VNIVERSIDAD NACIONAL AVFNOMA DE MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

CENTRO DE CIENCIAS GENÓMICAS

Functional Effects of Haplotype-Specific Open Chromatin Features in the MHC Region

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
LICENCIADA EN CIENCIAS GENÓMICAS

PRESENTA:

MARISOL ÁLVAREZ MARTÍNEZ

Tutora:
DRA. GOSIA TRYNKA

2016

CUERNAVACA, MORELOS

A quien la lea

ABSTRACT

Next generation sequencing analyses are designed for haploid genomes, which bear no representation of genetic variation. Thus, highly polymorphic loci have not been comprehensively investigated using high-throughput approaches, mainly due to read mapping biases. Here we provide a more accurate and extensive description of the open chromatin landscape of the MHC region by analysing the alternative loci of this genomic location. We propose a methodological framework that makes use of the variation aware aligner Bwakit and is able to integrate its output with established data resources and software designed for haploid genomes. By using this approach we reveal haplotypic differences in open chromatin and gene expression within the MHC class I and class II regions.

TABLE OF CONTENTS

I. INTRODUCTION

The vast majority of bioinformatic analyses assume a haploid and “linear” reference assembly model (Church *et al*. 2015), which represents the consensus genomic sequence of a species with no representation of variation whatsoever. Accepting such paradigm in the MHC region, the most polymorphic genomic location, will lead to biological information loss; however, the significance of this loss is unknown. Throughout this thesis we aim to explore, at different molecular levels, the answer of this interrogation. While doing so we explore and characterise the utility of the variation aware aligner Bwakit (Li 2016), and unravel haplotype-specific open chromatin regions; mainly near the *HLA-DRB1* gene.

This thesis is organised as follows, section II is dedicated to genetic variation and its representation in the reference assembly; likewise, an insight into the genetics of the MHC region is presented. Sections III, IV, and V describe in detail the goals of the project and the datasets analysed. In section VI, we examine the disparity between the annotated gene features in several assembled MHC haplotypes and highlight their relevance for functional next generation sequencing analyses . Next, in sections VII and VIII we evaluate a non-conventional read mapping software, a variation aware aligner. There are no current publications using the Bwakit; accordingly, we assessed its biases and integrated it into a pipeline for querying open-chromatin in polymorphic regions. In section IX we explore the functional implications at transcriptome level of the open chromatin regions that differ between the most scrutinised MHC haplotypes to date.

In section X we summarise the findings presented, as well we reflect on the limitations of our approach and offer some perspectives for the future developments in the MHC field. Furthermore, we discuss how the new sequencing technologies could aid in a deeper characterisation of this important genomic region.

II. THEORETICAL BACKGROUND

II.1. Reference Assemblies and Modern Bioinformatic Caveats

II.1.1 The Human Genome Project

The Human Genome Project (HGP) provided a high-quality reference assembly that allowed us to read the manuscript of nature to understand human life, from its organization to its origins and diversity. With it, the field of genomics had grown without precedents in the last decade. The first reference assembly was created out of the collapsed sequences from over 50 individuals. Each chromosome was represented in a "linear" haplotypic sequence, with a minuscule representation of sequence or structural variation. The most frequent errors in this assembly were due to complex structural variation, which hindered the overlap between clones (Church *et al.* 2015).

II.1.2 The Genome Reference Consortium and modern assemblies

In 2007 the administration of the reference assemblies changed to the Genome Reference Consortium (GRC); with it a new model of assemblies was proposed and GRCh37 was the first implementation of it. In this new design the addition of "alternative loci" was made in regions housing complex structural variation and extensive polymorphism. The alternative loci are aligned to the haploid primary assembly and provide alternative sequences to highly diverse regions; having as an outcome neither a completely haploid or diploid assembly (Church *et al.* 2015; Church *et al.* 2011). The newest implementation of the human reference assembly, GRCh38, represents 2.6 Mb more of novel sequence in alternative loci compared to GRCh37; nevertheless, not even in the newest assembly the alternative loci provide a complete catalogue of variation, they just represent an immediate solution to the lack of diversity represented in the primary assembly (Church *et al.* 2015).

Besides the alternative loci there are sequences whose location in a chromosome is not known, they are defined as unlocalized sequences/contigs. On the other hand, the term unplaced sequence/contig refers to those sequences whose location nor chromosome of origin is known (Church *et al.* 2011). The relevance of taking in consideration the alternative loci in next generation sequencing (NGS) analyses has become increasingly relevant as they can lead to off-target sequence alignments; and consequently errors in variant calling, underestimation of gene expression, etc. In fact when simulating reads from the alternative loci and re-mapping them only to the primary reference assembly 75% of them will map incorrectly (Church *et al.* 2015; Church *et al.* 2011). Such was the effect of off-target alignments that the Simons Genome Diversity Project generated a set of decoy sequences to lessen the bias. The decoy sequences comprise a set of sequences absent from the reference assembly GRCh38 but present in the assemblies of some of their samples, their main usage has been to minimize alignment errors (Mallick *et al.* 2016). In this document the primary reference assembly, refers to the haploid sequence that constitutes the chromosomes; on the contrary the extended reference assembly is the name assigned to the set of sequences comprising the primary assembly reference plus the alternative, unplaced, unlocalized, and decoy sequences and loci.

II.1.3 Bioinformatic hindrances to analysing alternative loci

The vast majority of bioinformatic algorithms and reporting formats were developed having in mind a linear haplotypic assembly model, therefore, they expect sequencing reads and genetic features to be localised in a single location in the reference assembly. As a matter of fact, several aligners penalize multimapping reads under the assumption that their location cannot be resolved due to paralogy; consequently, if the alternative loci are added carelessly, the alignment software is unable to distinguish between duplication arising from paralogy from allelic duplication introduced by the alternative loci (Church *et al.* 2015).

II.1.4 Biological importance of alternative loci

The existence of alternative loci is well documented, especially in the field of Population Genomics. For example, Li *et al*. recovered ~5 Mb of alternative loci absent from the HGP primary assembly; additionally, they could annotate coding regions in those alternative loci. Two thirds of those genes actually belonged to families of hypervariable genes such as mucins, olfactory receptors, HLA genes, etc (R. Li *et al.* 2010). Although higher levels of variation in these genes are appreciated between populations, the diversity across individuals of the same population is striking as well. The relevance of analysing the alternative loci and, therefore, development of computational tools is clear in disease association studies. In fact, the genomic region associated with most of immune mediated diseases is the major histocompatibility complex (MHC) region, which harbours the highly polymorphic human leukocyte antigen (HLA) genes (Traherne *et al.* 2006; Carapito, Radosavljevic, and Bahram 2016).

II.2 The Major Histocompatibility Complex (MHC) Region

The MHC region is located in the short arm of chromosome 6, comprising almost 4 Mb and ~0.6% out of the total number of genes in the human genome, making it the most gene-dense region (Trowsdale and Knight 2013; Carapito, Radosavljevic, and Bahram 2016). It is recognised as one of the most important genetic locations due to its association with numerous autoimmune disorders, susceptibility to infectious diseases, and its determining role in organ transplant compatibility (Traherne *et al*. 2006; Carapito, Radosavljevic, and Bahram 2016). The most prominent genes located inside the MHC region are the human leukocyte antigen (HLA) genes which are in responsible for antigen presentation, a process that consists of presenting small peptides to T cells, determining whether or not to initiate an immune response. In order to bind a wide variety of peptides, these genes are evolutionarily selected to be highly polymorphic; actually they are the most polymorphic genes in the entire genome. Functional HLA genes code for the glycoproteins that form the MHC protein

complex, which is formed by an α-chain and a β-chain (Marsh, Parham, and Barber 2000). The HLA genes classify the MHC region in three subregions: class I, II, and III (figure 1A).

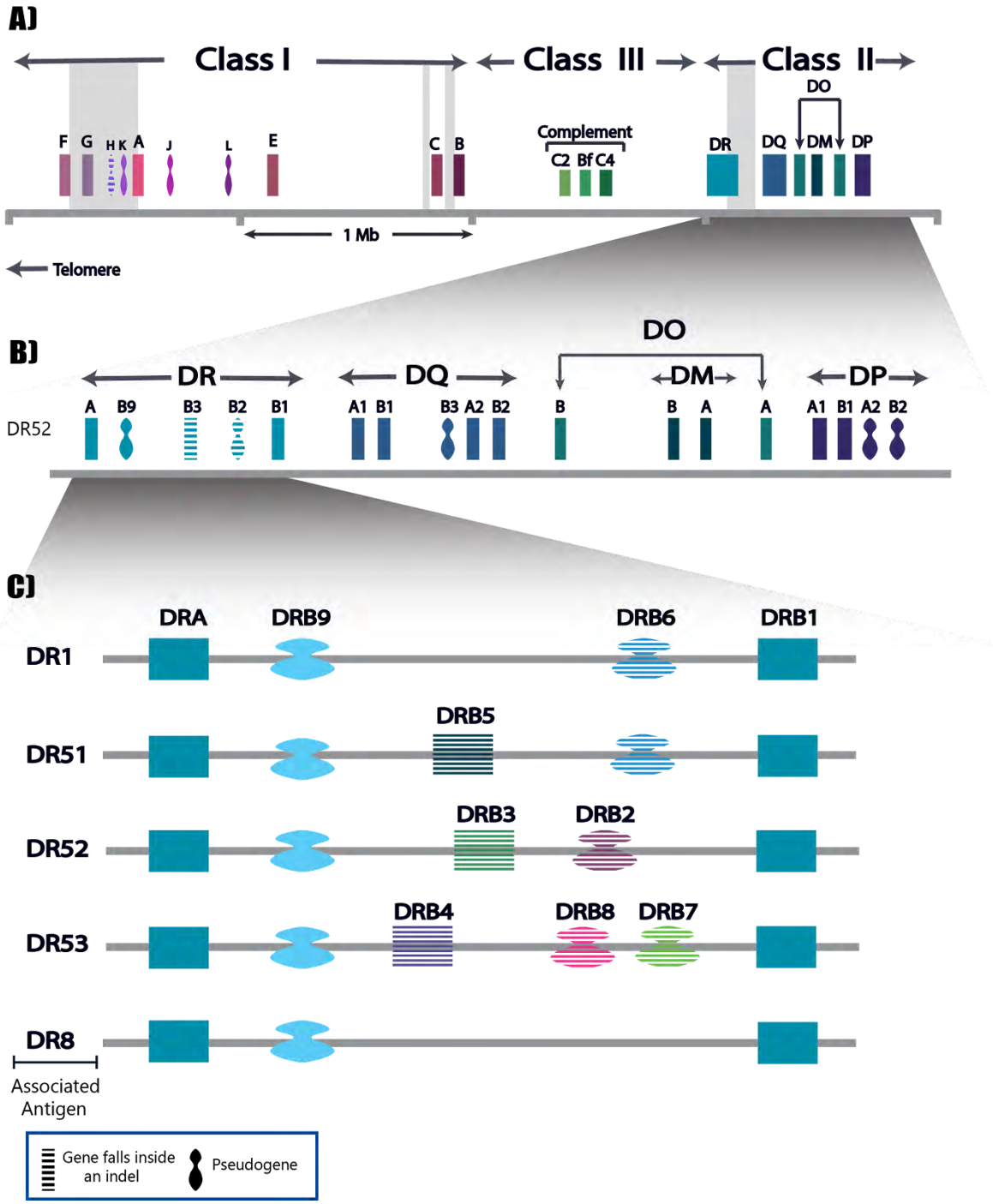

Figure 1. A) Depicts the organization of the HLA and Complement genes inside the MHC region, as well as the location of reported indels (shaded regions) (Norman *et al.* 2015). B) shows a detailed representation of the location of the class II HLA genes, in this figure the haplotype with the associated antigen "DR52" is represented. C) Examples of structural variants located in the DR region. Figure adapted from (Marsh, Parham, and Barber 2000) and (Norman *et al.* 2015).

II.2.1 MHC Class I region

The HLA class I functional genes: *HLA-A, -B, -C, -E, -F,* and *-G* are located in the MHC class I subregion, along with ~50 other genes unrelated to antigen presentation and HLA class I pseudogenes: *HLA-H, -J, -K, and -L*. A 50 kilobase deletion encompassing *HLA-H* has been reported in this region. MIC and HFE gene families, also located in this subregion, are class I-like genes involved in immune surveillance. While all the class I genes code for the α-chain, the complementary β-chain is located in chromosome 15 (Marsh, Parham, and Barber 2000).

II.2.2 MHC Class II region

HLA class II genes are classified in five isotypes: *HLA-DM, -DO, -DP, -DQ, and -DR*. Each of these isotypes has at least one gene coding for an α-chain and another for the β-chain. The pseudogenes that locate in this region include: *HLA-DQB3, HLA-DPA2, HLA-DPB2, HLA-DRB2*, and *HLA-DRB9*. The complexity of class II genes can be highlighted by the HLA-DR isotype, as the quantity of functional genes and pseudogenes present in the region varies between individuals (figure 1C). The functional genes *HLA-DRB5, -DRB4, -DRB3* might be deleted or present in some haplotypes in a mutually exclusive manner; the same goes for the pseudogenes *HLA-DRB2, -DRB6, -DRB7, -DRB8*. The indels that harbour these "optional" HLA-DR isotypes are always located between *HLA-DRB1* and the pseudogene *HLA-DRB9* (Deakin *et al.* 2006). There are few other non-HLA genes in this region, also with functions related to antigen presentation (Marsh, Parham, and Barber 2000).

II.2.3 MHC Class III region

MHC class III region is the most gene dense region, spans 700 kb, and is localized in between class I and class II regions. Although it does not contain HLA-genes it houses several genes involved in innate immunity along with other non-immune genes. MHC class III region is conserved across species, although gene-content

may vary slightly especially in highly specialized immune genes (Deakin *et al.* 2006; Trowsdale and Knight 2013) .

It is intuitive that the MHC region is the most disease-associated region due to its high levels of polymorphism and quantity of genes (Trowsdale and Knight 2013). The allele count for each of the HLA "classical" alleles goes above a thousand according to the IPD-IMGT/HLA database (Robinson *et al.* 2015); furthermore, linkage disequilibrium (LD) is high among these genes, consequently, the MHC region contains haplotypic blocks with reported functional consequences (Traherne *et al.* 2006; Vandiedonck *et al.* 2011). Although the high variability and LD are hallmarks of the HLA genes, these characteristics are not unique to these genes, as they have been reported in the MHC class III region as well (Vandiedonck *et al.* 2011; Yau *et al.* 2016).

Beyond the functional genomics analyses, there is a methodological barrier to analysing the MHC region, it cannot be properly studied with conventional genome-wide analyses, as the available methods do not take in consideration its variability. Thus, results from international efforts that characterise in a high-throughput manner the epigenome (ENCODE project (Kellis *et al.* 2014)), variation (1000 Genomes Project (Auton *et al.* 2015)), transcriptome (GTEx (GTEx Consortium 2015)) lack adequate accuracy in the MHC region.

