with personalized sequences corresponding to the HLA types detected in the individual using HLA-VBSeq (Nariai et al. 2015). This allowed us to build a set of personalized cDNA reference sequences, which were not affected by the large number of SNPs in the MHC region. cDNA sequences for each HLA type were retrieved from IPD-IMGT/HLA database release 3.30.0(Robinson et al. 2016). RSEM version 1.2.20 (Li and Dewey 2011) was used to quantify transcript abundance and to calculate transcripts per million (TPM) for each HLA type.

SNP-eQTL detection

For each individual with iPSCs and gene expression data, only one single sample was used (361 samples in total: 215 iPSCORE and 146 HipSci, Table S7). We obtained 83,969 SNPs with MAF >1% in the MHC region (and the surrounding 1 Mb) using bcftools (Li et al. 2009b) and decomposed multiallelic genotypes using vt (Tan et al. 2015). Out of the 383 Gencode v19 genes in the MHC locus (chr6:29640168-33115544), eQTL analysis was performed on 146 expressed genes (TPM $\geq$ 2 in at least 10 samples). Expression levels were quantile-normalized across all individuals. To detect SNP-eQTLs, normalized gene expression levels were adjusted for sex, age, batch (iPSCORE and HipSci), ethnicity, iPSC passage and ten PEER factors(Stegle et al. 2010) calculated on gene expression levels. We performed eQTL analysis for all 146 expressed genes and all the SNPs using Matrix eQTL (Shabalin 2012). The rsID for each SNP-eQTL was intersected with GWAS hits from the GWAS catalog (2018-12-21 release) (Buniello et al. 2019).

HLA-eQTL detection

For each sample, the HLA types were assigned dosage values and converted to VCF file format. Dosage was assigned as follows: 0 = the sample did not harbor the analyzed allele; 0.5 = the sample was heterozygous for the analyzed allele; 1 = the sample was homozygous for the analyzed allele. For each HLA gene, we investigated the associations of each single HLA type with gene expression. We performed eQTL analysis for all 146 expressed genes and all the HLA types (346 in total) using Matrix eQTL (Shabalin 2012).

Conditional HLA-eQTL analysis

To detect QTLs conditional on the top SNP-eQTL, for each gene we repeated HLA-eQTL and SNP-eQTL detection adding the genotype from the top SNP-eQTL as covariate. To detect HLA-eQTLs conditional on the top two independent SNP-eQTLs, we repeated HLA-eQTL detection using adding the genotype from the top SNP-eQTL as well as the genotype from the top conditional SNP-eQTL as covariates.

Accession numbers

Whole-genome sequencing data of 273 individuals in the iPSCORE cohort (Panopoulos et al. 2017) is publicly available through dbGaP: phs001325. RNA-seq data of 215 individuals in the iPSCORE cohort (DeBoever et al. 2017) is publicly available through dbGaP:phs000924. Whole-genome sequencing data of 377 individuals in the HipSci cohort is publicly available through EGA: PRJEB15299. RNA-seq data of 231 individuals in the HipSci cohort is publicly available through ENA: PRJEB7388.

Acknowledgements

We thank Ivan Carcamo-Orive and the other members of the i2QTL Consortium for their comments. This work was supported in part by a California Institute for Regenerative Medicine grant GC1R-06673 to KAF; and National Institutes

of Health grants HG008118 to KAF, HL107442 to KAF, DK105541 to KAF and DK112155 to KAF. These funding agencies played no role in the design or conclusions of this study.

Author information

KF and NN conceived the study. HM and DJ performed WGS analysis. JR and NN called and performed quality check on HLA types. MD performed gene expression and QTL analyses. MJB and OS Consortium contributed to WGS and RNA-seq data analysis. ENS, MJB and OS oversaw the QTL analysis. ADC performed iPSC culture and collected RNA for sequencing. MD, JR, NN and KAF prepared the manuscript.

