Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex

Claire Vandiedonck, Martin S Taylor, Helen E Lockstone, Katharine Plant, Jennifer M Taylor, Caroline Durrant, John Broxholme, Benjamin P Fairfax, Julian C Knight

Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex

Claire Vandiedonck, Martin S. Taylor, Helen E. Lockstone, et al.

Genome Res. published online May 31, 2011
Access the most recent version at doi:10.1101/gr.116681.110

Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex

Claire Vandiedonck,[1,2,3,5] Martin S. Taylor,[1,4] Helen E. Lockstone,[1] Katharine Plant,[1] Jennifer M. Taylor,[1] Caroline Durrant,[1] John Broxholme,[1] Benjamin P. Fairfax,[1] and Julian C. Knight[1,5]

[1] *Wellcome Trust Centre for Human Genetics, Oxford University, Oxford OX3 7BN, United Kingdom;* [2] *INSERM, UMRS-958, 75010 Paris, France;* [3] *Université Paris 7 Denis-Diderot, 75013 Paris, France;* [4] *MRC Human Genetics Unit, Edinburgh EH4 2XU, United Kingdom*

The human major histocompatibility complex (MHC) on chromosome 6p21 is a paradigm for genomics, showing remarkable polymorphism and striking association with immune and non-immune diseases. The complex genomic landscape of the MHC, notably strong linkage disequilibrium, has made resolving causal variants very challenging. A promising approach is to investigate gene expression levels considered as tractable intermediate phenotypes in mapping complex diseases. However, how transcription varies across the MHC, notably relative to specific haplotypes, remains unknown. Here, using an original hybrid tiling and splice junction microarray that includes alternate allele probes, we draw the first high-resolution strand-specific transcription map for three common MHC haplotypes (*HLA-A1–B8–Cw7–DR3*, *HLA-A3–B7–Cw7–DR15*, and *HLA-A26–B18–Cw5–DR3–DQ2*) strongly associated with autoimmune diseases including type 1 diabetes, systemic lupus erythematosus, and multiple sclerosis. We find that haplotype-specific differences in gene expression are common across the MHC, affecting 96 genes (46.4%), most significantly the zing finger protein gene *ZFP57*. Differentially expressed probes are correlated with polymorphisms between haplotypes, consistent with *cis* effects that we directly demonstrate for *ZFP57* in a cohort of healthy volunteers ($P = 1.2 \times 10^{-14}$). We establish that alternative splicing is significantly more frequent in the MHC than genome-wide (72.5% vs. 62.1% of genes, $P \leq 1 \times 10^{-4}$) and shows marked haplotypic differences. We also unmask novel and abundant intergenic transcription involving 31% of transcribed blocks identified. Our study reveals that the renowned MHC polymorphism also manifests as transcript diversity, and our novel haplotype-based approach marks a new step toward identification of regulatory variants involved in the control of MHC-associated phenotypes and diseases.

[Supplemental material is available for this article. The microarray data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE22455.]

The human major histocompatibility complex (MHC), located on chromosome 6p21 in humans, previously referred to as the "human leukocyte antigen (HLA) complex," plays a pivotal role in immune function (Dausset 1981). This region of 3.5 Mb is the most gene-dense of the genome, with 230 known genes and pseudogenes (Horton et al. 2004). It is classically divided into the class I region, which includes genes such as *HLA-A*, *HLA-B*, and *HLA-C*, and the class II region, including, for example, *HLA-DP*, *HLA-DQ*, and *HLA-DR*. These classical HLA genes encode molecules involved in antigen presentation and processing. The intervening MHC class III region notably includes genes encoding a variety of proteins involved in immunity including the Tumor Necrosis Factor (*TNF*) superfamily, components of the complement cascade, and molecular chaperones such as heat-shock proteins. The MHC is remarkable for its extensive polymorphism (de Bakker et al. 2006) and ranks first for the number of associations with immune and non-immune diseases (Shiina et al. 2004; Rioux et al. 2009). This has raised considerable interest across disciplines, from immunology and genetics, to medicine and evolutionary biology. Remarkable recent advances in our understanding of the genetic basis of common diseases have been achieved by genome-wide association studies (GWAS) (Wellcome Trust Case Control Consortium 2007; Manolio 2010), which have confirmed the preeminence of the MHC in terms of the magnitude of effect, statistical confidence, and the number of associations with autoimmune, infectious, and inflammatory diseases, together with cancer and adverse drug effects (Conde et al. 2010; Hamza et al. 2010; Hor et al. 2010; Singer et al. 2010).

The fine mapping of causal variants has proved challenging for the majority of complex traits, and we rarely understand the mechanisms through which DNA sequence polymorphisms operate (Knight et al. 2004). Their identification has been confounded by their multiplicity, their frequency in the general population, their modest effects, and linkage disequilibrium (LD). The latter is most remarkable in the MHC, where it may extend over several megabases (Ahmad et al. 2003; Yunis et al. 2003; Vandiedonck and Knight 2009). As a result, diseases are often found to be associated with common extended ancestral MHC haplotypes encompassing hundreds of genes, many of which are candidates.

[5] **Corresponding authors.**
E-mail claire.vandiedonck@inserm.fr.
E-mail julian@well.ox.ac.uk.

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.116681.110. Freely available online through the *Genome Research* Open Access option.

21:000–000 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

Gene expression levels are considered as relevant intermediate phenotypes in complex diseases (Vafiadis et al. 1997; Giraud et al. 2007; Cookson et al. 2009; Nica et al. 2010; Teslovich et al. 2010). These expression phenotypes are heritable (Yan et al. 2002) and can be mapped as quantitative traits (Emilsson et al. 2008; Cheung and Spielman 2009). Such studies have already highlighted *cis*- and *trans*-acting SNPs within the MHC (Dixon et al. 2007; Vandiedonck and Knight 2009). However, these studies and more recent RNA-seq-based expression quantitative trait analyses (Montgomery et al. 2010; Pickrell et al. 2010) were focused primarily on single-point mapping of gene expression and did not account for the extended haplotypes relating the associated polymorphisms. A particular allele can indeed be found on more than one haplotype. Thus, the reciprocal question of which genes are differentially expressed between MHC extended haplotypes remains essential to resolving functionally important genetic variants as one might expect to find the disease-related genes among the genes whose expression is specifically modified on the risk haplotype.

Here we sought to draw for the first time a map of transcription for the human MHC at a haplotypic resolution in which the consequences of genetic variation in phase for a large contiguous chromosomal region can be established. We investigated three important haplotypes that are common in northern European populations, are highly conserved, and show evidence of selection and important associations with diseases: *HLA-A1-B8-Cw7-DR3* (associated with type 1 diabetes, systemic lupus erythematosus and myasthenia gravis, together with other diseases including common variable immunodeficiency and infectious disease susceptibility) (Price et al. 1999); *A3-B7-Cw7-DR15* (associated with protection from type 1 diabetes and susceptibility to multiple sclerosis and systemic lupus erythematosus) (Barcellos et al. 2003; Larsen and Alper 2004); and *A26-B18-Cw5-DR3-DQ2* (associated with type 1 diabetes and Graves' disease) (Johansson et al. 2003). These haplotypes were fully resequenced as part of the MHC Project (Stewart et al. 2004; Traherne et al. 2006; Horton et al. 2008), but informative individuals carrying the specific haplotypes were not included in previous expression quantitative trait studies (Montgomery et al. 2010; Pickrell et al. 2010). In this study, we show how gene expression profiling of individuals homozygous for the region has allowed us to identify extensive haplotype-related transcriptional differences and highlight the importance of alternative splicing in this transcriptional diversity.

Results

The MHC array: design and validation

To understand more clearly the relationship between MHC sequence variation and gene expression, we aimed to investigate how transcription varies between commonly occurring haplotypes spanning the classical MHC, including resolution of strand-specific transcripts and alternative splicing. Conventional microarrays based on the human reference sequence are often confounded by sequence variation not accounted for in probe design (Walter et al. 2007), and, to date, the difficulties of mapping reads from high-throughput sequencing technologies to the highly polymorphic MHC have limited the application of RNA sequencing to this genomic region. Thus, we developed a hybrid microarray for the MHC (denoted "MHC array") that included alternate allele probes to account for known sequence diversity (Supplemental Methods). Our array design also aimed to resolve genic and intergenic transcription in a strand-specific manner at high resolution by including a strand-specific tiling path probe set together with probes specific to known and predicted splice junctions. We sought to use the MHC array to analyze transcription at haplotypic resolution using lymphoblastoid cell lines (LCLs) established from individuals MHC-homozygous for the three autoimmune disease-associated haplotypes of interest—COX (*HLA-A1-B8-Cw7-DR3*), PGF (*A3-B7-Cw7-DR15*), and QBL (*A26-B18-Cw5-DR3-DQ2*) (Horton et al. 2008).

The MHC array includes 505,686 probes of 25-mers interrogating 3.5 Mb of the classical MHC between coordinates chr6:29,748,239–33,231,091 (hg18), including 230 genes with a total of 2755 exons (Supplemental Fig. 1). One set of 398,626 overlapping probes (denoted the tiling path probe set) tiles both strands with a final resolution of 18 bases, allowing identification of any new transcript and its transcriptional orientation. A second set of 15,348 junction probes in four replicates aimed to monitor all known or predicted splice events, corresponding to 1043 junctions in the MHC class III region (12 overlapping probes on average per junction). For any junction or tiling probe, its reverse complement was also incorporated into the design. Importantly, alternate probes were specifically designed for all known SNPs or segmental duplications.

We first carried out experiments to assess the performance of the MHC array. Our design incorporated 10,572 shared probes with the Affymetrix Exon 1.0 ST array allowing comparison across platforms for these probes. We analyzed three biological replicate samples for each of three cell lines using the custom MHC array and the Affymetrix Exon 1.0 ST array. Intensity data from the shared probes were highly correlated for the nine samples hybridized to both platforms (Pearson correlation coefficients ranged from 0.83 to 0.91) (Supplemental Table 1), proving that our sample preparation and hybridization conditions were satisfactory. Interestingly, differences between cell lines were also correlated between platforms, suggestive of haplotypic differences (Supplemental Fig. 2). When all probes of the MHC array were considered, the correlation coefficient between culture replicates ranged between 0.96 and 0.98 (Supplemental Fig. 3). In addition to the usual standard quality controls for hybridization and sensitivity, we estimated the strand specificity as 84.7% (±4.1%) based on the observed ratios of expression between the two strands of known expressed housekeeping genes (Supplemental Methods). We also verified the coverage of full transcripts by comparing the signal intensities from probes tagging both ends of housekeeping genes (coefficient of variation, 13%).

To assess the signal specificity of alternate allele probes, we compared the signal intensity of the PGF samples in the transcribed regions (see below) measured on PGF-specific probes with that measured on COX- and QBL-specific probes. We found a significantly higher signal on PGF-specific probes (ANOVA, $P = 2.4 \times 10^{-5}$). We also compared the signal of the PGF samples on the 123 perfect match probes paired with probes carrying one mismatch corresponding to the COX path. The signal was consistently higher on perfect match probes (ANOVA with repeated measures, $P = 2 \times 10^{-4}$). We evaluated the junction probes' performance by using *CD79A* and *CD79B* genes that code for both main chains of the invariant component of the B-cell receptor complex and are expressed in LCLs. The comparison of array data and quantitative PCR data showed similar proportions between isoforms. For *CD79A*, we measured a ratio of 4.96 ± 0.17 between the long and the short isoforms using the array, and of 5.35 ± 0.46 by RT-PCR. For *CD79B*, we obtained a ratio of 2.71 ± 0.21 between the long and the short isoforms with the array, compared to 2.84 ± 0.48 by RT-PCR.

A high-resolution strand-specific MHC transcription map

Identification of transcriptionally active regions (TARs)

Using this validated platform, we initiated experiments in which we aimed to generate a high-resolution strand-specific transcriptomic map of the MHC. We first verified chromosome and MHC integrity of the selected homozygous cell lines by DNA-FISH and then analyzed RNA prepared from PGF, COX, and QBL cells grown in triplicate and hybridized to the MHC array. After preprocessing of all probes, we analyzed the tiling probes on each strand, in terms of the "shared paths" corresponding to probes shared and identical between the three haplotypic sequences, and the "alternate paths," which also include haplotype-specific probes for each haplotype. Hence, a total of eight sequence paths were considered (one shared and three haplotypic sequence paths for each strand). After signal smoothing, we determined transcriptionally active regions (TARs) (Bertone et al. 2004) as any 51-base windows with median signal intensity exceeding a threshold determined by permutation (Supplemental Methods). An overview of the signal across the entire region with the "shared paths" relative to the PGF reference assembly sequence is provided in Supplemental Figure 4. Overlapping TARs at a false discovery rate (FDR) of 1% were merged to define transcribed blocks, whose size range was similar between strands and haplotypes (from 51 to 1380 bases, mean = 108.2 bases). On average, there was one transcribed block per 1.4 kb. These are listed in Supplemental Table 2 including location relative to path, strand, and each transcript as annotated in Vega, currently the most comprehensive annotation of the MHC locus. Overall, we found that 6% of the MHC sequence is transcribed, with an equal distribution of 2% for transcribed blocks on the forward, reverse, or both strands.

Genic and intergenic transcription

We then sought to determine the extent of genic and intergenic transcription based on Vega gene annotations. We defined Vega genes as being transcribed based on the inclusion of at least one TAR using a 5% FDR on each "alternate path." Their proportion was similar between haplotypes and remarkably high, >92% for the genes and >70% for the pseudogenes, underscoring the accuracy of Vega annotations for the MHC (Supplemental Table 3). An overview of strand-specific gene transcription occurring across the MHC is provided in the associated Figure 1 (see foldout).

In terms of intergenic transcription, a remarkably high proportion (31%) of the transcribed blocks did not map to known genes. These intergenic blocks had an average size of 69 bases (range 51–367) in total, reaching 1.7% of the combined length of both strands, thus corresponding to 28.3% of overall transcribed genomic sequence length. When looking at the distribution of the distances of these TARs to known neighboring genes, the median was found to be 10 kb (Supplemental Fig. 5). One-half of the intergenic transcribed blocks thus mapping within 10 kb of annotated genes on either the 5′ or the 3′ side, could be new exons or regulatory elements as suggested by previous studies using either tiling arrays or RNA-seq (Bertone et al. 2004; Gaulton et al. 2010; van Bakel et al. 2010). The remaining 50% of intergenic transcribed blocks were more distant (>10 kb) and tended to cluster (>50% are <0.9 kb apart) in regions of lower gene density (65.1% in class I, 25.5% in class II, and 9.4% in class III). Most notably, 95% of them colocalized with repeat elements, 78% of which mapped to an *Alu* sequence. This is not simply a consequence of cross-hybridization with *Alu* sequences transcribed from elsewhere in the genome as only 37% of *Alu* repeats covered by the array design (necessitating probes to be of genome-wide unique sequence) overlapped a TAR. The same proportion was found when considering recent *Alu* subfamilies, *Alu*Y and *Alu*Sg. Similarly, this signal could not be attributed to edited RNAs from genome-wide *Alu* sequences that are widespread in human, as we found that only 0.1% of probes present in these distant intergenic TARs matched the A-to-I or C-to-U edited RNA sequences cataloged in the comprehensive DAtabase of RNa EDiting (Kiran and Baranov 2010). Altogether, our data support an abundant transcriptional activity from *Alu* sequences in the intergenic regions.

Haplotype-specific transcription

Numerous genes are differentially expressed between haplotypes

Using this high-resolution, strand-specific transcriptional map of the MHC, we addressed the issue of haplotypic-specific gene expression. First, we considered the highest resolution using TARs generated at a conservative FDR of 1% on the "shared paths." We found that 9%, 4.6%, and 11.1% of the TARs on PGF, COX, and QBL sequences, respectively, were identified in only one cell line, suggesting haplotype-specific expression.

We next tested quantitative differences in expression levels. To this end, we used the probes from the "alternate paths" matching exactly the haplotypic sequence of the corresponding cell line, grouped into metaprobesets based on Vega annotations. Moreover, these metaprobesets contain 4.13 times more probes per gene than in the Affymetrix Exon 1.0 ST array (Supplemental Fig. 6). The MHC array thus provides "individualized" gene levels with a high level of accuracy. As shown in Supplemental Table 4, this resulted in a somewhat different list of differentially expressed genes. Overall, using the MHC array, we identified 96 differentially expressed genes between the three cell lines (Fig. 1; Table 1). These included a number of classical HLA class I (*HLA-A*, *-B*, *-C*, and *-F*) and class II genes (*HLA-DQA2*, *-DQB2*, *-DPB1*) as well as class III genes including *TNF*, *LTA*, *NCR3*, and *LTB*. We selected 12 genes showing haplotypic differences in expression for study by quantitative RT-PCR and found expression level differences between the cell lines reaching statistical significance in nine of them (Supplemental Fig. 7) (see below for *ZFP57*, *HLA-DQB2*, and *HLA-C*).

This analysis allows candidate genes to be defined for specific haplotypes. For example, we determined genes, ordered on the chromosome from telomere to centromere, that were significantly differentially expressed (adjusted *P*-value < 0.05) between either COX and PGF/QBL for the *HLA-A1-B8-DR3* haplotype or between PGF and COX/QBL for the *HLA-A3-B7-DR15* haplotype. Only genes up- or down-regulated in the same direction were selected (Table 2). This highlights, for example, *ZFP57*, *LTA*, *TNF*, *HLA-DQA2*, and *HLA-DPB1* as showing greater than twofold differential expression with the *HLA-A1-B8-DR3* haplotype and as being important candidate genes to investigate further for this important disease-associated haplotype.

Colocalization of differentially expressed probes and polymorphic SNPs

That these differences could result from haplotype-specific sequence variation was supported by the correlation we found between the location of differentially expressed probes and polymorphisms between haplotypes along the chromosome for two sets of interval series of 10-kb windows shifted by 5 kb across the MHC (Fig. 2). This was particularly significant (as low as $P = 1.8 \times 10^{-6}$ between PGF and QBL, Spearman test) when the analysis was restricted to windows including at least one gene (Supplemental Table 5).

Table 1. Variation of gene expression between haplotypes

		$\log_2$ (fold change)			
Gene name	Class	COX vs. PGF	QBL vs. PGF	QBL vs. COX	Adjusted *P*-value
ZFP57	I	2.77	0.00	−2.76	1.22×10^{-14}
HLA-DPB2[a]	II	−3.19	−3.02	0.17	2.89×10^{-12}
HLA-DQA2	II	−2.45	−1.62	0.82	1.91×10^{-11}
HLA-DQB2	II	−2.74	−2.58	0.16	3.21×10^{-11}
HLA-U[a]	I	−2.52	0.36	2.87	1.32×10^{-10}
TNF	III	1.90	1.03	−0.87	4.79×10^{-10}
HLA-DPB1	II	−2.08	−0.90	1.18	6.44×10^{-10}
RPL32P1[a]	II	−1.52	−1.19	0.33	2.07×10^{-09}
HLA-B	I	−0.06	−1.19	−1.13	6.59×10^{-09}
HLA-A	I	−1.51	−1.86	−0.35	2.30×10^{-08}
HLA-L[a]	I	−1.29	−1.47	−0.18	2.30×10^{-08}
XXbac-BPG254F23.6	II	−1.59	−1.59	0.00	2.50×10^{-08}
HCG22	I	−1.56	−1.26	0.30	2.96×10^{-08}
XXbac-BPG254F23.5	II	−1.42	−1.61	−0.19	1.33×10^{-07}
LTA	III	1.32	0.57	−0.75	2.04×10^{-07}
NCR3	III	0.87	0.95	0.08	4.95×10^{-07}
HLA-F	I	0.15	−0.90	−1.05	4.95×10^{-07}
HLA-DOA	II	−1.32	−0.89	0.43	5.07×10^{-07}
TAP1	II	0.97	0.08	−0.89	6.86×10^{-07}
LTB	III	−0.95	−0.06	0.89	7.02×10^{-07}
LST1	III	−0.18	0.48	0.66	9.42×10^{-07}
DAQB-335A13.8	I	0.61	−0.02	−0.63	1.12×10^{-06}
TCF19	I	1.11	0.62	−0.49	1.49×10^{-06}
CLIC1	III	1.22	0.57	−0.66	1.49×10^{-06}
HLA-DMA	II	−0.57	−0.89	−0.33	3.52×10^{-06}
BRD2	II	0.78	0.27	−0.51	3.60×10^{-06}
NRM	I	0.77	0.39	−0.38	4.48×10^{-06}
HLA-C	I	0.05	1.11	1.06	4.98×10^{-06}
PSMB9	II	0.42	−0.29	−0.71	6.05×10^{-06}
HCG27	I	0.56	0.06	−0.50	7.01×10^{-06}

Top 30 genes showing significant differential expression between haplotypes after Benjamini-Hochberg adjustment. For each cell line, the gene level intensity was computed from the signal intensity of the probes matching uniquely and perfectly to its haplotype sequence.

