Defining tissue specific gene sets using Bayesian unsupervised clustering

GEN80436

Stephen Coleman
940309160050

supervised by
Bas Zwaan
Laboratory of Genetics, Wageningen University
and
Chris Wallace
Department of Medicine, Cambridge University

Abstract

A priori defined gene sets are key to gene set enrichment analysis [18] a powerful tool in genetic analysis. Gene sets are constructed through linking genes by some common feature. This can be a function, the location of the gene product, the participation of the product in some metabolic or signalling pathway, the protein structure, the presence of transcription-factor-binding sites or other regulatory elements, the participation in multiprotein complexes, or any one of several other definitions [19][18][6][1]. However, all of these criteria are tissue agnostic. We propose to produce tissue specific gene sets by applying multiple dataset integration [7] (a Bayesian unsupervised clustering method) to the gene expression data from the CEDAR cohort [20], a dataset of 9 tissue / cell types.

A thesis presented for the degree of
Master's in Bioinformatics

Wageningen University

1 Introduction

This project, which consists of applying a Bayesian unsupervised clustering method across multiple datasets to define tissue specific gene sets, is interesting on a number of fronts. It provides a chance to learn relevant, topical biology in understanding gene sets, the role context plays in gene expression and to learn the basics of immunology. From an informatics / statistics perspective, Bayesian inference, unsupervised clustering and the use of multiple datasets are all interesting. These are relevant skills to both industry and research that I wish to develop.

Beyond developing new skills, this project also offers the opportunity to be involved in relevant research. Gene sets are commonly used in genetic analyses, thus if we can produce sets that are informed by the context of interest, it could be relevant to many researchers. Hopefully by producing more informative gene sets we can help narrow the gap between biology and disease.

2 Theory

2.1 Mixture models

Given some data $X = (x_1, \ldots, x_n)$, we assume a number of unobserved processes generate the data, and membership to a process for individual i is represented using the latent variable c_i. It is assumed that each of the K processes can be modelled by a parametric distribution, $f(\cdot)$ with associated parameters θ and that the full model density is then the weighted sum of these probability density functions where the weights are the component proportions, π_k:

$$p(x_i) = \sum_{k=1}^{K} \pi_k f(x_i|\theta_k) \tag{1}$$

2.2 Bayesian inference

We carry out Bayesian inference of this model using Markov chain Monte Carlo methods. We sample first the component parameters, θ_k, and associated weights, π_k, from the associated distributions and then sample component membership.

Basically:

1. For each of K clusters sample θ_k and π_k from the associated distributions based on current memberships, c_i; and
2. For each of n individuals sample c_i based on the new θ_k and π_k.

For the mixture model we update the parameters after we allocate each observation to a cluster. For a given cluster with associated data X and parameter θ, we sample θ using Bayes' theorem from the distribution:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_\Theta p(X|\theta')p(\theta')d\theta'} \tag{2}$$

Here Θ is the entire sample space for θ.

- We refer to $p(\theta|X)$ as the *posterior* distribution of θ as it is the distribution associated with θ *after* observing X.
- $p(\theta)$ is the *prior* distribution of θ and captures our beliefs about θ before we observe X.
- $p(X|\theta)$ is the *likelihood* of X given θ, the probability of data X being generated given our model is true. It is the criterion we focus on in our model if we would use a frequentist approach to the inference; maximising this quantity in our model generates the curve that best describes the observed data.
- $\int_\Theta p(X|\theta')p(\theta')d\theta'$ is the *normalising constant*. This quantity is also referred to as the *evidence* [10] or *marginal likelihood* and is normally represented by Z. It is referred to as the marginal likelihood as we marginalise the parameter θ by integrating over its entire sample space.

In terms of sampling, the prior is very useful as a clever choice of prior can ensure that the posterior is always solvable, that we do not encounter singularities in our distribution.

2.3 Multiple dataset integration

Consider the case when we have observed paired datasets $X_1 = (x_{1,1}, \ldots, x_{n,1}), X_2 = (x_{1,2}, \ldots, x_{n,2})$, where observations in the ith row of each dataset represent information about the same individual. We would like to cluster individuals using information common to both datasets. One could concatenate the datasets, adding additional covariates for each individual. However, if the two datasets have different clustering structures this would reduce the signal of both clusterings and probably have one dominate. If the two datasets have the same structure but different signal-to-noise ratios this would reduce the signal in the final clustering. In both these cases independent models on each dataset would be preferable. Kirk et al. [7] suggest a method to carry out clustering on both datasets where common information is used but two individual clusterings are outputted. This method is driven by the allocation prior:

$$p(c_{i1}, c_{i2}|\phi) \propto \pi_{i1}\pi_{i2}(1 + \phi\mathbb{I}(c_{i1} = c_{i2})) \tag{3}$$

Here $\phi \in \mathbb{R}_+$ controls the strength of association between datasets. (3) states that the probability of allocating individual i to component $c_{i,1}$ in dataset 1 and to component $c_{i,2}$ in dataset 2 is proportional to the proportion of these components within each dataset and up-weighted by ϕ if the individual has the same labelling in each dataset. Thus as ϕ grows the correlation between the clusterings grow and we are more likely to see the same clustering emerge from each dataset. Conversely if $\phi = 0$ we have independent mixture models.

