BTRY 6830: Quantitative Genomics GWAS Project

Sruti Dammalpati (sd845)

May 12, 2020

Introduction

A Genome Wide Association Study or GWAS is an approach to identify genetic variants that are linked to underlying diseases or quantitative traits. GWAS examines SNPs (Single Nucleotide Polymorphisms) across the entire genome for many individuals to locate genes that are associated to phenotype. The application of GWAS can be quite useful in the detection, treatment and prevention of common diseases [8]. GWAS provides a useful unbaised method to infer causal relationships between genotypes and phenotypes since it does not involve any assumptions on the location of causal variants. This makes GWAS especially advantageous when no information about the position or function of causal genes is known [4]. However, like with any other association study, GWAS comes with its own limitations. One major concern is the result of several biological false positives, for which a high level of significance must be adopted. These concerns can also be addressed by using larger sample sizes or reducing the number of tests performed.

This project examines a subset of a publicly available data set from the Genetic European Variation in Health and Disease (gEUVADIS) [5]. The dataset provided for this analysis includes 334 samples with 50,000 SNP genotypes from 4 different European populations: CEU (Utah residents with European ancestry), FIN (Finns), GBR (British) and, TSI (Toscani). For each individual, we consider gene expression levels of five genes: MARCH7, FAHD1, PEX6, ERAP2, and GFM1. The expression levels for each sample are quantified through RNA sequencing of RNA levels generated from lympoblastoid cell lines (LCL). Each gene expression measurement is a phenotype for our GWAS analysis. Additionally, for our analysis, we incorporate information provided on the population and gender of each of these individuals, and information regarding the position of each gene and SNP in the genome. Using this dataset, a GWAS was attempted to locate positions of causal polymorphisms for the 5 expressed genes through rigorous statistical analysis.

Expression Quantitative Trait Locus (eQTL) Analysis

Expression Quantitative Trait Locus or eQTL Analysis is a GWAS when the phenotype measurements are gene expression levels, as in the case here. This section discusses exploratory and statistical data analysis methods used for the eQTL study.

Preparing Data for Analysis

The phenotype data was continuous with no missing information. The data consists of mRNA expression levels for 5 different genes for 334 individuals. Information regarding each genes involved in this study was gathered from UniProt [3]. MARCH7 or E3 ubiquitin-protein ligase is a protein involved in the pathway protein ubiquitination, which is part of Protein modification. FADH1 is a probable mitochondrial acylpyruvase which is able to hydrolyze acetylpyruvate and fumarylpyruvate in vitro. PEX6 is involved in peroxisome biosynthesis and forms heteromeric AAA ATPase complexes required for the import of proteins into peroxisomes. ERAP2 is an aminopeptidase that plays a central role in peptide trimming, a step required for the generation of most HLA class I-binding peptides. GFM1 is a mitochondrial GTPase that catalyzes the GTP-dependent ribosomal translocation step during translation elongation.

The phenotype data provided was analyzed to check whether it followed a normal distribution or if there were outliers present. This is important because a linear regression model is used here, and one of the assumptions prior to using this model is that the residuals follow a normal distribution. The presence of outliers can also skew the distribution. For this reason, histrograms for each phenotype are plotted and shown in Figure 1. A quick look indicates that the data follows a normal distribution and that there we no outliers present for any of the five genes.

