Genome-Wide Association Studies

Dominik Grimm

Machine Learning and Computational Biology Lab
Department of Biosystems Science and Engineering, ETH Zürich
dominik.grimm@bsse.ethz.ch

Introduction

Recent advances in sequencing technologies have made it possible for the first time to sequence and analyse the genomes of hundreds of individuals in both a cost-effective manner and in a reasonable amount of time. One of the primary applications of this data is to better understand and investigate the genetic basis of common phenotypes or diseases. Genome-wide association studies (GWASs) are an integral tool that are often used to identify loci that are associated with a phenotype or disease of interest. In other words, an association is a correlation between the allelic and the phenotypic differences of a cohort of independent individuals. Usually, single nucleotide polymorphisms (SNPs) are used as genetic markers for GWASs. The term phenotype is quite general and can be either an apparent characteristic (e.g. the height of a plant or human eye colour) or any quantifiable characteristic, such as having a disease or being a responder/non-responder to a certain drug.

Conducting a GWAS however is often a challenging endeavour: first, different types of hidden confounding factors, such as population structure, environmental or technical influences, could lead to spurious associations [*Listgarten et al.* 2010, *Novembre et al.* 2008, *Price et al.* 2006]. Second, it has been shown in several studies that associated loci often fail to explain much of the phenotypic variability – a phenomenon referred to as the problem of *missing heritability*. Third, the apparent success to infer surnames from anonymised human genetic data [*Gymrek et al.* 2013] led to many discussions about data privacy and about how to share human genetic data with other scientists and labs – if data sharing is an option at all.

Objectives

The objectives of this task are:

- to familiarise yourself with the concept of GWASs and its current challenges [*Bush and Moore*, 2012].
- to familiarise yourself with the cloud-service `easyGWAS` [*Grimm et al.* 2012].
- to conduct GWASs and meta-analyses using `easyGWAS`.

Genome-Wide Association Studies in Model Organisms

Due to data privacy constraints and the large size of the human genome we will use for this task publically available genotype [*Horton et al.* 2012] and phenotype data [*Atwell et al.* 2010] from the plant model organism *Arabidopsis thaliana*. Performing GWASs on model organisms is similar to performing GWASs on human data, however comes with some advantages compared to human data. Wild collected *Arabidopsis thaliana* individuals are selfed (crossed with itself) in the lab for multiple generations before sequencing to ensure homogeneity of these lines. This homogeneity, as well as the small size of the genome, makes *Arabidopsis thaliana* an ideal model organism to study. The homogeneity of these isogenic lines have non-negligible advantages, for example the same line can be grown repeatedly in the lab under controlled environmental conditions. Thus, it becomes possible to phenotype and study the same line under multiple controlled environments. In addition, significantly associated hits could be experimentally validated, e.g. via gene knockouts or knockdowns. Performing GWASs on data from model organisms is highly similar to performing GWASs on human data from a methodological point of view and can thus be easily transferred.

Figure 1 Arabidopsis thaliana

easyGWAS: A Cloud Platform for Genome-Wide Association Studies

We developed easyGWAS, a cloud and web-service, to provide a platform to facilitate the use of popular genome-wide association and meta-analyses methods [*Grimm et al.* 2012]. In addition, easyGWAS provides a platform to store, share and publish data, as well as results of GWASs and meta-analyses in a straightforward manner. Different step-by-step procedures (wizards) guide the user through all necessary steps to successfully conduct new experiments and analyses. Dynamic visualisations and annotations of GWAS results are offered to obtain more detailed information about specific regions of interest.

Task 1: Getting an Overview of easyGWAS

Task 1.1

Open the following URL in your web-browser (please note that we only support the latest versions of *Firefox*, *Chrome* and *Safari*. *Internet Explorer* is only supported in the newest version Edge): https://easygwas.tuebingen.mpg.de

Have a look at the web-application and familiarise yourself with the structure of the web-application (before you login). How is the web-application structured? Which information can you find in the publically available sections?

Task 1.2

Create an `easyGWAS` account using your e-mail address (it is enough if a single person of your group creates an account).
After a successful registration you have access to the private sections. Which differences can you observe?

Task 2: Performing GWAS using `easyGWAS`

Task 2.1

Next, you will learn how to perform genome-wide association studies using `easyGWAS`. For this purpose, go the the FAQ section at the `easyGWAS` website and follow the instructions in "Tutorial 1": https://easygwas.tuebingen.mpg.de/faq/

Task 2.2

After, you learned how to conduct GWASs using the `easyGWAS` wizard you are able to conduct our own experiments. You will perform two experiments for the phenotype *"FT GH"* [*Atwell et al.* 2010] using the *AtPolyDB* dataset [*Horton et al.* 2012] in *Arabidopsis thaliana* using a 10% minor allele frequency (MAF) filter. For the first experiment use a `Linear Regression` and for the second experiment `EMMAX` [*Kang et al.* 2010]. What differences could you observe between both experiments when looking at the Manhattan- and QQ-plots?