The generalised case for L datasets, $X_1 = (x_{1,1}, \ldots, x_{n,1}), \ldots, X_L = (x_{1,l}, \ldots, x_{n,l})$ for any $L \in \mathbb{N}$ is simply a matter of combinatorics. In this case, (3) extends to:

$$p(c_{i1}, \ldots, c_{iL}|\boldsymbol{\phi}) \propto \left[\prod_{l_1=1}^{L} \pi_{c_{il_1}l_1}\right] \left[\prod_{l_2=1}^{L-1} \prod_{l_3=l_2+1}^{L} (1 + \phi_{l_2l_3}\mathbb{1}(c_{il_2} = c_{il_3}))\right] \tag{4}$$

Here $\boldsymbol{\phi}$ is the $\binom{L}{2}$-vector of all ϕ_{ij} where ϕ_{12} is the variable ϕ in (3).

Thus MDI is an extension of mixture models to multiple datasets where correlated clustering structure is used to "upweigh" similar clusters across datasets. MDI has been applied to precision medicine, specifically glioblastoma sub-typing [17], in the past showing its potential as a tool.

2.4 Tissue specificity

Cell-type specific gene pathways are pivotal in differentiating tissue function, implicated in hereditary organ failure, and mediate acquired chronic disease [5]. More and more evidence is being accrued to highlight the cell-type specific level of gene expression [3][15][11].

We also see that there are many auto-immune disease, normally associated with a specific tissue type, that have strong genetic associations. Tissue specific isoforms and expression have also been observed [22]. This shows that genes have context-specific interactions that should be considered in analysis.

2.5 Gene sets

With the onset of microarrays and RNAseq, producing gene expression data in large quantities for a wide number of genes is increasingly enabled. Unfortunately the large amount of data gifted onto the genomics community by these methods is difficult to interpret and analyse. Gene Set Enrichment Analysis (GSEA) attempts to overcome some of these issues by analysing pre-defined gene sets and changes in the expression of the full set rather than considering each constituent member on an individual basis [13]. Consider, that in analysing gene sets as a group, the degree of perturbation in the expression of the full gene set due to the disease state / alternative phenotype that is required to be considered significant is much less than that

required in analysing each of its constituent members individually [2][23]. This use of gene sets can increase the power of the analysis.

Furthermore, we know from Genome Wide Association Studies (GWAS) that many diseases are polygenic in nature [13]. Furthermore, Subramanian et al. [18] highlight the importance of gene sets, claiming that within a single metabolic pathway an increase of 20% in all the associated gene products may be more important than a 20-fold increase in a single gene.

Thus clustering genes into groups known as "gene sets" is natural and useful from both a biological and statistical perspective - it can increase the interpretability and the power of an analysis [14][21].

However, the problem of defining gene sets is non-trivial with many variations in-use. There exist many databases of gene sets [1][6][19]. The Molecular Signature Database [18] (MSigDB) is one of the most popular resources for GSEA and encompasses many different gene sets defined under various criteria or generated from separate resources. However, none of these definitions of a "set" incorporate tissue specific information. We believe that this is an oversight as there is evidence that some genes are involved in tissue specific pathways (see section 2.4). Thus we propose defining tissue specific gene sets. Previous attempts to achieve this have used the Genotype Tissue Expression (GTEx) [4] database [9], but here the profiles are for human donors post-mortem. We suspect that the data derived from these cells may not contain the same information as that collected from living, active cells. Furthermore, the GTEx data is across many different tissues (144 are used in [9]), but we focus on cell types relevant to autoimmune disease in general (i.e. blood cells) and IBD in particular (intestinal samples). This restricted focus should offer relevant gene sets.

3 Data

The data is from the Correlated Expression and Disease Association Research (CEDAR) cohort [20]. We have 9 .csv files, one for each tissue / cell type present of normalised gene expression data for 323 individuals. These are healthy individuals of European descent; the cohort consists of 182 women and 141 men with an average age of 56 years (but ranging from 19 to 86). None of the individuals are suffering from any autoimmune or inflammatory disease and were not taking corticosteroids or non-steroid anti-inflammatory drugs (with the exception of aspirin).

With regards to tissue types, samples from six circulating immune cells types:

- CD4+ T lymphocytes;
- CD8+ T lymphocytes;
- CD14+ monocytes;
- CD15+ granulocytes;

- CD19+ B lymphocytes; and
- platelets.

Data from intestinal biopsies are also present, with sample taken from three distinct locations:

- the illeum;
- the rectum; and
- the colon.

Not every individual is present in every dataset. However, as we are clustering genes this should not present a problem.

Whole genome expression data were generated using HT-12 Expression Beadchips following the instructions of the manufacturer (Illumina). 29,464 autosomal probes (corresponding to 19,731 genes) were included across the datasets, but further thinning under various criteria reduced this further in each dataset. The fluorescence intensities were $\log_2$ transformed and Robust Spline Normalized with Lumi38.