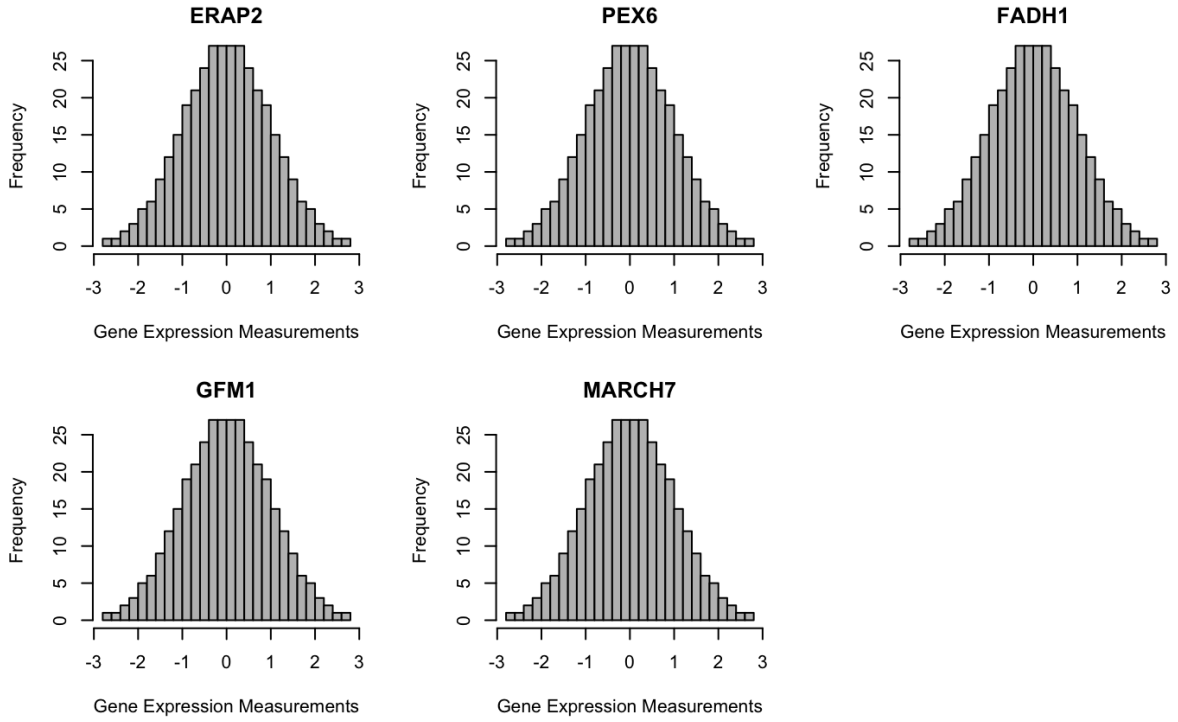

Figure 1: Histogram of Phenotype Data

The genotype data provided is encoded in 0s, 1s and 2s and is complete without any missing values. If 0s and 2s represent homozygous alleles and 1s represent heterozygous alleles, the independent variables can be dummy coded as:

$$X_a(0) = -1, X_a(1) = 0, X_a(2) = 1$$
$$X_d(0) = -1, X_d(1) = 1, X_d(2) = -1$$
$$A_1A_1 = 0, A_1A_2 = 1, A_2A_2 = 2$$

Modeling Strategy

The quantitative genetic model we use for the eQTL analysis is a multiple linear regression model with the following equation.

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + X_z\beta_z + \epsilon \tag{1}$$

$$\epsilon \sim N(0, \sigma_\epsilon^2) \tag{2}$$

Principal Component Analysis was performed on $X_aX_a^T$ to examine clusters in the genotype

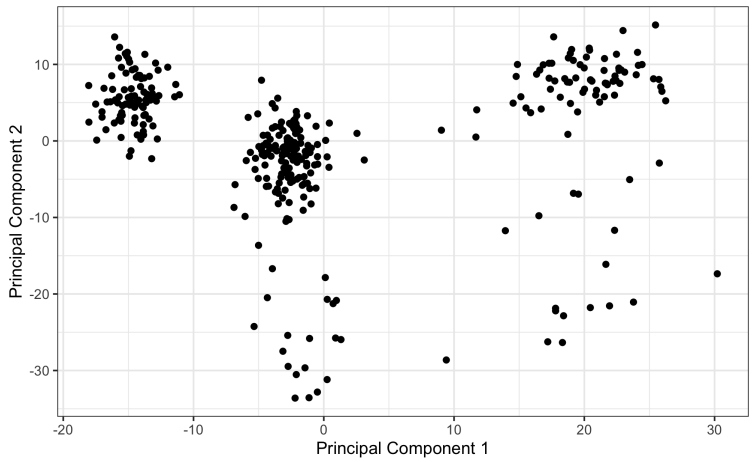

Figure 2: Principal Component Analysis

data and the results are shown in Figure 2. The clusters represent different sub-populations in our dataset. Three clusters are evident with some outliers which could represent other sub-populations. Thus, we can see five sub-populations which is in accordance with the problem description. To reduce the possibility of false positives given the strong variation among sub-populations, including covariates in our eQTL analysis is a good idea. Information capturing population structure and the gender of the individuals were provided which are included as covariates $X_{z,i}$ in the model.

