BTRY 6830 - Quantitative Genomics

Project Report

Abhishek Murti - am3248

12th May 2020

1 Introduction

Genome Wide Association Study (GWAS) is a technique of analyzing genetic sequences and identifying the alleles associated with a certain phenotype. This phenotype can be something trivial such as height or hair color. But the power of GWAS can be utilized to focus on more consequential phenotypes such as heart disease, diabetes, Alzheimer's disease, autism, Crohn's disease, breast cancer, bipolar disorder, asthma, and high cholesterol.

GWAS focus on single nucleotide polymorphisms (SNP's) in order to determine a correlation between a certain allele and a certain disease. It is important to understand that the results from a GWAS only provide an association between a certain allele and a certain phenotype rather than a definitive conclusion. This is because of the following reasons [5]-

- Most traits depend on a number of genes rather than a particular gene or allele
- Some traits are affected by factors other than genetics such as, lifestyle and environment
- Some dynamic factors such as age affect the probability of a person suffering the disease or the risk of a disease. These cannot be effectively incorporated in the GWAS study

2 Data

Among the recent large scale human genomics resources is Genetic European Variation in Health and Disease (gEUVADIS) with a samples from 4 different European populations (5 populations total). Each of these individuals were part of the 1000 Genomes project and their genomes were sequenced and analyzed to identify SNP genotypes. We have been provided is a small subset of these data that are publicly available. Specifically, we have been provided 50,000 of the SNP genotypes for 344 samples from the CEU (Utah residents with European ancestry), FIN (Finns), GBR (British) and, TSI (Toscani) population. For these same individuals, we have also been provided the expression levels of five genes.

2.1 Gene info

There are 5 genes present in the data

- MARCH7 - Encodes for the E3 ubiquitin-protein ligase. E3 ubiquitin ligases accept ubiquitin from an E2 ubiquitin-conjugating enzyme in the form of a thioester and then directly transfer the ubiquitin to targeted substrates[1]. They may also be involved in T-cell proliferation by regulating LIF secretion
- FAHD1 - Encodes for Fumarylacetoacetate hydrolase domain-containing protein 1. FAHD1 is a mitochondrial enzyme that hydrolyzes acetylpyruvate and fumarylpyruvate[3], as well as oxaloacetate[4]
- PEX6 - Encodes for Peroxisome assembly factor 2 protein. PEX6 (and PEX1) removes PEX5 from the peroxisomal membrane so that PEX5 may do additional rounds of peroxisomal import. Mutations in the genes encoding PEX6, along with PEX1, are the leading causes of peroxisomal biogenesis disorders.
- ERAP2 - Encodes for Endoplasmic reticulum aminopeptidase 2 protein. Aminopeptidases hydrolyze N-terminal amino acids of proteins or peptide substrates. Major histocompatibility complex (MHC) class I molecules rely on aminopeptidases such as ERAP1 and LRAP to trim precursors to antigenic peptides in the endoplasmic reticulum (ER) following cleavage in the cytoplasm by tripeptidyl peptidase II.[6]
- GFM1 - Encodes for Elongation factor G 1, mitochondrial protein. Eukaryotes contain 2 translational systems, one in the cytoplasm and the other in the mitochondria. In mitochondria, the elongation phase requires 3 elongation factors. GFM1 is one of those factors. This gene is associated with Combined oxidative phosphorylation deficiency 1.[2]

2.2 Genotype data

There were no missing values in the provided genotype data. There was also no need of removing values that were below a minimum allele frequency (MAF).

2.3 Phenotype Data

The phenotype data is the expression levels of 5 genes. Before beginning the actual GWAS analysis, it is important to check whether the phenotype data is normally distributed.

2.4 Covariates

There are three covariates in our data set. Population ID, region and sex.

3 GWAS

Since our phenotype is gene expression level we are going to perform a different type of GWAS called an 'expression Quantitative Trait Locus' or 'eQTL' analysis.

