Creation of a Computational Pipeline to Extract Genes from Quantitative Trait Loci for Diabetes and Obesity

Joseph Fox
Marquette University

CREATION OF A COMPUTATIONAL PIPELINE TO EXTRACT GENES FROM QUANTITATIVE TRAIT LOCI FOR DIABETES AND OBESITY

by

Joseph F. Fox

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Bioinformatics

Milwaukee, Wisconsin

May 2015

ABSTRACT
CREATION OF A COMPUTATIONAL PIPELINE TO EXTRACT GENES FROM QUANTITATIVE TRAIT LOCI FOR DIABETES AND OBESITY

Joseph F. Fox

Marquette University, 2015

Type 2 Diabetes is a disease of relative insulin deficiency resulting from a combination of insulin resistance and decreased beta-cell function. Over the past several years, over 60 genes have been identified for Type 2 Diabetes in human genome-wide association studies (GWAS). It is important to understand the genetics involved with Type 2 diabetes in order to improve treatment and understand underlying molecular mechanisms. Heterogeneous stock (HS) rats are derived from 8 inbred founder strains and are powerful tools for genetic studies because they provide a basis for high resolution mapping of quantitative trait loci (QTL) in a relatively short time period. By measuring diabetic traits in 1090 HS male rats and genotyping 10K single nucleotide polymorphisms (SNPs) within these rats, Dr. Solberg Woods' lab conducted genetic analysis to identify 85 QTL for diabetes and adiposity traits.

To identify candidate genes within these QTL, we propose creation of a bioinformatics pipeline that combines general gene information, information from the rat genome database including disease portals and Variant Visualizer as well as the Attie Diabetes Expression Database. My project has involved writing code to pull data from these databases to determine which genes within each QTL are potential candidate genes. I have scripted the code to analyze genes within a single QTL or multiple QTL simultaneously. The resulting output is a single excel file for each QTL, listing all genes that are found in the disease portals, all genes that have a highly conserved non-synonymous variant change and all genes that are differentially expressed in the Attie database. The program also highlights genes that are found in all three categories. After creating the pipeline, I ran the program for 85 QTL identified in my laboratory. The program identified 63 high priority candidate genes for future follow-up. This work has helped my laboratory rapidly identify candidate genes for type 2 diabetes and obesity. In the future, the code can be modified to identify candidate genes within QTL for any complex trait.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Joseph F. Fox

I would like to thank Dr. Leah Solberg Woods, Jack Littrell, Katie Holl, Dr. Hong He, Dr. Mary Shimmoya and Jeff De Pons for all of the help and guidance they provided me with as I worked towards completing my thesis work. I would also like to thank my family and girlfriend for all of their encouragement. Finally, I would like to thank the Graduate Schools of Marquette University and the Medical College of Wisconsin for making this opportunity possible.

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Type 2 diabetes and obesity are serious illnesses that are becoming more prevalent within our country. Type 2 Diabetes is a disease of relative insulin deficiency resulting from a combination of insulin resistance and decreased beta-cell function [17,19]. More than 25 million American adults have already been diagnosed with diabetes, while the CDC projects that as many as one in three U.S. adults could have diabetes by 2050 [4,5]. In addition, more than 80 percent of people who have diabetes are also overweight [7]. The American Diabetes Association released new research on March 6, 2013 estimating the total costs of diagnosed diabetes have risen to $245 billion in 2012 from $174 billion in 2007 [1]. Due to the substantial financial burden that diabetes and obesity impose on our country, new treatments and preventative measures are needed to help treat this increasing problem.

A person's genetic background has been shown to play a role in developing Type 2 diabetes and obesity, since studies of twins have shown that genetics play a very strong role in the development of type 2 diabetes [2]. Over the past several years, over 60 genes have been identified for Type 2 Diabetes in human genome-wide association studies (GWAS). [12,21,26]. Understanding the genes involved in these diseases will help reveal one's predisposition for developing either of these diseases and aid in developing novel therapeutics. The genomic differences between individuals are made up of in large part by changes in single nucleotide polymorphisms (SNPs) [3,13]. As a result, a common disease can develop from a combination of common SNP variants, which are responsible for quantitative variations of a common phenotypic trait [3,10,20]. There is a desire to

understand which SNP variants underlie a predisposition for the development of Type 2 diabetes and obesity in order to improve future prevention and treatment.

Genetic mapping is an effective tool for identifying quantitative trait loci (QTL) since genetic loci that contribute to specific traits can be measured and mapped to the genome. There are two approaches for carrying out genetic mapping in rats: the traditional F2 cross and the outbred Heterogeneous Stock (HS) approach. The traditional F2 cross consists of breeding two distinct strains of rats which give rise to offspring (F1). The F1 animals are then bred, leading to a limited number of recombination events. From here the desired traits can be mapped to regions on the genome. However, these regions are large and may contain hundreds of genes making it difficult to determine exactly which gene is truly responsible for the resulting phenotypic trait. The HS approach allows for narrower fine mapping of traits (only 2-3 Mb) in comparison to the typical F2 cross (generally 30-40 Mb). This approach works by using HS animals (mice or rats), which are created from 8 inbred founder strains that are bred for more than 50 generations to minimize inbreeding. As a result, the chromosomes of these offspring are a random mosaic of the founder strains, and the probability of decent from each founder can be traced, thus generating a narrower region for the trait of interest along with a smaller number of candidate genes [23].

Using HS rats, Dr. Solberg Woods' lab has conducted genome-wide association analysis to identify QTLs linked to Type 2 diabetes and obesity. They carried out their genetic mapping by measuring diabetic and adiposity traits in 1090 HS and genotyping these rats using a 10K single nucleotide polymorphisms (SNPs) array. Analysis was performed on these data using the programs Happy and Bagpipe. The Happy algorithm

uses hidden Markov models to estimate the expected proportions of founder haplotypes in each marker based on observed genotypes [16]. Association analysis is then implemented using Bagpipe, which assigns a level of association between the chromosomal location and the diabetic phenotype [24]. The results yield a collection of QTL for each chromosome where each QTL is assigned a respective -log P score.

The statistically significant QTL intervals Dr. Solberg Woods' lab focused on were those that generated a -log P score greater than or equal to a particular phenotype's significance threshold. This threshold was determine by parametric bootstrapping, which normally resulted in a -log P score lower than -5. The 1.5 LOD drop from the peak marker was then utilized to define the confidence intervals for each significant QTL [22]. Overall, this approach identified 85 QTL associated with diabetic and adiposity traits. The average size of each QTL is 2.2 Mb and contains on average 22 genes. Analyzing these genes more closely, our lab hopes to find candidate genes that play a significant role in the development of diabetes and obesity.