$$X_{z,1}(HG) = 1, X_{z,1}(NA) = 2$$
$$X_{z,2}(CEU) = 1, X_{z,2}(FIN) = 2, X_{z,3}(GBR) = 3, X_{z,4}(TSI) = 4$$
$$X_{z,3}(Male) = 1, X_{z,3}(Female) = 2$$

The total number of parameters in our model is 7: $\beta_\mu, \beta_a, \beta_d, \beta_{z,1}, \beta_{z,2}, \beta_{z,3}, \sigma_\epsilon^2$. The linear regression model was used to get maximum likelihood estimates of the parameters which were used to compute the F-statistic by dividing the mean square model by the mean square error. The p-value is calculated from the F-statistic. To control the type 1 error and reduce the number of statistically significant SNPs, Bonferroni correction is applied by dividing the p-value threshold (0.05) by the total number of tests (50,000), resulting in a corrected p-value of 10^-6. The p-values lying below

this threshold are considered statistically significant for which we could reject the null hypothesis and infer a causal relationship between that particular genotype and phenotype.

Results and Discussions

eQTL analysis was performed with and without covariates and the results showed no difference between the two cases. Manhattan Plots were used to visualize positions of causal polymorphisms for the five expressed genes by identifying peaks of high statistical significance (or low p-values). The negative log of p-values was plotted against position in the genome, colored by chromosome (grey and black). The thick black horizontal line in each Manhattan plot represents the Bonferroni-corrected p-value threshold. All p-values that peak above this threshold are locations in the genome that considered to be significant. The SNP and gene information provided in the dataset was used to map these locations to the exact position on the chromosome through respective rsIDs. A QQ plot which compares observed p-values with expected p-values is used for inference of the results from the analysis. A 45-degree line (shown in red) is drawn for each QQ plot and deviations from this line are checked for.

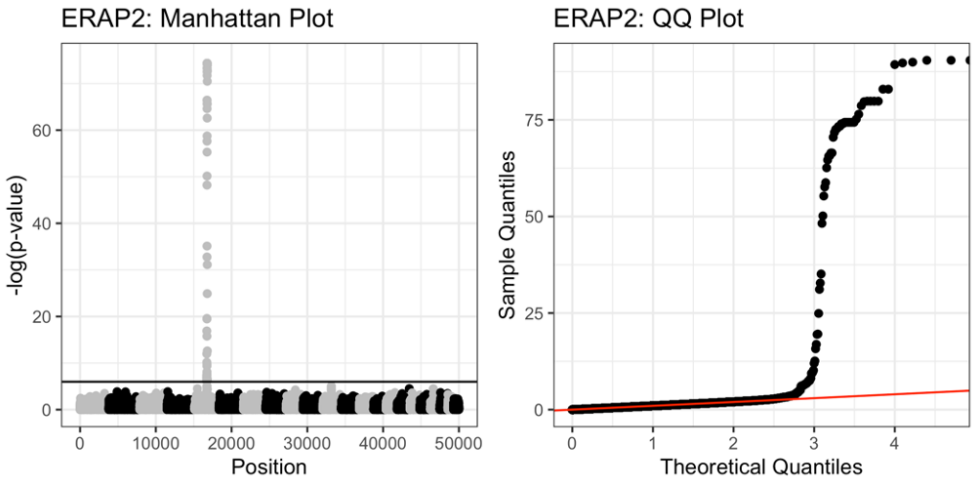

Figure 3: (left) Bonferroni-corrected Manhattan Plot and (right) QQ plot for ERP2

The Manhattan plots for ERAP2, FAHD1 and PEX6 show one peak each where the peak represents the presence of causal variants (refer Figures 3-5). In fact, the QQ plot validates this claim for the respective genes since the sample quantitles, or observed quantiles, hug the red line until it deviates at the tail. This is typically the case when there are significant hits in the Manhattan plot. Thus, we can infer that there are locations of causal genotypes. Or, in other words, we reject the null hypothesis for those significant genotypes. On the other hand, the Manhattan plots for MARCH7 and GFM1 have no points lying above the cutoff value (black line), which means that there are no causal genotypes (refer Figures 6-7). Accordingly, the QQ plots for MARCH7 and GFM1 follow the 45-degree line. This is because the expected p-values and observed p-values are almost the the same as there were no causal polymorphisms found. Thus, in other words, we cannot reject the null hypothesis for any of the genotypes.