II.2.4 The MHC Haplotype Project

Circa 2008 the completion of the sequencing, assembly, and annotation of eight MHC haplotypes of the homozygous cell lines: PGF, COX, QBL, MANN, APD, DBB, MCF, and SSTO the MHC Haplotype Project claimed to generate a comprehensive variation map of the MHC region for Europeans (Horton *et al.* 2008). The PGF haplotype is the one embedded in the human primary reference assembly in chromosome 6, the rest are considered alternative loci for MHC region and span ~4 Mb each. When comparing PGF, COX, and QBL haplotypes it was shown that the majority of variants fell in intergenic regions, with a high variation in repetitive

elements, and most of the coding variants fell in the HLA genes as expected (Traherne *et al.* 2006). The MHC Haplotype Project foresaw that the characterization of these haplotypes would provide a framework and a helpful resource for studying the MHC region (Horton *et al.* 2008). Nevertheless, the appropriate handling of all haplotypes in bioinformatic pipelines has not been established and most of the efforts from the GWAS and organ transplant community have focused on characterising variation in the classical HLA genes only (Carapito, Radosavljevic, and Bahram 2016). Establishing computational methods for analysing the non-coding regions of the MHC region will be essential for its complete understanding and, therefore, aid in the disentangling of the molecular mechanisms behind autoimmune disorders.

III. HYPOTHESIS AND MAIN OBJECTIVE OF THE PROJECT

The MHC region is one of the most polymorphic loci in the human genome, with a great density of genetic variants it is not feasible to be represented solely as a primary reference genome. Consequently, the MHC region has not been extensively explored through next-generation sequencing analyses. In this project we assessed how genetic variation in the MHC impacts gene expression regulation assessed through the chromatin accessibility profiles.

IV. PARTICULAR OBJECTIVES OF THE PROJECT

1. Assess the completeness of the primary reference assembly (PGF haplotype) in comparison to other Caucasian MHC haplotypes.
2. Evaluate the Bwakit, the variation aware version of the BWA-MEM aligner, for querying the epigenome within the MHC region.
3. Identification of haplotype-specific open chromatin features in the MHC region.
4. Describe haplotype-specific functional effects in the MHC region at gene expression level.

V. SAMPLES AND DATASETS ANALYSED

We analysed open chromatin profiles from assay for transposase accessible-chromatin followed by whole genome sequencing (ATAC-seq) generated from 24 lymphoblastoid cell lines (LCL) derived from British individuals (Kumasaka, Knights, and Gaffney 2016). Genotype data for these samples was available through the 1000 Genomes Project (Auton *et al.* 2015) and transcriptome data from the Geuvadis Project (Lappalainen *et al*. 2013).

V.1 Genetic Divergence From the Primary Reference Assembly

The haplotype of the PGF cell line is present in the primary reference assembly and we first sought to characterise how well the PGF haplotype represents each of our samples, and identify those that can benefit more from the usage of variation aware aligners. To do so, we defined the "genetic divergence score" (GDS), which is a quantitative measure of genetic distance between a sample and PGF haplotype. It takes advantage of the genotypes provided by the 1000 Genomes Project in a way that each variant can be a "reference" or an "alternative" allele, depending if the locus in question has the same nucleotide sequence than the primary reference assembly, in this case PGF, or not. The genetic divergence score is calculated in a bin-wise manner as follows:

$$GDS\ (Bin_x) = \frac{1}{2N} \sum_{i=0}^{N} (a_i + b_i)$$

Where *a* and *b* are boolean variables, each acquiring the value of 1 when the alternative allele is present for variant *i* in the first (*a*) and second (*b*) haplotypes of the diploid sample under analysis. The genotypes were queried as part of the 1000

Genomes project (Auton *et al.* 2015). *N* represents the size of the bin, in this case it comprises a thousand variants.

Figure 2 provides a visual representation of the genetic divergence score for 20 of our samples. As expected, the highest divergence scores are achieved in and near the most polymorphic HLA genes, and it varies extensively between samples, even though they come from the same population. It can be appreciated that for 17 out of 20 samples the use of alternative loci will be beneficial for an in-depth analysis of the MHC region. Additionally, we can conclude that each sample consists of a set of personalized variants.

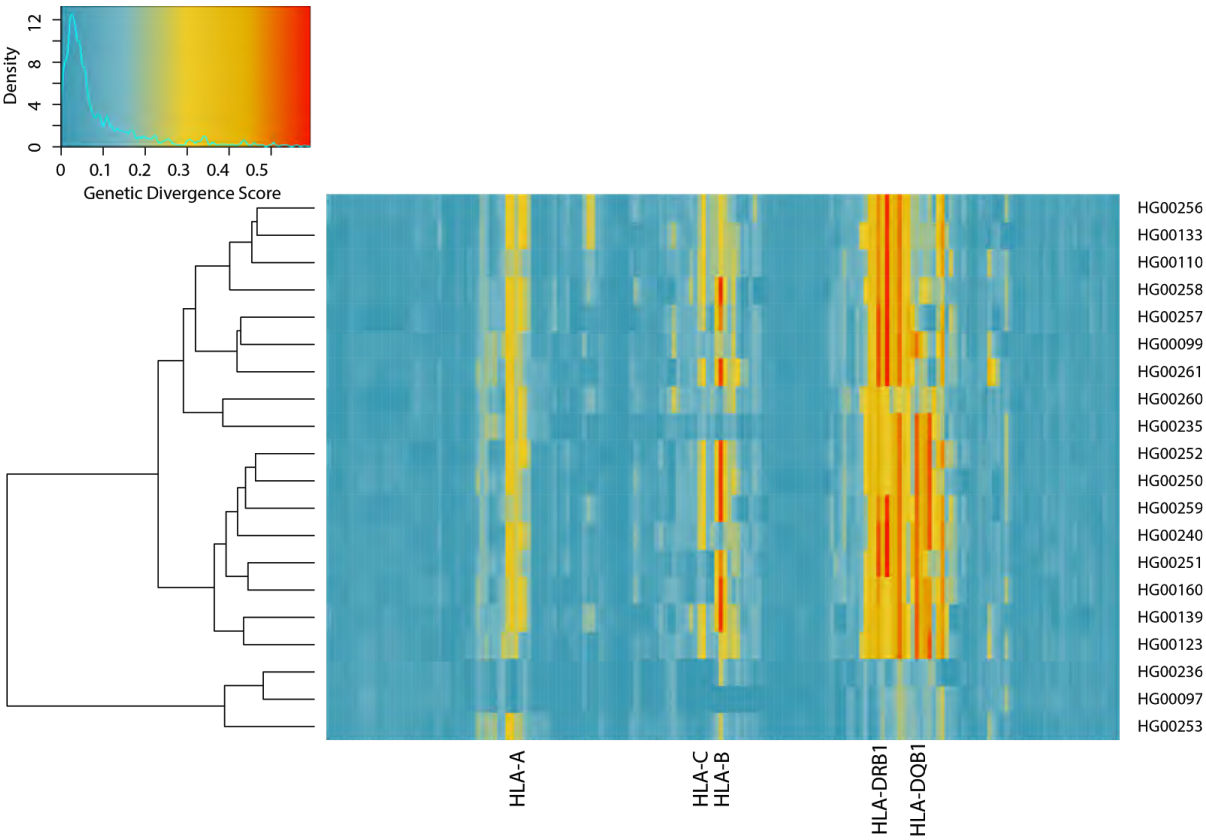

Figure 2. The Genetic Divergence Score was calculated per sample and represented as a heatmap. The highest values of the GDS are achieved in the genes HLA-A, -C, -B, -DRB, -DQB1 as expected.

VI. MHC REGION ASSESSED AT THE GENE FEATURES LEVEL

VI.1 Contrasts Between MHC Haplotypes at Gene Content Level

The aim of genome annotation is to identify functional elements. GENCODE annotations have stood out for their high quality by merging ENSEMBL and HAVANA annotations, this includes a manual annotation step (Aken *et al.* 2016). There have been differences reported in gene content between MHC haplotypes in previous HAVANA releases (Horton *et al.* 2008). We reassessed these differences in the newest version to date of GENCODE, release 25 (Harrow *et al.* 2012). The purpose of this analysis is to offer a better grasp on the biases that are encountered when just the primary reference assembly (PGF haplotype) is analysed. To do so, we retrieve genes that are putatively affected by the highly polymorphic nature of the MHC region.

VI.1.1 Method: Comparison at gene features level

The annotations used for the MHC haplotypes were taken from GENCODE (Harrow *et al.* 2012) release 25 for GRCh38.p7; the GFF file used was the “Comprehensive gene annotation” encompassing all the reference chromosomes, scaffolds and alternative loci. We defined in PGF the coordinates chr6:28702185-33451429 as the MHC region; and analyzed the following MHC alternative loci: COX, APD, DBB, MANN, SSTO, QBL, and MCF. The comparison of genetic features was done at gene symbol level, given that gene identifiers differ between alternative loci, as annotated by Ensembl release 85 (Aken *et al.* 2016). The final list of genes absent from PGF was subjected to manual curation.

VI.1.2 Results: Comparison at gene features level

Unsurprisingly, we observe that PGF has the highest quantity of unique genes out of all MHC alternative loci (figure 3A), as a consequence of being included in the primary reference assembly, *ergo* is more subjected to genome-wide analyses. Meaning as well that further refinement of gene annotation needs to be done in the MHC alternative loci. There are a few functional protein-coding genes absent from PGF as noted in figure 3C, in which known genes such as *HLA-DRB3* and *HLA-DRB4* are included. Nevertheless, processed transcripts are the most frequent biotype absent from PGF.

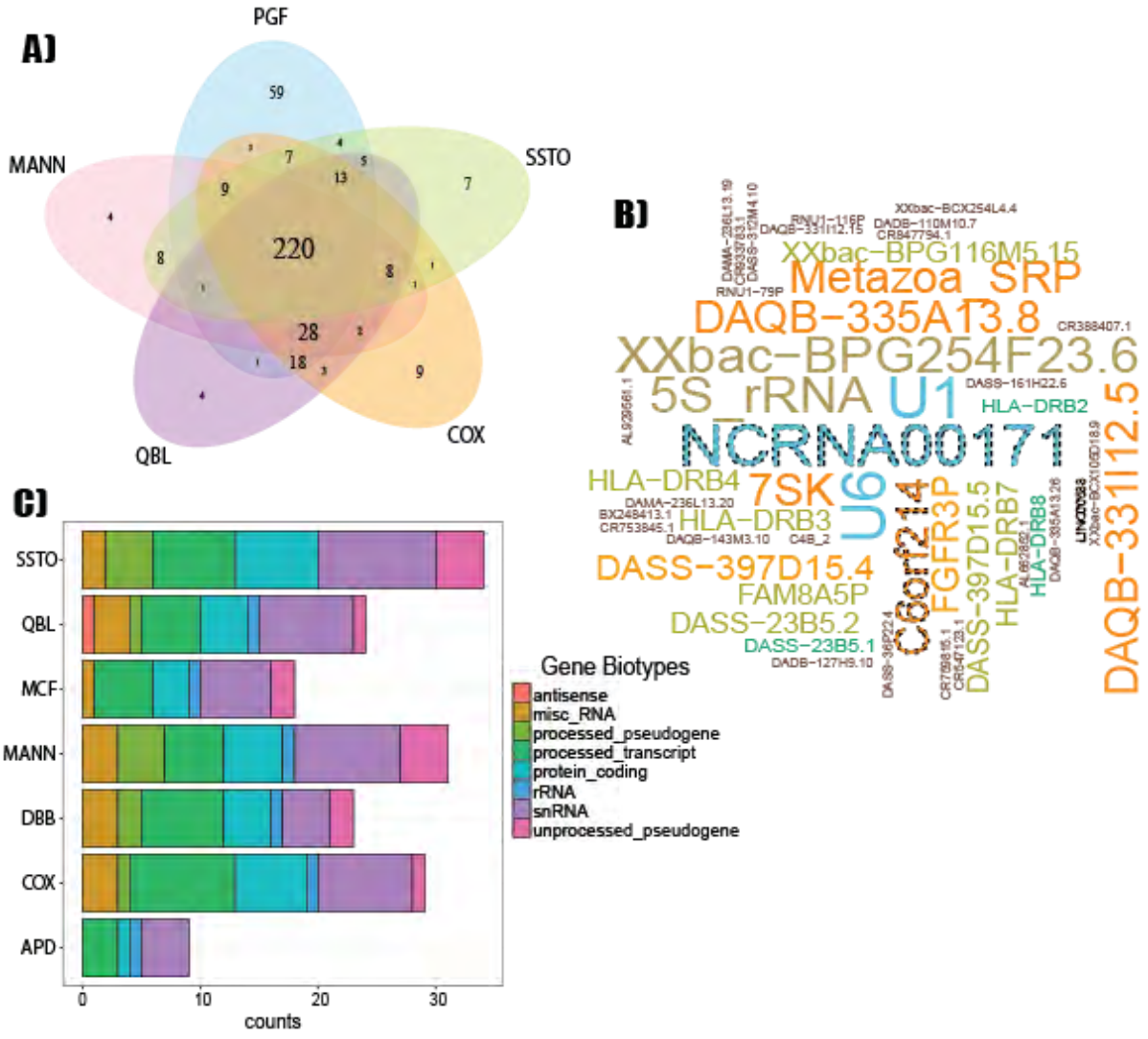

Figure 3. A) Venn diagram denoting the comparison of gene symbols annotated to five assembled MHC haplotypes: PGF, MANN, QBL, COX, and SSTO. B) Genes present in the alternative haplotypes of the MHC region but absent from the PGF haplotype; size and color correspond to frequency between the alternative MHC haplotypes. Genes with dashed margins (*LINC00533, C6orf214,* and *NCRNA00171*) have synonym genes in PGF but with different biotype C) Gene biotype counts of the genes that are not represented in PGF; biotypes were recovered from Ensembl (Aken *et al.* 2016).

VI.2 Expression Levels of Genes Absent from PGF

The aim of identifying genes absent from PGF is to highlight the importance of taking in consideration the alternative loci of the MHC region for genomic analysis. Pseudogenes, repeats, and polymorphic genes possess a challenge for genome annotation pipelines (Pei *et al.* 2012; Aken *et al.* 2016), all abundant in the MHC region; therefore, we do not expect that all genetic features missing from PGF are completely accurate. The quantification of gene expression of these features can provide validation of their existence, relevance, and putative functionality.

VI.2.1 Method: Estimating gene expression from genetic features absent from PGF

The measurement of genetic features absent from PGF cannot be made in a straightforward manner given that polymorphic genes are included in this analysis (figure 3B), *e.g. HLA-DRB3* and *HLA-DRB4*. Further, the *HLA-DRB2*, *-DRB3*, *-DRB4*, *-DRB5*, etc. genes are present in more than one haplotype, and with different alleles. The information about which of these genes and alleles are present in our 24 LCL samples is not available; thus, we cannot use the variation aware aligners available for RNA-seq data, as they expect just one true alternative allele (Wu and Nacu 2010).

