GEN80436 - Thesis proposal

Stephen Coleman

April 10, 2019

Abstract

A priori defined gene sets are key to gene set enrichment analysis (GSEA) [23]. 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, etc. [24][23][12] [1]. However, all of these criteria are tissue agnostic. Some attempts to include tissue-specific information has been proposed [7] [8], but these attempts have limitations. We propose to produce tissue specific gene sets by applying multiple dataset integration (MDI) [13] (a Bayesian unsupervised clustering method) to the CEDAR cohort [25].

1 Theory

1.1 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:

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

We carry out Bayesian inference of this model using MCMC methods. Specifically we use a Gibbs sampler. 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.

1.2 Bayesian inference

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, the distribution we sample θ from using Bayes' theorem:

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

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* [16] 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 it allows us to ensure that the posterior is always solvable, that we do not encounter singularities in our distribution.

1.3 Multiple dataset integration

If 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 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. [13] 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, $X_1 = (x_{1,1}, \dots, x_{n,1}), \dots, X_L = (x_{1,l}, \dots, x_{n,l})$ for any $L \in \mathbb{N}$ is simply a matter of combinatorics. In this case, (3) extends to:

$$p(c_{i,1}, \dots, c_{i,L} | \boldsymbol{\phi}) \propto \prod_{l_1=1}^{L} \pi_{c_{il_1} l_1} \prod_{l_2=1}^{L-1} \prod_{l_3=l_2+1}^{L} (1 + \phi_{l_2 l_3} \mathbb{1}(c_{il_2} = c_{il_3})) \tag{4}$$

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 [22], in the past showing its potential as a tool.

1.4 Expression quantitative trait loci

The last two decades have seen a huge body of research focused on genome variability due to its relevance in the risk of disease experienced by individuals. Fundamental to this study is understanding the effect different genome variants have; i.e. understanding how this change in genome translates to a different phenotype. This means we must investigate the change a variant effects within the cell. Ideally this information allows biological insight into the aetiology and nature of disease or of the phenotype. Genome-wide association studies (GWAS) [18] have shown that the majority of these variants are located within the non-coding regions of the genome [20] implying that they are involved in gene regulation. These sites that explain some of the phenotypic variance are referred to as expression quantitative trait loci (eQTL).

eQTLs have transformed the study of genetics. They provide a comprehensible, accessible and most importantly interpretable molecular link between genetic variation and phenotype. Standard eQTL analysis involves a direction association test between markers of genetic variation, typically using data collected from tens to hundreds of people.

This analysis can be proximal or distal.

- Proximal: immediately responsible for causing some observed result;
- Distal: (also *ultimate*) higher-level than proximal. The true cause for an event or result.

Consider the example of a ship sinking. This could have a *proximate* cause such as the ship being holed beneath the waterline leading to water entering the ship; this resulted in the ship becoming denser than water and it sank. However, the *distal* cause could be the ship hit a rock tearing open the hull leading to the sinking.

In terms of eQTLs, we designate proximal effects as *cis-eQTL* and distal causes as *trans-eQTL*. We normally consider an eQTL to be cis-regulating if it is within 1MB of the gene transcription start site (TSS) and trans-regulating if it is more than 5MB upstream or downstream of the TSS or if found on a different chromosome [20].

trans-eQTL are hard to find. They have weaker effects than cis-eQTL and thus require greater power in the experiment [5]. For some context, Burgess [4] claims that 449 donors provide low power in terms of finding trans-eQTL. As the power of experiments increases more trans-eQTL are observed and cis-eQTL are shown to be generally tissue agnostic [10]. Previous results suggested cis-eQTL would be have tissue-specific effects, but the increase in experimental power revealed that this is not the case [9]. The current power present in many genetic experiments is enough to observe trans-eQTLs and indicates these have tissue-specific properties [9][10]. It is possible that this result might be shown as an artefact of insufficient power, much as initial analysis suggested cis-eQTL had tissue-specific properties. However, for now we assume it is true and that trans-eQTL are more likely to display tissue-specific behaviour than cis-eQTL.

1.5 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. Clustering genes into groups known as "gene sets" increases both the interpretability and the power of an analysis [20][27]. A complete method of analysis known as Gene Set Enrichment Analysis (GSEA) is based upon this axiom [23] and well defined, meaningful gene sets are a prerequisite for GSEA [19]. 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 [6][28]. GSEA offers some insight into the biological meaning of a gene set; what gene products interact, the pathways and processes a gene is involved in. Thus gene sets offer a clearer connection between genotype and phenotype. they offer biological insight into a disease.

1.6 The importance of gene sets

If we can cluster genes together it is possible that we can find deeper biological interpretation, understanding the context of the gene products and what they interact with. This can offer some insight into the connection between the gene and the expressed phenotype. Furthermore, Nica and Dermitzakis [20] recommend investigating groups of cis-eQTL affecting a gene network that when perturbed results in a disease state. They claim this is far higher powered than the classical approach. This claim is supported by the findings of Võsa et al. [27] who found that associations between *polygenic risk scores* and gene expression (this association is referred to as "expression quantitative trait score" (eQTS) in [27]) contained the most biological information about disease in a comparison of cis-eQTL, trans-EQTL and eQTS. This finding is not unique to this paper [6][28]. More generally, gene set enrichment analysis (GSEA) [23][19] relies upon pre-defined gene sets. This method determines if gene sets have statistically significant, concordant differences between phenotypes and offers biological interpretation of the sets. Thus well-defined gene sets are required for informative, interpretable analysis of genomic information.

