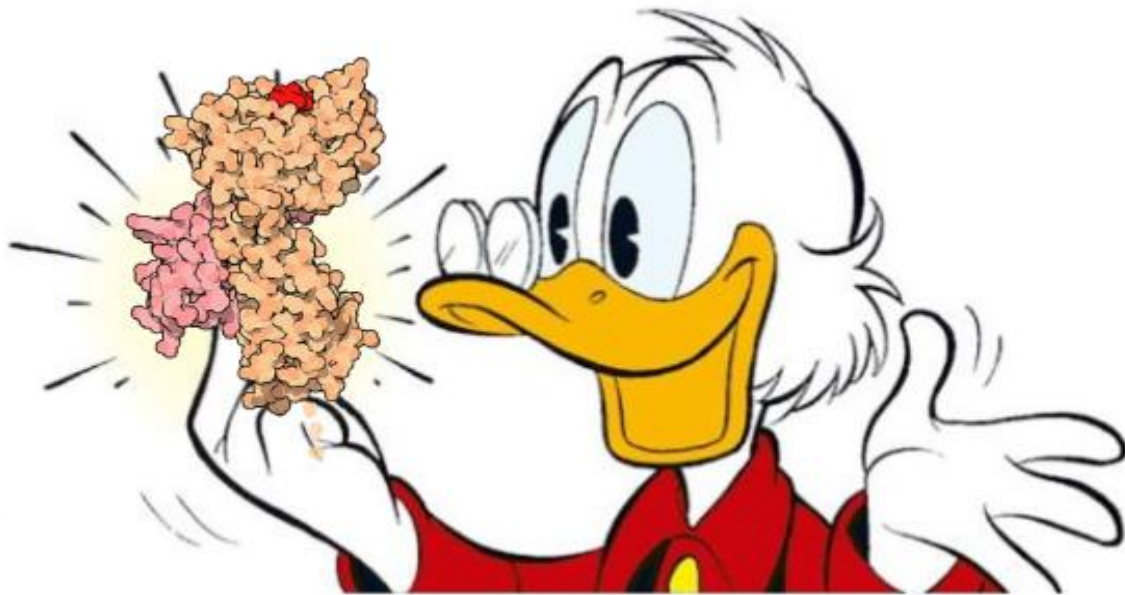


Identifying novel MHC pathway genes using a naïve Bayesian classifier



By: Dieter Gerrit Gijsbert Stoker
Student number: 4159853
Supervisor: Dr. Can Kesmir
Daily supervisor: Dr. T.J.P. van Dam
Second reviewer: Prof. Dr. R.J. de Boer



Abstract

Major histocompatibility complex (MHC) proteins present peptides at the cell surface and are at the heart of immunity. T cells constantly sample the expressed peptide repertoire, and can become activated upon recognition of non-self peptide-MHC pairs cognate to their TCR. Activated T cells proliferate and are an important part of the adaptive immune response against pathogens. Natural killer cells (NK cells) kill cells expressing aberrant amounts of MHC molecules. Unsurprisingly, MHC molecules are of utmost importance in autoimmunity, disease outcomes, and cancer therapy. Through decades of research, transcription factors (TFs), chaperones, peptide-loading complexes, and proteases involved in the MHC pathway have been discovered. Nevertheless, new proteins involved in this pathway have recently been discovered, and a systems understanding of the MHC pathway is lacking. In this manuscript, we shed light on this outstanding issue by integrating five different data sets using a naïve Bayesian classifier. Additionally, we uncover candidate TFs that might be important in regulation of MHC genes. We scored all human protein-coding genes on their similarity to 86 known MHC pathway genes in these data sets, and thereby generated an integrated MHC pathway-specific score. This score is better at identifying true MHC pathway genes than any separate data set, and pinpoints candidate genes for experimental follow-up.

Layman's summary

The human body is constantly challenged by pathogens. The immune system is responsible for recognising invaders and mounting appropriate immune responses. But how are pathogens recognised? An important part of the recognition machinery are the MHC (major histocompatibility complex) molecules. They are divided into two classes. MHC I molecules are present on every cell in the human body. They have slots that can be loaded with pieces of protein. Proteins are the workhorses of chemical reactions in the cell, and are destroyed when damaged or no longer necessary. In the process, small pieces of proteins, peptides, are created, and these are loaded by MHC I molecules. The MHC I molecules then move to the cell surface and present these peptides. There, immune cells that pass by sample these peptide-MHC pairs, looking for anything out of the ordinary. They have been trained not to respond to peptides that your cells normally present: 'self' peptides. Therefore, a healthy cell, which only presents self-peptides, will not produce an immune response. Upon viral infection, viruses create viral proteins in a cell. These, too, are broken down into peptides, and presented at the cell surface by MHC I. When an immune cell samples the MHC I molecule loaded with viral peptide, it recognises that this is non-self. This signals infection, and an immune response is mounted. The system is not limited to pathogens. Cancer cells carry many mutations, some of which lead to proteins that are different from normal self-proteins. Peptides from these proteins are presented, sampled by immune cells, and recognised as non-self. In this way, an immune response against cancer cells can be mounted.

MHC II molecules are different because they are only present on immune cells. Their focus is on presenting peptides that these cells take up from the environment. When your skin is scratched, bacteria can enter the wound. This causes inflammation. Special antigen-presenting immune cells move to the inflamed area, and take up bacterial peptides. These are presented on MHC II molecules. The antigen-presenting cells then present these loaded molecules to other immune cells in order to activate them and create immune responses that counter the invading bacteria.

Besides the MHC molecules themselves, many more genes are involved. They make sure the MHC molecules are loaded properly, folded properly, and are transported to the cell surface. Though MHC and related genes are at the core of immunity, many genes involved in this pathway are still unknown. To improve vaccination, anti-cancer therapies, and our understanding of disease and auto-immunity, it is crucial that we completely understand the MHC pathway genes. In this work, we integrated knowledge from five different published data sets, and used that knowledge to predict, for all human genes, whether they are likely to be involved in the MHC pathway. These predictions can inform lab experiments. In this way, we can specifically test the genes most likely to be involved, to increase our knowledge of this crucial pathway.

Table of contents

| | |
|--|-------------------------------------|
| Abstract | 1 |
| Layman's summary | 1 |
| Table of contents | Error! Bookmark not defined. |
| Introduction | 3 |
| Results | 4 |
| Discussion | 15 |
| Methods | 17 |
| Enrichment analyses | 20 |
| Construction of the final naïve Bayesian classifier | 22 |
| Literature analysis of candidates and TFBS motifs | 23 |
| References | 24 |
| Acknowledgements | 33 |
| Supplementary Material | 33 |
| Extended data 1: enriched TFBS motifs in promoters of MHC genes | 33 |
| Extended data 2: Literature study of the top ten negative set genes in the Bayesian classifier | 37 |
| Supplementary figures | 40 |

Introduction

The major histocompatibility complex (MHC) or HLA (human leukocyte antigen) in humans, is a set of two multi-subunit protein complexes that present peptides at the cell surface¹. By doing this, they allow the induction of adaptive immune responses if non-self peptides are presented, or if aberrant self-peptides are presented (as might be the case in cancer cells)¹. MHC molecules achieve these functions by interfacing with T-cell receptors (TCRs) and NK cell receptors^{1,2}. The MHC system is split into two distinct classes. The first is MHC class I, which is present on almost all nucleated cells in the human body³ (**Figure 1A**). It presents endogenous peptides, derived from the cleavage of proteins in the cell by the proteasome and other proteases, at the cell surface⁴. This requires transport of peptides over the ER membrane, stabilisation of the nascent MHC molecule by several chaperones, and a complex loading procedure that involves peptide editing (continuous switching in which peptide is loaded to find the most stably bound one)⁴⁻⁸. The loaded MHC I molecule is then transported to the cell surface, where it is presented for 3-7 days⁹. The MHC II molecule is different in that its expression is mostly restricted to immune cells¹⁰⁻¹² (**Figure 1B**). This molecule is capable of presenting exogenous peptides/antigens, which are taken up from the cell's environment^{1,13}. Dendritic cells, for example, traffic to sites of infection, become activated, take up extracellular antigens, traffic back to the lymph nodes and stimulate an adaptive immune response there¹⁴. Whereas the MHC I molecule features a complicated loading apparatus, the MHC II molecule is created with the so-called CLIP fragment that helps stabilise it in vesicles until exogenous antigen is introduced that can be loaded^{1,15}. Cathepsin S then degrades the CLIP fragment so that antigen can be loaded, and MHC II traffics to the cell surface^{1,16}.

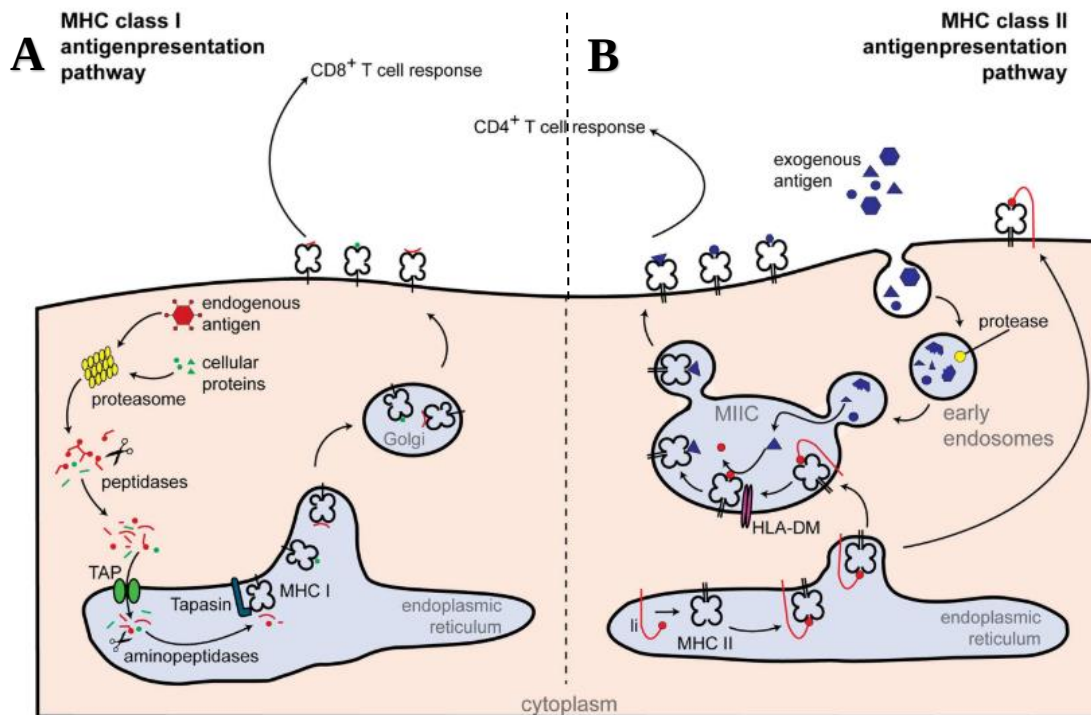


Figure 1. MHC I and MHC II pathways. A: MHC I is loaded with peptides derived from endogenous antigens and cellular proteins, which are broken down by the proteasome, and further processed by peptidases. They are then imported into the ER, sometimes further cleaved, and loaded onto MHC I molecules by means of many chaperones and the peptide loading complex^{1,9}. After peptide editing, the molecules are transported to the cell surface¹. B: MHC II is mainly expressed on immune cells, and present exogenous antigens which are taken up from the environment, processed in endosomes, and subsequently loaded on MHC II molecules, displacing the CLIP fragment of Ii that stabilised the unloaded MHC II. These molecules allow APCs to activate immune responses^{1,17}. Figure taken from¹³.

Given MHC's central role in adaptive immunity, it is not surprising that it is involved in cancer, autoimmune diseases, and highly associated with disease outcome in infectious diseases^{1,18–20}. It is, therefore, important to fully understand this major pathway to better grasp how to combat autoimmune diseases^{21,22}, to find novel therapeutic approaches to cancer¹, and how to best tailor vaccinations and disease treatments to specific HLA haplotypes²³. From many years of research, we now have a good understanding of the major players in the MHC pathway^{1,24}. The major components of the MHC I and II pathways are known, as are many of the chaperones involved, and we know much about the transcriptional regulation of these molecules^{3,8,25–28}. However, an important regulator of the MHC I genes (CITA/NLRC5) was only discovered in 2010^{3,13,29} and the peptide-editing function of TAPBPR was only recognised in 2012^{5,6,30}, which shows that many factors that play a role in the MHC I pathway are still unknown. Similarly for MHC II, a genome-wide siRNA knockout study published in 2011 identified many new factors governing MHC II transcription and its transport in dendritic cells²¹.

Given that the MHC genes are at the core of immunity, and that much about their regulation remains unknown, we developed a method to predict novel candidates of importance to the MHC pathway. Recent approaches using a Naive Bayesian classifier to integrate data sets and predict novel genes involved in mitochondria, the Rig I-like Receptor pathway (RLR-genes) or the cilium have had much success^{31–33}.

Therefore, in this work, we integrated five MHC-related data sets to predict novel genes belonging to the MHC pathway. These data sets contain information on measured protein-protein interactions between virus and host proteins, transcription factor binding sites (TFBS) in the promoters of MHC pathway genes, immune tissue overrepresentation of proteins, time-series gene expression profiles upon macrophage activation, and MHC II surface expression perturbation in an experimental genome-wide siRNA screen. We believe that the TFBS we find to be enriched in MHC pathway gene promoters offer novel insights into MHC transcriptional regulation, and that the candidates predicted by this classifier can serve to inform experimental studies into MHC pathway genes in order to further lift the veil on this highly important pathway.

Results

Defining an MHC pathway gene positive set and negative set

In order to extract features informative for being an MHC pathway-related gene, we need a set of known MHC pathway-related genes. To this end, we constructed a set of 86 genes which are known to be involved in MHC I and/or II transcription, surface presentation, and peptide loading (the latter also includes the subunits of the immunoproteasome) based on the KEGG database³⁴ and literature searches^{3,5,13,21,35} (**Supplementary file 86MHCPathwayPositiveSetGenes.csv**). We hypothesise that these genes have common features, and that those genes which are not in the positive set, but share many of these common features, are more likely to be MHC pathway-related.

The negative set was more difficult to define. Several publications have constructed specific negative sets^{31–33}, containing genes with known functions, and localisation, very different from the genes of interest. This, however, is a time-consuming ordeal, and it is very difficult to get a fair negative sample. Instead, we opted for a workable simplification. Given that we know that there are ~22,000 human genes, and we know that the verified positive set contains only ~90 genes, we can estimate the total number of MHC pathway genes to be at most double the known amount of genes, at 200. Given that that is such a minute fraction (<1%) of all genes, we can assume for simplicity that all genes, save those in the positive set, are not MHC pathway-related genes and therefore belong in the negative set. Our expectation is thus for positive set genes to behave alike in our different datasets: that they share some transcription factor binding sites, that they are upregulated in a similar fashion, and that their knockdown hinders MHC (II) expression. We expect negative set genes not to do this. Even though some of the genes in the negative set will be MHC pathway-related, the signal of these genes is vastly outnumbered by the genes that are not. If, however, a gene that we had assigned to the negative set

consistently behaves like a positive set gene in different data sets, then it might actually be an MHC pathway-related gene. We use a naïve Bayesian framework to objectively gauge which genes do so and thus find these likely MHC pathway candidates.

MHC pathway genes are enriched in protein-protein interactions with viral proteins

Viruses have evolved to subvert immune responses in diverse ways³⁶. Some can prevent chemokine signalling, others manipulate it outright by making their own chemokine mimics^{37,38}. Viruses can also modulate recognition and killing by natural killer cells (NK-cells) by mimicking NK-cell inhibitory ligands or reducing activating ligands, perturb endosomal function (hindering surface presentation), and reduce MHC molecules at the cell surface^{39,40}. Indeed, many viruses target the MHC I and II pathways specifically, to prevent the host mounting a successful (adaptive) immune response^{36,41}. This can be an insidious process: MHC I molecules, when completely absent, signal an abnormal cell state; these cells are killed by NK cells³⁶. Viruses, then, modulate MHC I expression to be much lower, but not completely absent, to escape both NK-cell and cytotoxic T-cell responses³⁶. Given this knowledge, we asked whether MHC pathway genes are enriched in reported interactions with viral proteins. Though viruses interfere with many immune pathways^{33,42}, part of the signal for being an MHC pathway gene could constitute being targeted by viral proteins more than other genes in the human genome. To investigate this question, we gathered data from three databases that record measured interactions between human and viral proteins^{43–45}. We proceeded with the union of these three databases (**Methods; Supplementary Figure 2**). We determined whether a protein had any recorded human-viral protein interactions, or not. We then asked whether MHC pathway genes were enriched in the proteins that have measured interactions with viral proteins. This is indeed the case, as expected (**Figure 2A**). About 70% of MHC pathway genes have interactions with viral proteins, as opposed to a mere 30% of other genes in the human genome.

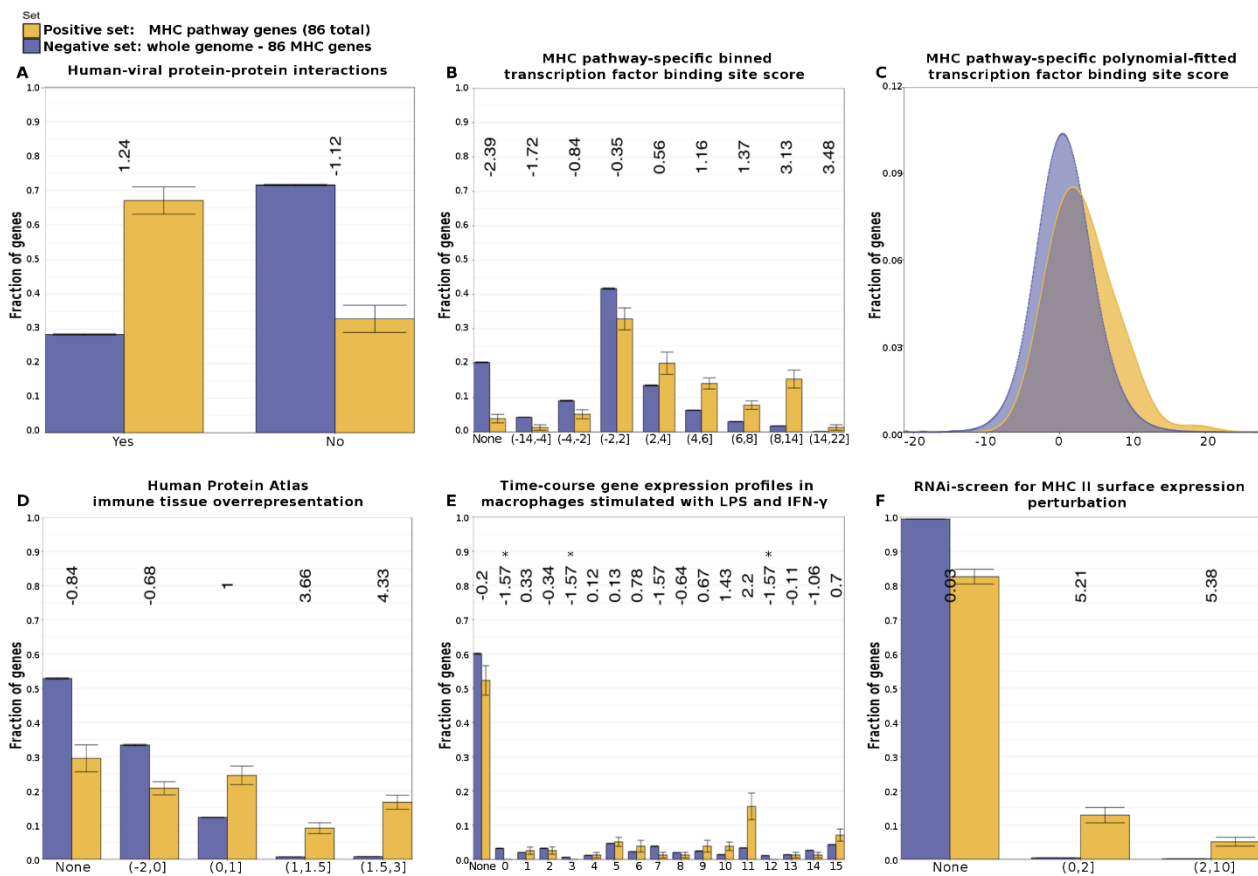


Figure 2. Data sets and log2 likelihood ratios (numbers above bins) used in the naïve Bayesian classifier. Median values from ten cross-validations. Purple: negative set genes. Yellow: MHC

pathway genes. Error bars: ± 2 sd. Genes without a value in a data set are reported as 'None' (see Methods). A: pooled host-viral protein-protein interactions as reported in three databases^{43–45}. B & C: MHC pathway gene specific additive Transcription Factor Binding Site (TFBS) motif score in promoters. TFBS motif data retrieved from⁴⁶. B: enrichment scores calculated via binning and calculation of log2 likelihood ratios. C: polynomial-fitted kernel density-smoothed distributions of the TFBS motif score in the negative and positive set. D: rank-based overrepresentation in immune tissues based on data from the Human Protein Atlas⁴⁷. E: Time-series gene expression profiles in macrophages stimulated with LPS and IFN- γ . Expression profiles based on microarray macrophage gene expression at 6, 12, and 24 hours post-activation, clustered using STEM^{48,49}. Asterisks: bins that contain no positive set genes, whose scores therefore become $-\infty$. These scores were corrected to the minimum scores in the data (log2 enrichment of -1.57 in this case) F: genome-wide siRNA screen with antibody fluorescence read-out of MHC II surface expression perturbation²¹. This data covers 276 genes. See Supplementary Material for additional information.

Discovery of TFBS motifs involved in regulating MHC pathway genes yields known regulators

Another defining feature for the MHC pathway genes could be the Transcription Factor Binding Sites (TFBS) found in their promoter regions. TFBS motifs are short stretches of DNA sequence (usually between 6-18 bp) that are bound by certain transcription factors^{50,51}. Transcription factors such as NF- κ B and IRF are known to be involved in regulation of many immune genes, and CIITA, CITA, CREB and NFY A-C are specifically involved in regulating MHC I and II genes^{3,26,52–55}. Previous naive Bayesian integration approaches have used TFBS motifs known to be involved in their pathways of interest, and scored other genes on whether their promoters contained them^{32,33}. However, transcriptional regulatory networks are extremely complex on many levels⁵⁶, and the same transcription factor can function in many different tissues, and affect many different processes^{57–59}. Accordingly, the regulators we know to be involved in pathways are often very important players, whose knockout leads to marked phenotypes or (almost) complete abrogation of a function. NF- κ B, a master regulator of immunity (among other processes) is one example⁶⁰. Given this complexity, and that functions of many transcription factors are as of yet undefined⁶¹, we aimed to be more broad in our search than previous naive Bayesian approaches.

Therefore, we decided to use a data set of genome-wide identification of TFBS motifs in the Encode human ChIP-seq data⁴⁶. This paper combines many motif-finding tools and literature to find high-confidence motifs throughout the human genome^{51,6246}. Using this data, we could determine for all TFBS whether they were over- or underrepresented in the promoters of MHC pathway genes. Briefly, we performed Fisher's exact test on counts of all motifs in MHC promoters and the whole genome and performed Benjamini-Hochberg FDR multiple testing correction. This yielded a sample estimate of under- or overrepresentation and a corrected p-value per TFBS. We used cross-validation to avoid recursiveness (**Methods**).

If a TFBS is overrepresented in MHC pathway gene promoters, having that TFBS apparently says something about being an MHC pathway gene. Conversely, genes that have TFBS which are significantly underrepresented in MHC pathway genes are unlikely to be involved in this pathway. This approach is a compromise between broad applicability and complexity. Of course, the different compartments of the MHC pathway (the ER chaperones, the peptide-loading machinery, the proteases and the immunoproteasomal component) will be regulated by different factors, but checking for enrichment in the entire MHC pathway gives us a data-driven broad MHC pathway TFBS signature, rather than a literature-driven one based only on known regulators. This approach in fact allows one to identify a TFBS signature for any pathway of interest, and can hence be used in other classifiers or to identify probable transcription factors (TFs) regulating a pathway.

A full list of TFBS motifs enriched in MHC pathway gene promoters in at least 5/10 cross-validations is provided below (**Table 1**). Additionally, we looked at the distribution of motif counts in all positive set genes that had an informative TFBS motif at least once (**Supplementary Figure 1**). Most motifs are present 1-4 times per promoter, but NF- κ B, IRF, and TFAP2 are notable exceptions. The most

broadly informative TFBS motifs are listed in the Supplementary Material along with descriptions of the known functions of TFs binding these motifs (**Extended data 1**). We find back many known regulators of immune pathways and the MHC pathway, such as IRF, NF- κ B, REL (a subunit of NF- κ B), NFY and STAT^{26,52,63,64} (**Table 1**). This shows that our approach works, but the true novelty and goal of our approach lies in the heretofore unknown candidates, which will be discussed in the next section.