[a] Pseudogene.

Conversely, no correlation was found in windows lacking genes or when testing genic windows against nonpolymorphic markers between the pairs of haplotypes. Altogether, these results are consistent with a role for *cis*-acting regulatory variants influencing levels of gene expression.

Cis control of MHC gene expression in LCLs and primary cells

To further test whether variation in expression could be attributed to haplotypic effects, we investigated the three most significant differentially expressed genes—*ZFP57*, *HLA-DQA2*, and *HLA-DQB2* (Table 1). *ZFP57* encodes a zinc finger protein involved in transcriptional regulation and DNA methylation (Li et al. 2008) and is located at the telomeric end of the MHC class I region. We mapped its quantitative expression in peripheral blood mononuclear cells (PBMCs) of 93 healthy volunteers using 45,237 SNPs genotyped on the Illumina HumanCVDv1 BeadChip (Keating et al. 2008). Strikingly, this showed a highly significant association between expression of *ZFP57* and the rs29228 SNP located 16.8 kb downstream from *ZFP57* ($P = 1.2 \times 10^{-14}$) (Fig. 3A,B). The COX cell line is homozygous for the minor allele of the SNP associated with expression and when we tested three additional LCLs, only those homozygous for the rare allele showed evidence of *ZFP57* expression (Fig. 3C). In addition, rs29228 is in complete linkage disequilibrium with rs3129073, which is also significantly associated with *ZFP57* expression (effect = −1.088; $P = 5.4 \times 10^{-30}$; rank = fourth) in LCLs from an independent familial asthma cohort (Dixon et al. 2007). There is evidence of association of the COX haplotype with type 1 diabetes, while mutations of *ZFP57* itself have been associated with transient neonatal diabetes (Mackay et al. 2008).

We also performed genome-wide eQTL mapping for *HLA-DQA2* and *HLA-DQB2* using the same cohort of healthy volunteers. For both genes, we found significantly associated SNP markers in the MHC. For *HLA-DQA2*, rs2269423 was the most significantly associated SNP in the MHC, located 653 kb away from the gene, and the sixth genome-wide ($P = 2.13 \times 10^{-4}$). Individuals possessing a copy of the A allele showed higher levels of expression with consistent results seen in the panel of six MHC-homozygous LCLs for this SNP (Supplemental Fig. 8A). Similarly, for *HLA-DQB2*, rs9469220 located 65 kb downstream from the gene is the best associated SNP in the MHC and the seventh genome-wide ($P = 1.01 \times 10^{-4}$) (Supplemental Fig. 8B).

We specifically investigated the SNP rs9264942 located 35 kb upstream of *HLA-C*, which was previously reported to be associated with expression of *HLA-C* in PBMCs (Thomas et al. 2009). Using the MHC array, we find that higher expression of *HLA-C* is seen in QBL, which is homozygous CC for this SNP compared to COX and PGF, which are homozygous TT, consistent with the previous report of higher expression associated with possession of the C allele. Moreover, when we genotyped our 96 healthy volunteers by Sanger sequencing and looked at expression of *HLA-C* at the transcript level in PBMCs, we found that possession of a copy of the C allele is associated with 22.6% higher expression of *HLA-C* (Mann Whitney, $P = 0.023$, two-tailed) (Supplemental Fig. 8C).

These results validate the use of homozygous LCLs to identify haplotype-specific expression patterns. Although our study does not rule out the involvement of *trans*-acting variants, the correlation of differential expression with adjacent polymorphisms and our findings from expression quantitative trait mapping are consistent with several studies reporting a majority of *cis* eQTLs (Cheung and Spielman 2009).

The extent of alternative splicing in the MHC

Alternative splicing is increased in the MHC compared to non–MHC genes

Alternative splicing (AS) is critical to the generation of transcriptomic diversity and is known to be modulated by sequence variation with important implications for disease (Wang and Cooper 2007; Keren et al. 2010). Here we sought to investigate the extent of haplotype-specific alternative splicing within the MHC. First, we used the Affymetrix Exon 1.0 ST array hybridized with the PGF samples (whose MHC sequence is the human reference) to establish the extent of AS in this region in comparison with the rest of the genome. Absolute exon normalized intensities (NI) were determined by subtracting the $\log_2$ exon intensity from the $\log_2$ gene

Table 2. Candidate genes for diseases associated with the *HLA-A1-B8-DR3* (susceptibility to type 1 diabetes, systemic lupus erythematosus, myasthenia gravis) and *HLA-A3-B7-DR15* (susceptibility to multiple sclerosis, protection against type 1 diabetes) haplotypes

		HLA-A1-B8-DR3		*HLA-A3-B7-DR15*	
Class	Gene name	Mean $\log_2$ (fold change) COX vs. PGF/QBL	Direction	Mean $\log_2$ (fold change) PGF vs. COX/QBL	Direction
I	*ZFP57*	2.77	Up		
	ZDHHC20P1[a]	0.43	Up		
	DAQB-335A13.8	0.62	Up		
	IFITM4P [a]	0.98	Up		
	HCG4 [a]			−0.43	Down
	MICG [a]			−0.49	Down
	HLA-G			0.39	Up
	HLA-T [a]			0.53	Up
	HLA-K [a]			−0.53	Down
	HLA-U [a]	−2.69	Down		
	HLA-A			1.68	Up
	HCG4P5 [a]			0.60	Up
	TRIM26	0.31	Up		
	HLA-L			1.38	Up
	HCG18	0.60	Up		
	RPP21	0.31	Up		
	RANP1 [a]			−0.62	Down
	PRR3			−0.37	Down
	NRM	0.57	Up	−0.58	Down
	FLOT1			0.40	Up
	IER3			0.58	Up
	DDR1	−0.44	Down		
III	*VARS2*			−0.36	Down
	HCG22			1.41	Up
	TCF19	0.80	Up	−0.86	Down
	HCG27	0.53	Up		
	XXbac-BPG299F13.14	0.49	Up		
	HLA-S [a]	0.52	Up		
	MICB	0.52	Up		
	MCCD1			0.37	Up
	DDX39B	0.53	Up		
	ATP6V1G2	−0.36	Down		
	LTA	1.03	Up	−0.94	Down
	TNF	1.38	Up	−1.46	Down
	LTB	−0.92	Down		
	LST1	−0.42	Down		
	NCR3			−0.91	Down
	AIF1	−0.63	Down		
	APOM			−0.56	Down
	CLIC1	0.94	Up	−0.90	Down
	HSPA1L			0.33	Up
	HSPA1A			2.13	Up
	DOM3Z	0.34	Up		
	PBX2	0.30	Up		
II	*HLA-DRA*			0.60	Up
	HLA-DRB1			0.68	Up
	HLA-DQB1			0.81	Up
	XXbac-BPG254F23.5			1.52	Up
	XXbac-BPG254F23.6			1.59	Up
	HLA-DQA2	−1.64	Down	2.03	Up
	HLA-DQB2			2.66	Up
	TAP2	0.79	Up		
	PSMB8	0.66	Up		
	XXbac-BPG246D15.8	0.67	Up	−0.56	Down
	PSMB9	0.57	Up		
	TAP1	0.93	Up		
	HLA-DMA			0.73	Up
	BRD2	0.65	Up	−0.53	Down
	XXbac-BPG181M17.4	0.40	Up		
	HLA-DOA	−0.87	Down	1.10	Up
	HLA-DPA1			0.56	Up
	HLA-DPB1	−1.63	Down	1.49	Up
	RPL32P1 [a]	−0.93	Down	1.35	Up
	HLA-DPB2			3.10	Up
	HLA-DPA3 [a]	−0.43	Down		

[a] Pseudogene.

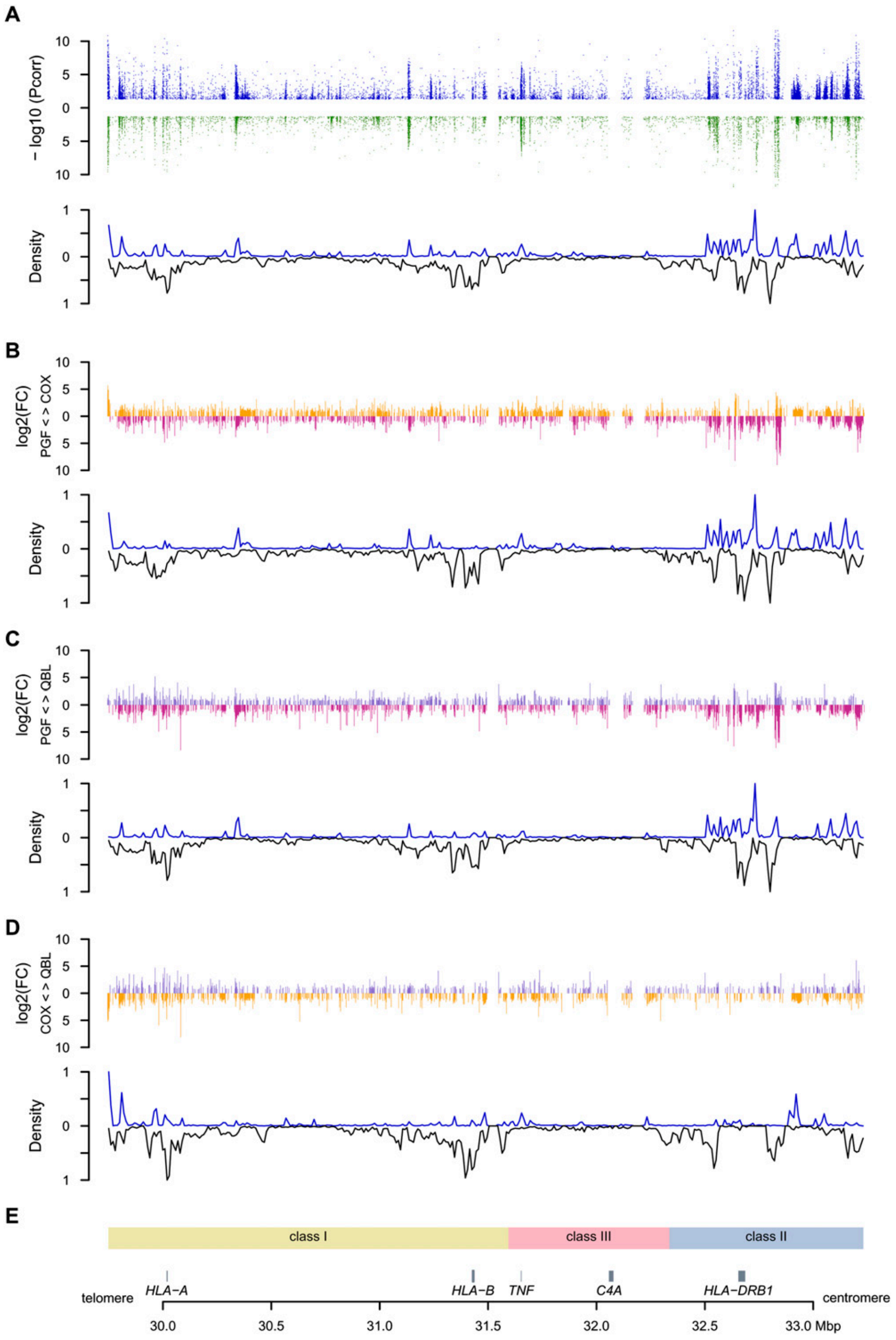


Figure 2. Distribution of differentially expressed (DE) probes versus polymorphic SNPs. Only probes shared by the three haplotypes were included. (*A*) Three-haplotype comparison. (*Upper* panel) Significance level of DE probes for either unstimulated (blue) or stimulated (green) cells. The $-\log_{10}$ of significant adjusted *P*-values are plotted against the genomic coordinates. (*Lower* panel) Density curve of DE probes normalized using the number of probes designed (upward) mirroring the density curve of polymorphic SNPs between the three cell lines (downward) for 350 10-kb windows spanning the MHC. Densities have been normalized. (*B–D*) Pairwise comparisons of COX versus PGF, QBL versus PGF, and QBL versus COX. For each pair, the $\log_2$ of the intensity fold change (FC) is represented in the *upper* panel. For example, when expression is higher in COX than in PGF, the FC is set positive and an orange bar is represented *above* the *x*-axis. Conversely, when expression is higher in PGF, the FC is negative and represented by a pink bar *below* the *x*-axis. The density curves of DE probes and of SNPs polymorphic between both cells are plotted in the *lower* panel. (*E*) Genomic context.

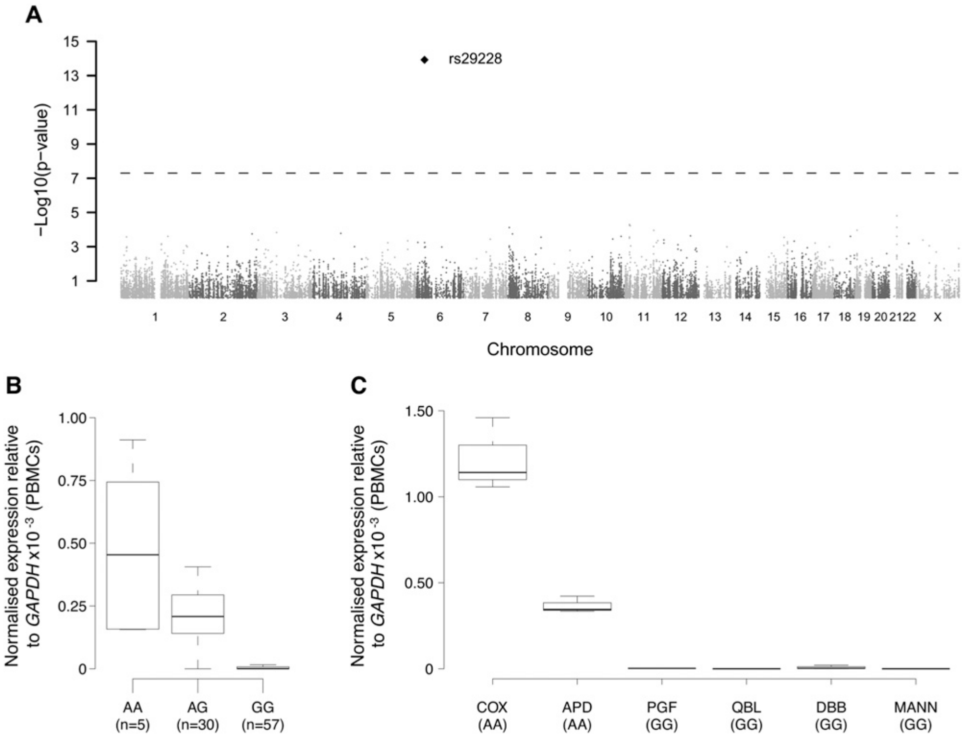


Figure 3. Expression quantitative trait mapping for *ZFP57*. Expression of *ZFP57* was determined by quantitative real-time RT-PCR in peripheral blood mononuclear cells of 93 healthy volunteers and analyzed for association using 45,237 SNPs enriched for immune and inflammatory genes. (*A*) Manhattan plot showing a highly significant association for an SNP, rs29228, 16.8 kb centromeric to *ZFP57*. The horizontal dashed line indicates the genome-wide threshold significance. The absence of other association with neighbor SNPs on chromosome 6 is not unexpected due to moderate SNP coverage in the region and low level of linkage disequilibrium. (*B*,*C*) Boxplots of *ZFP57* gene expression relative to *GAPDH* depending on rs29228 genotype in 92 successfully genotyped individuals (Kruskal-Wallis test on genotypes, $P = 6.7 \times 10^{-11}$) (*B*) or for MHC-homozygous lymphoblastoid cell lines (*C*).

intensity, positive and negative values indicating exon inclusion and exclusion, respectively, with an NI value >1 indicating that the exon is expressed at least twice more or less than the overall gene level. The proportion of exons with NI values different from zero was determined for MHC and other gene sets. This analysis revealed that AS events were strikingly enriched in MHC genes compared to non-MHC genes generally, or specifically to non-MHC genes with an immune function (Fig. 4). This was true whether we considered the number of spliced exons or the number of genes with at least one splice event. Overall, 72.5% of the MHC genes underwent AS of at least a twofold magnitude compared to 62.1% of the non-MHC genes (Fig. 4A). To avoid potential bias due to the number of annotated exons per gene, we determined the significance of these observations by permutations on genes with at least four annotated exons (median number in the genome) in either Vega (Fig. 4B) or Ensembl databases ($P \leq 1 \times 10^{-4}$ for MHC vs. non-MHC and MHC vs. non-MHC immune; not significant for non-MHC immune vs. non-MHC non-immune) (Supplemental Table 6). The extent of AS in the MHC could therefore be considered as a further means to increase diversity of gene expression in this genomic region already characterized by its extreme polymorphism.

The MHC haplo–spliceo–transcriptome

We next considered the extent to which alternative splicing varies between haplotypes. To do this, gene and exon level expression of genes were determined from the probes matching uniquely and perfectly to each haplotype. This analysis showed that these AS events also demonstrate haplotype-specific differences (Supplemental Table 7). In total, we found that 526 (23.9%) exons in the MHC showed haplotypic differences, notably affecting classical HLA genes such as *HLA-DPB2*, *HLA-DQB2*, *HLA-C*, or *HLA-G*. In the latter, AS has been described as playing a critical role in immunomodulation, susceptibility to preeclampsia, and sensitivity to tumor lysis by natural killer cells (Yao et al. 2005; Carosella et al. 2008). We complemented our analysis at the exon level with splice junction resolution in the class III region (where we designed junction probes). We computed junction level intensity values, which were then normalized against the gene intensity. We identified 27 out of 58 genes (46.6%) as showing haplotypic differences in AS (Supplemental Table 8). A number of genes in this region are known to undergo AS such as *AIF1* (Hara et al. 1999). We validated the array results for *AIF1* by RT-PCR, both in terms of exon normalized intensities and junction normalized intensities ($P < 0.02$ for all junctions, ANOVA) with evidence of haplotypic differences (Fig. 5).

Discussion

Our results provide the first high-resolution, strand-specific transcriptional map of the MHC. We find that both intergenic transcription and alternative splicing are abundant in the MHC and that the transcript diversity mirrors the unusually high level of polymorphism found in this region. Specifically, for common disease-associated haplotypes, we have been able to define transcription at haplotype-specific resolution using MHC-homozygous

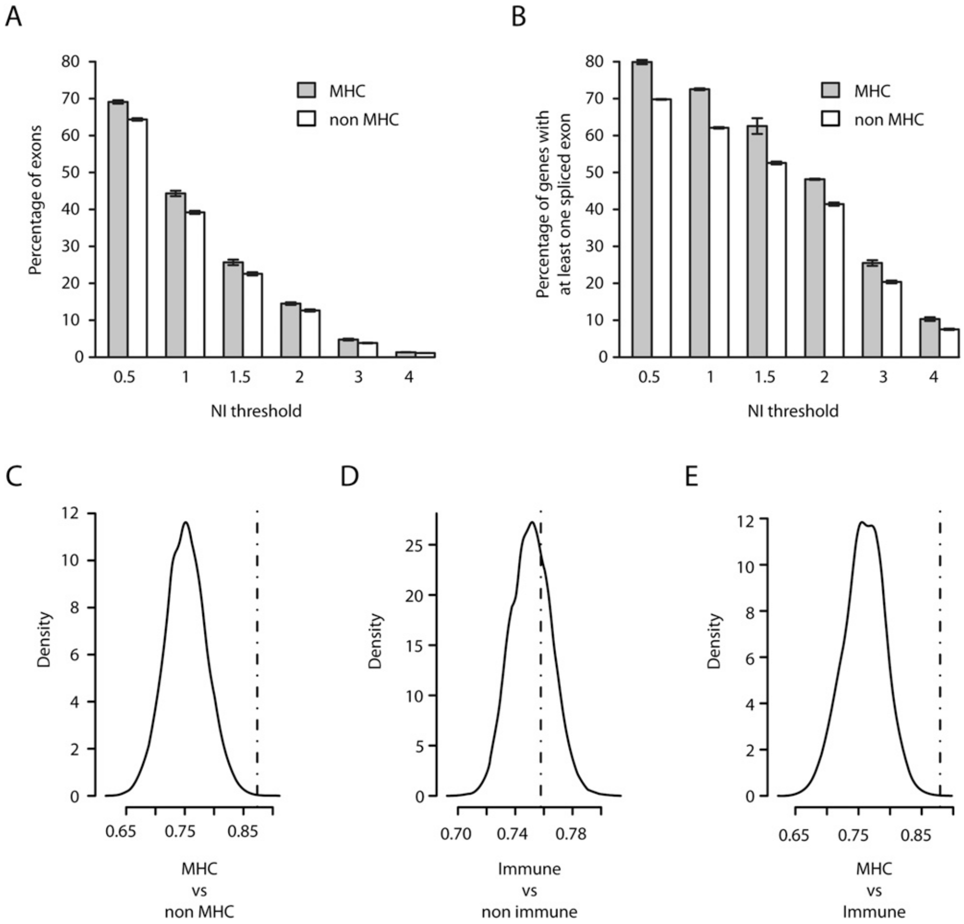


Figure 4. Extent of alternative splicing in the MHC. Absolute values of exon level intensities normalized against gene intensities [NI = $\log_2$(exon/gene)] were computed from the median signal of the three PGF sample replicates hybridized to the Affymetrix Exon 1.0 ST array. Thus, absolute NI > 1 indicates that the exon is expressed at least twice more or less than the overall gene level. Mean percentage of exons (*A*) and of genes with at least one exon (*B*) with NI value(s) exceeding the indicated thresholds for MHC (gray bars) and non-MHC genes (white bars). Error bars depict standard errors of the means of the three replicates (*C–E*) Comparisons of the median NIs (dashed vertical line) in the 131 MHC genes (*C,E*) or in 733 non-MHC immune genes (*D*) having at least four annotated exons in Vega with the density distribution of median NIs obtained in 10,000 random sets of similar numbers of non-MHC (*C*), non-MHC non-immune (*D*), and non-MHC immune (*E*) genes.

