The Journal of Immunology

RESEARCH ARTICLE | FEBRUARY 15 2018

Novel Transcriptional Activity and Extensive Allelic Imbalance in the Human MHC Region FREE

Elizabeth Gensterblum-Miller; ... et. al

J Immunol (2018) 200 (4): 1496–1503.
https://doi.org/10.4049/jimmunol.1701061

Novel Transcriptional Activity and Extensive Allelic Imbalance in the Human MHC Region

Elizabeth Gensterblum-Miller,* Weisheng Wu,† and Amr H. Sawalha*,‡

The MHC region encodes HLA genes and is the most complex region in the human genome. The extensively polymorphic nature of the HLA hinders accurate localization and functional assessment of disease risk loci within this region. Using targeted capture sequencing and constructing individualized genomes for transcriptome alignment, we identified 908 novel transcripts within the human MHC region. These include 593 novel isoforms of known genes, 137 antisense strand RNAs, 119 novel long intergenic noncoding RNAs, and 5 transcripts of 3 novel putative protein-coding human endogenous retrovirus genes. We revealed allele-dependent expression imbalance involving 88% of all heterozygous transcribed single nucleotide polymorphisms throughout the MHC transcriptome. Among these variants, the genetic variant associated with Behçet's disease in the *HLA-B/MICA* region, which tags *HLA-B*51*, is within novel long intergenic noncoding RNA transcripts that are exclusively expressed from the haplotype with the protective but not the disease risk allele. Further, the transcriptome within the MHC region can be defined by 14 distinct coexpression clusters, with evidence of coregulation by unique transcription factors in at least 9 of these clusters. Our data suggest a very complex regulatory map of the human MHC, and can help uncover functional consequences of disease risk loci in this region. ***The Journal of Immunology*****, 2018, 200: 1496–1503.**

The human MHC is a highly complex polymorphic genomic region containing many important immune-related genes. This region includes the HLA genes, involved in both self-tolerance and Ag presentation. Polymorphisms within HLA genes have been associated with over 100 autoimmune diseases and cancers, and allelotyping of translated genes has been the focus of extensive research (1–3). Intergenic variants within the MHC region, which may serve a role in gene regulation, have also been associated with several immune-related diseases (4, 5). However, the role of these intergenic variants is often not clear because regulation within the MHC is incompletely understood. The MHC contains a complex regulatory network including *cis*-acting and *trans*-acting regulation bridging inside and outside the MHC region (6, 7). Due to both the high rate of polymorphism and the complex regulatory networks within the MHC, the functional effects of specific disease-associated variants are difficult to elucidate.

Long intergenic noncoding RNAs (lincRNAs) have been extensively implicated in transcriptional regulation by recruitment of regulatory proteins. These proteins proceed to regulate gene expression by epigenetic modification, such as DNA methylation and chromatin modification (8, 9). Recruitment of transcription factors by lincRNAs has also been described previously (10). However, many lincRNAs are weakly expressed, and therefore may not be detected by RNA sequencing that spans the entire transcriptome.

Sequence-specific enrichment by magnetic bead pulldown has previously been used to sequence HLA genes for allelotyping, and to elucidate regulatory regions of individual genes (2, 4). We performed targeted enrichment of the entire MHC region in primary human monocytes using sequence-specific capture probes, followed by high-throughput sequencing of DNA and RNA (CaptureSeq) (11), to allow for deep sequencing coverage of the MHC region. We targeted the entire MHC, including both intergenic regions and known splice variants. We identified genetic variants, then constructed personalized genomes to accurately align RNA sequences. After enrichment and alignment to personalized genomes, we were able to detect low-expressed transcripts, and by including all genomic regions, we were able to identify novel intergenic transcripts. We also comprehensively assessed allelic expression imbalance and revealed extensive allele-specific expression throughout the entire MHC, indicating that polymorphism is a mechanism of complex transcriptional regulation in this region.

Materials and Methods

Probe design

Sequence-specific capture probes were designed to target the complete reference genomic sequence of the MHC region (chromosome 6: 28.5–33.5 Mb, hg19), as well as splice sites for known transcripts the region contains. By including intergenic and intronic genomic regions, sequences that overlap with previously unannotated regions could be captured and subsequently sequenced; moreover, the same set of probes was designed to enable us to capture both DNA and RNA. This pool of 75 base long capture probes was designed to target 35,895 sequences throughout the region. For the main

*Division of Rheumatology, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109; †Biomedical Research Core Facilities, Bioinformatics Core, University of Michigan, Ann Arbor, MI 48109; and ‡Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109

Received for publication August 2, 2017. Accepted for publication December 7, 2017.

This work was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award R01AR070148.

E.G.-M. contributed to experimental design, the performing of the experiments, analyzing the data, and drafting the manuscript; W.W. contributed to performing data analysis and editing the manuscript; and A.H.S. conceived the study and contributed to experimental design, data analysis, and drafting the manuscript.

The RNA sequencing data presented in this article have been submitted to the National Center for Biotechnology Information's Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108663) under accession number GSE108663.

Address correspondence and reprint requests to Dr. Amr H. Sawalha, University of Michigan, 5520 MSRB-1, SPC 5680, 1150 W. Medical Center Drive, Ann Arbor, MI 48109. E-mail address: asawalha@umich.edu

The online version of this article contains supplemental material.

Abbreviations used in this article: AI, allelic imbalance; CPAT, Coding Potential Analysis Tool; ERV, endogenous retroviral element; GATK, Genome Analysis Toolkit; lincRNA, long intergenic noncoding RNA; SNP, single nucleotide polymorphism.

reference allele, probes directly overlapped 75.9% of the genome, with 88% estimated total sequence coverage. However, because the MHC region is highly polymorphic, the seven alternative reference haplotypes for the MHC were used in addition to the reference allele to design probes targeting all reference genomic sequences in this region. In total, this region, including all alternative haplotypes, was 65.4% covered by the probes, and had a 75.7% estimated net coverage. Of the total target region, including alternate haplotypes, 10% was not covered due to shared homology with other parts of the genome, whereas 14.2% was not covered because of incomplete sequence information in the alternative haplotypes.

