Evolution of DNAase I Hypersensitive Sites in MHC Regulatory Regions of Primates

Yabin Jin,*,†,1 Rachel M. Gittelman,‡,1 Yueer Lu,* Xiaohui Liu,§ Ming D. Li, Fei Ling,*,2 and Joshua M. Akey‡,2**

*School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, P. R. China, †Clinical Research Institute, The First People's Hospital of FoShan (Affiliated FoShan Hospital of Sun Yat-sen University), 528000, China, ‡Department of Genome Sciences, University of Washington, Seattle, Washington 98125, §Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100094, P. R. China, and **State Key Laboratory of Diagnosis and Treatment of Infectious Diseases, First Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou 310009, P. R. China

ORCID ID: 0000-0002-3640-4280 (F.L.)

ABSTRACT It has been challenging to determine the disease-causing variant(s) for most major histocompatibility complex (MHC)-associated diseases. However, it is becoming increasingly clear that regulatory variation is pervasive and a fundamentally important mechanism governing phenotypic diversity and disease susceptibility. We gathered DNase I data from 136 human cells to characterize the regulatory landscape of the MHC region, including 4867 DNase I hypersensitive sites (DHSs). We identified thousands of regulatory elements that have been gained or lost in the human or chimpanzee genomes since their evolutionary divergence. We compared alignments of the DHS across six primates and found 149 DHSs with convincing evidence of positive and/or purifying selection. Of these DHSs, compared to neutral sequences, 24 evolved rapidly in the human lineage. We identified 15 instances of transcription-factor-binding motif gains, such as *USF*, *MYC*, *MAX*, *MAFK*, *STAT1*, *PBX3*, *etc*., and observed 16 GWAS (genome-wide association study) SNPs associated with diseases within these 24 DHSs using FIMO (Find Individual Motif Occurrences) and UCSC (University of California, Santa Cruz) ChIP-seq data. Combining eQTL and Hi-C data, our results indicated that there were five SNPs located in human gains motifs affecting the corresponding gene's expression, two of which closely matched DHS target genes. In addition, a significant SNP, rs7756521, at genome-wide significant level likely affects DDR expression and represents a causal genetic variant for HIV-1 control. These results indicated that species-specific motif gains or losses of rapidly evolving DHSs in the primate genomes might play a role during adaptation evolution and provided some new evidence for a potentially causal role for these GWAS SNPs.

KEYWORDS positive selection; purifying selection; DNase I hypersensitive sites; major histocompatibility complex; primate genome; regulatory variation; GenPred; Shared Data Resources; Genomic Selection

THE human major histocompatibility complex (MHC) contains ~260 genes in a ~4-Mb span on chromosomal region 6p21.3. As expected, nonhuman primate MHCs are similar to the human MHC, although they have different numbers of MIC (MHC class I chain)-related genes associated with class I. The MHC region is associated with more diseases (mainly autoimmune and infectious) than any other region of the genome (Trowsdale and Knight 2013). Interestingly, conditions other than infections and autoimmunity are also associated with the MHC, including some cancers and neuropathies (Trowsdale 2011; Trowsdale and Knight 2013). One of the major limitations in our understanding of how the MHC region contributes to diseases is our incomplete knowledge of the allelic variation of genes and considerable variation in regions flanking the classes I and II polymorphic regions (Spurgin and Richardson 2010). The extreme patterns of diversity in the MHC region are indicative of strong selection, and heterozygote advantage (balancing selection), frequency-dependent selection, and fluctuating selection due to continual change in pathogen type and abundance (Spurgin and Richardson 2010) have all been invoked as forces acting on the MHC region. Besides,

doi: https://doi.org/10.1534/genetics.118.301028
Manuscript received March 18, 2018; accepted for publication April 16, 2018; published Early Online April 18, 2018.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.6151235.
[1]These authors contributed equally to this work.
[2]Corresponding authors: School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, P. R. China. E-mail: fling@scut.edu.cn; and Department of Genome Sciences, University of Washington, Seattle, WA 98125. E-mail: jakey@princeton.edu

the extended linkage disequilibrium also makes it difficult to identify causal variants in MHC–disease associations. It is becoming increasingly clear that structural variation alone cannot fully account for disease associations in the MHC region, and there is an increasing interest in defining genetic variants that may modulate gene expression (Handunnetthi *et al.* 2010). Such a statement has gained support from recent expression quantitative trait loci (eQTL) studies, which have highlighted the impact of *cis*-acting genetic variation on expression of MHC genes such as *HLA-B*, *HLA-DQA1*, *HLA-DRB1*, *HLA-DPA1*, and *HLA-DQB* (Handunnetthi *et al.* 2010). These findings indicate that gene expression differences, rather than protein-coding changes, could underlie some of the observed disease associations. For instance, some viruses, as well as many tumors, employ strategies to down-modulate HLA expression to escape T cell recognition. The best example of this phenomenon is HLA-C levels in HIV infection (Trowsdale and Knight 2013). HLA-C expression levels, rather than specific protein-coding variants, may have the greatest influence on HIV control (Trowsdale and Knight 2013).

Additionally, regulatory variants have been linked to the susceptibility to various human diseases outside the MHC region, including infectious, autoimmune, psychiatric, neoplastic, and neurodegenerative disorders (Clop *et al.* 2013). A large number of genome-wide association (GWA) studies have also been performed yielding hundreds of novel genomic locations associated with phenotypic variation or disease susceptibility (Manolio 2010). In these studies, ~81% of associated SNPs are located in noncoding regions, although this is likely influenced by how SNPs were ascertained (Hindorff *et al.* 2009). In addition to pathogenic variants, regulatory variation has long been hypothesized to significantly contribute to evolutionary changes among human populations and between species (King and Wilson 1975; Fraser 2013). Numerous noncoding regions that are rapidly evolving in the human lineage have been identified, one of which shows enhancer function unique to the developing human forelimb (Prabhakar *et al.* 2006, 2008). In short, regulatory variation is pervasive and a fundamentally important mechanism governing phenotypic diversity and disease susceptibility.

The DNase I assay has been proven to be a highly successful and extensively validated methodology for discovery of *in vivo* regulatory sequences in complex genomes (Dorschner *et al.* 2004; Sabo *et al.* 2006; ENCODE Project Consortium *et al.* 2007; Hesselberth *et al.* 2009; ENCODE Project Consortium 2012). As part of the ENCODE Project and Roadmap Epigenomics Project, extensive maps of DHS have been created in over 140 cell types and high-resolution DNase I footprints in over 30 cell types. Previously, our research group has revealed new insights into conserved and adaptive regulatory DNA in humans and refined the set of genomic substitutions that distinguish humans from their closest living primate relatives (Gittelman *et al.* 2015).To highlight the types of inferences possible by superimposing evolutionary and functional genomics data sets, we analyzed DHS in the MHC region identified in the ENCODE Project (ENCODE Project Consortium 2004, 2012) and fibroblasts of three species from Crawford (Shibata *et al.* 2012), including human, chimpanzee, and macaque. Using the publicly available six primate (human, chimpanzee, gorilla, orangutan, macaque, marmoset) EPO alignments from Ensembl (Neph *et al.* 2012) to obtain sequence alignments for each DHS, we analyzed sequence conservation and human-specific acceleration of DHS and identified thousands of regulatory elements that have been gained or lost in the human or chimpanzee genomes since their evolutionary divergence. Polymorphic DNA bases in transcription factor motifs that we found in these regulatory elements may likely be responsible for the varied biological functions across species.