Table 1. TFBS motifs enriched in MHC pathway gene promoters.

All TFBS motifs found to be enriched in MHC pathway gene promoters in ≥ 5 cross-validations. Enrichments were calculated using a custom script. Data obtained from Kheradpour and Kellis⁴⁶.

| TFBS name | Number of cross-validations in which found enriched in MHC |
|-------------|--|
| IRF | 10 |
| NFKB | 10 |
| NFY | 10 |
| IRF4 | 10 |
| PRDM1 | 10 |
| RAD21 | 10 |
| LMX1B | 10 |
| HOXD13 | 10 |
| REL | 10 |
| TFAP2 | 10 |
| FOXD2 | 10 |
| ZSCAN16 | 9 |
| HOXB13 | 9 |
| MYC | 9 |
| LMX1A | 9 |
| PLAGL1 | 9 |
| STAT | 8 |
| HOXA13 | 8 |
| HSFY2 | 7 |
| HEY1 | 7 |
| EWSR1::FLI1 | 6 |
| ZNF281 | 6 |

Novel TFBS motifs enriched in MHC pathway gene promoters offer interesting new regulatory possibilities

Besides the TBFS motifs which can be linked to known regulators, we also found many new possibilities for regulation of MHC pathway genes. To assess these candidates, we performed a literature study to find relations to immune functions. We highlight three candidates here. All ten are discussed in the Supplementary Material (**Extended data 1; Table 1**).

One of the motifs found to be enriched is RAD21. RAD21 is a member of the cohesin complex, which binds the sister chromatids to keep them together before anaphase^{65,66}. However, it is also implicated in regulation of chromatin accessibility together with CTCF⁶⁶. It might be that this transcription factor regulates MHC pathway gene promoter accessibility.

Another promising candidate is the TFBS motif for HSFY2. HSFY2 is expressed solely in the testes, and is situated on the Y chromosome⁶⁷. Its deletion is linked to azoospermia⁶⁷. The testes are an immunoprivileged tissue, which means that immune reactions are suppressed⁶⁸. The testes instead have a specific innate immune system⁶⁸. Indeed, the testes also have a specific proteasome⁶⁹. Taken together, this suggests that this transcription factor is somehow involved in regulating MHC pathway genes in the testes, perhaps linked to the immunoprivileged conditions there.

TFBS for three HOX genes (HOXA13, HOXB13, and HOXD13) are also found enriched in the MHC pathway gene promoters. These genes are mostly known for their involvement in embryonic body patterning, and only HOXB13 has been linked to interleukin 6, which is a cytokine involved in regulation of immune responses⁷⁰⁻⁷².

The TFBS motif score is generalised, finds novel and known TFBS motifs in MHC genes, and is broadly applicable

We do not find motifs for all known transcriptional regulators. This is because the data set we use does not contain all possible motifs. Rather, it uses 427 ENCODE experiments on a total of 123 TFs, which constitutes 84 factor groups⁴⁶. As it integrates multiple motif-finding methods and yields only high-confidence motifs, not everything is included⁴⁶. Additionally, we grouped motifs together (**Methods**). This approach was chosen because the motifs were divided into previously known and newly discovered motif groups in the original data⁴⁶. This gave a very fine-grained resolution of the subtypes within motifs bound by a specific TF, but also limits enrichment approaches, since few members of any motif subtype exist in the genome. Therefore, we limited the resolution, but increased statistical power by grouping the motifs together based on which TF bound them. Given the fact that we manage to recover important players like IRF, NF- κ B, and NFY alongside novel motifs that are enriched in the MHC pathway, we believe that the TFBS motif score we build using these motifs is more defining than taking pre-existing literature motifs. Furthermore, the custom script we used for this analysis can be applied to calculate TFBS motif enrichment in promoters of any gene set of interest, relative to the whole-genome background.

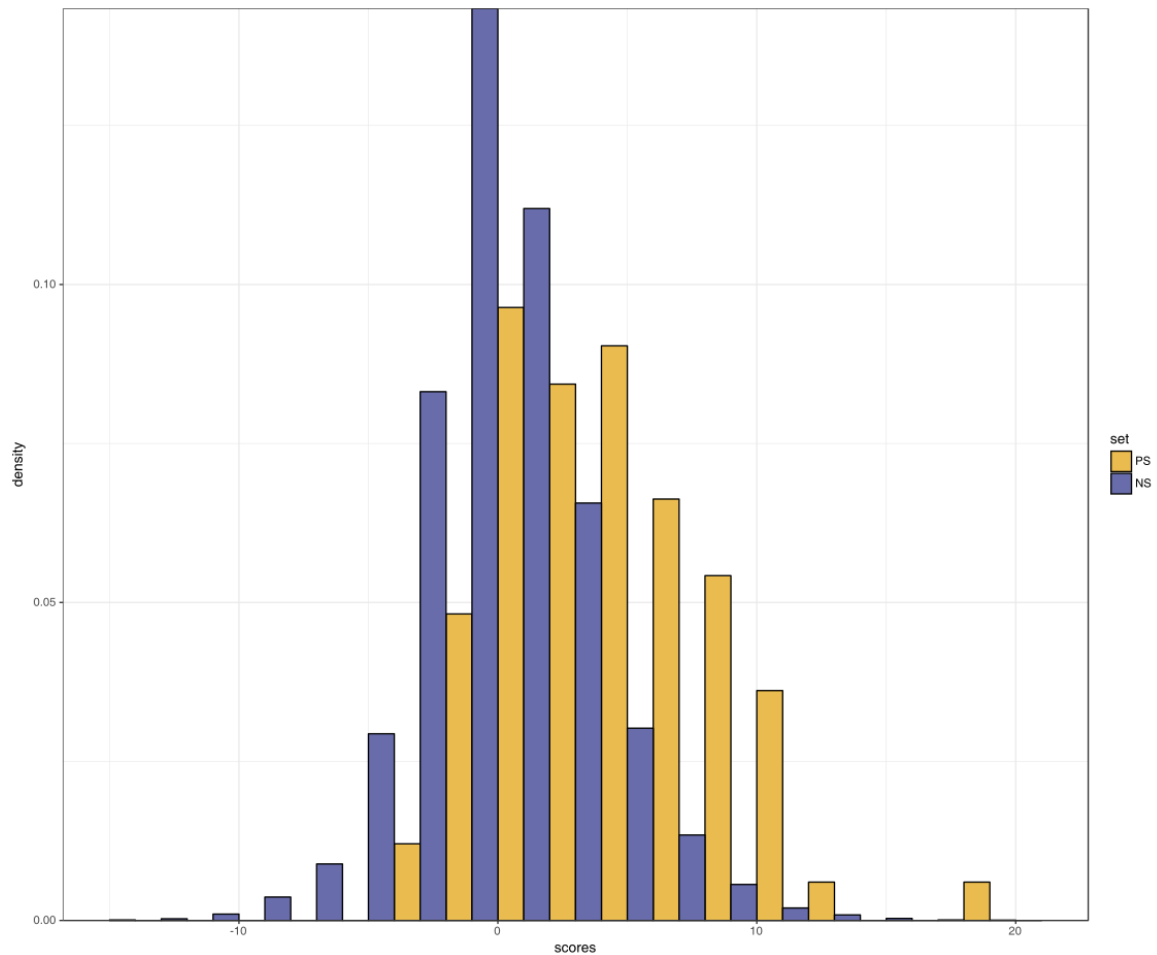


Figure 3. Histogram of the total additive MHC pathway-specific TFBS motif score for all protein-coding genes in the human genome. NS: negative set. PS: positive set (MHC pathway genes). Scores were binned in bins of width 2. Scores are combined data from ten cross-validations. Negative set genes have lower scores on average than positive set genes, showing that this score selects for MHC pathway genes.

A combined MHC pathway TFBS motif score greatly enriches for MHC pathway genes

Using the sample estimate scores for all TFBS motifs found to be enriched in the MHC pathway, we scored all protein-coding genes present in the human genome (with motifs assigned to them, see Methods) for their MHC pathway TFBS signature (**Figure 3**). For enriched TFBS motifs, the score metric takes into account the average number of motifs per MHC pathway gene. This means that if each gene in the MHC pathway has 4 NF-kB motifs on average, a gene that has just 1 NF-kB motif gets 1/4th of the score for that TFBS (see Methods). We chose our bins for the Bayesian classifier score such that every bin had an enrichment higher than the next, and contained at least one positive set genes. Our score metric achieves very good separation of negative set and positive set genes, indicating that the few negative set genes that are very high scorers might indeed be involved in the MHC pathway, based off their TFBS motifs (**Figure 2B**). See the Supplementary material for bins of width one for comparison with our custom bins (**Supplementary Figure 3**).

We wondered whether the distributions of values in the positive set and negative set could also be used more directly. Binning introduces a trade-off between bias and accuracy⁷³. For this reason, we aimed to fit the distributions directly. We reasoned that the dip in positive set log2 enrichment scores in TFBS scores between 2-8, the sharp rise in bin (8,14], and the complete absence of scores until bin (18,19] might be due to undersampling the positive set (**Figure 2B**; **Supplementary Figure 4**). Therefore, we smoothed the distributions before fitting a higher order polynomial to directly infer the log2 enrichment score per gene from the underlying distributions (**Figure 2C**). Alternatively, our

score metric might select for extremes in motif amounts in the promoters of MHC pathway genes, explaining the effect (see Discussion). Our approach results in a continuous TFBS motif score metric. Example scores along the range of the distribution can be found in the Supplementary Material (**Supplementary Figure 5**). At the fringes of the distribution, the scores are bounded because lack of data causes unrealistic spikes in the continuous score (**Supplementary Figure 6; Supplementary Figure 7; Supplementary Table 1**). The continuous TFBS motif score also gives good enrichment of positive set genes for higher TFBS scores. As an added benefit, it reflects the true score of each gene, rather than a binned average. We used this continuous score in the final classifier. In sum, our TFBS motif score selects for MHC pathway genes and those genes in the human protein-coding genome that are enriched in TFBS motifs found to be important in the promoter regions of the 86 MHC pathway genes. This score is more inclusive than using only literature-based TFBS, and achieves good separation of negative set and positive set genes.

MHC II-centric immune tissue enrichment score enriches for MHC pathway genes

The MHC I molecule has a large loading complex, chaperones, and a peptide transporter, while only the CLIP fragment is known to be crucial for MHC II loading. This indicates that more proteins might yet be involved. One of the defining features of MHC II is that it is expressed mostly in various immune cells such as dendritic cells and macrophages (though see refs 9 and 10), which can take up exogenous peptide and present it on MHC II^{1,74}. Therefore, if a gene or protein is overrepresented in immune cells or tissues relative to non-immune-related ones, it has a higher chance of being involved in the MHC II part of the pathway.

To investigate which genes fulfil these criteria and to score the enrichment, we used data from the Human Protein Atlas⁴⁷. This data was obtained using protein immunohistochemistry on 10,600 proteins, curated by experts and assigned a label from non-expressed to highly expressed, in multiple different tissues and cell types. We classified the tissue-cell type combinations into an immune set (for example, splenic cells, skin Langerhans cells and lung macrophages) and a non-immune set (all other tissue-cell type combinations), following earlier work²¹. We then performed a Mann Whitney U-test to determine the rank-based differences in means between the immune and non-immune sets for each protein in the data. The results were binned, and we performed enrichment analysis to see if MHC pathway genes were enriched in immune tissues and cell types.

MHC pathway genes are indeed enriched in immune tissues, relative to the negative set (**Figure 2D**). The last two bins have very high log₂ enrichment scores, as almost no negative set genes are so highly expressed in immune tissues (**Figure 2D**). The same holds true in bins of width one (**Supplementary Figure 8**). Also, as expected, not all positive set genes are overrepresented in immune tissues: about 20% of the MHC pathway genes are not enriched in immune tissues (**Figure 2D**). This is in line with knowledge on MHC I genes. While it should be noted that there is only data on about half of the protein-coding genes, the data that is present has been manually curated and is of high quality, strengthening our belief that it captures the true behaviour of these genes. In conclusion, this immune tissue enrichment score enriches for MHC pathway genes and hence captures part of the identity of MHC pathway genes.

MHC II-centric gene expression time course profiles in activated macrophages enrich for MHC pathway genes

In the previous sections, we showed that interacting with viral proteins, containing MHC pathway-enriched TFBS motifs and being more highly expressed in immune-related cell types and tissues are all characteristics that contribute to the MHC-ness of a gene. Another valuable insight can come from co-expression data. Genes that act together in the same pathway or process are often co-regulated^{75,76}. While this is in part due to specific shared TFBS motifs, it can also be due to higher-level processes, such as regulation of chromatin state, cell-scale reactions to nutrient availability, or distal enhancers^{56,76,77}. Co-expression profiling has also been successfully used to identify novel RLR-pathway genes with a naive Bayesian classifier^{32,33}. For these reasons, we sought to include co-expression data in our classifier. Macrophages and DCs upregulate MHC II after immune activation

to become efficient antigen-presenting cells (APCs)⁷⁸. We used a data set that contains microarray experiments on gene expression after macrophage activation with LPS and IFN- γ at 6, 12 and 24 hours post-activation⁴⁹. Following the strategy used by Derlindati *et al.*⁴⁹, we clustered the gene expression time-courses into several profiles using STEM (see Methods for details)⁴⁸. This resulted in 16 gene expression profiles, along with the profile 'None', which contains the genes that underwent no large changes in gene expression over the 24-hour period. We then queried these profiles for enrichment of MHC pathway genes (**Figure 2E**). Some profiles had no MHC pathway genes in them, which results in enrichment scores of $-\infty$. Therefore, their log2 enrichment score was set to the minimum enrichment score in the data set.

One can see that profile number 11, specifically, is highly enriched in MHC pathway genes (**Figure 2E**). This signal is consistent across all cross-validations. Genes in this profile are continuously upregulated following macrophage activation (**Figure 4**). Other profiles that are enriched in MHC genes are 6, 9 and 15 (**Figure 2E**). The majority of these profiles feature upregulation (**Figure 4**). A part of the MHC pathway gene identity is thus engendered in the upregulation upon macrophage activation, and genes that have this characteristic are somewhat more likely to be involved in the MHC pathway themselves.

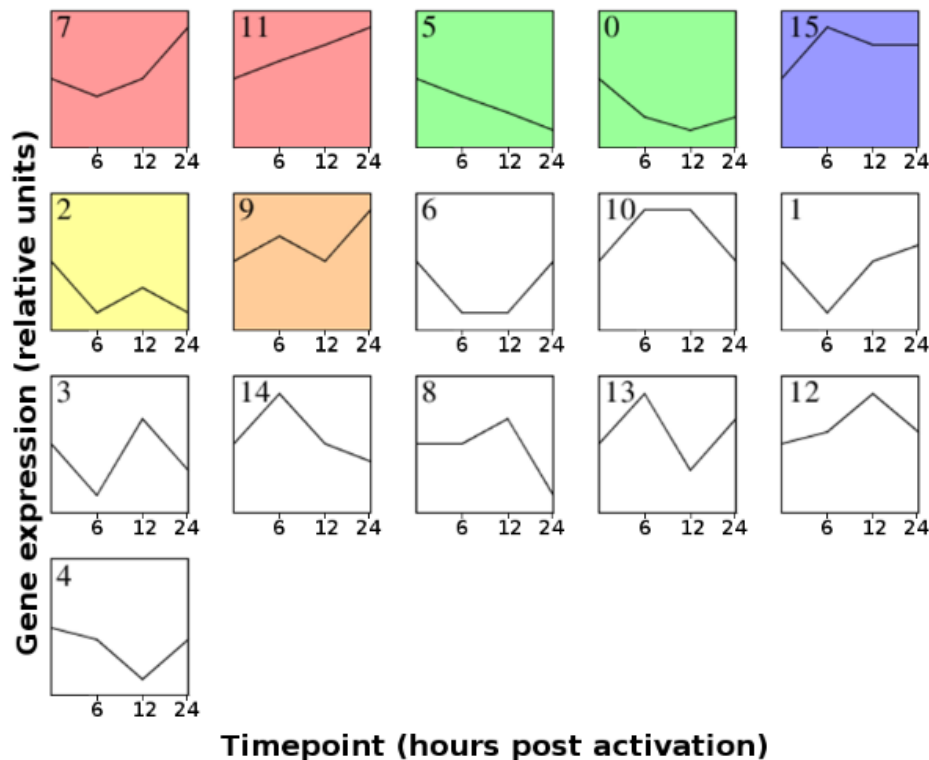


Figure 4. Activated macrophage gene expression profiles reported by STEM⁴⁸. Coloured backgrounds represent expression profiles which were present more often than expected (based on the amount of genes and a sample of random possible gene expression profiles). Profiles are ordered based on how much more often they were observed than expected. Data from Derlindati *et al.*⁴⁹.

An experimental genome-wide siRNA screen with read-out of MHC II cell surface perturbation pinpoints 276 high-confidence MHC pathway candidates

If a gene has been experimentally shown to influence the surface expression of MHC II, then this is powerful evidence supporting its involvement in the MHC pathway. Exactly such an experiment has been performed by Paul *et al.* from the Neefjes lab²¹. siRNAs against all human genes were introduced into MeIJuSo cells. These are not immune cells, but do express MHC II and CLIP molecules²¹. It was then assessed via antibody staining with two different antibodies whether the knockdown of a gene

altered either MHC II-peptide surface expression (which shows whether proper peptide loading occurs), or MHC II-CLIP surface expression (which shows whether something is amiss with peptide loading and/or surface expression)²¹. This allowed a functional read-out of effects on MHC II surface expression and peptide loading. Based on the distribution of all scores and known positive and negative controls, the authors could define which genes' knockdown significantly increased or decreased MHC II surface expression. This was done using a Z-score, which quantifies the deviation from the mean of a distribution. They performed resccreens on promising candidates, did deconvolution screens to correct for off-target effects, and determined via microarrays whether these genes were also expressed in actual immune cells²¹.

This genome-wide experimental screen of MHC II cell surface perturbation by siRNA knockdown of genes is the final data set we included in the classifier. It offers direct experimental evidence of involvement with the MHC II pathway, though in a sense broader than the candidate genes we aim to discover. For example, disturbance of proteins very basal to a healthy cell, such as metabolic processes, transport of vesicles, or the import of proteins into the ER, will also be reflected in lower scores of MHC II surface expression. These are not actual acting entities within the MHC pathway, but rather prerequisites for correct cellular function. It should be kept in mind that this data set will also capture these much broader definitions of 'being involved in the MHC pathway'. High scorers in the final Bayesian classifier, however, will score high in multiple data sets, which increases the chance that they are actual MHC pathway genes, rather than basal genes.

We used several forms of this data set: i) the initial results of screening all human genes (original), ii) the data that was rescreened to verify that hits were actual hits (rescreen only), and iii) the data that underwent deconvolution screens and was verified to be expressed in immune cells (high confidence). We are interested in uncovering novel candidate genes involved in the MHC pathway. We thus need not know whether knockdown increases or decreases surface expression, but only whether there is an effect on surface expression. Similarly, it is irrelevant whether knockdown affects MHC II-CLIP or MHC II-peptide surface expression specifically. We need only know whether a gene can affect surface expression of MHC II at all. Therefore, we took the absolute values of the Z-scores for both antibodies, and took the maximum of that value. This represents the maximum effect a gene can have on MHC II surface expression perturbation screen (MHC II SEPS).

MHC II is predominantly expressed in immune cells and of great importance to APCs (though see refs 9 and 10). Therefore, any candidate gene that influences its surface expression should be expressed in immune cells for it to be involved *in vivo*. Additionally, though their ROC curves looked promising (**Supplementary Figure 9**), the enrichment scores for the rescreened and original scores were lower (**Supplementary Figure 10 and 10**). For these reasons, we chose to use the 276 high confidence candidates, rescreened, deconvoluted, and checked for expression in one or more immune cell types, in the final Bayesian classifier.

The high confidence of these candidates is evidenced by their high enrichment scores (**Figure 2F**). Indeed, these are the highest scores of all data sets. This is as expected, since these genes have been conclusively shown to be involved in MHC II surface expression. See Supplementary Material for bins of width one to compare with our custom bins (Error! Reference source not found.-11).

In sum, the MHC II surface expression score identifies 276 genes that are highly likely to be involved in the MHC II part of the pathway. However, high scores in this data can occur because of disruption of basal processes, and we should be mindful of that fact in interpreting the classifier results.

Naive Bayesian data integration has higher recall and precision than its constituent data sets

We incorporated all data sets described above in our naive Bayesian classifier and trained it with a tenfold cross-validation. For every gene, we tallied up the scores for every data set and the prior likelihood ratio to arrive at the final Bayesian classifier score (**Methods**). We gauged the classification potential of all data sets and the integrated Bayesian score using an ROC curve (**Error! Reference source not found.B**). An ROC curve is a metric that is used to visualise the signal-to-noise ratio of a classifier⁷⁹.

From the ROC plot, it is clear that the combined Bayesian classifier score is more informative for MHC-ness than any of the data sets that constitute it alone (**Figure 5B**). Data integration has thus allowed us to more accurately pinpoint MHC pathway identity than any data set alone could. To test what the most vital data sets are for classifier performance, we left each data set out in turn and compared classifier performance with the full classifier. The immune tissue overrepresentation score and the virus-host PPI are most vital to classifier performance (**Supplementary Figure 13**). Conversely, the macrophage gene expression profiles are least important to classifier accuracy (**Supplementary Figure 13**). The ROC curve for the high confidence Z-scores from the MHC II surface expression screen candidates seems to perform close to random predictions (**Figure 5B**). That is, paradoxically, because this is a selection of very high confidence candidates. In a pool of candidates that are very likely involved in the MHC pathway and that have persisted through rescreening and deconvolution attempts, it is expected to find high scores to be divided about equally between known positive set genes and new candidates. Because of that, the signal-to-noise ratio in the ROC plot is close to random. In other words, all 276 genes are good candidates based on the data.

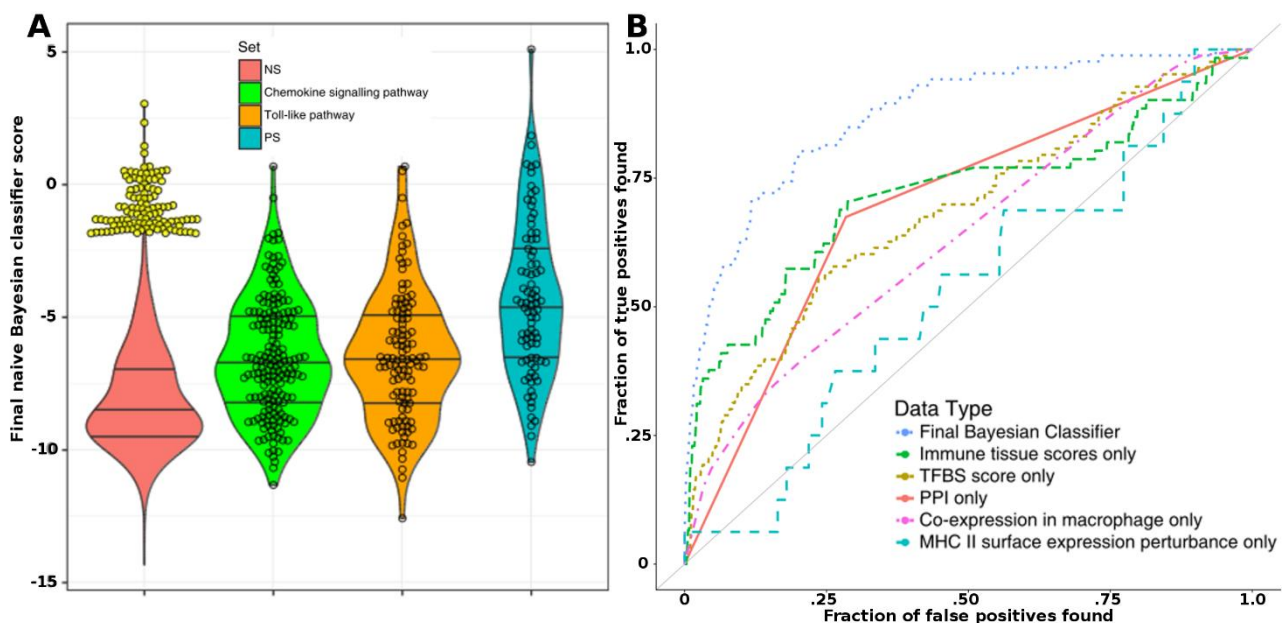


Figure 5. Final naïve Bayesian classifier scores and classifier performance. A: violin plot of the final scores for all protein-coding genes in the human genome. Red: negative set genes. Green and orange: chemokine signalling pathway genes and Toll-like pathway genes, respectively (taken from WikiPathways⁷⁸). Blue: MHC pathway genes (positive set). Yellow dots: top 0.5% scores in the negative set. B: ROC curve of the final naïve Bayesian classifier and the separate constituent data sets.