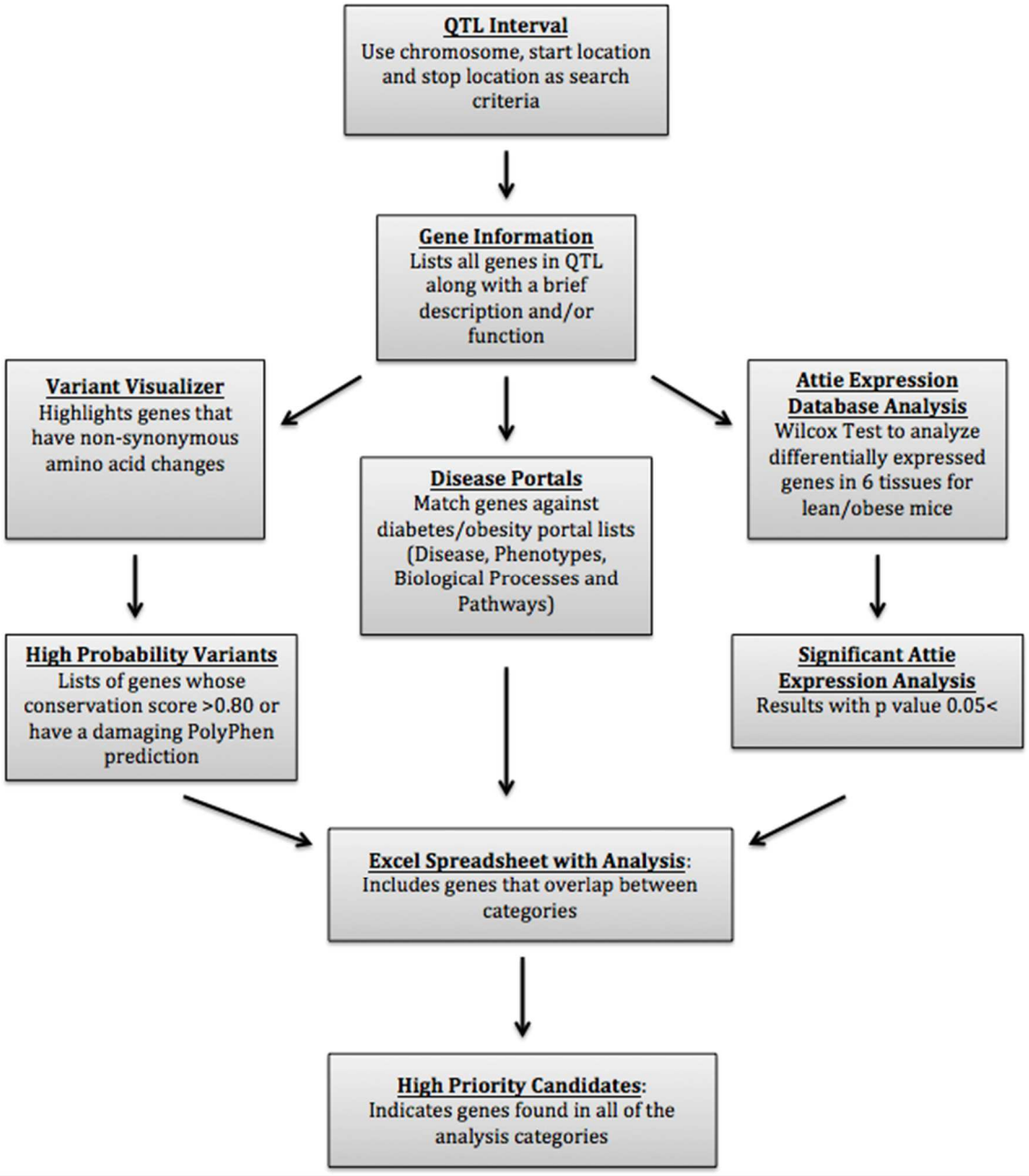
Despite having identified numerous statistically significant QTL intervals, the lab currently uses a manual process to identify potential candidate genes. The process consists of manually extracting important gene information for each QTL from two different websites: the Rat Genome Database (RGD) (http://rgd.mcw.edu/) and Attie Diabetes Expression Database websites (http://diabetes.wisc.edu/). The manual process is time-consuming and prone to error. The Diabetes/Obesity Disease Portals and the Variant Visualizer tool from the RGD website are used to determine if a gene has been previously shown to be associated with a particular disease or if a gene contains any non-synonymous variants that could be damaging. Information from the Attie Diabetes

Expression Database helps determine if any of those genes are differentially expressed in various tissues. Although possible to obtain manually, keeping track of all of this information for all of the genes of interest for each of the 85 QTL is a challenge due to the magnitude of the data. That is why a computational bioinformatics pipeline to automate this process was desired. Therefore, using the programming languages Perl and R, my present study was to create a computational bioinformatics pipeline to automate the processes listed above to identify potential candidate genes associated with the development of Type 2 diabetes and obesity.

METHODS

The programming languages Perl and R were used for the creation of this computational pipeline (for summary schematic, see Figure 1). The first part of the pipeline consists of a Perl script that uses the chromosome number and start/stop locations of a specific QTL as command line arguments to extract gene information found within that particular QTL. In addition, the user can also provide a text file that contains multiple QTLs to be analyzed. Once the QTL parameters are provided, the command line arguments and/or QTL text file are used to name and organize the results and analysis files that are generated by the Perl script. This script extracts general gene information from an available RGD rat gene file found on their FTP site as well as multiple disease portal gene lists which were manually copied from the RGD disease portal page. It then creates separate output files for the Disease Portals results, Variant Visualizer results, and Attie Diabetes Expression Database results and analysis using two Perl modules. The Perl module WWW::Mechanize was used to access and extract information from a website's URL when needed [8]. For example, the WWW::Mechanize module can be used to take a particular URL and obtain that URL page's resulting source code and content. From there, manual operations can be programmed to modify that page's source code and carry out a task, such as clicking the 'Download File' button. The Perl module WriteExcel::UseSpreadsheet was used to create one excel spreadsheet output with multiple sheets of analysis [15]. All of the results files that were originally created separately were incorporated into one excel analysis file with each file being represented by a separate sheet.

Figure 1: Schematic Summary of Computational Pipeline



In order to detect all known genes found within each QTL, the GENES_RAT.txt file generated by RGD was obtained via their FTP site and saved on our Dale server [9]. This file uses rat genome build version 3.4 and contains information for all of the active

genes known in the rat genome. As stated before, the chromosome number and start/stop locations of the QTL from the command line are used as filtering parameters to extract all of the genes found within that interval, along with obtaining a brief description for what each gene has been shown to encode and/or interact with. This gene information is saved under the "Gene Info" sheet in the excel analysis file.

To identify if a particular gene has been shown to be previously associated with diabetes or obesity, lists of genes that were found on RGD's disease portals web page were manually copied and saved into separate files on our Dale server for use (Diabetes_Biological_Processes.txt, Diabetes_Disease.txt, Diabetes_Pathways.txt, Diabetes_Phenotypes.txt, Obesity_Biological_Processes.txt, Obesity_Disease.txt, Obesity_Pathways.txt, and Obesity_Phenotypes.txt). These files were used within the Perl script to identify if any of the genes within a particular QTL have been shown to be associated with the diseases of diabetes or obesity. For each disease, these files were comprised of gene names that have been shown to be associated with the General Disease, Phenotype, Biological Processes, or Pathways related to that disease category. The genes identified from the GENES_RAT.txt file were used to determine if there was a match against any of these gene lists. This was done in Perl using hash keys and a foreach statement. Genes that found a match were printed out in the excel analysis file under the sheet "Portal Info" in addition to indicating which gene list there was a match with. Otherwise, 'no genes' was printed if there were no matches found within that list.

To utilize the Variant Visualizer tool from RGD's website, the Perl module WWW::Mechanize was used. For our analysis, we were interested in downloading a file that contains non-synonymous sequence polymorphism changes in HS founder strains for

genes found within a particular QTL. The HS rat strains are ACI/N, BN/Ssn, Buf/N, F344/N, M520/N, MR/N, WKY/N, and WN/N. Therefore, the names of these strains along with just non-synonymous polymorphism changes were selected on RGD's web page so that they were included within that particular URL so that the script downloaded a file with only this information. The URL was then modified within the Perl script so that the chromosome number and start/stop location command line arguments could be entered and applied to the URL in order to obtain that desired QTL's information. A particular QTL's Variant Visualizer results were then saved as a separate file as well as printed out in the "Variant Visualizer Results" sheet in the excel analysis file. In addition, a "High Probability Variants" sheet was created where genes had a conservation score greater than 0.8 or their PolyPhen prediction was either 'possibly damaging' or 'probably damaging'.

The results from the Attie Diabetes Expression Database website were obtained in a similar manner by utilizing the Perl module WWW::Mechanize. However, instead of manipulating the website's URL to account for the QTL's chromosome number and start/stop locations, a manual operation was performed to just enter in all of the genes found in a particular QTL into the page's source code as a Perl array. This array was obtained from an earlier process involving the extraction of gene information from the GENES_RAT.txt file. From here, code in the Perl script selected only intensity2 data (intensity of the transcript on the expression array) from an expression analysis done by the Attie lab and then clicked the "Download File" button. The resulting file has gene expression analysis for 6 key tissues that are evaluated in mice that differ in body weight (lean or obese), strain (B6 or BTBR mouse strains) and age (4 or 10 weeks). The BRTB

strain is susceptible and the B6 mouse is resistant to Type 2 diabetes. The six tissues are Islet, Adipose, Liver, Soleus, Gastrocnemius, and Hypothalamus. The results contain separate intensity2 data for B6 and BTBR tissue expression at 4 and 10-week time intervals. At each of these time intervals are data points from five individual animals. The results are saved in a separate file.