Figures

Figure 1

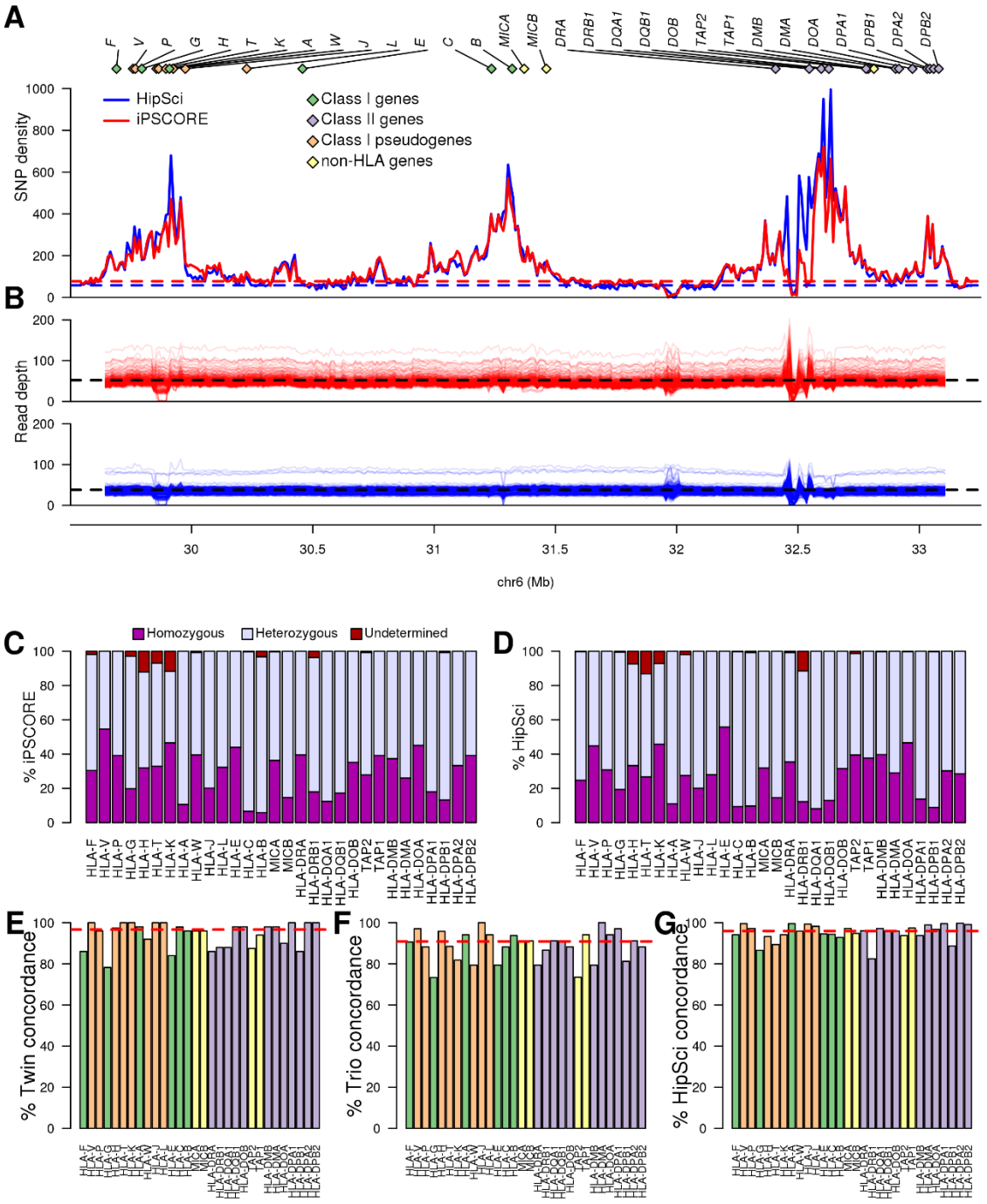

Figure 1. SNP density in HLA genomic region. (**A**) SNP density distribution in consecutive 10-kb bins in the HLA region (chr6: 29690551, 33102442). The dashed lines represent the genome-wide SNP density in the iPSCORE (blue) and HipSci (red) WGS samples. The thirty genes in the HLA region analysed in this study are represented by diamonds with green representing HLA Class I, purple HLA Class I pseudogenes, orange HLA Class II and yellow non-HLA genes. (**B**)

Read depth coverage of the HLA region for the iPSCORE WGS samples (top, red) and the HipSci WGS samples (bottom, blue) was calculated using consecutive 10-kb bins. Each line represents a single WGS sample, the black dashed lines represent the average coverage for the iPSCORE WGS samples (52X) and the HipSci WGS samples (38X) within the HLA region. (**C, D**) Fraction of the 273 iPSCORE (**C**) and 377 HipSci (**D**) WGS samples with homozygous, heterozygous and undetermined alleles for each of the 30 HLA genes. For each sample, if one of the two alleles for an HLA gene was undetermined, we considered the HLA genotype as "undetermined". (**E**) Twin concordance (iPSCORE:25 monozygotic twin pairs), (**F**) Mendelian inheritance concordance (iPSCORE:17 trios) and (**G**) Fibroblast-iPSC concordance (HipSci:231 fibroblast-iPSC pairs) of HLA types at eight-digit resolution. Genes were sorted based on their genomic position and colored according to their class, as shown in panel A. Twin and Mendelian inheritance concordance at two-, four- and six-digit are shown in Figure S1.