The naive Bayesian classification produces a score geared towards MHC pathway genes and identifies many high-scoring novel candidate genes in the genome

The goal of this project was to build a classifier that could pick out a signal for MHC-ness from a known positive set and identify novel candidates involved in the MHC pathway in all protein-

coding genes in the human genome. We have already shown that our classifier performs better than any individual data set in picking out the known positive set genes (**Figure 5B**). This implies that the signature arrived at is truly one that distinguishes MHC pathway genes. In confirmation of that fact, the genes from the Toll-like pathway and the chemokine signalling pathway, both involved in immune signalling and activation, score lower than the MHC pathway genes (**Figure 5A**). The classifier is primarily geared towards the MHC II part of the pathway (**Supplementary Figure 14**): MHC II genes receive higher scores than the MHC I component of the positive set. Interestingly, the immunoproteasomal subunits also receive high scores, despite only constituting a relatively small part of the positive set (15/86 genes).

The classifier reveals interesting candidates in the negative set. The top 0.5% scores in the negative set have scores that are higher than, or as high as, many of the genes in the actual positive set (**Figure 5A**). We did a literature study on the ten top-scoring candidates (**Extended data 2**). Three candidates are highlighted in the following sections: CD247, PPP1R18, and CHTOP.

Broader immune functionality of CD247: three avenues for MHC-relatedness

An interesting candidate is CD247, which has a total score of 1.06, making it about twice as likely to be involved in the MHC pathway relative to not being involved. It is a co-receptor of the TCR⁸⁰. At first sight, this seemingly makes it a bad candidate, as TCRs interact with MHC molecules on other cells (either class I or class II)¹: considering the molecule itself to be a part of the MHC pathway is counterintuitive. However, T cells express the MHC I molecule, as do almost all nucleated cells³⁵. More strikingly, with the exception of mice, T cells in many animals highly up-regulate MHC II upon activation⁸¹. This holds true for humans as well⁸¹. Therefore, it could well be that this molecule interacts with MHC in T cells. Indeed, CD247 is selected as a likely candidate because it is highly enriched in immune tissues (as expected), but also because it is continuously up-regulated in macrophages upon activation. However, its TFBS score is very low, and it was not one of the candidate genes in the MHC II surface expression screen. The upregulation in macrophages is striking, considering that this molecule is only widely known for its function in T-cell signalling⁸⁰. A recent study found that miRNAs in alternatively activated macrophages (M2 macrophages) might downregulate or interfere with CD247, which could contribute to their non-inflammatory phenotype⁸². This hints at some function in activated macrophages, which might be related to MHC.

Another interesting fact relates to the involvement of the immune system in (re)modelling of neurons⁸³. CD247 is expressed on mice retinal ganglion cells (RGCs), and knockout prevents the correct optical wiring from forming^{84,85}. Knockout of MHC I gives a similar phenotype^{84,85}. There are thus three possible links CD247 could have to MHC pathway genes: it might be something more than a TCR co-receptor and associate in some way with MHC (I or II) in T cells, it might be involved with MHC pathway genes in APCs, and it might have something to do with the non-classical functions of MHC pathway genes in neuronal remodelling. Nevertheless, it needs to be stressed that the high score in the classifier is mainly due to CD247's extreme enrichment in immune tissues, which is as expected, as many of them are populated with T cells. On the other hand, the upregulation of this molecule in activated macrophages is unexpected.

PP1R18 is downregulated in malignant breast cancers, is expressed in lymphocytes, and lies in a conserved region in-between HLA loci

PP1R18 has a total classifier score of 0.57, and codes for a protein known as phostensin⁸⁶. Phostensin is chiefly known as an actin-interacting compound that locates to the cell periphery. It is located between the HLA-C and HLA-E regions, a situation that is conserved in the swine homologue⁸⁷. It is overrepresented in immune tissues such as the spleen and lymph nodes, and expressed in immune cells⁸⁷. Furthermore, it is one of the genes differentially regulated between metastatic and non-tumorigenic breast cancer cell lines⁸⁷. Additionally, a recent report states that phostensin is invaluable in regulating actin and bone resorption in osteoclasts⁸⁸. Close ties between osteoclasts and the immune system have long been known⁸⁹. This strengthens the case for immune

function further. Note, however, that the high score of PP1R18 is based mainly on interaction with viruses, overrepresentation in immune tissues, and upregulation in macrophages. These are the three ‘general’ data sets, which pinpoint many different immune genes. In contrast, the TFBS score for PP1R18 is low, and it was not a high confidence candidate in the MHC II surface perturbation screen. Hence, the evidence for actual MHC pathway involvement is not ideal.

CHTOP scores well in all data sets except immune tissue overrepresentation

CHTOP’s total score in the classifier is 1.73. It is a gene that is involved in many processes, such as the export of mRNA from the nucleus as a component of the TREX complex, and, via membership of the SMN (survival of motor neuron) complex, transcriptional regulation, telomerase regeneration and cellular trafficking in all animal cells^{90,91}. It has a high score in all data sets except immune tissue overrepresentation, and is a high confidence candidate in the MHC II SEPS. However, this screen can also identify genes that tamper with processes very basal to proper presentation of proteins on the cell surface (such as correct mRNA export from the nucleus). The TFBS motif score is positive, which suggests potential for immune activation, but this might also be for upregulation of mRNA export during infection. As we found evidence in the virus-host PPI, upregulation in macrophages, MHC pathway TFBS motifs and impact on MHC II for involvement in the MHC pathway or immunity, this candidate merits further research.

Discussion

In this project, we sought to predict novel members of the MHC pathway. We were motivated to do so because many important players have been found and studied for many years, but important new genes have been discovered not too long ago^{92,93}, and a whole-genome siRNA study identified many new candidate proteins that influence MHC II surface expression and/or transcription²¹. To shed further light on this pathway, we have used a naïve Bayesian classifier. Through literature and KEGG pathway searches, we constructed a set of 86 positive genes known to be involved in the MHC pathway. We then scored other genes based on how much they behaved like MHC genes (our positive set) in all data sets to arrive at predictions for involvement with the MHC pathway.

Our data sets capture diverse and novel parts of what it means to be an MHC gene

As expected, MHC pathway genes were enriched in viral-human protein-protein interactions^{36,38,40,41}. We also discovered motifs enriched in MHC pathway gene promoters (2 kb window centred on the TSS)⁴⁶. Besides finding the motifs of known MHC-regulating transcription factors such as NF-κB, IRF, c-MYC and NFY, we identified enriched motifs for transcription factors that had not been previously associated with the MHC pathway (**Extended data 1**): HSFY2, for example, might curb MHC pathway gene expression in the testes. This example shows the potential of our approach, as, to the best of our knowledge, no interaction of HSFY2 with MHC pathway genes has been recorded.

The custom script we developed can be employed to find TFBS motifs of import to any set of genes. MHC pathway genes are enriched in all data sets (**Figure 2**). Every data set therefore manages to capture part of the MHC pathway gene identity.

The top ten candidates have interesting immune functionalities, but score high mainly in the three general data sets and lack direct links to MHC

Literature analysis of the top ten candidates shows that diverse immune-related genes dominate the top scores (**Extended data 2**). These genes have high scores in the three general data sets (immune tissue overrepresentation, macrophage expression profiles, virus-host PPI). These data sets score immune-relatedness, but are not as specific to the MHC pathway as the TFBS score and MHC II SEPS. For example, profile 11 is highly enriched in MHC pathway genes (**Figure 2E**), but many other genes are upregulated upon macrophage activation, and the vast majority of those are not

MHC pathway genes⁹⁴. These problems are further underlined by the gene with the tenth highest score, IFIH1 (**Extended data 2**). This gene is a dsRNA sensor of innate immunity⁹⁵. While it activates interferon signalling and thereby also activates MHC pathway genes, this gene does not conform to the definition of an MHC pathway gene as we mean it. Thus, based on these top ten candidates, it seems that the classifier selects more for involvement with immunity in lymphocytes and/or involvement in a basal process that, if perturbed, disturbs MHC II surface expression, rather than genes that are integral parts of the MHC pathway (**Extended data 2**). While this was not our aim, it should be noted that judging a classifier only on its top ten hits is not fair. Indeed, the ROC curve clearly shows that the combined data is better at picking out known MHC pathway genes than any individual data set (**Figure 5B**), and other immune pathways do score lower overall than known MHC pathway genes, even though some outliers score very high (**Figure 5A**). Therefore, a first step should be to analyse more candidates. Another approach would be to filter the top-ranking candidates on a high score in at least one of the MHC-specific data sets (TFBS score, MHC II surface expression). Currently, scores in the three general data sets are already high enough to secure a top score in the classifier, which is prevented in this way.

Another way to tackle this problem is to make a separate classifier for MHC I, MHC II, and immunoproteasomal genes. We chose to combine all three because we wanted a large enough positive set, but in the process we might well have lost specificity and important signals. Indeed, the fact that immunoproteasomal subunits have very high scores overall shows that the classifier is somehow geared towards finding immunoproteasome-like genes as well, something we did not expect (**Figure S2**). Since our custom script can easily accept different positive sets, separate classifiers are simple to make and could yield important insights. A final idea could be to make a more broad immune set, encompassing both innate and adaptive immune genes. We could then use the same data sets to train a classifier, and select candidates for experimental follow-up on the basis of a score that is higher in the MHC pathway gene classifier, relative to the broader immune classifier, effectively filtering for a true MHC signature.

Shortfalls of the (continuous) TFBS motif score

To calculate TFBS motif enrichment scores we used both the binning procedure used in previous works^{31–33}, and a direct fit of the positive and negative set score distributions (**see Methods**). The binned and continuous scores are similar, but the enrichment scores they award genes differ. We smoothed the actual TFBS motif score distributions because we attributed the decline and sharp rise in the positive set TFBS score to undersampling of the true MHC pathway gene repertoire. Our current score calculation method penalises genes that have less of the informative TFBS motifs in their promoters than the average amount in the positive set. However, having *more* motifs than MHC pathway genes is not penalised. Thus, a gene can be assigned a high TFBS motif score if it has many more motifs of a specific kind than known MHC pathway genes, which is contrary to the objective of selecting for genes that are most similar to MHC pathway genes. Therefore, one expects intermediate scores to dominate. Thus, the assumption behind smoothing may well be wrong. A solution to this problem is to get the distribution of the motif counts of all informative motifs in the MHC pathway promoters. The TFBS score can then be corrected for deviations from the distribution of motifs in the MHC pathway genes in either direction. This would make the MHC TFBS enrichment score better reflect true MHC identity.

As an additional point of improvement for the TFBS motif score, we now take the promoter region of a gene to be a 2 kb window surrounding its Transcription Start Site (TSS), and allow for motifs on both the plus and minus strand. In fact, promoters vary in length and position⁹⁶. Therefore, it would be an improvement to use ENCODE data to directly infer DNA occupancy around a gene, or to use data that pinpoints true promoter regions in a computational way⁹⁷, and search those sites for motifs. Nevertheless, our current, simpler, implementation finds many known motifs in addition to discovering heretofore unknown ones (**Extended data 1**).

Future perspectives

Constructing classifiers with only MHC I and MHC II genes as the positive set will be a priority, as this is easily implemented and will yield important insights relative to the combined classifier we now have. Moreover, one can pursue classifier voting or classifier ensembles, whereby one trains multiple separate classifiers on these data and pools their votes on whether a gene is an MHC gene or not, in order to increase accuracy^{98,99}. If different classifiers agree, it is more likely a gene is really involved. Alternative approaches to try would include SVMs and Random Forest classification, because of their ease of use, power, and ease of implementation^{100,101}. The ultimate goal would be an experimental screen of the top candidates in this classifier ensemble, in order to verify whether candidate genes are indeed MHC pathway genes.

Methods

Defining the positive (MHC pathway gene) and negative gene sets

In order to extract features informative for being an MHC pathway-related gene, we constructed a set of 86 genes known to be involved in the MHC pathway. This was based on the KEGG database and literature searches^{1,5,21,29,34}. We defined the negative set as all protein-coding genes in the human genome, bar the 86 known positive set genes. This negative set is thus both feasible and makes the classifier somewhat conservative in its estimates.

Transcription factor binding site data

Genes with Transcription Factor Binding Sites (TFBS) motifs that are found significantly more often in promoters of known MHC pathway genes than other genes are more likely themselves to be involved in the MHC-pathway. Two data sets capture that idea. The first is a genome-wide atlas of TFBS motifs which are evolutionarily constrained across 29 mammal species¹⁰². The second is a data set of genome-wide identification of motifs in human ChIP-seq data, incorporating both known literature motifs and *de novo* motif identification via various existing pipelines⁴⁶. Both datasets yielded the same core motifs from enrichment testing, such as IRF, NF- κ B and NF-Y (data not shown), but only the latter dataset was finally included in our analysis because most of the genes in our positive set were not in the 29 mammals data set, possibly because of low evolutionary constraint^{103–105}.

The motif dataset had 144,033,305 entries and included chromosome name, start and end site, and the strand for each identified motif. Data of all human protein-coding genes was downloaded from Ensembl (version Grch37p13)¹⁰⁶. Data consisted of these genes' Ensembl Gene ID, Gene Start (bp), Gene End (bp), Strand, Chromosome Name, Associated Gene Name and Description. Due to the size of the dataset, the motif file was split into 16 parts. A custom R script was used to map each motif to a 2-kb window around the Transcription Start Site of each gene (Gene Start (bp) in the Ensembl data). This 2-kb window was taken as an approximation of the promoter region. Then, the 16 separate resulting files were combined using a custom R script to achieve a final mapping of motifs to the promoter regions of all human protein-coding genes. This final data file gives, per gene, the amount of times a motif is present in it. This data file covers a total of 18,990 genes in the human genome.

The motifs called in the genome-wide ChIP-seq dataset were grouped into various transcription factor (TF) groups based on literature, and motifs were encoded as either newly discovered by motif scanning programmes (up to a maximum of 10 per motif TF group, and constrained by the fact that no single motif could be too close a match to any other motif within that TF group) or known from literature (as many as are known)⁴⁶. As a result, mappings in the data file were on the level of precise subspecies of motifs within a group bound by one TF. Generalising all sub-motifs to one

overarching motif species increases statistical power, allowing us to look at enrichment of an average overarching motif, rather than many subspecies. Using a custom R script, all motifs were generalised (removing the distinction between subtypes of a motif), which yielded the final dataset of generalised motifs. This final data was then used to calculate motif enrichments and scores for each TFBS (see TFBS enrichment analysis and genome-wide score calculations below)

Virus-host protein-protein interaction data

Data on protein-protein interactions (PPI) between host and viral proteins were obtained from three separate databases: PHISTO (Pathogen-Host Interaction Search Tool), VirHostNet 2.0, and HPIDB 2.0 (Host-Pathogen Interaction DataBase)^{43–45}. The three databases were downloaded, and filtered for virus-human interactions only. Obviously faulty entries were filtered out (for example, PHISTO featured 11 entries describing interactions with phages, but none of the mentioned source papers discussed protein-protein interaction with phages).

Protein-protein interactions were encoded with Uniprot Ids for the human proteins. These were translated to Ensembl IDs using BioMart¹⁰⁶. In cases where proteins mapped to multiple genes, we kept all genes. Selecting one correct mapping is difficult and a correct mapping need not exist due to multiple copies of a gene¹⁰⁷. Moreover, it is unlikely that double mappings are more prevalent in positive set genes than in negative set genes. Genes with one or more reported interaction(s) were classified as interactors, and those without as non-interactors. Genes not found in the databases were also classified as non-interactors. The union and intersect of these data sets was generated. Both data sets (union and enrichment) were then subjected to enrichment analysis (see virus-host protein-protein interaction enrichment analysis below)

Z-scores for affecting cell surface expression of peptide-loaded or CLIP-bound MHC II

Genes whose knockdown directly influences MHC II surface expression and/or peptide loading have a higher chance of being involved in the MHC pathway. The data set obtained from Paul *et al.* is a genome-wide multidimensional RNAi screen that was set up to assess the effect of knockout of 21,245 human genes on cell surface expression of MHC II, with four siRNA duplexes targetting each gene in a MeIJuSo cell line. Two antibodies were used to detect cell-surface effects of siRNA knockdowns on MHC II, one that binds peptide-loaded MHC II (L243), and one that binds to MHC II-CLIP (CerCLIP). siRNAs that target no genes were used as negative control, and siRNAs targetting known MHC II pathway genes (such as HLA-DM and HLA-DO) were used as positive control. A Z-score was then calculated for both L243 and CerCLIP fluorescence intensity. Extreme Z-scores indicated an effect of the target genes on MHC II cell surface expression. Target genes were rescreened, checked for expression in one or more actual immune cell types via microarray experiments, and siRNA effects were deconvoluted for a high confidence list of candidates that affect MHC II loading and surface expression in human immune cells²¹.

We need *one* score for the entire data set to include in the Bayesian classifier. Two separate scores would violate the conditional independence assumption. We took the maximum of the absolute of both Z-scores as the score for this data set. The absolute was chosen because the directionality of the effect is not important, and the maximum was chosen to estimate the strongest effect the gene could have. We used three versions of this data set: the initial Z-scores of all genes, the list of 789 target genes confirmed in rescreens, and the final list of 276 targets with expression in immune cells and verified on-target siRNA effects. All data sets were subjected to enrichment analysis (see Z-score enrichment analysis below). All genes in the human genome were sampled in this screen. Therefore, missing data (None) has meaning. Genes in this bin are those genes whose knockdown did not significantly affect MHC II surface expression. For the rescreened and high confidence scores, this bin also contains genes that were not significant in the rescreen, or that were not significant upon deconvolution and checking for immune cell expression, respectively. The final classifier is based on the final list of 276 target genes with known expression in immune cells, as

expression in immune cells is a vital characteristic of any candidate gene, and these scores performed better in terms of enrichment of the positive set.

Human macrophage activation microarray data

Data pertaining to up- and downregulation of genes in activated M1 macrophages (classically activated macrophages) and non-activated (control) macrophages over time was downloaded from the GEO database (Accession number: GSE57614), which originated from Derlindati *et al*⁴⁹. The authors performed time course experiments of the transcriptome of classically activated and non-activated macrophages using Agilent 4x44k Whole Human Genome Microarrays (version 2) at 6, 12, and 24 hours post-stimulation in triplicate. The activated macrophages showed a consistent pattern of up- and down-regulation of genes at all time points, and the authors thus chose to analyse the data using STEM (Short Time-series Expression Miner)⁴⁸ to define groups of genes that were similarly regulated over time. Following this approach, we calculated the normalised ratios of fluorescence of tested gene over reference data. For each replicate, the activated (M1) macrophage data was divided by the control macrophage data per time point. Log2 values of these data were used further. HGNC Gene Symbols for the Agilent probes were downloaded from Ensembl BioMart using the biomaRt package in R¹⁰⁸. Probes mapping to more than one gene were combined. The data was then loaded into STEM. STEM defines many random gene expression profiles and assigns genes to these profiles. It also determines whether certain expression profiles are found significantly more than expected. We used the same maximum number of profiles and unit changes (changes in gene expression level between time points) as Derlindati *et al.* (20 profiles; 3 unit changes)⁴⁹. However, we allowed a maximum correlation between profiles of 0.8 (above that, expression trajectories are merged) and used the FDR instead of Bonferroni correction, as the latter is often very strict¹⁰⁹. By default, the minimum gene expression change for a gene to be assigned a profile is 1. This means that genes that do not change their expression, or only very little, are not assigned a profile. The result was the assignment of every gene tested to one of 16 profiles found in the data, or an NA if the gene could not be assigned based on these criteria ('None' in Figure 2E). One gene was assigned to two profiles. We manually picked one and assigned it to that profile. The data generated was subjected to enrichment analysis (see macrophage activation expression changes below).

Human tissue immunohistochemistry data

Genes that are more highly expressed in immune cells or immune tissues than in other cells or tissues are more likely to be involved in immune pathways, such as the MHC (II) pathway, specifically the part of it that pertains to MHC II loading and expression. This data set incorporates that idea. Expression profiles for human proteins based on immunohistochemistry on tissue microarrays were downloaded from The Human Protein Atlas (<https://www.proteinatlas.org/about/download>; normal_tissue.tsv.zip; Human Protein Atlas version 18, based on Ensembl version 88.38)⁴⁷. This contains the expression of 10,618 human proteins in 82 cell types in 58 tissues. Expression levels are given in 4 categories (not detected, low, medium, and high) and reliability scores (Uncertain, Approved, Supported, Enhanced) are based on matching RNA expression, validation with independent antibodies, etc. Expression levels and the reliability score are manually curated. We selected on entries with a reliability score of Approved or above, and combined cell type and tissue into one variable to arrive at 131 unique combinations of a cell type and a tissue. Expression levels were encoded as a number from 0-3 based on the category, where 0 matches the score 'not detected'. We then defined immune tissue-cell type pairs (appendix lymphoid tissue, bone marrow hematopoietic cells, lung macrophages, lymph node germinal center cells, lymph node non-germinal center cells, skin 1 Langerhans, spleen cells in white pulp, spleen cells in red pulp, tonsil germinal center cells, tonsil squamous epithelial cells, tonsil non-germinal center cells, thymus medullary cells, thymus cortical cells) and labelled the rest as non-immune. We performed a Mann-Whitney U-test between the immune and non-immune tissues. We binned the

estimated rank differences to calculate enrichments (see human tissue immunohistochemistry enrichment analysis below)

Enrichment analyses

The below offers a detailed description of the procedures used to calculate enrichment of certain features in the positive set (known MHC pathway genes) relative to the negative set (whole genome – positive set). However, every implementation is derived from a common framework. We used Fisher’s exact test to calculate enrichment of a certain feature in a certain group of genes^{110,111}. The objective metric we use to assess how much more likely a certain gene is to be MHC pathway-related based on a certain dataset is the log2 likelihood ratio. This score can be directly used in the naïve Bayesian classifier (see Construction of the final naïve Bayesian classifier below). The score expresses the chance to be found in a certain bin if you are a positive set gene relative to if you are a negative set gene (odds ratio):

$$\text{Log}_2 \text{Likelihood Ratio} = \log_2 \left(\frac{\frac{\# \text{ of positive set genes in bin}}{\text{total \# of positive set genes}}}{\frac{\# \text{ of negative set genes in bin}}{\text{total \# of negative set genes}}} \right)$$

This log2 likelihood ratio is thus an objective measure of the strength of evidence that being in a certain score bin gives us about whether or not a gene might be MHC pathway-related.

TFBS motif enrichment analysis and genome-wide score calculations

We set out to find TFBS that were informative for being an MHC pathway-related gene. We checked, for all TFBS motifs present in the positive set, whether these were enriched or depleted relative to the negative set using Fisher’s exact test on the counts of these motifs. P-values were corrected for multiple testing by controlling the false discovery rate (FDR)¹⁰⁹. A TFBS motif was considered informative for MHC pathway identity if its corrected p-value was ≤ 0.10 , and the motif was present at least three times in the positive set. We took the sample estimate of enrichment of an informative TFBS motif as the score for that motif. We reasoned that similarity of motif count contains more evidence than mere presence/absence patterns. Therefore, we corrected the score for the average MHC pathway promoter count for enriched motifs:

$$\text{Enriched motif score} = \text{sample est. enrichment} * \frac{\text{motif count in gene}}{\text{average motif count in MHC pathway genes}}$$

Scores were tallied for all informative TFBS to arrive at a total additive MHC pathway TFBS score for each gene. This included unaltered negative scores for having a TFBS motif that is depleted in MHC pathway gene promoters.

4507 genes have a TFBS score of NA. In mapping the data from the paper, only 20,327 out of 22,837 Ensembl genes’ promoter regions could be found in the data. Of those 20,327 genes, 1,337 had no TFBS mapped to their promoter regions (defined by us as the 2 kb window surrounding their transcription start site, on both plus and minus strands). This left a total of 18,990 genes with one or multiple TFBS motifs in their promoter region. Thus, 3847 genes have no data. At present, the bin with NA values (“None”) is scored, because upon sampling all protein-coding genes, these are the only genes that have no TFBS motifs. Therefore, this constitutes a signal.