Nevertheless, the main source of bias is the high polymorphism of *HLA-DRB1*. In the case when the alleles of a sample are quite divergent from the ones of PGF, during the alignment process the reads belonging to *HLA-DRB1* would

not be mapping, or would mismap to some other HLA-DRB isotypes, consequently overestimating them. To overcome this bias we obtained from (Gourraud *et al.* 2014) the alleles for *HLA-A, -B, -C, -DRB1, -DQB1* for each of our samples analysed. This led us to design a personalized reference assembly and transcriptome for each of our samples, allowing us to decrease the biases originating from the most polymorphic genes in the MHC region. The nucleotide sequences coding for the peptide of each HLA gene and allele were retrieved from the IMGT/HLA database (Robinson *et al.* 2015) and added to the pertinent reference assembly of each sample, the GTF file was modified to include such sequences as well. The base reference transcriptome used was the "comprehensive annotation" of GENCODE release 25, which includes all the transcripts of the alternative loci.

When the allele of a gene could not be completely resolved for a sample and included more than one option in the data set of (Gourraud *et al.* 2014), a preliminary read mapping round was made with a reference assembly that included all possible allele sequences; the alignment was made with Tophat (Trapnell *et al.* 2012) in junction with Bowtie2 software (Langmead and Salzberg 2012) allowing multi-mapping of reads. The total concordant alignments were quantified for all the alleles tested. The most mapped allele was the one included in the personalized reference transcriptome.

Tophat and Bowtie2 were used once again to align to the personalized reference transcriptome allowing multi-mapping. Multi-mapping was allowed so reads mapping to genes present in more than one MHC alternative loci would not get discarded. Finally, the transcripts per million (TPM) values for each gene were retrieved by adapting the formula from (B. Li and Dewey 2011) in the following way:

$$TPM_n = \frac{X_n \cdot 10^6}{l_n}\left(\frac{1}{Y}\right)$$

$$Y = \sum_i^N \frac{x_i}{l_i}$$

Where *Y* will calculate the library size normalising factor; *N* is the total amount of genes in the transcriptome, *X* is the count of concordant reads mapping to a given gene , and *l* represents the effective length of the gene.

The genes absent from PGF were retrieved from the analysis described in section VI.1. When these genes appeared in more than one MHC alternative loci the highest TPM value among the alternative loci was taken. With this algorithm we do not aim to recover a sensitive measure of gene-expression, only an approximate that will point out whether a gene is expressed or not.

VI.2.2 Results: Estimating gene expression from genetic features absent from PGF

By measuring the expression of the genes retrieved as missing from PGF, we can curate false-positive results such as *NCRNA00171* which is expressed in all samples, a highly unlikely scenario for a gene not included in the primary reference assembly. Although this gene symbol is not present in PGF, we can find its name listed as a synonym for *ZNRD1ASP,* but these gene symbols are registered with a different biotype, this could be a consequence of not analysing the alternative loci as thoroughly as the PGF haplotype. On the other hand, *DASS-23B5.1* appears to be expressed in all samples and is not a synonym name for any gene in PGF. The fact that the latter gene is shown as expressed might be a side effect of misaligned sequenced reads given that this gene is annotated as a processed pseudogene.

Between the genes that would be worthy of experimental validation we can find: *XXbac-BPG254F23.6, CR847794.1, CR759815.1,* and *C4B_2* genes. *XXbac-BPG254F23.6* is annotated as a “processed transcript” and maps to the position of *HLA-DQB1*, no further data is available. *C4B_2* is involved in the complement system and is known to be a copy number variant that may also affect the length of the gene (Chung *et al.* 2002). On the other hand, *CR847794.1* and *CR759815.1* seem to be the same gene, their TPMs are very similar, and both are annotated as protein coding by Ensembl. From the functional genes *HLA-DRB3* and

HLA-DBR4 we retrieved a non-suspicious pattern of gene expression, as there are only a couple of samples expressing them. Further, the sample HG00104 do not express any of them, behaviour expected according to (Gourraud *et al.* 2014) genotype data. Such inference can be made because the sample HG00104 shares the same *HLA-DRB1* and *HLA-DQB1* alleles with PGF, which comprises a well conserved haplotypic block that does not contain *HLA-DRB3* nor *-DRB4 (*Traherne *et al. 2006)*. The rest of the expressed genes retrieved are more challenging to analyse as they are mostly pseudogenes and snRNAs.

In conclusion, the extent of the polymorphism that the MHC region hosts exceeds allelic variation (only within genes), and goes up to the introduction/removal of functional genes and pseudogenes between haplotypes. A median of four protein coding genes in the alternative MHC haplotypes are missing in the primary reference assembly, from which *HLA-DRB3, -DRB4*, *CR847794.1, CR759815.1,* and *C4B_2* are retrieved in the gene expression analysis (figure 4). Furthermore, a median of 24 genetic features annotated in the alternative MHC haplotypes are not present in PGF. Thus, the biological context does vary between MHC haplotypes and needs to be retrieved for an in-depth analysis of regulatory regions within this polymorphic genomic location. Most read aligners cannot manage high amounts of variation with minimum or no *a priori* knowledge, they mostly depend on genotype availability (Wu and Nacu 2010) ; or seek to integrate haplotypes into a joint space that is not yet suitable to be integrated in downstream analysis (Huang, Popic, and Batzoglou 2013; Dilthey *et al.* 2016), like peak calling and gene annotation. In the next section, we explore the Bwakit software that tackles read-mapping biases by integrating known alternative loci while mapping and can be integrated with already established NGS analyses with minor adjustments.

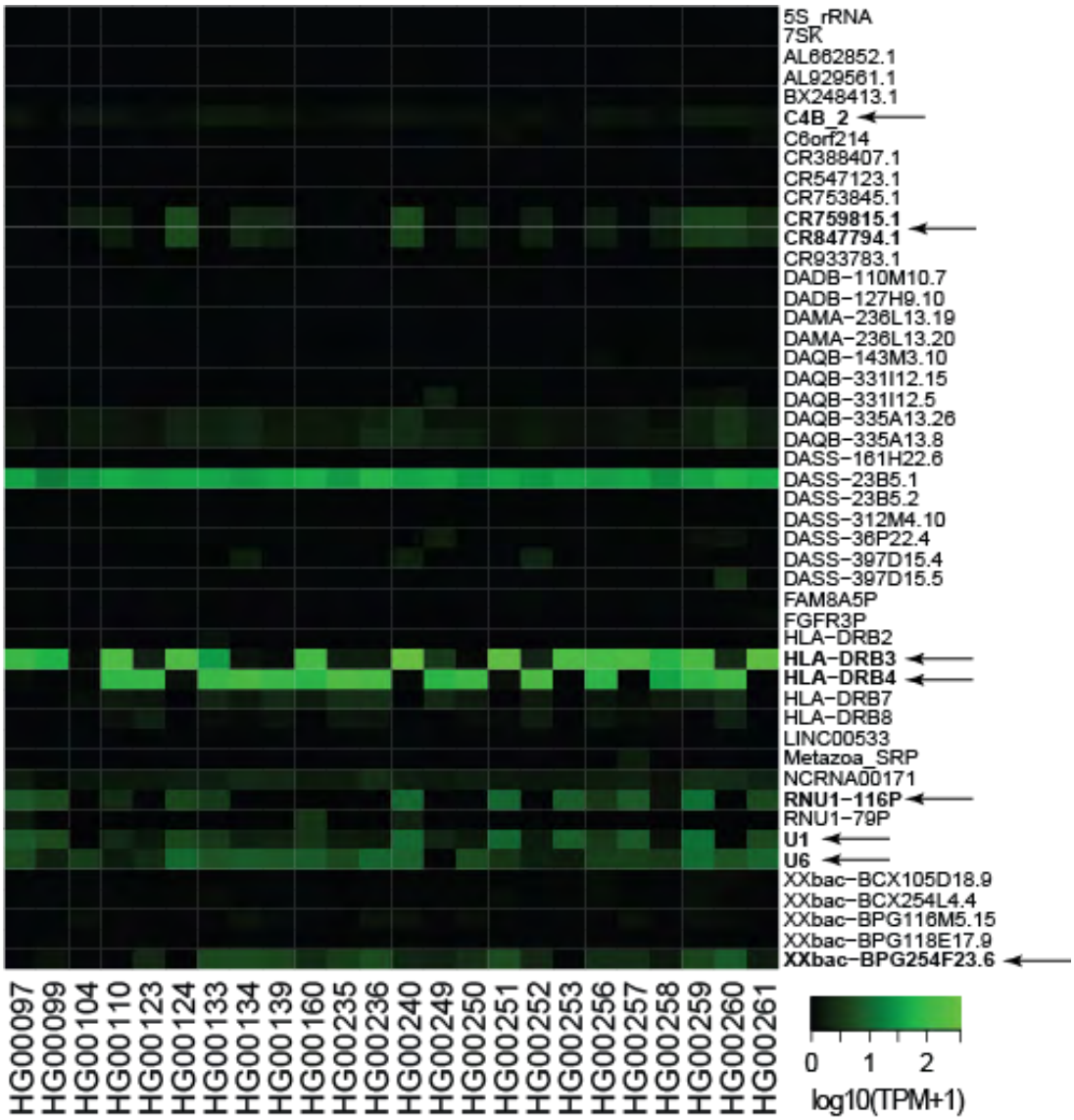

Figure 4. Heatmap of TPMs recovered from the gene features absent from the PGF haplotype, as retrieved in section VI.1. Given that the genes analysed vary between MHC haplotypes, a heterogeneous behaviour of expression is expected. The gene symbols that reflect such behaviour are marked with an arrow. On the other hand, some genes recover positive and uniform TPMs, this can be explained by errors in gene annotation or due to read mapping problems that several pseudogenes suffer from.

VII. USAGE OF VARIATION AWARE ALIGNERS FOR QUERYING THE EPIGENOME OF THE MHC REGION

VII.1 Read Mapping in Highly Polymorphic Loci

Read mapping is the process of aligning reads resulting from a sequencing run to a reference assembly to infer their genomic location. There is a broad range of available software to perform this task. However, software packages vary in speed and accuracy and their performance deteriorates in the presence of genetic variation (Lunter and Goodson 2011). High amounts of polymorphism generate a bias towards the reference allele, causing reads to be mapped to a wrong location, or not mapped at all; given that read mapping is the first step in any NGS analysis workflow, its accuracy impacts on the final results. For example, Brandt *et al*. reported a total of 18.6% incorrect genotype calls at the HLA genes in the 1000 Genomes Project (1kG Project) Phase I data, the hypothesised underlying cause was a mapping bias towards the PGF HLA alleles (Brandt *et al.* 2015). Few aligners have been created to deal with alternative alleles ((Buchkovich *et al.* 2015; Huang, Popic, and Batzoglou 2013; Dilthey *et al*. 2016)), most of them relying on available genotype data at an individual or population level, which is retrieved by interrogating a primary reference assembly. When dealing with minimal or absent structural variation these strategies are pertinent; nonetheless, for genomic locations where high density of polymorphism is a hallmark it may yield incomplete and misleading results. Novel approaches create graph structures to integrate variation information contained in the alternative loci and genotypes (Dilthey *et al.* 2016); nevertheless, the coordinate system they manage is not compatible with available software and data resources for genomes.

VII.1.1 Method: The Bwakit

The Bwakit software package based on BWA-MEM, both developed by Heng Li (Li 2013), is a read aligner that queries consciously and discriminately both reference and alternative loci/contigs (Alt-Ctgs) in a single run with no genotype data required. This makes it suitable to query genomic regions that span several megabases, have more than one alternative loci present in GRCh38 and do not have accurate genotype calls available, such as the MHC region. It could also be implemented to cases where genotypes are available and variation is less frequent, but sensitivity and specificity against other approaches under this scenario has not yet been assessed.

Initially, Bwakit maps the reads using the BWA-MEM software to the primary reference assembly plus the Alt-Ctgs. During this mapping process a read may have multiple hits (matches against the input sequence); BWA-MEM will tag a hit as an alternative (Alt-Hit) or not alternative hit (non-Alt-Hit) depending if the sequence where a read matched is part of an Alt-Ctg or not. The mapping quality (mapQ) of a non-Alt-Hit will be calculated considering all non-Alt-Hits for the read in question (the same as if no Alt-Ctgs were supplied) and the best non-Alt-Hit will be considered the "primary alignment" of the read. The mapQ of an Alt-Hit will be calculated considering both non-Alt-Hits and Alt-Hits for the read, and will be considered as a supplementary alignment (SAM flag 0x800); only when the read has solely Alt-Hits it will be considered as the primary alignment (Li 2016). BWA-MEM will output a SAM file with the best hits for a read (one primary alignment and, if existent, one supplementary alignment). If there were more viable hits they will be included as an XA tag in the primary alignment.

A second step is carried out to recalculate the mapQs of reads with both non-Alt-Hits and Alt-Hits (Li 2016). This step relies on the fact that coordinates in Alt-Ctg (Alt-Pos) have a corresponding coordinate assigned into the primary assembly (Chr-Pos), information stored as an alignment in a SAM file with the extension ".fa.alt" and is provided by the Bwakit. When re-estimating the mapQs for

primary and supplementary alignments of a given read, the Alt-Pos of Alt-Hits are lifted over to the primary assembly to obtain a Chr-Pos; at this point all the hits are in the same coordinate system (same start of contig), making them comparable and may be assessed if the multiple hits have a concordant Chr-Pos. Then, hits are grouped by their Chr-Pos, hits with overlapping positions belong to the same group. A group mapQ is assigned (the highest mapQ among the hits belonging to the group), and all the hits that do not belong to the group with the highest mapQ will be assigned a mapQ of zero. The mapQ of all the hits belonging to the group with the highest mapQ will be:

$$mapQ = 6 * (X[0] - X[1])$$

Where *X* is an array sorted decreasingly and containing the mapQ of each group. When a read has both primary and supplementary alignments, the output goes as follows:

1) one line for the primary alignment (mapQ recalculated) with the additional hits in the XA tag

2) one line with the supplementary alignment (mapQ recalculated)

3) if present, Alt-hits belonging to the group with the highest mapQ

The benefit of using this algorithm is clear when a read has multiple hits with discordant Chr-Pos and the most accurate hit position needs to be retrieved. In figure 5 the following example is provided: a read has four hits which can be grouped into two groups. The blue alignment is the first line in the output SAM file, present as the primary alignment; a mapQ of 0 is assigned to this alignment as it does not belong to the group of the read with the highest mapQ. The second line will be the best supplementary alignment, red, with a mapQ recalculated and bigger than 0, as it belongs to the group with the highest mapQ. Given that the yellow hit is grouped with the red hit, the yellow hit is outputted in the third line. The previously described mapQ recomputing algorithm is available in Heng Li's Github web-page: https://github.com/lh3/bwa/blob/master/bwakit/bwa-postalt.js under the name of bwa-postalt.js.