Isolation of primary monocyte DNA and RNA

PBMCs isolated from 12 healthy individuals were initially collected by density gradient centrifugation and immediately stored in liquid nitrogen. Cells were thawed, treated with 25 U/ml Benzonase, and incubated at 37°C in RPMI 1640/10% heat-inactivated FBS for 90 min. Thawed PBMCs had a minimum viability of 90%, with an average viability of 98.1 ± 3.6%, measured by trypan blue staining. Primary monocytes were then isolated from thawed PBMCs via negative selection using the Pan Monocyte Isolation Kit, following the manufacturer's instructions (Miltenyi Biotec, San Diego, CA). The remaining monocyte-depleted PBMCs were flushed from the magnetic column, and DNA was isolated using the DNeasy Blood and Tissue Kit (Qiagen, Germantown, MD). RNA was isolated from primary monocytes using the Direct-zol RNA Isolation Kit (Zymo Research, Irvine, CA), and then DNase treated using the TURBO DNA-free kit (Invitrogen, Carlsbad, CA). The purity of the isolated monocytes was measured by flow cytometry using the iCyt Synergy SY3200 cell sorter (Sony Biotechnology, San Jose, CA), staining with APC/Cy7 anti-CD14 (BioLegend, San Diego, CA). Monocyte purity was found to be >90%.

DNA and RNA sequencing

RNA integrity and concentration were verified using the Agilent Bioanalyzer (RIN > 8) (Agilent Technologies, Santa Clara, CA). A minimum of 500 ng of RNA was processed per sample. RNA was ribo-depleted using the NEBNext rRNA Depletion Kit (Human/Mouse/Rat) (New England Biolabs, Ipswich, MA), and sequencing libraries were prepared for every DNA and RNA sample. Sequence-specific magnetic bead capture was performed on DNA and RNA libraries according to the manufacturer's instructions, using the custom-designed probes (SeqCap EZ Choice XL Library System; Roche Nimblegen, Madison, WI). Samples were multiplexed, with four samples per capture reaction. All postcapture genomic DNA libraries were sequenced in one lane, whereas all postcapture cDNA libraries were sequenced in another. All samples were sequenced with the Illumina HiSEquation 2500, with paired 125 bp reads. The RNA sequencing data presented in this article have been submitted to the National Center for Biotechnology Information's Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108663) under accession number GSE108663.

Developing individualized genotypes

DNA reads were quality trimmed using Trimmomatic, then aligned to the hg19 reference sequence using the Burrows-Wheeler aligner (BWA MEM) (12). Duplicate sequences were then removed using Picard, and indels were processed using the Genome Analysis Toolkit (GATK) (13–15). From these aligned reads, SAMtools was used to generate an mpileup file, then VarScan mpileup2snp was used to identify single nucleotide polymorphisms (SNPs) (16, 17). For each individual, SNPs were called based on variation from the reference genome (hg19), and all called SNPs have a total read depth of at least eight and a maximum variant calling p value of 0.01. For all heterozygous SNPs, each allele also has a minimum variant-supporting read depth of two, a minimum average variant-supporting read base quality of 20, and a minimum allele frequency of 0.2. From these identified and quality filtered SNPs, individualized lists of variants were created for each sample. On average, per individual, 23,575 heterozygous variants were identified in the MHC region. The average read depth on heterozygous variants identified in all samples was 417 ± 95.2 with an average variant-supporting read base quality of 230.6 ± 11.7.

RNA alignment and assembly

RNA sequencing reads were quality trimmed using Cutadapt, then aligned to the human reference genome (hg19, chromosome 6, RefSeq transcriptome annotation) using GSNAP (18, 19). Alternate haplotypes for chromosome 6 in the reference genome were not used for alignment, to prevent errors from multimapped reads. RNA reads were aligned in a SNP-tolerant manner, meaning that variants that were called from the DNA sequencing alignment were not included in the mismatch penalty scores for RNA reads. Reads that successfully aligned to the target region were assembled into transcripts using StringTie, guided by the Ensembl transcriptome annotation (20).

Identification of novel transcripts

During RNA assembly, transcripts were annotated using an Ensembl reference. To identify novel transcripts, we used the following workflow (Supplemental Fig. 1). All transcripts that were successfully annotated using the Ensembl reference during alignment were excluded. Using CuffCompare, the remaining transcripts were annotated using Gencode Comprehensive v25 (hg19) as a reference (21, 22). All transcripts were assigned class codes based on their relation to transcripts in the reference. All transcripts that were assigned the class codes I (intronic), J (novel isoform), U (intergenic), and X (antisense) were identified as novel, whereas transcripts containing all other class codes were defined as not novel. The remaining novel transcripts were annotated with CuffCompare using the human lincRNA catalog (23). The transcripts that were found to be novel using all three references were next filtered to include only transcripts with fragments per kilobase of transcript per million mapped reads of 0.1 in two or more samples.

The coding potential of each novel transcript was analyzed using the Coding Potential Analysis Tool (CPAT) (24). The sequence of each novel transcript was determined using genomic coordinates determined by StringTie and the sequence of the reference genome, and these sequences were used to determine coding potential for each transcript. Transcripts with a coding probability of 0.364 or greater were defined as putatively coding, whereas all others were defined as noncoding. All novel transcripts were categorized based on their Gencode annotation class codes and by these coding predictions.

Predicted function of coding genes

Of the five putative coding transcripts that did not share exons with known genes, structure and function were predicted using IntFOLD3. Using the open reading frame predicted by CPAT, the putative amino acid sequence was determined from the transcript sequences using A Plasmid Editor. Using the IntFOLD3 pipeline, we predicted the tertiary structure of the novel peptides, guided by sequence homology with known proteins (25). In addition, putative ligand binding domains and gene ontology term annotation were predicted using the FunFOLD pipeline, which is integrated into IntFOLD3.

Retroviral element sequence alignment

The five novel putative coding intergenic transcripts described were categorized based on their alignment to retroviral sequences. The predicted open reading frames of each of the five transcripts were translated into a protein sequence. Two transcripts that were isoforms of the same gene shared an open reading frame, so four protein sequences were generated. These sequences were aligned to the human proteome using protein-protein BLAST with the nonredundant protein sequences database (26). Each of these four sequences aligned to known retroviral elements (E value $<1 \times 10^{-10}$). Sequence alignments were visualized using MView (27).

Allele-specific expression

Allele-specific expression of aligned transcript and genomic reads at each heterozygous SNP was assessed using GATK ASEReadCounter under the default settings, which includes a read downsampling step (13, 14). Only alignments with base quality and mapping quality no lower than 20 were used. The read counts for each transcript were then normalized to genomic allele-specific read counts derived from DNA reads using GATK ASEReadCounter under the default settings. The DNA allelic imbalance (AI) ratio was first calculated for both the reference and alternate allele of each variant as follows:

$$\mathrm{AI} = \frac{\text{Allele read counts}}{\text{Total DNA read counts}}.$$

Read counts for both alleles of each variant were then calculated from the following formula:

$$\text{Normalized read count} = \frac{\text{RNA Allele Specific Read Count}}{2 \times \mathrm{AI}}.$$

SNPs containing AI were defined by a χ^2 test, p value <0.05, calculated based on the normalized read counts. Relative AI for all expressed heterozygous SNPs was calculated as the reference SNP expression/total expression ratio. Heterozygous SNPs and relative reference allele

expression for all individuals were merged based on SNP position and reference allele. Allele specific expression at rs76546355 was also confirmed using the program bam-readcount (28).