We use the positive set to deduce TFBS motifs informative for MHC pathway identity and then score all genes. This means we also scores positive set genes based on a signature derived from them. This recursive characteristic is one form of overfitting¹¹². Therefore, we used a ten-fold cross-validation¹¹³. The 86 positive set genes and 22,271 negative set genes were randomly divided into ten sets: 6 sets with 9 positive set genes, and 4 with 8. Similarly: 9 negative sets with 2,227 genes, and 1 with 2,228 genes. We took 9 of the positive subsets and 9 of the negative subset, used these to calculate the which motifs were enriched in the positive set, and scored the unseen subset of both

the positive and negative set based on the presence of these TFBS in their promoter regions. The TFBS motif score thus obtained has a minimum of -13.85 and a maximum of 19.96, with a median of 0.5 (for the cross-validation seed '1234567890', values differ slightly for different cross-validation sets).

Bins were manually determined for this dataset such that each bin conferred a higher log2 likelihood of being an MHC pathway-related gene. This was done to avoid nonsense scores of $-\infty$ and because we expect that bins with higher scores should give progressively higher likelihoods, as these bins contain genes that should be increasingly MHC-like in their TFBS motif signature (though see Discussion). The bins chosen were: (-14, -4], (-4, -2], (-2, 2], (2, 4], (4, 6], (6, 8], (8, 14], (14, 22]. See Supplementary Material for differences in scoring between our custom bins and bins of width one.

To directly fit the distributions of positive set and negative set TFBS motif scores, we used kernel density estimation on the scores to remove putative undersampled regions (though see Discussion) and arrive at a conservative score (density function in R, $bw = 2.5$). We then fitted a 20th degree polynomial to the positive and negative score distributions. Using the continuous fitted function for the positive and negative set, we can determine a specific log2 likelihood ratio by dividing the values of the functions at that point by each other for each unique total TFBS score. Because the values in the extremes of the curves were extremely high or extremely low, scores were bounded (**Supplementary Table 1; Supplementary Figure 6Supplementary Figure 7**). A minimum was selected at the negative end of the score, and a maximum at the positive end of the score. Scores that were extremely high due to very low values of either fitted function at high/low values were bounded to this maximum/minimum. The final classifier incorporates the continuous scores.

Virus-host protein-protein interaction enrichment analysis

We performed enrichment analysis for every data set separately (PHISTO, HPIDB 2.0 and VirHostNet 2.0; data not shown) and for the union and intersect of all three data sets. The union was incorporated in the final Bayesian classifier.

Z-score enrichment analysis

The Z-scores for the three data sets (the initial results of the siRNA screen against all genes, the lower confidence predictions, which covers 789 genes that were rescreened, and the high confidence predictions, which cover 276 genes that are known to be expressed in immune cells, and were corrected for off-target effects) were binned in two bins: (0, 2] and (2, \rightarrow). There are no bins for negative scores because we took the absolute of the Z-score measured in the screen. Enrichment of positive set genes was calculated as log2 likelihood as above. We used the 276 high confidence candidates in the final Bayesian classifier.

Macrophage activation expression change enrichment analysis

Log2 enrichment calculations were performed as described. Since the programme was constrained to unit changes of at least one between time points to assign profiles, the 'None'-profile captures genes whose expression does not increase or decrease. Therefore, this profile is a valid category, and is treated as a bin in the enrichment analysis.

Human tissue immunohistochemistry enrichment analysis

Scores for rank-based overrepresentation in immune tissues were binned in custom bins: (-2, 0]; (0, 1]; (1, 1.5]; (1.5, 3]. Log2 enrichment analysis was performed as previously described. Currently, genes with NA values are assigned a score. These arise because of inexistent cell type-tissue combinations, and because these genes are simply absent from the data. The former are true missing values, whereas the latter might still have a signal. The current score for genes in the 'None'-bin was thus interpreted with caution.

Construction of the final naïve Bayesian classifier

The construction of the final Bayesian classifier was done by adding up, per gene, the log 2 likelihood ratio per data set. Median log2 enrichment scores from all cross-validations can be seen in Figure 2. The following explains exactly how Bayes' formula works in this context, and what we set as the prior for our classifier.

Using Bayes' theorem to generate an odds ratio of being an MHC pathway gene

Bayes' theorem states the following:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

This states that the chance of A given B is the same as the chance of the opposite (B given A) times the chance of A, divided by the chance of B. This formalism is derived from simple conditional probability. For predicting novel candidate genes in the MHC pathway, it can be rewritten as follows:

$$P(MHC\ gene|data) = \frac{P(data|MHC\ gene) * P(MHC\ gene)}{P(data)}$$

Given that we have a known set of MHC pathway genes, we can calculate the right hand side of this formula and therefore predict every gene's chance of being an MHC pathway gene.

$P(data | MHC\ gene)$ represents the chance of observing some data (for example, a measured host-viral PPI), given that you are an MHC gene. This is simply the amount of MHC genes that have measured host-viral PPI divided by the total number of MHC genes. $P(MHC\ gene)$ is the prior probability of being an MHC gene, which is the (expected) number of MHC genes divided by the total number of protein-coding genes in the human genome (see calculation of the prior below).


$P(data)$ is simply the chance of observing a data point at all. For the 'yes'-bin in the host-viral PPI, this would be the amount of genes in that bin, divided by the total amount of protein-coding genes in the data.

We have only two outcomes, either you are predicted to be an MHC gene, or you are not. We can thus use an odds ratio, which gives the relative chance that you are one or the other³². Note that we can do the exact same thing as above using 'not MHC gene':


$$P(not\ MHC\ gene|data) = \frac{P(data|not\ MHC\ gene) * P(not\ MHC\ gene)}{P(data)}$$

We can then divide these by each other, and thus receive the odds ratio of being an MHC gene:


$$\frac{P(MHC\ gene|data)}{P(not\ MHC\ gene|data)} = \frac{P(MHC\ gene)}{P(not\ MHC\ gene)} * \frac{P(data|MHC\ gene)}{P(data|not\ MHC\ gene)}$$



Odds of being
an MHC gene



Prior odds



Updating
based on data

Here we have separated the terms into the prior odds (relative chance of being an MHC gene

without using knowledge from data) and the Bayesian updating based on data. Note that $P(data)$ has disappeared from the equation in the process.

Determining the prior odds

The prior odds in the equation above need to be estimated. Given that there are 86 positive set genes, we estimated that at most about double the amount of MHC pathway genes might exist: 200. We thus set the prior chance of being an MHC gene to the following:

$$Prior = \frac{\frac{200}{22,357}}{\frac{(22,357-200)}{22,357}} \approx 0.009$$

The prior odds are the same for every gene, and thus do not influence the eventual ranks of genes in the classifier. That is based solely on data. We need this information only for later calculations of the false discovery rate, true discovery rate, sensitivity and specificity³². These calculations will be performed in future work.

Integration and final score

For ease of interpretation, we take the log2 of the odds ratio. This prevents carry-over of rounding errors (since adding log numbers is the same as multiplying their constituent numbers), and it allows an easy read-out: any bin with a negative log2 likelihood ratio is evidence to the contrary of being an MHC gene, while positive values mean an increase in the chance of being an MHC gene. Thus, for each gene, in each data set, we added up the log2 likelihood ratio of the bin (or the continuous log2 likelihood score) and added the prior odds to arrive at the final naïve Bayesian classifier score. For a very thorough description of Bayesian methodology, please consult reference ³².

Literature analysis of candidates and TFBS motifs

We followed up on the identification of important TFBS motifs in MHC pathway gene promoters and on the genes with the top ten scores in the negative set via a literature search. The gene name (or some of its synonyms, taken from GeneCards¹¹⁴) were entered in Google scholar, either on their own, or with additional keywords such as ‘MHC’, ‘immune’, ‘immunity’, ‘immune cells’, or ‘proteasome’. The results were scoured for mentions of the genes of interest and possible relations to immune functions. The results of this literature analysis can be found in the Supplementary Material (**Extended data 1 and 2**).

References

1. Rock, K. L., Reits, E. & Neefjes, J. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends Immunol.* **37**, 724–737 (2016).
2. Thielens, A., Vivier, E. & Romagné, F. NK cell MHC class I specific receptors (KIR): From biology to clinical intervention. *Curr. Opin. Immunol.* **24**, 239–245 (2012).
3. Downs, I., Vijayan, S., Sidiq, T. & Kobayashi, K. S. CITA/NLRC5: A critical transcriptional regulator of MHC class I gene expression. *BioFactors* **42**, 349–357 (2016).
4. Pamer, E. & Cresswell, P. MECHANISMS OF MHC CLASS I-RESTRICTED ANTIGEN PROCESSING. *Annu. Rev. Immunol.* **16**, 323–358 (1998).
5. Boyle, L. H. *et al.* Tapasin-related protein TAPBPR is an additional component of the MHC class I presentation pathway. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 3465–3470 (2013).
6. Morozov, G. I. *et al.* Interaction of TAPBPR, a tapasin homolog, with MHC-I molecules promotes peptide editing. *Proc. Natl. Acad. Sci.* **113**, E1006–E1015 (2016).
7. Zhang, Y. & Williams, D. B. Assembly of MHC class I molecules within the endoplasmic reticulum. *Immunol. Res.* **35**, 151–162 (2006).
8. Paulsson, K. & Wang, P. Chaperones and folding of MHC class I molecules in the endoplasmic reticulum. *Biochim. Biophys. Acta - Mol. Cell Res.* **1641**, 1–12 (2003).
9. Hulpke, S. & Tampé, R. The MHC I loading complex: a multitasking machinery in adaptive immunity. *Trends Biochem. Sci.* **38**, 412–420 (2013).
10. Kambayashi, T. & Laufer, T. M. Atypical MHC class II-expressing antigen-presenting cells: Can anything replace a dendritic cell? *Nat. Rev. Immunol.* **14**, 719–730 (2014).
11. van Eggermond, M. C. J. A., Tezcan, I., Heemskerk, M. H. M. & van den Elsen, P. J. Transcriptional silencing of RFXAP in MHC class II-deficiency. *Mol. Immunol.* **45**, 2920–2928 (2008).
12. Hume, D. A. Macrophages as APC and the Dendritic Cell Myth. *J. Immunol.* **181**, 5829–5835 (2008).
13. Neerincx, A., Castro, W., Guarda, G. & Kufer, T. A. NLRC5, at the heart of antigen presentation. *Front. Immunol.* **4**, 1–10 (2013).
14. Davoust, J. & Banchereau, J. Naked antigen-presenting molecules on dendritic cells. *Nat. Cell Biol.* **2**, E46–E48 (2000).
15. Andersson, T., Patwardhan, A., Emilson, A., Carlsson, K. & Scheynius, A. HLA-DM is expressed on the cell surface and colocalizes with HLA-DR and invariant chain in human Langerhans cell. *Arch. Dermatol. Res.* **290**, 674–680 (1998).
16. Nakagawa, T. Y. & Rudensky, A. Y. The role of lysosomal proteinases in MHC class II-mediated antigen processing and presentation. *Immunol. Rev.* **172**, 121–129 (1999).
17. Miller, M. A., Ganesan, A. P. V. & Eisenlohr, L. C. Toward a network model of MHC class II-restricted antigen processing. *Front. Immunol.* **4**, 1–8 (2013).
18. Lenz, T. L. *et al.* Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat. Genet.* **47**, 1085–1090 (2015).
19. Weiskopf, D. *et al.* HLA-DRB1 Alleles Are Associated with Different Magnitudes of Dengue Virus-Specific CD4+T-Cell Responses. *J. Infect. Dis.* **214**, 1117–1124 (2016).
20. Fitzmaurice, K. *et al.* Additive effects of HLA alleles and innate immune genes determine viral outcome in HCV infection. *Gut* **64**, 813–819 (2015).
21. Paul, P. *et al.* A genome-wide multidimensional RNAi screen reveals pathways controlling MHC class II antigen presentation. *Cell* **145**, 268–283 (2011).
22. Sharon, E. *et al.* Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat. Genet.* **48**, 995–1002 (2016).
23. Pellegrino, P. *et al.* The first steps towards the era of personalised vaccinology: Predicting adverse reactions. *Pharmacogenomics J.* **15**, 284–287 (2015).
24. van den Hoorn, T., Paul, P., Jongsma, M. L. M. & Neefjes, J. Routes to manipulate MHC class II antigen presentation. *Curr. Opin. Immunol.* **23**, 88–95 (2011).

25. Anderson, K. S. & Cresswell, P. A role for calnexin (IP90) in the assembly of class II MHC molecules. *EMBO J.* **13**, 675–82 (1994).
26. Zhu, X. S. *et al.* Transcriptional scaffold: CIITA interacts with NF- κ B, RFX, and CREB to cause stereospecific regulation of the class II major histocompatibility complex promoter. *Mol. Cell. Biol.* **20**, 6051–61 (2000).
27. Westerheide, J. M. *et al.* The MHC-Specific Enhanceosome and Its Role in MHC Class I and β 2-Microglobulin Gene Transactivation. *J Immunol Res.* **167**, 5175–5184 (2001).
28. Ellgaard, L. & Frickel, E.-M. Calnexin, calreticulin, and ERp57: teammates in glycoprotein folding. *Cell Biochem. Biophys.* **39**, 223–247 (2003).
29. Kobayashi, K. S. & van den Elsen, P. J. NLRC5: a key regulator of MHC class I-dependent immune responses. *Nat. Rev. Immunol.* **12**, 813–820 (2012).
30. Hermann, C., Trowsdale, J. & Boyle, L. H. TAPBPR: a new player in the MHC class I presentation pathway. *Tissue Antigens* **85**, 155–166 (2015).
31. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* **44**, D1251–D1257 (2016).
32. van Dam, T. J. P. *et al.* CiliaCarta: An Integrated And Validated Compendium Of Ciliary Genes. *bioRxiv* 123455 (2017). doi:10.1101/123455
33. van der Lee, R. *et al.* Integrative Genomics-Based Discovery of Novel Regulators of the Innate Antiviral Response. *PLOS Comput. Biol.* **11**, e1004553 (2015).
34. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
35. Kloetzel, P. M. & Ossendorp, F. Proteasome and peptidase function in MHC-class-I-mediated antigen presentation. *Curr. Opin. Immunol.* **16**, 76–81 (2004).
36. Schuren, A. B. C., Costa, A. I. & Wiertz, E. J. H. J. Recent advances in viral evasion of the MHC Class I processing pathway. *Curr. Opin. Immunol.* **40**, 43–50 (2016).
37. González-Motos, V., Kropp, K. A. & Viejo-Borbolla, A. Chemokine binding proteins: An immunomodulatory strategy going viral. *Cytokine Growth Factor Rev.* **30**, 71–80 (2016).
38. Heidarieh, H., Hernández, B. & Alcamí, A. Immune modulation by virus-encoded secreted chemokine binding proteins. *Virus Res.* **209**, 67–75 (2015).
39. Lučin, P., Mahmutefendić, H., Blagojević Zagorac, G. & Ilić Tomaš, M. Cytomegalovirus immune evasion by perturbation of endosomal trafficking. *Cell. Mol. Immunol.* **12**, 154–169 (2015).
40. Miller, D. M. *et al.* Human cytomegalovirus inhibits major histocompatibility complex class II expression by disruption of the Jak/Stat pathway. *J. Exp. Med.* **187**, 675–83 (1998).
41. Zuo, J. & Rowe, M. Herpesviruses placating the unwilling host: Manipulation of the MHC class II antigen presentation pathway. *Viruses* **4**, 1335–1353 (2012).
42. Alcamí, A. & Koszinowski, U. H. Viral mechanisms of immune evasion. *Trends Microbiol.* **8**, 410–418 (2000).
43. Ammari, M. G., Gresham, C. R., McCarthy, F. M. & Nanduri, B. HPIDB 2.0: a curated database for host–pathogen interactions. *Database* **2016**, baw103 (2016).
44. Guirimand, T., Delmotte, S. & Navratil, V. VirHostNet 2.0: Surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res.* **43**, D583–D587 (2015).
45. Durmuş Tekir, S. *et al.* PHISTO: Pathogen-host interaction search tool. *Bioinformatics* **29**, 1357–1358 (2013).
46. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).
47. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347**, 1260419–1260419 (2015).
48. Ernst, J. & Bar-Joseph, Z. STEM: A tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7**, 1–11 (2006).

49. Derlindati, E. *et al.* Transcriptomic analysis of human polarized macrophages: More than one role of alternative activation? *PLoS One* **10**, 1–17 (2015).
50. Shirley Liu, X., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* **20**, 835–839 (2002).
51. Eden, E., Lipson, D., Yogeve, S., Yakhini, Z. & Caldwell, C. Discovering Motifs in Ranked Lists of DNA Sequences. *PLoS Comput. Biol.* **3**, e39 (2007).
52. Hoesel, B. & Schmid, J. A. The complexity of NF- κ B signaling in inflammation and cancer. *Mol. Cancer* **12**, 86 (2013).
53. Huang, B., Qi, Z. T., Xu, Z. & Nie, P. Global characterization of interferon regulatory factor (IRF) genes in vertebrates: glimpse of the diversification in evolution. *BMC Immunol.* **11**, 22 (2010).
54. Vasquez, K. *et al.* IRF-7 is the master regulator of type-I interferon-dependent immune responses. *Nature* **434**, 772–777 (2005).
55. Lochamy, J., Rogers, E. M. & Boss, J. M. CREB and phospho-CREB interact with RFX5 and CIITA to regulate MHC class II genes. *Mol. Immunol.* **44**, 837–847 (2007).
56. Voss, T. C. & Hager, G. L. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.* **15**, 69–81 (2014).
57. Scott, V., Morgan, E. A. & Stadler, H. S. Genitourinary functions of Hoxa13 and Hoxd13. *J. Biochem.* **137**, 671–676 (2005).
58. Salsi, V., Vigano, M. A., Cocchiarella, F., Mantovani, R. & Zappavigna, V. Hoxd13 binds in vivo and regulates the expression of genes acting in key pathways for early limb and skeletal patterning. *Dev. Biol.* **317**, 497–507 (2008).
59. Zhao, G. N., Jiang, D. S. & Li, H. Interferon regulatory factors: At the crossroads of immunity, metabolism, and disease. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1852**, 365–378 (2015).
60. Gerondakis, S. *et al.* Unravelling the complexities of the NF- κ B signalling pathway using mouse knockout and transgenic models. *Oncogene* **25**, 6781–6799 (2006).
61. Sander, T. L. *et al.* The SCAN domain defines a large family of zinc finger transcription factors. *Gene* **310**, 29–38 (2003).
62. Bailey, T. L. *et al.* MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **37**, 202–208 (2009).
63. Shih, V. F.-S. *et al.* Control of RelB during dendritic cell activation integrates canonical and noncanonical NF- κ B pathways. *Nat. Immunol.* **13**, 1162–1170 (2012).
64. Shuai, K. & Liu, B. Regulation of JAK–STAT signalling in the immune system. *Nat. Rev. Immunol.* **3**, 900–911 (2003).
65. Nasmyth, K. & Haering, C. H. Cohesin: Its Roles and Mechanisms. *Annu. Rev. Genet.* **43**, 525–558 (2009).
66. Merckenschlager, M. Cohesin: A global player in chromosome biology with local ties to gene regulation. *Curr. Opin. Genet. Dev.* **20**, 555–561 (2010).
67. Mielnik, A.; Schlegel, P. N.; Paduch, D. A. TESTICULAR EXPRESSION ANALYSIS OF THE AZF GENES IN AZOOSPERMIC MEN SUGGESTS ESSENTIALITY AND SPECIFIC FUNCTION FOR DDX3Y, RPS4Y2, CDY2, AND HSFY. *Fertil. Steril.* **94**, S232 (2014).
68. Li, N., Wang, T. & Han, D. Structural, cellular and molecular aspects of immune privilege in the testis. *Front. Immunol.* **3**, 1–12 (2012).
69. Kniepert, A. & Groettrup, M. The unique functions of tissue-specific proteasomes. *Trends Biochem. Sci.* **39**, 17–24 (2014).
70. Norris, J. D. *et al.* The Homeodomain Protein HOXB13 Regulates the Cellular Response to Androgens. *Mol. Cell* **36**, 405–416 (2009).
71. Pérez-Cabrera, A., Kofman-Alfaro, S. & Zenteno, J. C. Mutational analysis of HOXD13 and HOXA13 genes in the triphalangeal thumb-brachyectrodactyly syndrome. *J. Orthop. Res.* **20**,

- 899–901 (2002).
72. Economides, K. D. Hoxb13 is required for normal differentiation and secretory function of the ventral prostate. *Development* **130**, 2061–2069 (2003).
73. Zadrozny, B. & Elkan, C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *Icml* 1–8 (2001).
74. Santambrogio, L. *et al.* Extracellular antigen processing and presentation by immature dendritic cells. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 15056–61 (1999).
75. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *October* **302**, 249–255 (2003).
76. Reja, R., Vinayachandran, V., Ghosh, S. & Pugh, B. F. Molecular mechanisms of ribosomal protein gene coregulation. 1942–1954 (2015). doi:10.1101/gad.268896.115.
77. Winter, D. R., Jung, S. & Amit, I. Making the case for chromatin profiling: a new tool to investigate the immune-regulatory landscape. (2015).
78. Boss, J. M. & Jensen, P. E. Transcriptional regulation of the MHC class II antigen presentation pathway. *Curr. Opin. Immunol.* **15**, 105–111 (2003).
79. Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Casp. J. Intern. Med.* **4**, 627–635 (2013).
80. Moran, M. M. *et al.* *Encyclopedia of Signaling Molecules*. *Nat Rev Drug Discov* **10**, (2013).
81. Holling, T. M., Schooten, E. & Van Den Elsen, P. J. Function and regulation of MHC class II molecules in T-lymphocytes: Of mice and men. *Hum. Immunol.* **65**, 282–290 (2004).
82. Fornari, T. A. *et al.* Comprehensive Survey of miRNA-mRNA Interactions Reveals That Ccr7 and Cd247 (CD3 zeta) are Posttranscriptionally Controlled in Pancreas Infiltrating T Lymphocytes of Non-Obese Diabetic (NOD) Mice. *PLoS One* **10**, 1–22 (2015).
83. Bombeiro, A. L., Hell, R. C. R., Simões, G. F., Castro, M. V. de & Oliveira, A. L. R. de. Importance of major histocompatibility complex of class I (MHC-I) expression for astroglial reactivity and stability of neural circuits in vitro. *Neurosci. Lett.* **647**, 97–103 (2017).
84. Xu, H. ping *et al.* The Immune Protein CD3ζ Is Required for Normal Development of Neural Circuits in the Retina. *Neuron* **65**, 503–515 (2010).
85. He, T., Mortensen, X., Wang, P. & Tian, N. The effects of immune protein CD3ζ development and degeneration of retinal neurons after optic nerve injury. *PLoS One* **12**, 1–18 (2017).
86. Kao, S. C. *et al.* Identification of phostensin, a PP1 F-actin cytoskeleton targeting subunit. *Biochem. Biophys. Res. Commun.* **356**, 594–598 (2007).
87. Lin, Y. S. *et al.* Immunolocalization of phostensin in lymphatic cells and tissues. *J. Histochem. Cytochem.* **59**, 741–749 (2011).
88. Matsubara, T. *et al.* The Actin-Binding Protein PPP1r18 Regulates Maturation, Actin Organization, and Bone Resorption Activity of Osteoclasts. 1–15 (2018).
89. Takayanagi, H. New immune connections in osteoclast formation. *Ann. N. Y. Acad. Sci.* **1192**, 117–123 (2010).
90. Izumikawa, K. *et al.* Chtop (Chromatin target of Prmt1) auto-regulates its expression level via intron retention and nonsense-mediated decay of its own mRNA. *Nucleic Acids Res.* **44**, gkw831 (2016).
91. Izumikawa, K., Ishikawa, H., Simpson, R. J. & Takahashi, N. Modulating the expression of Chtop, a versatile regulator of gene-specific transcription and mRNA export. *RNA Biol.* **0**, 1–29 (2018).
92. Hermann, C. *et al.* TAPBPR alters MHC class I peptide presentation by functioning as a peptide exchange catalyst. *Elife* **4**, e09617 (2015).
93. Meissner, T. B. *et al.* NLR family member NLRC5 is a transcriptional regulator of MHC class I genes. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 13794–9 (2010).
94. Murray, P. J. *et al.* Macrophage Activation and Polarization: Nomenclature and Experimental Guidelines. *Immunity* **41**, 14–20 (2014).
95. Rice, G. I. *et al.* Gain-of-function mutations in IFIH1 cause a spectrum of human disease

- phenotypes associated with upregulated type I interferon signaling. *Nat. Genet.* **46**, 503–509 (2014).
96. Ohler, U. & Niemann, H. Identification and analysis of eukaryotic promoters: Recent computational approaches. *Trends Genet.* **17**, 56–60 (2001).
 97. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
 98. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **33**, 1–39 (2010).
 99. Ruta, D. & Gabrys, B. Classifier selection for majority voting. *Inf. Fusion* **6**, 63–81 (2005).
 100. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R news* **2**, 18–22 (2002).
 101. Chorowski, J., Wang, J. & Zurada, J. M. Review and performance comparison of SVM- and ELM-based classifiers. *Neurocomputing* **128**, 507–516 (2014).
 102. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
 103. Hans, J. B., Bergl, R. A. & Vigilant, L. Gorilla MHC class I gene and sequence variation in a comparative context. *Immunogenetics* **69**, 303–323 (2017).
 104. Shiina, T. *et al.* Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* **173**, 1555–1570 (2006).
 105. Spurgin, L. G. & Richardson, D. S. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. R. Soc. B Biol. Sci.* **277**, 979–988 (2010).
 106. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016).
 107. Sudmant, P. H. *et al.* Diversity of Human Copy Number Variation and Multicopy Genes. *11184*, 2–7 (2010).
 108. Durinck, S. *et al.* BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
 109. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125**, 279–284 (2001).
 110. Rivals, I., Personnaz, L., Taing, L. & Potier, M. C. Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* **23**, 401–407 (2007).
 111. Upton, G. J. G. Fisher's Exact Test. *J. R. Stat. Soc.* **155**, 395–402 (1992).
 112. Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **44**, 1–12 (2004).
 113. Krstajic, D., Buturovic, L. J., Leahy, D. E. & Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* **6**, 1–15 (2014).
 114. Stelzer, G. *et al.* The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.* **2016**, 1.30.1–1.30.33 (2016).
 115. Stoep, N. Van Der, Quinten, E., Rezende, M. M. & Elsen, P. J. Van Den. E47, IRF-4, and PU.1 synergize to induce B-cell – specific activation of the class II transactivator promoter III (CIITA-PIII). *Society* **104**, 2849–2857 (2004).
 116. Stein, M. F. *et al.* Multiple interferon regulatory factor and NF-κB sites cooperate in mediating cell-type- and maturation-specific activation of the human CD83 promoter in dendritic cells. *Mol. Cell. Biol.* **33**, 1331–44 (2013).
 117. Chen, K., Liu, J. & Cao, X. Regulation of type I interferon signaling in immunity and inflammation: A comprehensive review. *J. Autoimmun.* **83**, 1–11 (2017).
 118. Serasanambati, M. & Chilakapati, S. R. Function of Nuclear Factor Kappa B (NF-κB) in Human Diseases-A Review. *South Indian J. Biol. Sci.* **2**, 368 (2016).
 119. Vallabhapurapu, S. & Karin, M. Regulation and Function of NF-κB Transcription Factors in the Immune System. *Annu. Rev. Immunol.* **27**, 693–733 (2009).
 120. Bonizzi, G. & Karin, M. The two NF-κB activation pathways and their role in innate and adaptive immunity. *Trends Immunol.* **25**, 280–288 (2004).
 121. Das, B. & Majumder, D. Interactions of Transcription Factors in HLA Class I Transcriptosome. **6**, 592–602 (2014).