1.7 Existing databases

There exist many databases of gene sets [1][12][24]. The Molecular Signature Database [23] (MSigDB) is one of the most popular resources for GSEA and encompasses many different gene sets defined under different criteria or generated from different resources. However, none of these definitions of a "set" incorporate tissue specific information.

1.8 Tissue specificity

Cell-type specific gene pathways are pivotal in differentiating tissue function, implicated in hereditary organ failure, and mediate acquired chronic disease [11]. More and more evidence is being accrued to highlight the cell-type specific level of gene expression [9][21][17]. As many gene set databases are summaries of multiple experiments across many different tissues, we attempt to create tissue specific gene sets. This seems particularly pertinent in the application of immunology where many diseases are tissue-specific and have strong associations to genetic pre-disposition [26][15][2][3]. Previous attempts to achieve this have used the Genotype Tissue Expression (GTEx) database [10][14], but this is a database that has a heavy focus on brain tissues and is also exclusively from tissues of dead people. We suspect that the data derived from these cells may not contain the same information as that collected from living, active cells.

2 Data

The data is from the Correlated Expression and Disease Association Research (CEDAR) cohort [25]. This is data collected from 323 healthy individuals of European descent visting the University of Liège with samples across 9 tissue types. The cohort consists of 182 women and 141 men with an average age of 56 years (the total range is 19-86). To ensure the integrity of the data all of the individuals are not suffering from any autoimmune or inflammatory disease and were not taking corticosteroids or non-steroid anti-inflammatory drugs (with the exception of aspirin). Samples for six circulating immune cells types (CD4+ T lymphocytes, CD8+ T lymphocytes, CD14+ monocytes, CD15+ granulocytes, CD19+ B lymphocytes and platelets) and from intestinal biopsies from three distinct locations (the illeum, rectum and some other one) are present for each individual. We initially explored the gene expression data corrected for sex, age, smoking and batch effects.

3 Methods

3.1 Data preparation

The first step is to acquire and prepare the data for MDI.

1. Download the data;
2. Transpose the data, remove NAs and apply vsn;
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.

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] Thomas M. Aune, Joel S. Parker, Kevin Maas, Zheng Liu, Nancy J. Olsen, and Jason H. Moore. Co-localization of differentially expressed genes and shared susceptibility loci in human autoimmunity. *Genetic Epidemiology*, 27(2):162–172, September 2004. ISSN 0741-0395, 1098-2272. doi: 10.1002/gepi.20013. URL http://doi.wiley.com/10.1002/gepi.20013.

[3] David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33(S3):228–237, March 2003. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng1090. URL http://www.nature.com/articles/ng1090z.

[4] Darren J. Burgess. Gene expression: Principles of gene regulation across tissues. *Nature Reviews Genetics*, 18(12):701–701, November 2017. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg.2017.94. URL http://www.nature.com/doifinder/10.1038/nrg.2017.94.

[5] Anna L Dixon, Liming Liang, Miriam F Moffatt, Wei Chen, Simon Heath, Kenny C C Wong, Jenny Taylor, Edward Burnett, Ivo Gut, Martin Farrall, G Mark Lathrop, Gonçalo R Abecasis, and William O C Cookson. A genome-wide association study of global gene expression. *Nature Genetics*, 39(10):1202–1207, October 2007. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng2109. URL http://www.nature.com/articles/ng2109.

[6] 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.

[7] H Robert Frost. Computation and application of tissue-specific gene set weights. *Bioinformatics*, 34(17):2957–2964, September 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty217. URL https://academic.oup.com/bioinformatics/article/34/17/2957/4962491.

[8] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, Daniel I Chasman, Garret A FitzGerald, Kara Dolinski, Tilo Grosser, and Olga G Troyanskaya. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569–576, June 2015. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3259. URL http://www.nature.com/articles/ng.3259.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] K. Maas, S. Chan, J. Parker, A. Slater, J. Moore, N. Olsen, and T. M. Aune. Cutting Edge: Molecular Portrait of Human Autoimmune Disease. *The Journal of Immunology*, 169(1):5–9, July 2002. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.169.1.5. URL http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.169.1.5.

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

[17] 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.

[18] Teri A. Manolio. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine*, 363(2):166–176, July 2010. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMra0905980. URL http://www.nejm.org/doi/10.1056/NEJMra0905980.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Timothy J Vyse and John A Todd. Genetic Analysis of Autoimmune Disease. *Cell*, 85(3):311–318, May 1996. ISSN 00928674. doi: 10.1016/S0092-8674(00)81110-1. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867400811101.

[27] 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.

[28] 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.