Materials and Methods

DHS data were obtained from 136 cell types published by the ENCODE Project group. More information about the 136 cell lines are available at http://genome.ucsc.edu/ENCODE/cellTypes.html. DHS data from another 15 cell types were published by the Crawford group at Duke University (Shibata *et al.* 2012), including primary skin fibroblast cells from three human, three chimpanzee, and three macaque individuals and LCLs, which are B cells immortalized with Epstein-Barr Virus, which were obtained from the same three human and three chimpanzee individuals but not macaque individuals as EBV did not reliably transfect macaque lymphocyte cells. All of these data were converted to hg19 using the UCSC liftover tool for consistency. We obtained DNase I peaks, footprints, and predicted motif locations from the ENCODE Project (http://genome.ucsc.edu/ENCODE/downloads.html) (ENCODE Project Consortium 2004). Peaks and footprints were empirically thresholded at a 1% false-discovery rate. For aggregate analyses over DHS across cell types, DHS peaks or footprints were merged across cell types using BEDOPS (Neph *et al.* 2012). We annotated DHS as intergenic, promoter, exon coding, intron, 5′UTR, or 3′UTR using annotation data from the UCSC genome browser. Because some DHSs are overlapped with multiple annotations, we chose the annotation with larger overlapped regions. As for merging overlapped DHS, we considered multiple overlap cutoffs, including 90, 50, 10, 5, and 1%, and finally set the cutoff at 50%, meaning that two DHSs needed to overlap by 50% in order to be merged.

The approaches used for evolution analysis were the same as those described in our previous study (Gittelman *et al.* 2015). First of all, we obtained the EPO (Enredo-Pecan-Ortheus) six primates alignment of each DHS from Ensembl, repeats and human polymorphic bases in which the derived allele was at <95% in the 1000 Genomes Phase I data, as well as CpGs that were masked in each DHS. Then we obtained the ±50-kb alignments of each DHS from EPO to establish its neutral model. Within these alignments, repeats, human polymorphic bases, CpGs, exons splice sites, promoters (500 bp upstream of TSS), other DHS, segmental duplications, and phastCons conserved regions were masked to prevent confounding our estimate of the neutral model. <90% of the remaining unmasked bases in DHS or <15 kb of the

remaining unmasked bases in the defined neutral region were filtered out. After filtering, we calculated a neutral model of evolution based on neutral alignment with phyloFit (PHAST package), with the parameters: -nrates 4 – subst -mod SSREV -EM. Three LRT tests were run using Phylop in the PHAST package (http://compgen.bscb.cornell.edu/phast/) (Pollard *et al.* 2010), which identified DHSs that were evolving at a different rate than the neutral model across all species, faster on the human branch or at a different rate than the neutral model across all species expect human, corresponding to the CONACC LRT test, ACC LRT test, and CON LRT test. The detailed parameters of three LRT tests were -method LTR -mode CON, -method LTR -mode ACC -subtree homo_sapiens, and -method LTR -mode CON after removing the human sequence from the alignment, respectively.

The potential transcription factor motifs were predicted by FIMO, version 4.6.1, using a *P*-value threshold of $\leq 1 \times 10^{-5}$ (Grant *et al.* 2011) and were obtained from ChIP-seq data in UCSC. Motif models were obtained from TRANSFAC (Wingender *et al.* 2000), version 2011.1, JASPAR database, and ChIP-seq data. For these analyses, all motifs were intersected with footprint data using BEDOPS. The eQTL data were obtained from the GTEx Analysis V6p database in UCSC. Hi-C data were obtained from the published research (Maurano *et al.* 2012; Thurman *et al.* 2012). To detect the homininae gains variation, we used the "liftover" tools in UCSC to obtain the homologous sequences of other species and compared them.

Data availability

The data sets supporting the results of this article are included within the article and its Supplemental Materials. Figure S1: A total of 4867 DHSs were detected in the MHC region (±500 kb from 29570004 to 33290793 in chr6) in 136 human cells. Of these DHSs, 2763, 1595, and 509 were detected in MHC class I, MHC class II, and other MHC regions, respectively; Figure S2: DHSs in MHC and other regions of all human cells from UCSC; Figure S3: Distribution of DHSs in MHC and non-MHC regions; Figure S4: A total of 1885, 2507, and 1736 DHSs were detected in the MHC region from human, chimpanzee, and macaque fibroblasts, respectively; Figure S5: The distance of the common DHSs to the nearest gene was closer than the species-specific DHS; Figure S6: Using different criteria for overlap across three species; Table S1: 149 DHSs underwent significantly positive selection compared to neutral model; Table S2: TF gains or loss across species; Table S3: Six SNPs from the 1000 Genomes Project located in five DHSs; Table S4: seven GWAS SNPs in six DHSs with significant conservation across primates. Supplemental material available at Figshare: https://doi.org/10.25386/genetics.6151235.

Results

Distribution of DHSs in MHC region

We detected 4867 DHSs in the MHC region on chromosome 6 across 136 human cell types. Of these DHSs, 2783, 1595, and 509 were detected in MHC class I, MHC class II, and other MHC regions, respectively (Figure S1). Ninety-three percent of DHSs fell in noncoding regions with respect to GENECODE gene annotations. Approximately 50% of DHSs were located in intergenic regions, which was consistent with the previous results (Shibata *et al.* 2012; Thurman *et al.* 2012). The MHC class I region had the highest proportion of intergenic DHSs (55.8%; Figure S2).

We next compared DHSs across species. We detected 1885, 2507, and 1736 DHSs in the MHC region from human, chimpanzee, and macaque fibroblasts, respectively. Interestingly, the same number of DHSs was also observed in separate human fibroblast data sets (data not shown). Surprisingly, there was a lower proportion of macaque intergenic DHSs compared with human and chimpanzee fibroblasts. The proportion was 47.5, 59.2, and 57.8% in macaque, human, and chimpanzee fibroblasts, respectively (Figure S4). Such difference is probably due to MHC class I duplication in macaque, especially high duplication of the Mamu MHC-B gene.

Using stringent criteria (50% quartile) for overlap across three species, we identified 815 human-specific DHSs, 1402 chimpanzee-specific DHSs, 820 macaque-specific DHSs, and 1284 common DHSs (Figure 1, also see Figure S6 for the results from different criteria). We also found that species-specific DHSs are reduced in promoter regions relative to common DHSs and enriched in distal intergenic regions and introns. The distances of the common DHSs to the nearest gene were closer than the species-specific DHSs (Figure S5).