122. Zhang, Q., Lenardo, M. J. & Baltimore, D. 30 Years of NF- κ B: A Blossoming of Relevance to Human Pathobiology. *Cell* **168**, 37–57 (2017).
123. Sachini, N. & Papamatheakis, J. NF- κ B and the immune response: Dissecting the complex regulation of MHC genes. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1860**, 537–542 (2016).
124. Choi, N. M., Majumder, P. & Boss, J. M. Regulation of major histocompatibility complex class II genes. *Curr. Opin. Immunol.* **23**, 81–87 (2011).
125. Ren, G., Cui, K., Zhang, Z. & Zhao, K. Division of labor between IRF1 and IRF2 in regulating different stages of transcriptional activation in cellular antiviral activities. *Cell Biosci.* **5**, 17 (2015).
126. Tanaka, H. *et al.* Epigenetic Regulation of the Blimp-1 Gene (Prdm1) in B Cells Involves Bach2 and Histone Deacetylase 3. *J. Biol. Chem.* **291**, 6316–6330 (2016).
127. Gateva, V. *et al.* A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat. Genet.* **41**, 1228–1233 (2009).
128. Raychaudhuri, S. *et al.* Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nat. Genet.* **41**, 1313–1318 (2009).
129. Kim, S. J. *et al.* Increased cathepsin S in Prdm1-/- dendritic cells alters the T FH cell repertoire and contributes to lupus. *Nat. Immunol.* **18**, 1016–1024 (2017).
130. Fang, F. *et al.* A PAX5–OCT4–PRDM1 developmental switch specifies human primordial germ cells. *Nat. Cell Biol.* (2018). doi:10.1038/s41556-018-0094-3
131. Nitzsche, A. *et al.* RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS One* **6**, (2011).
132. Doucet-Beaupré, H. *et al.* Lmx1a and Lmx1b regulate mitochondrial functions and survival of adult midbrain dopaminergic neurons. *Proc. Natl. Acad. Sci.* **113**, E4387–E4396 (2016).
133. Doucet-Beaupré, H., Ang, S. L. & Lévesque, M. Cell fate determination, neuronal maintenance and disease state: The emerging role of transcription factors Lmx1a and Lmx1b. *FEBS Lett.* **589**, 3727–3738 (2015).
134. Rasclé, A. *et al.* The LIM-homeodomain transcription factor LMX1B regulates expression of NF-kappa B target genes. *Exp. Cell Res.* **315**, 76–96 (2009).
135. Muragaki, Y., Mundlos, S., Upton, J. & Olsen, B. R. Altered Growth and Branching Patterns in Synpolydactyly Caused by Mutations in HOXD Author (s): Yasuteru Muragaki , Stefan Mundlos , Joseph Upton and Bjorn R . Olsen Published by : American Association for the Advancement of Science Stable URL : <http://>. **272**, 548–551 (1996).
136. Perkins, N. D. The diverse and complex roles of NF- κ B subunits in cancer. *Nat. Rev. Cancer* (2012). doi:10.1038/nrc3204
137. Hailfinger, S. *et al.* Malt1-dependent RelB cleavage promotes canonical NF-kappaB activation in lymphocytes and lymphoma cell lines. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 14596–14601 (2011).
138. Orso, F. & Taverna, D. TFAP2A (transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)). *Atlas Genet. Cytogenet. Oncol. Haematol.* **14**, 735–738 (2011).
139. Jackson, B. C., Nebert, D. W. & Vasilou, V. Update of human and mouse forkhead box (FOX) gene families. *Hum. Genomics* **4**, 194–201 (2010).
140. Jonsson, H. & Peng, S. L. Forkhead transcription factors in immunology. *Cell. Mol. Life Sci.* **62**, 397–409 (2005).
141. Johansson, C. C. *et al.* A winged helix forkhead (FOXD2) tunes sensitivity to cAMP in T lymphocytes through regulation of cAMP-dependent protein kinase RI α . *J. Biol. Chem.* **278**, 17573–17579 (2003).
142. Shah, N. *et al.* HOXB13 mediates tamoxifen resistance and invasiveness in human breast cancer by suppressing ER α and inducing IL-6 expression. *Cancer Res.* **73**, 5449–5458 (2013).
143. HEINRICH, P. C. *et al.* Principles of interleukin (IL)-6-type cytokine signalling and its

- regulation. *Biochem. J.* **374**, 1–20 (2003).
144. Dang, C., Le, A. & Gao, P. MYC-induced Cancer Cell Energy Metabolism and Therapeutic Opportunities. *Clin. Cancer Res.* **15**, 6479–6483 (2009).
 145. Rathmell, J. C. T Cell Myc-metabolism. *Immunity* **35**, 845–846 (2011).
 146. Wang, R. *et al.* The Transcription Factor Myc Controls Metabolic Reprogramming upon T Lymphocyte Activation. *Immunity* **35**, 871–882 (2011).
 147. Stephanie C. Casey, Ling Tong, Yulin Li, Rachel Do, Susanne Walz, Kelly N. Fitzgerald, Arvin M. Gouw, Virginie Baylot, Ines Gütgemann, Martin Eilers, D. W. F. MYC regulates the antitumor immune response through CD47 and PD-L1. **9935**, (2016).
 148. Versteeg, R., Noordermeer, I. A., Krüse-Wolters, M., Ruiter, D. J. & Schrier, P. I. c-myc down-regulates class I HLA expression in human melanomas. *EMBO J.* **7**, 1023–9 (1988).
 149. Peltenburg, L. T. C. & Schrier, P. I. Transcriptional suppression of HLA-B expression by c-Myc is mediated through the core promoter elements. *Immunogenetics* **40**, 54–61 (1994).
 150. Yuasa, S. *et al.* Zac1 is an essential transcription factor for cardiac morphogenesis. *Circ. Res.* **106**, 1083–1091 (2010).
 151. Abdollahi, A. *et al.* LOT1 (PLAGL1/ZAC1), the candidate tumor suppressor gene at chromosome 6q24-25, is epigenetically regulated in cancer. *J. Biol. Chem.* **278**, 6041–6049 (2003).
 152. Quagliata, L. *et al.* Long noncoding RNA HOTTIP/HOXA13 expression is associated with disease progression and predicts outcome in hepatocellular carcinoma patients. *Hepatology* **59**, 911–923 (2014).
 153. Pockley, A. G. Heat shock proteins as regulators of the immune response. *Lancet* **362**, 469–476 (2003).
 154. Srivastava, P. Roles of Heat-Shock Proteins in Innate and Adaptive Immunity. *Nat. Rev. Immunol.* **2**, 185–194 (2002).
 155. Sharff, K. A. *et al.* Hey1 basic helix-loop-helix protein plays an important role in mediating BMP9-induced osteogenic differentiation of mesenchymal progenitor cells. *J. Biol. Chem.* **284**, 649–659 (2009).
 156. Sakamoto, M., Hirata, H., Ohtsuka, T., Bessho, Y. & Kageyama, R. The Basic Helix-Loop-Helix Genes *Hesr1/Hesr2* and *Hey1/Hesr2* Regulate Maintenance of Neural Precursor Cells in the Brain. *J. Biol. Chem.* **278**, 44808–44815 (2003).
 157. Belandia, B. *et al.* Hey1 , a Mediator of Notch Signaling , Is an Androgen Receptor Corepressor Hey1 , a Mediator of Notch Signaling , Is an Androgen Receptor Corepressor. **25**, 1425–1436 (2005).
 158. Fischer, A., Schumacher, N., Maier, M., Sendtner, M. & Gessler, M. The Notch target genes Hey1 and Hey2 are required for embryonic vascular development. *Genes Dev.* **18**, 901–911 (2004).
 159. Hu, X. *et al.* Integrated Regulation of Toll-like Receptor Responses by Notch and Interferon- γ Pathways. *Immunity* **29**, 691–703 (2008).
 160. Hertzog, P. A Notch in the Toll Belt. *Immunity* **29**, 663–665 (2008).
 161. Hu-Lieskovan, S., Heidel, J. D., Bartlett, D. W., Davis, M. E. & Triche, T. J. Sequence-specific knockdown of EWS-FLI1 by targeted, nonviral delivery of small interfering RNA inhibits tumor growth in a murine model of metastatic Ewing’s sarcoma. *Cancer Res.* **65**, 8984–8992 (2005).
 162. Hahm, K. B. *et al.* Repression of the gene encoding the TGF- β type II receptor is a major target of the EWS-FLI1 oncoprotein. *Nat. Genet.* **23**, 222–227 (1999).
 163. Hahn, S. & Hermeking, H. ZNF281/ZBP-99: A new player in epithelial-mesenchymal transition, stemness, and cancer. *J. Mol. Med.* **92**, 571–581 (2014).
 164. Zhou, H. *et al.* ZNF281 enhances cardiac reprogramming by modulating cardiac and inflammatory gene expression. *Genes Dev.* **31**, 1770–1783 (2017).
 165. Speiser, D. E. *et al.* A regulatory role for TRAF1 in antigen-induced apoptosis of T cells. *J. Exp. Med.* **185**, 1777–1783 (1997).

166. Xie, P. TRAF molecules in cell signaling and in human diseases. *J Mol Signal* **8**, 7 (2013).
167. Locksley, R. M., Killeen, N. & Lenardo, M. J. The TNF and TNF receptor superfamilies: Integrating mammalian biology. *Cell* **104**, 487–501 (2001).
168. Léveillé, C., Chandad, F., Al-Daccak, R. & Mourad, W. CD40 associates with the MHC class II molecules on human B cells. *Eur. J. Immunol.* **29**, 3516–26 (1999).
169. Guo, F. *et al.* TRAF1 is involved in the classical NF- κ B activation and CD30-induced alternative activity in Hodgkin's lymphoma cells. *Mol. Immunol.* **46**, 2441–2448 (2009).
170. Lee, S. Y., Kandala, G., Liou, M. L., Liou, H. C. & Choi, Y. CD30/TNF receptor-associated factor interaction: NF-kappa B activation and binding specificity. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 9699–703 (1996).
171. Duckett, C. S. *et al.* Induction of nuclear factor kappaB by the CD30 receptor is mediated by TRAF1 and TRAF2 . Induction of Nuclear Factor B by the CD30 Receptor Is Mediated by TRAF1 and TRAF2. *Mol Cell Biol* **17**, 1535–1542 (1997).
172. Song, H. Y., Rothe, M. & Goeddel, D. V. The tumor necrosis factor-inducible zinc finger protein A20 interacts with TRAF1/TRAF2 and inhibits NF-kappaB activation. *Proc. Natl. Acad. Sci.* **93**, 6721–6725 (1996).
173. Schwenzer, R. *et al.* The Human Tumor Necrosis Factor (TNF) Receptor-associated Factor 1†Gene (TRAF1) Is Up-regulated by Cytokines of the TNF Ligand Family and Modulates TNF-induced Activation of NF-kappa B and c-Jun N-terminal Kinase. *J. Biol. Chem.* **274**, 19368–19374 (1999).
174. Tsitsikov, E. N. *et al.* TRAF1 is a negative regulator of TNF signaling: Enhanced TNF signaling in TRAF1-deficient mice. *Immunity* **15**, 647–657 (2001).
175. Boyd, K. & Farnham, P. Myc versus USF: discrimination at the cad gene is determined by core promoter elements. *Mol. Cell. Biol.* **17**, 2529–37 (1997).
176. Huang, M. & Graves, L. M. De novo synthesis of pyrimidine nucleotides; emerging interfaces with signal transduction pathways. *Cell. Mol. Life Sci.* **60**, 321–336 (2003).
177. Lucas-Hourani, M. *et al.* Original Chemical Series of Pyrimidine Biosynthesis Inhibitors That Boost the Antiviral Interferon Response. **61**, 1–17 (2017).
178. Evans, M. E., Jones, D. P. & Ziegler, T. R. Glutamine inhibits cytokine-induced apoptosis in human colonic epithelial cells via the pyrimidine pathway. *Am. J. Physiol. Gastrointest. Liver Physiol.* **289**, G388-96 (2005).
179. Richmond, A. L. *et al.* The nucleotide synthesis enzyme CAD inhibits NOD2 antibacterial function in human intestinal epithelial cells. *Gastroenterology* **142**, 1483–1492.e6 (2012).
180. Radstake, T. R. D. J. *et al.* Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. *Nat. Genet.* **42**, 426–429 (2010).
181. Martins, M. *et al.* Genetic association of CD247 (CD3 ζ) with SLE in a large-scale multiethnic study. *Genes Immun.* **16**, 142–150 (2015).
182. Lin, J. X. & Leonard, W. J. The role of Stat5a and Stat5b in signaling by IL-2 family cytokines. *Oncogene* **19**, 2566–2576 (2000).
183. Hennighausen, L. & Robinson, G. W. Interpretation of cytokine signaling through the transcription factors. *Genes Dev.* 711–721 (2008). doi:10.1101/gad.1643908.GENES
184. Lai, P.-S. *et al.* A STAT inhibitor patent review: progress since 2011. *Expert Opin. Ther. Pat.* **25**, 1397–1421 (2015).
185. Howard, M. *et al.* Formation and Hydrolysis of Cyclic ADP-Ribose Catalyzed by Lymphocyte Antigen CD38. **262**, 1056–1059 (1993).
186. Laubach, J. P. & Richardson, P. G. CD38-Targeted immunochemotherapy in refractory multiple myeloma: A new horizon. *Clin. Cancer Res.* **21**, 2660–2662 (2015).
187. Partida-Sánchez, S. *et al.* Regulation of dendritic cell trafficking by the ADP-ribosyl cyclase CD38: Impact on the development of humoral immunity. *Immunity* **20**, 279–291 (2004).
188. Mehta, K. & Shahid, U. Human CD38, a Cell-Surface Protein With Multiple Functions. **10**, 1408–1417 (2017).
189. Wurdak, M., Schneider, M., Iftner, T. & Stubenrauch, F. The contribution of SP100 to

cottontail rabbit papillomavirus transcription and replication. 344–354 (2018).
doi:10.1099/jgv.0.001012

190. Sadigh, Y. & Nair, V. A Review on PML Nuclear Bodies and their Interaction with Herpesviruses. **3**, 1027 (2017).
191. Lallemand-Breitenbach, V. & de Thé, H. PML nuclear bodies : from architecture to function. *Curr. Opin. Cell Biol.* **52**, 154–161 (2018).
192. Habiger, C., Jager, G., Walter, M., Iftner, T. & Stubenrauch, F. Interferon Kappa Inhibits Human Papillomavirus 31 Transcription by Inducing Sp100 Proteins. *J. Virol.* **90**, 694–704 (2015).
193. Chelbi-Alix, M. K. & De Thé, H. Herpes virus induced proteasome-dependent degradation of the nuclear bodies-associated PML and Sp100 proteins. *Oncogene* **18**, 935–941 (1999).
194. Everett, R. D., Parada, C., Gripon, P., Sirma, H. & Orr, A. Replication of ICP0-null mutant herpes simplex virus type 1 is restricted by both PML and Sp100. *J. Virol.* **82**, 2661–2672 (2008).
195. Hasham, A. *et al.* Genetic analysis of interferon induced thyroiditis (IIT): Evidence for a key role for MHC and apoptosis related genes and pathways. *J. Autoimmun.* **44**, 61–70 (2013).
196. Ihara, H. & Gao, C. Gene Section. *Atlas Genet. Cytogenet. Oncol. Haematol.* **17**, 266–268 (2011).
197. Delgado-Vega, A. M. *et al.* Bcl-2 antagonist killer 1 (BAK1) polymorphisms influence the risk of developing autoimmune rheumatic diseases in women. *Ann. Rheum. Dis.* **69**, 462–465 (2010).
198. Takeuchi, O. *et al.* Essential role of BAX,BAK in B cell homeostasis and prevention of autoimmune disease. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 11272–11277 (2005).
199. Miner, J. J. & Diamond, M. S. MDA5 and autoimmune disease. *Nat. Genet.* **46**, 418–419 (2014).
200. Von Herrath, M. Diabetes: A virus-gene collaboration. *Nature* **459**, 518–519 (2009).
201. Gorman, J. A. *et al.* The A946T variant of the RNA sensor IFIH1 mediates an interferon program that limits viral infection but increases the risk for autoimmunity. *Nat. Immunol.* **18**, 744–752 (2017).
202. Graham, D. S. *et al.* Association of NCF2, IKZF1, IRF8, IFIH1, and TYK2 with systemic lupus erythematosus. *PLoS Genet.* **7**, (2011).
203. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. a. Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. **333**, 387–389 (2009).

Acknowledgements

This work could not have been performed without a great deal of help and support. I am very grateful to John van Dam for guiding me through the planning stages of this work, for his open door policy, and for consistently pointing out the positive aspects of my work. Without his optimistic view of my progress, I might at times have lost hope. Furthermore, John supplemented in-depth knowledge of data wrangling with in-depth knowledge of brewing one's own beer and wonderful pictures of foreign lands for those oft-needed escapes from continuous programming. It was an absolute pleasure to have John as a daily supervisor.

Can Kesmir has been supportive of me for close to five years, and has been a bastion of positivity throughout the project. I chose this project in part because I wanted to be sure that I had a great supervisor, and so it was. I have yet to meet someone more invested in seeing students succeed. Additionally, the very thorough reviews by Can of both this manuscript and the manuscript for the presentation of this work at the BioSB conference helped immensely to increase the quality of the writing (or so I hope). She also bakes a mean brownie.

I am grateful to Prof. Dr. Rob de Boer for agreeing to be my second reviewer, but also for his clear, concise, and riveting explanation of journal articles in the immunology journal club. Leading by example in the truest sense, an example I have since tried to follow (to some apparent success). Additionally, Rob bestowed upon me the honour of crafting a zucchini soup at the TBB group outing, which was, of course, a magical moment.

I wish to thank Prof. Dr. Paulien Hogeweg for her insightful suggestion to test which individual data set is most crucial to the Bayesian classifier, and for introducing me to the wonders of computational biology. I wish to thank Dr. Rutger Hermesen for an insightful discussion on Bayesian integration of continuous scores. I also wish to thank Lotte Pronk for her continuous support, lovely cooking, and willingness to listening to me droning on about such interesting topics as distribution fitting, programming errors in nested for-loops, and debugging problems. She is amazing. Lastly, I owe Jan Kees van Amerongen thanks for managing the IT infrastructure of TBB, and thereby providing me an excellent place to work. 42.

Supplementary Material

Extended data 1: enriched TFBS motifs in promoters of MHC genes

We list here the results of a literature study for known functions of the transcription factors that bind TFBS motifs found to be informative for MHC pathway gene identity in more than five of the ten cross-validation runs. The number in brackets denotes in how many cross-validations the TFBS motif was found to be informative.

IRF (10)

The mammalian Interferon Regulatory Factor (IRF) TF family consists of 9 members. They all have an N-terminal DNA-binding domain of 120 aa with 5 tryptophan-rich repeats. Their carboxy-terminal regions have greater diversity and allow for interactions with different co-factors and TFs, giving them their specificity and different functions. They are involved in regulating innate and adaptive immunity, but also in regulating metabolism⁵⁹.

Zooming in on adaptive immunity and the link with MHC, IRF4 and IRF 8 play important roles in stimulating maturation of dendritic cells (Dcs) and the pro-inflammatory maturation of macrophages. IRFs also play roles in (sustained) activation of T and B cells^{59,115}. IRFs and NF- κ B are major players in the antiviral and innate immune response. Their pathways co-regulate, with the NF- κ B pathway directly targetting IRF pathway genes and vice versa¹¹⁶. As the name suggests, a major component of their function is regulating interferon, which itself is very important in various immune processes¹¹⁷.

NFKB (10)

NF- κ B is present in the cytoplasm of every cell in its inactive state, and is conserved from *Drosophila* to humans. It can be activated by a large variety of factors, including stress, cytokines, bacteria and viruses. Five NF- κ B subunits are known: NF- κ B1, NF- κ B2, RelA, RelB and c-Rel. Upon activation of the classical pathway, two inactive subunits (NF- κ B 1 and 2) are cleaved, and the NF- κ B complex translocates to the nucleus. There, it regulates the expression of many immunity-, growth- and inflammation-related genes¹¹⁸.

The different NF- κ B subunits operate in about 15 different complexes, and there are multiple activation pathways for NF- κ B complexes¹¹⁹. One of those activating pathways is important for proper function of adaptive immunity. In mice, mRNA for specific NF- κ B subunits is higher in specific areas of the spleen, and knockout of different subunits causes different defects in proper B- and T-cell maturation processes¹²⁰. NF- κ B complexes containing RelA (See REL below) are important activators of HLA¹²¹. In sum, NF- κ B is a powerful regulator of both innate and adaptive immunity, and its role in MHC-pathway genes is therefore not unexpected¹²².

NFY (10)

NFY has three subunits (A-C) and is one of the factors in the MHC enhanceosome, a multiprotein complex that is necessary, but not sufficient, to drive MHC II gene expression¹²³. CIITA is also part of this enhanceosome¹²⁴. The enrichment of this TFBS is therefore as expected.

IRF4 (10)

IRF4 is a member of the IRF family of TFs (see IRF above). IRF4 specifically competes with activating IRFs to counter constitutive activation of a pro-inflammatory state as a result of TLR signalling¹²⁵.