Coexpression analysis

Coexpression networks were assigned using the R package weighted correlation network analysis (29). This package clusters every sequenced transcript based on the normalized read counts (fragments per kilobase of transcript per million mapped reads) in all 12 samples, using a weighted correlation network analysis. For initial quality filtering, all transcripts that were missing from more than one half of all samples were removed from analysis; 320 transcripts were removed. Samples were then clustered according to transcription patterns to remove any outlier samples; however, no outlier samples were observed. From this filtered set of 2753 transcripts, a coexpression network was created, with a soft-thresholding power of seven, a dendrogram cut height of 0.25, and a minimum cluster size of 30 transcripts. All transcripts were categorized within an expression dendrogram, then successfully assigned to a coexpression cluster. A total of 14 clusters were defined. The genomic localization of each cluster was visualized using Circos (30).

Transcription factor binding site enrichment analysis

Transcription factor binding site enrichment analysis was performed for each of the 14 coexpression clusters using Genome*Runner* Web (31), which compares the genomic coordinates of each transcript to the genomic positions of known transcription factor binding sites, using a database that includes the non–cell-specific binding patterns of 161 transcription factors, measured via transcription factor chromatin immunoprecipitation–seq distributed by ENCODE. The coordinates for the promoter region of each gene in each coexpression cluster was used as input, defined as the 1500 bp preceding transcriptional start sites. As a background for enrichment analysis, we included the promoter region of every gene within the MHC region, as annotated by the University of California, Santa Cruz known genes list, and also included the novel genes that we have described. The University of California, Santa Cruz known gene list contains an aggregation of gene annotations from across the RefSeq, GenBank, CCDS, Rfam, and tRNAscan-SE databases. Transcription factor enrichment was calculated for each coexpression cluster individually, and a cluster was called enriched for a specific transcription factor when an increased frequency of the target was observed in the cluster compared with the background (OR > 1, χ^2 $p < 0.05$). In total, 9 of the 14 coexpression clusters were enriched for specific transcription factors.

HLA allelotypes

HLA allelotypes for each sample were determined using a BWAkit. This pipeline calls types by aligning reads to each HLA gene using the BWA-MEM algorithm, and comparing the exons of each gene to alleles defined by IMGT/HLA. The called types (Supplemental Table I) are defined as the alleles that have minimal exonic mismatch with the individual's sequence.

Sanger sequencing

Allele-specific expression was validated by Sanger sequencing for the target variant rs76546355. RNA was saved from each individual before sequence-specific capture and was converted into cDNA using the Verso cDNA Synthesis Kit (Thermo Fisher Scientific, Waltham, MA). This cDNA was then amplified via PCR, using primers that flank the target SNP (forward primer: 5′-TGCTTGCCTGTTGTGAGATG-3′, reverse primer: 5′-AAGCAACAGTAATTTGGATCTTCC-3′). The proportion of each allele represented in this PCR product was estimated using a Sanger sequencing trace file for each sample.

Results

Targeted genome and transcriptome sequencing in the human MHC region

We performed deep targeted genome and transcriptome sequencing of the human MHC region [chromosome 6 (hg19): 28.5–33.5 Mb] in primary human monocytes. Constructing individualized genomes for aligning RNA sequencing reads generated by deep targeted transcriptome sequencing improved transcript alignment and characterization in this complex polymorphic region. DNA sequence reads aligned against the reference genome human MHC region with a mean read depth of 334.8 ± 84.3 in all samples (Supplemental Fig. 2). Genetic polymorphisms in each sample were identified and an individualized MHC genome in each sample was constructed. A total of 65,289 genetic variants relative to the reference genome were identified, including 62,449 genetic variants that are heterozygous in at least one sample.

Targeted RNA sequencing was performed following rRNA depletion, allowing for high density coverage with an average 36.3 million alignments to the MHC region, and a mean read depth of 594.5 ± 87.2 per gene in this region in all samples. RNA sequence alignment was performed in an individualized SNP-tolerant mode using DNA sequencing data from each sample to allow alignment to polymorphic loci identified in each corresponding sample. This strategy significantly enhanced successful alignment of transcript reads to the polymorphic MHC region, which, coupled with highly dense targeted RNA sequencing, allowed for accurate identification of known and novel transcripts in the MHC region, including transcripts with low expression levels.

Identification and classification of novel transcripts within the MHC region

We identified a total of 3072 transcripts aligned to the human MHC region in human primary monocytes. Of these, 908 were identified as novel transcripts that were present in at least two independent samples (Supplemental Table II). This includes 517 and 76 novel coding and noncoding transcript isoforms of known genes, respectively. In addition, we identified 137 novel antisense strand transcripts, 119 lincRNA transcripts, 54 intronic noncoding RNA transcripts, and 5 transcripts of 3 novel coding genes (Fig. 1).

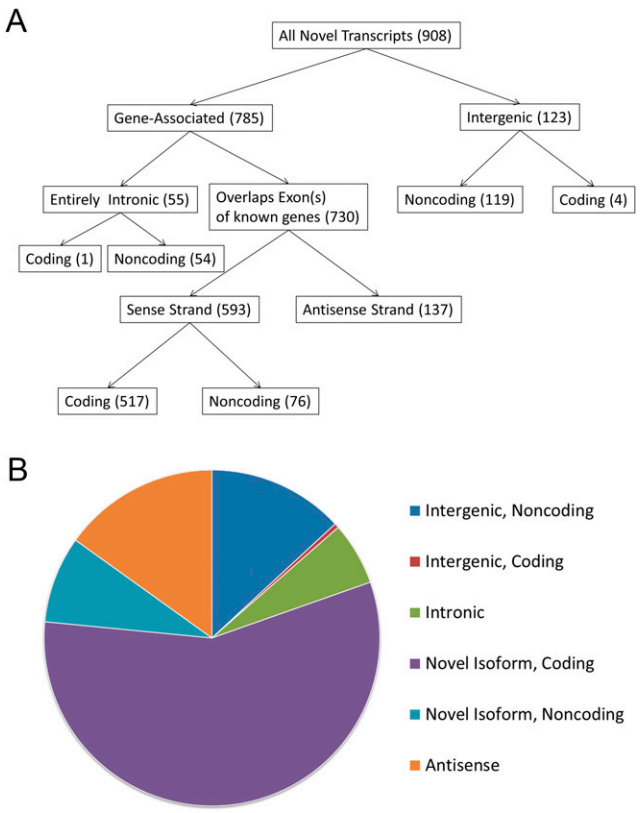

FIGURE 1. A flowchart (**A**) and pie chart (**B**) depicting and summarizing the filtering categories used to classify novel transcripts identified in this study. The categories for intronic, novel isoforms, antisense, and intergenic transcripts were defined via a CuffCompare annotation using the Comprehensive Gencode Resease 25 annotation (hg19) reference transcriptome. Coding potential of novel transcripts was predicted using CPAT.