Species-specific DHS sites show evidence of selection in MHC region

The functional interpretations of common and species-specific DHSs led to predictions about the operation of natural selection. We used a neutral model of evolution to test the DHSs for evidence of selection. First, we obtained an alignment for each DHS using the six Primates EPO alignments from Ensembl. To estimate a separate local neutral model for each DHS, we also obtained an alignment of the 50 kb surrounding each DHS. After alignment filters, 301 DHSs remained to be tested with the PHAST package (http://compgen.bscb.cornell.edu/phast/). Using the phyloP program from the PHAST package, we conducted three likelihood ratio tests on each DHS: lineage-specific rate acceleration on the human branch, positive selection in all six branches, and conservation. We found that a total of 149 DHSs had undergone significantly positive selection or conservation. Of these DHSs, relative to the neutral sequence, 1 was accelerated across all six primates, 118 were conserved across all six primates, 131 were conserved across all five primates except human, and 24 evolved rapidly on the human lineage (Figure 2). Reportedly, the clustering of antigen-processing and antigen-presenting genes in the MHC is consistent with the idea that the region evolved from a block of duplicated immune system genes. Among these 149 DHSs, 43 were highly correlated ($r > 0.7$) with the promoter of at least one nearby gene across cell types, including 11, 13, and 4 genes related to MHC class I, MHC class II, and antigen-processing genes, respectively. Importantly, 8 DHSs that evolved rapidly on the human lineage

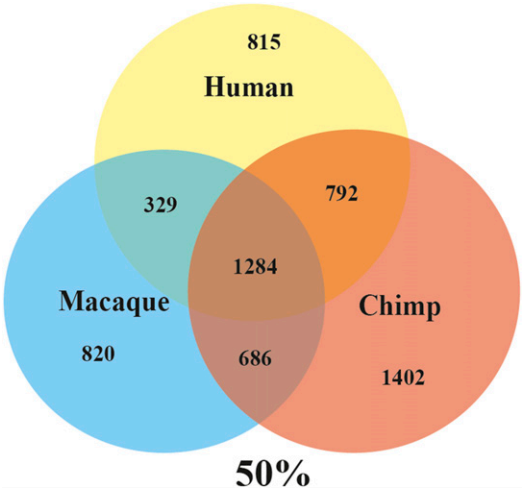

Figure 1 Using stringent criteria (50% quantile) for overlap across three species, we identified 815 human-specific DHSs, 1402 chimpanzee-specific DHSs, 820 macaque-specific DHSs, and 1284 common DHSs.

were correlated with nearby gene promoters, including 4, 6, and 1 gene(s) relevant to MHC class I, class II, and antigen processing (Table 1), respectively. The expression of the four MHC class I genes *HLA-A*, *-E*, *-F*, and *-G* may thus be evolving rapidly compared to *HLA-B* and *-C*. It is likely that the expression of the five MHC class II genes *HLA-DPB1*, *-DQA1*, *-DQB1*, *-DQB2*, and *-DMB*, and *HLA-DRB6* may thus be evolving rapidly compared to the other MHC class II genes (Table 1).

Specific motif analysis of rapidly evolving DHSs in the primate genomes

Next, we analyzed transcription factor motifs in the sets of species-specific and rapidly evolving DHSs by using eQTL, Hi-C, and ChIP-seq data from UCSC. The potential transcription factor motifs were predicted by FIMO (see *Materials and Methods* section) and were obtained from ChIP-seq data in UCSC. We used both the JASPAR database and ChIP-seq data to define binding site gains as cases where the sequence was a perfect match to the motif consensus sequence in one species only. Similarly, losses were defined as matching the consensus sequence in all but one species. As a result, we identified a total of 26 gained or lost binding sites including 15 human gains, 2 macaque gains, 1 orangutan gain, and 8 homininae gains (Table S2). Figure 3 shows examples of (Transcription factors) gain. All 15 human gains, which included 9 potential motifs (*SOX9*, *OLF1*, *ER*, *SP3*, *HIC1*, *GATA1*, *TP53*, *GLI3*, and *BEN*) from FIMO and 6 motifs obtained from ChIP-seq data (*USF*, *MYC*, *MAX*, *MAFK*, *STAT1*, and *PBX3*), occurred in DHSs that were rapidly evolving on the human lineage (Table 1). Importantly, by analyzing polymorphic data from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012), we found two SNPs in *SOX9* and *ER* motifs, respectively (Table S3). Interestingly, both SNPs were located in a key base pair of the two motifs. We found that there was low frequency (0.3%) of the SNP located in the *ER* motif, and it was only detected in the EAS population, suggesting that the SNP has been fixed in the human population. However, the frequencies of another SNP, rs7761563, identified in the *SOX9* motif with human gains differed among five populations, including EAS, AMR, AFR, EUR, and SAS populations, suggesting that the SNP rs7761563 has not yet been fixed in the human population. Importantly, analysis of eQTL data showed that the SNP rs7761563 located in the *SOX9* motif with human gains affected the expression of *HCG9* in human tissue. We also observed that SNP rs7741418 happened to be the core base (C) in *USF/MYC/MAX* motifs (Table 1), and the frequency of allele C is 68%, while the frequency of ancestral allele T is 32% in the human population. Notably, in other primates this variant was T.

The mutation of the ancestral allele T into C in humans implied that the binding affinity of the *USF/MYC/MAX* transcription factor differed in primates, resulting in the differential expression of DHS target genes. Additionally, analysis of eQTL data showed that rs7741418 affected the expression of *HCG9* in heart and lung tissues. Therefore, we proposed that the human-specific *SOX9/USF/MYC/MAX* motifs within human accelerated DHS may play an important role in the development of the disease by affecting *HCG9* expression. HCG9 is noncoding, but its function has not been determined and is positively associated with Behcet syndrome, glomerulonephritis, immunoglobulin A, lupus erythematosus, systemic sclerosis, multiple sclerosis, and vitiligo. Similarly, we observed that SNP rs114565826 was located in the core base of human-gain *PBX3* motif *UA2*, for which the allele frequency of C is 12% and for T is 98%, whereas in other primates this variant was a C. So, during adaptive evolution, we conjecture that the human ancestral base mutated to the T allele at this locus, obtaining the binding ability of *PBX3* (pre-B-cell leukemia transcription factor 3), an essential for leukemia development and a putative biomarker of aggressive prostate cancer. This resulted in the expression of the downstream target gene, suggesting that the SNP has been nearly fixed in the human population. Bizarrely, SNP rs9261450, with an allele frequency of 79% for A and of 21% for G, was located in a noncore base of the human-gain *STAT1* motif. According to eQTL data analysis, this mutation was found to affect the expression of *TRIM31/TRIM40* in testis, a target gene for the corresponding DHS. It is likely that the *STAT1* motif with A allele genotype showed more binding affinity than the G allele. In the DHS (chr6:30063120-30063775), there were two more human SNPs (rs576022961 and rs7765810) in addition to SNP rs7741418, which were located in the core base of TF motifs, indicating the possibility that these SNPs may affect the binding affinity of these TFs, thereby affecting target gene expression. In rs576022961, the G allele frequency was 0.02% and located in a *FOX* motif. *FOX* is a transcriptional activator for liver-specific transcripts such as albumin and transthyretin. But according to SNP databases, the core base corresponding to rs576022961 was A, and the target genes of this DHS included *TRIM31*. With data from UCSC, we predicted that rs576022961 was associated