PRDM1 (10)

This transcription factor, also known as B lymphocyte-induced maturation protein 1, is most prominently known as a master regulator of B cell differentiation into antibody-producing plasma cells¹²⁶. It is also a known risk factor for the autoimmune diseases lupus and rheumatoid arthritis^{127,128}. It is known specifically as a regulator of cathepsin S (important for MHC II antigen presentation and a member of our positive set), and its conditional knockout in DCs gives a lupus-like phenotype that can be reversed upon administering cathepsin S-blocking agents^{16,129}. It is, however, also involved in switching pluripotent stem cells to primordial germ cells in human development, so its profile is not solely immunity-related¹³⁰.

RAD21 (10)

RAD21 is mainly known as a member of the cohesin complex, a ring-like structure that envelops DNA and helps keep the sister chromatids together before anaphase⁶⁵. It also works in DNA repair⁶⁵. Recent evidence also implicates its function in gene regulation, especially via chromatin-mediated gene regulation, as it often co-localises with the chromatin-bounding factor CTCF¹³¹. No literature references to the MHC pathway or immunity were found. It might modulate chromatin state of MHC genes.

LMX1B (10)

LMX1B is a homeobox transcription factor that is important to dorsal patterning and organ formation (for example, the kidneys), and is also known for its role in the proper generation of (midbrain) dopaminergic neurons^{132,133}. It has been published before that this transcription factor is involved in regulating NF- κ B target genes¹³⁴, which links it to MHC pathway regulation¹³⁴.

HOXD13 (10)

HOXD13 is a homeobox gene that is important in body patterning. Mutations in this gene are associated with synpolydactyly, and it is important for early skeletal and body patterning^{58,135}. HOXD 11-13, together with HOXA 11-13 are important for distal limb structure development⁷¹. Specific immune-related functions are not found in the literature, making this an interesting candidate for further research.

REL (10)

REL is a subunit of NF-kB. RelA, RelB and c-Rel are the three transcriptionally active subunits that can make up the NF-kB heterodimer^{136,137}. RelA, specifically, is important for HLA expression¹²¹. It is therefore wholly expected to find it enriched in such immune-related genes as are present in the MHC-pathway.

TFAP2 (10)

Also called AP-2 transcription factor. Part of a group of transcription factors that form either homodimers or heterodimers and have a wide variety of functions¹³⁸. It can regulate the expression of c-MYC¹³⁸, which is also found as an important TFBS in the positive set (see below). No mention of the importance of this TFBS to immune processes has been found in the literature.

FOXD2 (10)

FOXD2 belongs to the forkhead box family of transcription factors, which has 40 members in humans. They play disparate roles, influencing development, organogenesis, metabolism, and immunity¹³⁹. The exact function of FOXD2 is unknown, though it is expressed in monocytes and T-cells, but not in B-cells, and knocking it out makes T cells less responsive to inhibition of proliferation by cAMP^{140,141}. A complete picture of its activities is missing, and the discovery of its importance here is interesting in that light.

ZSCAN16 (9)

ZSCAN16 is part of the family of zinc-finger transcription factors that contain a SCAN domain and comprises 71 members. Most of the functions of this family of TFs are undefined, with only involvement in regulating growth factors and lipid metabolism specifically known⁶¹. This find seems to be the first that suggests a possible function for ZSCAN16.

HOXB13 (9)

HOXB13 is a homeobox transcription factor, like HOXD13 above, and HOXA13 below. It regulates the response to androgens, and is required for correct formation and function of the ventral prostate^{70,72}. Overexpression makes breast cancers unresponsive to tamoxifen, which was found to happen via upregulation by HOXB13 of IL-6 (interleukin 6)¹⁴². Given that IL-6 is a cytokine that can mediate inflammation¹⁴³, this is a tantalising connection to immune function.

MYC (9)

MYC is also called proto-oncogene c-MYC. It contributes to tumorigenesis in a variety of ways, for example, by switching metabolism into a different state that facilitates proliferation¹⁴⁴. Intriguingly, a similar effect of c-MYC is observed in T-cells, where switching of metabolism by c-MYC upon activation is necessary for proper proliferation^{145,146}. MYC additionally controls CD47 and PDL-1, factors on the cell surface that, when absent, promote immune system action against cells¹⁴⁷. Importantly, c-MYC has been shown to decrease MHC I expression levels^{148,149}. The fact that this is enriched in the whole MHC pathway in 9/10 cross-validations suggests that c-MYC might be involved in regulating the entire pathway.

LMX1A (9)

LMX1A is very closely related to LMX1B above. It, too, functions in dorsal patterning, organ formation, and the formation of dopaminergic neurons^{132,133}. However, the NF- κ B-related activities are ascribed solely to LMX1B, though the sequence identity of the two is quite high, especially in the homeobox and LIM domains¹³².

PLAGL1 (9)

PLAGL1 is a tumor-suppressor gene, essential for cardiac morphogenesis, and also a maternally imprinted gene^{150,151}. Relation to immune processes could not be found in the literature.

STAT (8)

STAT is a family of transcription factors with many immune functions, exemplified by the JAK/STAT pathway which plays a major role in cytokine signalling⁶⁴. Cytomegalovirus disrupts the JAK/STAT pathway, thereby perturbing MHC II gene expression⁴⁰. Finding it in MHC genes is therefore expected.

HOXA13 (8)

HOXA13, like HOXB13 and HOXD13, is a homeobox gene that is important in body patterning. HOXA 11-13 and HOXD 11-13 are involved in distal limb structure development⁷¹. Mutations in HOXA13 cause hand-foot-genital syndrome, with incorrect development of the genitourinary tract and distal limbs⁵⁷. HOXA13 is associated with gastric cancer progression and hepatocellular carcinoma (via a long non-coding RNA transcribed from the HOXA locus: HOTTIP)¹⁵². Involvement of HOXA13 in immune genes is, as far as we are aware, undocumented.

HSFY2 (7)

It is interesting to find HSFY2 here. This is a member of the heat shock factor family of proteins, which have many (stress-related) functions, among which are immune functions^{153,154}. However, HSFY2 is special in that it is based on the Y-chromosome, and is a candidate gene for azoospermia since it is sometimes missing in infertile males⁶⁷. It is also supposedly only expressed in the testes⁶⁷. The fact that the testes are an immunoprivileged region where adaptive responses are not enforced⁶⁸ gives a tentative link to the MHC pathway. Additionally, the testes have a specific proteasome⁶⁹. It might be that HSFY2 somehow influences MHC transcription in the testes, or it might suppress expression of the normal proteasome.

HEY1 (7)

HEY1 is involved in the cardiovascular development of the embryo, in osteogenic differentiation of mesenchymal progenitors, maintenance of neural precursor cells, and functions as an androgen receptor corepressor^{155–158}. It has known functions in immunity, for it has been noted to act in a TLR-induced feedback inhibitory loop with Notch signalling in macrophages¹⁵⁹. Thus, this TF is a feedback inhibitor of TLR-induced Notch signalling¹⁶⁰. It might also regulate MHC pathway genes in this capacity as a feedback inhibitor.

EWSR1::FLI1 (6)

EWSR1 and FLI1 are transcription factors that normally lie on different chromosomes, but which are recurrently fused in cancers such as Ewing's sarcoma, forming an aberrant transcription factor that promotes tumorigenesis^{161,162}.

While it might be the case that EWSR1::FLI1 fusion transcription factors influence MHC pathway genes, this is not the case in normal regulation (i.e. when the genes for these transcription factors are not fused). It is still interesting to see that this oncogenic protein could, in theory, regulate MHC pathway genes. As cancers need to evade immune response, perhaps this is one of the ways in which Ewing's sarcoma manages to do that.

ZNF281 (6)

ZNF281 is phylogenetically conserved among mammals, and has a confirmed role in regulating gastrin (a peptide that promotes gastric acid release) and HIS3 (an enzyme in the histidine biosynthesis pathway). It is closely related to ZBP-89, which has functions in apoptosis, proliferation, differentiation and tumorigenesis. ZNF281 interacts with c-MYC, which is also found to be important (see above). It is also involved in regulating stemness¹⁶³. Furthermore, ZNF281 plays an important role in the conversion of pre-cardiac cells into cardiac cells. Interestingly, in this capacity it downregulates many inflammatory genes¹⁶⁴. Given that ZNF281 is present at detectable levels in all tissues¹⁶³ and can modulate inflammatory genes, perhaps it is also able to regulate the MHC (I) pathway genes.

Extended data 2: Literature study of the top ten negative set genes in the Bayesian classifier

TRAF1 (ENSG00000056558, score: 2.8)

TRAF1 is a member of the family of TNF receptor-associated factors, seven of which are known^{165,166}. These proteins were first discovered to help mediate responses to TNF receptor activation. TNF receptors are cytokine receptors that are characterised by their ability to bind to tumor necrosis factors (TNFs)¹⁶⁷. Most of the TNF-TNFR pairs are employed in immunity, where they are essential for coordinating proliferation and protective functions of pathogen-reactive cells, though they are also involved in processes such as organogenesis¹⁶⁷. As a particularly interesting example, TRAFs are involved in downstream activation of NF- κ B by the CD40 receptor¹⁶⁸. This receptor associates with MHC II on the surface of B-cells, and it has been found that this association increases B-cell proliferation. It might be that the TRAFs are somehow involved¹⁶⁸. In the past twenty years, it has been found that TRAFs actually have functions in many more pathways, and they are implicated in autoimmune disease, cancer, and immunodeficiencies¹⁶⁶.

Four TRAFs, including TRAF1, are involved in TLR signalling, with TRAF1 apparently functioning as a repressor of NF- κ B activation^{166,169}. Three TRAFs (though not TRAF1) are involved in the signal transduction cascade from NLRs (intracellular sensors of pathogens, similar to TLRs). TRAFs also function in the RLR pathway of innate immunity, and in cytokine receptor signalling (e.g. for the pro-inflammatory IL-17)¹⁶⁶. TRAF1 further inhibits CD-3 induced NF- κ B2 activation and proliferation in T cells¹⁶⁶. As would be expected, TRAFs are also frequently thwarted by viruses in their immune functions¹⁶⁶. Many more effects of TRAFs on B-cell and T-cell function are known, too numerous to report in detail here¹⁶⁵⁻¹⁷⁴.

CAD (ENSG00000084774, score: 2.12)

The CAD gene codes for an enzyme that catalyzes the first three rate-limiting steps in de novo pyrimidine biosynthesis¹⁷⁵. It is mainly a cytoplasmic protein, though nuclear and mitochondrial localisations are also known¹⁷⁶. It is known that Myc (c-Myc) regulates growth-induced CAD activation^{175,176}. c-Myc TFBS are also found enriched in the MHC pathway gene promoter regions (**Extended data 1**). More and more evidence points towards an integration of signalling pathways and pyrimidine biosynthesis¹⁷⁶. In fact, the most recent evidence indicates that CAD suppressors are potent anti-viral compounds¹⁷⁷. Not, as was first thought, because this depletes the resources for viral replication, but because of stimulation of innate immunity when pyrimidines are low¹⁷⁷. This process requires IRF1, a transcription factor driving the expression of antiviral genes. RIG-I is also induced upon inhibition of CAD¹⁷⁷. Besides this function, pyrimidine biosynthesis is involved with preventing cytokine-induced apoptosis in epithelial cells¹⁷⁸, a negative regulator of the antibacterial protein NOD2 in epithelial cells¹⁷⁹, and a possible drug target for Crohn's disease¹⁷⁹. Given its role in innate immune signalling, it would be interesting to investigate whether CAD is also involved in the (induction of the) MHC pathway.

CHTOP (ENSG00000160679, score: 1.73)

CHTOP or Chromatin target of PRMT1 is involved in the regulation of estrogen receptor target genes, downregulating fetal gamma globin during the switch to adult hemoglobin, the export of mRNA from the nucleus as a component of the TREX complex, and is a component of the SMN (survival of motor neuron) complex, which regulates transcription, telomerase regeneration and cellular trafficking in all animal cells⁹⁰. It is involved in transcriptional regulation of many groups of genes⁹¹. Given that it functions in mRNA export as a component of the TREX complex, it might well be that this is a gene selected on the basis of its broad function in ensuring correct production of proteins through its effect on mRNA export (i.e. it is mostly a housekeeping gene). It was identified in the MHC II SEPS, but that could well be because of just this reason (**see Results**).

CD247 (ENSG00000198821, score: 1.06)

CD247 has been established, along with MHC and STAT4, as a risk locus for the autoimmune disease systemic sclerosis¹⁸⁰. It has also been associated with lupus¹⁸¹. CD247 is a component of the T-cell receptor, which interfaces with (peptide-loaded) MHC molecules to see what is presented and starts an appropriate reaction^{1,181}. It is known that activated T cells from many species, save mice, express MHC II molecules on their cell surface⁸¹. In fact, human T cells express both MHC I and MHC II^{35,81}, and the MHC pathway could thus be linked to CD247. It is interesting to note that mice retinal ganglion cells (RGCs) also express CD247, and its knockout impairs proper formation of the RGCs^{84,85}. Knockout of MHC I has a similar effect. Thus, CD274 is not only an important part of the TCR, but has other functions linked to non-classical functions of MHC^{84,85}. CD247 might thus be linked to MHC in T cells, or via non-classical activities of immune molecules in neurons.

STAT5A (ENSG00000126561, score: 0.98)

STAT5A and STAT5B, together known as STAT, play an important role in the signal transduction of IL-2 family cytokines, which are involved in immunity and modulation of lymphocyte activities during immune responses¹⁸². Other than its immune functionality, it also functions in regulating milk proteins in the mammalian milk gland, and in generating functioning milk glands during pregnancy¹⁸³. 6 STAT family members exist, and they are activated in varying combinations¹⁸⁴. STATs act in the canonical JAK/STAT signalling pathway⁶⁴. STAT motifs are also highly enriched in the MHC pathway gene promoter regions (**Results and Extended data 1**).

CD38 (ENSG0000004468, score: 0.79)

CD38 is a protein that is found in large quantities on B- and T-cells, but is expressed in many other tissues as well⁸⁰. Its extracellular domain catalyses the conversion from NAD⁺ to cyclic ADP-ribose (cADPR)¹⁸⁵. cADPR enhances the proliferative response of B cells, and is involved in mobilising calcium. It is also involved in receptor adhesion-mediated signalling¹⁸⁶. In mice, CD38 has been found to influence DC trafficking¹⁸⁷. Binding CD38 with antibodies produces many effects in hematopoietic cells, such as growth stimulation, induction or prevention of apoptosis, stimulation of cytokine production, activation of kinases, and protein phosphorylation¹⁸⁸. CD38 is known to be highly present on myeloma cells, and it is therefore targeted by many antibody therapies¹⁸⁶. Its many functions point to a broad integration of signals, rather than sole involvement in the MHC pathway.

SP100 (ENSG00000067066, score: 0.68)

Humans encode four different isoforms of SP100¹⁸⁹. It is one of the principal constituents of ND10 components/PML nuclear bodies (PNBs), which are multi-protein spherical compartments in mammalian nuclei¹⁹⁰. Their functions are not completely known, but they are thought to function in transcriptional regulation, epigenetic regulation, DNA repair, and innate immunity, among other

processes^{190,191}. These functions are performed by recruiting transiently associated client proteins to the inside of the spherical protein complexes¹⁹¹. SP100 is an interferon beta- and interferon kappa-inducible gene, and has been implicated in control of viral replication^{190,192–194}. For example, interferon kappa induces SP100 which restricts replication of high risk human papillomavirus (hr-HPV). HPV-infected cells down-regulate interferon kappa, so viruses can influence SP100 production¹⁹². In rabbits, however, SP100 increases cottontail rabbit papillomavirus replication¹⁸⁹. It thus seems SP100 has a complex antiviral function, but sometimes also increases viral replication (perhaps through hijacking). Interestingly, three risk loci for developing interferon-induced thyroiditis (ITT), a complication that arises when virus-infected individuals are treated with interferons, contain the genes for three SP100 family proteins, HLA, and TAP1¹⁹⁵. The latter two are both MHC pathway genes. However, given that ITT is an autoimmune disorder, this is not unexpected, and SP100 might work via a different pathway all together¹⁹⁵. Nevertheless, the evidence points to functions of ITT in (auto)immunity. As the function of PNBs, and SP100 in particular, become clearer, it will be interesting to see whether MHC pathway genes are somehow involved.

BAK1 (ENSG00000030110, score: 0.68)

BAK1 is a member of the Bcl-2 family of proteins and is a pro-apoptotic protein whose gene is situated close to the MHC II region¹⁹⁶. BAK1 is present in mitochondrial membranes and undergoes conformational changes upon induction of apoptosis¹⁹⁶. A small amount of BAK1 is also present in the ER membrane¹⁹⁶. BAK1 and the related BAX are important for tissue homeostasis and regulation of development, and BAK1 is implicated in the development of autoimmune rheumatic diseases in women¹⁹⁷. Deletion of both BAK1 and BAX in mice results in the development of severe autoimmune disease, as the mice accumulate excess memory T- and B-cells^{196,198}.

PP1R18 (ENSG00000146112, score: 0.57)

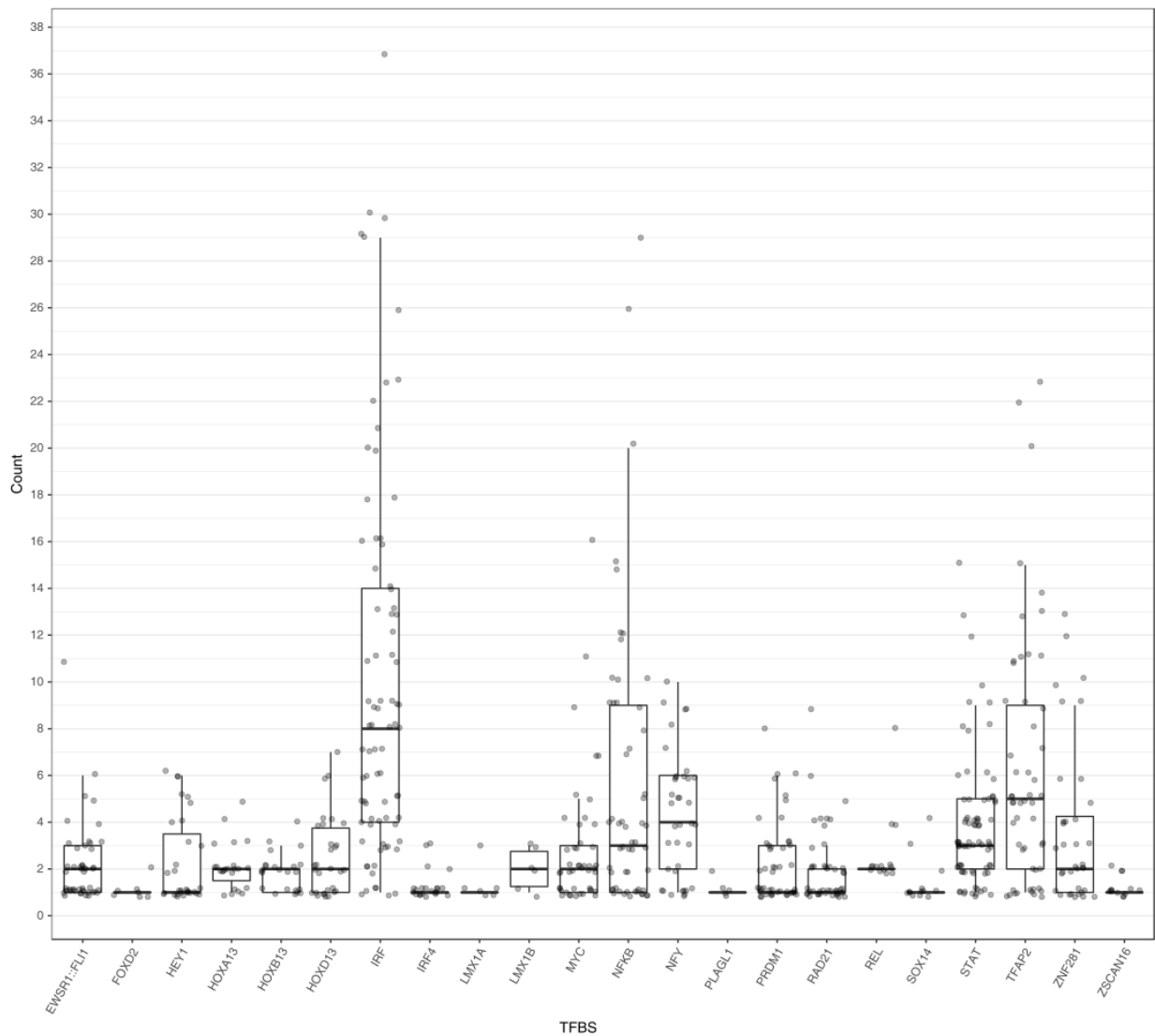
PP1R18 codes for a protein, also called phostensin, that is closely associated with actin, and is mainly expressed in lymphocytes, monocytes and macrophages (though also present in other cells)^{86,87}. It localises mainly at the cell periphery⁸⁷. It is more highly expressed in lymphoid tissues such as the spleen and thymus, and might have important immune functions⁸⁷. Furthermore, it is located between HLA-C and HLA-E on the genome, and this position is preserved in the swine homologue, which provides some further support for a possible immune function⁸⁷. Many breast cancers downregulate phostensin/PPP1R18 and it has been shown to be one of the differentially regulated genes between non-tumorigenic and metastatic breast cancer cell lines⁸⁷. PPP1R18 has also recently been shown to regulate actin organisation (actin ring formation) and bone resorption activity of osteoclasts⁸⁸. This further hints at an immune connection, as osteoclasts and the immune system have long been known to be intimately related⁸⁹. Taken together, this is certainly a high-profile candidate to look into, given its proximity to the HLA region and immune tendencies.