Figure 2

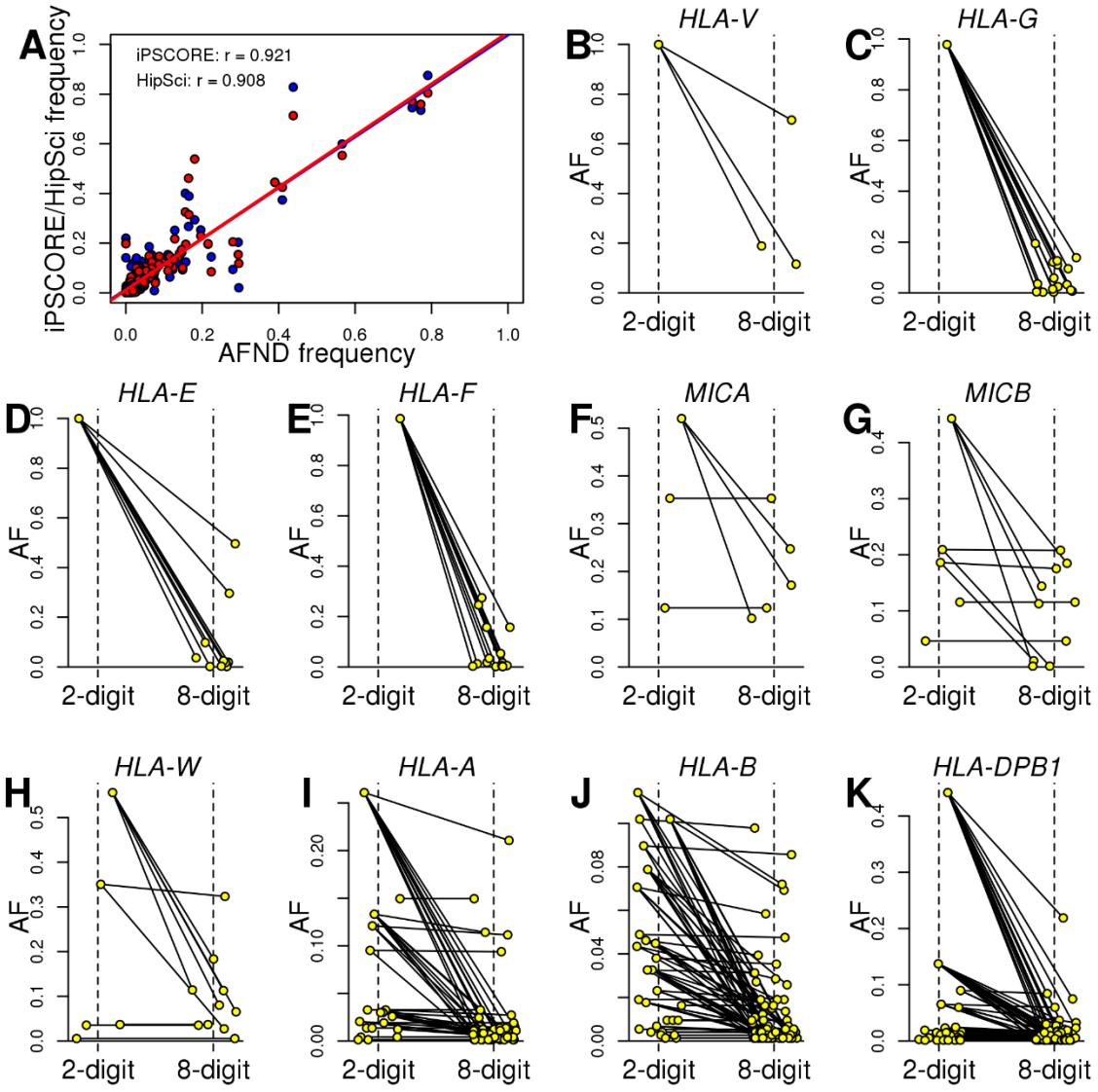

Figure 2. Allele frequency of predicted HLA types. (**A**) Correlation between HLA allele frequency at the 4-digit resolution between individuals in the Allele Frequency Net Database (AFND) and either iPSCORE (red) or HipSci (blue). Each dot represents the frequency for a given HLA allele. (**B-K**) Associations between allele frequency (AF) of HLA types for ten genes at 2-digit (left) and 8-digit resolution (right). Lines link all 8-digit alleles their parent 2-digit allele. Figure S2 shows the same associations for the other 20 genes.