3.1 Transformation of genotype data

Since the available data was coded as 0,1 and 2, it was transformed into the general X_a coding i.e.,

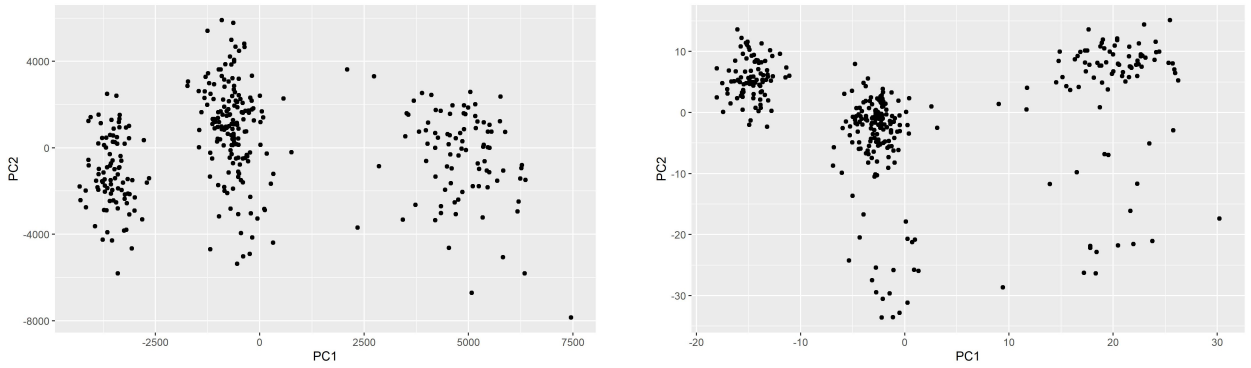
$$X_a(A_1A_1) = -1, X_a(A_1A_2) = 0, X_a(A_2A_2) = 1$$

Correspondingly the X_d matrix was coded as:

$$X_d(A_1A_1) = -1, X_d(A_1A_2) = 1, X_d(A_2A_2) = -1$$

3.2 Data exploration

Principal component analysis was performed on $X_aX_a^T$ and X_a matrices. The transformed data was plotted along the first two principal components. From the PCA on $X_aX_a^T$ we see three different clusters with a few outliers. But the PCA on X_a results in 5 clusters. This is in line with the given data, that has a 5 different populations.



(a) PCA on $X_aX_a^T$

(b) PCA on X_a

Figure 1: Principal Component Analysis

3.3 Checking Phenotype Data

Our model is based on linear regression. On of the most important characteristics of the data on which linear regression is performed is that the residuals should be normally distributed. Our phenotype data is normally distributed which means that we can proceed with our analysis.