LCLs. This has highlighted the extent of variation in gene expression that exists between haplotypes when a large contiguous homozygous sequence is analyzed, in this case, a 3.5-Mb region spanning the classical MHC. Haplotype-specific analysis is critical to advance our understanding of the nature and consequences of genetic variation within the diploid human genome. Given its biological significance and wealth of disease associations, the full haplotype-specific sequence for eight common haplotypes spanning the MHC were defined by the MHC Project (Stewart et al. 2004; Traherne et al. 2006; Horton et al. 2008). Here we complement this data with a haplotype-specific map of transcription in which differentially expressed genes were quantified in the context of phased sequence variants, allowing any regulatory variants to exert allele-specific effects in the naturally occurring genomic context. The use of MHC-homozygous LCLs avoids the confounding effects normally encountered in analysis of the diploid genome and provides a route map for further fine mapping and functional analysis of observed MHC disease associations.

The transcriptional landscape we have described is likely to vary in a context-specific manner, and it will be important to extend our approach to relevant cell/tissue types and conditions for specific MHC-associated diseases. Nevertheless, the three LCLs we used were our material of choice for these first studies of the MHC transcription at a haplotype-specific resolution. Firstly, these cell lines are MHC-homozygous for the disease-associated haplotypes of interest, and we benefited from having available the full MHC sequence for each, facilitating our approach as if we had been studying mouse strains, thus avoiding the issue of recombination. Secondly, linkage of expression phenotypes was first demonstrated in LCLs, proving these phenotypes are not artifactual. Since then, most eQTL mapping studies have investigated that material with reproducible findings (Stranger et al. 2007; Cheung and Spielman 2009; Montgomery et al. 2010; Pickrell et al. 2010). Thirdly, in the MHC we found that 87% of the genes in the region are expressed in that cell type, making it a relevant choice for investigating differential expression between individuals. Finally, we proved with *ZFP57*, *HLA-DQA2*, *HLA-DQB2*, and *HLA-C* that our findings could be replicated in primary peripheral blood cells.

We have shown how a high-density tiling path array design incorporating sequence diversity and splice junctions is a powerful

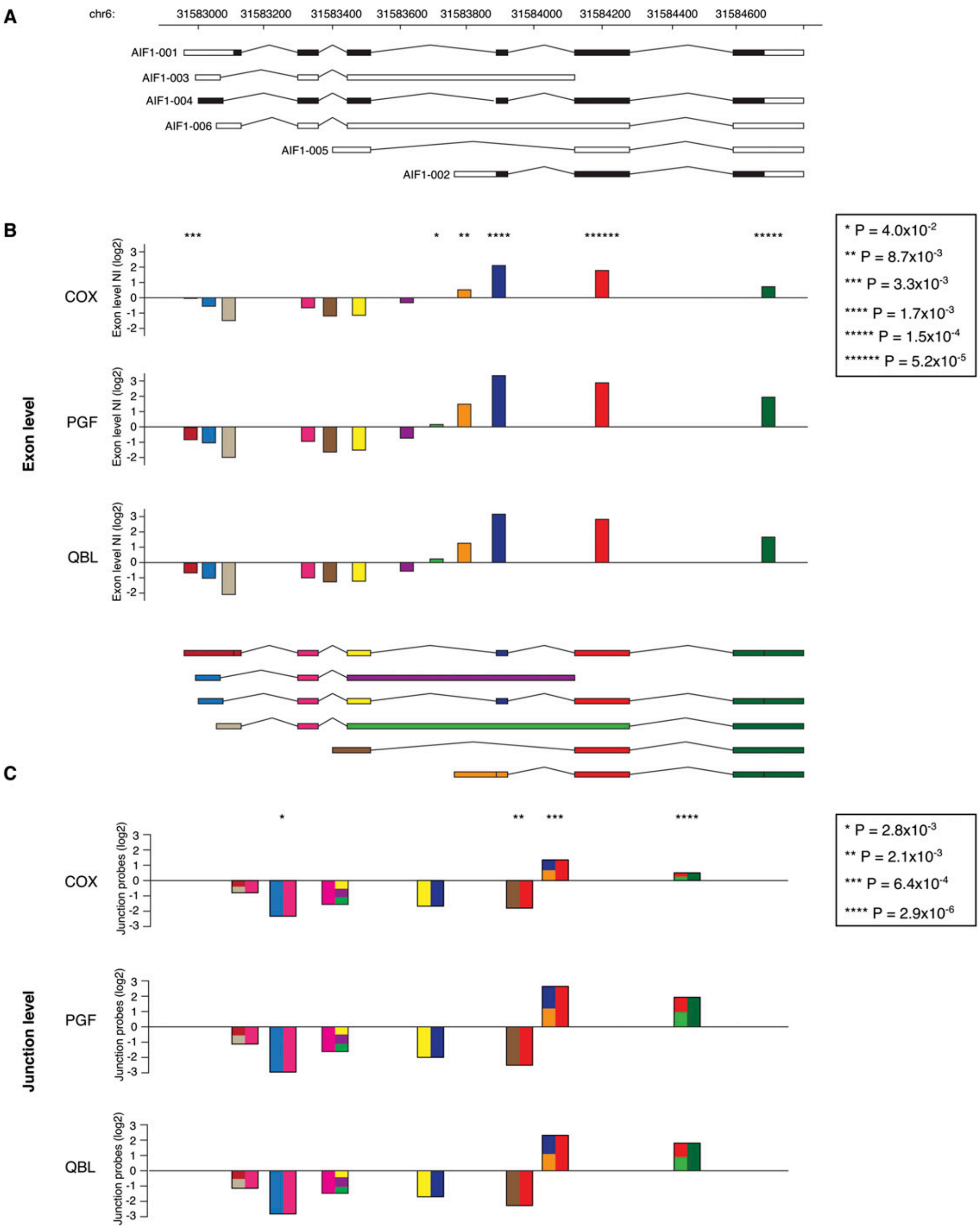


Figure 5. Variation of splicing events in *AIF1* between haplotypes. (*A*) Gene transcripts as they are annotated in Vega. (*B*) Barplots of all exon normalized intensities (NI) for each cell line. The color code for each exon is indicated in the transcript scheme *underneath*. (*C*) Barplots for the junction normalized intensities (JNI). Donor and acceptor exons are represented on each half of the junction with the same color code as in *B*. If the junction is shared between different transcripts, the corresponding site is depicted as a composite of all possible exons. (*B,C*) Asterisks *above* barplots indicate the level of significance, as listed in the caption, for differential expression between the three cell lines. For example, the isoform AIF1-002 tagged by the exon in orange is proportionally more represented in QBL and PGF than in COX. Conversely, the isoform AIF1-005 characterized by the junction between the brown and the red exons is better represented in COX than in PGF and QBL.

tool to help dissect the haplo-spliceo-transcriptome (Graveley 2008) of a large genomic region of interest. Like RNA-seq, our array overcomes some major issues associated with commercial expression arrays; notably, it accounts for underlying sequence polymorphism, allows for identification of new transcribed regions, and monitors splice events. Moreover, these microarrays currently provide a much less costly tool than RNA-seq to assess the transcription status of a chromosomal region the size of the MHC, although we recognize the greater dynamic range and allele-specific resolution of this technology (Wang et al. 2009). We have applied the MHC array to MHC-homozygous individuals, but the custom array should also be informative when applied to heterozygous samples. Here it will be necessary to know the DNA sequence or relevant genotypic information to select the correct probes for analysis and interpret the data correctly. The tiling path probe set can then be defined across the MHC by individual. At positions where the individual is heterozygous, the average of the two informative allele-specific probes can be taken.

Our findings that only 6% of the region is transcribed might appear as a low figure. This, however, is the same order of magnitude as, for example, 4.6% of the total length of ENCODE regions screened on tiling arrays (ENCODE Project Consortium et al. 2007). That 31% of the TARs, representing 28.3% of total transcribed MHC sequence, were found in intergenic regions is also in the line with the 25% recent estimate of "dark matter transcripts" obtained by RNA-seq (Ponting and Belgard 2010) and is, however, of considerable interest, given that the MHC region is the most gene-dense region of the human genome. Their expression was overall low, which might explain why we failed to detect haplotype-specific differences and correlation with SNPs localization, unlike for TARs in genic areas. The biological significance of these TARs is unclear. We have not investigated their processing, but the fact that of the 50% localizing distantly (>10 kb) from annotated genes, the majority colocalize with *Alu* sequences is particularly intriguing. It is known that *Alu* sequences can be actively transcribed and may contribute to the emergence of alternative splicing or even new genes and pseudogenes (Deininger et al. 2003). Some classical HLA genes have been postulated to derive from such repeat elements. The intergenic TARs we detect might therefore reflect an ongoing process of exonization of transposed elements leading to the emergence of new MHC genes, also important for the regulation of existing genes and therefore eventually for MHC-associated pathology.

It has been suggested that alternative splicing is a key modulator of immune gene expression (Lynch 2004), possibly leading to antagonist effects as seen for *MYD88* or *CD44*. A previous study has revealed that up to 94% of human genes are alternatively spliced across 15 tissues tested from different individuals (Wang et al. 2008). We found that in a single cell type from a single individual, 72.5% of the MHC genes are alternatively spliced. Moreover, alternative splicing is enriched among MHC genes compared to non-MHC immune genes. We also demonstrate that alternative splicing is related to the haplotypic structure. In the context of common ancestral MHC haplotypes, one can thus imagine that alternative splicing is used by evolution to generate more transcript diversity in the MHC while preserving some of the haplotypic structure. Consequences of such splicing patterns can lead to dramatic consequences as already highlighted by mutations in the *BTLN2* gene associated with sarcoidosis (Valentonyte et al. 2005).

Our study presents a proof of principle that, beyond standard SNP-based eQTL mapping studies, it is possible to directly study haplotype-specific gene expression at a high resolution for a 3.5-Mb region and find striking differences. Our approach will be of value in a generic sense for characterizing other genomic intervals identified by GWAS or other approaches. Risk haplotypes are ultimately associated to the phenotypes, and identifying genes differentially expressed can reduce the number of genes to study at the disease locus region.

For the MHC, this is of particular interest given the remarkable number of associations with common diseases reported, while the fine mapping of functionally important regulatory variants remains a challenge. That such important differences in gene expression could be detected by investigating only three haplotypes supports the hypothesis that they play a role in the autoimmune diseases associated with these haplotypes. Our data suggest a number of candidate variants and gene transcripts for further characterization. For example, Figure 1 presents a graphical overview of the locus showing differentially expressed genes by haplotype, while Table 2 lists 37 candidate genes potentially accounting for the association of the *HLA-A1-B8-DR3* haplotype with numerous diseases.

For the MHC, both structural and regulatory genetic variants are important in determining disease susceptibility, and our approach to this region needs to consider such variants if causal relationships are to be established. Our analysis has provided new insights into how transcription differs between individuals across the classical MHC, and our custom array can be used to quantify haplotype-specific differences in related contexts such as DNA methylation or chromatin accessibility based on DNase hypersensitivity (Sabo et al. 2006; Weber et al. 2007). As our knowledge of the complexities of gene regulation continues to grow, it is important to acknowledge how much remains to be understood and the need for a more complete picture of gene expression beyond transcript level analysis. At a mechanistic level, much attention has focused on modulation of transcriptional initiation, but sequence diversity will impact in multifaceted ways on the whole process of transcription and translation, as well as how chromatin is packaged and gene expression coordinated at a local and global level within the nucleus. It will be critical to establish the nature and basis of individual epigenetic variation, defining how this may be modulated by underlying DNA sequence variation as well as environmental factors relevant to disease. We believe our analysis opens the door to such studies and provides an important further step in our quest to define the functional basis of the remarkable disease associations found for this region of the genome.

Methods

Full methods and any associated references are available in the Supplemental Material.

Samples

Lymphoblastoid cell lines (LCL)

COX was obtained both from The International Histocompatibility Working Group (IHW, ref 0922) and by the generosity of S. Marsh and N. Mayor (Anthony Nolan Research Institute, UK). PGF and QBL were purchased from the European Cell Culture Collection (Salisbury, UK; ref 94050342 and 94070713). Chromosome integrity was checked by FISH. In addition to chromosome 6 painting, the AF129756 BAC (The Sanger Institute) encompassing most of the class III region was used as a second probe to verify MHC integrity. Genotypes of HLA classical molecules (HLA-A, B, C, DR, and DQ) were verified by the Tissue Typing Laboratory in Oxford (Dr. Barnardo Martin), while the homozygosity and genotypes of microsatellites along the class III region (D6S272, D6S2800, MICA, TNFb, and D6S2789) were checked as described before (Vandiedonck et al. 2004). Apart from D6S272, which showed heterozygosity for PGF, all other markers showed the expected genotypes. Genotypes for 410 SNPs in the MHC region were also verified for COX, PGF, and QBL using a cardiovascular gene-centric 50 K SNP array (humanCVD bead array; Illumina) (Keating et al. 2008). With one exception (rs562047 found G/C in QBL), all genotypes were those expected. To follow up results on *ZFP57*, *HLA-DQA2*, *HLA-DQB2*, and *HLA-C* expression, three additional MHC-homozygous cell lines—MANN/MOU, DBB, and APD (IHW9050, 9052, 9291)—were studied, and their MHC genotypes were also checked with the cardiovascular array (Keating et al. 2008).

PBMCs from healthy volunteers

PBMCs from healthy volunteers were recruited with cDNA prepared as described in Fairfax et al. (2010). Their genomic DNA was extracted using Puregene kits (Gentra Systems, Inc.). Genotyping on the humanCVD bead array was performed using genomic DNA from the volunteers and homozygous LCLs DNA, as previously described (Fairfax et al. 2010). For two specific SNPs not included on the array—rs2269423 and rs9264942—genotyping was performed by direct Sanger sequencing (primer sequences available on request). For some genes subsequently interrogated by expression quantitative trait mapping, genotyping and/or gene expression data were not available for all volunteers. The total numbers of volunteers included for each gene analyzed are shown in the associated figure legends.

Design of the MHC array

The MHC array was designed for the Affymetrix platform (Affymetrix) using ad hoc algorithms as described in detail in the Supplemental Methods. Criteria of uniqueness against the genome and transcriptome and of structural conformation were considered. Known polymorphisms and segmental duplications have been incorporated into the design.

Experimental procedures

Cell culture

Lymphoblastoid cells were grown in triplicate at a minimum density of 6 × 10^5 cells/mL in RPMI 1640 (Sigma, lot 16K2379) supplemented with 10% Fetal Calf Serum Gold (PAA, lot A64095-0537) and 2 mM L-glutamine (PAA, lot M00406-0102) at 37°C in a 5% CO_2 wet environment. Cultures were stimulated or not for 6 h with 200 nM phorbol 12-myristate 13 acetate (PMA; Sigma) and 125 nM ionomycin (Sigma) and harvested at 8 × 10^5 to 1 × 10^6 cells/mL in log growth phase. Volunteers peripheral blood mononuclear cells (PBMC) were prepared as previously described (Fairfax et al. 2010).

RNA extraction

Total RNA was isolated using RNeasy midiprep kits (QIAGEN) including on column DNase I digestion. Quantifications were done by Nanodrop (ThermoScientific), and integrity was verified using a 2100 Bioanalyzer (Agilent). All samples had a RNA integrity number >9. Genomic DNA contamination was checked by real-time PCR and was <0.1%.

Array experimental design

We hybridized samples from the unstimulated and stimulated triplicate cultures of COX, PGF, and QBL LCLs to custom MHC arrays, while only unstimulated samples were hybridized to commercial Affymetrix Exon 1.0 ST arrays.

Sample labeling and array hybridization

We used the GeneChip Whole Transcript (WT) Sense Target Labeling kit (Affymetrix), following the manufacturer's instructions, starting with 1.5 μg of total RNA and including the ribosomal RNA depletion step (Ribominus kit; Invitrogen). Then, cDNA was synthesized using random hexamers tailed with a T7 promoter to avoid 3′ bias, and the complementary strand of RNA was generated by an in vitro transcription reaction. Subsequently, a new first strand of cDNA was synthesized, complementary to the initial cDNA, in the same orientation as the mRNA and denoted "ccDNA." It was fragmented (range 40–70 bases), end-labeled, and hybridized to the MHC and GeneChip Exon 1.0 ST arrays (Affymetrix) for 16 h at 45°C following the manufacturer's instructions. Reduced RNA, cRNA following IVT, and fragmented ccDNA were verified on a 2100 Bioanalyzer. Hybridized arrays were then washed and stained on a GeneChip Fluidics 450 workstation (Affymetrix) using the FS450_0001 protocol. The arrays were scanned on a GCS3000 Scanner (Affymetrix).

cDNA synthesis and RT–PCR for validation

cDNA was synthesized using random hexamers and Superscript III Reverse Transcriptase (RT) (Invitrogen) as per the manufacturer's instructions including control reactions without reverse transcriptase for each sample. Quantitative PCR was performed on three technical replicates using SYBR Green Supermix (Bio-Rad) on an iQ Cycler (Bio-Rad). PCR efficiency was determined using serial dilutions of pooled cDNAs from COX, PGF, and QBL cells. Melt curve analysis was performed for gene-specific primer sets. Relative gene transcript levels were determined by the ΔΔCt method. Primer sequences are available upon request.

Array signal processing

Custom MHC array signals

Custom MHC array signals were processed using an in-house pipeline under R and Bioconductor environment and using Perl scripts as detailed in the Supplemental Methods. Briefly, after preprocessing all probes, tiling and junction probes were analyzed independently. Tiling path analysis was conducted to determine the extent of transcription on the shared path and on each of the alternate paths. Transcription within a gene was assessed by the inclusion of at least one TAR at a FDR of 5%. Alternative splicing was evaluated on each of the alternate paths both at the exon level and, for the MHC class III region, at the splice junction level. Exon and junction intensities were normalized against the gene level intensities.

Exon array signals

Exon array signals were processed using Affymetrix Power Tools, and R scripts (see Supplemental Methods). Briefly, probe-level analysis was carried out for cross-platform validation, while gene level and alternative splicing analyses were performed using custom CDF files from the Microarray Lab (http://brainarray.mbni.med.umich.edu/).

Quality controls

Quality controls are given in the Supplemental Methods.

Statistical analyses

All statistical analyses, including distribution of the TARs; comparison of expression between haplotypes; correlation of differentially expressed probes with SNP distribution; eQTL mapping for *ZFP57*, *HLA-DQA2*, and *HLA-DQB2*; and analysis of the extent of the MHC splicing were performed using R, Perl, and PLINK as provided in detail in the Supplemental Methods.