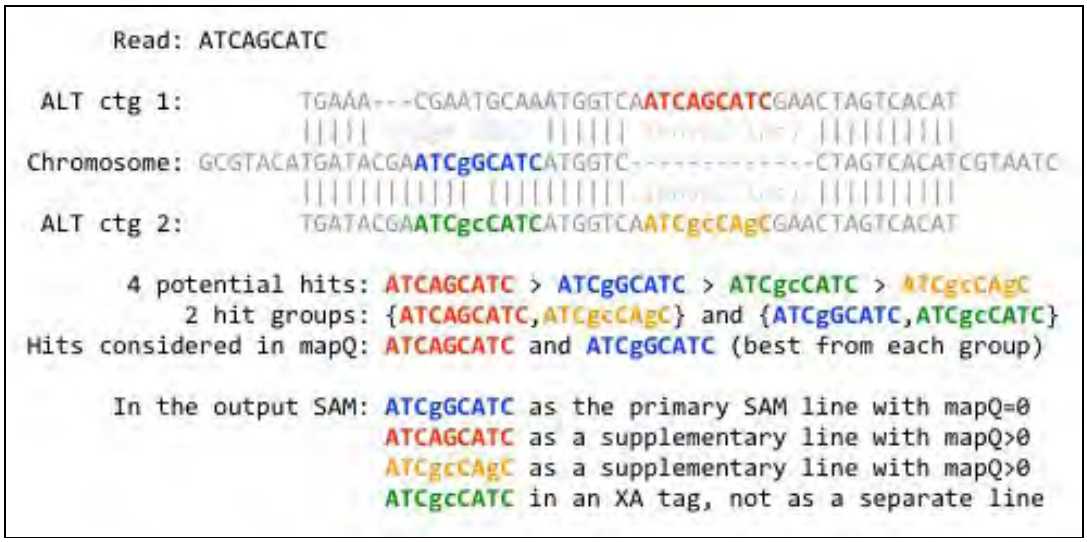

Figure 5. Example of the mapQ recalculating algorithm of Bwakit, in this particular case when alignments with two different Chr-Pos occur. Figure taken from (Li 2016)

VII.1.2 Method: Processing the output of Bwakit

Bwakit will output all the hits of the group with the highest mapQ and all with the same mapQ, so this cannot be accepted as the final output when we are interested in the supplementary alignments. Taking advantage of the order of the output SAM file by bwa.kit, the primary alignment and the first supplementary alignment were kept for further processing; this was achieved by modifying Heng Li's javascript code "bwa-postalt.js". Further filtering was done using *samtools view* (Li *et al.* 2009) to remove unmapped reads and only selecting alignments when both read pairs mapped concordantly; to reduce memory usage reads mapping to mitochondrial DNA (~%50 of ATAC-seq reads) were removed. To further resolve a read's alignment, a python script was developed. This script compares the mapQ and the editing distance of the primary and supplementary alignments and decides which one is the best using these metrics, if there is no best both alignments are kept.

VII.2 Using Variation Aware Aligner Bwakit to Query the Open Chromatin Landscape of the MHC Region

VII.2.1 Method: Input data for Bwakit

Data analysed were ATAC-seq pair-end reads belonging to 24 lymphoblastoid cell lines (LCLs) (Kumasaka, Knights, and Gaffney 2016). Further quality control (QC) was done with *Trimmomatic* (Bolger, Lohse, and Usadel 2014) by trimming bases with a phred score < 35, keeping reads with a minimum length of 50bp, and when both pairs remain after the QC.

The reference assembly for the ATAC-seq reads was the extended GRCh38.p7 plus an extra set of sequences provided by Bwakit in the file hs38DH-extra.fa (Li 2016), which includes decoys retrieved by GenBank (Benson *et al.* 2013) and the Simons Genome Diversity Project (SGDP) (Mallick *et al.* 2016), and the HLA-alleles retrieved from the IMGT/HLA database version 3.18.0 (Robinson *et al.* 2015).

VII.2.2 Result: Analysis of Bwakit alignments

In addition to the Alt-Ctgs, the Bwakit also includes a set of “decoy sequences”, which comprises structural variants that are highly variable in presence, serve to hinder reads from mapping inaccurately into the reference assembly (Mallick *et al.* 2016). To prove that BWA-MEM is factual while assigning mapQs to the supplementary alignments Figure 6A shows the mapQ assigned to the supplementary alignments. The mapQ distribution of the reads mapping to the decoy sequences is below our acceptance threshold for reliable alignments (mapQ >= 30). On the other hand the distribution for the alignments to the Alt-MHC-Hap is above it, and centered in the maximum value, which is 60. Although the Alt-MHC-Hap and the decoy sequences comprise structural variants and polymorphism absent from the reference assembly, the Alt-MHC-Hap contain sequences that are popular amongst

British individuals, therefore with higher similarity to our samples, while the decoy sequences provide a more global set of sequences and not thoroughly characterised.

Finally. we also assessed the location of the reads mapping in MHC-Alt-Hap. Figures 6B and 6C represent the number of alignments in a 10 kb bin located within the MHC region. A strong enrichment of reads mapping to an Alt-MHC-Hap was detected in MHC class II for most samples, accompanied by a gap formed by the absence of reads mapping to PGF, as depicted in figure 6B for sample HG00099. For a sample less genetically divergent from PGF (refer to figure 2), such as HG00097, the gap in PGF is absent, and a pileup of reads near MHC class II has a weaker depth. We therefore concluded, that the Bwakit is sensitive to polymorphic regions, as further explored in this thesis.

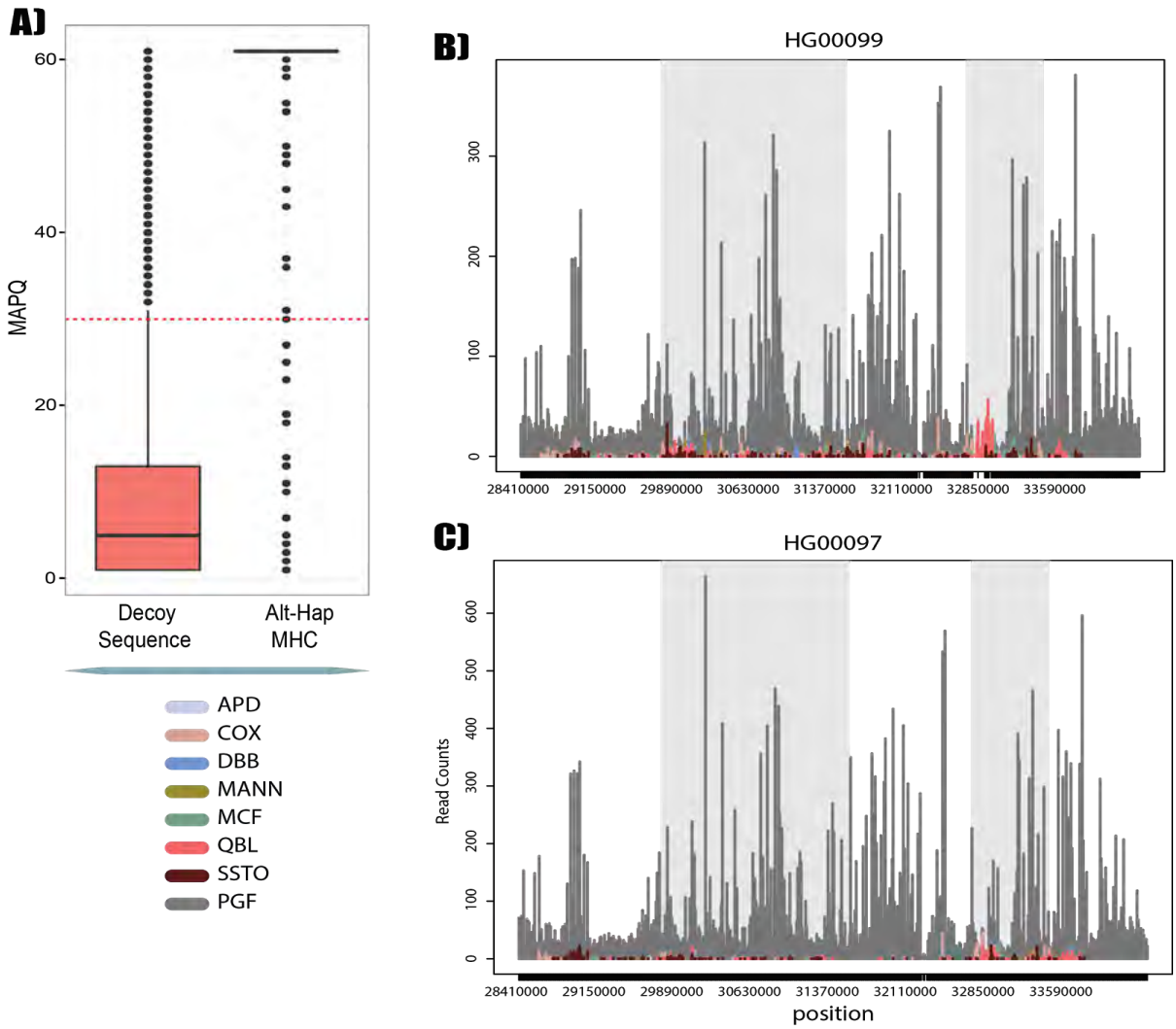

Figure 6. A) Box plot depiction of the mapQs assigned to the supplementary alignments, as defined by the Bwakit; specifically of the alignments done to the alternative loci of the MHC region and the sequences annotated as decoy by the SGDP. The decoy sequences represent recently discovered alternative loci recovered by the assembly of multiethnic genomes; while the alternative loci of MHC represent the most frequent MHC haplotypes in Caucasians. As expected the sequencing reads align better to sequences pertaining to Caucasian population. B) and C) represent the read counts falling into each of the thoroughly characterised MHC haplotypes. Shaded regions represent MHC class I and class II regions. In B) the sample HG00099 is genetically divergent to PGF, thus, several gaps are noted in the x-axis and a pileup of reads mapping to QBL haplotype is appreciated. Sample HG00097 in C) is less divergent to PGF as it can be noted by the absence of gaps in the x-axis and weaker read pile-ups in the alternative loci of the MHC.

VIII. BIASES OF VARIATION AWARE ALIGNERS FOR QUERYING THE EPIGENOME OF THE MHC REGION

VIII.1 Assessing Biases During Peak-Calling in the MHC Region

To ensure that the reads aligning to the MHC-Alt-Hap were not randomly distributed and followed the distribution expected from ATAC-seq reads, the MACS2 software (Zhang *et al*. 2008) was used. MACS2 uses a Poisson based model with a dynamic λ_{local} parameter which captures biases (chromatin structure, sequencing bias, etc.) in a locus specific manner (Zhang *et al*. 2008). Given that a Poisson distribution underlies ChIP-seq and ATAC-seq tags, a level of statistical significance can be assigned to a peak through a *p*-value, and if corrected for multiple-testing a *q*-value.

VIII.1.1 Mehtod: Further processing of Bwakit output for MACS2

During the read mapping process the output from the Bwakit is processed as described in the previous section; however further quality control is needed before running MACS2. At this stage we keep only the concordant alignments with a mapQ greater than 29 (the maximum being 60) and eliminated the supplementary alignment flag (0x800) out of the Alt-MHC-Hap in order to be compatible with MACS2 restrictions.

VIII.1.2 Results: Bwakit biases during peak calling

We tested whether the peaks called at MHC-Alt-Hap (hereafter referred as Alt-Peaks) possessed certain biases compared to the peaks called at PGF (PGF-Peaks). The variables tested were peak length, fold-enrichment of the peak, and the *q-value*. For all the metrics tested the Alt-Peaks are of lower quality than the PGF-Peaks, in fold-enrichment the difference between the median(Alt-Peaks) and

the median(PGF-Peaks) is of -0.301, the q-values are less significant with $\Delta log_{10}(qval)$ equal to -1.933, and the peaks are shorter resulting in the $\Delta(length(peak))$ of -115.99 (for all, the *p-value*<.015 using a Mann-Whitney U test); the metric that is more affected is the length of the peaks, nevertheless, we do retrieve good quality peaks as they are not skewed towards the minimum acceptance threshold for peak-calling (qval(peak)<=.01 in this analysis), as shown in figure 7.

Additionally, it can be noted from figure 7 that the proportion of ATAC-seq peaks within the Alt-MHC-Hap is smaller than the proportion of peaks in PGF, this is due to the fact that only when the sequence underlying an Alt-Peak is polymorphic from PGF sequence an Alt-Peak will be called. To assess this in more detail, a modified version of the genetic divergence score was defined, the "peak divergence score". The peak divergence score (PDS) defines the deviation of a sample from the PGF haplotype and is specific to analysed samples. It is defined as follows:

$$PDS\ (Bin_z) = \frac{1}{X} \sum_{s=1}^{X} \frac{1}{2M} \sum_{i=0}^{M} (a_{s,i} + b_{s,i})$$

Where *X* is equal to the number of analysed samples; *M* represents the total of variants genotyped by 1000 Genomes project in Bin_z; *a* and *b* are boolean variables that are equal to 1 when the variant *i* in sample *s* has the alternative allele in haplotype 1 or 2, respectively, and each bin comprises of 500 bp. As shown in figure 8 no Alt-Peaks are called in a bin when the PDS is equal to 0; there is a weak negative correlation of *r* = -0.12 (*p-val*<0.003) between peak count of PGF-Peaks and PDS. A statistically insignificant correlation is made in the opposite direction for Alt-Peaks, *r* =0.145 (*p-val<0.25*). Both correlations behave as expected, as the amount of polymorphism increases, the PGF haplotype will be mapped less as it does not represent correctly the underlying sequence; on the other hand, an Alt-MHC-Hap can represent the sequence better than PGF, thus, the reads align to it.