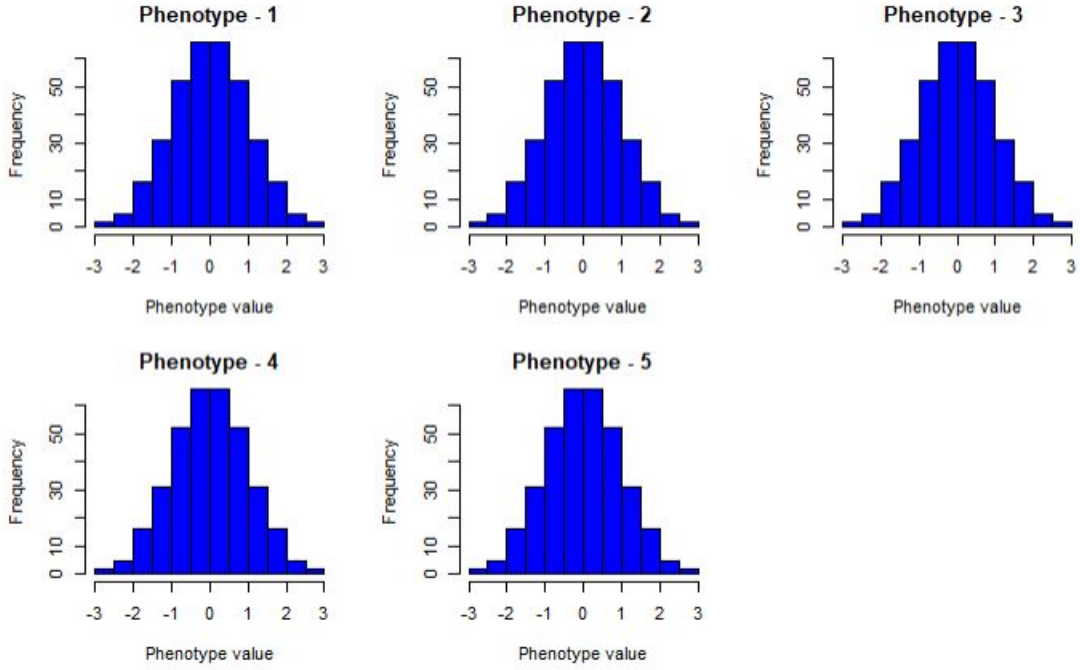


Figure 2: Phenotype data distribution

3.4 Linear Regression Model

The linear regression model is defined as-

$$y = X\beta + \epsilon \tag{1}$$

$$MLE(\hat{\beta}) = (x^T x)^{-1} x^T y \tag{2}$$

$$\hat{y_i} = \hat{\beta_\mu} + x_{i,a}\hat{\beta_a} + x_{i,d}\hat{\beta_d} \tag{3}$$

3.5 Incorporating covariates

The three types of covariates were coded as follows-

- Population ID - $X_{z,1}(HG) = -1$; $X_{z,1}(NA) = 1$
- Population Region - $X_{z,2}(GBR) = 1$; $X_{z,2}(FIN) = 2$; $X_{z,2}(CEU) = 3$; $X_{z,2}(TSI) = 4$
- Population Sex - $X_{z,3}(Male) = -1$; $X_{z,3}(Female) = 1$

The equation for modeling covariates has a few modifications

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \Sigma_{i=i}^{n} X_{z,i}\beta_{z,i} + \epsilon \tag{4}$$

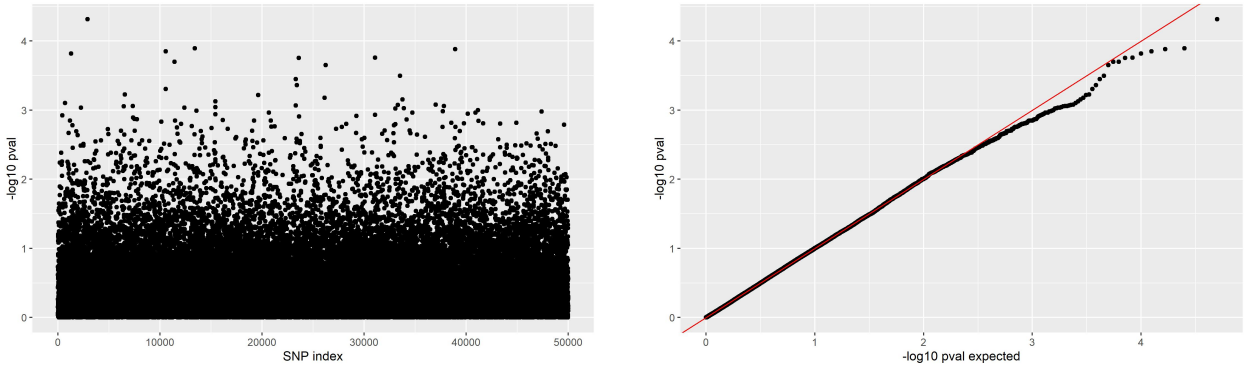
4 Results

After running the eQTL analysis the following results were obtained. For calculating the number of significant SNP's a cut-off value of $\alpha = 0.05$ was used.

4.1 Manhattan and QQ Plots

- MARCH7:

  There were 2438 significant SNP's for this gene. But by observing the scale of the Manhattan plot, we can clearly see that there are no significant hits. This is also indicated by the QQ Plot in figure 3(b). The QQ plot doesn't resemble the ideal GWAS scenario and nor does it have a tail. This means that it would be wrong to interpret the results of the GWAS for this phenotype. After applying a Bonferroni correction, we can see that there are no significant SNP's below the new cut-off value. ($\alpha = 0.05/N$; N is the total number of SNP's)