Acknowledgments

This study was funded by the Wellcome Trust (grant number 074318/075491/Z/04) and the Multiple Sclerosis Society (grant number 875/07). C.V. has been supported by the Fondation pour la

Recherche Médicale since July 2009. We are indebted to T. Watt and C. Blancher for technical assistance, N. Wilson and E. Volpi for performing the FISH analysis, B. Martin and H.J. Garchon labs for genotyping classical MHC genes and microsatellites, and S. Marsh and N. Major for generously providing the COX cell line. We are grateful to the volunteers who agreed to participate. We are grateful for critical discussions to P. Donnelly, G. McVean, D. R. Campbell, R. Mott, C. Julier, and H.J. Garchon. We thank all other members of the Knight laboratory for technical assistance.

Authors' contributions: C.V. and J.K. designed the study; M.S.T. and C.V. designed the array; C.V. prepared samples and performed the array experiments and validations; K.P. performed the volunteers experiment; B.F. collected the volunteers; C.V., H.L., C.D., J.M.T., J.B., and J.K. analyzed the array data; K.P. and B.F. analyzed the volunteers data; all authors edited and proofread the manuscript; C.V. and J.K. wrote the manuscript.

References

Ahmad T, Neville M, Marshall SE, Armuzzi A, Mulcahy-Hawes K, Crawshaw J, Sato H, Ling KL, Barnardo M, Goldthorpe S, et al. 2003. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum Mol Genet* **12:** 647–656.

Barcellos LF, Oksenberg JR, Begovich AB, Martin ER, Schmidt S, Vittinghoff E, Goodin DS, Pelletier D, Lincoln RR, Bucher P, et al. 2003. HLA-DR2 dose effect on susceptibility to multiple sclerosis and influence on disease course. *Am J Hum Genet* **72:** 710–716.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242–2246.

Carosella ED, Favier B, Rouas-Freiss N, Moreau P, Lemaoult J. 2008. Beyond the increasing complexity of the immunomodulatory HLA-G molecule. *Blood* **111:** 4862–4870.

Cheung VG, Spielman RS. 2009. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* **10:** 595–604.

Conde L, Halperin E, Akers NK, Brown KM, Smedby KE, Rothman N, Nieters A, Slager SL, Brooks-Wilson A, Agana L, et al. 2010. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nat Genet* **42:** 661–664.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. *Nat Rev Genet* **10:** 184–194.

Dausset J. 1981. The major histocompatibility complex in man. *Science* **213:** 1469–1474.

de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, et al. 2006. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet* **38:** 1166–1172.

Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. 2003. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* **13:** 651–658.

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, et al. 2007. A genome-wide association study of global gene expression. *Nat Genet* **39:** 1202–1207.

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al. 2008. Genetics of gene expression and its effect on disease. *Nature* **452:** 423–428.

ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Fairfax BP, Vannberg FO, Radhakrishnan J, Hakonarson H, Keating BJ, Hill AV, Knight JC. 2010. An integrated expression phenotype mapping approach defines common variants in LEP, ALOX15 and CAPNS1 associated with induction of IL-6. *Hum Mol Genet* **19:** 720–730.

Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, Panhuis TM, Mieczkowski P, Secchi A, Bosco D, et al. 2010. A map of open chromatin in human pancreatic islets. *Nat Genet* **42:** 255–259.

Giraud M, Taubert R, Vandiedonck C, Ke X, Levi-Strauss M, Pagani F, Baralle FE, Eymard B, Tranchant C, Gajdos P, et al. 2007. An IRF8-binding promoter variant and AIRE control CHRNA1 promiscuous expression in thymus. *Nature* **448:** 934–937.

Graveley BR. 2008. The haplo-spliceo-transcriptome: common variations in alternative splicing in the human population. *Trends Genet* **24:** 5–7.

Hamza TH, Zabetian CP, Tenesa A, Laederach A, Montimurro J, Yearout D, Kay DM, Doheny KF, Paschall J, Pugh E, et al. 2010. Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet* **42:** 781–785.

Hara H, Ohta M, Ohta K, Nishimura M, Obayashi H, Adachi T. 1999. Isolation of two novel alternative splicing variants of allograft inflammatory factor-1. *Biol Chem* **380:** 1333–1336.

Hor H, Kutalik Z, Dauvilliers Y, Valsesia A, Lammers GJ, Donjacour CE, Iranzo A, Santamaria J, Peraita Adrados R, Vicario JL, et al. 2010. Genome-wide association study identifies new HLA class II haplotypes strongly protective against narcolepsy. *Nat Genet* **42:** 786–789.

Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, et al. 2004. Gene map of the extended human MHC. *Nat Rev Genet* **5:** 889–899.

Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, Forbes S, Gilbert JG, Halls K, Harrow JL, et al. 2008. Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. *Immunogenetics* **60:** 1–18.

Johansson S, Lie BA, Todd JA, Pociot F, Nerup J, Cambon-Thomsen A, Kockum I, Akselsen HE, Thorsby E, Undlien DE. 2003. Evidence of at least two type 1 diabetes susceptibility genes in the HLA complex distinct from HLA-DQB1, -DQA1 and -DRB1. *Genes Immun* **4:** 46–53.

Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT, Galver L, Barrett JC, Grant SF, Farlow DN, et al. 2008. Concept, design and implementation of a cardiovascular gene-centric 50 K SNP array for large-scale genomic association studies. *PLoS ONE* **3:** e3583. doi: 10.1371/journal.pone.0003583.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11:** 345–355.

Kiran A, Baranov PV. 2010. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics* **26:** 1772–1776.

Knight JC, Keating BJ, Kwiatkowski DP. 2004. Allele-specific repression of lymphotoxin-alpha by activated B cell factor-1. *Nat Genet* **36:** 394–399.

Larsen CE, Alper CA. 2004. The genetics of HLA-associated disease. *Curr Opin Immunol* **16:** 660–667.

Li X, Ito M, Zhou F, Youngson N, Zuo X, Leder P, Ferguson-Smith AC. 2008. A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. *Dev Cell* **15:** 547–557.

Lynch KW. 2004. Consequences of regulated pre-mRNA splicing in the immune system. *Nat Rev Immunol* **4:** 931–940.

Mackay DJ, Callaway JL, Marks SM, White HE, Acerini CL, Boonen SE, Dayanikli P, Firth HV, Goodship JA, Haemers AP, et al. 2008. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. *Nat Genet* **40:** 949–951.

Manolio TA. 2010. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363:** 166–176.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464:** 773–777.

Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET. 2010. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6:** e1000895. doi: 10.1371/journal.pgen.1000895.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464:** 768–772.

Ponting CP, Belgard TG. 2010. Transcribed dark matter: meaning or myth? *Hum Mol Genet* **19:** R162–R168.

Price P, Witt C, Allcock R, Sayer D, Garlepp M, Kok CC, French M, Mallal S, Christiansen F. 1999. The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol Rev* **167:** 257–274.

Rioux JD, Goyette P, Vyse TJ, Hammarstrom L, Fernando MM, Green T, De Jager PL, Foisy S, Wang J, de Bakker PI, et al. 2009. Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc Natl Acad Sci* **106:** 18680–18685.

Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* **3:** 511–518.

Shiina T, Inoko H, Kulski JK. 2004. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens* **64:** 631–649.

Singer JB, Lewitzky S, Leroy E, Yang F, Zhao X, Klickstein L, Wright TM, Meyer J, Paulding CA. 2010. A genome-wide study identifies HLA alleles associated with lumiracoxib-related liver injury. *Nat Genet* **42:** 711–714.

Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P, Dunham I, Forbes S, Halls K, Howson JM, et al. 2004. Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res* **14:** 1176–1187.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, et al. 2007. Population genomics of human gene expression. *Nat Genet* **39:** 1217–1224.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466:** 707–713.

Thomas R, Apps R, Qi Y, Gao X, Male V, O'Huigin C, O'Connor G, Ge D, Fellay J, Martin JN, et al. 2009. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet* **41:** 1290–1294.

Traherne JA, Horton R, Roberts AN, Miretti MM, Hurles ME, Stewart CA, Ashurst JL, Atrazhev AM, Coggill P, Palmer S, et al. 2006. Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet* **2:** e9. doi: 10.1371/journal.pgen.0020009.

Vafiadis P, Bennett ST, Todd JA, Nadeau J, Grabs R, Goodyer CG, Wickramasinghe S, Colle E, Polychronakos C. 1997. Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. *Nat Genet* **15:** 289–292.

Valentonyte R, Hampe J, Huse K, Rosenstiel P, Albrecht M, Stenzel A, Nagy M, Gaede KI, Franke A, Haesler R, et al. 2005. Sarcoidosis is associated with a truncating splice site mutation in BTNL2. *Nat Genet* **37:** 357–364.

van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8:** e1000371. doi: 10.1371/journal.pbio.1000371.

Vandiedonck C, Knight JC. 2009. The human Major Histocompatibility Complex as a paradigm in genomics research. *Brief Funct Genomics Proteomics* **8:** 379–394.

Vandiedonck C, Beaurain G, Giraud M, Hue-Beauvais C, Eymard B, Tranchant C, Gajdos P, Dausset J, Garchon HJ. 2004. Pleiotropic effects of the 8.1 HLA haplotype in patients with autoimmune myasthenia gravis and thymus hyperplasia. *Proc Natl Acad Sci* **101:** 15464–15469.

Walter NA, McWeeney SK, Peters ST, Belknap JK, Hitzemann R, Buck KJ. 2007. SNPs matter: impact on detection of differential expression. *Nat Methods* **4:** 679–680.

Wang GS, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8:** 749–761.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63.

Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39:** 457–466.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661–678.

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. 2002. Allelic variation in human gene expression. *Science* **297:** 1143.

Yao YQ, Barlow DH, Sargent IL. 2005. Differential expression of alternatively spliced transcripts of HLA-G in human preimplantation embryos and inner cell masses. *J Immunol* **175:** 8379–8385.

Yunis EJ, Larsen CE, Fernandez-Vina M, Awdeh ZL, Romero T, Hansen JA, Alper CA. 2003. Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks. *Tissue Antigens* **62:** 1–20.

Received October 31, 2010; accepted in revised form April 15, 2011.

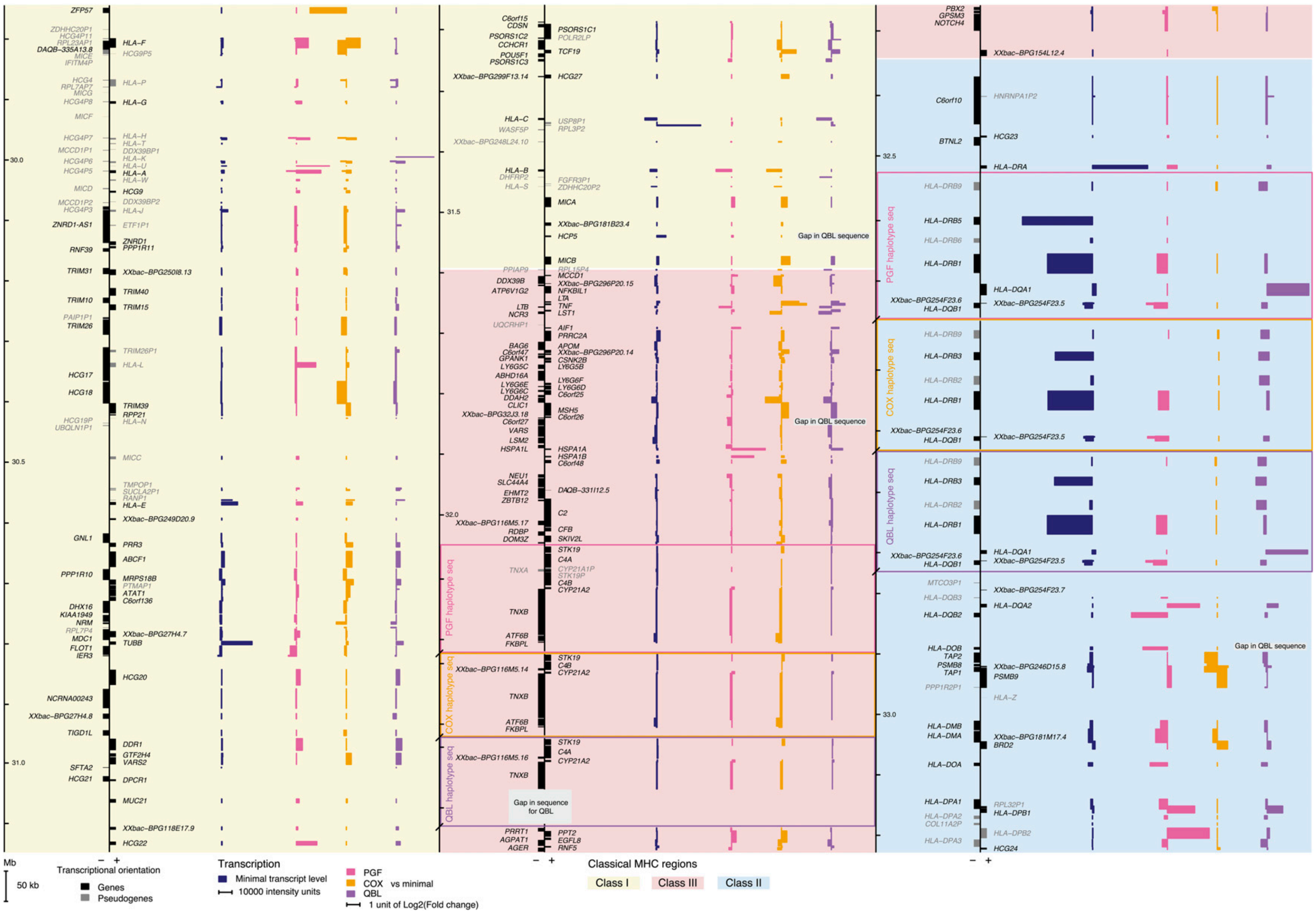


Figure 1. The first transcriptional map of the human MHC: Haplotype-specific pattern of expression. Each cell line transcription level against the minimal level of the three cell lines is displayed; (pink) PGF, (orange) COX, and (purple) QBL. Bars corresponding to genes on the minus strand are displayed *leftward* and those for the plus strand *rightward*. Boxes in the class III and class II regions indicate haplotype-specific segmental duplications. Overall, the MHC class II region presents the highest minimal expression (dark blue). The most differentially expressed gene is *ZFP57*, located at the beginning of the class I region.

Supplemental Information for

Pervasive haplotypic variation of the spliceo-transcriptome of the human Major Histocompatibility Complex

Claire Vandiedonck[1,2,3±], Martin S Taylor[1,4], Helen E Lockstone[1], Katharine Plant[1], Jennifer M Taylor[1], Caroline Durrant[1], John Broxholme[1], Benjamin P. Fairfax[1] and Julian C Knight[1±]

[1] *Wellcome Trust Centre for Human Genetics, Oxford University, Roosevelt Drive, OX3 7BN, UK.*
[2] *INSERM UMRS-958, 10 avenue de Verdun, 75010 PARIS, France*
[3] *Université Paris 7 Denis-Diderot, Paris, France*

[4] *MRC Human Genetics Unit, Edinburgh, EH4 2XU, UK*

[±] to whom correspondence should be addressed :
claire.vandiedonck@inserm.fr and julian@well.ox.ac.uk

Supplemental Material includes:

Supplemental Methods

Supplemental Figures 1 to 8

Supplemental Tables 1 to 8

Supplemental References

TABLE OF CONTENTS:

1. SUPPLEMENTAL METHODS

1.1. The MHC Array

The MHC array was manufactured by Affymetrix using their 49-7875 format with 11 µm features and all probes were 25 nucleotides in length. The design incorporated tile probes and splice junction probes as well as alternate allele versions in both the tile and splice junction sets. For every Watson strand probe the corresponding Crick probe was also included on the array, for both tile and junction probes. For the junction set, we spotted each probe in four technical replicates on the array.

1.1.A. Target regions and annotation

The reference sequence for the MHC array design was based on the NCBI:35 (aka. UCSC:hg17, May 2004) reference human genome assembly. Except where otherwise stated, the annotation used in the array design was from the hg17 UCSC Genome Browser annotation database (September 2006). Regions targeted for tile probe design correspond to chr6:29748239-33231091. We were able to construct tile paths at an average resolution of one (Watson strand) probe per 18 nucleotides (nt) across 86% of this target region. The sequence of the remaining 14% was not sufficiently unique in the genome for discriminatory probes to be designed. A more tightly defined region of chr6:31593635-32482819 was additionally targeted for the design of splice junction probes. The complete extent of the hg17 reference assembly alternate HLA_HAP1 and HLA_HAP2 haplotype assemblies were also targeted for tile probe design.

Subsequently, we remapped all the probes using the hg18 reference sequence and analyzed and presented the data using this build. Importantly, the MHC reference sequence (that is from PGF) has remained unchanged between hg17, hg18 and the current build, hg19. The latter two builds differ only by a constant shift in the numbering that is of 107,079 nt on the reference sequence, of 63,673 nt on COX sequence and of 46,053 nt on QBL sequence.

1.1.B. Probe types

Tile probes. For the targeted genomic intervals, every possible tile probe was evaluated (*n*-25+1 Watson strand probes, where *n* is the length of the target interval in nt). Our target density was one probe per 18 nucleotides of target sequence, resulting in a tile path of overlapping probes of unprecedented resolution.

Splice junction probes. A genome wide, database of splice junction sequences was produced, comprising 850,552 unique junctions. This was based on annotation of splice donors and acceptors derived from pre-computed alignments of UCSC known genes, RefSeq transcripts, VEGA annotation, public mRNA sequences and Acembly clustered ESTs on the reference genome. All alignment data were obtained from UCSC hg17 database tables. Splice junction sequences used for the design typically comprised 24 nt exonic splice donor concatenated with 24 nt exonic splice acceptor, though exonic sequences were truncated at points of overlap with other annotated splice sites. Splice junctions mapping to the hg17 reference sequence coordinates chr6:31593635-32482819 were targeted for the design of splice junction traversing probes. Our aim was six overlapping, Watson strand, splice junction traversing probes per targeted splice junction.

Alternate allele probes. Where a candidate probe overlapped a known single nucleotide (substitution) polymorphism, an alternate allele version of that probe was also designed. In the cases where multiple known polymorphisms overlap the probe position, alternate allele probes were designed that represented every permutation of SNP phasing.

Affymetrix exon probes. To enable direct comparison with gene expression measured by Affymetrix human Exon array 1.0 ST (HuEx1), we incorporated all probes from HuEx1 that mapped into our tiling array target regions, regardless of their quality score or uniqueness. Many of these HuEx1 probes are not genome-wide unique, only unique probes were included in the tile-path analyses reported in this work.

Control probes. A total of 19,184 control perfect match (PM) probes were selected from the Affymetrix arrays: 99 for spikes (antisense), 640 for housekeeping genes (both strands), 138 for *Alu* (both strands) and 264 for polyA (both strands) from the Human Genome U133 Plus 2.0 array, as well as 1,100 probes corresponding to introns and exons of housekeeping genes and 16,943 antigenomic with different content of GC from the Exon 1.0 ST array. We also included 10,973 probes, only PM and in sense orientation, corresponding to 994 probesets of 371 non-MHC genes involved in alternative splicing, immune response, cell cycle, signaling, as well as human tissue specific genes. In addition, 17 genomic regions (range 212 nt to 19.5 kb), considered as positive and negative controls for future array applications, were incorporated into the tiling and/or splicing design.

1.1.C. Probe selection

Each candidate probe was scored for its genome-wide uniqueness, sequence properties (probe quality), and overlap with polymorphic sites. For each of these measures we defined an optimum (opt) value as well as thresholds for acceptable minima (min) and maxima (max).

Uniqueness. We considered two measures of uniqueness for candidate probe sequences. First, the number of nucleotide identities between the candidate probe and it's best non-self match in the genome or whole-genome splice junction dataset (opt=0, min=0, max=24; i.e. we required the candidate probe to be at least one substitution away from any other sequence). The second uniqueness measure was the number of highly similar (up to three substitutions from the candidate probe) sequences in the genome (opt=0, min=0, max=9). For tile probes, alignments were calculated using a customized version of Olly (Jim Kent, UCSC, Unpublished; customization to stop searching for matches after the tenth match is found) that finds all nearly-identical matches, up to a specified substitution distance, in ungapped alignment between a query sequence and a target genome. This allowed the uniqueness measures to be calculated for non-repeat masked genome, considered desirable as diverged repetitive sequences often contain genome-wide unique 25 mers, adding to the target sequence coverage, particularly in regions that are otherwise difficult to assay.