To determine if there are genes within each QTL that are differentially expressed between strains, the second part of the pipeline is made up of an R script to perform statistical testing on expression data from the Attie Diabetes Expression file described above. This R script is called from within the Perl script using the system() function. We used the Wilcox Test, which is a non-parametric statistical hypothesis test for the comparison of the means between 2-paired samples [14]. In our case, the 2-paired samples are between lean and obese mice for B6 and BTBR tissue expression at the 4 and 10-week time intervals. The script's analysis provides the means for each of the 5 data points for B6 and BTBR tissue expression at the 4 and 10-week time intervals for both lean and obese mice as well as the p value for the hypothesis test comparison. In addition, means and p values were provided for combined B6 and BTBR tissue expression at both of the 4 and 10-week time periods. This analysis is saved as a separate file and is also incorporated in the "Attie Analysis" sheet in the excel analysis file. Another sheet named "Significant Attie Analysis" was created which includes genes that have a p value lower than 0.05 for any of the B6 and BTBR tissue expression at 4 and 10-week time intervals.

Finally, a comparison for all the genes found within each of the results was performed to highlight which genes are potential high priority candidate genes associated

with Type 2 diabetes and obesity. These comparisons were performed by comparing hash keys corresponding to gene names found within each of the respective result categories. Multiple combinations of comparisons were done to indicate which genes are more noteworthy.

RESULTS

The present study yielded a computational pipeline, comprised of Perl and R written scripts that provided informational output regarding disease portals, non-synonymous variants and expression analysis for all genes within each QTL. This resulting excel file contains multiple sheets that provide all of the desired information. Each analysis file is made up of 7 excel sheets, which are Gene Info, Portal Info, Variant Visualizer Results, High Probability Variants (described below), Attie Analysis, Significant Attie Analysis (described below) and Gene Comparisons.

The Gene Info sheet provides information for all of the genes found within a particular QTL of interest. This information consists of the RGD gene ID number, gene symbol, chromosome, start location, stop location and a general description (see Figure 2). The general description is particularly helpful because it describes what the gene encodes, interacts with, is associated with, is involved in and/or what it participates in.

The Portal Info sheet indicates which genes in a certain QTL have been shown to be associated with diabetes or obesity and/or related pathways and the disease process. It does this by listing which genes are associated with the general disease, phenotype, biological processes or pathways for that specific disease (see Figure 3). If no genes have been shown to be associated with that disease, then ‘no genes’ will be listed instead.

The Variant Visualizer sheet lists all of the genes in the QTL that have non-synonymous amino acid changes in one or more of the HS founder strains. In addition, this sheet also provides the chromosome, position, conservation score, gene symbol, reference nucleotide, founder strain nucleotides, accession ID, reference amino acid,

variant amino acid and PolyPhen prediction for each variant (see Figure 4). The conservation score helps determine how much a particular gene is conserved, with the scores ranging from 0-1 and 1 being highly conserved. The reference nucleotide and 8-founder strain nucleotide information indicate exactly which founders have the variants. Finally, the PolyPhen prediction helps determine which of those variants could be potentially damaging.

The High Probability Variants sheet provides the genes that more likely to be candidate genes. The genes listed on this sheet have either a conservation score greater than 0.8 or their PolyPhen prediction is possibly/probably damaging (see Figure 5).

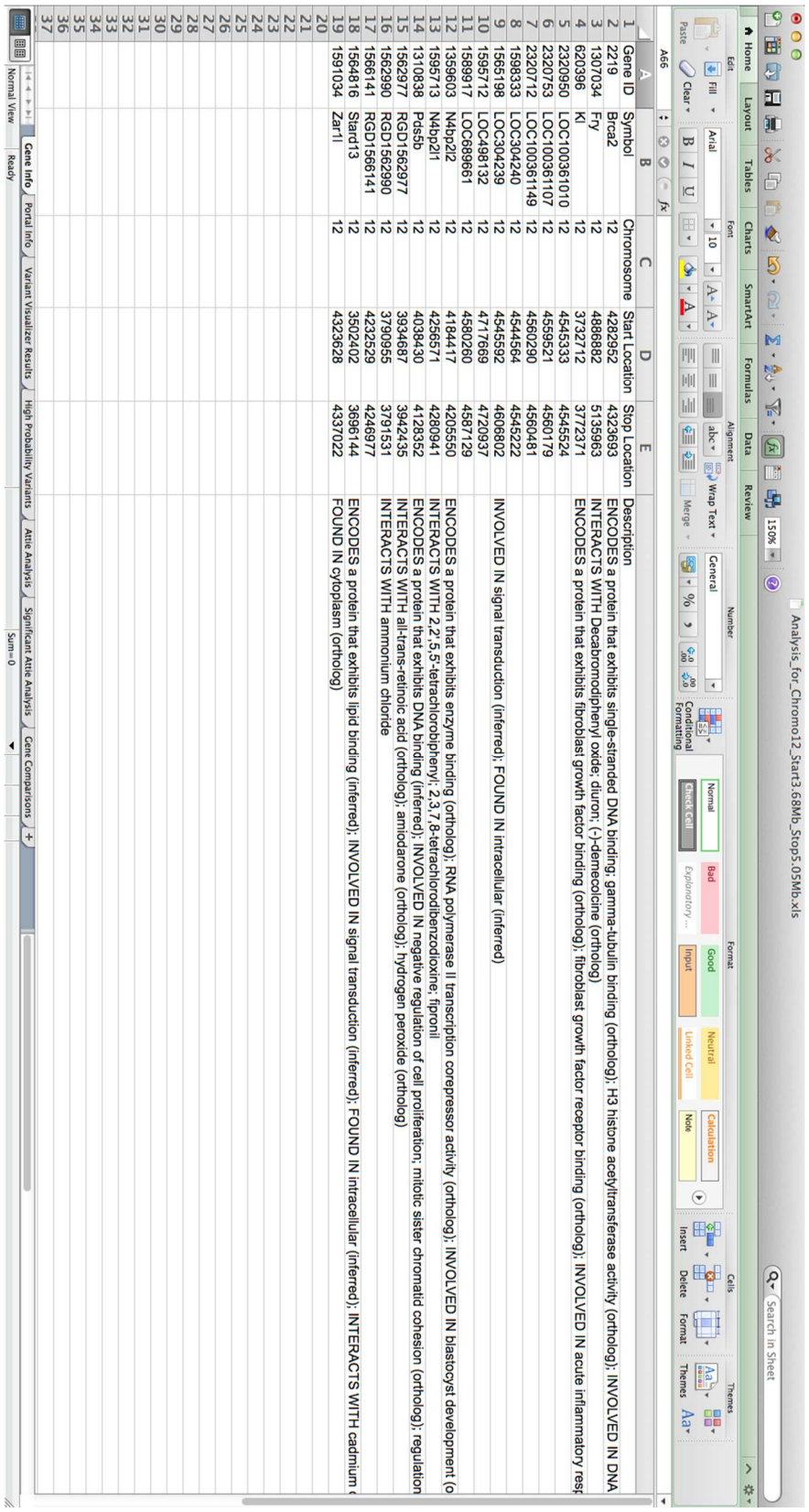
The Attie Analysis sheet provides a list of genes within a QTL and their Wilcox Test results for tissue expression in B6 and BTBR tissue at 4 and 10 week time periods in lean and obese mice as well as combined B6 and BTBR tissue at the 4 and 10 week time periods. For each gene, there is an average score for lean and obese mice at each tissue and time period as well as a p value indicating the significance of the expression differences. In addition, each gene is further categorized by tissue as well as its particular start location in case multiple locations were used (see Figure 6).

