Preliminary Genome-wide Association Analysis of Insulin Resistance in Americans

by

Ashley Choi

A Capstone Project

Presented to the Department of Medical Informatics & Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Biomedical Informatics

August 2016

School of Medicine

Oregon Health & Science University

Master of Biomedical Informatics

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone Project of

Ashley Choi

"Preliminary Genome-wide Association Analysis of Insulin Resistance in Americans"

Has been approved

_____

Guanming Wu, Ph.D.
Capstone Advisor

Table of Contents

Acknowledgments

I would like to express my gratitude to God for blessing me to complete this internship.

I am grateful to my capstone advisor Dr. Guanming Wu for his support and guidance. This internship would not have been possible without the help from my internship coordinator Diane Doctor. I also would like to acknowledge Dr. Shannon McWeeney and Dr. Christina Zheng for guiding me through my academic years at OHSU.

I would like to thank my internship supervisor Dr. Anna Lyubetskaya for the challenging and exciting project. I would like to acknowledge Dr. Evan Rosen and Dr. Linus Tsai for this internship opportunity.

Lastly, I thank my family and friends for their unwavering faith, love and support.

Introduction

According to American Diabetes Association, approximately 30 million Americans are diagnosed with diabetes as of 2012 [1]. This means that more than ten percent of the American population is diagnosed with diabetes[1]. Diabetes is a disease that affects not only Americans, but also people worldwide. The International Diabetes Federation reports that the number of people with diabetes is expected to increase from 415 million to 642 million by 2040 [2]. There are different types of diabetes. Type 1 diabetes patients produce limited amounts of insulin or none at all, which results in increased glucose levels in the bloodstream [3,4]. Type 2 diabetes patients are capable of producing insulin normally, however the insulin does not function properly, resulting in increased glucose levels in the bloodstream [3,4]. This is called insulin resistant. 10% of diabetes patients are type 1, while 90% of patients are type 2 [4]. There is clear evidence that both genetic and environmental factors increase the susceptibility to type 2 diabetes; however, previous studies have shown that epigenetic factors are also involved in type 2 diabetes. Epigenetic factors include DNA methylation, histone modification and miRNA[5]. I participated in a large on-going epigenetic analysis of insulin resistance in human adipocytes. The aim of the project was to discover and study genetic variants that are associated with insulin resistance, and examine how the single nucleotide polymorphisms (SNPs) look different between insulin resistance and insulin sensitive patients.

Data

I worked with sequencing (RNA-seq and ChIP-seq) data and microarray genotyping data collected and processed from the adipose tissue of 47 female patients. The main reason for using adipose tissue is it allows scientists to separate adipocytes from oil to obtain purified adipocytes. These purified adipocytes are used for RNA-sequencing, DNA methylation and genotyping. I worked with RNA-seq and ChIP-seq (H3K27ac, H3K4me1, and H3K4me3) to establish pipelines for GATK. For genotyping data, I worked with only Caucasian samples. Among the total of 43 Caucasian samples, 26 of them were insulin resistant and 17 of them were insulin sensitive.

Methods/Tools

1) **GATK for Sequencing data**

To establish the GATK piplines, I performed the following steps:

1. Build an index for the Bam files using Picard.
2. Reorder the Bam files using Picard.
3. Run the SNP discovery tool, GATK.
4. Filter the SNPs based upon annotations using GATK.

The following are the tools that I used for SNP calling the sequencing data.

Picard [6]: was used to prep input files for GATK. The two main features used in Picard were:

- The "BuildBamIndex" feature builds an index for the Bam files. The index is the required input for running variant calling (GATK).
- The "ReorderSam" feature reorders Bam files.

GATK [7]: was for calling SNPs and filteration. The main GATK features used were:

- "HaplotypeCaller" was used for calling SNPs.
- "VariantFiltraton" was used for filtering variant called based on annotations.

2) Imputation for genotyping data

The following are the steps I used for genotype imputation.

1. Resort the Bim File.
2. Use SNPFLIP to find SNPs that have reverse strands.
3. Run PLINK to flip the SNPs that have the reverse strands, and convert genotyping data into a VCF file.
4. Run ShapeIt for quality control and for phasing.
5. Run IMPUTE2 for genotype imputation.
6. Run SNPTest for SNP analysis.