(a) Manhattan plot

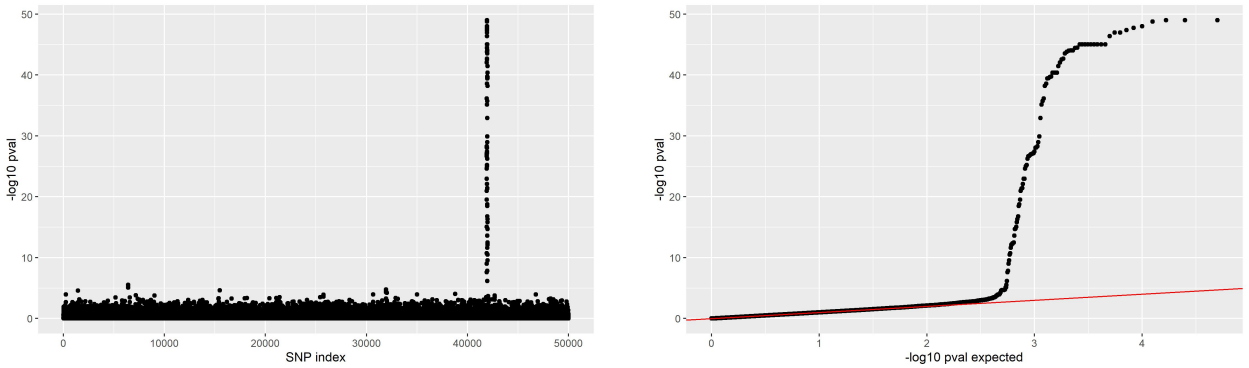
(b) QQ Plot

Figure 3: MARCH7

- FAHD1:

  The Manhattan plot shows that there are a few significant hits. After applying a Bonferroni correction, there were 90 significant SNP's. The QQ plot has a tail which means that the results of this GWAS can be trusted to suggest a correlation between the genotype and the particular phenotype.

Most significant SNP's (Lowest p-value)			
SNP index	Chromosome	Position	ID
41938	16	1829958	rs11644748
41942	16	1832761	rs140254902
41944	16	1836231	rs9652776

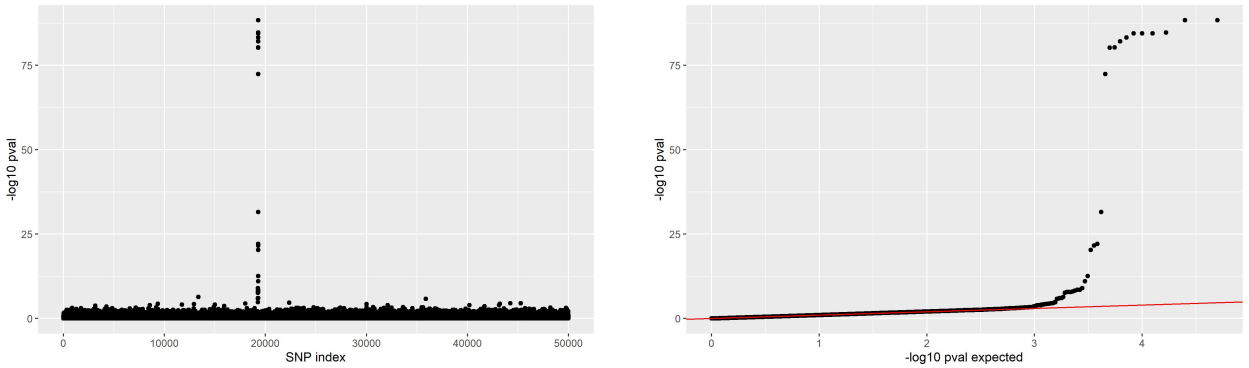


(a) Manhattan plot

(b) QQ Plot

Figure 4: FAHD1

- PEX6: After applying a bonferroni correction, we obtain 29 significant SNP's. This can be observed in the Manhattan plot and the QQ plot suggests that the analysis can be trusted.