Figure 3

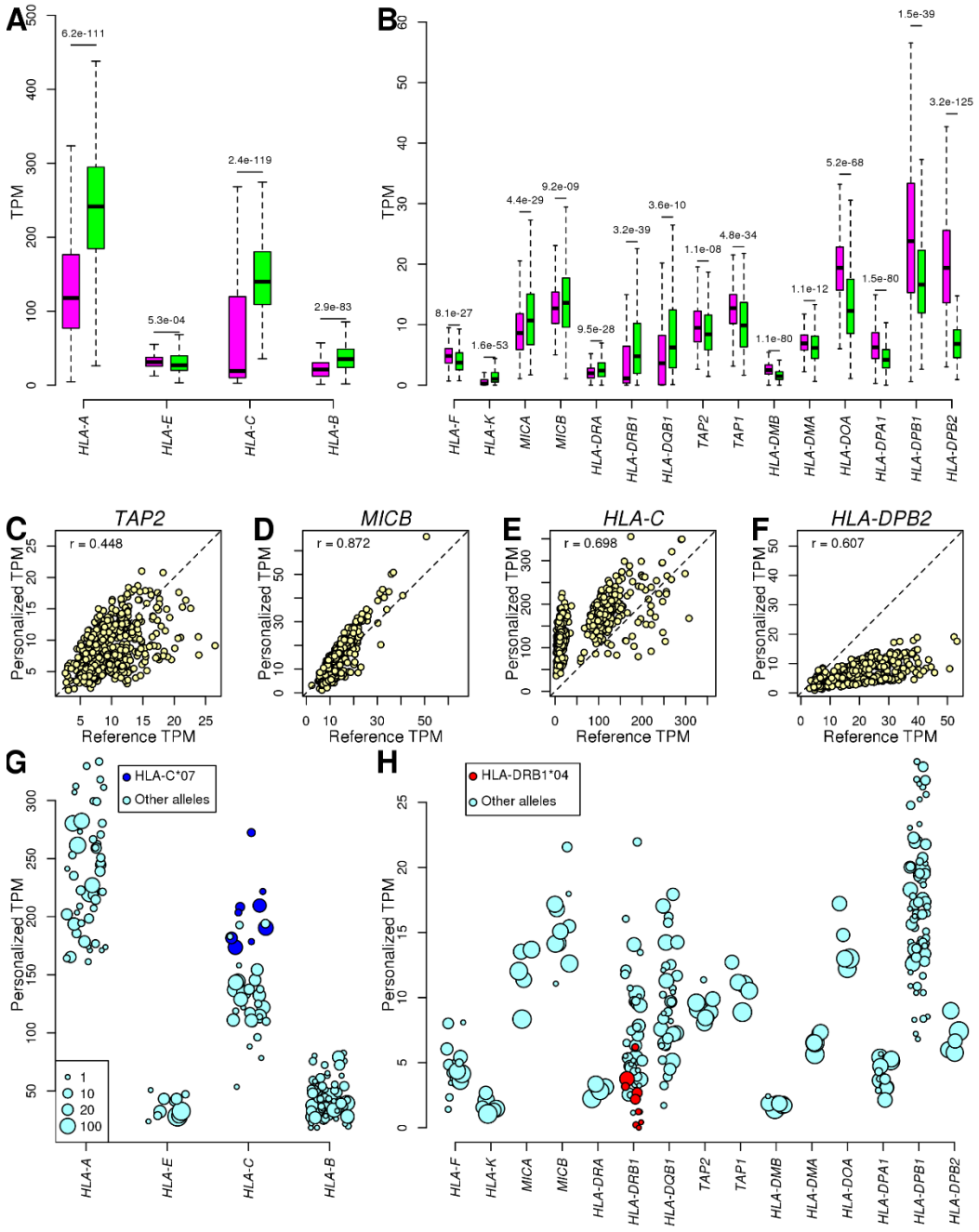

Figure 3. Allele-specific expression of HLA genes in iPSCs. (**A-B**) For each of the 19 genes expressed in iPSCs, boxplots showing the distribution of TPM calculated using reference transcripts (purple) and personalized transcripts (green) is shown. P-values are shown for each gene (calculated using paired t-test between reference and personalized TPM in each sample). (**C-F**) Scatterplots showing the differences between TPM calculated using reference transcripts (X axis) and personalized transcripts (Y axis). Scatterplots for the other 15 genes are shown in Figure S3. (**G-H**) Mean personalized TPM expression across HLA types. Each point represents one allele. Eight-digit HLA-types that map to the same two-digit HLA-C*07 and HLA-DRB1*04 alleles are shown in dark blue and red, respectively. Point size represents the number of individuals carrying each allele. Shown are all HLA types, although differential gene expression analysis (Table S4) was performed only on HLA types carried by at least two individuals.