Following is the list of tools that were used for genotyping imputation:

SNPFLIP [8]: was used to find SNPs with reverse strands.

PLINK [9]: was for quality control and reformatting the data file.

- To correct SNPs with the stand order, "flip" was used.
- "Recode" converted genotyping ped files to VCF files.

SHAPEIT[10]: was for quality control and phasing.

- It checked to remove any missing, misaligned and duplicated SNPs
- Phasing determined which SNPs are inherited together and determined which nucleotide belongs to which chromosome.

IMPUTE2[11]: was for genotype imputation

- For imputation on phased genotyping data, "use_prephased_g" was used. Imputation is performed to impute any missing genotype information.

SNPTest[12]: was for SNP analysis

- It was used for computing summary statistics using "summary_stats_only"

Although SNPTest performs the SNP association test using Frequentist and Bayes' statistics, I decided to use the Euclidean distance as a straightforward intuitive method to calculate the number of SNPs that are significant different between insulin resistance and insulin sensitive patients instead. I calculated the Euclidean distance using genotyping counts. Table 1 is a summary output of SNPTest results for genotyping data. It produces genotyping counts for each SNPs. Figure 1 shows how I used information from Table 1 to calculate the Euclidean distance. I calculated the Euclidean distance for all imputed SNPs, and the number of total imputed SNPs was 77,355,572.

Results

Figure 2 is a histogram of SNPs with a Euclidean distance higher than 10. The number of SNPs that have Euclidean distance higher than 10 was 5,917,485. We can clearly see that this histogram is positively skewed. I was interested in investigating the SNPs at the end tail of this distribution, the area within the red box in figure 2. Figure 3 is the histogram of SNPs that have Euclidean distance higher than 50. By increasing the Euclidean distance from 10 to 50, the number of SNPs has decreased to 75,189 after filtering. I wanted to explore how SNPs look different between insulin resistant patients

and insulin sensitive patients. So, I chose two SNPs with high Euclidean score, and examined how their genotyping look different. Figure 4 shows how a SNP look different between insulin resistant and insulin sensitive patients. This particular SNP has a Euclidean score of 96.2. Even with a short glimpse, it is clear that there is a significant difference between two groups of patients. The majority of insulin resistant patients were homozygous AA, while the majority of insulin sensitive patients were heterozygous TA. Figure 5 is another example of a SNP with significant difference between the two groups. This SNP also has a very high Euclidean distance, 93. While all insulin sensitive patients were homozygous AA, only 31% of insulin sensitive patients were homozygous AA. To further investigate the SNPs that have a Euclidean distance higher than 50, I intersected these SNPs with insulin resistant and insulin sensitive differentially expressed genes. 284 SNPs were within the regions of differentially expressed genes. Figure 6 is a Venn diagram that shows the number of differential SNPs with a Euclidean distance higher than 50 in differentially expressed genes. I plotted a histogram for the 284 SNPs that are within the regions of differentially expressed genes. The maximum number of differential SNPs in one gene was 16, and the gene was called GRID1, glutamate ionotropic receptor delta type subunit 1. Figure 7 is the histogram of differential SNPs in differential expressed genes.

Conclusion

The initial scope of this project was to discover SNPs using GATK with sequencing data, perform genotype imputation with microarray genotyping data, and

integrate these datasets together to compare SNPs to insulin-resistant profiles. Listed below is my original plan, which was supposed to start on July 1st.

Initial Proposed Timeline:

1. Set-Up (7/1)
2. Write a wrapper script for the GATK pipeline (7/29)
3. Apply the script to ChIP and RNA-seq data (8/10)
4. Compare GATK results with the experimental genotyping data (8/16)
5. Match SNPS from both genotyping and GATK to insulin-resistant profiles (8/24)
6. Prepare a presentation and present to the group (8/26)
7. Write aproject report (9/2)

However, due to a delay during the process of contract negotiations with Beth Israel and a delay with access permission to the Broad Institute server, I was officially able start this project on July 19th. The timeline below is the final version that I followed to complete this project.