For the candidate junction probes, alignments to the reference genome and the genome wide splice junction dataset were performed using BLASTN (NCBI blastall with options -FF -w7) which allows for alignment gaps as well as substitutions. Self matches and partial self matches (where splice junctions share either a splice donor or acceptor exonic sequence) were filtered out. For uniqueness scoring, we considered alignment identity as with Olly based alignments, and ignored alignment gaps.

Probe quality score. The Affymetrix probe score as detailed in (Mei et al. 2003) was applied to candidate probes and averaged over the Watson and Crick strands (opt=0.8, min=0.08, max=0.8). This score incorporates measures of sequence composition, secondary structure and hybridization thermodynamics.

Overlapping polymorphisms. To minimize the number of permutations of alternate allele probes and to simplify the interpretation of hybridization results between haplotypes we sought to minimize the number of known polymorphisms (dbSNP build 126) overlapping a probe (opt=0, min=0, max=2).

Probe tiling/spacing. Tracts of target genome sequence (> 24 nt) for which no unique probes could be designed, and Affymetrix HuEx1 derived probes already incorporated into the design (see above) provided a natural punctuation with which to constrain the choice of the probe tiling path. Between these fixed positions, we calculated an optimal tiling path of probes using a scoring system that incorporated the uniqueness, probe quality and polymorphism overlap measures described above as well as relative spacing measured as the distance between the midpoint of adjacent probes (opt=18, min=12, max=24), all transformed onto a unified scale (*D*) and equally weighted using equation 1:

$$D = \frac{\sum_{i=1}^{N} \left(\frac{(Obs_i - Opt_i)}{(Opt_i - (Min_i|Max_i)+p)} \right)^2}{N}$$

Where Obs, Opt, Min and Max were respectively: observed values, optimal values and minimal and maximal allowed values. Min_i is used where Obs_i < Opt_i, otherwise Max_i was used. p is a small constant (0.0001) that overcomes the problem of divisions by zero. This rescaled score was averaged (sum divided by the number *N* of measures *i*) over each of the probe measures (*i* represents the probe measures: best non-self match, number of non-self matches, probe quality score, overlapping polymorphisms and probe spacing). As the probe spacing parameter was dependent on adjacent probes we applied Dijkstra's dynamic programming algorithm (Aho et al. 1983) to select an optimal tile path.

Segmental duplications. There is a known, high identity segmental duplication within the targeted MHC region, defined as hg17 coordinates chr6:32056288-32089047 and chr6:32089084-32121844. Specifically for these regions, uniqueness criteria were adjusted so that a probe match to either region was considered a self-match, allowing tile-paths to be designed through these regions.

Alternate haplotypes. Initial probe selection (including alternate allele probe design) was based on the PGF haplotype in the main reference assembly. After mapping designed probes (including alternate allele versions) into the alternate COX and QBL haplotypes using a hash table, remaining gaps in the tile path were closed by selecting an optimal tile path as described above. As with the segmentally duplicated regions, uniqueness criteria were adjusted over the targeted alternate haplotype intervals, so that matches in any of the haplotypes were considered self-matches.

Junction probe specific parameters. Splice junction traversing probes were selected using the same set of parameters as for tile probes, but the probe spacing was adjusted to (opt=3, min=1, max=24). We also constrained selection to only consider probes that had nucleotides corresponding to at least 4 nucleotides from both exons of the splice junction.

1.2. Array signal processing

1.2.A. Custom MHC array processing. Custom array signals were processed using an in-house pipeline under the R and Bioconductor environment, and using Perl scripts.

Preprocessing. Affymetrix data intensities (.CEL files) were read using the "affy" Bioconductor software package. Signal intensity was first background corrected for each array by subtracting the median intensity of 2,524 blanks. Using control antigenomic GC bins, we corrected probe intensities by subtracting the corresponding GC bin median intensity. Experimental probe GC content ranged from 3 to 21. Between array normalization was done using vsn (Huber et al. 2002), which also stabilizes the variance by transforming intensities to a generalized log scale to base 2. Finally, probes showing excessive dispersion between probe replicates first, and then between biological replicates, were filtered out. In all cases, they correspond to the first percentile of the distribution of the ratio standard deviation/mean.

Tiling path analysis. The smoothing process involved two steps (Sabo et al. 2006). First, using a moving window of 100 bases, we averaged the signal of the probes that were above the background set at the fifth percentile. Then the signal was weighted for distance from the centre of the window using a Gaussian function whose σ was defined at 30 as being the expected standard deviation of the size of the fragmented hybridized samples, ranging from 40 to 70 bases. Transcriptionally active regions (TARs) were determined by identifying windows where the signal exceeded a threshold determined by permutation of probe signal intensities. We considered 51-base (25 bases on each side of the tested position) windows with at least 3 probes (corresponding to the 18 nt average resolution of the tiling) above background. The median intensities across all probes above background in each window were computed. To determine thresholds, we generated 1,000 random datasets of 1,000 consecutive probes with a randomly attributed intensity, similarly considered 51-base windows and used the 99th and 95th percentiles of the generated median distributions. Thresholds corresponding to 1% or 5% false discovery rate (FDR) were used depending on purpose. Eight tiling paths were considered, including the "shared paths" on each strand, corresponding to probes mapping uniquely in all three haplotypes, and the "alternate paths" on each strand and also including probes specific to each haplotype. Overlapping TARs were merged into transcribed blocks.

Gene and exon level computations were performed on each of the haplotypic "alternate paths". To this end, we grouped probes into metaprobesets corresponding to Vega annotations per haplotypic path. At the exon level, they comprise all probes associated with each exon of each transcript. At the gene level, they include all unique probes corresponding to exons of each gene, such that if an exon is present in different transcripts of a given gene, the corresponding probes are only counted once. Probes corresponding to introns were ignored. Within a gene or exon, the hybridization to each probe can differ depending on the affinities and transcript diversity. To account for these variations, we computed the median rather than the mean intensities of all probes in the given metaprobeset. For each cell line, this was done using its corresponding "alternate path" metaprobeset. This provided the most complete and robust information on gene and exon expression in each cell line.

Splicing computation was performed on each of the "alternate paths" per gene. We first considered every exon with reference to their transcript and generated normalized intensities (NI) (defined as the log2 ratio of exon level intensities and gene intensities). A null value indicates the exon is not alternatively spliced in the gene. An exon can be represented with different names in different transcripts of a given gene, but the underlying probes are the same and therefore the NI values are identical. To better estimate the transcript abundance, we considered a second complementary level by computing junction normalized intensities (JNI). To this end, we generated junction metaprobesets per "alternate path" by grouping all junction probes targeting a common splice junction of a transcript. Note that if a splice junction is shared between several transcripts or even genes, a metaprobeset exists for each transcript. Junction probes were present in four technical replicates on the array. Their signal intensities were averaged to obtain a more accurate estimate of intensities. Junction intensities were next computed as the median of all probes in the junction metaprobeset and normalized against the gene level intensity.

1.2.B. Exon array processing. Exon array signals were processed using Affymetrix Power Tools (APT) and R.

Cross-platform correlation with the MHC array:

Probe level data for the entire exon array were first extracted using apt-cel-extract in Affymetrix Power Tools (apt-1.8.0), with a GC-based background correction. The 9 unstimulated samples were then normalized using vsn under the R/Bioconductor environment, so processing was the same as for MHC array data. A set of 10,572 probes from the Affymetrix Human Exon 1.0 ST array were included in the MHC array, each targeting a unique genomic location. An expression data matrix (10,572 x 9) for each platform was created and the correlation across all probes for the same sample on the two different platforms was measured (positive Pearson correlation test). Consistency between MHC and exon array data was also assessed by looking at differences between haplotypes. We computed fold changes between each pair of haplotypes for the 2,129 probes varying between haplotypes (defined as standard deviation across all samples > 0.5 on both platforms) and performed positive Pearson correlation tests, on all probes or after filtering out low intensity probes.

Gene, exon and splicing computation:

We used probeset definitions based on custom CDF libraries from the Microarray Lab (University of Michigan) (Dai et al. 2005) corresponding to the Vega and Ensembl gene and exon annotations (version 11 - November 12, 2008 which is based on the hg18 build). Probe level data from PGF, COX and QBL unstimulated samples were extracted, background corrected, normalized and summarized at the probeset level (exon or gene) using Robust Multichip Average (RMA) with the apt -probeset-summarize -a rma-sketch command and the appropriate CDF in APT (apt-1.8.0). Exon normalized intensities (NI) were computed for PGF by subtracting the log2 gene intensity from the log2 exon intensity (each first summarized as the median across the 3 replicates).

1.2.C. Quality controls.

We first assessed hybridization efficiency using spikes and polyA signals provided by Affymetrix.

The specificity of the strandedness was estimated by computing the ratio of the difference between the sense and antisense probeset signals to the sense probeset signal using 5 known expressed housekeeping genes (*ACTB*, *GAPDH*, *IRF9*, *RN28S1* and *RN18S1*). We provide the mean of all these ratios with its confidence interval at 95%.

The coverage of full length transcript was estimated by comparing signal intensities from the same housekeeping genes for the 5', middle located and 3' metaprobesets. We computed the coefficient of variation (standard deviation/mean).

Similarly, we computed the coefficient of variation of the 4 replicates of the junction probes spotted on the array. This was 0.0373 (95% confidence interval, 0.0362-0.0384).

We computed the Pearson correlation coefficient for all probes between each pair of samples and performed a hierarchical clustering on these coefficients (shown as a heatmap in Supplemental Fig. 3).

The sensitivity of junction probes was checked by comparing array data and quantitative real-time PCR data for the two main transcript isoforms of *CD79A* and *CD79B* genes.

The perfect match design specificity was estimated first by a global analysis comparing the signal of PGF ccDNA on PGF-, COX- and QBL-specific probes. The analysis was restricted to probes included in TARs to filter out background signal. Second, we carried a probe-wise comparison of the signal produced by the PGF samples on the 123 perfect match probes paired with probes carrying one mismatch corresponding to the COX path. Again, only probes included in TARs were considered. We used analysis of variance with repeated measures to account for the three biological replicates.

1.3. Statistical analyses

They were performed using R, Perl and PLINK.

1.3.A. TARs distribution:

Localization of TARs was studied with reference to the strand orientation. If a TAR was present on the sense path, it was expected to correspond to a transcript on the forward strand and vice versa. Hence, TARs tracks available at GEO (GSE22455) are given with respect to the transcript orientation for the strand column, although their names correspond to the path considered. We searched for overlap of at least one base inside gene boundaries obtained from Vega gene or pseudogene annotations. Percentage overlap, overlapped gene name and distance to neighboring genes relative to strand are provided in Supplemental Table 2. Intergenic TARs were defined by the absence of gene overlap on both strands. Median distance between Vega genes is 18.09 kb. Distal intergenic TARs were defined as being at least 10 kb away from the nearest gene. Their distribution was compared to that of genomic features extracted from UCSC hg18, including repeat elements from RepeatMasker (rmsk table in UCSC), CpG islands (cpgIslandEx table), open chromatin (EncodeDukeDNaseSeqPeaksGm12891V2 table where GM12891 carries *HLA-DR3/DR2* and is therefore heterozygous for COX/QBL and PGF alleles) and conservation (phastConsElements17way table). Colocalization was defined by an overlap of at least one base.

1.3.B. Differential expression:

Probe level comparisons were performed using the “shared path” probes only (269,678 of 381,916 tiling probes). Among the 230 genes and pseudogenes present in the MHC region, 206 are found on the three haplotype sequence annotations and were considered for gene level comparison using their respective “alternate path”. Similarly, after filtering for redundancy between transcripts, we considered the normalized values for 2,198 exons and 591 junctions present in all three sequence annotations, thus directly comparable using their respective “alternate path”.

All differential expression analyses were performed with the Bioconductor software package limma (Smyth 2004) which uses linear models and empirical Bayesian methods. We used a group-means parameterization with the model formula ~0+groups, where groups was a factor describing the 18 samples in terms of the 6 haplotype-condition combinations. Contrasts were used to extract the comparisons of interest and Benjamini-Hochberg's adjustment was used to control the false discovery rate. Adjusted p-values below 0.05 were considered significant. First, a brief assessment of overall stimulation, haplotype and interaction effects was performed on the processed probe-level data: of the 269,678 probes, 9,232 probes showed a significant response to stimulation in at least one haplotype; 16,250 probes differed significantly between haplotypes in either stimulated or unstimulated conditions, and just 24 probes showed a significant interaction effect (i.e. the response to stimulation differed between haplotypes). The detailed analyses of haplotypic differences at the gene level, exon level NI values and junction JNI values presented in the current manuscript are based on the unstimulated samples. The same approach was used for the gene level analyses using data from the Affymetrix Exon 1.0 ST array.

1.3.C. Correlation of differentially expressed (DE) probes with SNP distribution:

The lists of polymorphic SNPs between the three haplotypes or haplotype pairs were generated using hg18 snp129 table at UCSC. Altogether there were 23,556 polymorphic SNPs: 16,787 between PGF and COX, 15,993 between PGF and QBL, and 13,734 between COX and QBL. They were compared to the lists of differentially expressed probes obtained with limma using the "shared path". To this end, we took the MHC reference sequence and removed genomic segments corresponding to contig gaps in the QBL sequence. The MHC region was then segmented into two series of 10 kb windows shifted by 5 kb. The counts of DE probes per window were normalized by the number of probes designed in each window. Correlation between the normalized counts of DE probes and counts of polymorphic SNPs across windows was tested using the Pearson test (after log-transformation) and nonparametric Spearman tests. This was done either using all the windows ("_poly.all" sheet of Supplemental Table 5) or after filtering out windows with low SNP counts in the lower quartile of the SNP count distribution ("_poly.quartile" sheet).

To assess the relevance of the correlation tests, we restricted the comparison to the windows overlapping with genes (and to control windows lacking genes). As an additional control, we performed the correlation tests similarly with nonpolymorphic

SNPs ("_notpoly.*" sheets). Findings were corroborated using a robust scheme, by performing nonparametric Spearman tests using equidistant bins of size 5, 10 or 20 for the distribution of counts of DE probes (always normalized with the number of designed probes per window) and of polymorphic (or non polymorphic) SNPs ("_poly.bin" and "_notpoly.bin" sheets).

1.3.D. eQTL mapping for *ZFP57*, *HLA-DQA2* and *HLA-DQB2*:

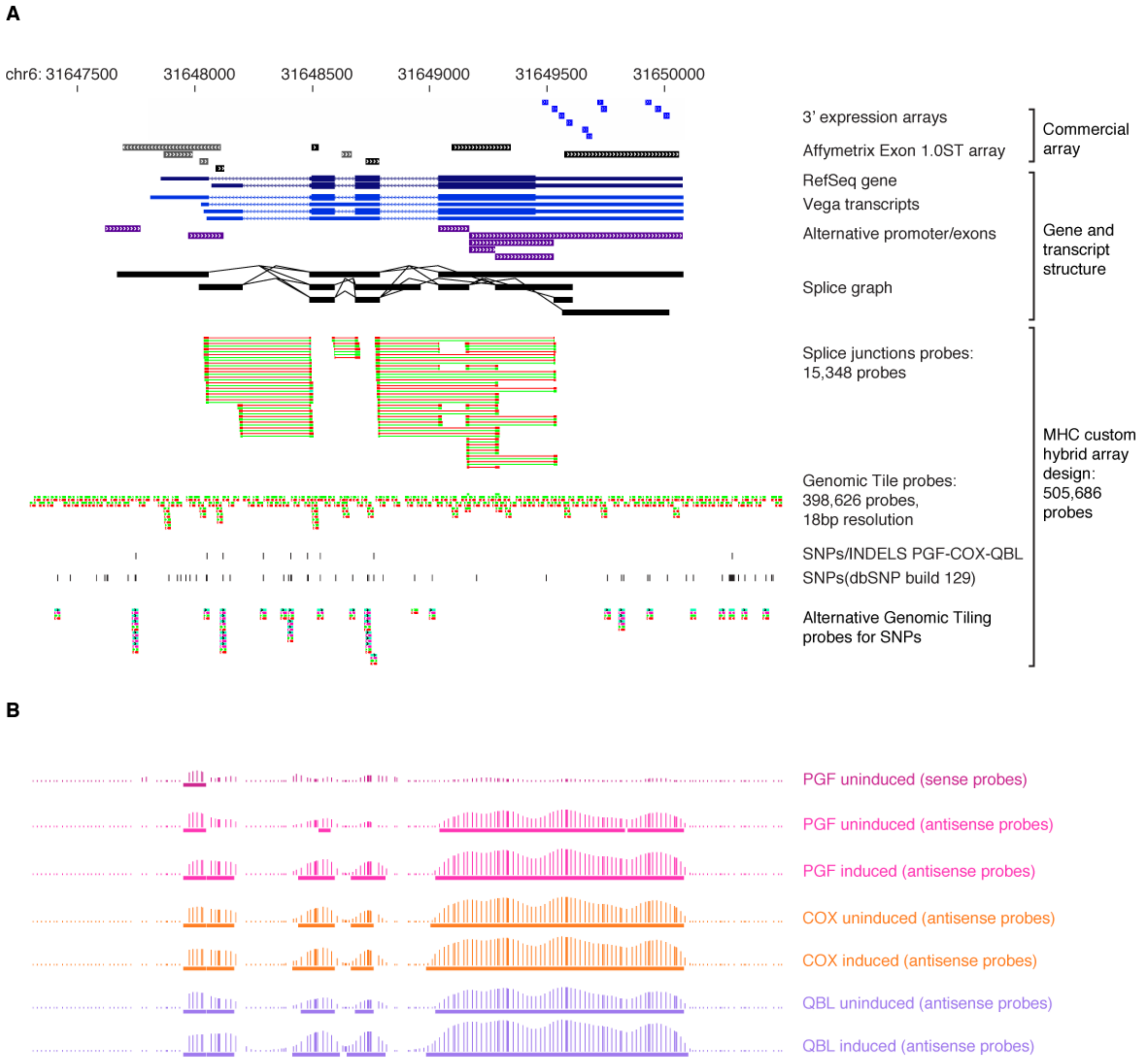
Standard QC measures and filtering on the volunteer data were done as described in (Fairfax et al. 2010). The quantitative trait association was conducted using PLINK with 4 maximum per-SNP missing genotypes (GENO 0.1) and MAF 0.03. For each SNP, PLINK generates a phenotypic mean for the three genotypic states and compares these means using the Wald test statistic to generate a P-value. The Wald test is useful especially in this instance, since it does not require that the data fit a normal distribution.

1.3.E. MHC splicing extent analysis in PGF sample:

Exon intensities normalized to the corresponding gene intensity (NIs) were computed as described earlier (see Splicing Computation). A negative NI value indicates the exon may be spliced out whereas a positive NI value indicates the exon may be included. We thus computed the percentage of exons with an absolute NI value different from zero exceeding various thresholds. This was done for both the 226 MHC genes annotated in Vega on the PGF sequence and the 28,454 non-MHC annotated Vega genes in the CDF files.

To test whether the observed difference between the two lists of genes was statistically significant, we used a permutation method to exclude the possibility of influence from any differences in gene structure. We analyzed the 131 MHC genes and the 15,659 non-MHC genes having at least 4 exons, a number that corresponds to the median number of exons per gene in the human genome. First, for the 131 MHC genes, we generated 1,000 samples of 4 exons per gene, drawn randomly with replacement. Using the NI of the selected MHC exons (n=131*4*1,000=524,000), we obtained the median and mean NI for MHC exons. For each sample, we also cumulatively enumerated the "spliced" exons, i.e. exons with an absolute NI above a given threshold (varying from 1 to 4). We then generated bootstrap distributions of the median and mean NI values and of the cumulative counts of spliced exons in non-MHC genes. To this end, we produced 10,000 sets of 131 genes randomly drawn from the 15,659 non-MHC genes having at least four exons. For each set, we generated 1,000 samples of 4 exons per gene and computed the median and mean NI and the total count of spliced exons for each threshold. We used these distributions to assess whether the NI values and the number of spliced exons were significantly different in MHC and non-MHC genes. We similarly compared non-MHC immune genes (733 genes with 4 exons) with non-MHC non-immune ones (sampling from 14,926 genes with 4 exons), and MHC genes with non-MHC immune genes. Immune-related genes were identified based on Gene Ontology classifications (GO:0002376). The comparison for the NI medians is shown graphically in Fig 4.C-D and for the number of spliced exons for different threshold in Supplemental Table 6. Repeating the whole analysis with Ensembl annotations yielded similar results.