Figure 4

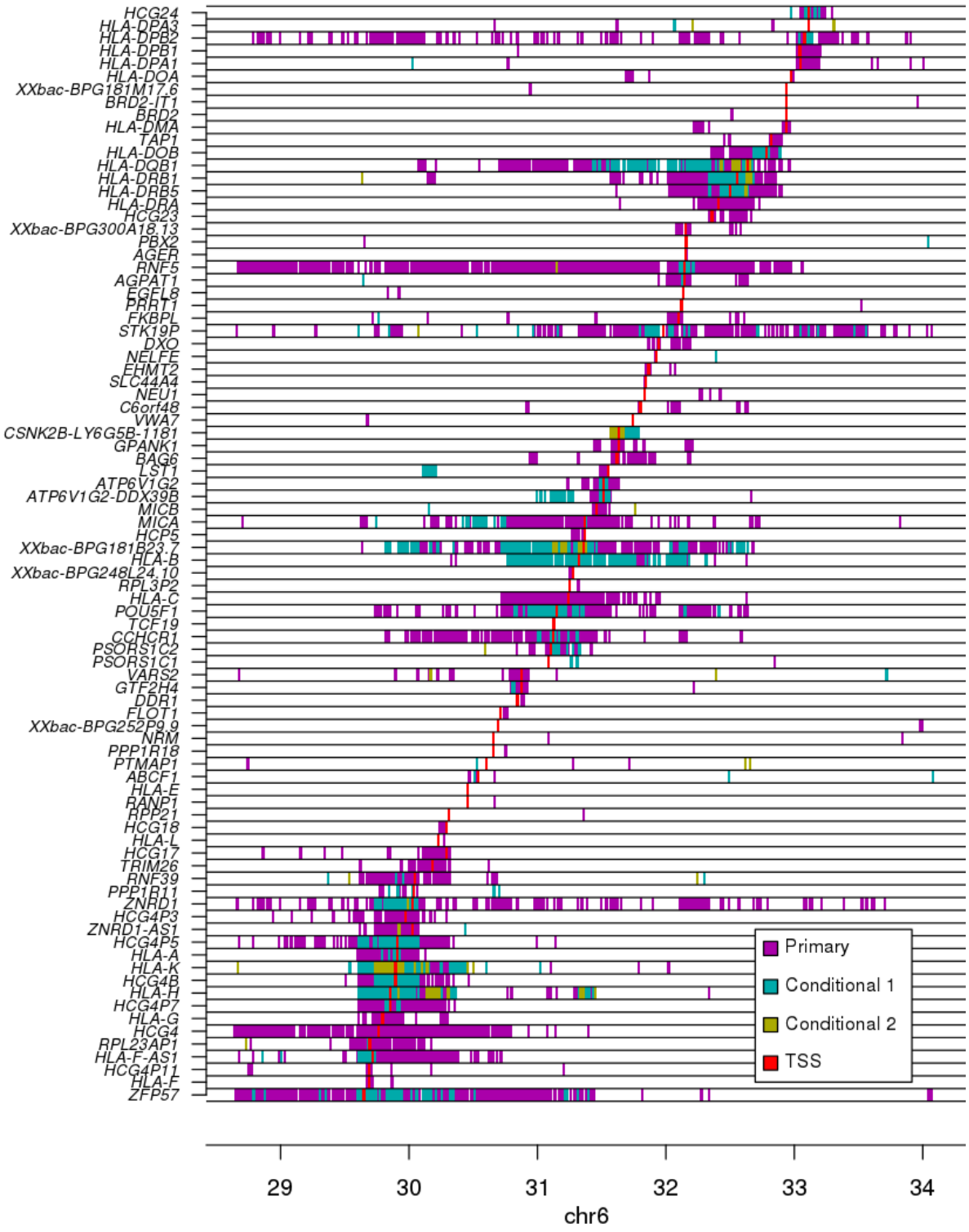

Figure 4. Associations between SNPs and gene expression levels in the MHC region. SNP-eQTL associations for 83 genes in the MHC region. Colors represent whether each SNP-QTL is primary or conditional. TSS for each gene is shown in red.

Figure 5

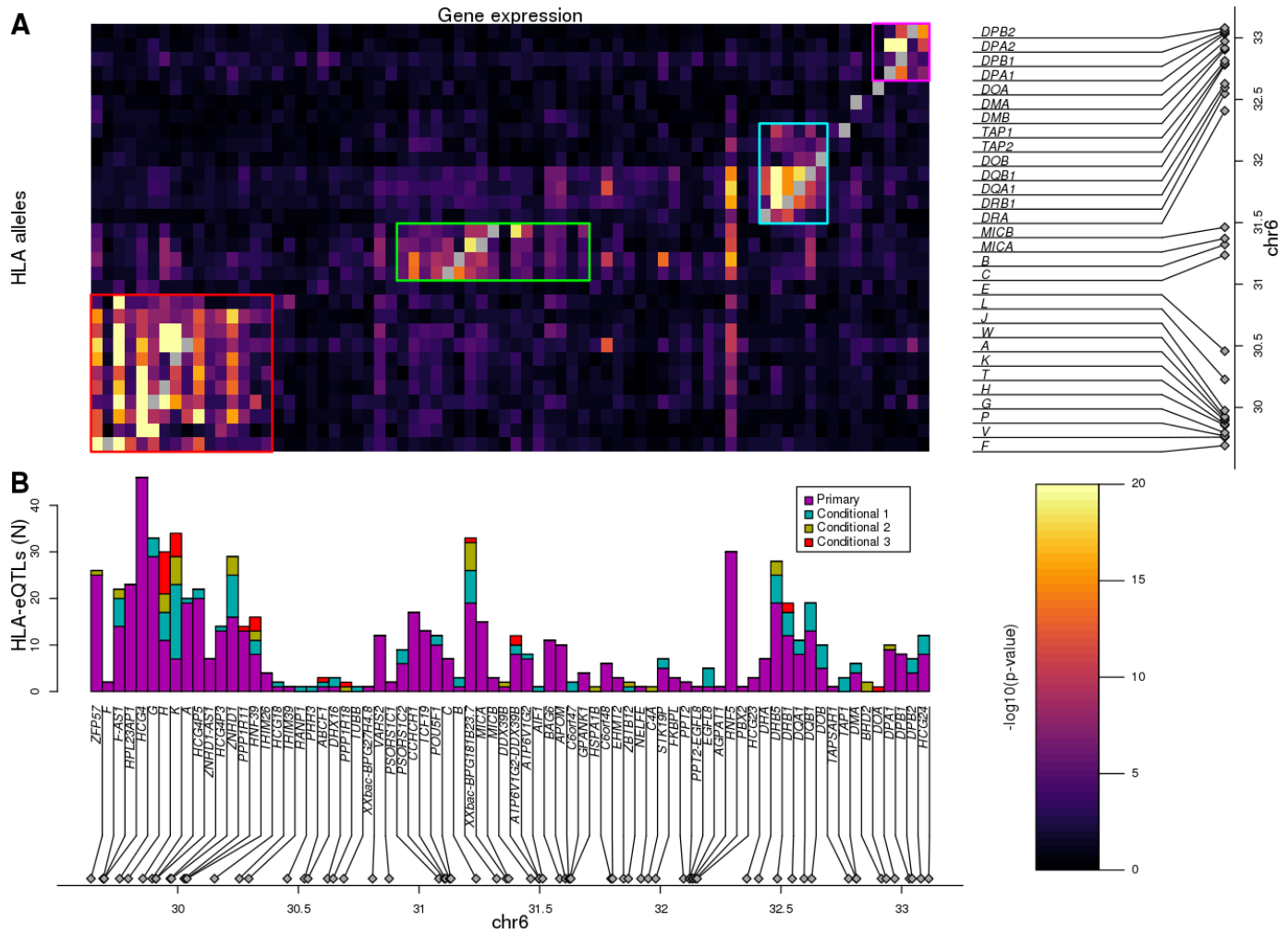

Figure 5. Associations between HLA types and gene expression levels in the MHC region. (**A**) Heatmap showing the associations between HLA types at 8-digit resolution (Y axis) and the expression of 77 genes in the MHC region (X axis). For each of the 30 HLA genes, the p-value of the HLA type with the most significant association is shown. Associations between HLA types of the same gene were not considered (grey squares). Four different groups of genes with shared HLA-eQTLs are outlined in different colors. (**B**) Barplot showing for each of the 77 genes in the heatmap in (**A**), the number of primary and conditional associations (adjusted based on the top SNP-eQTL, the top two independent SNP-eQTLs, or the top three independent SNP-eQTLs) between HLA types and gene expression levels.