The Significant Attie Analysis is similar to the High Probability Variants sheet in that it provides quick look into the noteworthy differential expression results from the Attie Analysis sheet. This sheet provides a list of genes that have a p value lower than 0.05 for lean and obese mice expression in the B6 and BTBR tissues at the 4 and 10 week time periods. Underneath each tissue and time period combination lists the genes and specific tissues that are significant (see Figure 7).

The Gene Comparisons sheet indicates which genes appear in more than one of the analysis sheets. The lists of genes from each of the analysis sheets are compared to one another to determine which genes have multiple lines of evidence suggesting they are involved in the phenotype. The comparisons of these genes lists are between the High Probability Variants, Portal Info and Attie Significant Analysis sheets (see Figure 8). The genes found in all three outputs are genes that our lab finds particularly interesting for future research.

This computational pipeline was used to analyze the 85 QTL that our lab was interested in. The program identified 63 high priority candidate genes for future follow-up. Out of the 63 genes identified, 44 were associated with adiposity phenotypes (see Table 1) and 19 with diabetic phenotypes (see Table 2). This program has made it possible to rapidly identify genes associated with diabetes and obesity.

Figure 2: Gene Info Sheet



Gene ID	Symbol	Chromosome	Start Location	Stop Location	Description
2219	Brca2	12	4282952	4323693	ENCODES a protein that exhibits single-stranded DNA binding; gamma-tubulin binding (ortholog); H3 histone acetyltransferase activity (ortholog); INVOLVED IN DNA
1307034	Fry	12	4886882	5135963	INTERACTS WITH Decabromodiphenyl oxide; diuron; (-)-demecolcine (ortholog)
620396	Kl	12	3732712	3772371	ENCODES a protein that exhibits fibroblast growth factor binding (ortholog); fibroblast growth factor receptor binding (ortholog); INVOLVED IN acute inflammatory resp
2320950	LOC100361010	12	4545333	4545524	
2320753	LOC100361107	12	4559521	4560179	
2320712	LOC100361149	12	4560290	4560481	
1598333	LOC304240	12	4544564	4545222	
1565198	LOC304239	12	4545592	4606802	
1595712	LOC498132	12	4717669	4720937	
1589917	LOC689661	12	4580260	4587129	
1359603	N4bp2l2	12	4184417	4205550	ENCODES a protein that exhibits enzyme binding (ortholog); RNA polymerase II transcription corepressor activity (ortholog); INVOLVED IN blastocyst development (o
1595713	N4bp2l1	12	4256571	4280941	INTERACTS WITH 2,2',5,5'-tetrachlorobiphenyl; 2,3,7,8-tetrachlorodibenzodioxine; fipronil
1310838	Ptds5b	12	4038430	4128352	ENCODES a protein that exhibits DNA binding (inferred); INVOLVED IN negative regulation of cell proliferation; mitotic sister chromatid cohesion (ortholog); regulation
1562977	RGD1562977	12	3934687	3942435	INTERACTS WITH all-trans-retinoic acid (ortholog); amiodarone (ortholog); hydrogen peroxide (ortholog)
1562990	RGD1562990	12	3790955	3791531	INTERACTS WITH ammonium chloride
1566141	RGD1566141	12	4232529	4246977	
1564816	Stard13	12	3502402	3696144	ENCODES a protein that exhibits lipid binding (inferred); INVOLVED IN signal transduction (inferred); FOUND IN intracellular (inferred); INTERACTS WITH cadmium c
1591034	Zar1l	12	4323628	4337022	FOUND IN cytoplasm (ortholog)

INVOLVED IN signal transduction (inferred); FOUND IN intracellular (inferred)

Figure 3: Portal Info Sheet

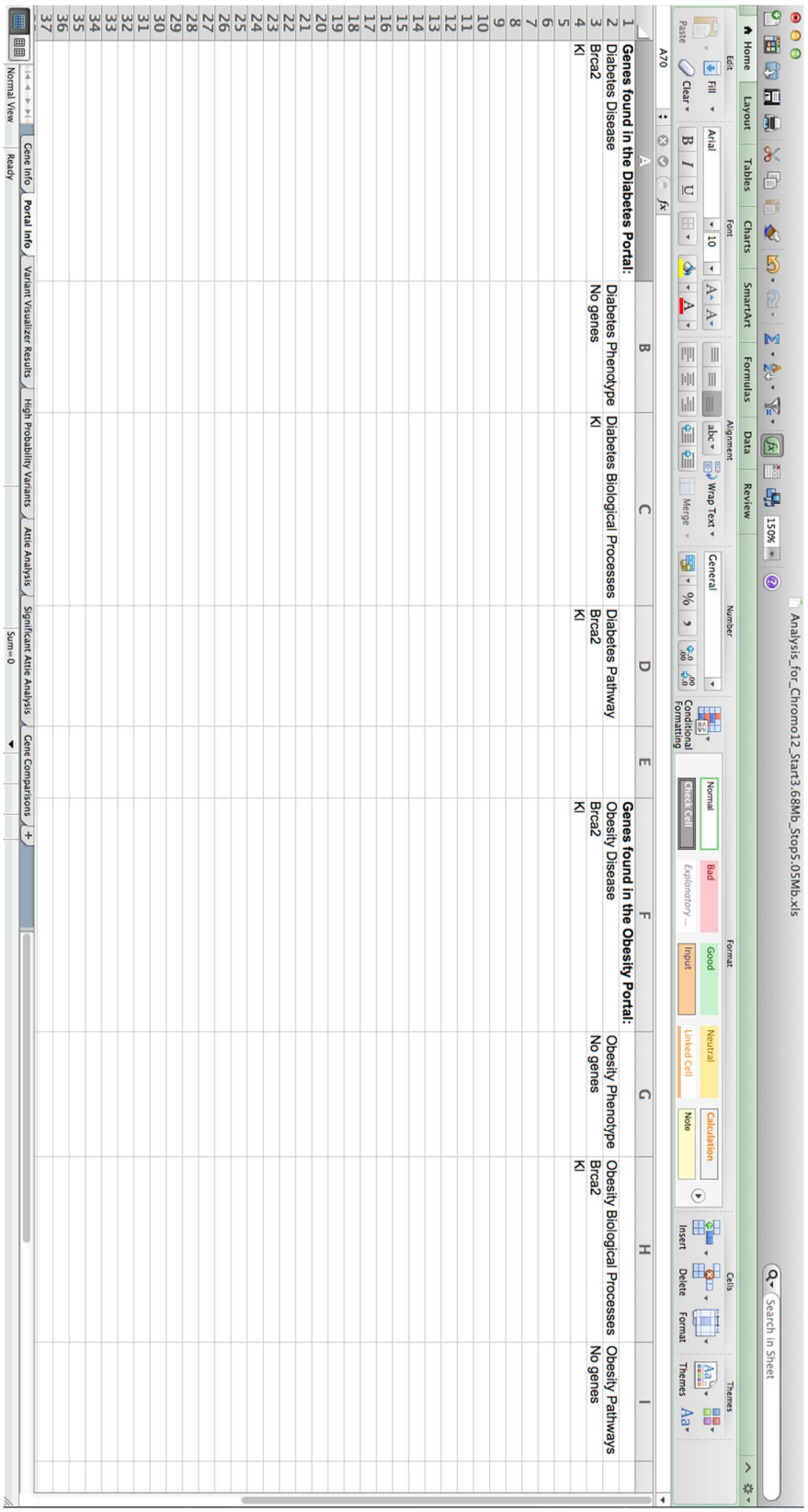
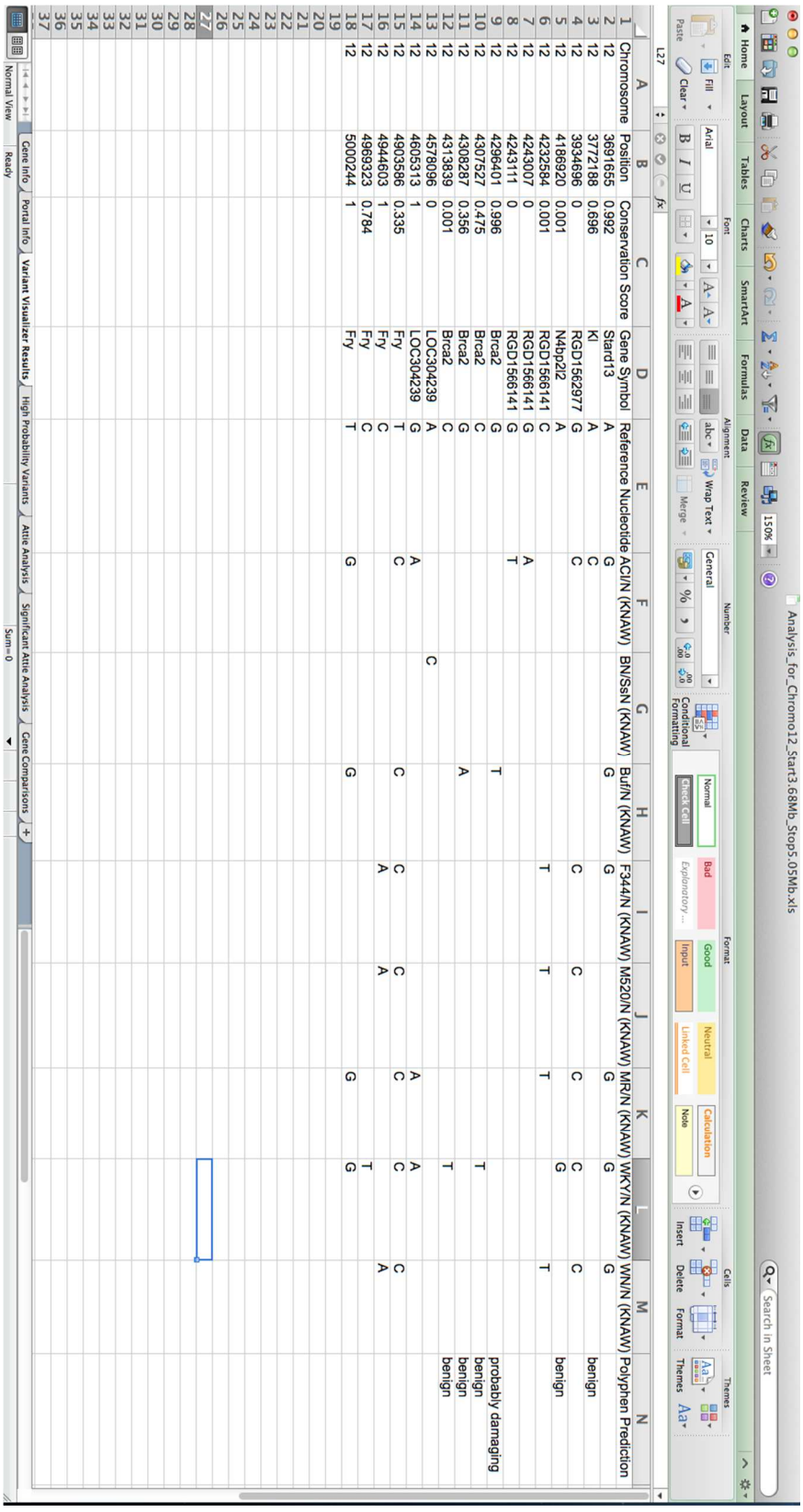