2. SUPPLEMENTAL FIGURES AND TABLES


A
chr6: 31647500
31648000
31648500
31649000
31649500
31650000
3' expression arrays
Affymetrix Exon 1.0ST array
Commercial array
RefSeq gene
Vega transcripts
Alternative promoter/exons
Splice graph
Gene and transcript structure
Splice junctions probes: 15,348 probes
Genomic Tile probes: 398,626 probes, 18bp resolution
SNPs/INDELS PGF-COX-QBL
SNPs(dbSNP build 129)
Alternative Genomic Tiling probes for SNPs
MHC custom hybrid array design: 505,686 probes
B
PGF uninduced (sense probes)
PGF uninduced (antisense probes)
PGF induced (antisense probes)
COX uninduced (antisense probes)
COX induced (antisense probes)
QBL uninduced (antisense probes)
QBL induced (antisense probes)

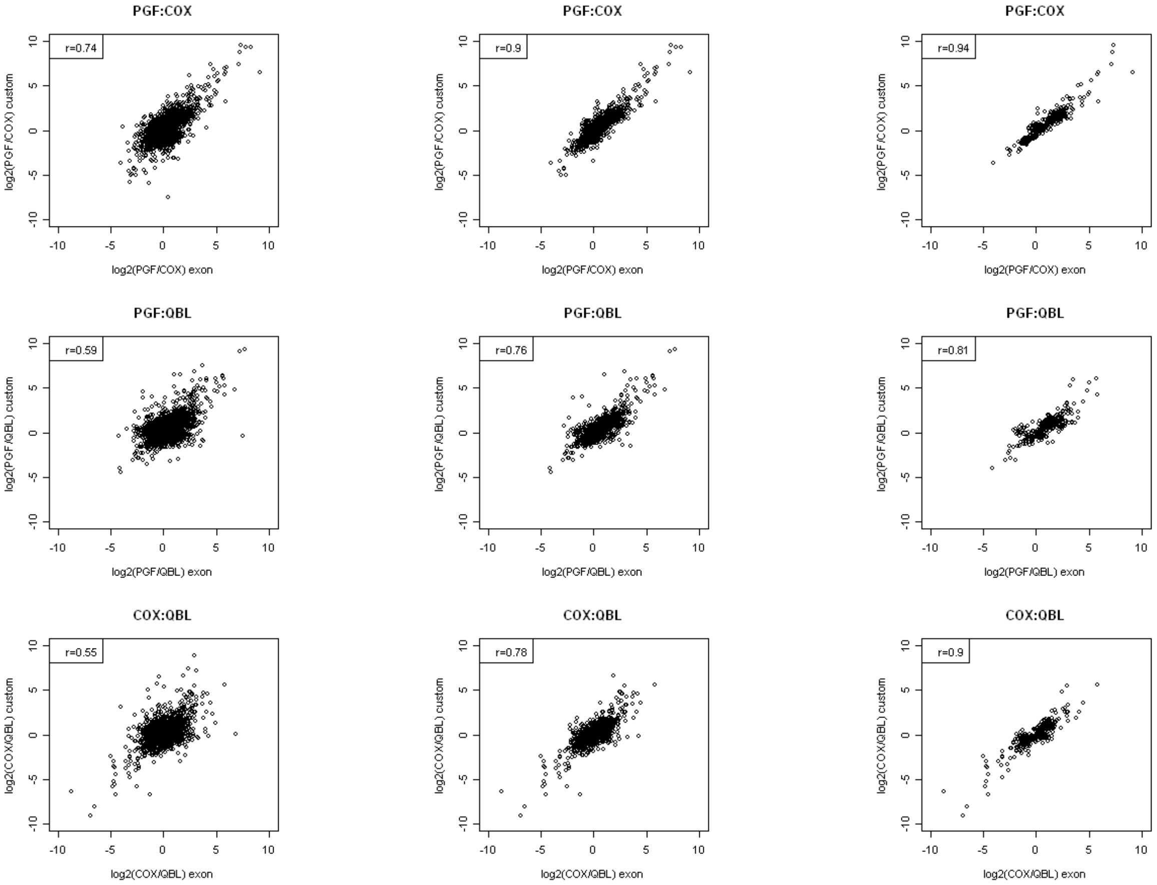
Supplemental Figure 1. Illustration of array design strategy and generated tracks for the *LTA* gene.

(A) To characterize the expression and splicing phenotypes of the MHC, as depicted here for the *LTA* gene, available commercial arrays with probes at the 3' end of transcripts were inappropriate. Recently, the Affymetrix GeneChip Human Exon 1.0 ST Array with probes covering the full length of transcripts was released, but its design does not account for the full complexity of gene expression, including alternative promoters or exons. Furthermore, commercial expression arrays are designed to the human reference assembly sequence and do not account for genetic diversity. Therefore, a custom array on the Affymetrix platform was conceived, with an original hybrid combination of two main probe sets. A set of 15,348 probes present in 4 copies aims at monitoring any possible known or predicted splice events. On average 6 probes were designed at the centre of each junction (range 2-42), corresponding to 1,043 junctions of 78 genes. For each single probe, its reverse complement was also designed (red versus green color). In addition, a tiling set of 398,626 overlapping probes covers both strands of the genomic MHC region with a final resolution of 18 bases, enabling accurate transcript profiling and discovery. Strand specificity can be determined as reverse-complement tiling probes were also designed. The design also takes into account genetic polymorphism as alternate probes were specifically designed for any SNPs or haplotypic segmental duplications in the region. The array also includes 26,484 probes for relevant non-MHC genes involved in alternative splicing or immune response. Finally, 19,184 control probes were included for signal processing (for calibration assessment, background correction) and for other specific applications of the array. The complete array comprises 505,686 probes. Criteria, including uniqueness against the genome and transcriptome, and structural conformation were carefully considered during the design process (cf. Material and Methods).

(B) Example of custom tracks for shared probes between the three haplotypes as they appear in the UCSC browser. For each sample, the smoothed intensity signal is displayed with the corresponding TARs at a FDR of 1% shown below. This matches the known exon structure of the *LTA* gene. As the *LTA* gene is transcribed on the forward strand, the signal is expected to be on antisense probes as observed for example in the PGF sample. For the three cell lines, antisense probes tracks are shown for either unstimulated or stimulated conditions, revealing a more pronounced induction of *LTA* gene expression for PGF and QBL than in COX whose unstimulated expression was already higher.

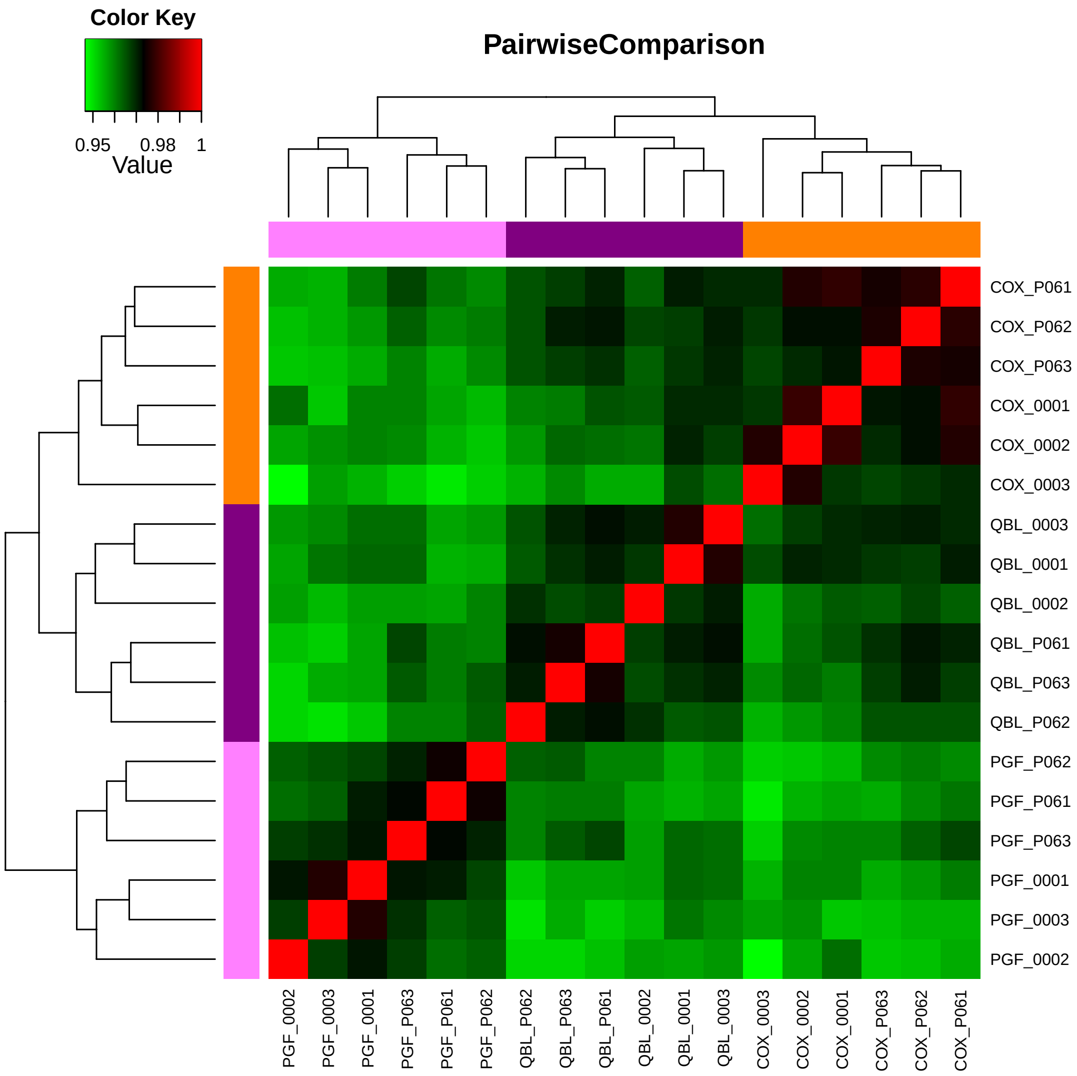
Supplemental Table 1. Correlation of intensity data for 10,572 probes shared between the MHC and the Affymetrix Exon 1.0 ST arrays.

Platform	**Comparison**	**Number of correlations**	**Median correlation coefficient**	**Minimum correlation coefficient**	**Maximum correlation coefficient**
MHC	Within haplotypes	9	0.98	0.97	0.98
Exon	Within haplotypes	9	0.92	0.90	0.94
MHC	Between haplotypes	27	0.95	0.93	0.96
Exon	Between haplotypes	27	0.89	0.84	0.91
Exon-MHC	Within haplotypes	27	0.88	0.83	0.91
Exon-MHC	Between haplotypes	27	0.86	0.80	0.89
Exon-MHC	Same Sample	9	0.88	0.83	0.91



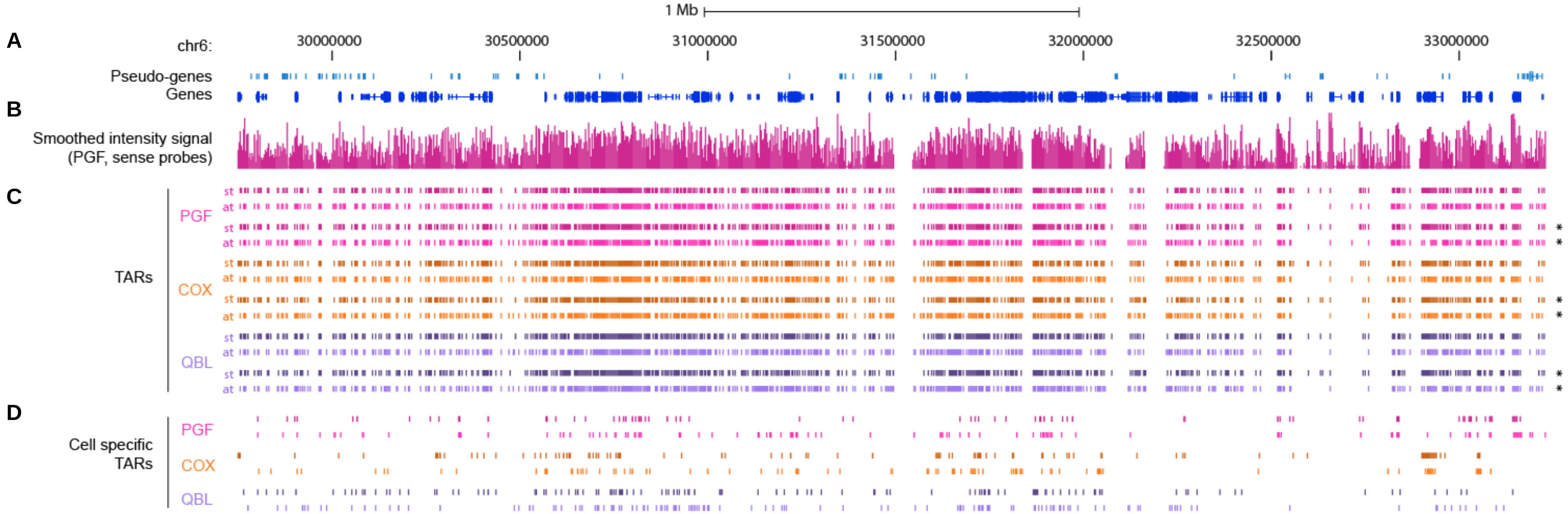
Supplemental Figure 2. Correlation of the haplotypic differences between the MHC array and the Affymetrix Exon 1.0 ST Array.

The fold changes between PGF-COX (top), PGF-QBL (middle) and COX-QBL (bottom) was computed using Exon 1.0 ST array data (x axis) and plotted against the same fold changes computed using MHC array data (y axis). Left column includes data for the 2,129 most varying probes (standard deviation greater 0.5 in both Exon 1.0 ST and MHC array datasets); middle column for 924 probes additionally having mean intensity >6 on Exon array; right column for 324 probes additionally having mean intensity >8 on the Exon array. The strong positive Pearson correlation indicates that similar results are obtained on both platforms, and increases when low intensity probes are removed.



Supplemental Figure 3. Heatmap of between-array pairwise Pearson correlation coefficients.

The color key for the correlation coefficient is given on the top left-hand corner. The dendograms illustrate the relationship between samples. The bars below the dendograms are colored in orange for COX samples, in purple for QBL samples and in pink for PGF samples. All biological replicates from each cell line are found clustered together. The names of the samples are displayed on each row and column. The suffixes stand for the culture condition (000=unstimulated; P06=stimulated), followed by one digit for the biological replicate number.



Supplemental Figure 4. Transcriptionally active regions across the MHC on the “shared path” relative to the human reference sequence.

(A) Vega genes (dark blue) and pseudo-genes (light blue). **(B)** Smoothed intensity signals from probe hybridization with single-strand ccDNA of PGF unstimulated cells on the forward strand. **(C)** Transcriptionally active regions (TAR) in unstimulated or stimulated (asterisk on the right-hand side) PGF (pink), COX (orange), and QBL (purple) in antisense (dark hue) or sense (light hue) orientation. TARs are defined by the presence of at least 3 probes above background per 51-base window whose median intensity is above thresholds corresponding to a false discovery rate of 1%. **(D)** Cell-specific TARs found only in one cell line (color code as in c). Number of cell-specific TARs for PGF, COX and QBL unstimulated cells are respectively: 419, 519 and 289 in antisense orientation, and 661, 544 and 270 in sense orientation.

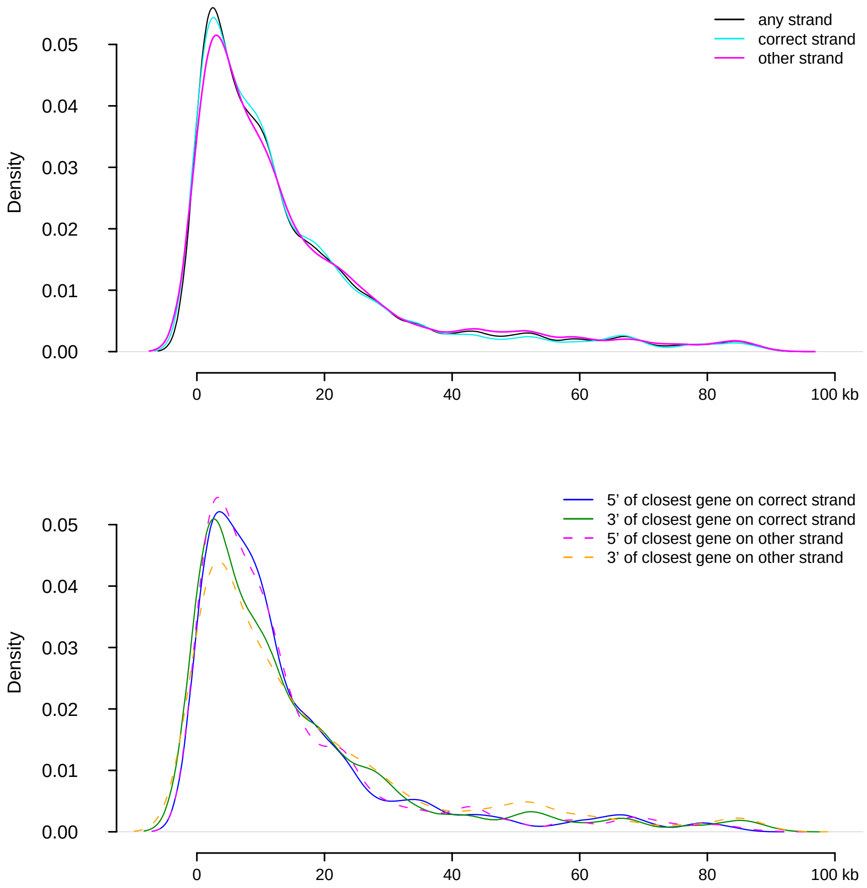
Supplemental Table 2. List of transcribed blocks on each haplotypic path

This supplemental element made of an excel workbook of 7 spreadsheets is provided to the manuscript as an independent zip file.

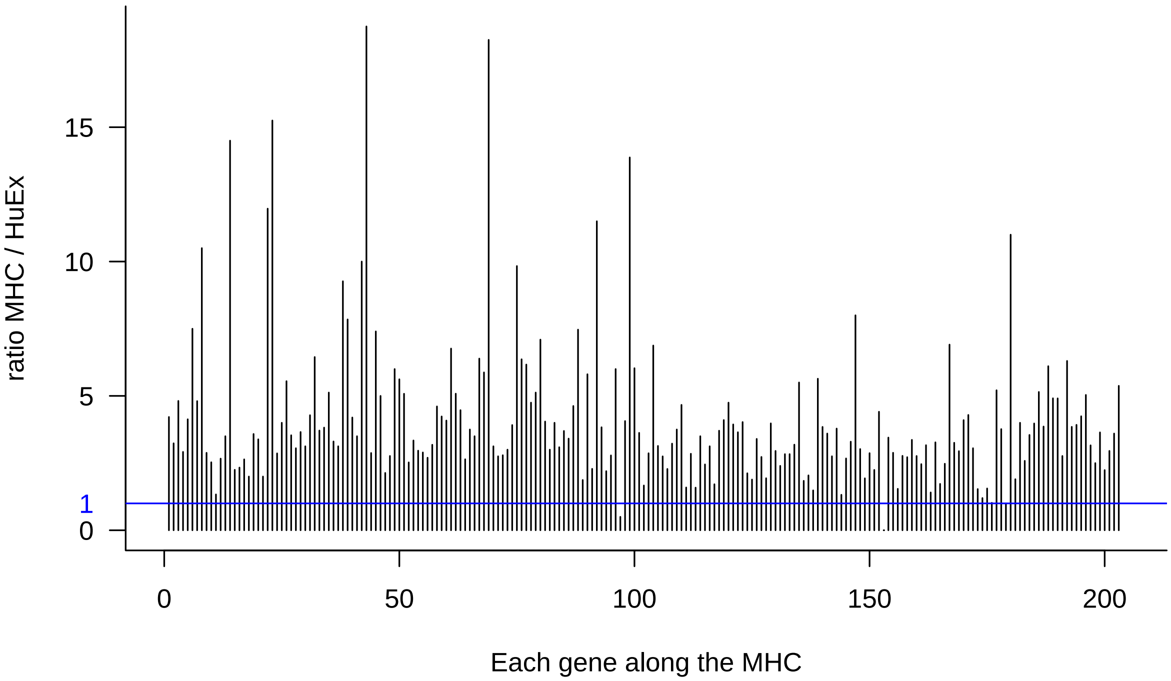
Supplemental Table 3. Statistics of gene annotations including at least one TAR at a FDR of 5%.

