Bayesian data integration to predict novel genes involved in the MHC pathway

Dieter Stoker[1], Teunis J.P. van Dam[1] and Can Kesmir[1]
[1]Theoretical Biology and Bioinformatics, University of Utrecht, Utrecht, The Netherlands
E-mail: d.g.g.stoker@students.uu.nl, t.j.p.vandam@uu.nl, c.kesmir@uu.nl

1. Introduction

Peptide presentation on MHC molecules is essential for initiation of T cell responses, yet its winding paths are still not fully charted. Though relatively much is known about the basic biology of MHC I and II[1–3], many regulatory pathways and interactors remain unknown. Understanding these interactions is paramount for better (tumor) immunotherapies and transplant acceptance[2]. In order to discover more candidate genes that might play a role in these pathways, we integrated five complementary data sets to develop a naive Bayesian classifier. Such naive Bayesian classifiers have recently been successfully used to predict ciliary, RLR, and mitochondrial genes[4–6].

2. Approach

We first designated a set of 86 genes known to be important for MHC antigen presentation (positive set). Rather than explicitly specifying a negative set, we assumed that all other protein-coding genes in Ensembl Grpch38.p13 (22,237 genes) were non-MHC-related.

We selected five data sets with complementary information on involvement with the MHC pathway. The first dataset is a genome-wide assignment of Transcription Factor Binding (TFBS) motifs based on Encode data[7]. With this data, we calculated enrichment (or underrepresentation) of TFBS in the promoters of positive set genes, and defined an additive TFBS score that captures how much a gene's TFBS motifs correspond to those in the MHC pathway. To improve ubiquitous binning approaches[4,5], we also calculated scores based on fitting the kernel density-smoothened distributions of these scores. This method is able to separate the positive set genes and negative set genes more smoothly than the traditional binning approach, and thereby increases the differentiation among genes in the final naïve Bayesian classifier (Figure 1).

We calculated enrichment of MHC pathway genes in a similar way using four more data sets: i) a time-course of microarray gene expression in an activated M1 macrophage[8], ii) three viral protein-human protein interaction databases for measured interactions with viral proteins[9–11], iii) immunohistochemistry data from the Human Protein Atlas[12] and iv) a genome-wide siRNA screen for disturbance of MHC II peptide loading and surface expression[13].

3. Results

The data sets individually give strong log2 likelihood ratios for bins containing MHC pathway genes. The genes in our positive set are enriched among viral interactors, are among genes consistently upregulated in M1-activated macrophages, and are enriched in bins with higher expression in immune tissues. Thus, each data set captures a part of what it means to be an MHC pathway gene, and therefore can predict other candidate genes involved in this pathway. During our presentation, we will discuss the candidate genes we discovered with our naive Bayesian classifier trained using this combined data in more detail, and look towards future prediction improvement with classifier voting.

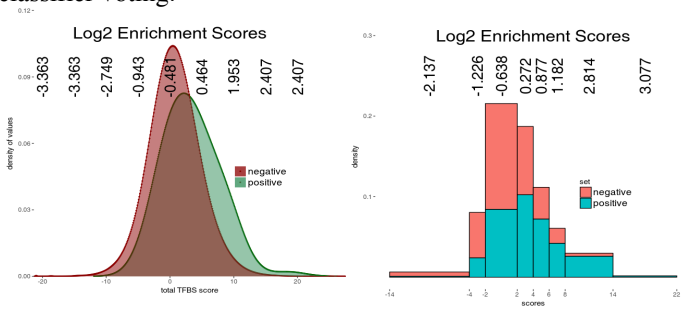

Figure 1. Side-by-side comparison of the smoothened TFBS scores and the usual binning approach.

4. References

1. 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).

2. Rock, K. L., Reits, E. & Neefjes, J. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends Immunol.* **37,** 724–737 (2016).

3. Anderson, D. A. *et al.* Revisiting the specificity of the MHC class II transactivator CIITA in classical murine dendritic cells in vivo. *Eur. J. Immunol.* (2017). doi:10.1002/eji.201747050

4. van der Lee, R. *et al.* Integrative Genomics-Based Discovery of Novel Regulators of the Innate Antiviral Response. *PLOS Comput. Biol.* **11,** e1004553 (2015).

5. van Dam, T. J. P. *et al.* CiliaCarta: An Integrated And Validated Compendium Of Ciliary Genes. *bioRxiv* 123455 (2017). doi:10.1101/123455

6. 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).

7. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42,** 2976–2987 (2014).

8. Derlindati, E. *et al.* Transcriptomic analysis of human polarized macrophages: More than one role of alternative activation? *PLoS One* **10,** 1–17 (2015).

9. 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).

10. Durmuş Tekir, S. *et al.* PHISTO: Pathogen-host interaction search tool. *Bioinformatics* **29,** 1357–1358 (2013).

11. 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).

12. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347,** 1260419–1260419 (2015).

13. Paul, P. *et al.* A genome-wide multidimensional RNAi screen reveals pathways controlling MHC class II antigen presentation. *Cell* **145,** 268–283 (2011).