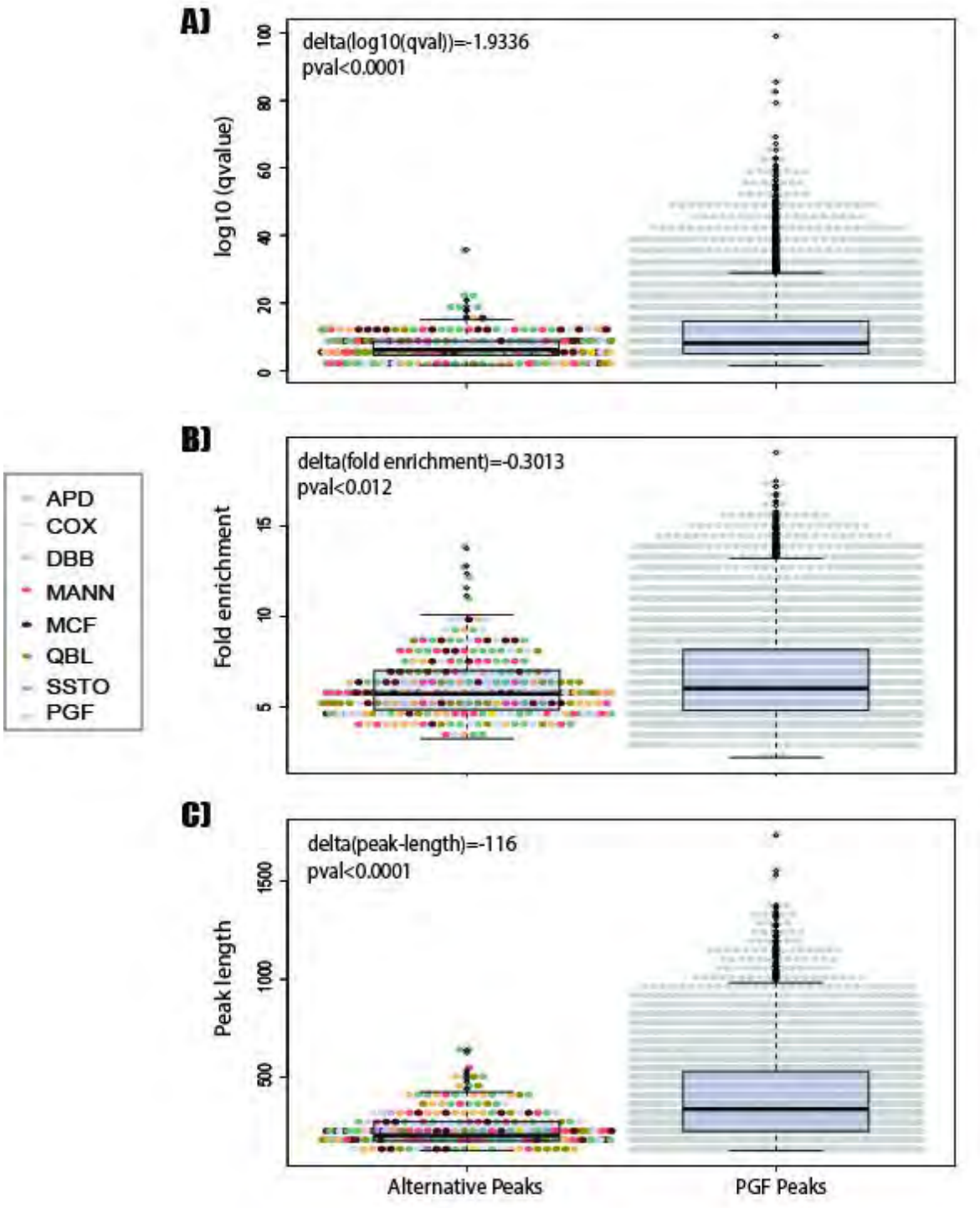

Figure 7. Box plot depiction of q-values (A), fold enrichment (B), and peak length (C) for each peak called by MACS2 software, the color of each dot represents the MHC haplotype from where the peak was called. For all metrics the Alternative Peaks were of lesser quality; nevertheless, not skewed towards quality control thresholds.

When correlating the PDS with the median *q-value* of the peaks in a bin, for PGF-peaks there is a weak negative correlation of $r = -0.1$ with a *p-value*<0.018. In Alt-Peaks a stronger and significant negative correlation is obtained, $r = -0.378$ with a *p-value*<0.002. The later result is unexpected, given that if a sequencing read will not map to PGF due to sequence divergence, then the read will map to an Alt-MHC-Hap. There are two reasons why such correlation might be observed: 1) the PDS is very sensitive to the number of variants in a bin. For example, there is just one variant in a 500 bp bin and all samples are heterozygous in such position, then the PDS equals to 0.5. The issue of such assumption is that high PDS are very unlikely to occur, as it is shown in figure 8C. High values of PDS are only frequent in the MHC class I and class II, but these regions are densely genotyped by the 1000 Genomes Consortium (supplementary figure 1). Second possibility is that the sequence underlying an Alt-Peak is divergent from PGF, but not with multiple Alt-MHC-Hap, consequently resulting in read-mapping problems.

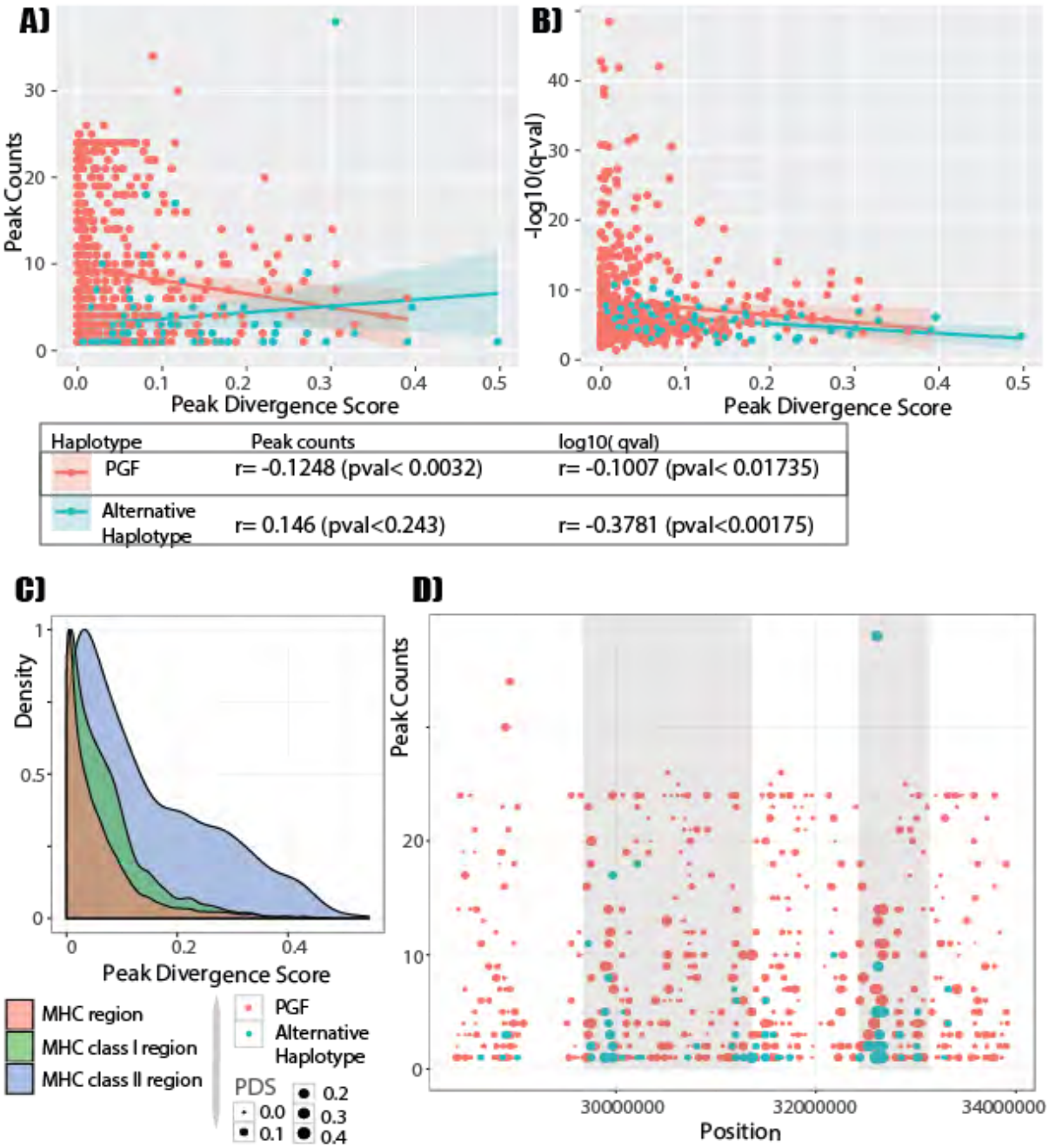

Figure 8. Correlations between the peak divergence score (PDS) and number of peaks called at PGF and the alternative loci of MHC (A) and the q-value obtained from the peak calling analysis (B). No peaks were called with a PDS of 0, meaning that in absence of variation no significant amount sequencing reads will map to PGF. C) represents the frequency of PDS values in the MHC (chr6:28400000-34000000), class I (chr6:29672223-31357223) and class II (chr6:32432223-33132223) regions. The highest PDS scores are achieved in the class II region, and even a shift to a PDS> 0 as the most common value is appreciated, restating the high amounts of polymorphism encountered in the class II regions. D) shows the amount of peaks called at certain position in a 500 bp bin among all samples and its PDS. The highest amount of Alternative Peaks is achieved in MHC class II; care should be taken in interpreting the counts of peaks, as in a Chr-Pos more than one peak per sample might be called if the comprised sequence is polymorphic and under balancing selection (traits of the MHC region). The size of each dot in figure D is associated with the PDS for a given peak.

IX. THE OPEN CHROMATIN LANDSCAPE OF THE MHC REGION

A multicellular organism is comprised of cells with, mostly, identical genetic information. The phenotype of a cell will be determined by which transcripts they express, protein-coding or not; further, gene expression needs to be adaptable and heritable. Epigenetic mechanisms are the key regulators in modulating gene expression. The epigenetic mechanisms can be summarised as follows: 1) covalent modification of DNA, *e.g.* methylation of cytosines is associated with promoter repression, 2) covalent modification of histones, *e.g.* the acetylation of lysine 4 on histone H3 (H3K4Ac) is associated with active genes, 3) and non-coding RNAs, *e.g.* the long non-coding RNA gene *Xist* is responsible for the X chromosome inactivation. (Tough *et al*. 2016). Thus, controlling the binding of transcription factors in regulatory regions. The transposase-accessible chromatin using sequencing (ATAC-seq) is a sensitive method for the interrogation of the open chromatin landscape in a cell-type and condition specific manner (Buenrostro *et al*. 2013). Knowing that there are gene content differences between haplotypes, we aimed to seek haplotype-specific differences in the open chromatin landscape. Our analytical framework allows to call ATAC-seq peaks in the different MHC alternative loci, and make their coordinates comparable between each other.

IX.1 Categorising Open Chromatin Features Within the MHC Region

IX.1.1 Method: Lifting-over of Alt-MHC-Hap coordinates to the human primary reference assembly GRCh38

To make the peaks called at alternative MHC haplotypes (Alt-MHC-Hap) and PGF peaks comparable, a custom script was created to lift-over coordinates from the Alt-MHC-Haps to GRCh38. The script works by parsing the CIGAR string of the alignment of an Alt-MHC-Hap to GRCh38. In this way it is possible to retrieve comparable coordinates, as well as insertions and deletions to PGF.

IX.1.2 Method: ATAC-seq peak classification

The Alt-Peaks were classified in the following manner, and in the hierarchy shown in figure 9:

- **PGF-Alt Peak**: there is an ATAC-seq peak called at PGF and another in an Alt-MHC-Hap within a sample and they overlap in coordinates
- **Alt-Alt Peak**: there are two overlapping ATAC-seq peak called at two different Alt-MHC-Hap within a sample. As it was not assigned to the PGF-Alt category, this indicates that the PGF haplotype is absent in this sample.
- **Recovery Peak**: a list of “PGF consensus peaks” is created. For a peak to be included in such a list it must overlap with at least 8 other PGF-Peaks located in other samples. This might imply a presence of a polymorphic site that is better described in an Alt-MHC-Hap for a given sample
- **Insertion Peak**: an Alt-Peak is localized in a genetic location absent from PGF and with a minimum length of 150 bp. This is defined by the alignment of the Alt-MHC-Hap in question with PGF.
- **Low Recovery Peak**: same as the Recovery Peak classification, except that the threshold to be considered as a consensus peak is greater or equal to one sample and less than eight samples.
- **Absent from PGF**: the Alt-Peak does not overlap with a PGF-Peak called in any of the samples analysed. This could implicate highly polymorphic sequence to the PGF haplotype, introducing a binding site.

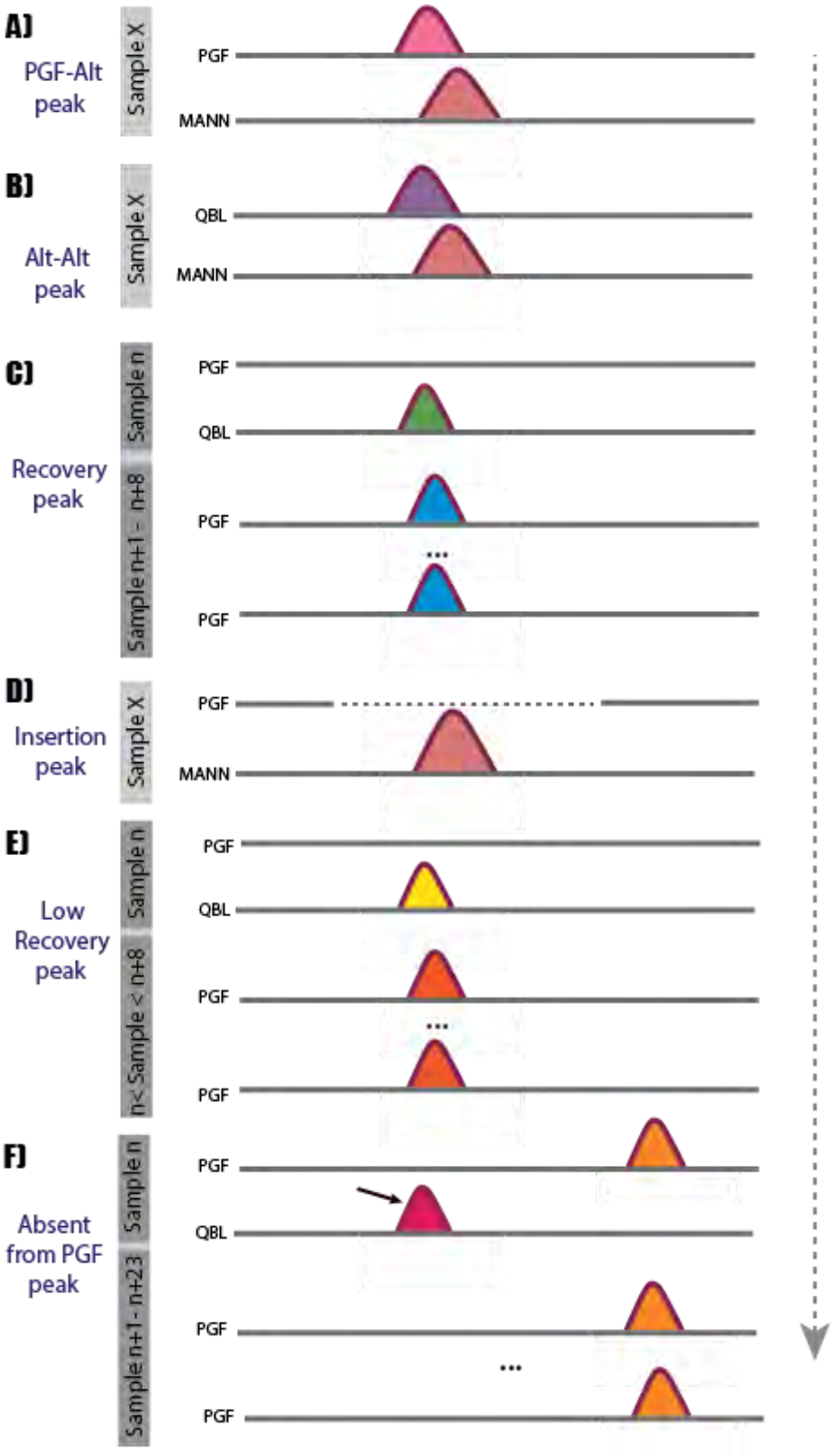

Figure 9. Alt-Peak categorization. Detailed description in the main text.