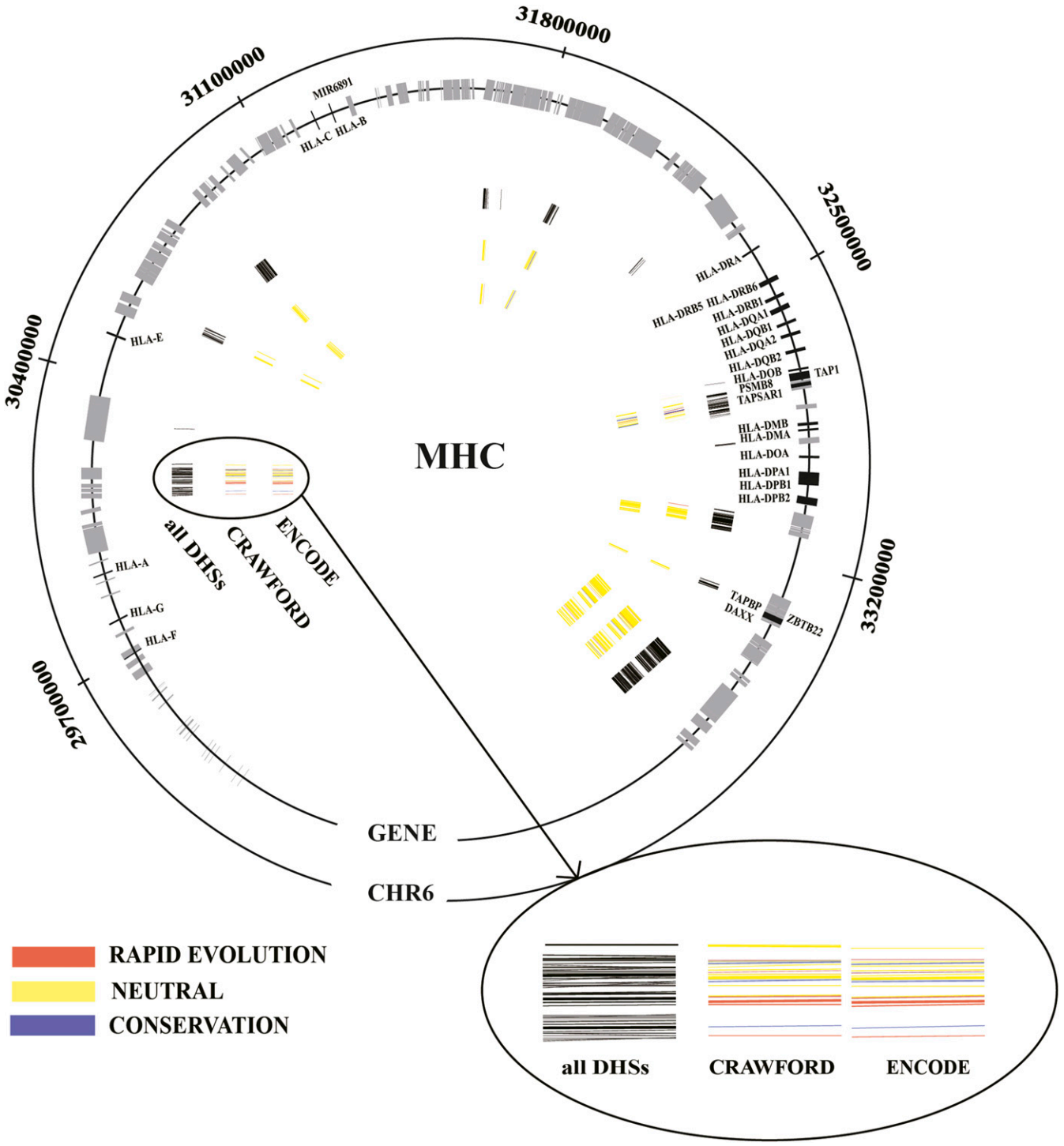

Figure 2 A total of 158 DHSs have undergone significantly positive selection. Of these DHSs, compared to neutral sequence, 1 was accelerated across all six primates, 127 were conserved across all six primates, 139 were conserved across all five primates except human, and 24 evolved rapidly on the human lineage.

with *TRIM31*. Whether this SNP would affect the expression of TRIM31 still needs further experimental verification. In addition, a high-frequency mutation site, rs7765810 (C > A, A: 31.929%; C: 68.071%), was located in the noncore base of the *HNF4* motif. However, at this locus, we found an eQTL that affects *HCG9* expression, suggesting that the mutation of the C allele to the A allele in the human branch may decrease the binding affinity of *HNF4* and therefore affects the expression of *HCG9*. But in our analysis, this gene was not the target gene for the corresponding DHS. Therefore, whether the SNP mutation affects the expression of *HCG9* through the binding of HNF4 still needs further experimental verification. In human

Table 1 The DHSs with rapid evolution in human lineage compared to all other species

Start	End	Target gene from HI-C data	TF from ChIP-seq data	Human-gain TF motif from FIMO and ChIP--seq data	eQTL	SNP located at the TF motif
30063120	30063775	ABCF1,TRIM15,TRIM31	FOSL2, FOS, JUND, RXRA, HNF4G, HNF4A, FOXA2, FOXA, MAX, MYC, USF1, USF2	FIMO: SOX9;	rs7761563/ HCG9, rs7741418/ HCG9	rs7741418/USF, MYC
				ChIP-seq: USF, MYC/MAX	rs7765810/HCG9	rs7765810/HNF4 rs576022961/FOXA
30094800	30095910	HLA-W, HCG4P3, TRIM40, TRIM39, DAQB-335A13.8, HCG4P8, HLA-G, HLA-U	none	FIMO: GATA1, HIC1, OLF1	rs9261447/ TRIM31, TRIM40	none
30097545	30100218	TRIM15, HLA-E, HLA-A, HLA-F, TRIM26, TRIM26P1, TRIM31, TRIM39, TRIM40	STAT1, MAFK, CTCF, CTCFL, PBX3	FIMO: ER(ESR1, ESR2); ChIP-seq: STAT1, MAFK, UA2	rs9261450/ TRIM31, TRIM40,STAT1	rs9261450/STAT1 rs984319427/UA2 rs114565826/UA2
30102968	30104164	HLA-P	HNF4A, FOXA1, FOXA2, STAT1	none	none	none
30128980	30129442	none	GATA1	none	none	none
30140442	30140975	TRIM26P1	none	none	none	none
30568140	30569070	HCG22, NRM	none	FIMO: SP3(3'UTR)	none	none
30573680	30574016	TRIM39, TRIM39-RPP21, HLA-E,	none	none	none	none
32808378	32808930	HLA-DRB6, HLA-DMB, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DQB2	RUNX3, EBF1	FIMO: GLI3(3'UTR)	none	rs112080182/RUNX1 rs554021812/EBF1
33664209	33665085	none	EBF1, PBX3, ZNF263, FOS, JUN	FIMO: BEN(3'UTR)	none	none
33693670	33695328	none	ZNF263	FIMO: P53_DECAMER, BEN	none	none
33760926	33761554	none	REST	none	none	rs114889417/REST rs779529174/REST rs758088611/REST

accelerated DHSs (chr6:32808378-32808930), we found that SNP rs112080182, with 0.719% A allele frequencies and 99.281% C allele frequencies, was located in the core of the *RUNX1* (as a tumor suppressor) motif. Individuals carrying the A allele may have different binding affinity of *RUNX1*, resulting in the down-expression of the target gene.