Figure 4: Variant Visualizer Results Sheet



Chromosome	Position	Conservation Score	Gene Symbol	Reference Nucleotide	AC/N (KNAW)	BN/SsN (KNAW)	Buf/N (KNAW)	F344/N (KNAW)	M520/N (KNAW)	MR/N (KNAW)	WKY/N (KNAW)	WN/N (KNAW)	Polyphen Prediction
12	3691655	0.992	Stard13	A	G		G	G	G	G	G	G	benign
12	3772188	0.696	Kl	C				C	C	C	C	C	
12	3934696	0	RGD1562977	G	C			C	C	C	C	C	benign
12	4186920	0.001	N4bp2l2	A									
12	4232584	0.001	RGD1566141	C				T	T	T	T	T	benign
12	4243007	0	RGD1566141	G	A								
12	4243111	0	RGD1566141	G	T								
12	4296401	0.996	Brca2	G			T						probably damaging
12	4307527	0.475	Brca2	C							T		benign
12	4308287	0.356	Brca2	G			A						benign
12	4313839	0.001	Brca2	C							T		benign
12	4578096	0	LOC304239	A		C							
12	4605313	1	LOC304239	G	A					A	A		
12	4903586	0.335	Fry	T	C		C	C	C	C	C	C	
12	4944603	1	Fry	C				A	A			A	
12	4969323	0.784	Fry	C							T		
12	5000244	1	Fry	T	G		G			G	G		