Final Project Timeline:

1. Set-Up (7/11)
2. Broad Institute access (7/19)
3. Run GATK on RNA-seq data for one patient sample (8/1)
4. Prep microarray genotyping data (8/5)
5. Phase genotyping data using ShapeIT (8/10)
6. Impute genotyping data using IMPUTE2 (8/22)

7. Perform Analyses using SNPTest (8/25)
8. Analyze results from SNPTest (8/30)
9. Prepare presentation slides (9/2)

For this project, I had an opportunity to work with a variety of different files. I learned how to unify and reformat different formats of files, so that files are compatible with each other as well as to bioinformatics tools. I also had an opportunity to work in Linux environment, and I am glad that I now feel more comfortable writing and running commands to manipulate and analyze data sets using Linux commands. Throughout this internship at Beth Israel, I was able to attend weekly computational meetings, where multidisciplinary scientists discussed actual biomedical research difficulties and obstacles. I am very glad to have obtained a real "hands-on" experience and participate on a global-scale project to better understand the underlying mechanism of insulin resistance.

Appendix

Table 1. Output of SNPTest results

Chr	position	Insulin Resistance				Insulin Sensitive			
Chr1	11008	CC	CG	GG	ALL	CC	CG	GG	ALL
		17.9	**7.8**	**0.32**	**26**	**13.3**	**3.6**	**0.09**	**17**