IX.1.3 Method: ATAC-seq peak annotation

To obtain a further understanding of the open chromatin landscape within the MHC we designed a gene assignment/annotation framework for the ATAC-seq peaks. This annotation algorithm will designate to a peak all genes that overlap or are downstream an ATAC-seq peak within 6 kb; if no gene was assigned to a peak in this first step, then the nearest gene is assigned to it (figure 10C). The ChIPpeakAnno package (Zhu *et al.* 2010) within the R software was used to make the annotation. The annotation is made in an alternative loci specific manner based on the genetic features included in the release 25 of GENCODE (Harrow *et al.* 2012). Finally, each gene was weighted by the frequency of its assignment to ATAC-seq Alt-Peaks and represented as a word cloud diagram.

IX.1.3 Result: ATAC-seq peak annotation

The assigned genes are shown in a word cloud graphic in figure 10B, the most assigned gene is *HLA-DRB1* followed by *TRIM26* and *MICD*; furthermore, genes absent from PGF make also an appearance, *e.g. HLA-DRB3 and HLA-DRB4*. Importantly, all Alt-Peaks have a sequence bearing polymorphisms when compared to PGF (section XIII); thus, it is not surprising that genes located in highly polymorphic regions are assigned to a peak, *e.g.* the HLA genes.

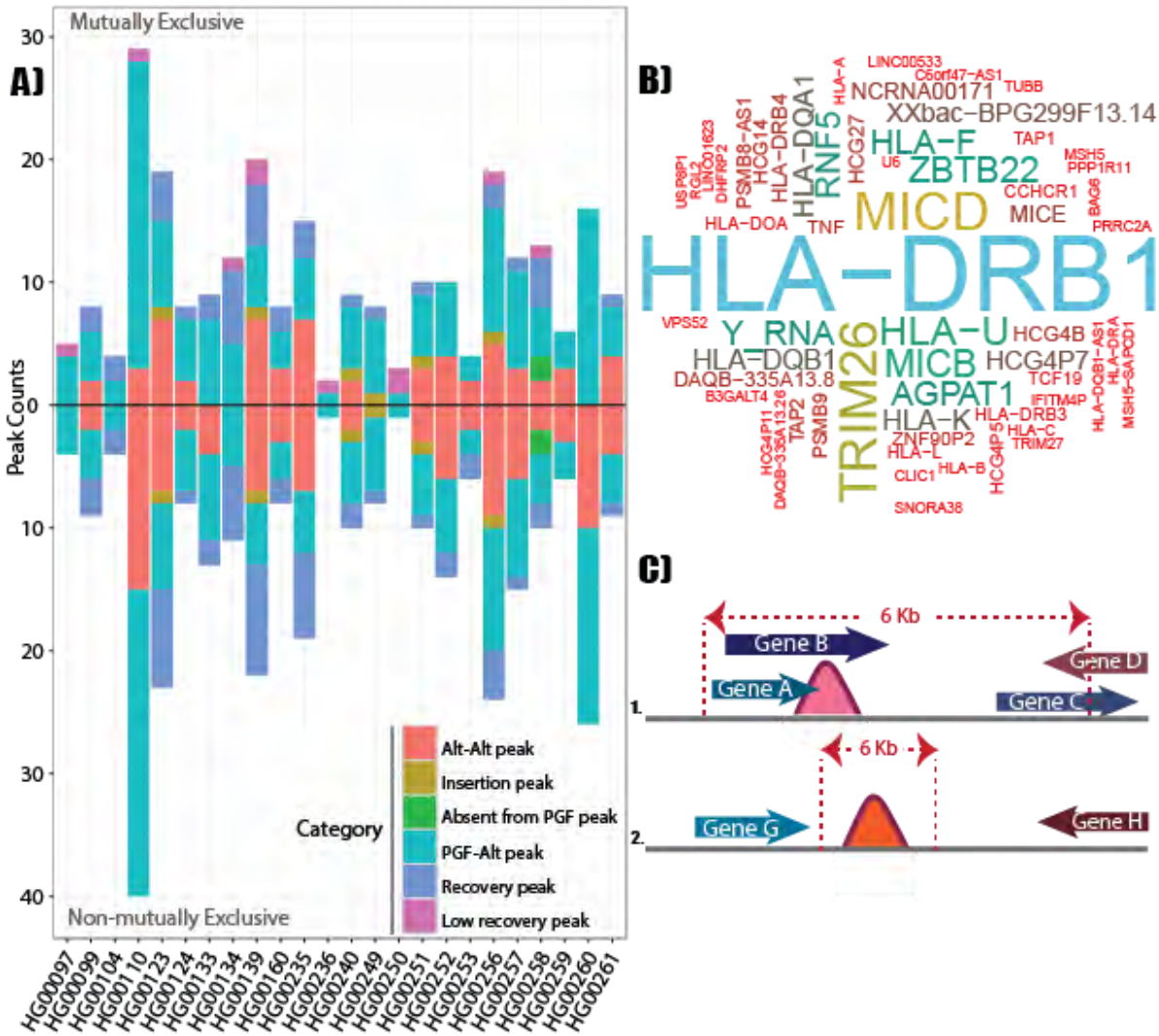

Figure 10. A) Number of Alt-Peaks that fall within each category; the amount of genes vary between samples, as it depends on the genetic divergence score (section V.1) and to the availability of alternative loci. The plot on the top assigns to each peak one category, as defined in figure 9; each bar sums up to the total of Alt-Peaks found in each sample. The plot on the bottom depicts the number of Alt-Peaks per category when one peak is allowed to have more than one category. For the latter the "Low recovery peak" category was not made. B) Genes associated with Alt-Peaks, colour and font size correspond to the frequency of association. C) Framework for peak annotation. 1. All genes upstream and overlapping a peak within 6 kb are annotated to a given peak; in the example above genes A,B, and C are assigned to the pink peak. 2. If a peak had no gene assigned in 1. the closest gene to the peak is annotated for the peak in question; in this case gene G is assigned to the orange peak.

IX.2 Analysing Haplotype-Specific Effects in the MHC Region

IX.2.1 Result: Functional interpretation of open chromatin features absent from PGF

There are six ATAC-seq peaks among all samples falling into sequences not present in PGF (figure 10A), to address their functionality we assessed which genes were annotated to each of them. Out of the six peaks, only four were assigned a gene in less than 60 bp, for these cases *HLA-DRB4* a functional DRB gene is annotated. The remaining two peaks are annotated to the pseudogene *RNU1-79P* which is near *HLA-DRB3*, both absent from PGF. Table 1 gives further information in respect of these cases.

ENSEMBL Gene ID	Haplotype	Start	End	Peak Length	Sample	Distance to Gene	Gene Symbol
ENSG00000227826	MANN	3855323	3855531	209	HG00139	-40	*HLA-DRB4*
ENSG00000227826	MANN	3855322	3855540	219	HG00249	-44	*HLA-DRB4*
ENSG00000227826	MANN	3855348	3855532	185	HG00123	-53	*HLA-DRB4*
ENSG00000227826	MANN	3855359	3855519	161	HG00256	-52	*HLA-DRB4*
ENSG00000265223	COX	3923945	3924121	177	HG00240	7311	*RNU1-79P*
ENSG00000265223	COX	3924000	3924176	177	HG00251	7256	*RNU1-79P*

Table 1. Description of peaks that are categorised as Insertion peaks. The Alt-MHC-Hap from which they were called is described, as well as the coordinated they fall into.

ENSEMBL Gene ID	Haplotype	Start	End	Start in PGF	End in PGF	Sample	Distance to Gene	Gene Symbol
ENSG00000225691	MCF	2616092	2616237	31273378	31273523	HG00258	-1315	*HLA-C*
ENSG00000226050	MCF	2616092	2616237	31273378	31273523	HG00258	-2113	*USP8P1*
ENSG00000229074	MANN	4017382	4017527	32604619	32604763	HG00258	5589	*HLA-DRB1*

Table 2. Description of peaks that are categorised as absent from PGF. The Alt-MHC-Hap from which they were called is described. The coordinates in the Alt-MHC-Hap from which the peak was called, as well as the position they were lifted to in PGF.

There are two peaks in the category of peaks “not present in PGF”, both coming from the same sample. One of these peaks is annotated with *HLA-C* and *USP8P1*, the other is annotated with *HLA-DRB1*; however, these genes are annotated by a PGF-Peak in some samples which means that these open chromatin features are not exclusive to the MHC-Alt-Hap; furthermore they have been reported as enhancers. In *HLA-C*, the difference between the Alt-Peak and PGF-Peak lies on the distance to the annotated gene, as shown by table S1. For example, while the PGF-Peaks lie in a distance minor to -130 bp, the Alt-Peak is -1315 bp of distance to HLA-C. It remains to be tested if these enhancers differ in genetic sequence, if so they may house different binding sites for TFs and may have functional consequence at transcriptomic level. On the other hand, some PGF-Peaks annotate *HLA-DRB1* to the same distance as the reported Alt-Peak, this might be to the presence of CNVs that differ in both haplotypes.

Furthermore, an estimate of the expression of *HLA-DRB4* was assessed using the same method as described in the section V1.2.1. As previously stated, *HLA-DRB4* is a gene absent from PGF, and it is in proximity of an “Insertion peak” located in the MANN haplotype in 4 of 24 samples; therefore, we hypothesized that the presence of this open chromatin feature will be an indicator of the expression of *HLA-DRB4*. Figure 11 shows the TPMs recovered for the *HLA-DRB4* gene present in different alternative loci of MHC, plus a control gene *MS4A1* located in chromosome 11; the *MS4A1* gene is not highly polymorphic, located outside the MHC region and according to the GTEx portal it is only expressed in LCLs among all the cell types they queried (GTEx Consortium 2015). It can be noted that all the samples which had the insertion peak express *HLA-DRB4*, and nine out of the twenty-four samples do not express *HLA-DRB4* at all. The remaining eleven samples do express *HLA-DRB4* and do not have a peak annotated to *HLA-DRB4*. Based on the TPMs it appears that the *HLA-DRB4* expressed for the latter samples is more

similar to the allele present in the MCF and SSTO haplotypes; nevertheless the reasons why there is no “Insertion Peak” for this samples remains unknown.

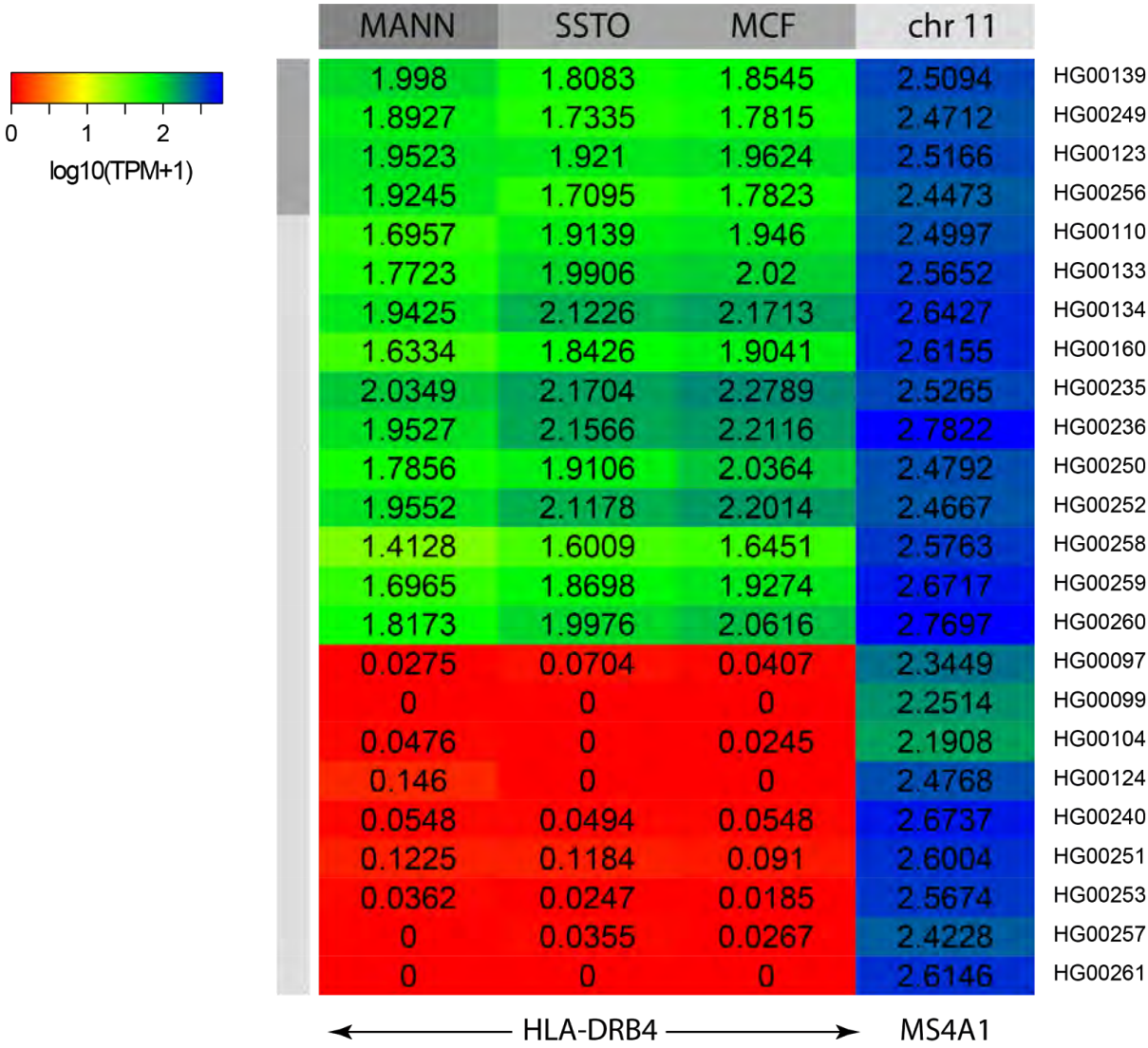

Figure 11. Heatmap representation for the gene *HLA-DRB4*, located between *HLA-DRB9* and *HLA-DRB1*. Each column represents a different haplotype analysed; the haplotypes MANN, SSTO, and MCF were selected given that all of them have a functional *HLA-DRB4* gene. The shaded rows in the left-most bar correspond to those samples in which the Insertion peak was called.