Figure 5: High Probability Variants Sheet

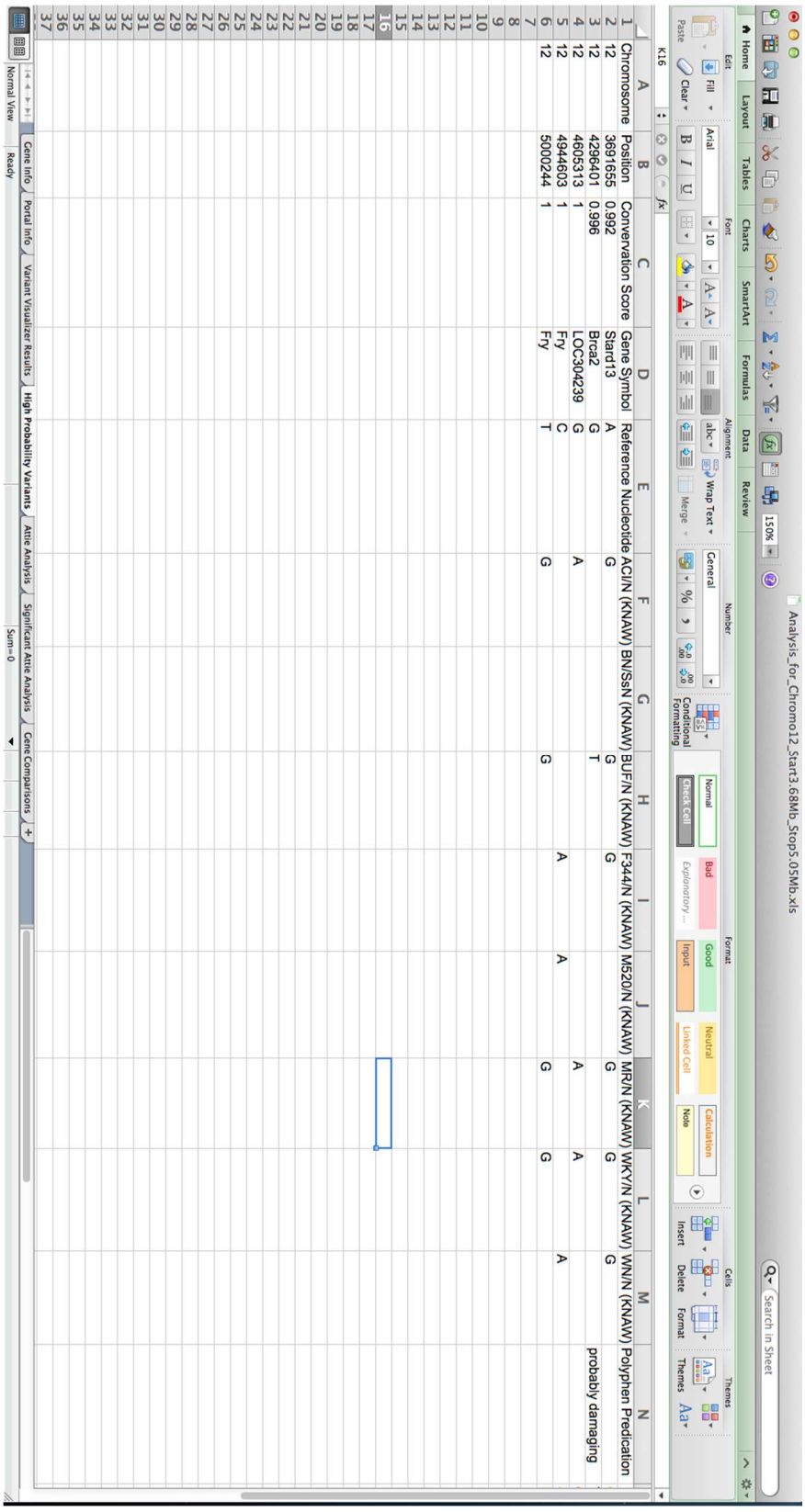


Figure 6: Attie Analysis Sheet

Figure 7: Significant Attie Analysis Sheet

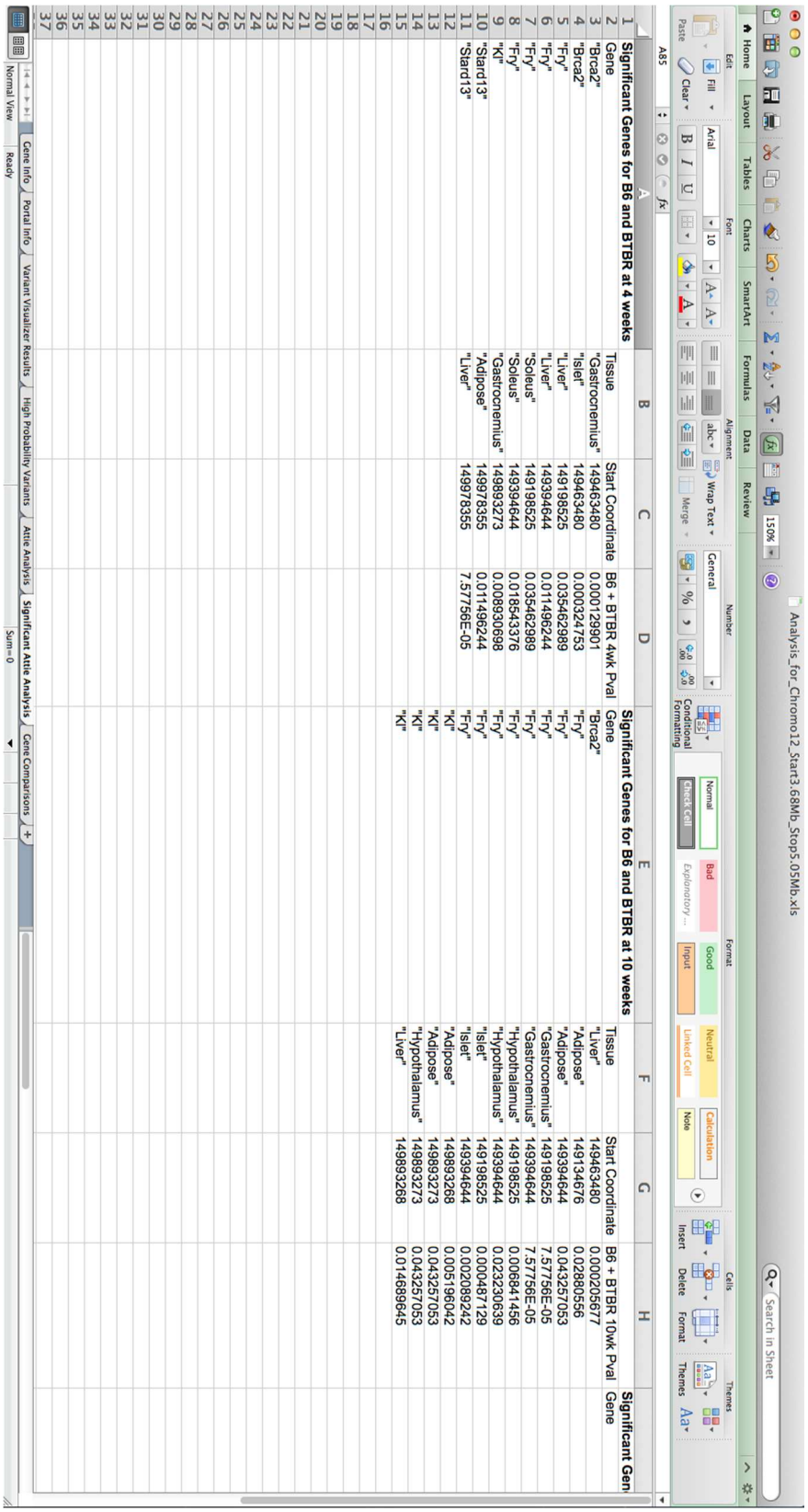


Figure 8: Gene Comparisons Sheet

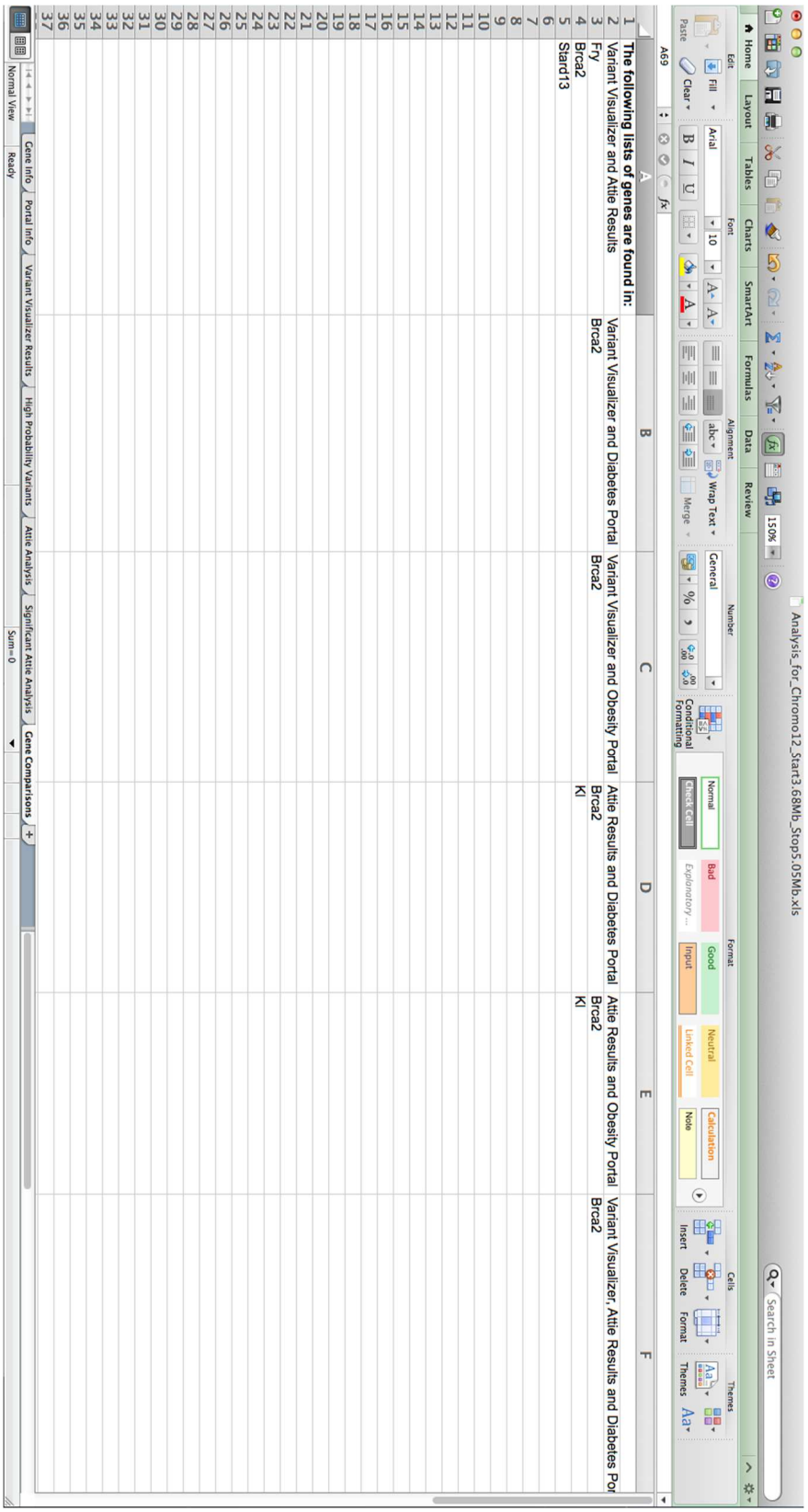


Table 1: Results for 66 Adiposity QTL Confidence Intervals

Phenotypes	Chromosome	Start Location (Mb)	Stop Location (Mb)	Number of Genes in QTL	Number of Significant Genes	Variant Visualizer, Attie Results and Diabetes Portal	Variant Visualizer, Attie Results and Obesity Portal
BMI_Butt	13	42.47	44.23	26	0		
BMI_Butt/BMI_Tail	17	5.7	10.94	54	0		
BMI_Tail	1	262.9	266.46	31	1	Pnlip	Pnlip
BMI_Tail	2	62.23	70.54	23	0		
BMI_Tail	3	37.82	40.32	14	0		
BMI_Tail	6	14.64	19.11	8	1	Nrxn1	Nrxn1
BMI_Tail	6	39.8	44.78	43	2	Ywhaq, Odc1	
BMI_Tail	6	111.68	113.91	19	1	Snw1	Snw1
BMI_Tail	7	36.44	40.32	21	0		
BMI_Tail	7	112.74	113.87	39	0		
BMI_Tail	8	123.18	123.91	3	1	Itga9	
BMI_Tail	10	28.72	29.23	7	1	Adra1b	Adra1b
BMI_Tail	10	50.07	51.2	5	0		
BMI_Tail	10	83.78	84.76	22	0		
BMI_Tail	13	72.22	73.11	8	0		
BMI_Tail	15	98.81	105.6	31	1	Abcc4	
BMI_Tail	17	40.12	44.15	39	1	Uqcrfs1	
BMI_Tail	18	66.33	69.91	17	2	Mbd2, Tcf4	Mbd2, Tcf4
BMI_Tail	20	34.43	35.42	5	0		
BMI_Tail	X	53.04	55.93	24	1	Pdha1	Pdha1
BW_Gained	2	75.27	76.16	1	0		
Cholesterol	3	117.89	119.99	49	2	Cdc25b, Ptpra	
Cholesterol	5	114.48	115.25	7	0		
Cholesterol	7	102.94	106.54	21	1	Ndrg1	Ndrg1
Cholesterol	13	54.1	59.46	21	0		
Cholesterol	14	105.28	106.71	3	0		
Cholesterol	19	41.54	42.46	18	1	Bcar1	
Cholesterol	20	40.69	44.22	30	1	Hs3st5	
Cholesterol	X	38.04	40.25	26	0		
EpiFatg_to_sacBWg	2	235.7	236.32	16	3	Adh7, Adh4	Adh7, Adh4, Adh1
EpiFatg_to_sacBWg	3	31.02	32.55	9	0		
EpiFatg_to_sacBWg	3	117.1	117.29	4	0		
EpiFatg_to_sacBWg	5	38.5	39.58	5	0		
EpiFatg_to_sacBWg	5	169.5	172.1	39	2	Pank4, Prkcz	Prkcz
EpiFatg_to_sacBWg	6	22.44	24.41	17	0		

EpiFatg_to_sacBWg	7	69.06	70.71	15	1	Pop1	
EpiFatg_to_sacBWg	7	103.63	104.64	10	1	Ndrg1	Ndrg1
EpiFatg_to_sacBWg	8	83.06	84.1	15	0		
EpiFatg_to_sacBWg	8	97.78	98.31	1	0		
EpiFatg_to_sacBWg	8	116.03	117.6	6	0		
EpiFatg_to_sacBWg	10	39.55	40.81	24	1	Gm2a	Gm2a
EpiFatg_to_sacBWg	10	52.94	53.67	8	0		
EpiFatg_to_sacBWg	10	102.64	103.62	12	1	Cog1	Cog1
EpiFatg_to_sacBWg	11	86.89	87.08	3	1	Prkdc	Prkdc
EpiFatg_to_sacBWg	13	86.77	88.37	64	4	Atp1a2, Usp21, Ndufs2, B4galt3	Atp1a2, Usp21, Ndufs2, B4galt3
EpiFatg_to_sacBWg	15	0.67	1.72	2	1		Kcnma1
EpiFatg_to_sacBWg	15	32.77	38.27	151	2	Myh6, Pck2	Myh6, Pck2
EpiFatg_to_sacBWg	15	98.28	99.49	3	0		
EpiFatg_to_sacBWg	19	25.44	27.07	46	1	Gipc1	
EpiFatg_to_sacBWg	20	49.14	50.34	7	0		
EpiFatg_to_sacBWg	X	37.89	40.25	27	0		
RetroFatg_to_sacBWg	3	30.76	31.68	11	0		