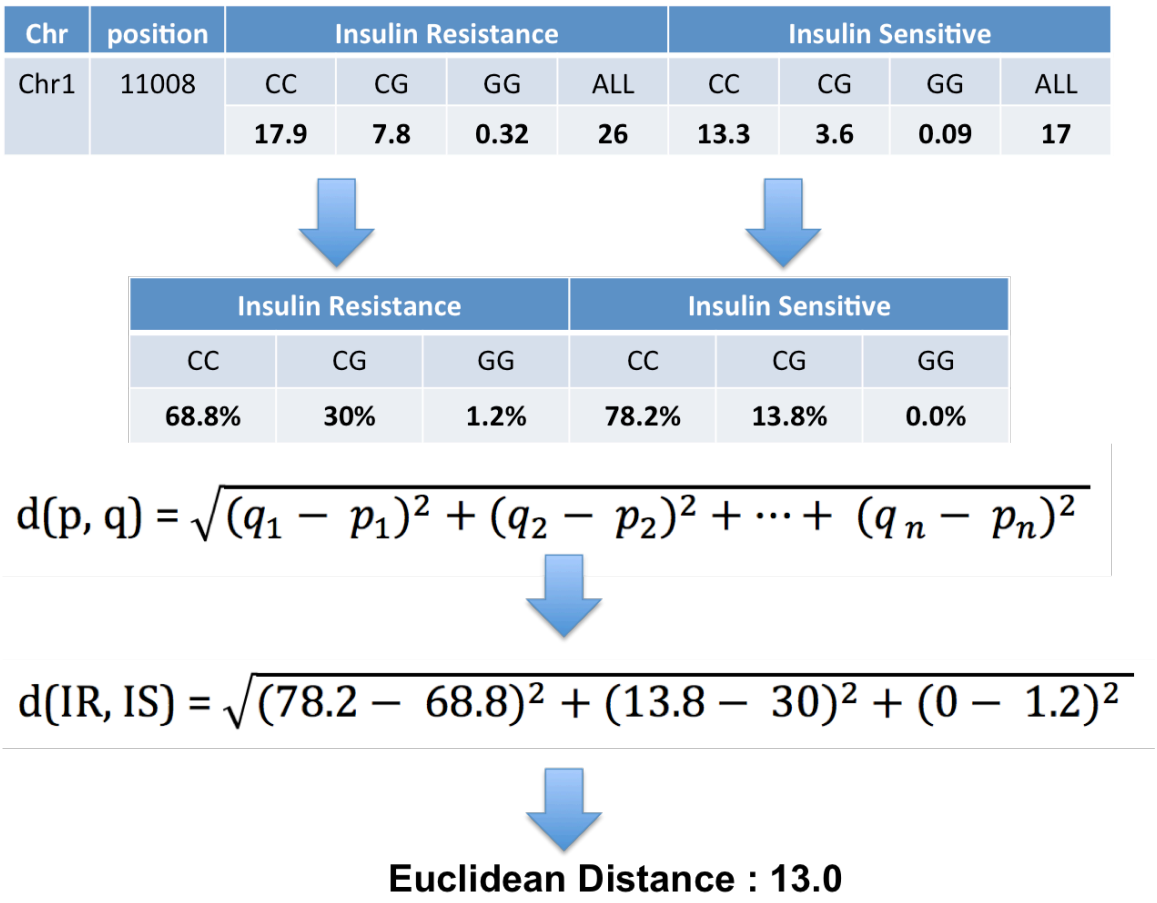


Figure 1. Calculating Euclidean distance for each SNP