It should be noted that some datasets are less information rich than others (for instance the platelets dataset has only around 8 thousand probes present).

4 Methods

We intend to follow this pipeline to produce the clusters:

1. Transpose the data to have rows associated with gene probes and columns associated with individuals;
2. Remove NAs either imputing values using the minimum expressed value (as missingness is not random) or if above a threshold of missingness removing the column;
3. Inspect the data by PCA and remove outlier individuals for each dataset in each gene set;
4. To apply MDI we require that each dataset have the same row names in the same order, so we re-arrange our datasets to have common order of probes and include rows of 0's for probes entirely missing from a given dataset; and
5. Apply MDI [12].

To validate our clusters we intend to check if some well-annotated gene sets (such as the interleukin pathways) cluster appropriately.

5 Results

To check if the algorithm ran and as an initial exploration of the data, we implemented the steps described in 4 applying MDI to all 9 datasets. This was done twice - on the first occasion probes missing from a dataset or containing NAs were dropped (resulting in a total dataset of 4,964 probes in each dataset) and on the second occasion using an imputed value of 0 for missing probes (on this occasion we dropped probes that had NAs in some columns in all datasets reducing the dataset from 18,523 to 18,517 probes).

The algorithm was capable of running for 10,000 iterations with a thinning factor of 25 over both these set of data.

We used the modal clustering as the predicted clustering as the labels became fixed and did not vary for the majority of iterations. We did not use the clustering implied by the posterior similarity matrix (PSM) as the clusters were very defined and thus the uncertainty captured by the PSM was not necessary for the predicted clustering. Furthermore, the computational cost of calculating the PSM, particularly for the larger dataset, was quite high (the PSM is a $n \times n$ matrix).

MDI begins with 50 clusters (as an approximation of a Dirichlet process - note that we can change the number of clusters present). In the 9 datasets the number of occupied clusters stabilised around 10 (ranging from 8 - 13).

We inspected the resulting clusters using the *pheatmap* package [8] in R [16].

We can see from figure 1 that some genes cluster across all datasets (the beige band about 0.25 along the vertical axis). Between the combination of a visual inspect and the hierarchical clustering visible in the tree at the top of figure 1 combined with the information in figure 2, one can see that the platelets behave significantly differently to the other datasets - very few rows align with other datasets. We can see that we have 4 distinct groups of datasets here:

1. CD14, CD4 and CD8;
2. IL and RE;
3. CD15, CD19 and TR; and
4. PLA.

However, there is too much information in figure 1 for serious analysis and we must use subsets of the data to better understand the information contained here. Based on the clusters of datasets mentioned in , we visualise the clustering in these groups.

From figures 3 and 4, one can see that inspecting the clusters in subsets of the datasets makes it easier to see similarity in clustering.