RetroFatg_to_sacBWg	4	18.92	20.42	3	0		
RetroFatg_to_sacBWg	6	99.61	101.1	10	2	Gphn, Fut8	Gphn, Fut8
RetroFatg_to_sacBWg	8	84.49	85.8	12	1	Col12a1	
RetroFatg_to_sacBWg	10	7.5	8.82	2	0		
RetroFatg_to_sacBWg	13	11.93	18.86	42	1	Serpinb7	Serpinb7
RetroFatg_to_sacBWg	13	30.33	34.34	35	0		
RetroFatg_to_sacBWg	17	83.5	84.06	3	0		
RetroFatg_to_sacBWg	19	23.43	25.8	84	1	Cacna1a	Cacna1a
RetroFatg_to_sacBWg	19	38.64	39.64	11	1	Hp	Hp
SacBWg	12	3.68	5.05	18	1	Brca2	Brca2
SacBWg	14	22.59	23.47	16	0		
SacBWg	15	104.44	108.42	40	0		
SacBWg	19	47.73	50.2	20	2	Plcg2, Mlycd	Plcg2, Mlycd
Triglycerides	8	49.26	50.98	9	0		

Table 2: Results for 18 Diabetic Confidence Intervals

Phenotypes	Chromosome	Start Location (Mb)	Stop Location (Mb)	Number of Genes in QTL	Number of Significant Genes	Variant Visualizer, Attie Results and Diabetes Portal	Variant Visualizer, Attie Results and Obesity Portal
Gluc0	9	78.65	79.97	14	1	Mrpl44	Mrpl44
Gluc1	18	68.5	69.91	1	0		
GTotalAUC	1	209.36	213.37	136	3	Mta2, Cox8a, Bad	Bad
GTotalAUC	10	46.94	53.41	70	2	Ncor1, Akap10	Ncor1
GTotalAUC	19	47.73	49.52	4	1	Plcg2	Plcg2
IGI15	3	67.17	68.02	16	0		
IGI15	7	115.58	117.12	42	0		
IGI15	13	57.75	59.46	7	0		
IGI15	18	25.44	29.28	50	2	Apc, Brd8	
Ins0	1	174.8	176.59	9	0		
Ins0	1	199.9	203.48	139	4	Dusp8, Paox, Ctsd	Cd81, Ctsd
Ins0	2	158.1	158.7	1	0		
Ins0	9	16.83	21.56	35	0		
Ins0	18	26.7	27.9	22	2	Apc, Brd8	
ITotalAUC	1	263.39	266	20	1	Pnlip	Pnlip
ITotalAUC	5	41.43	43.05	3	0		
ITotalAUC	6	22.44	24.41	17	0		
ITotalAUC	18	27.11	28.14	12	0		
QUICKI	1	174.8	176.4	7	0		
QUICKI	1	202.81	203.48	13	1		Cd81
QUICKI	2	158.1	158.7	1	0		
QUICKI	9	16.83	21.56	35	0		
QUICKI	18	26.7	28.14	22	2	Apc, Brd8	

DISCUSSION

Identifying candidate genes that underlie each of the 85 QTL will aid in identifying novel genes involved in Type 2 diabetes and obesity. Utilizing the present work's computational pipeline, our laboratory was able to rapidly identify 66 genes located within these QTL that we consider to be high priority candidate genes. We used criteria consisting of : 1) containing a highly conserved non-synonymous variant changes in one of the founder strains (Variant Visualizer), 2) have been shown to be associated with diabetes or obesity in prior research (RGD portal information) and 3) are differentially expressed in obese versus lean mice (Attie database). Having identified 66 candidate genes associated with adiposity and diabetic traits, future work in the lab will consist of determining which of these genes play a causal role in the phenotype, followed by understanding the exact underlying genetic mechanisms and roles these genes have in the development of diabetes and obesity.