Evidence for extensive cis *allele-specific expression within the human MHC*

Next, we evaluated the extent of allele-specific expression imbalance in MHC region transcripts that overlap with heterozygous single nucleotide polymorphisms identified using DNA sequencing. We show that 88% of heterozygous transcribed SNPs within the MHC region are associated with significant allele-dependent transcriptional imbalance, with 43% demonstrating extreme allele-dependent expression (>95% expression on either the reference or alternative allele) (Fig. 2A). Indeed, AI is observed in over 69% of all heterozygous SNPs identified in our study within the MHC region (Supplemental Tables III, IV). This remarkably extensive allele-specific expression pattern is nonstochastic and consistent across independent samples in heterozygous SNPs that are present in two or more samples (Fig. 2B, 2C). Whereas the overall number of heterozygous SNPs with evidence of allelic expression imbalance was highest in the HLA class II gene region within the MHC, the frequency of transcribed SNPs with AI relative to all transcribed SNPs was consistent throughout the HLA regions within the MHC (Fig. 3, Supplemental Table V).

To demonstrate AI in a disease-relevant locus in the MHC region, we examined the expression of novel transcripts that overlap with and include the SNP rs76546355 (rs116799036) localized between *HLA-B* and *MICA*. This genetic variant tags the most robust genetic association in Behçet's disease (5). We show that rs76546355, previously thought to be intergenic, is expressed within four lincRNA transcripts we identified between *HLA-B* and *MICA*. Importantly, these four transcripts are exclusively expressed from the haplotype with the disease-protective allele in this SNP. There was no expression of these transcripts from the haplotype with the disease risk allele in heterozygous individuals (Fig. 4). These data suggest evidence for haploinsufficiency involving the expression of novel lincRNAs, induced by a disease risk variant within the MHC region in a complex polygenic disease.

Coexpression patterns and transcription factor binding analyses identify transcriptional clusters in the human MHC transcriptome

We characterized the expression patterns of the transcripts within the MHC using a coexpression network analysis. We defined a

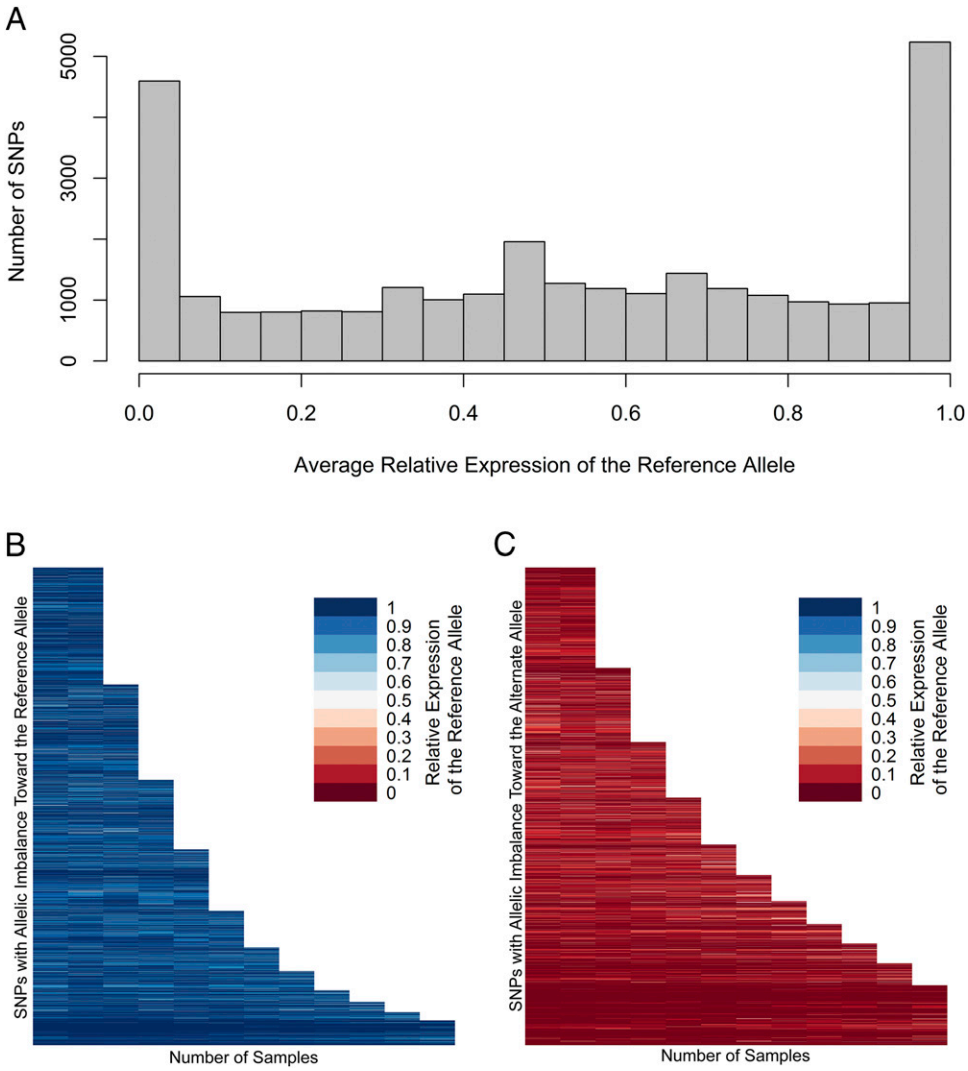

FIGURE 2. (**A**) Frequency distribution histogram of instances of allele-specific expression. The average relative expression (proportion of reads containing the reference allele) was calculated for all transcribed heterozygous SNPs identified in our study. Each bin spans a relative expression range of 0.05. (**B**) Variants in which the average relative expression of the reference allele is >0.5, and in which the average relative expression is <0.5 (**C**). The relative expression of the reference allele in each SNP with AI (binomial $p < 0.05$) in two or more samples is represented on the *y*-axis. The reference allele is defined by the genotype of the reference genome, which is consistent across all samples. Relative expression ranges from 0 (red) to 1 (blue). The AI of specific SNPs is shown to be highly consistent across individuals.