IX.2.2 Method: Functional effects of the alternative peaks in the MHC in non-HLA genes

Obtaining a quantitative measure of the expression of the highly polymorphic HLA genes without an adequate experimental setup cannot be achieved nowadays with the NGS analyses available to date. The biggest obstacles for an accurate gene

expression quantification are paralogy, polymorphism, and ploidy of the human genome. There are two alleles per gene present in a given sample, overestimation of TPMs will be highly coupled to resemblance of both alleles, the more alike the higher the probability of a read being shared between alleles. Therefore, we sought haplotypic differences at transcriptomic levels in non-HLA genes; eliminating the biases of high genetic divergence from PGF, paralogy, and ploidy.

We used the RSEM software (B. Li and Dewey 2011) to quantify gene expression of features included on the reference chromosome as annotated by GENCODE 25 (Harrow *et al*. 2012). All the genes that were annotated by an Alt-peak were included in analyses only if they were expressed. A gene was defined as expressed if they had a TPM value above 75; eight genes passed these thresholds. A sample and its correspondent TPM value for a given gene was classified according to the alternative loci where the Alt-Peak was identified; furthermore, a classification was made if at least two samples fell in it, otherwise, it was discarded from further analysis.

IX.2.2 Result: Functional effects of the haplotypic structure of the MHC in non-HLA genes

TRIM26, localised in MHC class I region , was the only gene that showed a statistically significant differential expression between haplotypes (pval <0.009, with a Kruskal-Wallis test). It is noteworthy that most of the genes tested had a small sample size which hindered the making of more inferences (figure 12). *TRIM26* has shown to be linked to schizophrenia, along with *RNF5* and *HLA-DRB3* (de Jong *et al.* 2012); moreover, *TRIM26* has shown to have a critical role in the regulation of the innate immune response (Wang *et al*. 2015; Ran *et al.* 2016).

In brief, the analytical framework proposed is able to retrieve haplotype specific events such as open chromatin and transcript abundance (of not highly polymorphic genes) in alternative loci. No genotype data is required for the analysis,

thus, providing an advantage to current proposes available. Further, the same methodology can be applied to query covalent modification of DNA and histones.

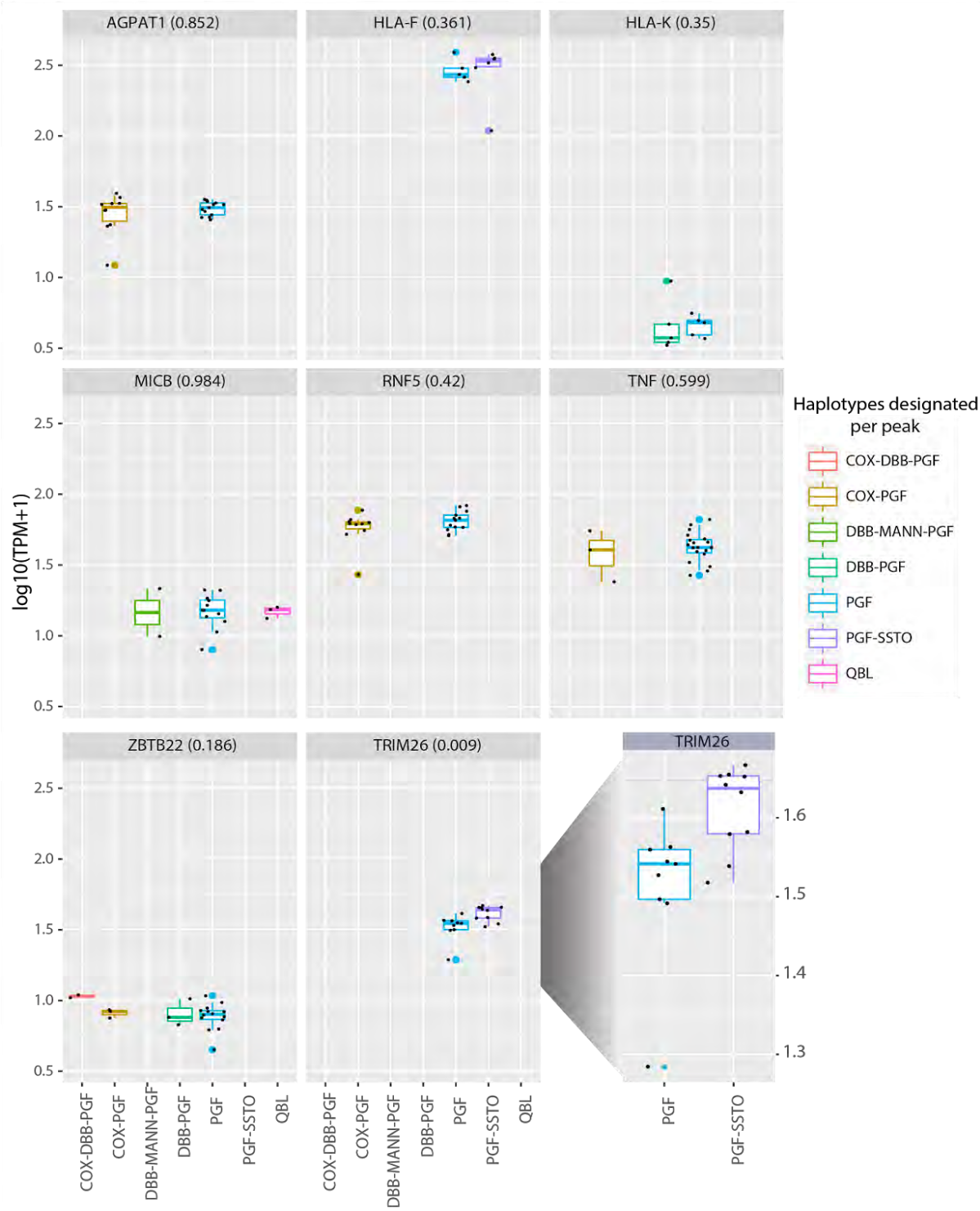

Figure 12. In the y-axis the TPM values are plotted for each of the expressed genes that were annotated by an alternative peak. Each TPM value was categorised, in the x-axis, accordingly to the haplotypes from which the peaks associated to the gene were called. Differential expression of a gene was tested between haplotypes, p-values are indicated between parenthesis next to the gene symbol.

X. CONCLUSIONS AND PERSPECTIVES

Throughout this thesis we characterised the usage of the variation-aware aligner, Bwakit, for querying one of the most polymorphic genomic locations in the human genome: the MHC region. The main goal was to uncover open-chromatin features not represented in the primary reference assembly and assess their impact on the gene expression profiles in 24 lymphoblastoid cell lines.

First, we familiarised the reader with the current status of data availability for the MHC region, while highlighting the importance, extent, and consequences of polymorphism in this locus. We assessed the dissimilarities between eight MHC haplotypes by comparing the gene sets annotated by GENCODE in each of them. Even though the highest quantity of differing genes are non-coding, the integration of them is of vital importance as they can impact and produce bias and misinterpretation. For example, (Dilthey *et al.* 2016) used population reference graph to type HLA alleles; they noted that the omission of the *HLA-DRB5* gene from their analyses resulted in the misalignment of reads to *HLA-DRB1*. The assembly and annotation of more population inclusive haplotypes is urged to get a better understanding of variation within the MHC region. Likewise, The Simons Genome Diversity Project assembled genomes from 142 populations from which they retrieved several alternative loci, most of them pertaining to the MHC and IgH locus (Mallick *et al*. 2016). Third generation sequencing technologies will play a pivotal role in such ambitious task; with the usage of long reads, greater accuracy in assembly of haplotypes and HLA allele typing are expected to benefit of these technologies (Carapito, Radosavljevic, and Bahram 2016).

Knowing that the biological context differs between haplotypes, we implemented into our pipeline a variation aware aligner that allowed us to retrieve such differences and include them in our analysis, and even integrate it with other software implemented for linear haplotypes. The pipeline herein referred to is

suitable for ATAC-seq and for ChIP-seq, although data from the latter technique was not presented. A remarkable advantage of this pipeline presented is that no genotype data is required; however, it is best suited for Caucasian individuals. Additionally, we evaluated the biases found in the read mapping and in the peak-calling downstream step. In general, reads were aligned only to the alternative haplotypes of the MHC when the location mapped underlied polymorphic sites. In terms of peak quality, obtained from MACS2 software, the peaks called at the alternative haplotypes of MHC were of lesser quality, nevertheless, they did not skewed towards quality control thresholds implemented.

Next, we assigned a putative functionality to each peak called in the alternative haplotypes of the MHC, given that each peak represents an open-chromatin feature and, therefore, may have a role in gene expression. Among the 24 samples, we called six peaks that fell into a sequence absent from PGF, “Insertion Peaks”; these peaks located in a known indel between *HLA-DRB9* and *HLA-DRB*1. These six peaks could be divided into two groups: 1) peak associated with *HLA-DRB4* and 2) peak associated with *HLA-DRB3*. Out of the 24 samples, four had the peak associated with *HLA-DRB4* and we could recover positive gene expression metrics for *HLA-DRB4* in these samples. Next we assessed if there were any differences between gene expression values amongst genes annotated with peaks falling into different haplotypes; for the eight genes queried, *TRIM26*, showed differential expression between haplotypes. This last analysis was limited by the small sample size analysed and restricted to genes low in polymorphism; even though the HLA genes were the most assigned to an alternative peak. To overcome the constraints of querying accurately the expression of polymorphic HLA genes, data sets extracted from homozygous cell lines with known types for all HLA genes and pseudogenes could be sequenced with third generation machines.

All together, we encourage the development of resources, in the form of data and software, to appropriately investigate the MHC region, without disregarding the rich biological context it hosts; valuable analyses could benefit ranging from

evolution and population genomics, to studying complex traits in association studies. After all the main question is: amongst all the variation, which is functional.

XI. REFERENCES

Aken, Bronwen L., Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, *et al*. 2016. "The Ensembl Gene Annotation System." *Database: The Journal of Biological Databases and Curation* 2016 (June). doi:10.1093/database/baw093.

Auton, Adam, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2013. "GenBank." *Nucleic Acids Research* 41 (Database issue): D36–42.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.

Brandt, Débora Y. C., Vitor R. C. Aguiar, Bárbara D. Bitarello, Kelly Nunes, Jérôme Goudet, and Diogo Meyer. 2015. "Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data." *G3* 5 (5): 931–41.

Buchkovich, Martin L., Karl Eklund, Qing Duan, Yun Li, Karen L. Mohlke, and Terrence S. Furey. 2015. "Removing Reference Mapping Biases Using Limited or No Genotype Data Identifies Allelic Differences in Protein Binding at Disease-Associated Loci." *BMC Medical Genomics* 8 (July): 43.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods* 10 (12): 1213–18.

Carapito, Raphael, Mirjana Radosavljevic, and Seiamak Bahram. 2016. "Next-Generation Sequencing of the HLA Locus: Methods and Impacts on HLA Typing, Population Genetics and Disease Association Studies." *Human Immunology*, April. doi:10.1016/j.humimm.2016.04.002.

Chung, Erwin K., Yan Yang, Robert M. Rennebohm, Marja-Liisa Lokki, Gloria C. Higgins, Karla N. Jones, Bi Zhou, Carol A. Blanchong, and C. Yung Yu. 2002. "Genetic Sophistication of Human Complement Components C4A and C4B and RP-C4-CYP21-TNX (RCCX) Modules in the Major Histocompatibility Complex." *American Journal of Human Genetics* 71 (4): 823–37.

Church, Deanna M., Valerie A. Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, *et al.* 2011. "Modernizing Reference

Genome Assemblies." *PLoS Biology* 9 (7): e1001091.

Church, Deanna M., Valerie A. Schneider, Karyn Meltz Steinberg, Michael C. Schatz, Aaron R. Quinlan, Chen-Shan Chin, Paul A. Kitts, *et al.* 2015. "Extending Reference Assembly Models." *Genome Biology* 16 (January): 13.

Deakin, Janine E., Anthony T. Papenfuss, Katherine Belov, Joseph G. R. Cross, Penny Coggill, Sophie Palmer, Sarah Sims, Terence P. Speed, Stephan Beck, and Jennifer A. Marshall Graves. 2006. "Evolution and Comparative Analysis of the MHC Class III Inflammatory Region." *BMC Genomics* 7 (November): 281.

Dilthey, Alexander T., Pierre-Antoine Gourraud, Alexander J. Mentzer, Nezih Cereb, Zamin Iqbal, and Gil McVean. 2016. "High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs." *PLoS Computational Biology* 12 (10): e1005151.

Gourraud, Pierre-Antoine, Pouya Khankhanian, Nezih Cereb, Soo Young Yang, Michael Feolo, Martin Maiers, John D. Rioux, Stephen Hauser, and Jorge Oksenberg. 2014. "HLA Diversity in the 1000 Genomes Dataset." *PloS One* 9 (7): e97282.

GTEx Consortium. 2015. "Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." *Science* 348 (6235): 648–60.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, *et al*. 2012. "GENCODE: The Reference Human Genome Annotation for The ENCODE Project." *Genome Research* 22 (9): 1760–74.

Horton, Roger, Richard Gibson, Penny Coggill, Marcos Miretti, Richard J. Allcock, Jeff Almeida, Simon Forbes, *et al*. 2008. "Variation Analysis and Gene Annotation of Eight MHC Haplotypes: The MHC Haplotype Project." *Immunogenetics* 60 (1): 1–18.

Huang, Lin, Victoria Popic, and Serafim Batzoglou. 2013. "Short Read Alignment with Populations of Genomes." *Bioinformatics* 29 (13): i361–70.

Jong, Simone de, Kristel R. van Eijk, Dave W. L. H. Zeegers, Eric Strengman, Esther Janson, Jan H. Veldink, Leonard H. van den Berg, *et al*. 2012. "Expression QTL Analysis of Top Loci from GWAS Meta-Analysis Highlights Additional Schizophrenia Candidate Genes." *European Journal of Human Genetics: EJHG* 20 (9): 1004–8.

Kellis, Manolis, Barbara Wold, Michael P. Snyder, Bradley E. Bernstein, Anshul Kundaje, Georgi K. Marinov, Lucas D. Ward, *et al.* 2014. "Defining Functional DNA Elements in the Human Genome." *Proceedings of the National Academy of Sciences of the United States of America* 111 (17): 6131–38.

Kumasaka, Natsuhiko, Andrew J. Knights, and Daniel J. Gaffney. 2016. "Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq." *Nature Genetics* 48 (2): 206–13.
Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, *et al*. 2013. "Transcriptome and Genome

Sequencing Uncovers Functional Variation in Humans." *Nature* 501 (7468): 506–11.

Li, Heng. 2016. "lh3/bwa." *GitHub*. Accessed June 16. https://github.com/lh3/bwa.

Li, Bo, and Colin N. Dewey. 2011. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome." *BMC Bioinformatics* 12 (August): 323.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Li, Ruiqiang, Yingrui Li, Hancheng Zheng, Ruibang Luo, Hongmei Zhu, Qibin Li, Wubin Qian, *et al.* 2010. "Building the Sequence Map of the Human Pan-Genome." *Nature Biotechnology* 28 (1): 57–63.