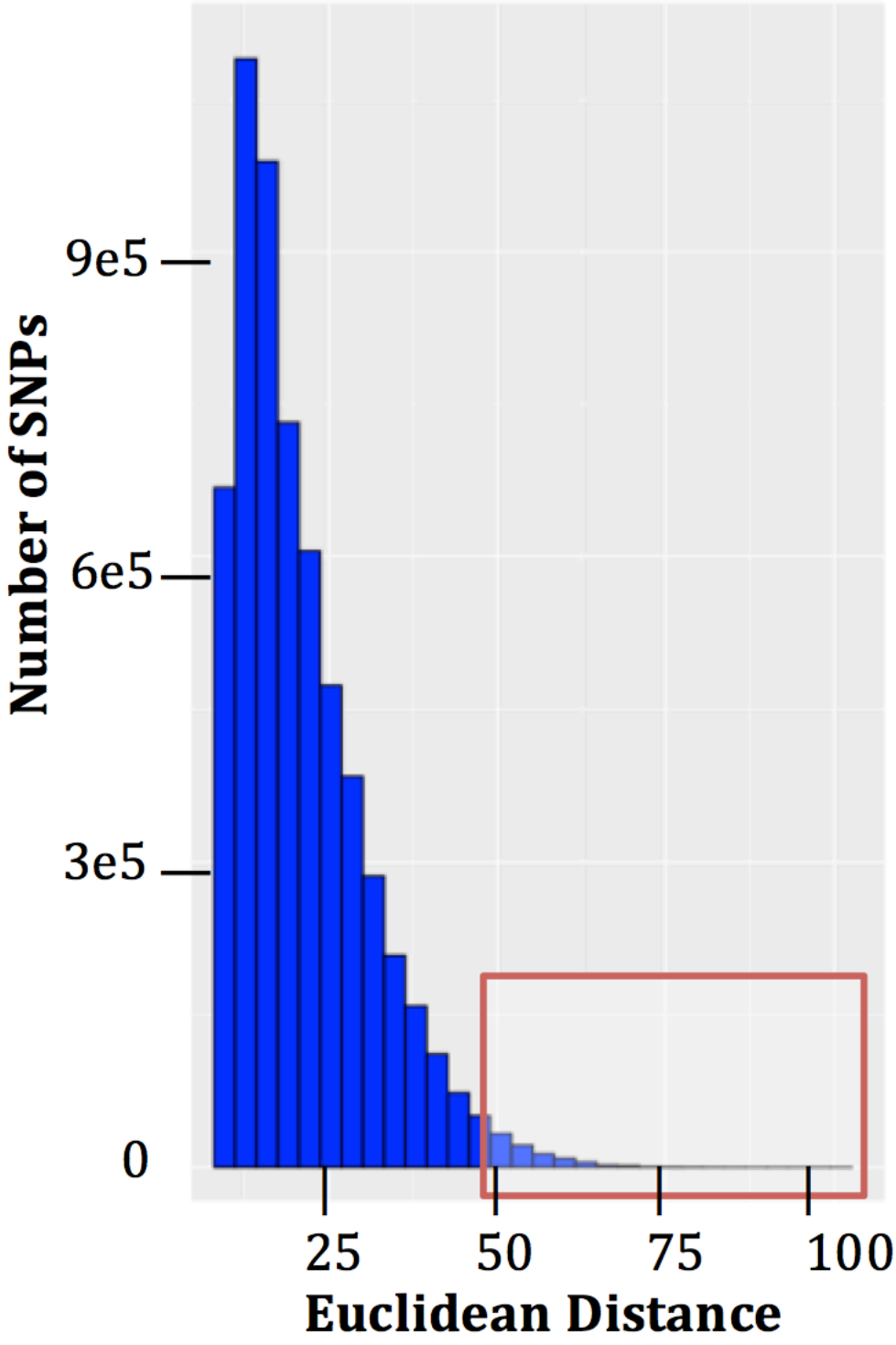


Figure 2. Histogram of SNPs with Euclidean distance higher than 10