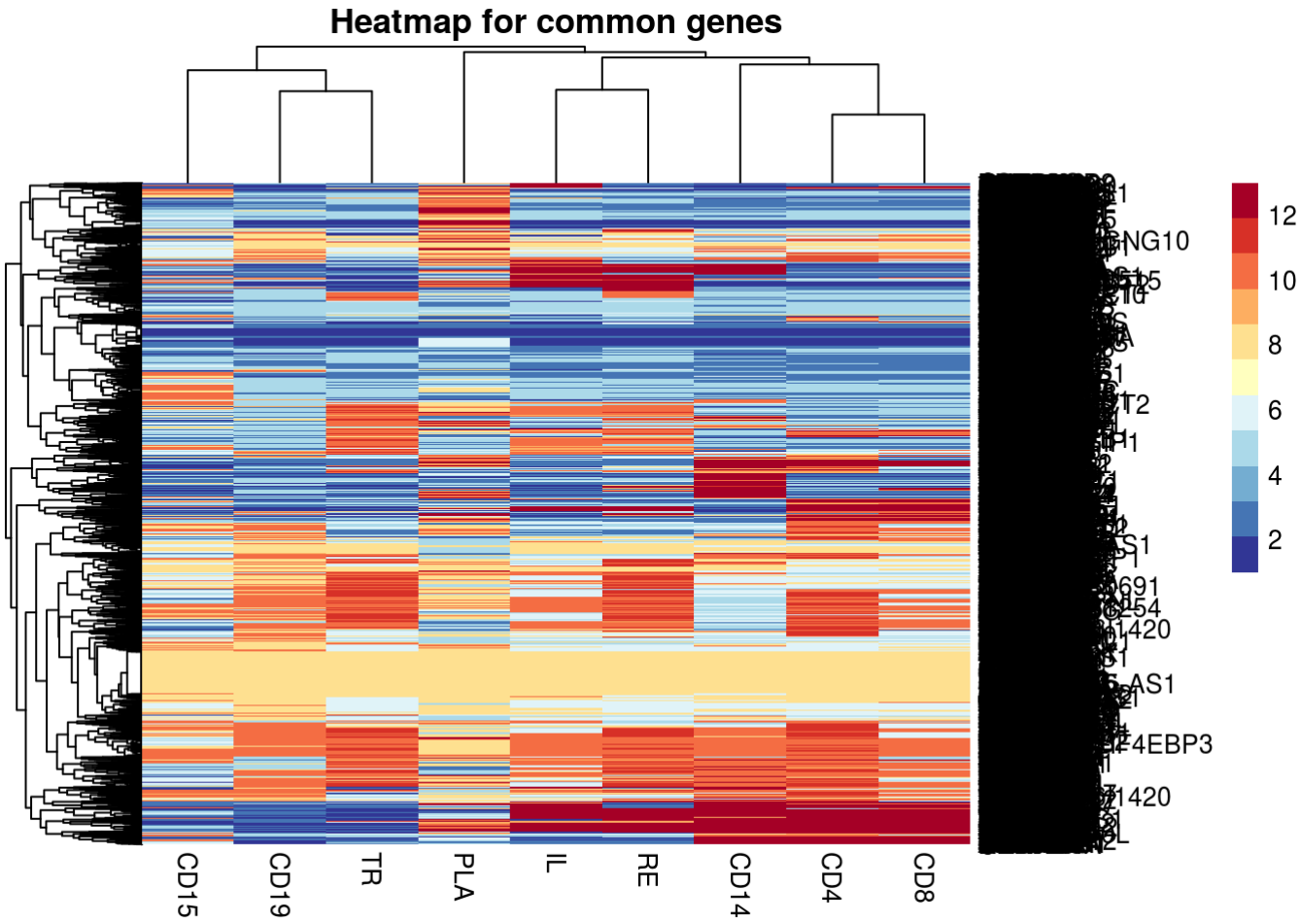

Figure 1: Predicted clusters for MDI applied to 9 datasets for common probes with datasets as columns and probes as rows.

References

[1] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1061-4036, 1546-1718. doi: 10.1038/75556. URL http://www.nature.com/articles/ng0500_25.

[2] Frank Dudbridge. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9(3):e1003348, March 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003348. URL https://dx.plos.org/10.1371/journal.pgen.1003348.

[3] Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, James Nisbett, Magdalena Sekowska, Alicja Wilk, So-Youn Shin, Daniel Glass, Mary Travers, Josine L Min, Sue Ring, Karen Ho, Gudmar Thorleifsson, Augustine Kong, Unnur Thorsteindottir, Chrysanthi Ainali, Antigone S Dimas, Neelam Hassanali, Catherine Ingle, David Knowles, Maria Krestyaninova, Christopher E Lowe, Paola Di Meglio, Stephen B Montgomery, Leopold Parts, Simon Potter, Gabriela Surdulescu, Loukia Tsaprouni, Sophia Tsoka, Veronique Bataille, Richard Durbin, Frank O Nestle, Stephen O'Rahilly, Nicole Soranzo, Cecilia M Lindgren, Krina T Zondervan, Kourosh R Ahmadi, Eric E Schadt, Kari Stefansson, George Davey Smith, Mark I McCarthy, Panos Deloukas, Emmanouil T Dermitzakis, and Tim D Spector. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089, October 2012. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.2394. URL http://www.nature.com/articles/ng.2394.

[4] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, October 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature24277. URL http://www.nature.com/articles/nature24277.

[5] Wenjun Ju, Casey S. Greene, Felix Eichinger, Viji Nair, Jeffrey B. Hodgin, Markus Bitzer, Young-suk Lee, Qian Zhu, Masami Kehata, Min Li, Song Jiang, Maria Pia Rastaldi, Clemens D. Cohen, Olga G. Troyanskaya, and Matthias Kretzler. Defining cell-type specificity at the transcriptional level in human disease. *Genome Research*, 23(11):1862–1873, November 2013. ISSN 1088-9051. doi: 10.1101/gr.155697.113. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.155697.113.

[6] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1):D590–D595, January 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gky962. URL https://academic.oup.com/nar/article/47/D1/D590/5128935.

[7] Paul Kirk, Jim E. Griffin, Richard S. Savage, Zoubin Ghahramani, and David L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, December 2012. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/bts595. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts595.

[8] Raivo Kolde. *pheatmap: Pretty Heatmaps*, 2019. URL https://CRAN.R-project.org/package=pheatmap. R package version 1.0.12.

[9] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young,

Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, June 2013. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.2653. URL http://www.nature.com/articles/ng.2653.

[10] David J C MacKay. *Information Theory, Inference, and Learning Algorithms*. September 2003.

[11] T Maniatis, S Goodbourn, and J. Fischer. Regulation of inducible and tissue-specific gene expression. *Science*, 236(4806):1237–1245, June 1987. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.3296191. URL http://www.sciencemag.org/cgi/doi/10.1126/science.3296191.

[12] Samuel A. Mason, Faiz Sayyid, Paul D.W. Kirk, Colin Starr, and David L. Wild. MDI-GPU: accelerating integrative modelling for genomic-scale data using GP-GPU computing. *Statistical Applications in Genetics and Molecular Biology*, 15(1), January 2016. ISSN 1544-6115, 2194-6302. doi: 10.1515/sagmb-2015-0055. URL https://www.degruyter.com/view/j/sagmb.2016.15.issue-1/sagmb-2015-0055/sagmb-2015-0055.xml.