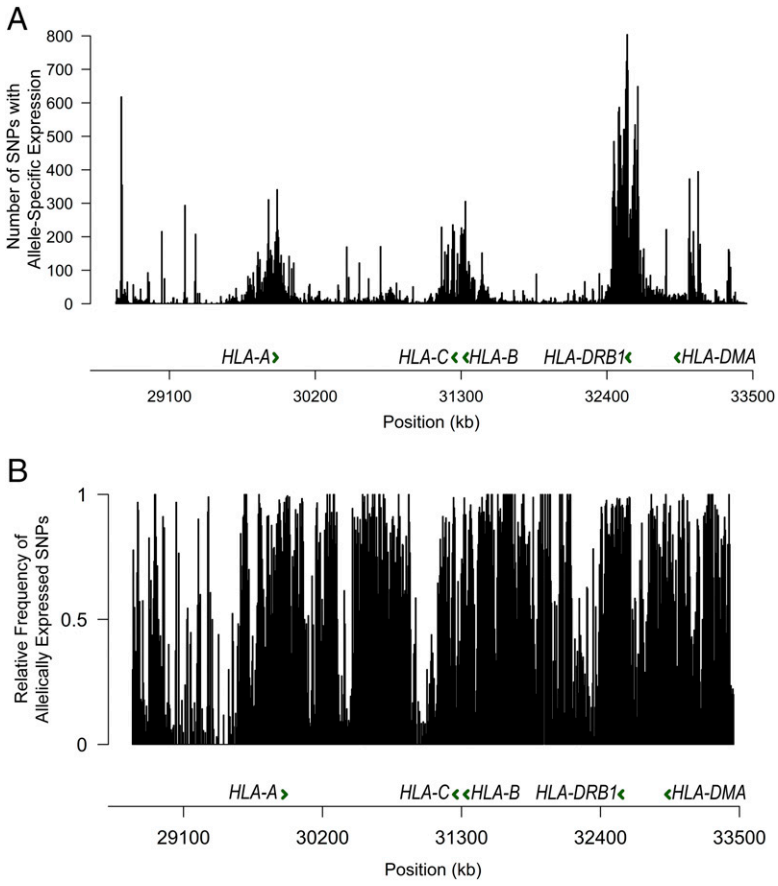

FIGURE 3. Histograms depicting the number of SNPs with allelic expression imbalance (**A**), and frequency of SNPs with allelic expression imbalance relative to all heterozygous SNPs detected in the MHC region (**B**). Each bin spans 5000 bp.

coexpression network including all aligned transcripts, based on the normalized read counts across all 12 samples, using a weighted correlation network analysis (29). Based on this network, the transcripts were grouped into 14 coexpression clusters, which do not localize to specific genomic regions within the MHC (Fig. 5). Nevertheless, coexpression remains highly aggregated within individual clusters, and there is a high degree of separation between each cluster within the network (Supplemental Fig. 3).

We further described transcription factor binding site enrichment in each cluster (Supplemental Table VI). Of the 14 coexpression clusters, 9 were enriched for specific transcription factors (OR $>$ 1, $p < 0.05$). For these clusters, the transcription factor binding sites most significantly enriched were TCF3 (OR: 2.17, p: 8.06×10^{-7}), ESR1 (OR: 2.72, p: 2.21×10^{-5}), RFX5 (OR: 1.68, p: 4.27×10^{-5}), SMARCA4 (OR: 2.56, p: 2.10×10^{-4}), GATA1 (OR:2.08, p: 2.25×10^{-4}), IKZF1 (OR: 4.80, p: 6.92×10^{-4}), GRp20 (OR: 2.57, p: 2.39×10^{-3}), CEBPD (OR: 1.52, p: 5.44×10^{-3}), and SMARCA4 (OR: 3.68, p: 6.74×10^{-3}). The enrichment of these specific transcription factor binding sites suggests that these nine clusters may show coexpression due to transcription factor–dependent coregulation.

Identification of novel retroviral genes with the human MHC

We identified three novel genes with an open reading frame that are predicted to be protein coding within the human MHC region, and demonstrate gene expression at the mRNA level. Using protein function and structure prediction algorithms, two of the three coding genes we identified are predicted with very high and moderate certainty to be novel endogenous retroviral *pol* and *gag* genes, respectively (Fig. 6). The structure and function of the third gene could not be predicted. The predicted amino acid sequences of all three genes were aligned to the human proteome using protein-protein BLAST (26). Based on the homology between each novel sequence and the human endogenous retroviruses to which it is aligned, we predict that these genes are retroviral *pol*, *gag*, and *gag* proteins, respectively (Supplemental Table VII).

Discussion

Variation within the MHC contributes to genetic risk of immune and inflammatory disease. However, this region is characterized by complex variation patterns that complicate identifying causal variants and their direct effects on disease etiology (32). Moreover, these complex variation patterns play a role in the complex alternative splicing and gene regulation networks that have been described in this region (7, 33). Quantification of MHC transcription by RNA sequencing has been limited by both the high rate of polymorphism and the high rate of splice variants, resulting in limitations in RNA sequence alignment (18). Using individualized genomes to map RNA sequencing reads, we accurately measured gene and splice variant expression within the MHC, which can be used to further elucidate the functional effects of variations relevant to disease.

Sequence variation can affect the expression of transcripts by interfering with *cis*-regulation, such as altering promoter or enhancer activity, altering DNA methylation patterns, or altering the sequence of regulatory RNAs. Variants linked to these *cis*-effects (*cis*–expression quantitative trait loci) affect expression in an allele-specific manner. Haplotype-specific gene expression within the HLA, and AI linked to *cis*–expression quantitative trait loci in autoimmunity have been previously described (4, 34). Our findings suggest extensive allele-specific expression throughout the MHC, which involves 88% of all transcribed SNPs in this region.