Figure 6

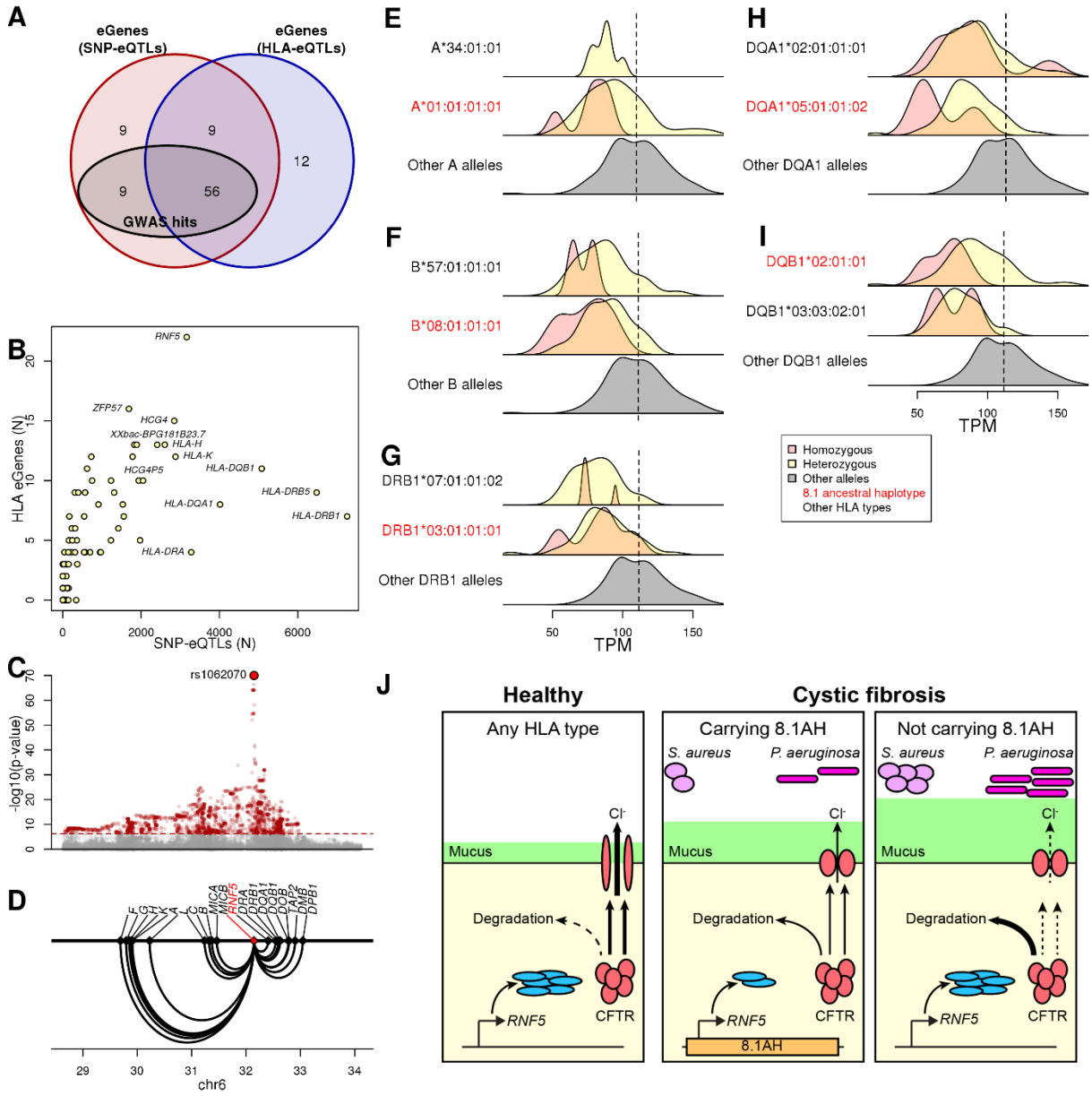

Figure 6. Associations between HLA types and *RNF5* expression. (**A**) Venn diagram showing the overlap between eGenes with SNP-eQTLs, HLA-eQTLs and SNP-eQTLs that were GWAS hits. (**B**) Scatterplot showing, for each of the 95 MHC eGenes, the number of SNP-QTLs and the number of HLA genes whose alleles are HLA-QTLs. (**C**) Scatterplots showing *RNF5* SNP-eQTLs for all the SNPs in the MHC region. Significant SNP-eQTLs (Bonferroni-corrected p-values < 0.05) are shown in red. (**D**) Interactions between *RNF5* (indicated in red) expression and HLA genes whose alleles are HLA-eQTLs. (**E-I**) Distributions of *RNF5* expression levels in samples that are homozygous (red) and heterozygous (yellow) for significant HLA-eQTLs in six HLA genes. Each gene has one significant HLA type included in 8.1AH (the 8.1AH allele is highlighted in red) as well as one other significant HLA type. Gene expression distribution for all samples that do not carry any of the significant HLA-QTL alleles is shown in grey. Gene expression distributions for all other HLA types significantly associated with *RNF5* expression are shown in Figure S5. (**J**) Cartoon depicting our hypothesis of

the molecular mechanisms underlying the associations between 8.1AH, *RNF5* expression and CFTR function in CF. On the left, in healthy individuals CFTR is correctly folded in the airway epithelium and Cl^- ions can be secreted. On the right, in CF patients not carrying 8.1AH, CFTR is misfolded and high levels of RNF5 result in its degradation which, in turn, results in decreased Cl^- secretion, mucus hypersecretion and colonization by *S. aureus* and *P. aeruginosa*[52]. In the middle, in CF patients carrying 8.1AH, *RNF5* is expressed at low levels, resulting in the lower degradation of the misfolded mutated CFTR protein, improved Cl^- secretion, lower mucus secretion and delayed colonization by *S. aureus* and *P. aeruginosa*.