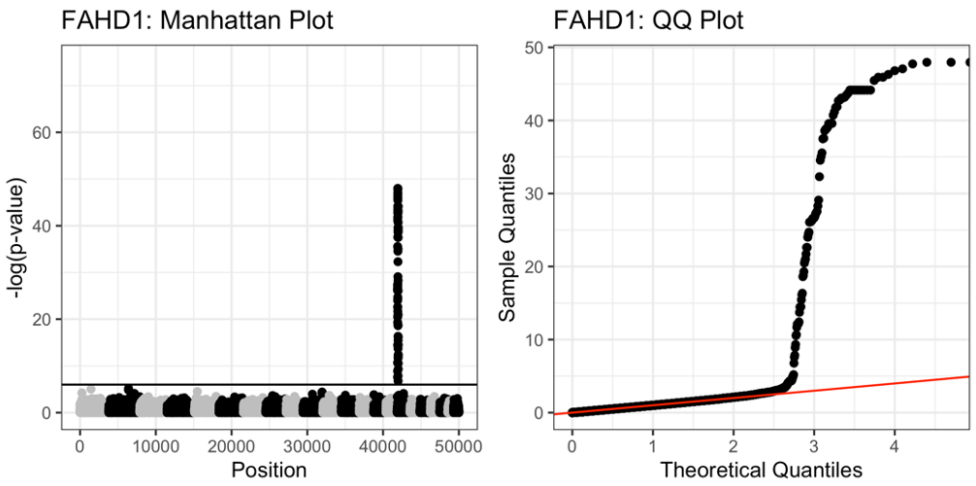

Figure 4: (left) Bonferroni-corrected Manhattan Plot and (right) QQ plot for FAHD1

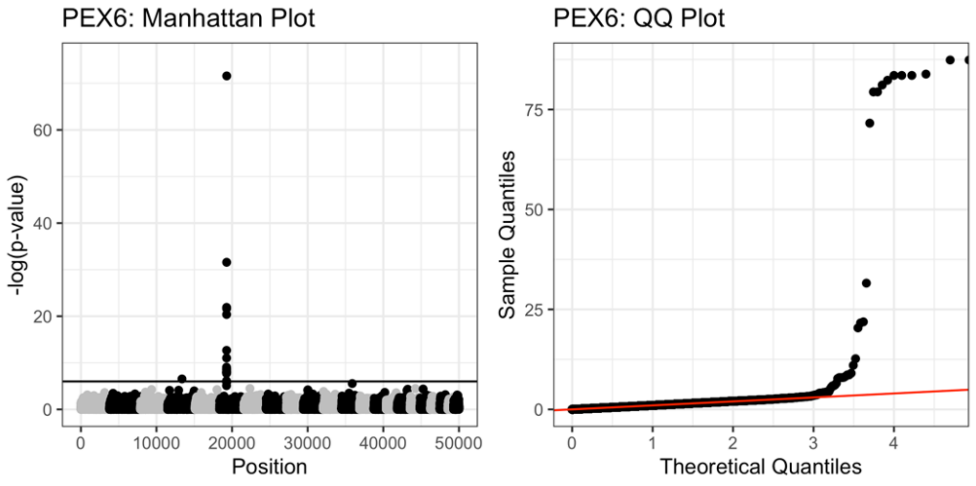

Figure 5: (left) Bonferroni-corrected Manhattan Plot and (right) QQ plot for PEX6

Table 1: Positions of causal polymorphism for each gene

Gene	**From Position**	**To Position**	**Most Significant**
ERAP2	Position: 96772432 on Chr 5	Position: 97110808 on Chr 5	969* (45338,53255,82440)
FAHD1	Position: 98486048 on Chr 4	Position: 43108015 on Chr 6	429*(64461,72496)[a]
PEX6	Position 1524250 on Chr 16	Position: 1929366 on Chr 16	18*(29958,32761,36231)
MARCH7	-	-	None
GFM1	-	-	None