Interestingly, there were three consecutive SNPs (rs758088611, rs779529174, rs11488941) in the DHS (chr6:33760926-33761554), all of which were located within the core of the *REST* motif, implying that if an individual is carrying the mutation allele, it will most likely affect the binding affinity of the transcription factor, resulting in the expression of downstream target genes. Since *REST* acts as a master negative regulator of neurogenesis and as a tumor suppressor (predisposing to Wilms tumor), it is speculated that a mutation in human beings means that the binding capacity of the transcription factor is decreased or lost, resulting in the abnormal expression of the target gene. Surprisingly, we also observed loss of two motifs in this set of DHSs. *NRSF*, one motif loss, represses transcription by binding a DNA sequence element called the neuron-restrictive silencer element, and is known to be involved in Alzheimer's disease and Down syndrome (Lu *et al.* 2014, 2016). *LMO2COM*, the other motif loss, is an essential transcriptional regulator in hematopoiesis, whose inappropriate regulation frequently contributes to the development of leukemia (El Omari *et al.* 2011). Loss of both motifs suggested a possible role for the DHSs during evolution of MHC regulation. Of the eight instances of hominiae gains, one was an *HNF1* transcription factor binding motif in the DHS accelerated across all six primates (Table S2). *HNF1* regulates the insulin gene and glucose transport and metabolism (Lee *et al.* 1998; Pontoglio 2000), suggesting that the DHS with hominiae gains may play a role in obesity. Collectively, species-specific motif gains or losses of rapidly evolving DHSs in primate genomes could play a role during adaptation evolution by affecting the target gene's expression

GWAS SNPs harboring DHSs provide insight into the causative variation underlying diseases associated with the MHC region

The MHC region is associated with a variety of diseases and other traits, but it has proven difficult to identify causal variants because of the high gene density, strong linkage disequilibrium, and complex interactions among MHC genes within the region. Reportedly, disease- and trait-associated variants are concentrated in regulatory DNA (Maurano *et al.*

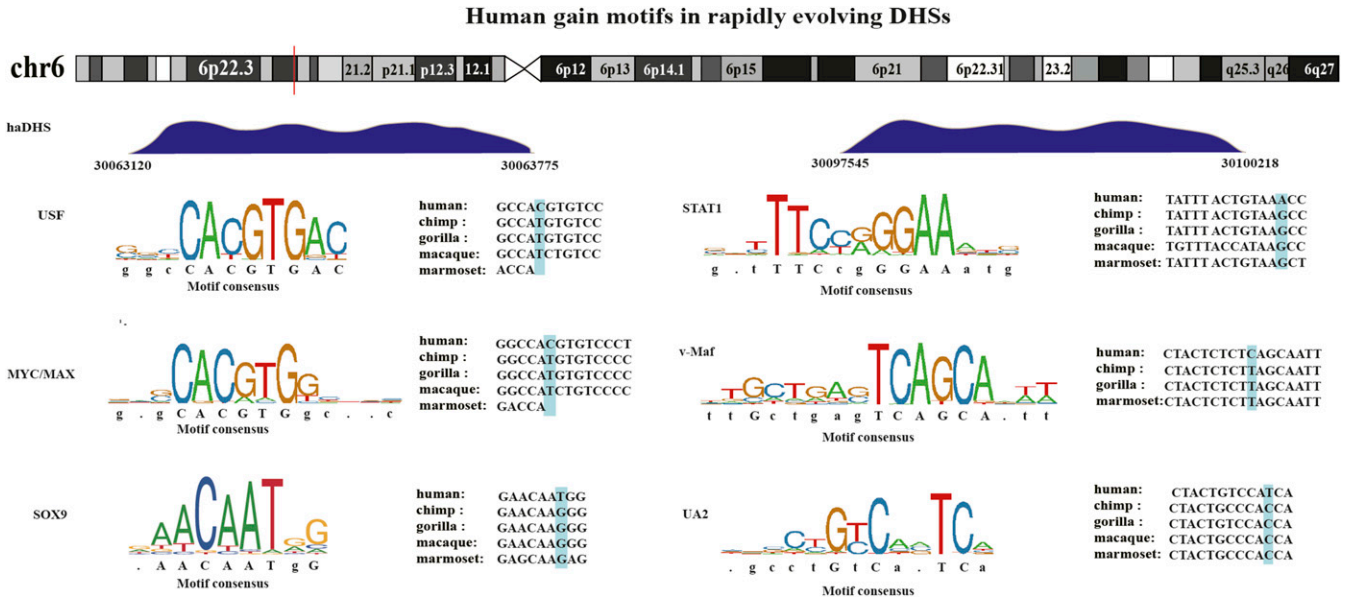

Figure 3 A list of gained TFs in human genome.

2012). We employed the EMBL-EBI GWAS database to identify GWAS variants that localize to MHC DHSs. We identified seven GWAS SNPs in six DHSs with significant conservation across primates, which were associated with control of HIV infection (HIV-1 control), age-related macular degeneration, low-density lipoprotein (LDL) cholesterol, height, educational attainment, and obesity-related traits (Table S4).

Among these SNPs, we found SNP rs541862 and rs2254287. One SNP associated with age-related macular degeneration was mapped in the CFB (complement factor B) gene located in the DHS correlated with *CFB*, suggesting that the DHS could play a key role in age-related macular degeneration. It is likely that the other SNP associated with LDL cholesterol was mapped in the *COL11A2* gene, which is located in the DHS correlated with *COL11A2*, suggesting that the DHS could play a key role in LDL cholesterol.

In addition, we found 141 GWAS SNPs in 129 DHSs in human, chimp, and macaque fibroblasts (Figure 4). Of these SNPs, a total of 29 SNPs just located in DHSs correlated with target genes, in which we found there were 16 GWAS SNPs associated with diseases by analyzing eQTL, Hi-C, and GWAS SNPs (Table 2), including 11 cases that a target gene of the DHS closely matched the association reported for its GWAS SNP, six cases that GWAS SNPs affected target genes expression, and two cases that a target gene of eQTL closely matched the association reported for its GWAS SNP.

For example, one very interesting GWAS SNP rs7756521 associated with HIV-1 control is located in Discoidin domain receptor tyrosine kinase 1 (*DDR1*), and is also located in a DHS that is correlated with the intergenic region of *DDR1*, and also affected proximal *DDR1* expression in human tissues combined on eQTL data. This indicated that the GWAS SNP rs7756521 is a causative genetic variant, and this DHS could play a key role in HIV-1 control (Table 2). Similarly, GWAS SNP rs2074488 in the DHS associated with chronic obstructive pulmonary disease probably affected *POU5F1* gene expression in human tissue according to eQTL data, and *POU5F1* is also located in this DHS. Further, the SNP rs2074488 characterized in *HLA-C* is located in a DHS common to the three species whose target genes were *HLA-C*, suggesting that the DHS may play a key role and the GWAS SNP rs2074488 could be a causative genetic variant in chronic obstructive pulmonary disease.

Like GWAS SNP rs2074488, the GWAS SNP rs2071278 associated with Complement C3 and C4 levels in *HLA-DRA* is located in a DHS that targets *HLA-DRA* and *CYP21A2*, while SNP rs2071278 affects *CYP21A2* gene expression in human tissue according to eQTL data. This observation suggests that this DHS could play an important role, and the GWAS SNP rs2071278 could be a causative genetic variant in Complement C3 and C4 levels. In the other example, eQTL data showed that SNP rs2736172 in the DHS likely affected proximal *BAG6*, *ATF6B*, and *PRRC2A* gene expression in human tissue; all three genes are also the DHS target genes, indicating that SNP rs2736172 was the causative genetic variant in psychosis (atypical).

Moreover, GWAS SNP rs1480380 in DHS likely affected the gene expression of proximal *HLA-DMA* and *BRD2* in human tissues including brain. *BRD2* has been implicated in juvenile myoclonic epilepsy, a common form of epilepsy that becomes apparent in adolescence. *HLA-DMA* and *BRD2* were also the DHS target genes. Thus, we speculated that rs1480380 is a causative genetic variant by affecting the gene expression of proximal *HLA-DMA* and *BRD2*. In summary, combined with eQTL, Hi-C, and GWAS SNPs data, our results suggest that six GWAS SNPs in DHSs could be causative genetic variants by likely affecting proximal gene expression in human tissues.