References

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.

Bai Y, Ni M, Cooper B, Wei Y, Fury W. 2014. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC genomics* **15**: 325.

Bauer DC, Zadoorian A, Wilson LO, Thorne NP. 2016. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Brief Bioinform* doi:10.1093/bib/bbw097.

Blackwell JM, Jamieson SE, Burgner D. 2009. HLA and infectious diseases. *Clin Microbiol Rev* **22**: 370-385, Table of Contents.

Bodis G, Toth V, Schwarting A. 2018. Role of Human Leukocyte Antigens (HLA) in Autoimmune Diseases. *Rheumatol Ther* **5**: 5-20.

Boegel S, Lower M, Schafer M, Bukur T, de Graaf J, Boisguerin V, Tureci O, Diken M, Castle JC, Sahin U. 2012. HLA typing from RNA-Seq sequence reads. *Genome medicine* **4**: 102.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**: D1005-D1012.

DeBoever C, Li H, Jakubosky D, Benaglio P, Reyna J, Olson KM, Huang H, Biggs W, Sandoval E, D'Antonio M et al. 2017. Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell* **20**: 533-546.e537.

Dechecchi MC, Tamanini A, Cabrini G. 2018. Molecular basis of cystic fibrosis: from bench to bedside. *Ann Transl Med* **6**: 334.

DeGiorgio M, Lohmueller KE, Nielsen R. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS genetics* **10**: e1004561.

Dendrou CA, Petersen J, Rossjohn J, Fugger L. 2018. HLA variation and disease. *Nature reviews Immunology* **18**: 325-339.

Dilthey AT, Gourraud PA, Mentzer AJ, Cereb N, Iqbal Z, McVean G. 2016. High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. *PLoS computational biology* **12**: e1005151.

Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, Bonder MJ, Fu J, Deelen P, Groen HJ, Smolonska A et al. 2011. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS genetics* **7**: e1002197.

Gambino CM, Aiello A, Accardi G, Caruso C, Candore G. 2018. Autoimmune diseases and 8.1 ancestral haplotype: An update. *Hla* **92**: 137-143.

Gonzalez-Galarza FF, Takeshita LY, Santos EJ, Kempson F, Maia MH, da Silva AL, Teles e Silva AL, Ghattaoraya GS, Alfirevic A, Jones AR et al. 2015. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic acids research* **43**: D784-788.

Gough SC, Simmonds MJ. 2007. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr Genomics* **8**: 453-465.

Gregersen PK, Silver J, Winchester RJ. 1987. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis and rheumatism* **30**: 1205-1213.

Holoshitz J. 2013. The quest for better understanding of HLA-disease association: scenes from a road less travelled by. *Discovery medicine* **16**: 93-101.

Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I. 2013. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. *BMC genomics* **14**: 355.

Huang Y, Yang J, Ying D, Zhang Y, Shotelersuk V, Hirankarn N, Sham PC, Lau YL, Yang W. 2015. HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome medicine* **7**: 25.

Jensen JM, Villesen P, Friborg RM, Danish Pan-Genome C, Mailund T, Besenbacher S, Schierup MH. 2017. Assembly and analysis of 100 full MHC haplotypes from the Danish population. *Genome research* **27**: 1597-1607.

Ka S, Lee S, Hong J, Cho Y, Sung J, Kim HN, Kim HL, Jung J. 2017. HLAscan: genotyping of the HLA region using next-generation sequencing data. *BMC bioinformatics* **18**: 258.

Kennedy AE, Ozbek U, Dorak MT. 2017. What has GWAS done for HLA and disease associations? *International journal of immunogenetics* **44**: 195-211.

Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, Bala S, Bensaddek D, Casale FP, Culley OJ et al. 2017. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**: 370-375.

Kim HJ, Pourmand N. 2013. HLA typing from RNA-seq data using hierarchical read weighting [corrected]. *PloS one* **8**: e67885.

Klein L, Kyewski B, Allen PM, Hogquist KA. 2014. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature reviews Immunology* **14**: 377-391.

Laki J, Laki I, Nemeth K, Ujhelyi R, Bede O, Endreffy E, Bolbas K, Gyurkovits K, Csiszer E, Solyom E et al. 2006. The 8.1 ancestral MHC haplotype is associated with delayed onset of colonization in cystic fibrosis. *Int Immunol* **18**: 1585-1590.

Lee H, Kingsford C. 2018. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome biology* **19**: 16.

Levine JE, Yang SY. 1994. SSOP typing of the Tenth International Histocompatibility Workshop reference cell lines for HLA-C alleles. *Tissue antigens* **44**: 174-183.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**: 323.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009b. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Lyczak JB, Cannon CL, Pier GB. 2002. Lung infections associated with cystic fibrosis. *Clin Microbiol Rev* **15**: 194-222.

Mall MA, Hartl D. 2014. CFTR: cystic fibrosis and beyond. *Eur Respir J* **44**: 1042-1054.

Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Fernandez-Vina M, Geraghty DE, Holdsworth R, Hurley CK et al. 2010. Nomenclature for factors of the HLA system, 2010. *Tissue antigens* **75**: 291-455.

Matzaraki V, Kumar V, Wijmenga C, Zhernakova A. 2017. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome biology* **18**: 76.

McNicholas CM. 2017. Beyond cystic fibrosis transmembrane conductance regulator (CFTR) single channel kinetics: implications for therapeutic intervention. *J Physiol* **595**: 1015-1016.

Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, Morrison J, Whittaker P, Lander ES, Cardon LR et al. 2005. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *American journal of human genetics* **76**: 634-646.

Munder A, Tummler B. 2015. Origins of cystic fibrosis lung disease. *N Engl J Med* **372**: 1574.

Mungall AJ Palmer SA Sims SK Edwards CA Ashurst JL Wilming L Jones MC Horton R Hunt SE Scott CE et al. 2003. The DNA sequence and analysis of human chromosome 6. *Nature* **425**: 805-811.

Nariai N, Kojima K, Saito S, Mimori T, Sato Y, Kawai Y, Yamaguchi-Kabata Y, Yasuda J, Nagasaki M. 2015. HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. *BMC genomics* **16 Suppl 2**: S7.

Norman PJ, Norberg SJ, Guethlein LA, Nemat-Gorgani N, Royce T, Wroblewski EE, Dunn T, Mann T, Alicata C, Hollenbach JA et al. 2017. Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. *Genome research* **27**: 813-823.

Okada Y, Eyre S, Suzuki A, Kochi Y, Yamamoto K. 2018. Genetics of rheumatoid arthritis: 2018 status. *Annals of the rheumatic diseases* doi:10.1136/annrheumdis-2018-213678.

Okada Y, Momozawa Y, Ashikawa K, Kanai M, Matsuda K, Kamatani Y, Takahashi A, Kubo M. 2015. Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. *Nature genetics* **47**: 798-802.

Oldstone MB. 1998. Molecular mimicry and immune-mediated diseases. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **12**: 1255-1265.

Panopoulos AD, D'Antonio M, Benaglio P, Williams R, Hashem SI, Schuldt BM, DeBoever C, Arias AD, Garcia M, Nelson BC et al. 2017. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Reports* **8**: 1086-1100.

Pier GB, Grout M, Zaidi TS, Goldberg JB. 1996. How mutant CFTR may contribute to Pseudomonas aeruginosa infection in cystic fibrosis. *Am J Respir Crit Care Med* **154**: S175-182.

Robinson J, Soormally AR, Hayhurst JD, Marsh SG. 2016. The IPD-IMGT/HLA Database - New developments in reporting HLA variation. *Human immunology* **77**: 233-237.

Schlott F, Steubl D, Ameres S, Moosmann A, Dreher S, Heemann U, Hosel V, Busch DH, Neuenhahn M. 2018. Characterization and clinical enrichment of HLA-C*07:02-restricted Cytomegalovirus-specific CD8+ T cells. *PloS one* **13**: e0193554.

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**: 1353-1358.

Sondo E, Falchi F, Caci E, Ferrera L, Giacomini E, Pesce E, Tomati V, Mandrup Bertozzi S, Goldoni L, Armirotti A et al. 2018. Pharmacological Inhibition of the Ubiquitin Ligase RNF5 Rescues F508del-CFTR in Cystic Fibrosis Airway Epithelia. *Cell Chem Biol* **25**: 891-905 e898.

Souza de Lima D, Morishi Ogusku M, Porto Dos Santos M, de Melo Silva CM, Alves de Almeida V, Assumpcao Antunes I, Boechat AL, Ramasawmy R, Sadahiro A. 2016. Alleles of HLA-DRB1*04 Associated with Pulmonary Tuberculosis in Amazon Brazilian Population. *PloS one* **11**: e0147543.

Stegle O, Parts L, Durbin R, Winn J. 2010. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology* **6**: e1000770.

Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. 2014. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**: 3310-3316.

Tan A, Abecasis GR, Kang HM. 2015. Unified representation of genetic variants. *Bioinformatics* **31**: 2202-2204.

Tomati V, Sondo E, Armirotti A, Caci E, Pesce E, Marini M, Gianotti A, Jeon YJ, Cilli M, Pistorio A et al. 2015. Genetic Inhibition Of The Ubiquitin Ligase Rnf5 Attenuates Phenotypes Associated To F508del Cystic Fibrosis Mutation. *Scientific reports* **5**: 12138.

Trowsdale J, Knight JC. 2013. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet* **14**: 301-323.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**: 11 10 11-33.

Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, Holt RA. 2012. Derivation of HLA types from shotgun sequence datasets. *Genome medicine* **4**: 95.

Xie C, Yeo ZX, Wong M, Piper J, Long T, Kirkness EF, Biggs WH, Bloom K, Spellman S, Vierra-Green C et al. 2017. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc Natl Acad Sci U S A* **114**: 8059-8064.

Yin L, Dai S, Clayton G, Gao W, Wang Y, Kappler J, Marrack P. 2013. Recognition of self and altered self by T cells in autoimmunity and allergy. *Protein & cell* **4**: 8-16.

Zhang C, de Smith AJ, Smirnov IV, Wiencke JK, Wiemels JL, Witte JS, Walsh KM. 2017. Non-additive and epistatic effects of HLA polymorphisms contributing to risk of adult glioma. *J Neurooncol* **135**: 237-244.