	PGF		**COX**		**QBL**	
	transcribed	untranscribed	transcribed	untranscribed	transcribed	untranscribed
Total genes and pseudogenes	197	29	192	36	187	27
Total genes	150	9	146	11	141	9
KNOWN_processed_transcript	11	0	11	0	10	0
KNOWN_protein_coding	122	4	118	6	117	3
NOVEL_processed_transcript	7	2	7	3	6	0
NOVEL_protein_coding	1	0	1	0	1	2
PUTATIVE_processed_transcript	9	3	9	2	7	4
Total pseudogenes	47	20	46	18	46	18
processed_pseudogene	17	8	17	8	17	8
transcribed_processed_pseudogene	0	0	1	0	1	0
transcribed_unprocessed_pseudogene	5	1	3	0	3	0
unprocessed_pseudogene	24	11	24	10	24	10
KNOWN_polymorphic_pseudogene	1	0	1	0	1	0

Note:
The selection for TARs using a FDR of 5% rather than 1% was adopted here as we were not aiming to discover new transcripts but to define expression of known and predicted transcripts, which allowed us to be less stringent on threshold inclusion.



Supplemental Figure 5. Distance of intergenic TARs from closest gene.



Supplemental Figure 6. Increased number of probes per gene in the MHC array compared with the Human Exon 1.0 ST array.

For each of the 204 genes of the MHC (x-axis, from telomere to centromere) present in the CDF libraries from the Microarray Lab (University of Michigan), we computed the ratio of the number of probes in our MHC array versus that in the Exon 1.0 ST array. The horizontal blue line indicates an equal number of probes.

Supplemental Table 4. Top 30 MHC genes showing significant differential expression between haplotypes after Benjamini-Hochberg adjustment using the Affymetrix Exon 1.0 ST array.

Pseudogenes are indicated by an asterisk.

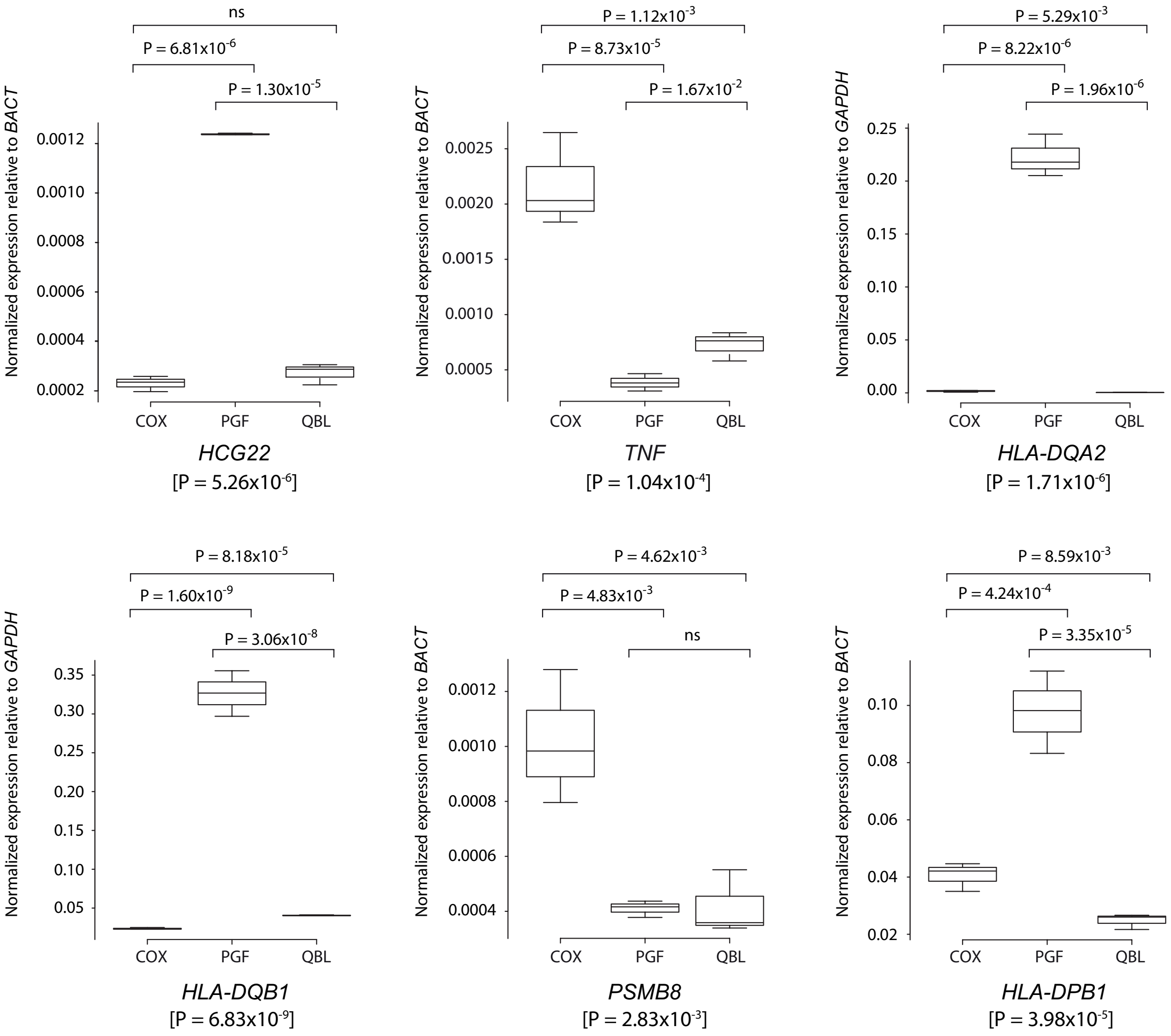
rank at genome scale [a]	**rank with MHC array** [b]	**Gene Name** [c]	**Class** [c]	**Log2 (Fold Change)** [c]			**Adj.P.Val** [c]
				COX vs PGF	**QBL vs PGF**	**QBL vs COX**	
13	NA	*HLA-DRB5*	II	-3.62	-3.19	0.43	3.55x10^{-8}
23	5	*HLA-U* *	I	-3.53	-0.91	2.61	7.81x10^{-8}
104	7	*HLA-DPB1*	II	-1.56	-0.88	0.68	1.30x10^{-6}
105	8	*RPL32P1* *	II	-1.64	-1.05	0.59	1.30x10^{-6}
144	4	*HLA-DQB2*	II	-1.77	-1.87	-0.10	3.91x10^{-6}
235	13	*HCG22*	I	-1.59	-1.59	0.00	1.53x10^{-5}
263	NA	*HLA-DOB*	II	-2.04	-1.16	0.88	1.82x10^{-5}
286	18	*HLA-DOA*	II	-1.26	-1.13	0.13	1.82x10^{-5}
331	54	*HLA-DRA*	II	-1.00	-1.15	0.15	2.45x10^{-5}
387	10	*HLA-A*	I	-0.32	-0.82	-0.50	2.45x10^{-5}
354	2	*HLA-DPB2* *	II	-1.00	-0.94	0.06	2.57x10^{-5}
372	40	*HLA-DRB1*	II	-1.81	-0.94	0.06	3.59x10^{-5}
471	25	*HLA-DMA*	II	-0.62	-0.83	-0.22	4.90x10^{-5}
589	80	*HLA-DMB*	II	-0.94	-0.37	0.57	1.15x10^{-4}
561	1	*ZFP57*	I	1.75	0.17	-1.58	1.15x10^{-4}
639	62	*IER3*	I	-0.75	-0.98	-0.23	1.15x10^{-4}
730	11	*HLA-L* *	I	-0.87	-1.14	-0.27	1.16x10^{-4}
923	29	*PSMB9*	II	0.47	-0.18	-0.65	3.80x10^{-4}
1094	28	*HLA-C*	I	0.02	-0.78	-0.79	7.45x10^{-4}
1185	55	*HLA-DPA1*	II	-0.57	-0.52	0.05	7.77x10^{-4}
1090	6	*TNF*	III	1.47	0.41	-1.06	7.77x10^{-4}
1149	3	*HLA-DQA2*	II	-1.06	-0.92	0.14	8.68x10^{-4}
1294	17	*HLA-F*	I	-0.08	-0.70	-0.62	1.06x10^{-4}
1346	NA	*FLOT1*	I	-0.47	-0.64	-0.17	1.06x10^{-4}
1254	15	*LTA*	III	0.92	0.05	-0.87	1.06x10^{-4}
1332	24	*CLIC1*	III	0.56	-0.10	-0.66	1.06x10^{-4}
1277	39	*AIF1*	III	-1.21	0.56	0.65	1.10x10^{-4}
1677	129 ns	*GPSM3*	III	0.02	-0.42	-0.44	1.17x10^{-4}
1526	42	*HLA-DQB1*	II	-0.98	-1.18	-0.20	1.18x10^{-4}
1617	19	*TAP1*	II	0.53	-0.38	-0.90	1.19x10^{-4}

Notes:

[a] rank obtained after comparing differential expression between unstimulated cell lines at the genome scale

[b] For comparison, this is the rank obtained after comparing differential expression between unstimulated samples hybridized on the MHC array. Twelve genes that had been found with a significant differential expression using the MHC array are not among the top 30 genes using the Exon array but were found as follows: *HLA-B* (43rd with Adj.P = 0.015); *XXbac-BPG254F23.6* (61st, with Adj.P = 0.038); *XXbac-BPG254F23.5* (No metaprobesets available); *NCR3* (not significant (n.s)); *LTB* (n.s); *LST1* (n.s); *DAQB-335A13.8* (n.s); *TCF19* (39th, Adj.P = 0.009); *BRD2* (48th, Adj.P = 0.017) ; *NRM* (64th, Adj.P = 0.044); *HCG27* (n.s). NA: not applicable, concerns 3 genes that are not present on the three annotated haplotype sequences, and thus that were not considered when running the comparison between haplotypes with the MHC array.

[c] results obtained comparing differential expression between unstimulated cell lines on the extracted MHC genes only.



HCG22
[P = 5.26x10^{-6}]

TNF
[P = 1.04x10^{-4}]

HLA-DQA2
[P = 1.71x10^{-6}]

HLA-DQB1
[P = 6.83x10^{-9}]

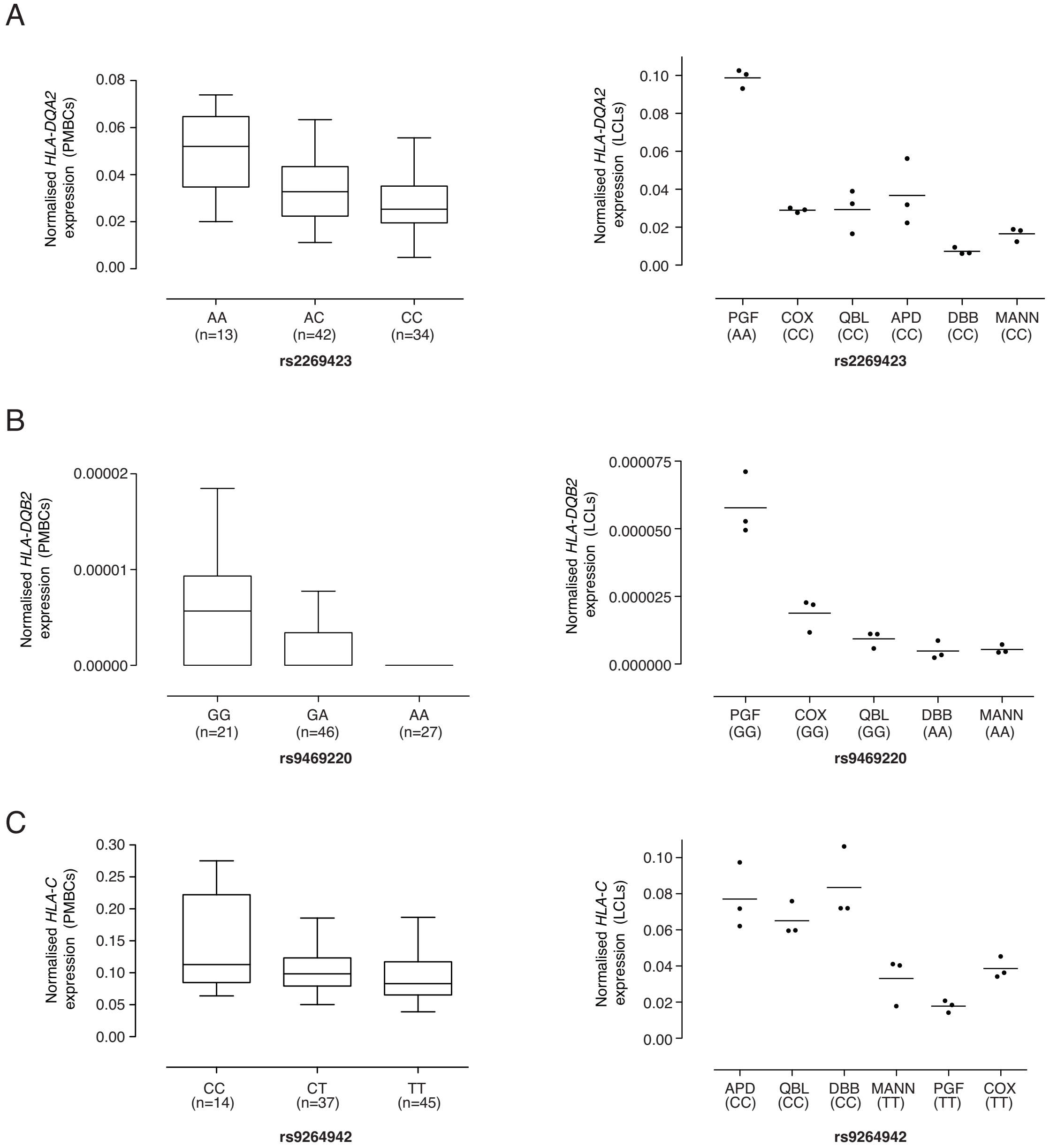
PSMB8
[P = 2.83x10^{-3}]

HLA-DPB1
[P = 3.98x10^{-5}]

Supplemental Figure 7. Validation of differential expression between COX, PGF and QBL samples. Three other genes were also tested, *NCR3*, *CLIC1* and *TCF19* but although there was a similar pattern of haplotypic expression as in the array, this was not significant.

Supplemental Table 5. Correlation between distribution of differentially expressed probes and polymorphic SNPs

This supplemental element made of an excel workbook of 9 spreadsheets is provided to the manuscript as an independent zip file.



Supplemental Figure 8. Validation data for differentially expressed genes.

(A) Expression of *HLA-DQA2* (relative to *ACTB*) determined by quantitative real time RT-PCR in peripheral blood mononuclear cells of 89 healthy volunteers plotted by genotype for rs2269423 (Kruskal-Wallis test on genotypes P=1.5x10-3) and for MHC-homozygous lymphoblastoid cell lines. **(B)** Expression of *HLA-DQB2* (relative to *ACTB*) determined by quantitative real time RT-PCR in peripheral blood mononuclear cells of 94 healthy volunteers plotted by genotype for rs9469220 in 94 healthy volunteers (Kruskal-Wallis test on genotypes P=7x10^{-4}) and for MHC-homozygous lymphoblastoid cell lines. **(C)** Expression of *HLA-C* (relative to *GAPDH*) determined by quantitative real time RT-PCR in peripheral blood mononuclear cells of 96 healthy volunteers plotted by genotype for rs9264942 (Kruskal-Wallis test on genotypes P=0.034) and for MHC-homozygous lymphoblastoid cell lines. Among the lymphoblastoid cell lines, genotypes that were unknown for rs2269423 (DBB cell line) and for rs9264942 (APD) were determined by direct sequencing.

Supplemental Table 6. Permutation results to evaluate the extent of alternative splicing between the MHC genes and non MHC genes.

For the mean and median SI values, we counted the number of sets with a higher value than in the tested set. For the different NI thresholds we counted how many of the 10,000 compared sets had more exons above the given threshold than in the tested set.

		MHC versus non-MHC			Non-MHC Immune versus Non-MHC Non-Immune			MHC versus non-MHC immune		
		Value for MHC set	Number of sets with a value > MHC set	*P*-value	Value for Immune set	Number of sets with a value > Immune set	*P*-value	Value for MHC set	Number of sets with a value > MHC set	*P*-value
Vega-based probesets	Mean NI	1.13	1	0.0001	1.00	4753	0.4753	1.13	0	<0.0001
	Median NI	0.88	3	0.0003	0.76	3148	0.3148	0.88	0	<0.0001
	NI >1	226619	40	0.0040	1131879	4567	0.4576	226736	28	0.0028
	NI>1.5	135367	78	0.0078	646512	5551	0.5551	135789	40	0.0040
	NI>2	76672	306	0.0306	360474	6098	0.6098	76958	188	0.0188
	NI>3	32409	8	0.0008	111275	5789	0.5789	32647	0	<0.0001
	NI>4	12188	43	0.0043	35556	3886	0.3886	12263	3	0.0003
Ensembl-based probesets	Mean NI	1.08	0	<0.0001	0.95	2877	0.29	1.08	0	<0.0001
	Median NI	0.79	228	0.0228	0.72	2711	0.27	0.79	164	0.0164
	NI >1	159168	86	0.0086	603940	1970	0.20	159316	137	0.0137
	NI>2	54684	52	0.0052	185148	1502	0.15	54735	68	0.0068
	NI>3	22443	0	<0.0001	45523	8888	0.89	22509	0	<0.0001
	NI>4	10421	0	<0.0001	14986	5794	0.58	10583	0	<0.0001

Supplemental Table 7. Variation of splicing between haplotypes.

Top 30 exons showing significant differential expression between haplotypes after Benjamini-Hochberg adjustment are listed. For each biological replicate of each cell line, the exon level intensity was normalized against the gene level intensity to generate the normalized intensity (NI). Intensities are expressed using the log2 scale. Gene and exon levels of genes shared by the three haplotypes were computed from the signal intensity of the probes matching uniquely and perfectly to their haplotype sequence. Pseudogenes are indicated by an asterisk.