Many lincRNA transcripts are expressed at low levels, rendering them undetectable without sequence enrichment. By targeting the

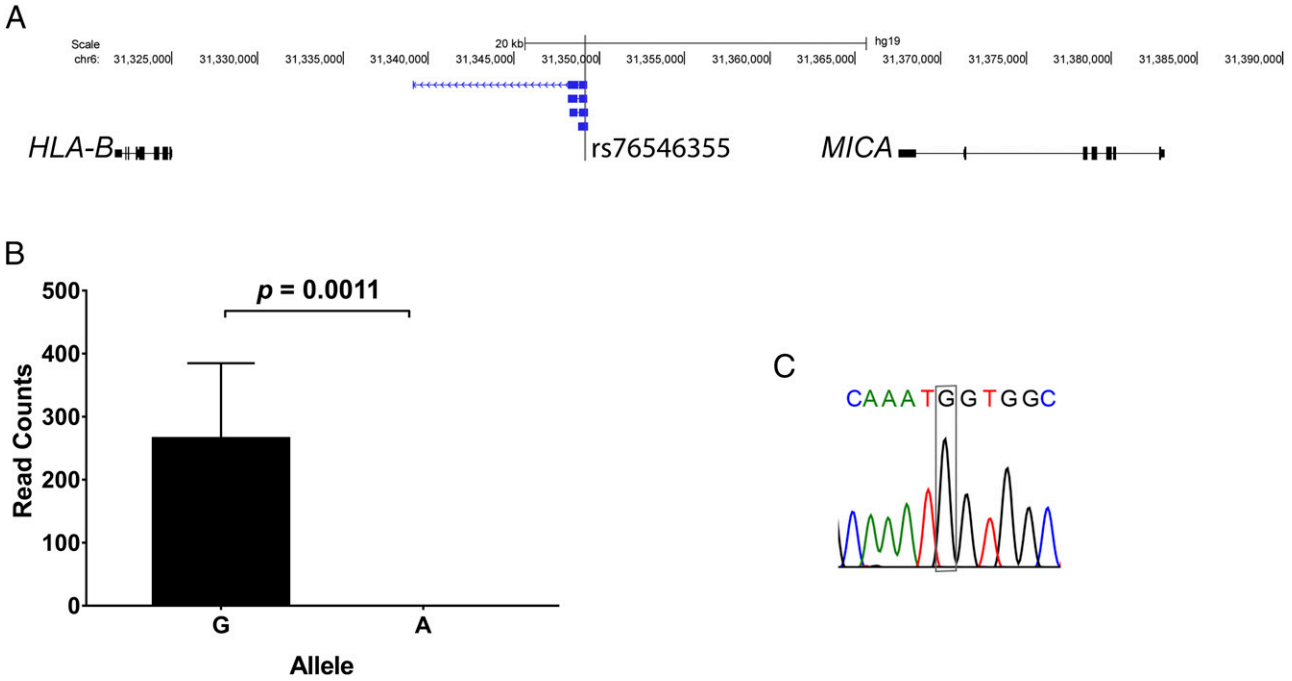

FIGURE 4. (**A**) The genetic variant rs76546355 (rs116799036), which explains the most robust genetic association for Behçet's disease and previously thought to be in a nontranscribed genetic region, is expressed within four lincRNA transcripts between *HLA-B* and *MICA*. (**B**) RNA sequencing revealed that these lincRNA transcripts are exclusively expressed from the disease protective allele (allele G), and no expression was detected from the disease risk allele (allele A) in heterozygous samples. RT-PCR followed by Sanger sequencing confirmed expression of the novel lincRNA transcripts in this locus, and allele expression imbalance in rs76546355 (a representative chromatogram of seven heterozygous samples is shown) (**C**).

MHC region using sequence-specific capture probes, we identified novel noncoding transcripts throughout the region. As lincRNAs have been implicated in transcriptional regulation, this suggests a far more complex regulatory network within the MHC than has been previously described. Variation within the MHC further affects the patterns of transcription regulation, due to AI as we demonstrate.

The genetic association of polygenetic diseases within the HLA is complex, and often the identification of causal genetic variants is complicated by the extensive linkage disequilibrium within this region. Although specific amino acid residues and classical HLA allelotypes have been considered to contribute to disease pathogenesis in several immune-mediated diseases, our data highlight the importance of including regulatory effects of these disease-associated polymorphisms to better understand the functional role of genetic variants within the HLA.

When we compare the expression patterns of all transcripts across all 12 sequenced individuals, a pattern of coexpression is observed. Although the coexpression of genes does not intrinsically imply coregulation, regulation by the same transcription factors is one mechanism by which coexpression can occur. After quantifying the enrichment of the transcription factors binding to the

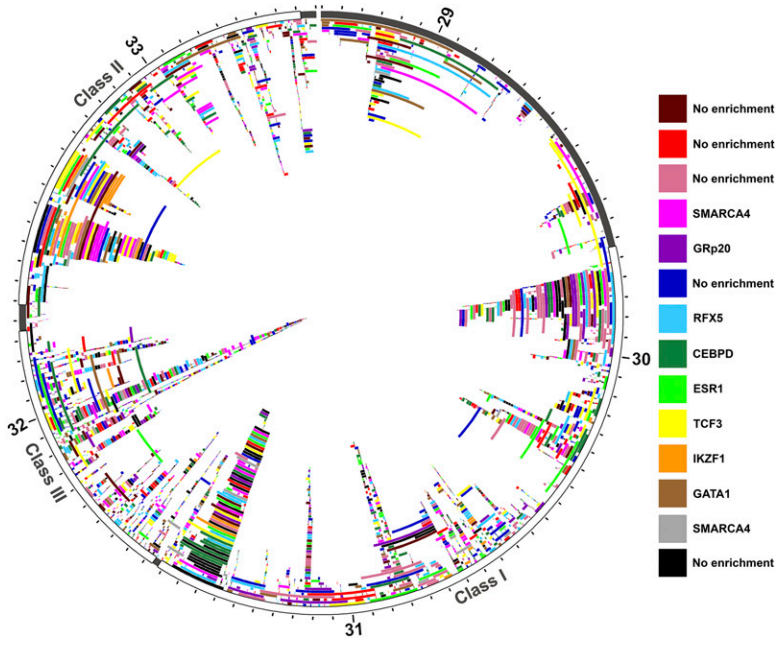

FIGURE 5. All unique transcripts plotted according to genomic position within the MHC region [chromosome 6: 28.7–33.5 Mb (hg19)]. Chromosome 6 position (labeled in megabases) is plotted on the outer ideogram, and each MHC class is marked. Each aligned transcript, including novel transcripts, was grouped into coexpression clusters using the normalized read counts from each sequenced individual ($n = 12$). Every transcript is plotted according to position, and colored according to cluster identity (red, dark red, orange, yellow, lime green, green, light blue, dark blue, purple, magenta, pink, black, gray, and brown). Multiple isoforms of the same gene can be found in the same coexpression cluster, but this is not a requirement and is never the case across all isoforms of a gene. There is no evidence for colocalization based on genomic positions of transcripts within individual coexpression clusters. Nine of the fourteen were enriched for specific transcription factors; the most significantly enriched transcription factor for each cluster is listed.