During our analysis of several cases where a target gene of the DHS closely matched the association reported for its

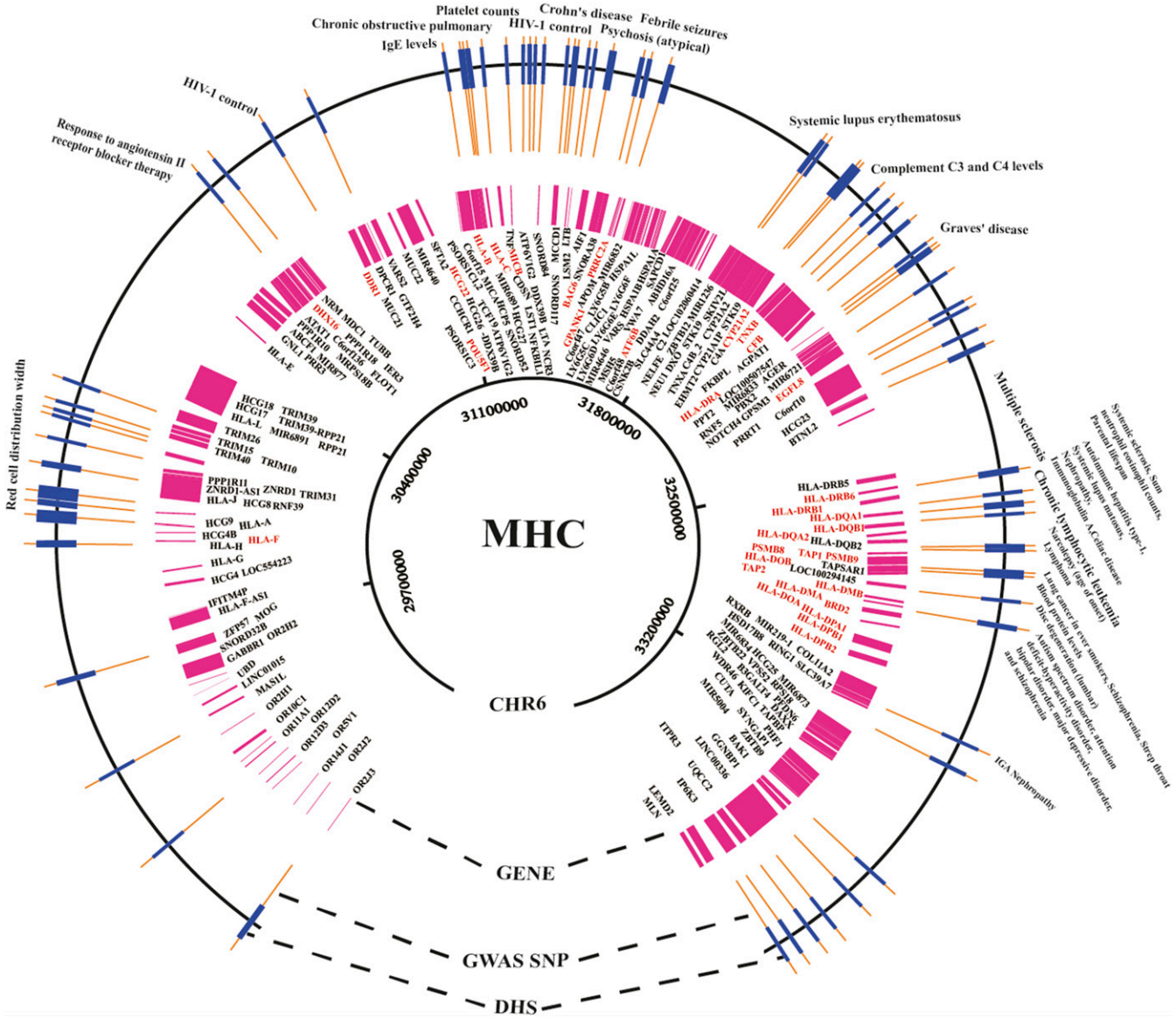

Figure 4 The diseases associated with the SNP in the DHS detected from three species based on the GWAS database and the distribution of genes in MHC region. The blue color means the diseases associated with the remarkable DHS in positive selection test. The red color means the diseases associated with the DHS detected in our three species and also reported have relations with GENEs (>0.7) in other research.

GWAS SNP, we found one GWAS SNP (rs3130618) that was associated with febrile seizures (MMR vaccine-related), and idiopathic membranous nephropathy was mapped in the *GPANK1* gene. Interestingly, one of the corresponding DHS target genes was *GPANK*1, which was located in a cluster of HLA-B-associated transcripts and was included in the human major histocompatibility complex III region. This gene encodes a protein that is thought to play a role in immunity. Based on eQTL data, we also observed that this variant affected *CYP21A2* expression in human tissues. Like GWAS SNP rs3130618, two GWAS SNPs were present: rs2187668 and rs9274477. For example, in DHS (chr6:32605800-32605972), the GWAS SNP rs2187668, which is associated with autoimmune hepatitis type 1, systemic lupus erythematosus, nephropathy, immunoglobulin A, and celiac disease, affected *CYP21A2* expression according to eQTL data. However, a target gene of the DHS closely matched the association reported for its GWAS SNP rs2187668 (Table 2). Another GWAS SNP, rs9274477, associated with narcolepsy (age of onset), was observed to affect *HLA-DQB2* expression while the SNP mapped at HLA-DQB1 closely matched a target gene of the DHS.

Besides, two GWAS SNPs (rs7383287 and rs2071747) were found at a DHS. The SNP rs7383287, which has been linked to lung cancer and schizophrenia, was mapped in *HLA-DOB*. Coincidentally, one of the target genes for this DHS was also *HLA-DOB*. Like the above SNP, we also observed a couple of SNPs, including GWAS SNP rs17220241 associated with blood protein levels, GWAS SNP rs1883414 associated with IGA nephropathy, and GWAS SNP rs3819299 associated with platelet counts. Interestingly, we observed that the mapping

Table 2 Analysis of GWAS SNPs located in DHSs on MHC region from human, chimpanzee, and macaque fibroblasts