[13] Michael A. Mooney and Beth Wilmot. Gene set analysis: A step-by-step guide. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(7):517–527, October 2015. ISSN 15524841. doi: 10.1002/ajmg.b.32328. URL http://doi.wiley.com/10.1002/ajmg.b.32328.

[14] A. C. Nica and E. T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362–20120362, May 2013. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2012.0362. URL http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2012.0362.

[15] Chin-Tong Ong and Victor G. Corces. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12(4):283–293, April 2011. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2957. URL http://www.nature.com/articles/nrg2957.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

[17] Richard S. Savage, Zoubin Ghahramani, Jim E. Griffin, Paul Kirk, and David L. Wild. Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. *arXiv:1304.3577 [q-bio, stat]*, April 2013. URL http://arxiv.org/abs/1304.3577. arXiv: 1304.3577.

[18] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0506580102. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0506580102.

[19] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, January 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gky1131. URL https://academic.oup.com/nar/article/47/D1/D607/5198476.

[20] The International IBD Genetics Consortium, Yukihide Momozawa, Julia Dmitrieva, Emilie Théâtre, Valérie Deffontaine, Souad Rahmouni, Benoît Charloteaux, François Crins, Elisa Docampo, Mahmoud Elansary, Ann-Stephan

Gori, Christelle Lecut, Rob Mariman, Myriam Mni, Cécile Oury, Ilya Altukhov, Dmitry Alexeev, Yuri Aulchenko, Leila Amininejad, Gerd Bouma, Frank Hoentjen, Mark Löwenberg, Bas Oldenburg, Marieke J. Pierik, Andrea E. vander Meulen-de Jong, C. Janneke van der Woude, Marijn C. Visschedijk, Mark Lathrop, Jean-Pierre Hugot, Rinse K. Weersma, Martine De Vos, Denis Franchimont, Severine Vermeire, Michiaki Kubo, Edouard Louis, and Michel Georges. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nature Communications*, 9(1):2427, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04365-8. URL http://www.nature.com/articles/s41467-018-04365-8.

[21] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, Natalia Pervjakova, Isabel Alvaes, Marie-Julie Fave, Mawusse Agbessi, Mark Christiansen, Rick Jansen, Ilkka Seppälä, Lin Tong, Alexander Teumer, Katharina Schramm, Gibran Hemani, Joost Verlouw, Hanieh Yaghootkar, Reyhan Sönmez, Andrew A. Andrew, Viktorija Kukushkina, Anette Kalnapenkis, Sina Rüeger, Eleonora Porcu, Jaanika Kronberg-Guzman, Johannes Kettunen, Joseph Powell, Bernett Lee, Futao Zhang, Wibowo Arindrarto, Frank Beutner, BIOS Consortium, Harm Brugge, i2QTL Consortium, Julia Dmitrieva, Mahmoud Elansary, Benjamin P. Fairfax, Michel Georges, Bastiaan T Heijmans, Mika Kähönen, Yungil Kim, Julian C. Knight, Peter Kovacs, Knut Krohn, Shuang Li, Markus Loeffler, Urko M Marigorta, Hailang Mei, Yukihide Momozawa, Martina Müller-Nurasyid, Matthias Nauck, Michel Nivard, Brenda Penninx, Jonathan Pritchard, Olli Raitakari, Olaf Rotzschke, Eline P. Slagboom, Coen D.A. Stehouwer, Michael Stumvoll, Patrick Sullivan, Peter A.C. 't Hoen, Joachim Thiery, Anke Tönjes, Jenny van Dongen, Maarten van Iterson, Jan Veldink, Uwe Völker, Cisca Wijmenga, Morris Swertz, Anand Andiappan, Grant W. Montgomery, Samuli Ripatti, Markus Perola, Zoltan Kutalik, Emmanouil Dermitzakis, Sven Bergmann, Timothy Frayling, Joyce van Meurs, Holger Prokisch, Habibul Ahsan, Brandon Pierce, Terho Lehtimäki, Dorret Boomsma, Bruce M. Psaty, Sina A. Gharib, Philip Awadalla, Lili Milani, Willem H. Ouwehand, Kate Downes, Oliver Stegle, Alexis Battle, Jian Yang, Peter M. Visscher, Markus Scholz, Gregory Gibson, Tõnu Esko, and Lude Franke. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. preprint, Genomics, October 2018. URL http://biorxiv.org/lookup/doi/10.1101/447367.

[22] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, November 2008. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature07509. URL http://www.nature.com/articles/nature07509.

[23] Naomi R. Wray, Sang Hong Lee, Divya Mehta, Anna A.E. Vinkhuyzen, Frank Dudbridge, and Christel M. Middeldorp. Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10):1068–1087, October 2014. ISSN 00219630. doi: 10.1111/jcpp.12295. URL http://doi.wiley.com/10.1111/jcpp.12295.