[a]for chromosome 6
Total number of significant SNPs: ERAP2: 73; FAHD1: 28; PEX6: 90; MARCH7: 0; GFM1: 0

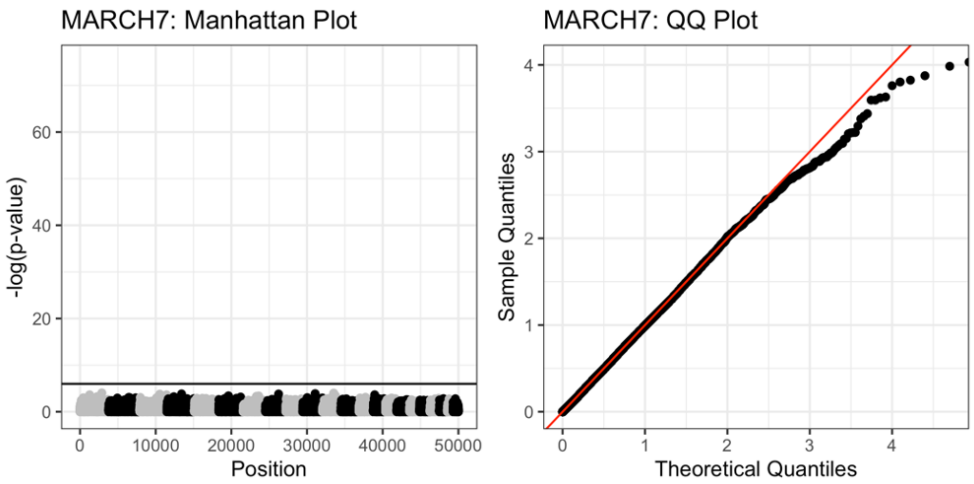

Figure 6: (left) Bonferroni-corrected Manhattan Plot and (right) QQ plot for MARCH7

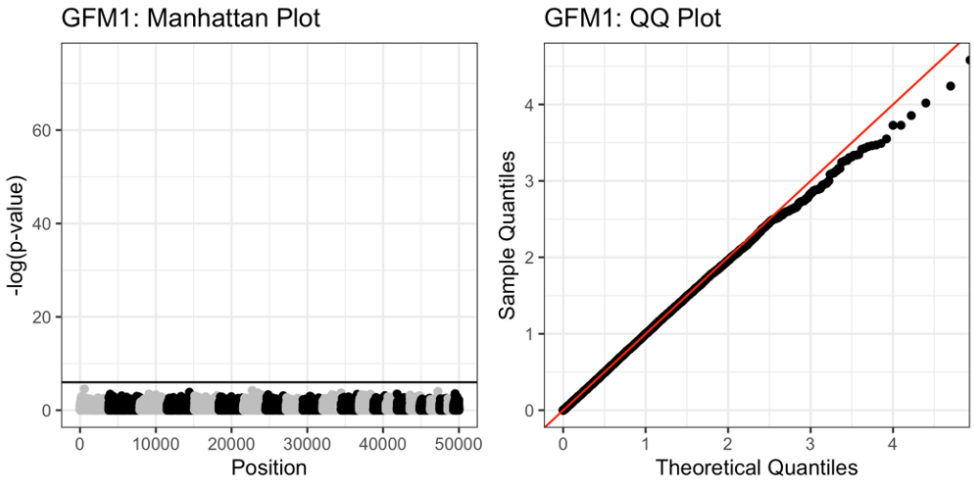

Figure 7: (left) Bonferroni-corrected Manhattan Plot and (right) QQ plot for GFM1

Since we cannot determine the precise location of the causal polymorphism by just identifying the most significant genotype, a range is defined within which the causal polymorphism could lie. Table 1 presents this range for the three genes with significant hits and also reports the position with the most significant genotype. Follow-up investigations would need to be performed to find the causal variants in these regions. FAHD1 shows a significant SNP in chromosome 4 and chromosome 6. It is likely that the hit in chromosome 4 is a false positive because of a relatively smaller peak, arising due to several reasons. It would be advisable to investigate chromosome 6 around the most significant SNP instead for a causal polymorphism for follow-up analysis.