Task 2.3

Have a look at Figure 2 from the paper from Susanna Atwell et al. (Nature 2010): "Genome-Wide Association Study of 107 Phenotypes in *Arabidopsis thaliana* Inbred Lines" [*Atwell et al.* 2010]. Try to reproduce the findings from this figure using a method of your choice that is implemented in `easyGWAS`.
Which algorithm have you used and why? Could you find a SNP in the genomic region surrounding the gene *RPM1* (Note: you can zoom into Manhattan plots; additional information about genes can be obtained in the *SNP Annotation* view and the provided links in that view)? How does your QQ-plot look like, what is the genomic control factor of your experiment and what can you conclude from it?

Task 3: Conducting Meta-Analyses with `easyGWAS`

The apparent success to infer surnames from anonymised human genetic data [*Gymrek et al.* 2013] led to many discussions about data privacy and about how to share genetic data with other scientists and labs – if sharing is an option at all. Because of concerns about privacy, researchers tend to be overzealous to protect the data and only share summer statistics of the GWASs they conducted (e.g. p-values or effect estimates). To still investigate different types of diseases large genetic consortia consisting of many different labs in various countries have

been created. These consortia often use a technique called meta-analysis to combine the results from several independent GWASs conducted at different nodes of the consortia.
In this task we will show how to conduct a meta-analysis on two different sets of individuals using a phenotype in *Arabidopsis thaliana*.

Task 3.1

First, download the following zip file and unpack it:
https://polybox.ethz.ch/public.php?service=files&t=7e6ba8443a700861bad3357c2fb51a63
The archive contains two phenotype files. Upload the two phenotypes *4W.AtPolyDB.txt* and *4W.1001.txt* using the `easyGWAS` upload manager. Upload the phenotype 4W.AtPolyDB.txt to the *AtPolyDB* dataset and the phenotype *4W.1001.txt* to the *1001 Genomes* dataset. The FAQ at the `easyGWAS` website gives a detailed description of how to upload phenotypes to `easyGWAS`: https://easygwas.tuebingen.mpg.de/faq/
In the private data section, you can have a look at the uploaded phenotypic data and its distributions. How many individuals/samples do the two phenotypes have? Are the phenotypes normally distributed?

Task 3.2

Go to the `easyGWAS` wizard and conduct two GWASs using the newly uploaded phenotypes. For both GWASs select a MAF filter of 10% and use `EMMAX` for performing the analysis. For the phenotype *"4W_AtPolyDB"* select the *AtPolyDB* dataset and for the phenotype *"4W_1001"* the *1001 Genomes* dataset. Decide yourself if you like to normalise your phenotypes and explain your decision! Have a look at the results of both experiments. What do you observe?

Task 3.3

To perform a meta-analysis, you first have to store the conducted GWASs permanently (how to store experiments permanently is explained in the FAQ section of `easyGWAS`). Next, navigate to the *Meta-Analysis* wizard and conduct a meta-analysis using the latter two experiments. Next, select *Stouffer's Z weighted* method and submit your analysis. What can you observe?

References

Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, et al. "Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines." *Nature* 465, no. 7298 (March 2010): 627–31. doi:10.1038/nature08800.

Bush, William S., and Jason H. Moore. "Chapter 11: Genome-Wide Association Studies." *PLoS Comput Biol* 8, no. 12 (December 2012): e1002822. doi:10.1371/journal.pcbi.1002822.

Grimm, Dominik, Bastian Greshake, Stefan Kleeberger, Christoph Lippert, Oliver Stegle, Bernhard Schölkopf, Detlef Weigel, and Karsten Borgwardt. "easyGWAS: An Integrated Interspecies Platform for Performing Genome-Wide Association Studies." *arXiv:1212.4788 [cs, Q-Bio, Stat]*, December 2012. http://arxiv.org/abs/1212.4788.

Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. "Identifying Personal Genomes by Surname Inference." *Science* 339, no. 6117 (January 18, 2013): 321–24. doi:10.1126/science.1229566.

Horton, Matthew W., Angela M. Hancock, Yu S. Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N. Wayan Muliyati, et al. "Genome-Wide Patterns of Genetic Variation in Worldwide Arabidopsis Thaliana Accessions from the RegMap Panel." *Nat Genet* 44, no. 2 (February 2012): 212–16. doi:10.1038/ng.1042.

Kang, Hyun Min, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. "Variance Component Model to Account for Sample Structure in Genome-Wide Association Studies." *Nat Genet* 42, no. 4 (April 2010): 348–54. doi:10.1038/ng.548.

Listgarten, Jennifer, Carl Kadie, Eric E. Schadt, and David Heckerman. "Correction for Hidden Confounders in the Genetic Analysis of Gene Expression." *Proceedings of the National Academy of Sciences* 107, no. 38 (September 21, 2010): 16465–70. doi:10.1073/pnas.1002425107.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, et al. "Genes Mirror Geography within Europe." *Nature* 456, no. 7218 (November 2008): 98–101. doi:10.1038/nature07331.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nat. Genet.* 38, no. 8 (August 2006): 904–9. doi:10.1038/ng1847.