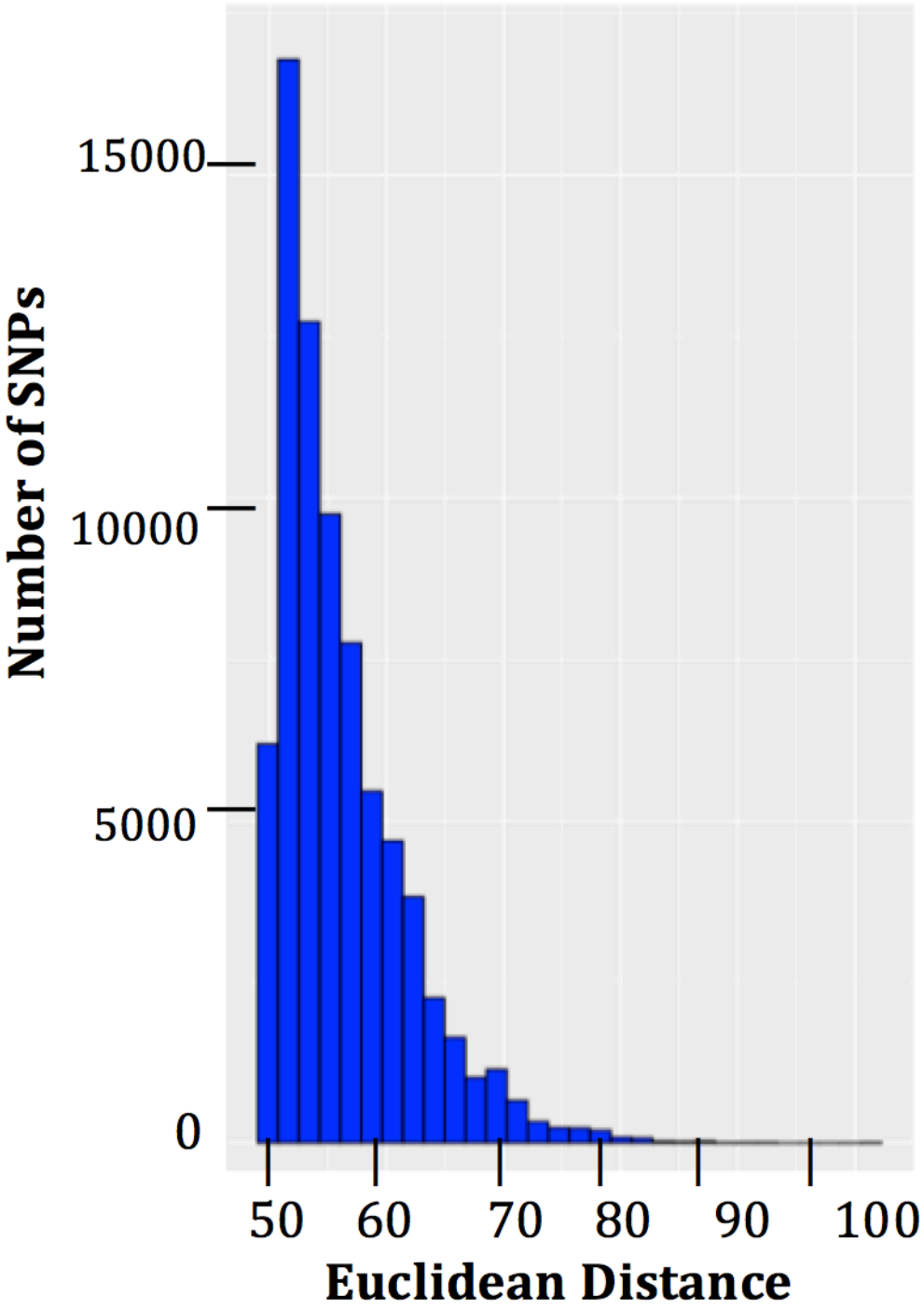
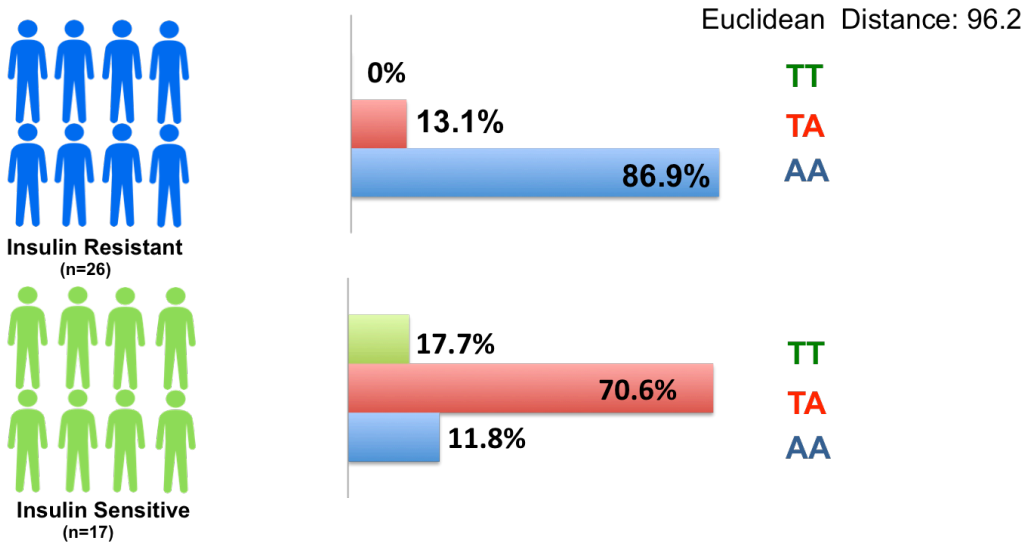
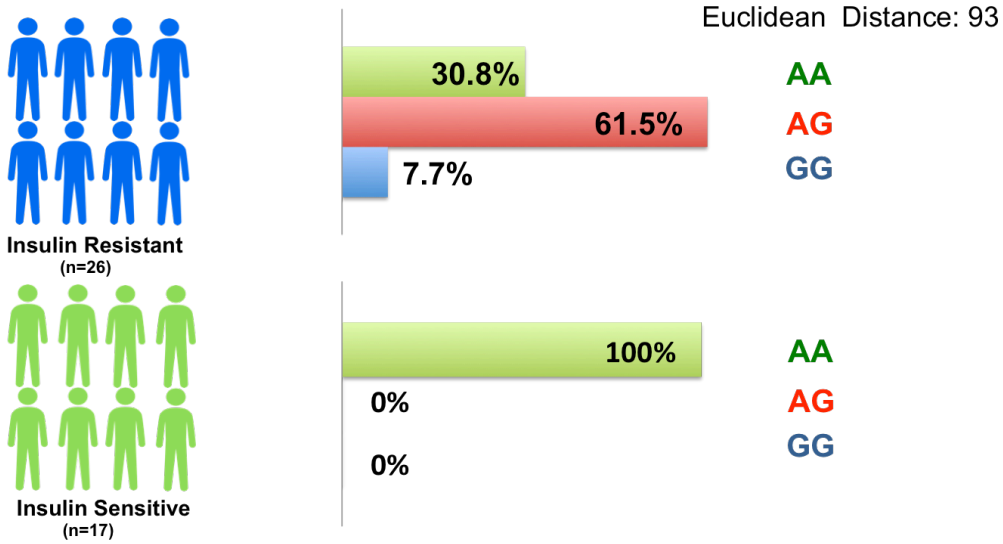