Lunter, Gerton, and Martin Goodson. 2011. "Stampy: A Statistical Algorithm for Sensitive and Fast Mapping of Illumina Sequence Reads." *Genome Research* 21 (6): 936–39.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, *et al.* 2016. "The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations." *Nature* 538 (7624): 201–6.

Marsh, Steven G. E., Peter Parham, and Linda D. Barber. 2000. *The HLA Facts Book*. Elsevier.

Norman, Paul J., Steve J. Norberg, Neda Nemat-Gorgani, Thomas Royce, Jill A. Hollenbach, Melissa Shults Won, Lisbeth A. Guethlein, Kevin L. Gunderson, Mostafa Ronaghi, and Peter Parham. 2015. "Very Long Haplotype Tracts Characterized at High Resolution from HLA Homozygous Cell Lines." *Immunogenetics* 67 (9): 479–85.

Pei, Baikang, Cristina Sisu, Adam Frankish, Cédric Howald, Lukas Habegger, Xinmeng Jasmine Mu, Rachel Harte, *et al.* 2012. "The GENCODE Pseudogene Resource." *Genome Biology* 13 (9): R51.

Ran, Yong, Jing Zhang, Li-Li Liu, Zhao-Yi Pan, Ying Nie, Hong-Yan Zhang, and Yan-Yi Wang. 2016. "Autoubiquitination of TRIM26 Links TBK1 to NEMO in RLR-Mediated Innate Antiviral Immune Response." *Journal of Molecular Cell Biology* 8 (1): 31–43.

Robinson, James, Jason A. Halliwell, James D. Hayhurst, Paul Flicek, Peter Parham, and Steven G. E. Marsh. 2015. "The IPD and IMGT/HLA Database: Allele Variant Databases." *Nucleic Acids Research* 43 (Database issue): D423–31.

Tough, David F., Paul P. Tak, Alexander Tarakhovsky, and Rab K. Prinjha. 2016. "Epigenetic Drug Discovery: Breaking through the Immune Barrier." *Nature Reviews. Drug Discovery*, October. doi:10.1038/nrd.2016.185.

Traherne, James A., Roger Horton, Anne N. Roberts, Marcos M. Miretti, Matthew E. Hurles, C. Andrew Stewart, Jennifer L. Ashurst, *et al*. 2006. “Genetic Analysis of Completely Sequenced Disease-Associated MHC Haplotypes Identifies Shuffling of Segments in Recent Human History.” *PLoS Genetics* 2 (1): e9.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. 2012. “Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks.” *Nature Protocols* 7 (3): 562–78.

Trowsdale, John, and Julian C. Knight. 2013. “Major Histocompatibility Complex Genomics and Human Disease.” *Annual Review of Genomics and Human Genetics* 14 (July): 301–23.

Vandiedonck, Claire, Martin S. Taylor, Helen E. Lockstone, Katharine Plant, Jennifer M. Taylor, Caroline Durrant, John Broxholme, Benjamin P. Fairfax, and Julian C. Knight. 2011. “Pervasive Haplotypic Variation in the Spliceo-Transcriptome of the Human Major Histocompatibility Complex.” *Genome Research* 21 (7): 1042–54.

Wang, Peng, Wei Zhao, Kai Zhao, Lei Zhang, and Chengjiang Gao. 2015. “TRIM26 Negatively Regulates Interferon-β Production and Antiviral Response through Polyubiquitination and Degradation of Nuclear IRF3.” *PLoS Pathogens* 11 (3): e1004726.

Wu, Thomas D., and Serban Nacu. 2010. “Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads.” *Bioinformatics* 26 (7): 873–81.

Yau, Anthony C. Y., Jonatan Tuncel, Sabrina Haag, Ulrika Norin, Miranda Houtman, Leonid Padyukov, and Rikard Holmdahl. 2016. “Conserved 33-Kb Haplotype in the MHC Class III Region Regulates Chronic Arthritis.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (26): E3716–24.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, *et al.* 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology* 9 (9): R137.

Zhu, Lihua J., Claude Gazin, Nathan D. Lawson, Hervé Pagès, Simon M. Lin, David S. Lapointe, and Michael R. Green. 2010. “ChIPpeakAnno: A Bioconductor Package to Annotate ChIP-Seq and ChIP-Chip Data.” *BMC Bioinformatics* 11 (May): 237.

XII. SUPPLEMENTARY MATERIAL

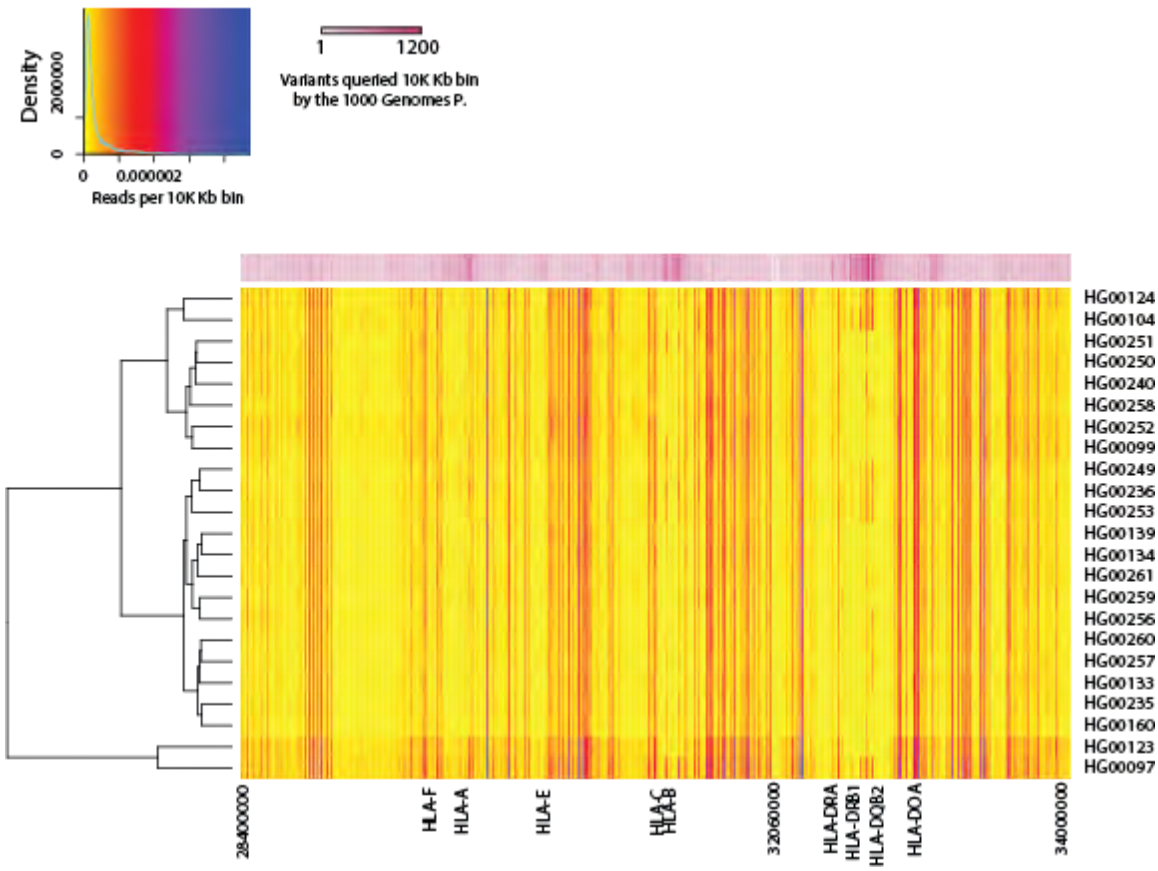

Figure S1. Heatmap of reads (normalised by library size) per 10kb bin. Only reads aligning to PGF with a mapQ greater than 29 are shown. Upper bar shows the quantity of variants queried by the 1000 Genomes Project in each bin.

GeneID	Haplotype	Start	End	Sample	Distance to Gene	Gene Symbol
ENSG00000225691	MCF	2616092	2616237	HG00258	-1315	HLA-C
ENSG00000204525	PGF	31272003	31272511	HG00253	-127	HLA-C
ENSG00000204525	PGF	31271993	31272466	HG00249	-100	HLA-C
ENSG00000204525	PGF	31272101	31272327	HG00139	-84	HLA-C
ENSG00000204525	PGF	31272125	31272302	HG00240	-84	HLA-C
ENSG00000204525	PGF	31272099	31272313	HG00133	-76	HLA-C
ENSG00000204525	PGF	31272083	31272302	HG00260	-62	HLA-C
ENSG00000204525	PGF	31272100	31272274	HG00134	-57	HLA-C
ENSG00000204525	PGF	31272101	31272270	HG00110	-56	HLA-C
ENSG00000204525	PGF	31272004	31272361	HG00257	-52	HLA-C
ENSG00000204525	PGF	31272066	31272298	HG00123	-52	HLA-C
ENSG00000204525	PGF	31272002	31272309	HG00160	-26	HLA-C
ENSG00000204525	PGF	31271884	31272418	HG00097	-21	HLA-C
ENSG00000204525	PGF	31271945	31272343	HG00235	-14	HLA-C
ENSG00000204525	PGF	31271933	31272344	HG00236	-8	HLA-C
ENSG00000204525	PGF	31271866	31272376	HG00251	9	HLA-C
ENSG00000204525	PGF	31271735	31272454	HG00099	36	HLA-C

Table S1.1. Peaks annotating gene *HLA-C,* including the Alt-Peak categorised as “Absent from PGF”

GeneID	Haplotype	Start	End	Sample	Distance to Gene	Gene Symbol
ENSG00000196126	PGF	32589693	32590266	HG00104	-132	HLA-DRB1
ENSG00000196126	PGF	32589775	32590068	HG00260	-74	HLA-DRB1
ENSG00000206240	COX	4028582	4028755	HG00259	-16	HLA-DRB1
ENSG00000196126	PGF	32589701	32589994	HG00124	0	HLA-DRB1
ENSG00000196126	PGF	32589631	32589911	HG00097	77	HLA-DRB1
ENSG00000196126	PGF	32589661	32589867	HG00249	84	HLA-DRB1

ENSG00000206240	COX	4028401	4028732	HG00252	86	HLA-DRB1
ENSG00000196126	PGF	32589650	32589863	HG00253	92	HLA-DRB1
ENSG00000229074	MANN	4022828	4023073	HG00252	93	HLA-DRB1
ENSG00000229074	MANN	4022804	4023060	HG00139	111	HLA-DRB1
ENSG00000206306	QBL	3809330	3809677	HG00099	117	HLA-DRB1
ENSG00000206240	COX	4028366	4028693	HG00099	122	HLA-DRB1
ENSG00000206306	QBL	3809258	3809732	HG00261	126	HLA-DRB1
ENSG00000228080	SSTO	4012047	4012390	HG00236	134	HLA-DRB1
ENSG00000196126	PGF	32589561	32589857	HG00236	139	HLA-DRB1
ENSG00000228080	SSTO	4012035	4012384	HG00160	142	HLA-DRB1
ENSG00000206306	QBL	3809307	3809645	HG00251	145	HLA-DRB1
ENSG00000229074	MANN	4022719	4023048	HG00110	159	HLA-DRB1
ENSG00000206306	QBL	3809284	3809638	HG00240	160	HLA-DRB1
ENSG00000229074	MANN	4022792	4022950	HG00256	172	HLA-DRB1
ENSG00000229074	MANN	4022596	4023132	HG00123	179	HLA-DRB1
ENSG00000206240	COX	4028376	4028555	HG00235	186	HLA-DRB1
ENSG00000228080	SSTO	4011948	4012368	HG00260	194	HLA-DRB1
ENSG00000236884	DBB	3848238	3848661	HG00256	195	HLA-DRB1
ENSG00000206240	COX	4028254	4028657	HG00251	196	HLA-DRB1
ENSG00000206306	QBL	3809168	3809682	HG00235	196	HLA-DRB1
ENSG00000206306	QBL	3809212	3809632	HG00259	199	HLA-DRB1
ENSG00000206240	COX	4028272	4028630	HG00261	201	HLA-DRB1
ENSG00000228080	SSTO	4011933	4012358	HG00250	206	HLA-DRB1
ENSG00000236884	DBB	3848193	3848679	HG00123	209	HLA-DRB1
ENSG00000236884	DBB	3848279	3848591	HG00110	210	HLA-DRB1
ENSG00000228080	SSTO	4011819	4012457	HG00110	214	HLA-DRB1
ENSG00000228080	SSTO	4011873	4012394	HG00258	218	HLA-DRB1
ENSG00000206240	COX	4028239	4028627	HG00240	219	HLA-DRB1
ENSG00000228080	SSTO	4012049	4012216	HG00123	220	HLA-DRB1

ENSG00000228080	SSTO	4011909	4012352	HG00256	222	HLA-DRB1
ENSG00000236884	DBB	3848219	3848625	HG00139	223	HLA-DRB1
ENSG00000206240	COX	4028300	4028537	HG00257	234	HLA-DRB1
ENSG00000228080	SSTO	4011801	4012426	HG00134	238	HLA-DRB1
ENSG00000229074	MANN	4022672	4022936	HG00160	239	HLA-DRB1
ENSG00000228080	SSTO	4011937	4012280	HG00259	244	HLA-DRB1
ENSG00000236884	DBB	3848229	3848558	HG00160	251	HLA-DRB1
ENSG00000206306	QBL	3809228	3809502	HG00258	256	HLA-DRB1
ENSG00000206306	QBL	3809259	3809425	HG00124	279	HLA-DRB1
ENSG00000228080	SSTO	4011966	4012174	HG00257	282	HLA-DRB1
ENSG00000228080	SSTO	4011923	4012119	HG00097	331	HLA-DRB1
ENSG00000228080	SSTO	4011880	4012130	HG00251	347	HLA-DRB1
ENSG00000228080	SSTO	4011916	4012090	HG00253	349	HLA-DRB1
ENSG00000228080	SSTO	4011876	4012109	HG00235	360	HLA-DRB1
ENSG00000206306	QBL	3809156	3809347	HG00257	369	HLA-DRB1
ENSG00000196126	chr6	32584061	32584607	HG00097	5514	HLA-DRB1
ENSG00000229074	MANN	4017382	4017527	HG00258	5589	HLA-DRB1
ENSG00000196126	PGF	32583911	32584573	HG00124	5606	HLA-DRB1
ENSG00000196126	PGF	32584011	32584462	HG00236	5612	HLA-DRB1
ENSG00000196126	PGF	32583834	32584601	HG00104	5630	HLA-DRB1
ENSG00000196126	PGF	32583997	32584204	HG00249	5748	HLA-DRB1
ENSG00000196126	PGF	32583868	32584136	HG00253	5846	HLA-DRB1

Table S1.2 Peaks annotating gene *HLA-DRB1,* including the Alt-Peak categorised as “Absent from PGF”