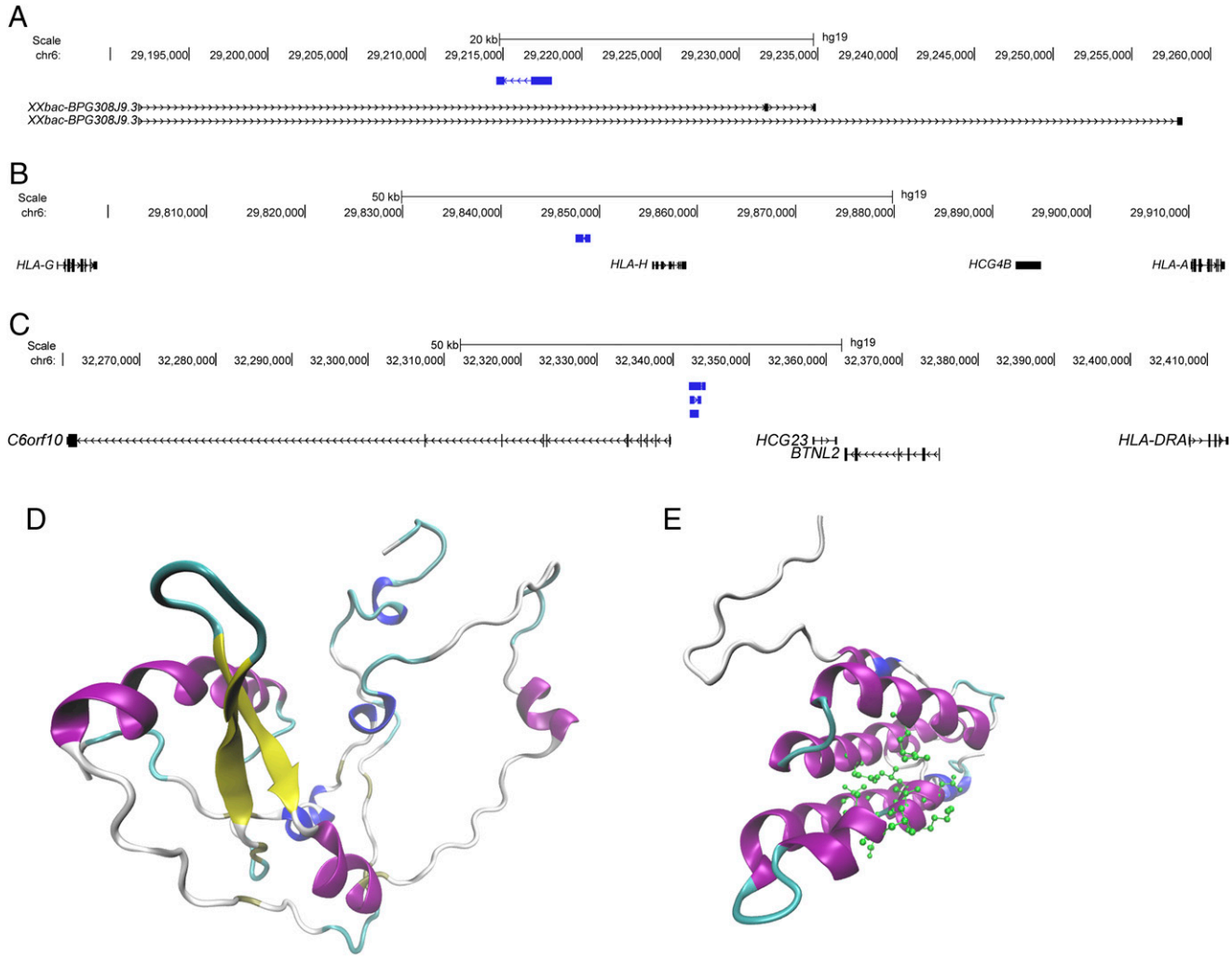

FIGURE 6. Genomic position (hg19) and predicted protein structure of the three novel protein-coding genes identified in this study. (**A**) One protein-coding novel transcript (blue) is within the intronic region of the gene *XXbac-BPG308J9.3*. (**B**) and (**C**) depict novel protein coding transcripts (blue) in intergenic regions, near *HLA-A* and *HLA-DRA*, respectively. Each of these transcripts shares homology with ERVs. (**D**) The predicted protein structure of transcript A (prediction p value = 1.17×10^{-4}). This structure shares homology with an endogenous retroviral pol protein, and no predicted ligands are available. (**E**) The predicted protein structure of transcripts C (prediction p value = 0.037). This structure shares homology with a retroviral gag protein, and is predicted to bind to a leucine residue (predicted active site amino acids are shown in green). In both (D and E), coloring is based on secondary structure: α helices are purple, 3–10 helices are blue, β sheets are yellow, β bridges are tan, turns are cyan, and coils are white.

promoter regions of the transcripts in each cluster, we found that 9 of the 14 clusters were enriched for specific transcription factors. This suggests that regulation by these transcription factors may play a role in the expression patterns of the transcripts in each cluster. Some of the enriched transcription factors identified play a role in specific immunological processes. For example, one of these coexpression clusters was found to be enriched for RFX5, a transcription factor that activates MHC class II expression by enhancing CIITA activity (35). Another transcription factor, enriched in a different cluster, CEBPD, is directly involved in promoting macrophage activation, M1 macrophage polarization, and proinflammatory cytokine production in macrophages (36). The transcription factor GATA1 is involved in dendritic cell differentiation and survival (37). Each of these enriched transcription factors has a unique role in monocyte differentiation, suggesting that they have a role not only in determining coexpression patterns of transcripts, but also in downstream determination of cellular phenotypes.

We found five novel putative coding transcripts, identifying three novel human endogenous retroviral elements (ERVs). ERVs comprise 8% of the human genome (38). Though mutations have silenced the expression of the majority of these elements, ~7% of all known ERVs are transcriptionally active (39). Moreover, mutations in these elements have been linked to diseases, including cancer (33) and multiple sclerosis (40). Translated ERVs have been shown to play a role in lymphocyte activation, and transcribed ERVs play a role in transcriptional regulation (41). The exact function of these novel ERVs, and their precise effects on transcription and immune function, has yet to be fully elucidated.

In summary, we performed deep sequencing of both the genome and the transcriptome, targeting the MHC region with sequence-specific capture probes in human monocytes. We accurately identified and quantified the expression of 908 novel transcripts in this region, including 123 transcripts aligning to regions previously thought to be intergenic. In addition, we uncovered extensive allele-specific expression imbalance within the MHC region, which appears to be nonstochastic, suggesting complex *cis*-acting transcriptional regulation throughout the human MHC. This AI can have functional consequences upon disease risk loci within the MHC region.