Figure 3.Histogram of SNPs with Euclidean distance higher than 10



Chr	Gene	position	IR			IS		
Chr1	HIVEP3	42365326	TT	TA	AA	TT	TA	AA
			0%	13.1%	86.9%	17.7%	70.6%	11.8%

Figure 4. Differential SNPs between Insulin Resistant and Insulin Sensitive 1



Chr	Gene	position	IR			IS		
Chr3	LSAMP	115618796	AA	AG	GG	AA	AG	GG
			30.8%	61.5%	7.7%	100%	0%	0%

Figure 5. Differential SNPs between Insulin Resistant and Insulin Sensitive 2

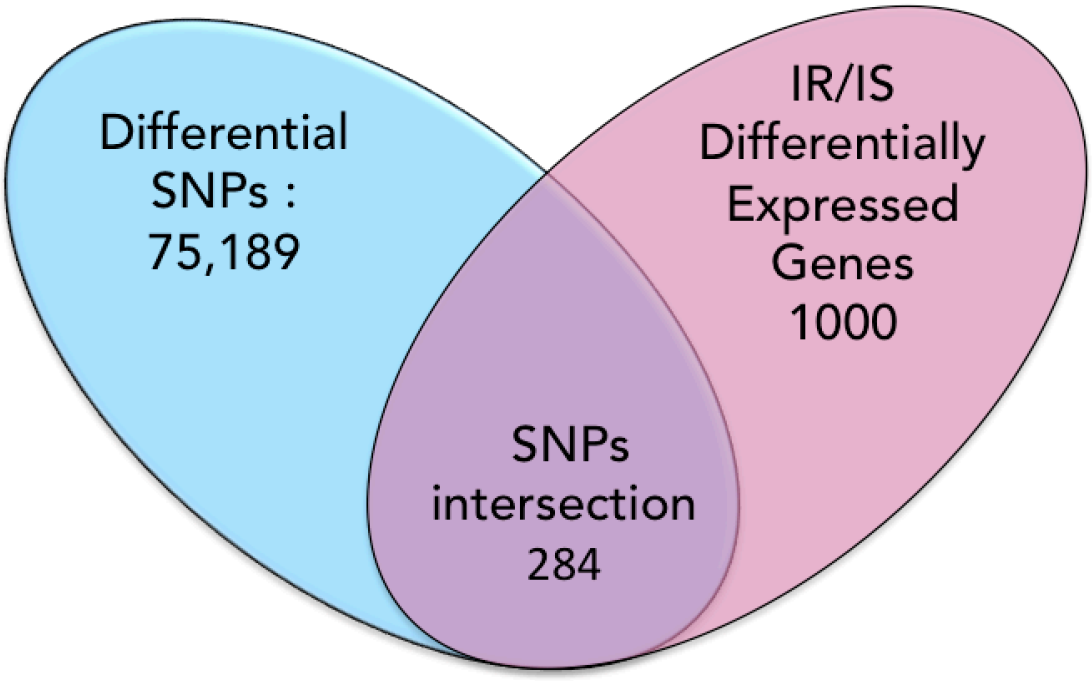


Figure 6. Venn Diagram of intersection of differential SNPs and differentially expressed genes