(a) Manhattan plot

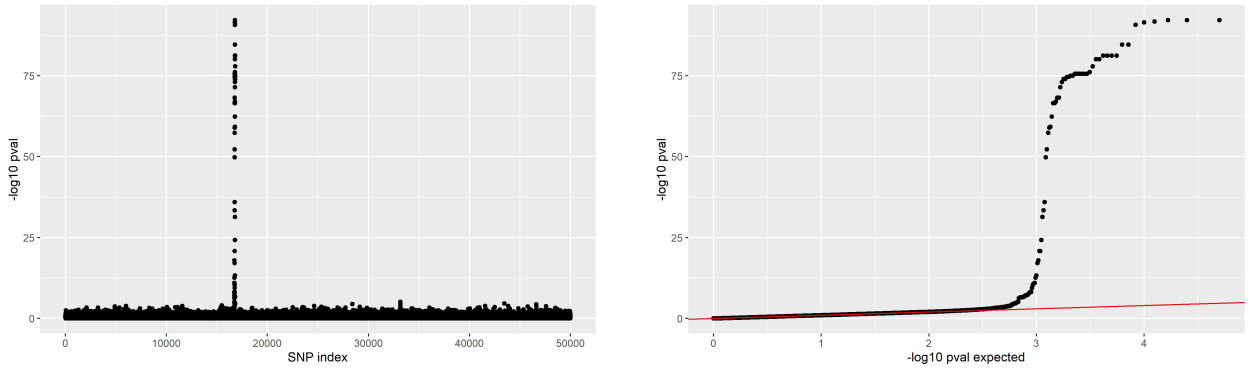
(b) QQ Plot

Figure 5: PEX6

Most significant SNP's (Lowest p-value)			
SNP index	Chromosome	Position	ID
19286	6	42964461	rs1129187
19288	6	42972496	rs10948061

- ERAP2:

  We obtain 73 significant SNP's for this particular phenotype. The QQ plot has a tail which means the analysis is correct and can be used to make a conclusion.



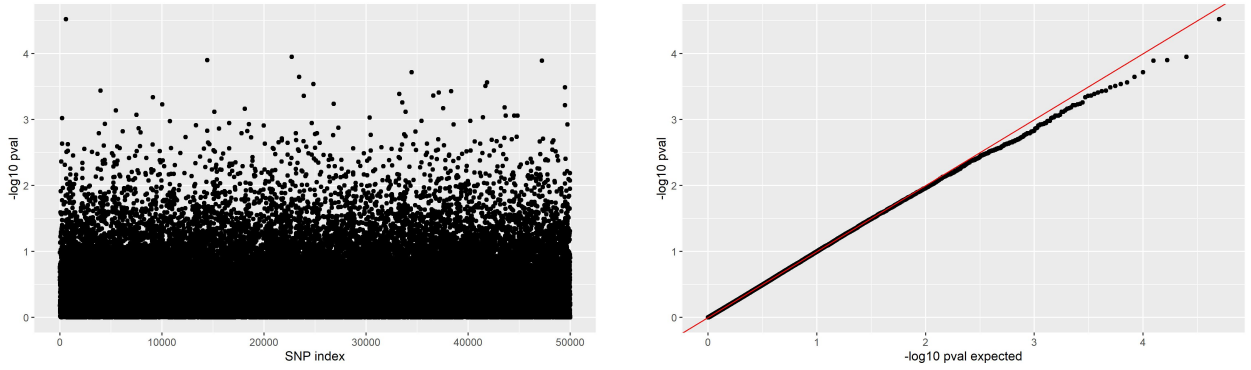
(a) Manhattan plot (b) QQ Plot

Figure 6: ERAP2

Most significant SNP's (Lowest p-value)			
SNP index	Chromosome	Position	ID
16784	5	96945338	rs7726445
16786	5	96953255	rs7731592
16791	5	96982440	rs2161548

- GFM1:

  There were 2475 p-values below the cut-off of 0.05. But after correcting the p-values, there were zero significant SNP's. This is in line with the QQ plot which cannot be used to make a conclusion about this analysis.