Start	End	SNP	Species-specific	SNP associating with disease (GWAS)	Mapped gene	Reported gene	DHS target gene from HI-C data	eQTL
30848154	30848345	rs7756521	Cgain	HIV-1 control	DDR1	DDR1, VARS2, DPCR1	DDR1	DDR1
31197381	31197644	rs3130941	Hloss	IgE levels	TRNAI25	HLA-C, HCG27	HCG22	HCG27
31233725	31241873	rs2074488	Common	Chronic obstructive pulmonary disease	HLA-C	HLA-C	HLA-C, POU5F1	PSORS1C3, POU5F1
31322298	31322508	rs3819299	Closs	Platelet counts	HLA-B	HLA-B	HLA-B, HLA-C	none
31431385	31431813	rs2395029	Closs	HIV-1 control	HCP5	HLA-B, HCP5	HLA-C, HLA-B, MICB	MICB (moderate negative)
31542063	31542726	rs1799964	Common	Crohn's disease	LTA;TNF	LTA,HLA-DQA2,TNF,LST1,LTB	HLA-C, HLA-B	LST1
31587350	31591070	rs2736172	Common	Psychosis (atypical)	PRRC2A; SNORA38	MICB, TNF	PRRC2A, BAG6, ATF6B	BAG6, ATF6B, PRRC2A
31631600	31634790	rs3130618	Common	Febrile seizures (MMR vaccine-related)	GPANK1	GPANK1	HLA-B, HLA-C, GPANK1	CYP21A2
32165291	32165513	rs2071278	Common	Complement C3 and C4 levels	NOTCH4	HLA-DRA	HLA-DRA, CYP21A2	CYP21A2, C4B
32590725	32591250	rs9271588, rs3129763, rs34831921	Mloss	Systemic sclerosis, sum neutrophil eosinophil counts, parental lifespan Sjogren's syndrome	TRNAI25	HLA-DRA, HLA-DQA1, HLA-DRB1 HLA-DRB5, HLA-DRB6	HLA-DQA1	HLA-DRB6
32605800	32605972	rs2187668	Hgain	Autoimmune hepatitis type 1, systemic lupus erythematosus, nephropathy, immunoglobulin A, celiac disease	HLA-DQA1	HLA-DQA1, HLA-DR3, HLA-DRB1, HLA-DQB1	HLA-DQA1, HLA-DRB1, HLA-DQB1	CYP21A1P
32633678	32634221	rs9274477	Closs	Narcolepsy (age of onset)	HLA-DQB1	NR	HLA-DQB1	HLA-DQB2
32781842	32785278	rs2071747, rs7383287	Common	Lung cancer in ever smokers, Schizophrenia, strep throat	HLA-DOB,	HLA-DOB	HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, HLA-DRB1, HLA-DRB5, HLA-DRB9	none
32820065	32823157	rs17220241	Common	Blood protein levels	TAP1, PSMB9	HLA-Z, PSMB8, PSMB9, TAP1, TAP2, XXbac-BPG246D15.9,	HLA-DMA, HLA-DMB, HLA-DPA1, HLA-DPB1, HLA-DPB2, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DRA, HLA-DRB1, PSMB8, PSMB9, TAP1, TAP2, XXbac-BPG246D15.9	none
32913110	32913295	rs1480380	Mloss	Autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia	TRNAI25	NR	HLA-DRB1, HLA-DQA1, HLA-DMA, HLA-DMB, BRD2	HLA-DMA, BRD2, HLA-DMB (combined eQTL)
33086255	33087057	rs1883414	Mloss	IGA nephropathy	TRNAI25	HLA-DPA1, HLA-DPB1, HLA-DPB2	HLA-DPA1, HLA-DPB1, HLA-DPB2	none

genes of two GWAS SNPs were the same as the target genes of DHS as well as the same as those genes whose expression was probably affected by the same SNPs in human tissues based on eQTL data, including one GWAS SNP rs2736172 associated with psychosis (atypical) and another GWAS SNP rs1480380 associated with autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia. In summary, several GWAS SNPs were observed in DHSs, providing insight into the causative variation underlying disease associated with the MHC region.

Discussion

In this study, we identified 15 instances of transcription factor binding motif gains located in the rapidly evolving 24 DHSs in the MHC region, nine of which were *SOX9*, *OLF1*, *ER*, *SP3*, *HIC1*, *GATA1*, *TP53*, *GLI3*, and *BEN* predicted by FIMO, as well as *USF*, *MYC*, *MAX*, *MAFK*, *STAT1*, and *PBX3* obtained from UCSC ChIP-seq data. Combining eQTL and Hi-C data, we found that there were five SNPs located in human gains motifs affecting the corresponding gene's expression, two of which closely matched DHS target genes. According to our results, the human gains of *STAT1*, *MAFK*, and *PBX3* transcription factor binding sites during evolution led to different expression of the target genes of DHSs with rapid evolution, including five members of the TRIM family, as well as HLA-E, HLA-A, and HLA-F, especially *TRIM31* and *TRIM40* based on eQTL data. TRIM31, as a tumor suppressor in non-small-cell lung cancer, shows altered expression in certain tumors and may be a negative regulator of cell growth, including up-regulation in gastric adenocarcinomas, and is associated with Behcet syndrome and cardiomegaly (from UCSC data). TIRM40, which is associated with multiple sclerosis, may play a role as a negative regulator against inflammation and carcinogenesis in the gastrointestinal tract. Therefore, it needs further investigation to build the relationship between human-specific binding of *STAT1*, *MAFK*, and *PBX3* transcription factors and TRIM31/TRIM40. Likewise, in another rapidly evolving DHS, genetic variations in the core motifs of *SOX9*, *USF*, *MYC*, and *MAX* with human gains were observed affecting the expression of *HCG9* (but not the target gene of the DHS). A possible explanation was that HCG9 was also a potential target gene of DHS but had not been identified by the current Hi-C. We still need to perform more experiments to explore the relationship between the DHS target genes and *SOX9*, *USF*, *MYC*, and *MAX* motifs with human gains. In addition, the *SOX9*, *USF*, *MYC*, and *MAX* with human-gained binding sites in this rapidly evolving DHS may undergo adaptive evolution through these transcription factors affecting the expression of target genes. However, this needs further study in the future.

In the comprehensive analysis of GWAS SNP, DHS target gene, and eQTL data in the DHS of the MHC region, we found the interesting GWAS SNP rs7756521 (C: 41.494%, T: 58.506%), which is associated with HIV-1 control, located in *DDR*, which was also a target gene of the contained DHS. Moreover, the GWAS SNP likely affects *DDR* expression. In our study, this DHS was defined as human loss compared to fibroblast cells from two other primates. Together with the current UCSC data, there was no DHS peak signal at the position of the SNP but the presence of a DHS peak signal in chimpanzee and macaque fibroblasts. We observed that the position was the T allele in other nonhuman primates. So, we speculated that when the *DDR* (chr6: 30848253C> T) position was the T allele, this DHS was active in chimpanzees and macaque and human beings carrying the T allele, and when this position was the C allele, the DHS was turned off and was inactive, affecting *DDR* expression. However, there was no TF motif identified in the corresponding location in UCSC ChIP-seq database. Therefore, it is a challenge to speculate that the variant affected the expression of the target gene.

We also found that SNP rs531359461 (C: 99.980%, G: 0.020%) was located at the *RFX5* motif. While mutated to the G allele, the mutated allele may affect TF binding, resulting in the down-expression of MHC class II genes. *RFX5* mutations are seen in cases of a severe immunodeficiency syndrome called MHC-II deficiency [also known as bare lymphocyte syndrome (BLS)]. These mutations prevented the RFX complex from binding to the X box in MHC-II promoters, resulting in the decrease in MHC-II expression. But we have not yet found a eQTL to be present.

Our study has limitations. First, the DHS data we used were mostly generated from human tissues and only few came from nonhuman primates (Shibata *et al.* 2012). Second, the activity of a *cis*-element may be highly developmental time point and cell type specific, and depends on the coordination of additional regulatory elements; hence, more extensive *in vivo* experiments would be more powerful. Therefore, understanding the genetic basis of uniquely human traits will ultimately require *in vivo* functional studies of specific loci, individually and in combination. By using methods such as iPSCs and humanized mouse models with CRISPR/Cas9 and introducing a uniquely human genetic change into a living nonhuman organism or mouse we will be able to isolate and determine its effects on gene expression, regulation, and phenotype. This strategy will allow us to link genetic changes in human evolution directly to changes in human traits, revealing the evolutionary events that shaped our species.