Conclusions

The results from the eQTL analysis reveal causal genotypes present in chromosome 5, chromosome 6 and chromosome 16 which are associated with genes ERAP2, FAHD1 and PEX6, respectively. There seem to be no causal genotype that is associated with the genes MARCH7 and GFM1. The genotypes close to the causal variant on the chromosome are correlated with the causal variant through linkage disequilibrium which results in a locus of highly significant genotypes. Although each locus could have more than one causal variant, we can assume that there is a single casual variant in this case as generally observed in humans. With the help of these findings, further analysis through follow-up experiments should focus on discerning the precise causal variant from the identified loci.

In fact, our findings agree well with that reported in the literature. The genes were found in the list of human protein sequence entries encoded on their respective chromosomes [3]. Gene ERAP2 encodes for protein Endoplasmic reticulum aminopeptidase 2, and other GWA studies have demonstrated associations of these proteins with immune-mediated diseases such as ankylosing spondylitis, psoriasis, Behçet's disease, inflammatory bowel disease and type I diabetes [1]. Gene PEX6 encodes for the protein Peroxisome assembly factor 2 and mutations in this gene can result in genetic diseases such as Zellweger Syndrome spectrum, infantile Refsum disease, and neonatal adrenoleukodystrophy. These genetic diseases are autosomal recessive and occur in 1 of every 50,000 births [2]. FAHD1 gene encodes for Fumarylacetoacetate hydrolase domain-containing 1 protein and diseases associated with this gene include Ecthyma and Cerebral Lymphoma [7]. Although these genes are known to us, GWAS analysis can cognize the function and relevance of other previously unsuspected genes, and experimental follow-up of loci can lead to the discovery of novel biological mechanisms.

A point to note here is that GWAS indentifies causal variants using information in the population database accessed for the study and the results cannot always be extrapolated to other diverse worldwide populations. Since most GWAS have been performed primarily in populations of European descent, as in this project, the identified causal variants cannot be translated to other non-European populations [6]. For this reason, alternative approaches such as linkage analysis or even complete genome sequencing can be used to circumvent some of the shortcomings of traditional GWAS. Nonetheless, GWAS has the ability to advance genetic mapping of complex diseases by identifying areas of exploration and probing unsought directions for research on disease mechanisms.

References

[1] Nitish Agrawal and M Brown. Genetic associations and functional characterization of m1 aminopeptidases and immune-mediated diseases. *Genes and immunity*, 15, 08 2014.

[2] Nancy Braverman, Gerald Raymond, William Rizzo, Ann Moser, Mark Wilkinson, Edwin Stone, Steven Steinberg, Michael Wangler, Eric Rush, Joseph Hacia, and Mousumi Bose. Peroxisome biogenesis disorders in the zellweger spectrum: An overview of current diagnosis, clinical manifestations, and treatment guidelines. *Molecular Genetics and Metabolism*, 117, 12 2015.

[3] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 11 2018.

[4] Joel Hirschhorn and Mark Daly. Genome-wide association studies for common diseases and complex traits. *Nature reviews. Genetics*, 6:95–108, 03 2005.

[5] Tuuli Lappalainen, Michael Sammeth, Marc Friedländer, Peter Hoen, Jean Monlong, Manuel Rivas, Mar Gonzàlez Porta, Natalja Kurbatova, Thasso Griebel, Pedro Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, and Emmanouil Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501, 09 2013.

[6] Noah Rosenberg, Lucy Huang, Ethan Jewett, Zachary Szpiech, Ivana Jankovic, and Michael Boehnke. Genome-wide association studies in diverse populations. *Nature reviews. Genetics*, 11:356–66, 05 2010.

[7] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Stein, Ron Nudel, Iris Lieder, Yaron Mazor, Sergey Kaplan, Dvir Dahary, David Warshawsky, Yaron Guan-Golan, Asher Kohn, Noa Rappaport, Marilyn Safran, and Doron Lancet. *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses: The GeneCards Suite*, volume 54, pages 1.30.1–1.30.33. 06 2016.

[8] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 05 2019.