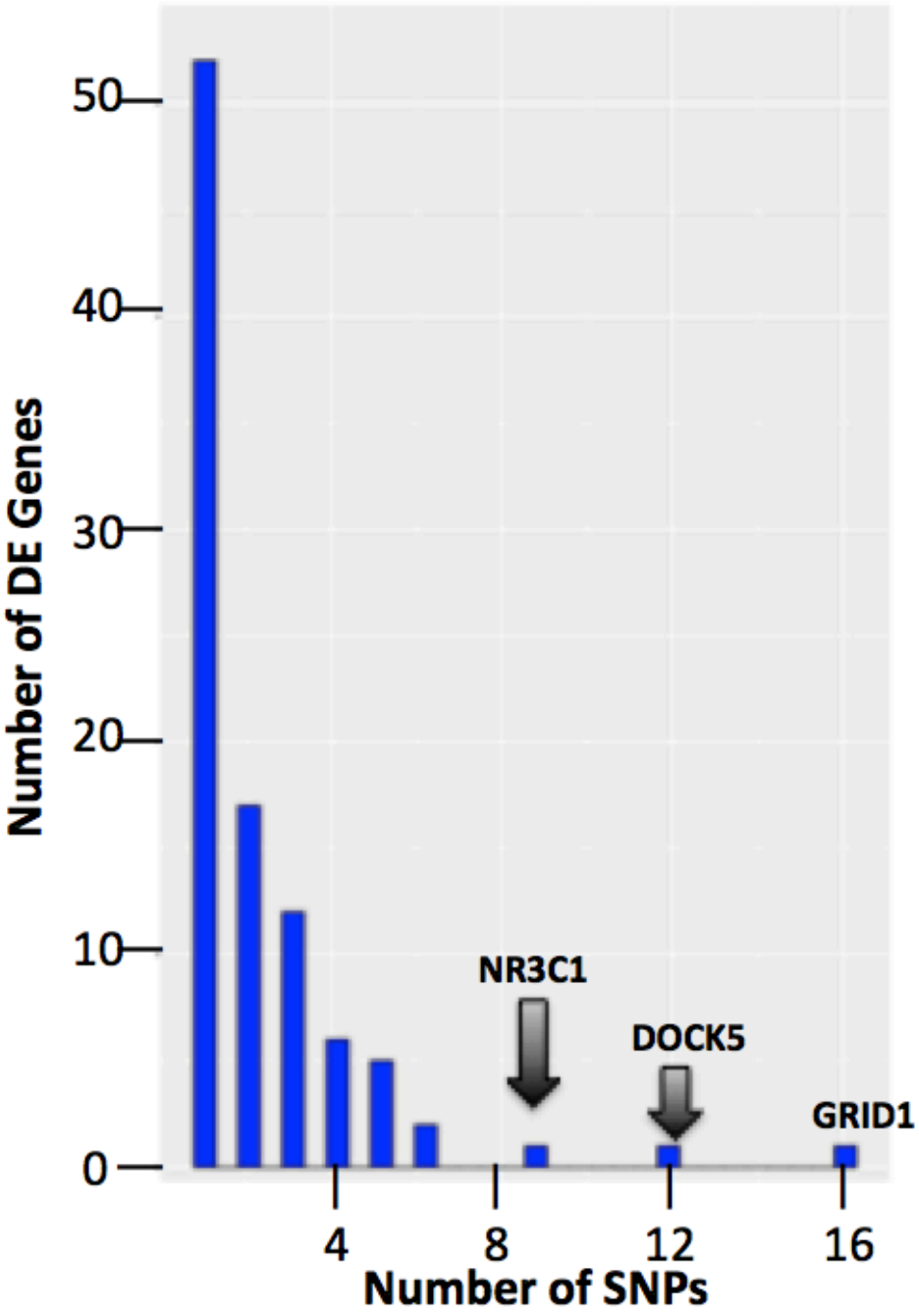


Figure 7. Histogram of the number of differential SNPs in differentially expressed genes

Reference

1. "Statistics About Diabetes". *American Diabetes Association*. N.p., 2016. Web. 9 Sept. 2016.
2. "Diabetes: Facts And Figures". *International Diabetes Federation*. N.p., 2016. Web. 9 Sept. 2016.
3. "Types Of Diabetes | NIDDK". *National Institute of Diabetes and Digestive and Kidney Diseases*. N.p., 2016. Web. 9 Sept. 2016.
4. "Diabetes: Differences Between Type 1 And 2-Topic Overview". *WebMD*. N.p., 2016. Web. 9 Sept. 2016.
5. Ling C, Groop L. Epigenetics: A molecular link between environmental factors and type 2 diabetes. Diabetes. 2009;58(12):2718–25.
6. "Picard Tools - By Broad Institute". *Broadinstitute.github.io*. N.p., 2016. Web. 10 Sept. 2016.
7. McKenna, A. et al. "The Genome Analysis Toolkit: A Mapreduce Framework For Analyzing Next-Generation DNA Sequencing Data". *Genome Research* 20.9 (2010): 1297-1303.
8. "Endrebak/Snpflip". *GitHub*. N.p., 2016. Web. 10 Sept. 2016.
9. Purcell, Shaun et al. "PLINK: A Tool Set For Whole-Genome Association And Population-Based Linkage Analyses". *The American Journal of Human Genetics* 81.3 (2007): 559-575. Web.
10. Delaneau, Olivier et al. "Haplotype Estimation Using Sequencing Reads". *The American Journal of Human Genetics* 93.4 (2013): 687-696.

11. Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. "A Flexible And Accurate Genotype Imputation Method For The Next Generation Of Genome-Wide Association Studies". *PLoS Genetics* 5.6 (2009): e1000529.
12. "Snptest". *Mathgen.stats.ox.ac.uk*. N.p., 2016. Web. 11 Sept. 2016.