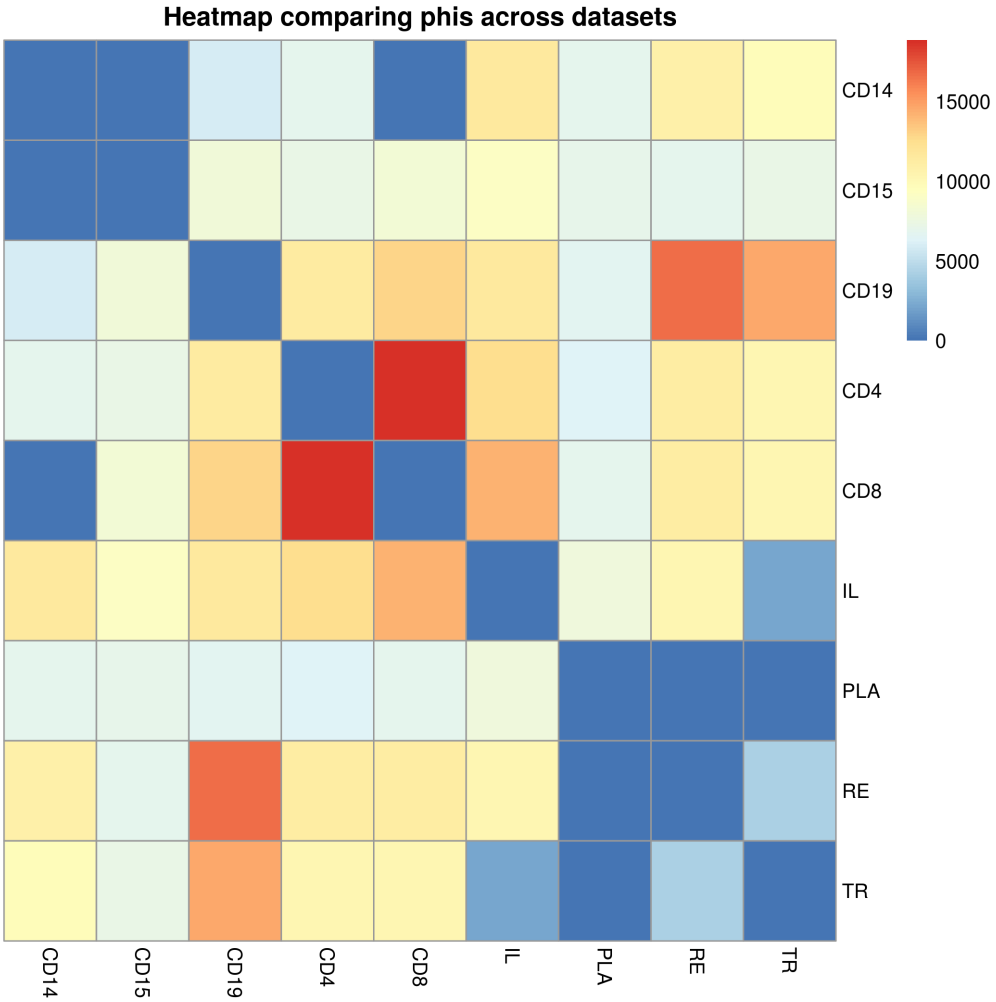

Figure 2: Mean ϕ value between datasets across iterations. ϕ can be considered a measure of similarity between datasets - the greater $\phi_{i,j}$ is, the more correlated the clustering in datasets i and j is.