In conclusion, we found that common DHSs were more proximal to their nearest gene than species-specific DHSs. Using tests for positive selection, we identified 24 DHSs that evolved rapidly on the human lineage. Of these 24 DHSs, we characterized 15 transcription factor motifs with human gains, nine of which are *SOX9*, *OLF1*, *ER*, *SP3*, *HIC1*, *GATA1*, *TP53*, *GLI3*, and *BEN* predicted by FIMO, as well as *USF*, *MYC*, *MAX*, *MAFK*, *STAT1*, and *PBX3* obtained from UCSC ChIP-seq data. In addition, we found 141 GWAS SNPs in 129 DHSs in human, chimp, and macaque fibroblasts. Of these SNPs, just 25 were located in DHSs correlated with target genes, within which we found 16 interesting GWAS SNPs associated with diseases by analyzing eQTL, Hi-C, and GWAS SNPs. One of the most interesting GWAS SNPs was rs7756521, which is associated with HIV-1 control; this probably affects DDR expression and could be a causative genetic variant for HIV-1

control. Our results provide a novel insight into the genetic basis of diseases or traits associated with variation in the MHC region.

Acknowledgments

We thank the Crawford research group from Duke University. This work was supported by joint funding from the School of Bioscience and Bioengineering, South China University of Technology and BGI-ShenZhen as well as funding of 2014 major projects from the Education Department of Guangdong Province (2014KZDXM009). We declare that there are no competing interests.

Author contributions: F.L. and J.M.A. conceived and designed the study. Y.L., R.M.G., X.L., and Y.J. analyzed the data. F.L. and M.D.L. wrote and edited the paper.

Literature Cited

Clop, A., A. Bertoni, S. L. Spain, M. A. Simpson, V. Pullabhatla *et al.*, 2013 An in-depth characterization of the major psoriasis susceptibility locus identifies candidate susceptibility alleles within an HLA-C enhancer element. PLoS One 8: e71690. https://doi.org/10.1371/journal.pone.0071690

Dorschner, M. O., M. Hawrylycz, R. Humbert, J. C. Wallace, A. Shafer *et al.*, 2004 High-throughput localization of functional elements by quantitative chromatin profiling. Nat. Methods 1: 219–225. https://doi.org/10.1038/nmeth721

El Omari, K., S. J. Hoosdally, K. Tuladhar, D. Karia, P. Vyas *et al.*, 2011 Structure of the leukemia oncogene LMO2: implications for the assembly of a hematopoietic transcription factor complex. Blood 117: 2146–2156. https://doi.org/10.1182/blood-2010-07-293357

ENCODE Project Consortium, 2004 The ENCODE (ENCyclopedia of DNA elements) project. Science 306: 636–640. https://doi.org/10.1126/science.1105136

ENCODE Project Consortium, 2012 An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74. https://doi.org/10.1038/nature11247

ENCODE Project ConsortiumBirney, E., J. A. Stamatoyannopoulos, A. Dutta, R. Guigo *et al.*, 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816. https://doi.org/10.1038/nature05874

Fraser, H. B., 2013 Gene expression drives local adaptation in humans. Genome Res. 23: 1089–1096. https://doi.org/10.1101/gr.152710.112

Gittelman, R. M., E. Hun, F. Ay, J. Madeoy, L. Pennacchio *et al.*, 2015 Comprehensive identification and analysis of human accelerated regulatory DNA. Genome Res. 25: 1245–1255. https://doi.org/10.1101/gr.192591.115

Grant, C. E., T. L. Bailey, and W. S. Noble, 2011 FIMO: scanning for occurrences of a given motif. Bioinformatics 27: 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Handunnetthi, L., S. V. Ramagopalan, G. C. Ebers, and J. C. Knight, 2010 Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. Genes Immun. 11: 99–112. https://doi.org/10.1038/gene.2009.83

Hesselberth, J. R., X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom *et al.*, 2009 Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods 6: 283–289. https://doi.org/10.1038/nmeth.1313

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta *et al.*, 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 106: 9362–9367. https://doi.org/10.1073/pnas.0903103106

King, M. C., and A. C. Wilson, 1975 Evolution at two levels in humans and chimpanzees. Science 188: 107–116. https://doi.org/10.1126/science.1090005

Lee, Y. H., B. Sauer, and F. J. Gonzalez, 1998 Laron dwarfism and non-insulin-dependent diabetes mellitus in the Hnf-1alpha knockout mouse. Mol. Cell. Biol. 18: 3059–3068. https://doi.org/10.1128/MCB.18.5.3059

Lu, T., L. Aron, J. Zullo, Y. Pan, H. Kim *et al.*, 2014 REST and stress resistance in ageing and Alzheimer's disease. Nature 507: 448–454. https://doi.org/10.1038/nature13163

Lu, T., L. Aron, J. Zullo, Y. Pan, H. Kim *et al.*, 2016 Addendum: REST and stress resistance in ageing and Alzheimer's disease. Nature 540: 470. https://doi.org/10.1038/nature20579

Manolio, T. A., 2010 Genomewide association studies and assessment of the risk of disease. N. Engl. J. Med. 363: 166–176. https://doi.org/10.1056/NEJMra0905980

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen *et al.*, 2012 Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190–1195. https://doi.org/10.1126/science.1222794

Neph, S., M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman *et al.*, 2012 BEDOPS: high-performance genomic feature operations. Bioinformatics 28: 1919–1920. https://doi.org/10.1093/bioinformatics/bts277

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, 2010 Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20: 110–121. https://doi.org/10.1101/gr.097857.109

Pontoglio, M., 2000 Hepatocyte nuclear factor 1, a transcription factor at the crossroads of glucose homeostasis. J. Am. Soc. Nephrol. 11: S140–S143.

Prabhakar, S., J. P. Noonan, S. Paabo, and E. M. Rubin, 2006 Accelerated evolution of conserved noncoding sequences in humans. Science 314: 786. https://doi.org/10.1126/science.1130738

Prabhakar, S., A. Visel, J. A. Akiyama, M. Shoukry, K. D. Lewis *et al.*, 2008 Human-specific gain of function in a developmental enhancer. Science 321: 1346–1350. https://doi.org/10.1126/science.1159974

Sabo, P. J., M. S. Kuehn, R. Thurman, B. E. Johnson, E. M. Johnson *et al.*, 2006 Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. Nat. Methods 3: 511–518. https://doi.org/10.1038/nmeth890

Shibata, Y., N. C. Sheffield, O. Fedrigo, C. C. Babbitt, M. Wortham *et al.*, 2012 Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. 8: e1002789. https://doi.org/10.1371/journal.pgen.1002789

Spurgin, L. G., and D. S. Richardson, 2010 How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc. Biol. Sci. 277: 979–988. https://doi.org/10.1098/rspb.2009.2084

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano *et al.*, 2012 The accessible chromatin landscape of the human genome. Nature 489: 75–82. https://doi.org/10.1038/nature11232

Trowsdale, J., 2011 The MHC, disease and selection. Immunol. Lett. 137: 1–8. https://doi.org/10.1016/j.imlet.2011.01.002

Trowsdale, J., and J. C. Knight, 2013 Major histocompatibility complex genomics and human disease. Annu. Rev. Genomics Hum. Genet. 14: 301–323. https://doi.org/10.1146/annurev-genom-091212-153455

Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich *et al.*, 2000 TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. 28: 316–319. https://doi.org/10.1093/nar/28.1.316

Communicating editor: R. Nielsen