IFIH1 (ENSG00000115267, score: 0.55)

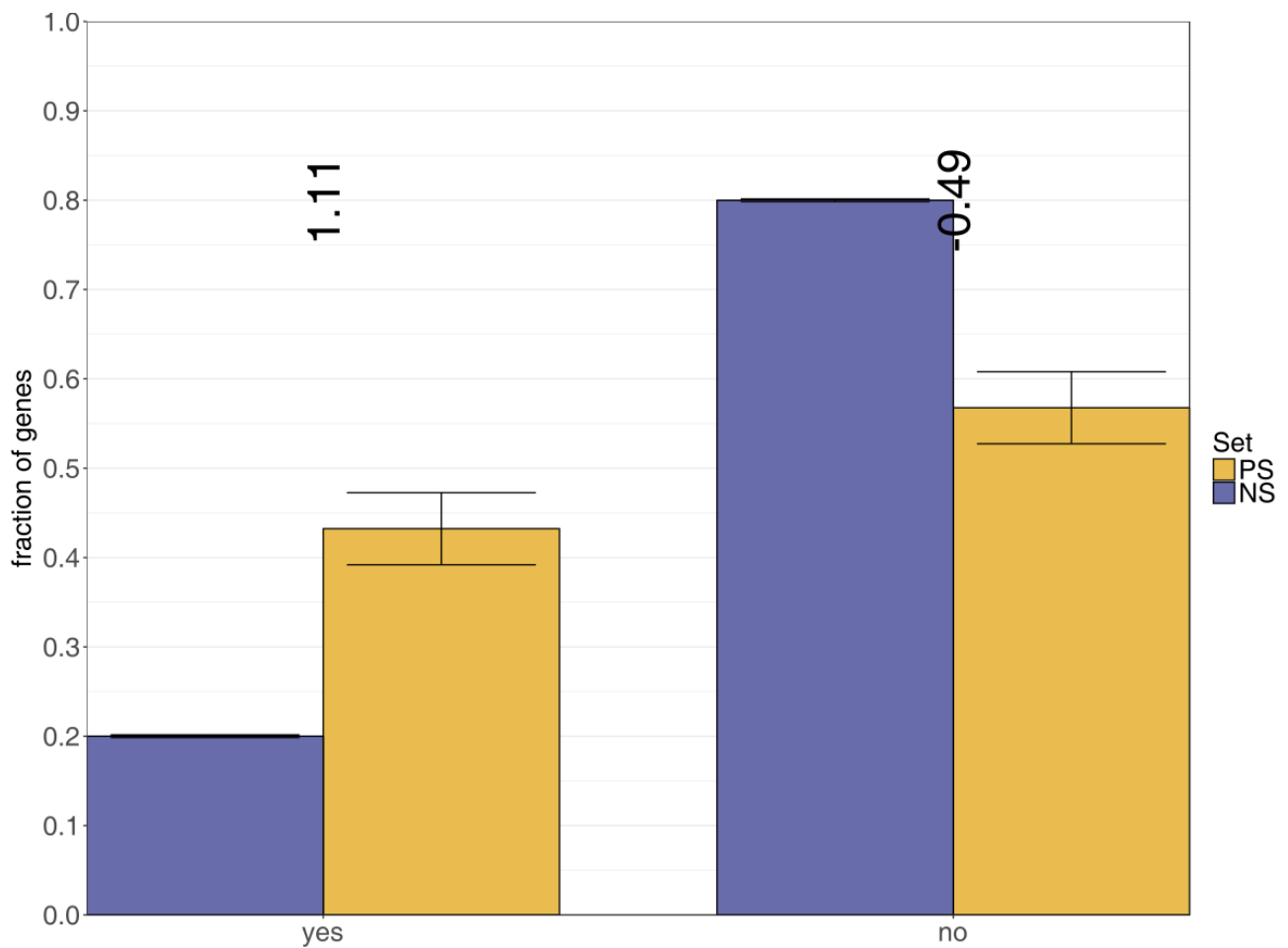
IFIH1 or MDA5 is a cytosolic double-stranded RNA receptor gene, a RIG I-like protein that is part of innate immune defence against RNA viruses^{33,95}. It senses (viral) dsRNA and activates NF-kB, which activates interferon I responses, which in turn activate inhibitory IFN-regulated genes that limit viral replication^{199,200}. Gain-of-function mutations in this gene are associated with autoimmune disease, specifically Aicardi-Goutières syndrome (AGS)^{95,199}. It is interesting to note that some of the mutations causative for AGS are mutations in the TREX complex, of which CHTOP (identified as the third highest scorer, see above) is also a part^{91,95}. An important autoantibody in idiopathic inflammatory myositis (another autoimmune disease) is targeted against IFIH1²⁰¹, and it is also involved in lupus²⁰². Rare variants of the gene protect against type 1 diabetes, which has also been linked with viral infections that precede the generation of autoantibodies^{200,203}. Interestingly, the upregulation of interferons also leads to higher expression of MHC on the cell surface, increasing visibility to lymphocytes²⁰⁰. Based on prior knowledge, it is highly doubtful that this is an MHC

pathway gene is highly doubtful. It is a sensor of viral dsRNA and therefore a component of innate immunity and the RIG I-like genes³³. Downstream effects do upregulate MHC molecules, but in this sense every gene that upregulates interferons would be an MHC pathway gene, which is a definition too broad to be useful

Supplementary figures

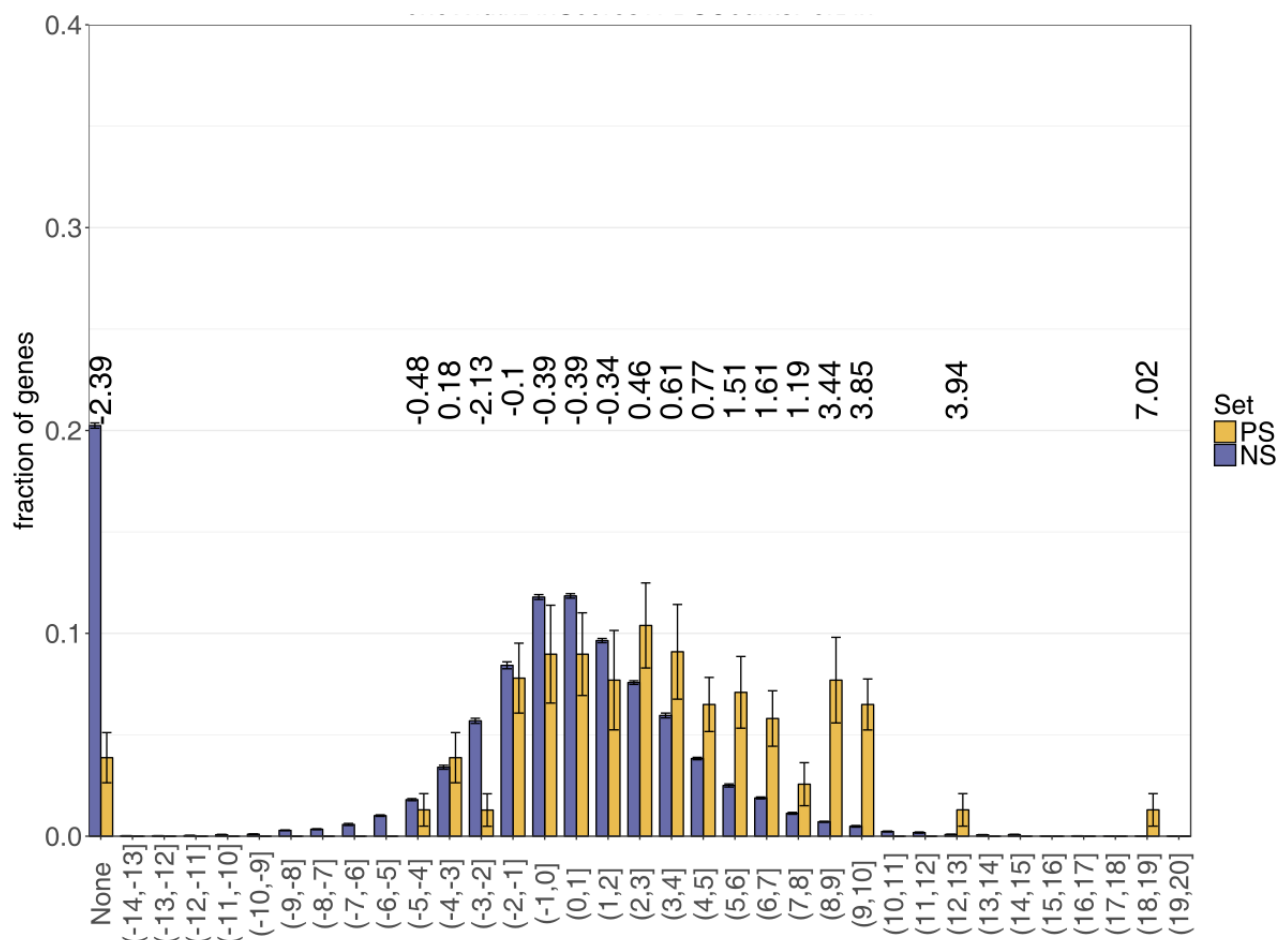


Supplementary Figure 1. Counts of informative TFBS motifs in positive set genes whose promoters have at least one of those motifs. Most TFBS motifs are present 1-4 times per promoter, but IRF, NF-kB and TFAP2 are notable exceptions to the rule.

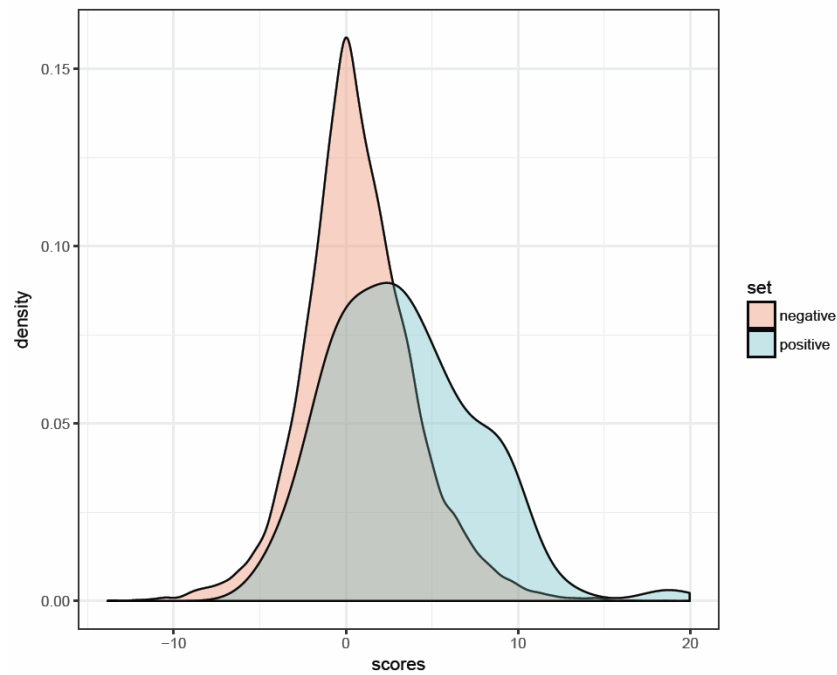


Supplementary Figure 2. Enrichment of MHC pathway genes in viral-host PPI (intersect).

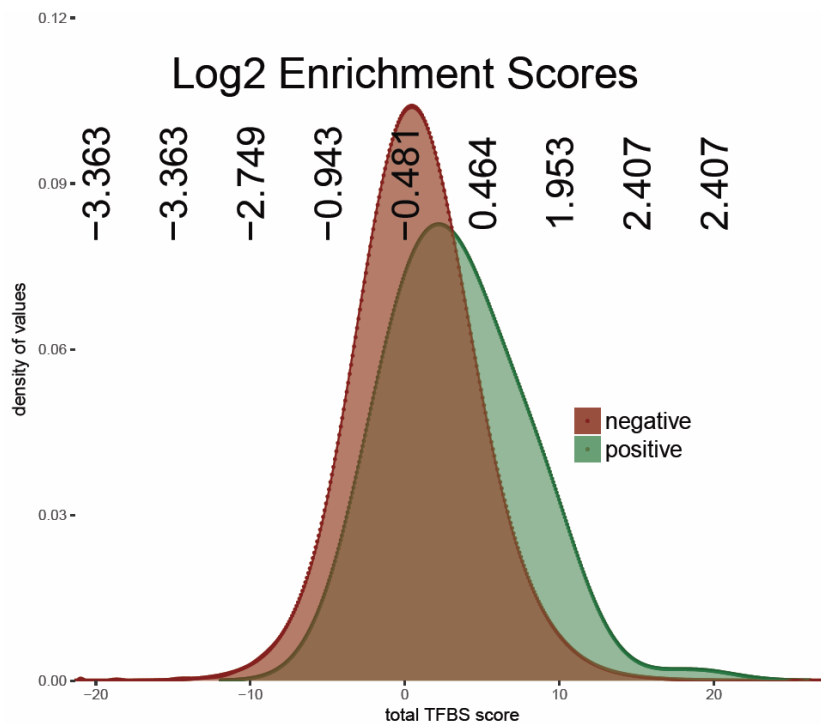
Median scores for all cross-validations of the host-viral PPI on the intersect of all three virus-host databases. Numbers above bins: median log2 likelihood ratios of all 10 cross-validations. Error bars: ± 2 sd.



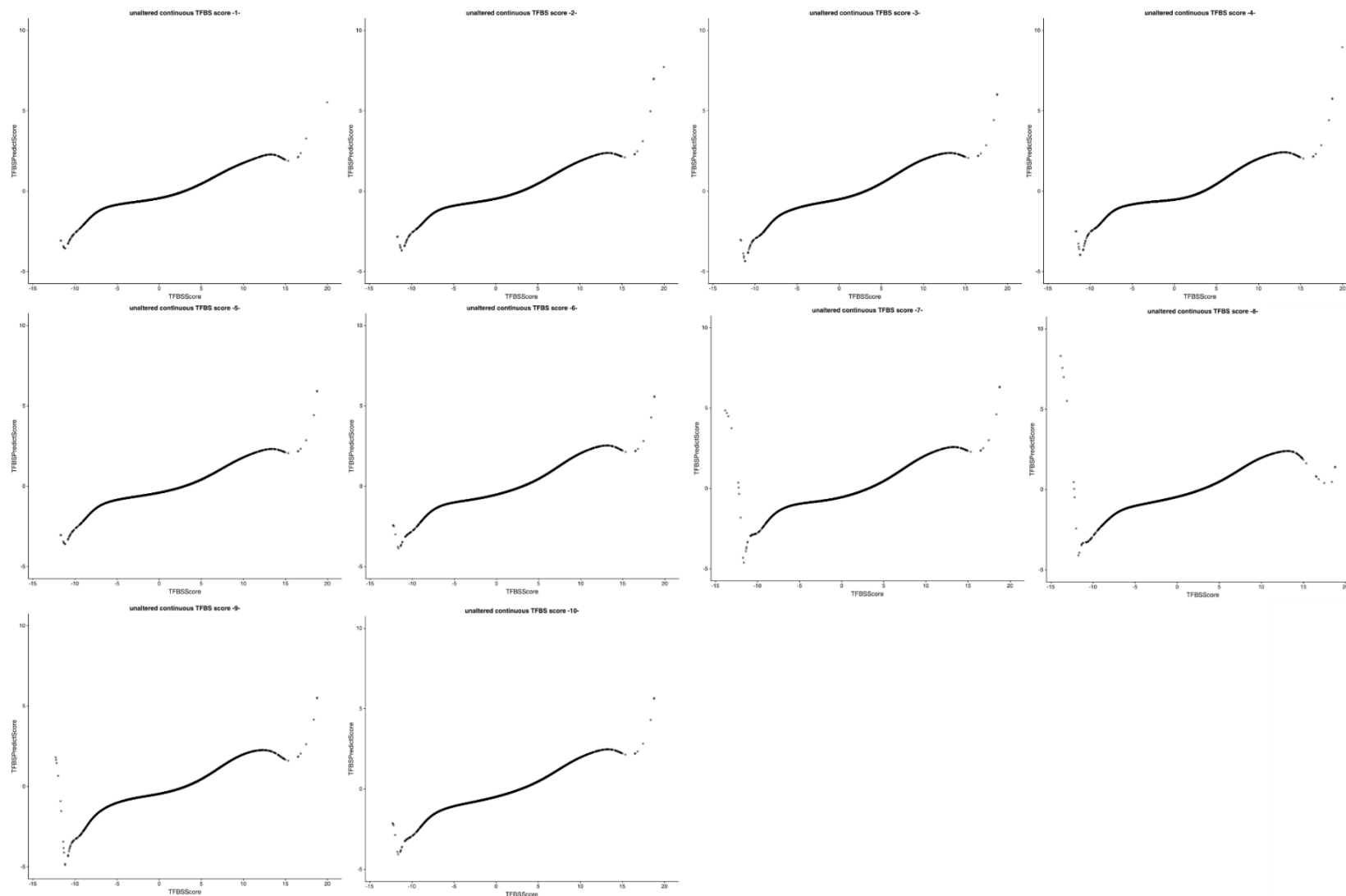
Supplementary Figure 3. MHC pathway gene TFBS motif score (bins of width one). Median MHC pathway gene TFBS motif scores for all cross-validations. Numbers above bins: median log2 likelihood ratios of all 10 cross-validations. Bins without log2 likelihood ratios had genes of only one set, so no score could be calculated. Error bars: ± 2 sd. PS: positive set. NS: negative set.



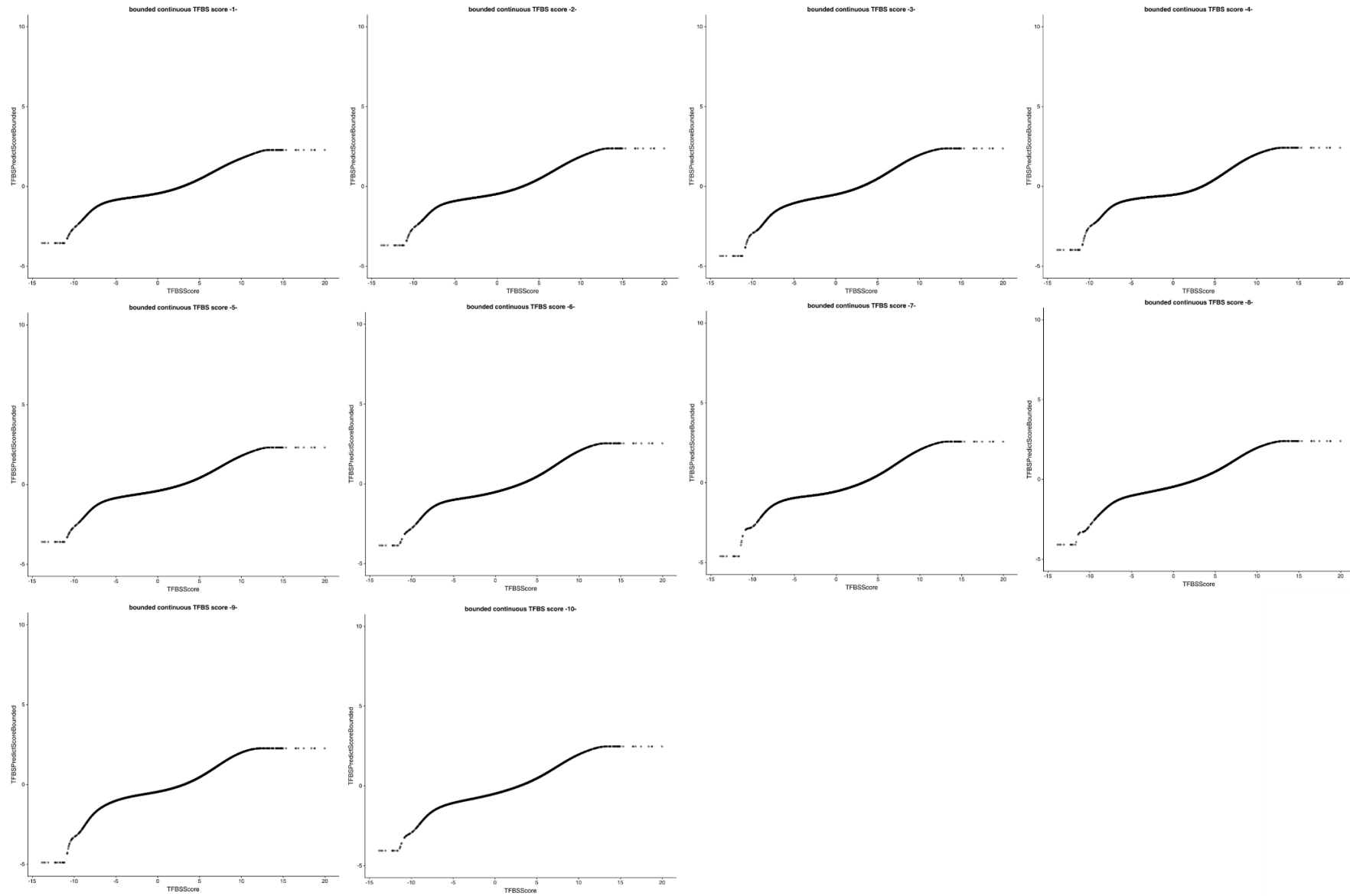
Supplementary Figure 4. Non-smoothed MHC pathway TFBS motif score distribution. There is a sharp increase and rapid decrease in the score. Negative: negative set. Positive: positive set.



Supplementary Figure 5. Smoothened, fitted MHC pathway TFBS motif score distributions of the negative and positive set. Representative log2 enrichment scores are plotted along the range of the scores. Notice the effect of the bounding at both extremes. Negative: negative set. Positive: positive set.



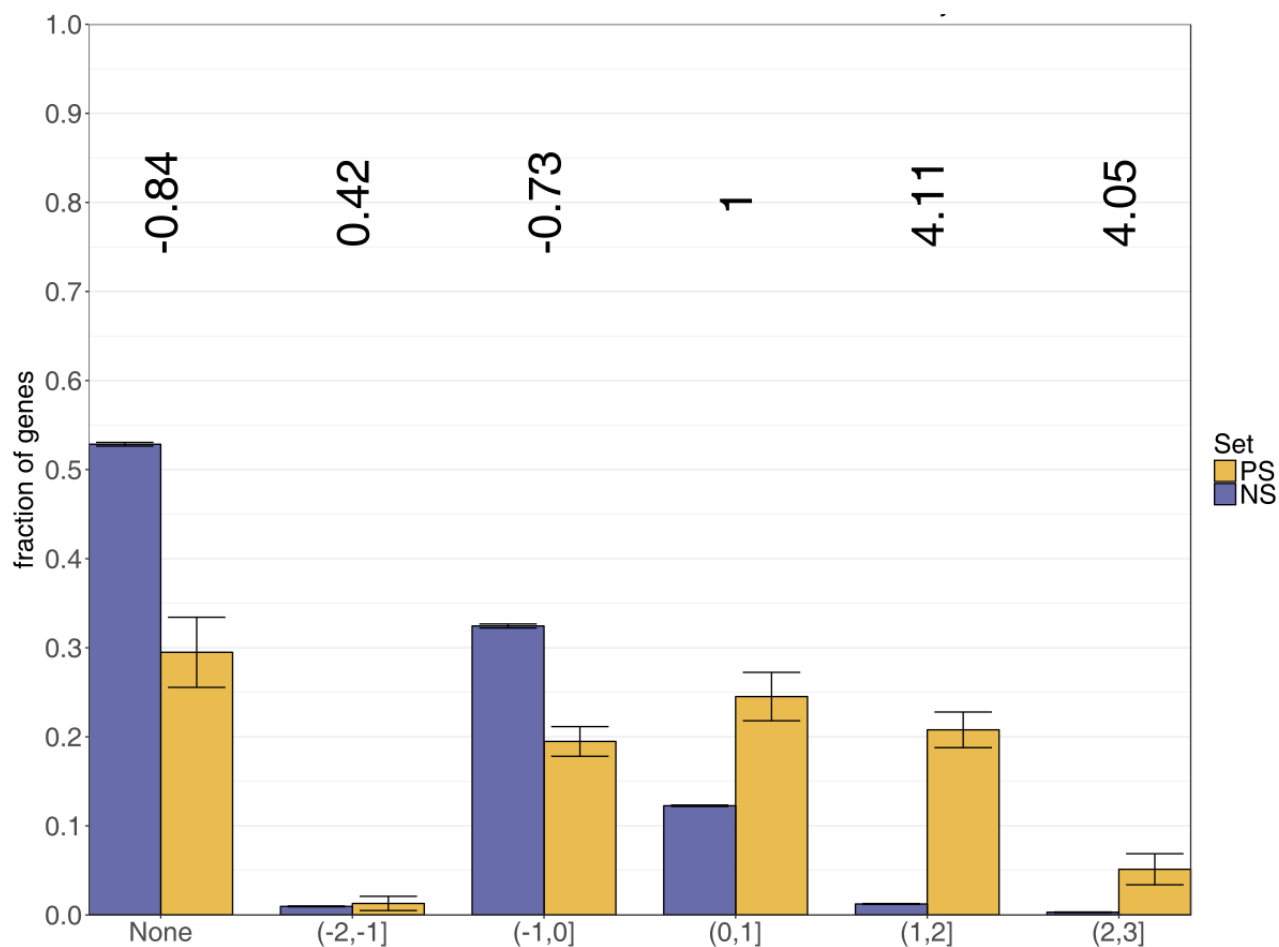
Supplementary Figure 6. Unbounded continuous MHC pathway TFBS motif scores for all ten cross-validations. Note the very high scores at the extremes.



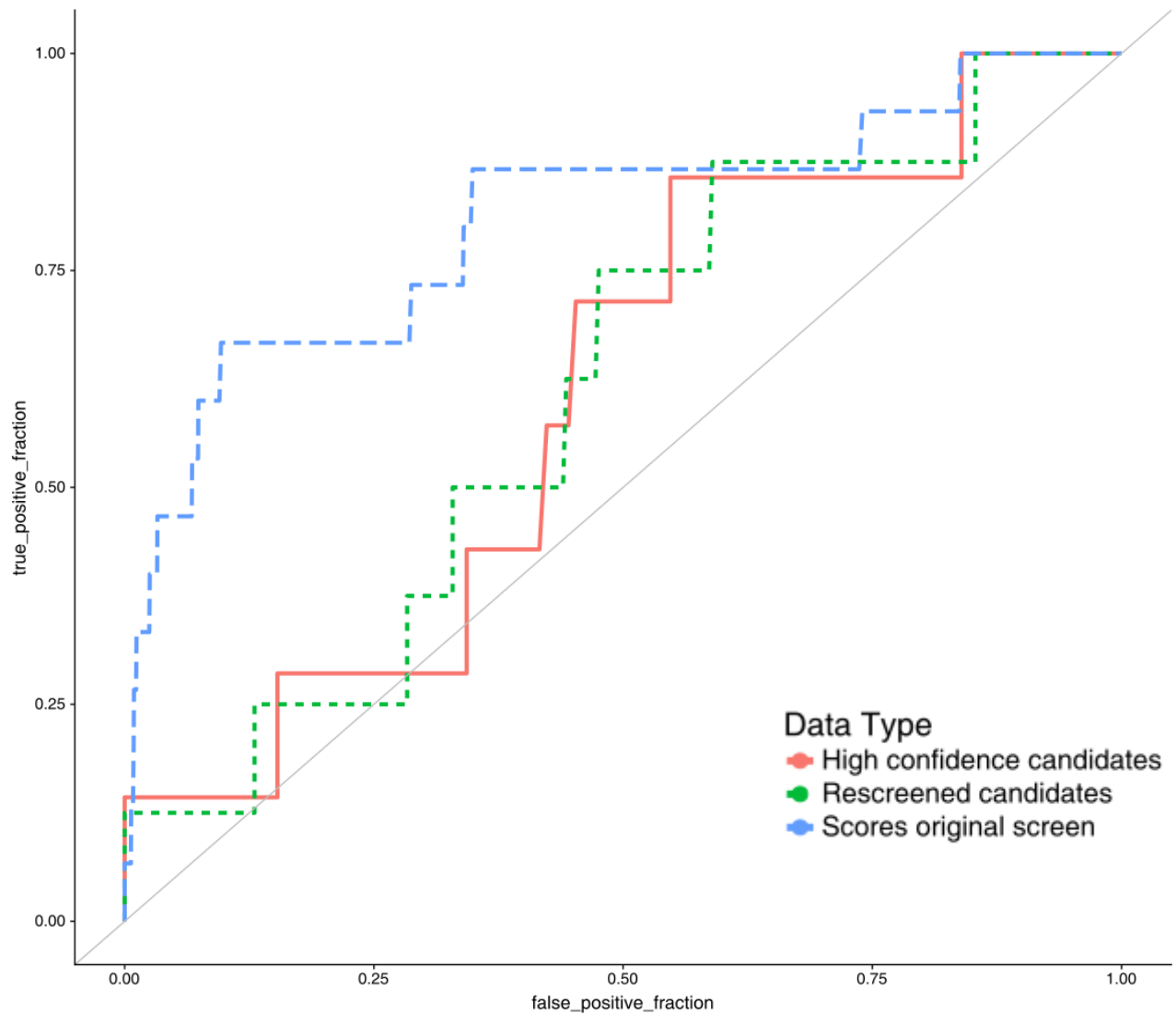
Supplementary Figure 7. Bounded continuous MHC pathway TFBS motif score for all ten cross-validations. Extreme values at the edges of the positive and negative set distributions have been bounded.