Gene Name	**Exon**	**Class**	**Gene Intensity**			**Exon Intensity**			**NI**			**Adj.P.Val**
	(Transcript.Exon ID)		**PGF**	**COX**	**QBL**	**PGF**	**COX**	**QBL**	**PGF**	**COX**	**QBL**	
HLA-C	012.2	I	13.15	13.20	14.26	14.35	14.69	9.84	1.20	1.49	-4.42	7.76x10^{-17}
HLA-G	004.4/001.5/006.3/002.6/005.4	I	10.21	9.83	9.81	14.63	10.19	8.46	4.42	0.36	-1.35	5.70x10^{-16}
*HLA-T**	001.2	I	8.43	7.85	7.95	9.49	15.03	11.29	1.06	7.18	3.34	2.77x10^{-14}
*HLA-DPB2**	001.4	II	11.02	7.83	8.00	12.49	14.55	14.88	1.47	6.72	6.88	2.26x10^{-13}
HLA-DQA2	001.4	I	9.81	7.36	8.19	13.97	15.27	15.54	4.16	7.91	7.35	3.14x10^{-13}
HLA-DPA1	001.5	II	11.03	10.42	10.53	15.98	10.54	15.55	4.95	0.12	5.02	4.74x10^{-13}
CYP21A2	008.1/001.1/009.1	III	8.31	8.10	8.21	12.60	6.77	9.76	4.29	-1.33	1.55	7.04x10^{-13}
*HLA-P**	001.3	I	7.35	7.26	7.36	14.21	10.16	10.15	6.86	2.90	2.79	4.41x10^{-12}
HLA-G	001.4/004.3/005.3/002.5	I	10.21	9.83	9.81	10.74	13.68	8.88	0.53	3.85	-0.93	4.41x10^{-12}
LSM2	004.1	III	11.45	11.47	11.45	11.11	6.71	11.23	-0.34	-4.76	-0.22	5.34x10^{-12}
HLA-B	004.4	I	13.50	13.44	12.31	10.57	10.09	12.88	-2.93	-3.35	0.57	5.93x10^{-12}
*HLA-DPB2**	001.3	II	11.02	7.83	8.00	14.43	13.76	14.03	3.41	5.93	6.03	1.06x10^{-11}
HLA-C	001.1	I	13.15	13.20	14.26	12.66	12.86	9.67	-0.49	-0.34	-4.59	1.25x10^{-11}
HLA-DQB2	004.4	II	12.12	9.38	9.54	14.65	7.77	10.26	2.53	-1.61	0.72	1.87x10^{-11}
EHMT2	003.19	III	9.86	10.00	9.96	11.33	11.57	9.12	1.47	1.57	-0.84	8.64x10^{-11}
*DDX39BP1**	001.2	I	6.88	7.38	7.27	10.71	14.68	14.68	3.83	7.30	7.41	1.34x10^{-10}
HLA-A	001.2	I	12.75	11.24	10.89	6.82	7.63	8.81	-5.93	-3.61	-2.08	1.91x10^{-10}
HLA-C	005.1	I	13.15	13.20	14.26	12.96	12.94	10.88	-0.19	-0.26	-3.38	2.16x10^{-10}
HLA-DQA2	001.3	II	9.81	7.36	8.19	15.24	14.84	15.20	5.43	7.48	7.01	7.88x10^{-10}
HLA-C	007.2	I	13.15	13.20	14.26	14.46	15.36	14.19	1.31	2.16	-0.07	8.83x10^{-10}
HLA-C	010.2	I	13.15	13.20	14.26	9.96	11.68	9.96	-3.19	-1.52	-4.30	1.40x10^{-09}
HLA-A	005.2/006.2/007.2	I	12.75	11.24	10.89	14.50	11.35	10.59	1.75	0.11	-0.30	1.56x10^{-09}
EHMT2	003.15	III	9.86	10.00	9.96	9.73	10.25	8.31	-0.13	0.25	-1.65	1.82x10^{-09}
EHMT2	008.23	III	9.86	10.00	9.96	12.11	12.25	10.96	2.25	2.25	1.00	2.02x10^{-09}
EHMT2	009.14	III	9.86	10.00	9.96	9.98	10.31	8.31	0.12	0.31	-1.65	2.09x10^{-09}
HLA-B	008.1	I	13.50	13.44	12.31	11.51	11.65	11.96	-1.99	-1.79	-0.35	2.18x10^{-09}
HLA-B	007.1	I	13.50	13.44	12.31	11.29	11.40	11.52	-2.21	-2.04	-0.79	5.02x10^{-09}
*HLA-DPA3**	001.1	II	8.11	7.72	8.18	9.20	11.28	9.54	1.09	3.56	1.36	7.60x10^{-09}
ZFP57	001.2	I	7.59	10.36	7.60	7.39	8.99	7.52	-0.20	-1.37	-0.08	8.88x10^{-09}
HLA-G	004.1/005.1/003.1/006.1	I	10.21	9.83	9.81	12.04	13.48	14.30	1.83	3.65	4.49	1.13x10^{-08}

Supplemental Table 8. Variation of junction intensities between haplotypes.

Genes showing a differential NI and having at least one junction showing differential expression between haplotypes are listed. The other significant junctions of these genes are also displayed, several being mutually exclusive (same acceptor or donor site, but not both). For each biological replicate of each cell line, the junction level intensity was normalized against the gene level intensity to generate the junction normalized gene intensity (JNI). Gene and junction levels of genes shared by the three haplotypes were computed from the signal intensity of the probes matching uniquely and perfectly to their haplotype sequence. Names of 5' donor exons and 3' acceptor exons are listed with reference to their transcript name (Transcript.ExonID). P values after Benjamini Hochberg adjustment are displayed. Overall, 31 genes had a junction showing differential expression between haplotypes, 27 of them also showed a differential exon NI. These are: *LST1*, *PPT2*, *AIF1*, *LTA*, *DDX39B*, *CLIC1*, *SLC44A4*, *XXbac-BPG296P20.15*, *EHMT2*, *C6orf25*, *DDAH2*, *CSNK2B*, *C6orf48*, *BAG6*, *PRRT1*, *DAQB-331I12.5*, *NOTCH4*, *ABHD16A*, *APOM*, *STK19*, *LSM2*, *TNF*, *PRRC2A*, *SKIV2L*, *EGFL8*, *NCR3* and *NFKBIL1* (ordered by the minimal Adj.P. Value).

Gene	**Donor site**	**Acceptor site**	**5' exon^3' exon**	**JNI**			**Log2 Fold Change**			**Adj.P.Val**
				PGF	**COX**	**QBL**	**COX vs PGF**	**QBL vs PGF**	**QBL vs COX**	
LST1	31663074	31663699	017.1^2, 011.2^3, 020.2^3, 004.2^3, 006.1^2	-1.91	-1.85	-0.50	0.06	1.41	1.34	1.83x10^{-8}
	31663074	31663396	013.2^3, 022.1^2, 001.2^3, 002.2^3, 018.2^3, 016.2^3, 005.2^3, 019.1^2, 023.1^2	-0.67	-1.36	-2.19	-0.69	-1.52	-0.83	1.83x10^{-6}
	31662872	31662955	016.1^2, 014.1^2, 020.1^2, 005.1^2	0.30	-0.26	0.72	-0.56	0.42	0.98	5.16x10^{-6}
	31663722	31664252	022.3^4, 006.2^3	0.72	0.82	0.06	0.10	-0.66	-0.76	1.41x10^{-5}
	31663074	31664273	014.2^3, 003.2^3, 015.1^2, 012.2^3	-0.01	0.10	0.76	0.11	0.77	0.66	1.98x10^{-5}
	31662862	31662955	018.1^2	-0.39	-0.60	0.18	-0.21	0.56	0.77	1.32x10^{-4}
	31663722	31664318	004.3^4, 019.3^4	2.74	3.15	2.41	0.41	-0.33	-0.74	6.04x10^{-4}
	31663074	31664318	008.2^3	-0.11	-0.13	0.66	-0.02	0.77	0.79	6.11x10^{-4}
	31663489	31664252	002.3^4	-2.17	-1.96	-2.53	0.21	-0.36	-0.57	1.46x10^{-3}
	31663489	31663699	022.2^3, 019.2^ 3, 016.3^4, 001.3^4	-0.43	-1.06	-1.29	-0.63	-0.86	-0.23	0.002

	31663489	31664273	005.3^4	-0.80	-0.55	-1.13	0.25	-0.33	-0.58	0.012
	31662071	31662955	011.1^2, 004.1^2, 012.1^2, 001.1^2, 010.1^2, 009.1^2, 008.1^2	-1.53	-1.37	-1.88	0.16	-0.35	-0.51	0.013
	31663722	31664273	011.3^4, 16.4^5, 021.2^3, 017.2^3, 001.4^5, 020.3^4	1.26	0.52	0.41	-0.74	-0.85	-0.11	0.015
	31663489	31663571	018.3^4	-1.36	-1.57	-1.75	-0.21	-0.39	-0.18	0.016
	31663489	31664318	023.2^3	0.16	-0.14	-0.40	-0.30	-0.56	-0.26	0.019
	31663078	31663306	010.2^3	-2.21	-2.13	-2.61	0.08	-0.40	-0.48	0.025
PPT2	32231538	32231625	001.4^5, 005.4^5, 010.4^5, 009.3^4, 003.4^5, 002.5^6, 007.5^6, 008.3^4, 006.4^5, 004.4^5, 011.3^4	1.06	1.32	-0.10	0.26	-1.16	-1.42	1.25x10^{-6}
	32233679	32238322	014.2^3, 004.7^8, 001.7^8, 003.7^8, 005.7^8, 006.7^8, 002.8^9, 016.2^3	1.17	1.34	0.45	0.17	-0.72	-0.89	5.16x10^{-6}
	32238377	32238561	004.8^9, 003.8^9, 001.8^9, 002.9^10	-0.35	0.10	-0.73	0.45	-0.38	-0.83	9.28x10^{-5}
	32231733	32233391	005.5^6, 001.5^6, 011.4^5, 010.5^6, 009.4^5, 003.5^6, 002.6^7, 008.4^5, 006.5^6, 004.5^6	0.53	0.69	0.07	0.16	-0.46	-0.62	0.0016
	32229391	32229753	002.1^2	-2.05	-2.14	-1.64	-0.09	0.41	0.50	0.0020
	32233475	32233594	014.1^2, 005.6^7, 001.6^7, 003.6^7, 002.7^8, 009.5^6, 006.6^7, 015.1^2, 016.1^2, 008.5^6, 004.6^7	-0.97	-0.70	-1.31	0.27	-0.34	-0.61	0.0025
	32229391	32230341	008.1^2, 001.1^2	-1.44	-1.79	-1.04	-0.35	0.40	0.75	0.0029
	32229391	32229950	007.1^2	-2.04	-2.10	-1.65	-0.06	0.39	0.45	0.029

	32230001	32230341	004.1^2	2.43	2.46	2.95	0.03	0.52	0.49	0.0047
	32247312	32247477	006.20^21	1.90	0.26	0.68	-1.64	-1.22	0.42	0.0061
	32245760	32246169	006.19^20	-1.91	-1.99	-1.55	-0.08	0.36	0.44	0.066
	32245256	32245662	006.18^19	-1.82	-2.13	-1.76	-0.31	0.06	0.37	0.021
AIF1	31692262	31692571	001.5^6, 004.5^6, 005.2^3, 006.3^4, 002.2^3	1.93	0.51	1.80	-1.42	-0.13	1.29	2.89x10^{-6}
	31691901	31692099	004.4^5, 001.4^5, 002.1^2	2.64	1.35	2.31	-1.29	-0.33	0.96	6.37x10^{-4}
	31691492	31692099	005.1^2	-2.51	-1.80	-2.27	0.71	0.24	-0.47	0.0021
	31691049	31691276	004.1^2, 003.1^2	-2.96	-2.33	-2.82	0.63	0.14	-0.49	0.0028
LTA	31648058	31648489	002.1^2, 003.1^2	-0.82	-2.22	-1.54	-1.40	-0.72	-0.68	1.14x10^{-5}
DDX39B	31617389	31616290	016.1^2	-2.97	-3.10	-1.84	-0.13	1.13	1.26	1.45x10^{-5}
	31616908	31616420	012.2^3	-3.62	-4.43	-3.75	-0.81	-0.13	0.68	2.44x10^{-5}
	31617151	31616420	018.1^2	-3.54	-4.16	-3.45	-0.62	0.09	0.71	5.28x10^{-4}
	31614902	31614695	013.3^4, 010.3^4	-1.00	-1.72	-1.72	-0.72	0.00	0.00	7.78x10^{-4}
	31617705	31616420	010.1^2, 003.3^4, 019.1^2, 001.1^2	0.85	-0.08	0.37	-0.93	-0.48	0.45	0.0012
	31606534	31606206	004.9^10, 001.10^11, 022.5^6, 002.10^11, 006.9^10, 003.12^13, 007.3^4	3.84	3.24	3.79	-0.60	-0.05	0.55	0.0013
	31617588	31616420	014.1^2	-3.16	-3.96	-3.45	-0.80	-0.28	0.52	0.0020
	31606808	31606705	006.8^9	-3.44	-4.20	-3.53	-0.76	-0.09	0.67	0.0022
	31617151	31616290	017.1^2	-3.52	-4.06	-3.40	-0.54	0.12	0.66	0.0025
	31621864	31617904	003.2^3	-1.07	-1.64	-1.08	-0.57	-0.01	0.56	0.011
	31610241	31608667	020.1^2	-2.87	-3.27	-2.80	-0.40	0.07	0.47	0.025
CLIC1	31812106	31815019	003.2^003.3	-4.41	-5.68	-5.11	-1.27	-0.70	0.57	2.39x10^{-5}
	31812106	31812551	002.1^002.2	-4.25	-5.73	-5.18	-1.48	-0.93	0.55	3.32x10^{-5}
	31809717	31809909	004.3^004.2, 003.4^003.5, 002.3^002.4, 001.2^001.3	3.23	2.44	2.78	-0.78	-0.45	0.33	1.91x10^{-4}
	31808154	31809321	004.5^004.6, 001.4^001.5, 003.6^003.7, 002.5^002.6	1.85	1.34	1.44	-0.51	-0.41	0.11	1.72x10^{-3}

	31806759	31807972	004.6^004.7, 002.6^002.7, 001.5^001.6, 003.7^003.8	1.80	1.10	1.48	-0.70	-0.32	0.38	8.88x10^{-3}
SLC44A4	31947317	31950215	004.7^004.8, 001.7^001.8	1.81	2.32	2.84	0.51	1.03	0.52	2.44x10^{-5}
	31945393	31946567	006.1^006.2	0.49	1.04	0.70	0.55	0.21	-0.34	0.030
	31950602	31950684	004.5^004.6, 001.5^001.6, 002.5^002.6, 003.5^003.6	1.48	1.95	1.30	0.47	-0.18	-0.65	0.04
XXbac-BPG296P 20.15	31618152	31618464	002.1^002.2	0.46	0.65	-0.57	0.18	-1.03	-1.22	3.72x10^{-5}
EHMT2	31968318	31968449	001.6^001.7, 003.4^003.5, 010.2^010.3, 008.5^008.6, 009.5^009.6, 007.6^007.7	0.49	0.34	1.09	-0.15	0.60	0.75	3.72x10^{-5}
	31956020	31956428	007.26^007.27, 001.27^001.28, 005.8^005.9, 008.26^008.27, 009.26^009.26, 004.9^004.10, 003.25^003.26, 006.8^006.9	1.09	1.00	0.79	-0.10	-0.30	-0.21	0.047
C6orf25	31799742	31800501	004.2^004.3	1.26	1.54	-1.42	0.28	-2.69	-2.96	9.28x10^{-5}
			003.2^003.3	-4.60	-4.26	-4.73	0.35	-0.12	-0.47	0.046
DDAH2	31804280	31804401	006.2^006.3, 010.2^010.3, 001.3^001.4, 008.3^008.4, 002.2^002.3, 004.3^004.4, 007.3^007.4	-0.91	-1.69	-2.07	-0.79	-1.16	-0.37	3.93x10^{-4}
C6orf10	32407809	32411191	002.14^002.15, 001.15^001.16	0.60	0.03	0.00	-0.57	-0.60	-0.03	4.32x10^{-4}
CSNK2B	31744928	31745074	001.5^001.6, 007.5^007.6, 008.5^008.6, 004.5^004.6, 002.5^002.6, 009.5^009.6	2.13	1.71	2.29	-0.43	0.15	0.58	0.0012
	31742243	31742576	009.1^009.2	-3.00	-3.20	-2.59	-0.19	0.42	0.61	0.0043
	31744410	31744852	001.4^001.5, 007.4^007.5, 008.4^008.5, 004.4^004.5, 002.4^002.5, 009.4^009.5	3.46	3.17	3.54	-0.28	0.09	0.37	0.047
C6orf48	31913191	31915298	007.4^007.5, 001.3^001.4, 004.3^004.4, 008.2^008.3, 006.4^006.5, 009.1^009.2, 005.3^005.4, 002.2^002.3, 003.4^003.5	3.63	2.72	3.03	-0.91	-0.60	0.31	0.0015
	31910956	31911166	006.1^006.2, 003.1^003.2, 007.1^007.2, 001.1^001.2	0.77	-0.07	0.27	-0.84	-0.50	0.34	0.0030

	31911207	31912991	008.1^008.2, 004.2^004.5, 005.2^005.3, 001.2^001.3	1.21	0.34	0.69	-0.87	-0.52	0.35	0.0127
	31911207	31912054	006.2^006.3	-2.50	-3.04	-2.67	-0.54	-0.17	0.37	0.049
BAG6	31719950	31720062	004.12^004.13	0.46	0.43	-0.49	-0.03	-0.95	-0.92	0.0015
	31714982	31715255	031.1^031.2, 029.3^029.4, 002.24^002.25, 003.24^003.25, 001.24^001.25	1.86	1.46	1.31	-0.39	-0.55	-0.15	0.0022
	31714982	31715954	028.5^028.6, 030.3^0.30^4	0.52	0.89	1.10	0.37	0.58	0.21	0.0077
	31716062	31716400	0.30.2^030.3	-3.38	-3.32	-2.88	0.06	0.50	0.44	0.0147
PRRT1	32227930	32229491	005.2^005.3	2.56	2.01	1.95	-0.55	-0.62	-0.06	0.0030
	32229591	32229992	008.1^008.2, 012.1^012.2, 005.1^005.2	3.76	3.18	3.42	-0.59	-0.34	0.25	0.0041
	32225153	32225291	006.2^006.3, 001.5^005.2, 002.3^002.4, 003.2^003.3, 007.1^007.2	-3.39	-2.00	-3.12	1.39	0.27	-1.12	0.0145
DAQB-331I12.5	31959567	31959638	001.1^001.2	-0.01	-1.07	-0.14	-1.06	-0.13	0.93	0.0077
NOTCH4	32296039	32296159	001.6^001.7, 002.6^002.7	-1.27	-0.36	-1.08	0.91	0.19	-0.73	0.0088
	32272176	32272679	005.1^005.2, 003.8^003.9, 001.28^001.29	0.21	-0.14	-0.43	-0.35	-0.64	-0.29	0.0159
	32277255	32277830	001.21^001.22.	1.79	2.32	1.87	0.54	0.08	-0.46	0.039
	32278354	32279524	001.20^001.21	-1.86	-1.67	-1.16	0.19	0.70	0.51	0.046
ABHD16A	31767436	31767559	005.6^005.7, 004.6^004.7, 001.8^001.9, 013.2^013.3, 002.7^002.8, 006.1^006.2, 003.8^003.9	0.28	0.70	0.28	0.42	0.00	-0.42	0.014
	31767674	31768782	005.5^005.6, 004.5^004.5, 013.1^013.2, 002.6^002.7, 001.7^001.8, 003.7^003.8	0.56	1.23	0.83	0.67	0.27	-0.40	0.049
APOM	31728307	31732227	003.1^003.2, 002.1^002.2	-1.13	-1.57	-1.60	-0.44	-0.47	-0.03	0.015
STK19	32048675	32054658	010.1^010.2	-1.38	-1.62	-1.23	-0.25	0.14	0.39	0.017
	32048267	32048376	004.2^004.3, 007.1^007.2, 001.2^001.3, 005.1^005.2, 002.2^002.3	5.20	4.70	5.31	-0.49	0.11	0.60	0.031
	32054754	32055169	003.2^003.3, 010.2^010^3, 009.2^009.3, 001.4^001.5,	-1.83	-1.98	-1.55	-0.15	0.28	0.43	0.034

			008.2^008.3, 005.4^005.5							
LSM2	31881898	31882510	002.1^002.2, 001.2^001.3, 003.1^003.2	-0.79	-1.06	-0.60	-0.27	0.20	0.47	0.024
TNF	31652570	31652871	001.3^001.4	-0.65	-0.23	-0.30	0.42	0.35	-0.06	0.032
PRRC2A	31711505	31711722	002.24^002.25, 001.24^001.25	1.80	1.39	1.50	-0.41	-0.30	0.11	0.038
SKIV2L	32044294	32044378	010.1^010.2, 002.21^002.22	0.49	-0.01	0.36	-0.50	-0.13	0.37	0.046
EGFL8	32242375	32242455	001.3^001.4, 002.3^002.4, 003.1^003.2, 005.2^005.3, 004.3^004.4	-0.95	-1.05	-1.46	-0.51	-0.51	-0.41	0.047
NCR3	31665389	31665537	006.2^006.3, 001.2^001.3, 007.2^007.3, 002.2^002.3, 003.2^003.3, 004.3^004.4	1.38	1.83	1.74	0.45	0.36	-0.09	0.048
NFKBIL1	31622729	31623918	003.1^003.2	-1.35	-0.98	-0.96	0.37	0.39	0.02	0.049

3. SUPPLEMENTAL REFERENCES

Aho AV, Hopcroft JE, Ullman JD. 1983. *Data Structures and Algorithms*. Pearson Education

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H et al. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33**(20): e175.

Fairfax BP, Vannberg FO, Radhakrishnan J, Hakonarson H, Keating BJ, Hill AV, Knight JC. 2010. An integrated expression phenotype mapping approach defines common variants in LEP, ALOX15 and CAPNS1 associated with induction of IL-6. *Hum Mol Genet* **19**(4): 720-730.

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**: S96-104.

Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen MM, Lu G, Fang J, Liu WM, Ryder T et al. 2003. Probe selection for high-density oligonucleotide arrays. *Proc Natl Acad Sci U S A* **100**(20): 11237-11242.

Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* **3**(7): 511-518.

Smyth GK. 2004. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1): art3.