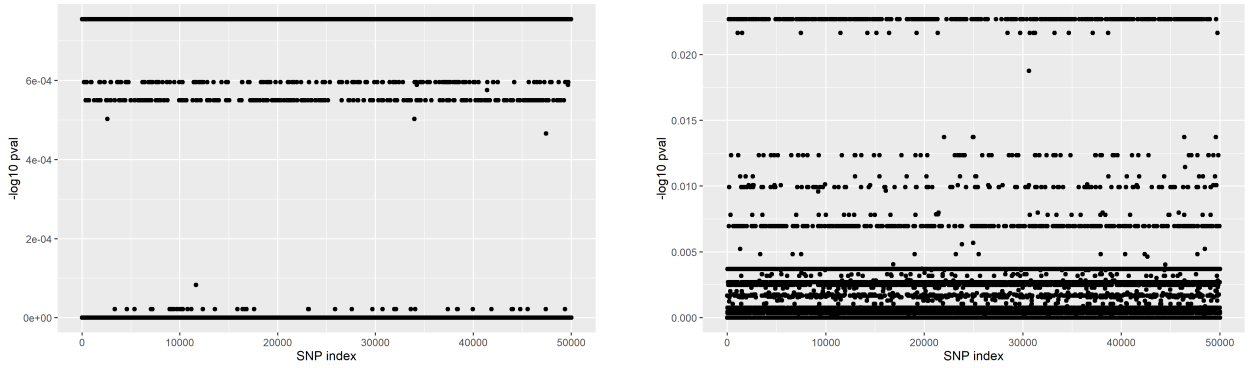


(a) Manhattan plot (b) QQ Plot

Figure 7: GFM1

4.2 Corrected p-values

Although for calculating the number of significant SNP's the α was changed by incorporating a bonferroni correction, other methods of correcting the p values were also employed in order to be thorough. The p-values were corrected using a FDR correction as well. There was no difference in the Manhattan plots for ERAP2, PEX6 and FADH1. But the scale of the Manhattan plots for the corrected values for MARCH7 and GFM1 further supported our finding that the GWAS did not yield any significant SNP's.



(a) Manhattan plot for GFM1

(b) Manhattan plot for MARCH7

Figure 8: FDR correction

4.3 Effect of covariates

Three covariates namely, population ID, region and sex were added to the GWAS. The number of significant SNP's for MARCH7 and GFM1 were still zero. The number of significant SNP's for ERAP2, PEX6 and FAHD1 changed from 73, 29 and 90 to 72, 28 and 90, respectively. There was virtually no effect of covariates on the number of signifcant SNP's for any of the phenotypes.

Along with the given covariates, the first two principal components of the $X_a X_a^T$ were also added as covariates but the there was so noteworthy effect of that as well. The number of significant SNP's for ERAP2 changed from 72 to 50, but the change wasn't significant enough to warrant any conclusion.

5 Discussion

The expression levels of three genes, namely, ERAP2, PEX6 and FAHD1 can be influenced by the identified single nucleotide polymorphisms. Even after incorporating the effect of covariates the results were unaffected. This was expected because we could distinctly identify the peaks in the Manhattan plots which clearly demonstrate the presence of few significant single nucleotide polymorphisms, for all three genes. The tails in the corresponding QQ plots provide further conviction in our results of the eQTL analysis. The influence of those SNP's can be further verified by the fact that even after correcting the p-values using several methods

such as Bonferroni and FDR, there was no change in the Manhattan plots or the number of significant SNP's. Neither did adding covariates influence the results. For the other two genes, MARCH7 and GFM1, the Manhattan plots indicate no significant hits. The enormous deviation from the 45°line in the QQ plot is indicative of the same. In order to be comprehensive, covariates were added to this analysis but with seemingly no effect. Therefore we can safely suggest that expression levels of MARCH7 and GFM1 are not influenced by any particular set of alleles.

References

[1] Flierman, D., Coleman, C. S., Pickart, C. M., Rapoport, T. A., and Chau, V. E2-25k mediates us11-triggered retro-translocation of mhc class i heavy chains in a permeabilized cell system. *Proceedings of the National Academy of Sciences 103*, 31 (2006), 11589–11594.

[2] Online Mendelian Inheritance in ManOMIM. Omim database. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) https://omim.org/.

[3] Pircher, H., Straganz, G. D., Ehehalt, D., Morrow, G., Tanguay, R. M., and Jansen-Dürr, P. Identification of human fumarylacetoacetate hydrolase domain-containing protein 1 (fahd1) as a novel mitochondrial acylpyruvase. *Journal of Biological Chemistry 286*, 42 (2011), 36500–36508.

[4] Pircher, H., von Grafenstein, S., Diener, T., Metzger, C., Albertini, E., Taferner, A., Unterluggauer, H., Kramer, C., Liedl, K. R., and Jansen-Dürr, P. Identification of fah domain-containing protein 1 (fahd1) as oxaloacetate decarboxylase. *Journal of Biological Chemistry 290*, 11 (2015), 6755–6762.

[5] Stranger, B. E., Stahl, E. A., and Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics 187*, 2 (2011), 367–383.

[6] Tanioka, T., Hattori, A., Masuda, S., Nomura, Y., Nakayama, H., Mizutani, S., and Tsujimoto, M. Human leukocyte-derived arginine aminopeptidase the third member of the oxytocinase subfamily of aminopeptidases. *Journal of Biological Chemistry 278*, 34 (2003), 32275–32283.