Supplementary Table 1. Statistics on the continuous TFBS score bounding procedure for each cross-validation. Shown is the minimum and maximum log2 enrichment score bounded to, and the number of values set to each of these bounds. There is quite a difference in the minimum log2 enrichment scores per cross-validation.

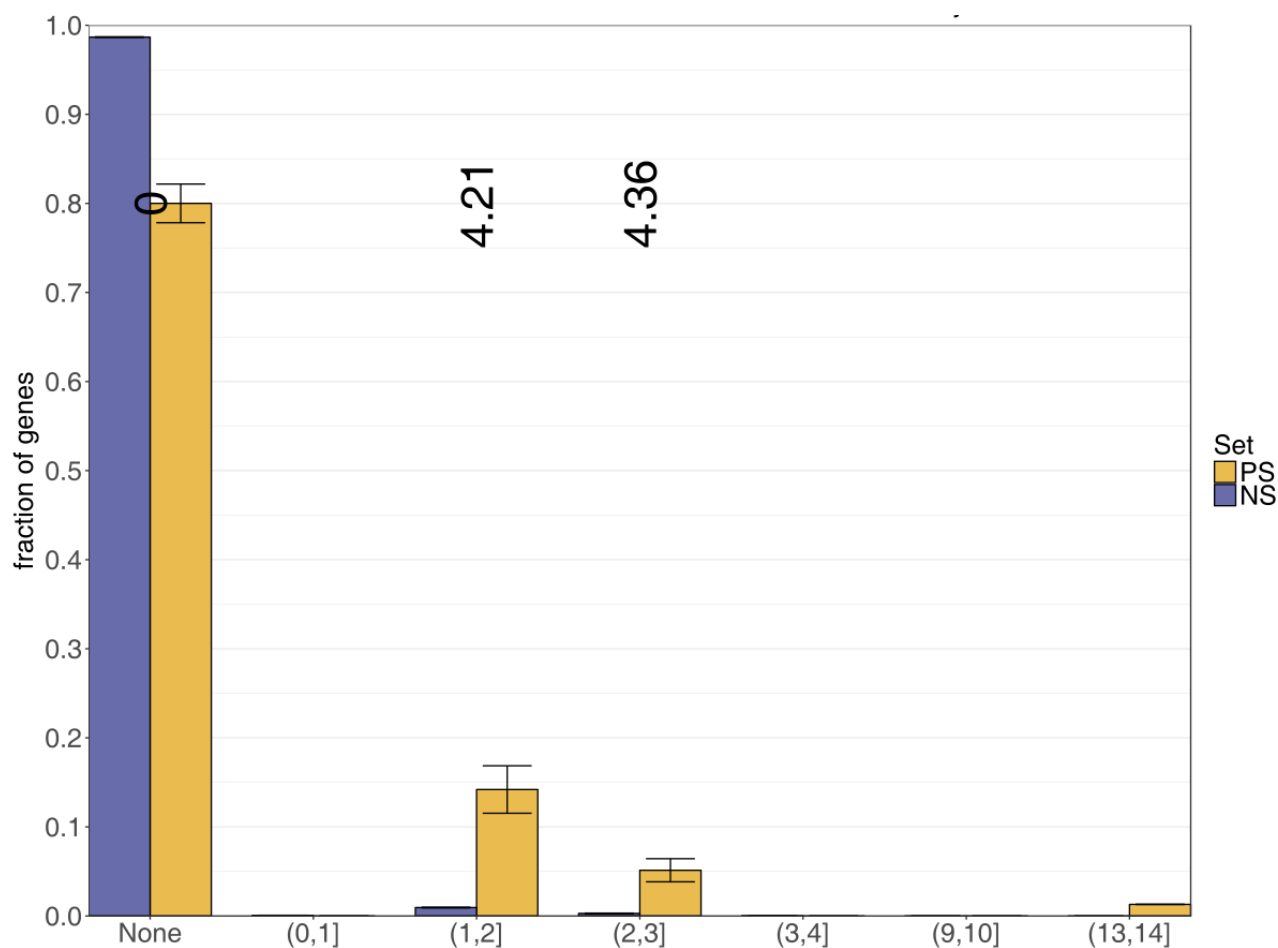
| <i>cross-validation</i> | <i>minimum bound value</i> | <i>maximum bound value</i> | <i># of values bounded to the minimum value</i> | <i># of values bounded to the maximum value</i> |
|--------------------------------|-----------------------------------|-----------------------------------|--|--|
| 1 | -3.5558 | 2.283299 | 13 | 37 |
| 2 | -3.69123 | 2.380559 | 14 | 36 |
| 3 | -4.35711 | 2.378434 | 14 | 39 |
| 4 | -3.97848 | 2.423146 | 14 | 41 |
| 5 | -3.59612 | 2.325407 | 13 | 34 |
| 6 | -3.86176 | 2.537486 | 9 | 40 |
| 7 | -4.60801 | 2.580176 | 9 | 36 |
| 8 | -4.08456 | 2.400968 | 8 | 40 |
| 9 | -4.88673 | 2.266579 | 14 | 52 |
| 10 | -4.05316 | 2.474879 | 9 | 36 |



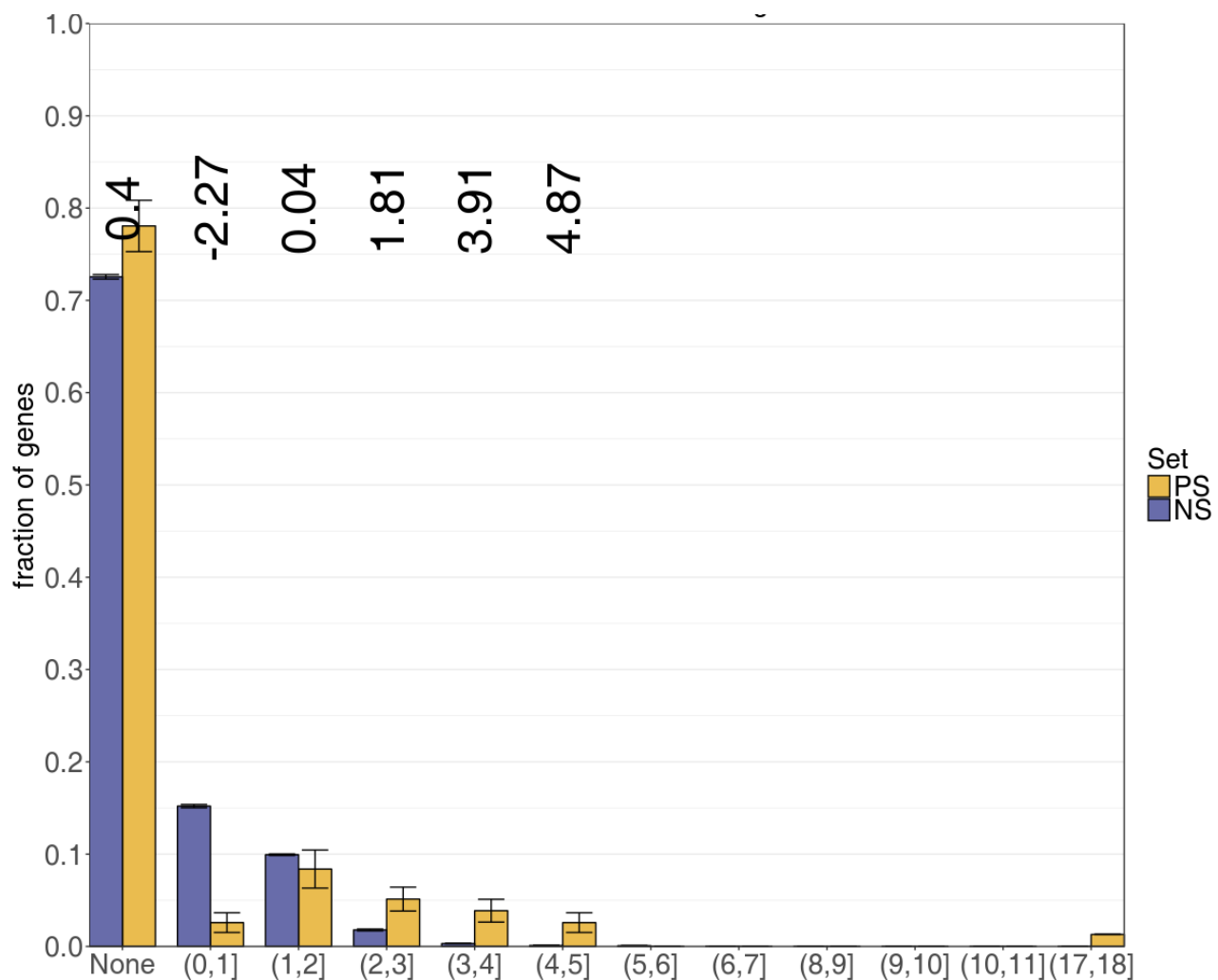
Supplementary Figure 8. Immune tissue overrepresentation score (bins of width one). Median score calculated using Mann-Whitney U rank-based differences between immune and non-immune tissue-cell type combinations for all cross-validations. Numbers above bins: median log2 likelihood ratios of all 10 cross-validations. Error bars: ± 2 sd. PS: positive set. NS: negative set.



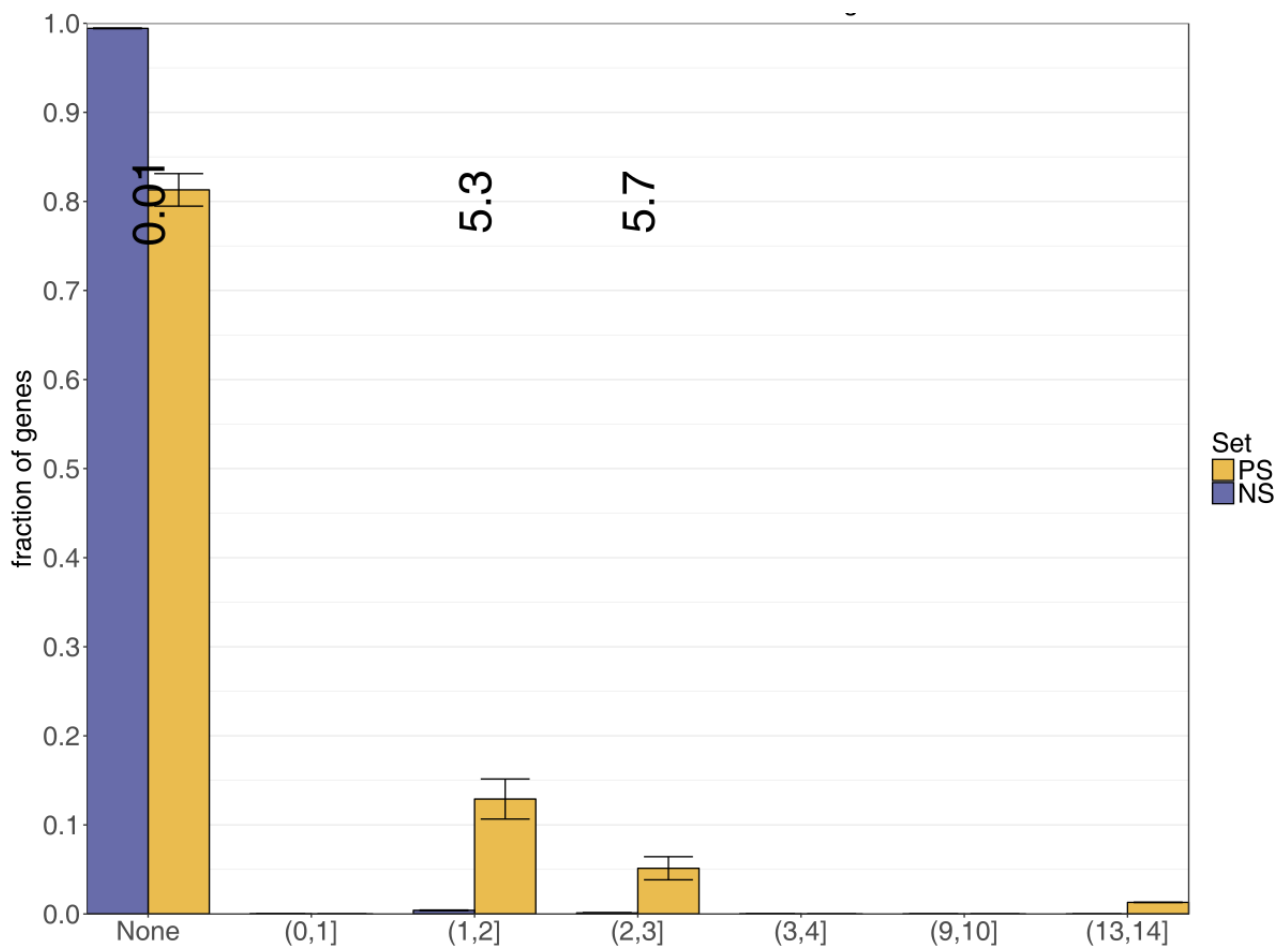
Supplementary Figure 9. ROC curve for the three different MHC II surface expression perturbation data sets. High confidence candidates: 276 genes that were rescreened, deconvoluted, and checked for expression in immune cells. Rescreened candidates: candidates that were rescreened only. Scores original screen: scores from the initial genome-wide siRNA screen. Data from²¹.



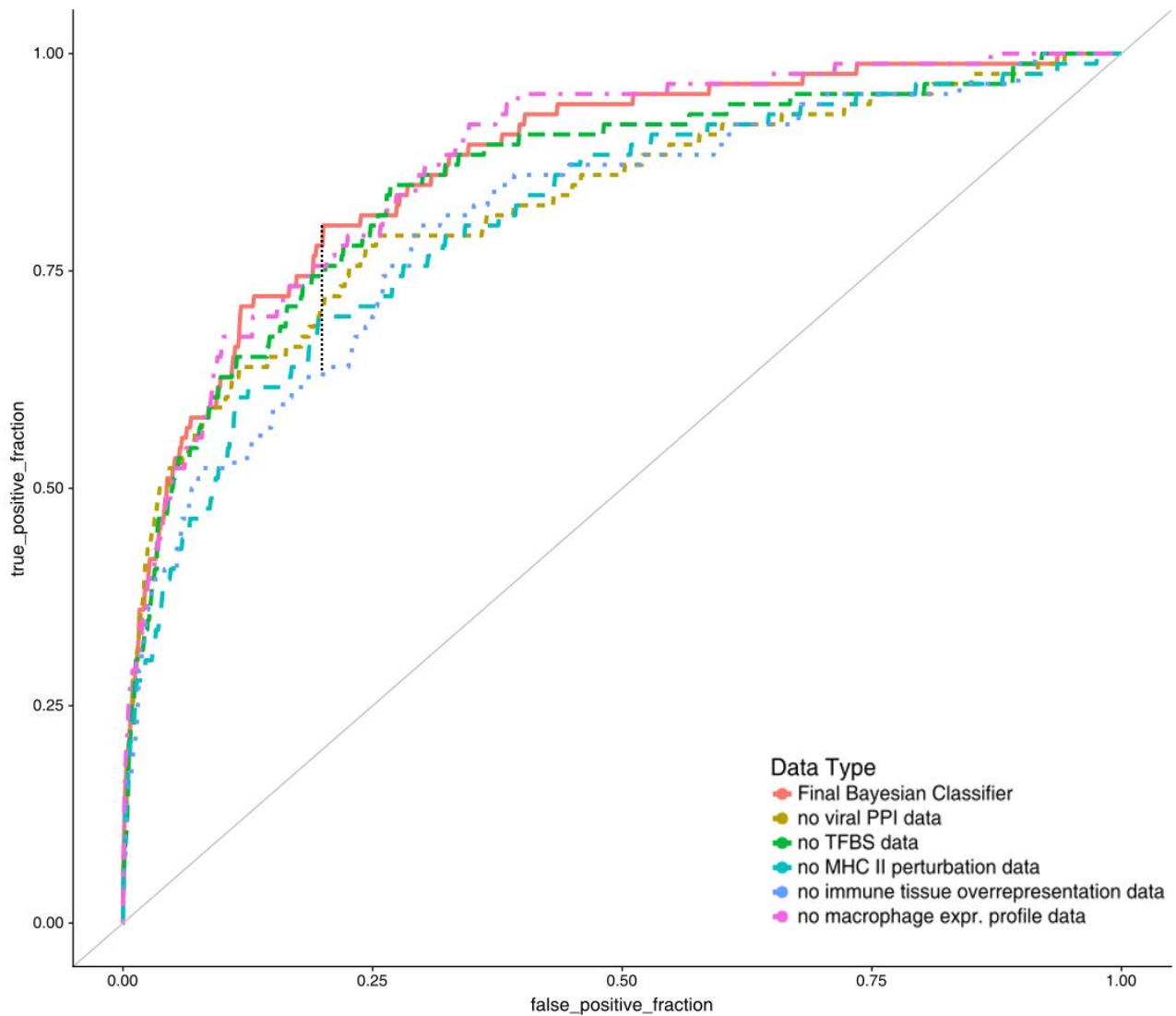
Supplementary Figure 10. Rescreen MHC II surface expression perturbation scores (bins of width one). Z scores from the Paul et al. screen, rescreened only, for all cross-validations. Numbers above bins: median log2 likelihood ratios of all 10 cross-validations. Bins without log2 likelihood ratios had genes of only one set, so no score could be calculated. Error bars: ± 2 sd. PS: positive set. NS: negative set.



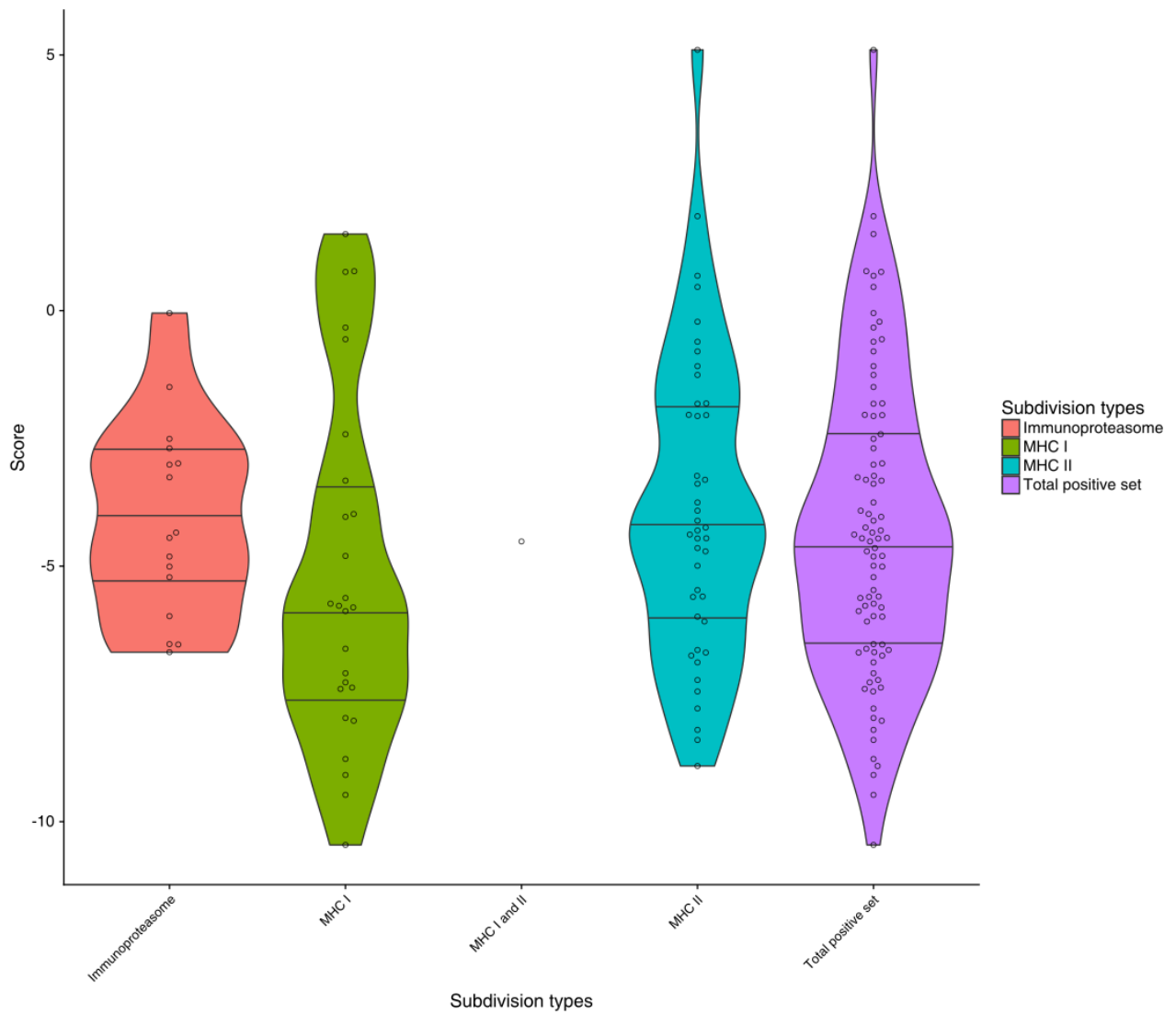
Supplementary Figure 11. Original MHC II surface expression perturbation scores (bins of width one). Original Z scores from the Neefjes screen, for all cross-validations. Numbers above bins: median log2 likelihood ratios of all 10 cross-validations. Bins without log2 likelihood ratios had genes of only one set, so no score could be calculated. Error bars: ± 2 sd. PS: positive set. NS: negative set.



Supplementary Figure 12. High confidence MHC II surface expression perturbation scores (bins of width one). Z scores from the Paul et al. screen, rescreened, corrected for off-target effects, and tested for expression in immune cells for all cross-validations. Numbers above bins: median log2 likelihood ratios of all 10 cross-validations. Bins without log2 likelihood ratios had genes of only one set, so no score could be calculated. Error bars: ± 2 sd. PS: positive set. NS: negative set.



Supplementary Figure 13. Classification potential of five integrated data sets versus all permutations of four data sets. The classification accuracy of the final classifier was compared with the classification potential of a classifier based on all permutations of 4/5 data sets. Macrophage expression profiles seem the least informative (the pink classifier line performs almost equal to the final classifier). Immune tissue overrepresentation data is the most crucial data set. When it is lacking, the true positive fraction at a specific cut-off can differ by up to 0.18 (dashed line). Viral PPI is also very informative, especially in the latter half of the curve.



Supplementary Figure 14. Violin plot of final naïve Bayesian classifier scores of different subsets of the MHC pathway genes. Generated using custom R scripts. Data on subdivisions from KEGG and literature^{5,13,21,34,35}. MHC II and immunoproteasomal components score higher than MHC I genes.