Disclosures

The authors have no financial conflicts of interest.

References

1. Trowsdale, J., and J. C. Knight. 2013. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics Hum. Genet.* 14: 301–323.
2. Hosomichi, K., T. Shiina, A. Tajima, and I. Inoue. 2015. The impact of next-generation sequencing technologies on HLA research. *J. Hum. Genet.* 60: 665–673.
3. Deitiker, P., and M. Z. Atassi. 2015. MHC genes linked to autoimmune disease. *Crit. Rev. Immunol.* 35: 203–251.
4. Raj, P., E. Rai, R. Song, S. Khan, B. E. Wakeland, K. Viswanathan, C. Arana, C. Liang, B. Zhang, I. Dozmorov, et al. 2016. Regulatory polymorphisms modulate the expression of HLA class II molecules and promote autoimmunity. *Elife* 5: e12089.
5. Hughes, T., P. Coit, A. Adler, V. Yilmaz, K. Aksu, N. Düzgün, G. Keser, A. Cefle, A. Yazici, A. Ergen, et al. 2013. Identification of multiple independent susceptibility loci in the HLA region in Behçet's disease. *Nat. Genet.* 45: 319–324.
6. Fairfax, B. P., S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, and J. C. Knight. 2012. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* 44: 502–510.
7. Fehrmann, R. S., R. C. Jansen, J. H. Veldink, H. J. Westra, D. Arends, M. J. Bonder, J. Fu, P. Deelen, H. J. Groen, A. Smolonska, 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 Genet.* 7: e1002197.
8. Davidovich, C., and T. R. Cech. 2015. The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA* 21: 2007–2022.
9. Zhao, Y., H. Sun, and H. Wang. 2016. Long noncoding RNAs in DNA methylation: new players stepping into the old game. *Cell Biosci.* 6: 45.
10. Bumgarner, S. L., G. Neuert, B. F. Voight, A. Symbor-Nagrabska, P. Grisafi, A. van Oudenaarden, and G. R. Fink. 2012. Single-cell analysis reveals that noncoding RNAs contribute to clonal heterogeneity by modulating transcription factor recruitment. *Mol. Cell* 45: 470–482.
11. Mercer, T. R., M. B. Clark, J. Crawford, M. E. Brunck, D. J. Gerhardt, R. J. Taft, L. K. Nielsen, M. E. Dinger, and J. S. Mattick. 2014. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* 9: 989–1009.
12. Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.
13. McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303.
14. DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43: 491–498.
15. Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 43: 11.10.1–11.10.33.
16. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.
17. Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22: 568–576.
18. Wu, T. D., and S. Nacu. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26: 873–881.
19. Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10–12.
20. Pertea, M., G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, and S. L. Salzberg. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33: 290–295.
21. Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7: 562–578.
22. Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE project. *Genome Res.* 22: 1760–1774.
23. Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25: 1915–1927.
24. Wang, L., H. J. Park, S. Dasari, S. Wang, J. P. Kocher, and W. Li. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41: e74.
25. McGuffin, L. J., J. D. Atkins, B. R. Salehe, A. N. Shuid, and D. B. Roche. 2015. IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Res.* 43(W1): W169–W173.
26. Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.
27. Brown, N. P., C. Leroy, and C. Sander. 1998. MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics* 14: 380–381.
28. Koboldt, D. C., D. E. Larson, and R. K. Wilson. 2013. Using VarScan 2 for germline variant calling and somatic mutation detection. *Curr. Protoc. Bioinformatics* 44: 15.4.1–15.4.17.
29. Langfelder, P., and S. Horvath. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9: 559.
30. Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res.* 19: 1639–1645.
31. Dozmorov, M. G., L. R. Cara, C. B. Giles, and J. D. Wren. 2016. GenomeRunner web server: regulatory similarity and differences define the functional impact of SNP sets. *Bioinformatics* 32: 2256–2263.
32. Horton, R., R. Gibson, P. Coggill, M. Miretti, R. J. Allcock, J. Almeida, S. Forbes, J. G. Gilbert, K. Halls, J. L. Harrow, et al. 2008. Variation analysis and gene annotation of eight MHC haplotypes: the MHC haplotype project. *Immunogenetics* 60: 1–18.
33. Goering, W., K. Schmitt, M. Dostert, H. Schaal, R. Deenen, J. Mayer, and W. A. Schulz. 2015. Human endogenous retrovirus HERV-K(HML-2) activity in prostate cancer is dominated by a few loci. *Prostate* 75: 1958–1971.
34. Vandiedonck, C., M. S. Taylor, H. E. Lockstone, K. Plant, J. M. Taylor, C. Durrant, J. Broxholme, B. P. Fairfax, and J. C. Knight. 2011. Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. *Genome Res.* 21: 1042–1054.
35. Reith, W., S. LeibundGut-Landmann, and J. M. Waldburger. 2005. Regulation of MHC class II gene expression by the class II transactivator. *Nat. Rev. Immunol.* 5: 793–806.
36. Ko, C. Y., W. C. Chang, and J. M. Wang. 2015. Biological roles of CCAAT/enhancer-binding protein delta during inflammation. *J. Biomed. Sci.* 22: 6.
37. Gutiérrez, L., T. Nikolic, T. B. van Dijk, H. Hammad, N. Vos, M. Willart, F. Grosveld, S. Philipsen, and B. N. Lambrecht. 2007. Gata1 regulates dendritic-cell development and survival. *Blood* 110: 1933–1941.
38. Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al; International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. [Published erratum appears in 2001 *Nature* 411: 720; 412: 565.] *Nature* 409: 860–921.
39. Oja, M., J. Peltonen, J. Blomberg, and S. Kaski. 2007. Methods for estimating human endogenous retrovirus activities from EST databases. *BMC Bioinformatics* 8(Suppl. 2): S11.
40. Brütting, C., A. Emmer, M. Kornhuber, and M. S. Staege. 2016. A survey of endogenous retrovirus (ERV) sequences in the vicinity of multiple sclerosis (MS)-associated single nucleotide polymorphisms (SNPs). *Mol. Biol. Rep.* 43: 827–836.
41. Suntsova, M., A. Garazha, A. Ivanova, D. Kaminsky, A. Zhavoronkov, and A. Buzdin. 2015. Molecular functions of human endogenous retroviruses in health and disease. *Cell. Mol. Life Sci.* 72: 3653–3675.