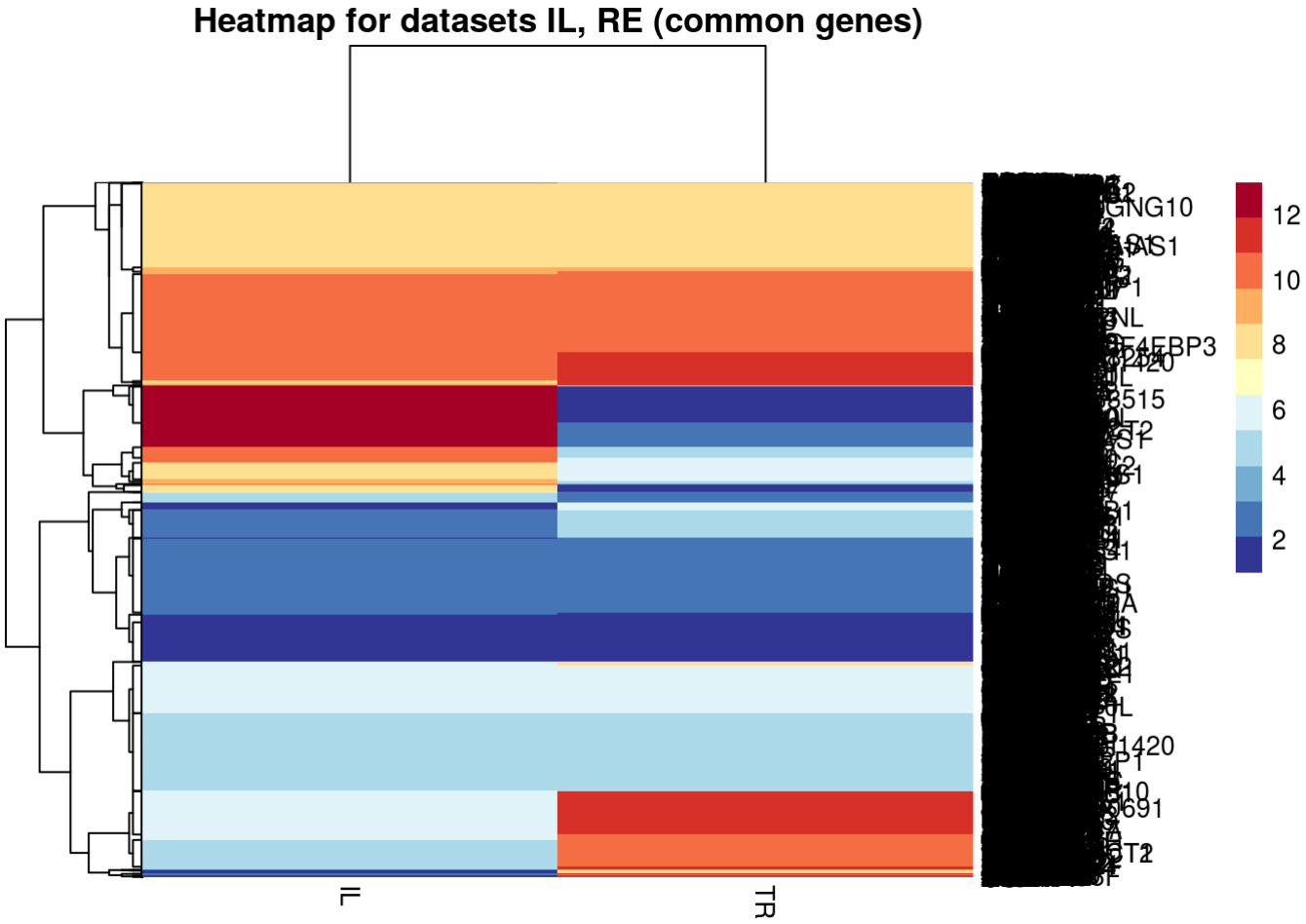

Figure 3: Predicted clusters for MDI applied to 9 datasets for common probes with datasets as columns and probes as rows.

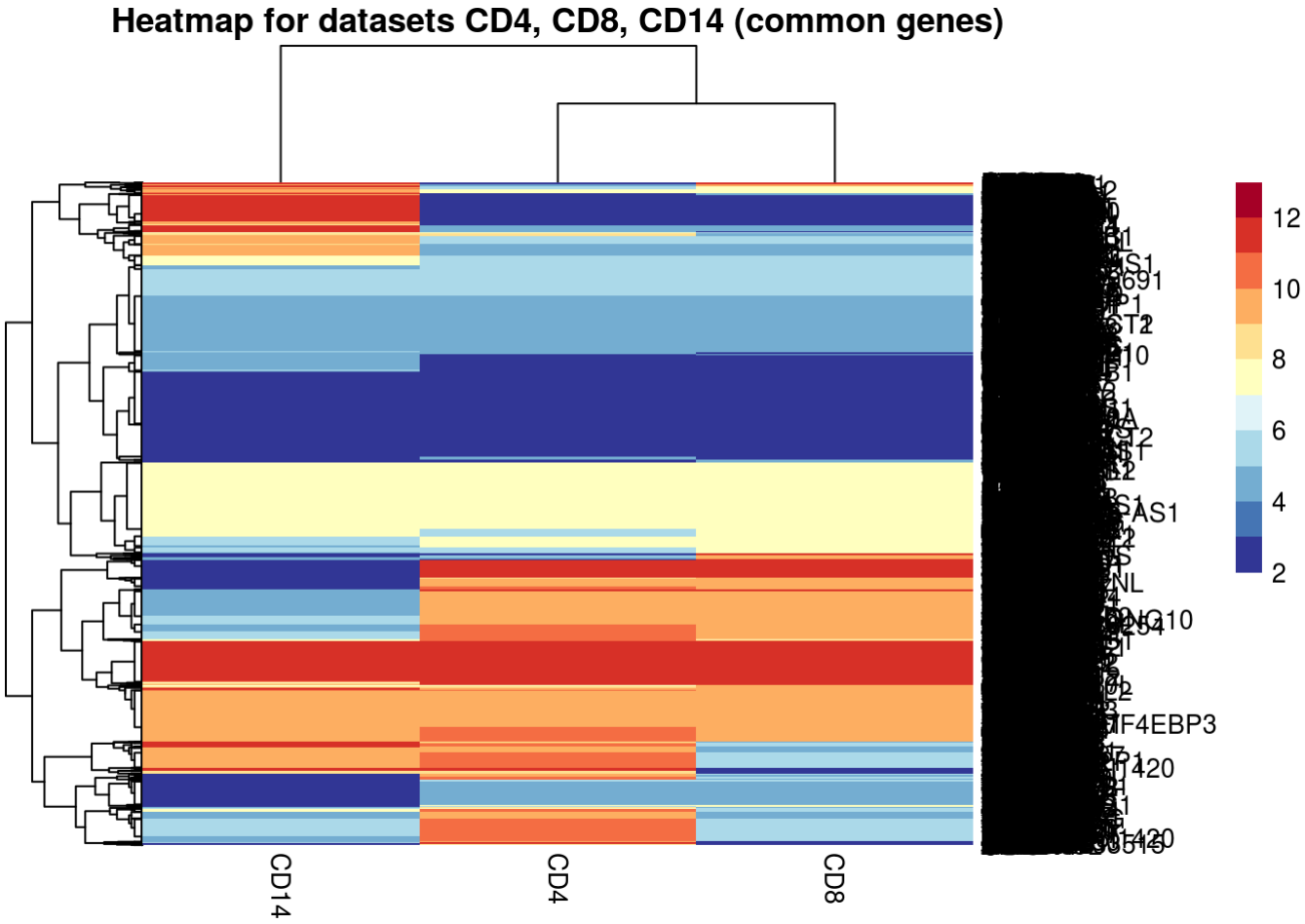

Figure 4: Predicted clusters for MDI applied to 9 datasets for common probes with datasets as columns and probes as rows.

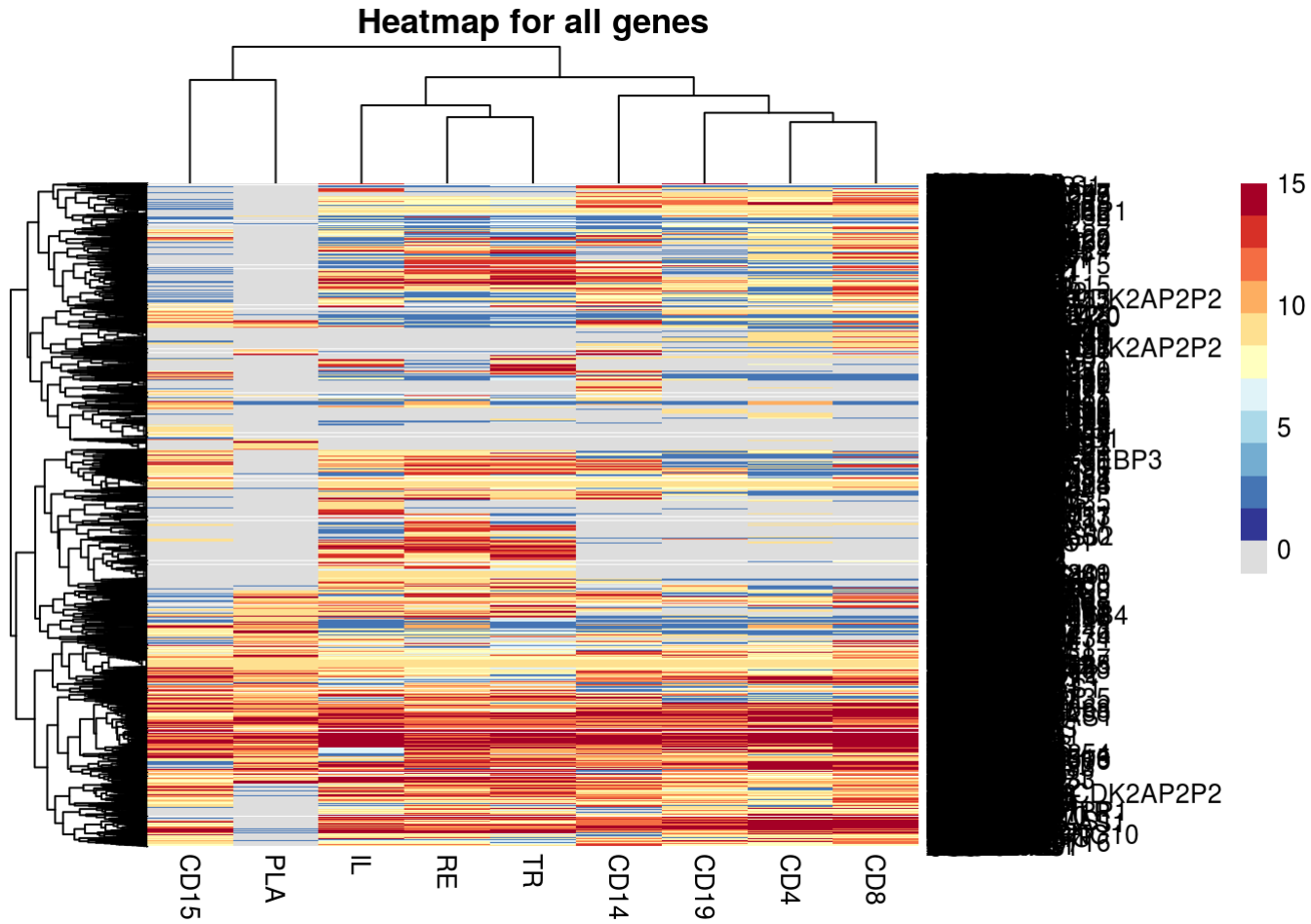

Figure 5: Predicted clusters for MDI applied to 9 datasets for all probes.