The computational pipeline created for this present study was created with diabetes and obesity traits in mind. However, future work could also be implemented on the code to modify it so that it could identify potential candidate genes for any complex trait and not just diabetic or obesity traits. Future steps could also include other on-line information such as gene ontologies or expression and sequence information from other databases.

BIBLIOGRAPHY

[1] American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2012. Diabetes Care. March 2013.

[2] American Diabetes Association. Genetics of Diabetes. May 2014. URL: http://www.diabetes.org/diabetes-basics/genetics-of-diabetes.html

[3] Collins FS, Guyer MS, Chakravarti A: Variations on a theme: cataloging human DNA sequence variation. Science278 :1581 –1581,1997

[4] Centers for Disease Control and Prevention. National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States, 2011. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011.

[5] Centers for Disease Control and Prevention. Number of Americans with diabetes projected to double or triple by 2050. [Press Release] (accessed March 2011).

[6] Diabetes Disease Portal, Rat Genome Database Web Site, Medical College of Wisconsin, Milwaukee, Wisconsin. World Wide Web (URL: http://rgd.mcw.edu/). [January 2015].

[7] Do You Know the Health Risks of Being Overweight? In National Institutes of Diabetes and Digestive and Kidney Diseases (accessed April 2007).

[8] Etheridge K. WWW-Mechanize-1.74. January 2015. (URL: https://github.com/libwww-perl/WWW-Mechanize.git)

[9] GENES_RAT.txt, Rat Genome Database Web Site, Medical College of Wisconsin, Milwaukee, Wisconsin. World Wide Web (URL: http://rgd.mcw.edu/). [January 2015].

[10] Gura T: Can SNPs deliver on susceptibility genes? Science293 :593 –595,2001

[11] Keller MP, Choi Y, Wang P, Davis DB, Rabaglia M, Oler A, Stapleton D, Argmann C, Schueler K, Edwards S, Steinberg HA, Neto EC, Kleinhanz R, Turner S,

[12] Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang W, Frossard P, Been LF, Chia KS, Dimas AS, Hassanali N, Jafar T, Jowett JB, Li X, Radha V, Rees SD, Takeuchi F, Young R, Aung T, Basit A, Chidambaram M, Das D, Grundberg E, Hedman AK, Hydrie ZI, Islam M, Khor CC, Kowlessur S, Kristensen MM, Liju S, Lim WY, Matthews DR, Liu J, Morris AP, Nica AC, Pinidiyapathirage JM, Prokopenko I,

Rasheed A, Samuel M, Shah N, Shera AS, Small KS, Suo C, Wickremasinghe AR, Wong TY, Yang M, Zhang F, Abecasis GR, Barnett AH, Caul- field M, Deloukas P, Frayling TM, Froguel P, Kato N, Katulanda P, Kelly MA, Liang J, Mohan V, Sanghera DK, Scott J, Seielstad M, Zimmet PZ, Elliott P, Teo YY, McCarthy MI, Danesh J, Tai ES, Chambers JC. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. NatGenet 43: 984 –989, 2011.

[13] Lander ES: The new genomics: global views of biology. Science274 :536 –539,1996

[14] Logos T. Wilcoxon singed rank test. R bloggers. July 2009. (URL: http://www.rbloggers.com/wilcoxon-signed-rank-test/

[15] McNamara J. Spreadsheet-WriteExcel-2.40. November 2013. (URL: http://github.com/jmcnamara/spreadsheet-writeexcel)

[16] Mott R, Talbot CJ, Turri MG, Collins AC and Flint J. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci.* 97:12640 12654, 2001.

[17] Muoio DM, Newgard CB. Mechanisms of disease: molecular and met- abolic mechanisms of insulin resistance and beta-cell failure in type 2 diabetes. Nat Rev Mol Cell Biol 9: 193–205, 2008.

[18] Obesity Disease Portal, Rat Genome Database Web Site, Medical College of Wisconsin, Milwaukee, Wisconsin. World Wide Web (URL: http://rgd.mcw.edu/). [January 2015].

[19] Polonsky KS, Sturis J, Bell GI. Seminars in Medicine of the Beth Israel Hospital, Boston. Non-insulin-dependent diabetes mellitus - a genetically programmed failure of the beta cell to compensate for insulin resistance. N Engl J Med 334: 777–783, 1996.

[20] Risch NJ: Searching for genetic determinants in the new millennium. Nature 405 :847–856,2000

[21] Saxena R, Elbers CC, Guo Y, Peter I, Gaunt TR, Mega JL, Lanktree MB, Tare A, Castillo BA, Li YR, Johnson T, Bruinenberg M, Gilbert- Diamond D, Rajagopalan R, Voight BF, Balasubramanyam A, Bar- nard J, Bauer F, Baumert J, Bhangale T, Bohm BO, Braund PS, Burton PR, Chandrupatla HR, Clarke R, Cooper-DeHoff RM, Crook ED, Davey-Smith G, Day IN, de Boer A, de Groot MC, Drenos F, Ferguson J, Fox CS, Furlong CE, Gibson Q, Gieger C, Gilhuijs Pederson LA, Glessner JT, Goel A, Gong Y, Grant SF, Grobbee DE, Hastie C, Humphries SE, Kim CE, Kivimaki M, Kleber M, Meisinger C, Kumari M, Langaee TY, Lawlor DA, Li M, Lobmeyer MT, Maitland-van der Zee AH, Meijs

MF, Molony CM, Morrow DA, Murugesan G, Musani SK, Nelson CP, Newhouse SJ, O'Connell JR, Padmanabhan S, Palmen J, Patel SR, Pepine CJ, Pettinger M, Price TS, Rafelt S, Ranchalis J, Rasheed A, Rosenthal E, Ruczinski I, Shah S, Shen H, Silbernagel G, Smith EN, Spijkerman AW, Stanton A, Steffes MW, Thorand B, Trip M, van der Harst P, van der AD, van Iperen EP, van Setten J, van Vliet-Ostaptchouk JV, Verweij N, Wolffenbuttel BH, Young T, Zafarmand MH, Zmuda JM, Boehnke M, Altshuler D, McCarthy M, Kao WH, Pankow JS, Cappola TP, Sever P, Poulter N, Caulfield M, Dominiczak A, Shields DC, Bhatt DL, Zhang L, Curtis SP, Danesh J, Casas JP, van der Schouw YT, Onland-Moret NC, Doevendans PA, Dorn GW 2nd, Farrall M, FitzGerald GA, Hamsten A, Hegele R, Hingorani AD, Hofker MH, Huggins GS, Illig T, Jarvik GP, Johnson JA, Klungel OH, Knowler WC, Koenig W, März W, Meigs JB, Melander O, Munroe PB, Mitchell BD, Bielinski SJ, Rader DJ, Reilly MP, Rich SS, Rotter JI, Saleheen D, Samani NJ, Schadt EE, Shuldiner AR, Silverstein R, Kottke Marchant K, Talmud PJ, Watkins H, Asselbergs FW, de Bakker PI, McCaffery J, Wijmenga C, Sabatine MS, Wilson JG, Reiner A, Bowden DW, Hakonarson H, Siscovick DS, Keating BJ. Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. Am J Hum Genet 90: 410–425, 2012.

[22] Solberg Woods LC, KL Holl, D Oreper, Y Xie, S Tsaih and W Valdar. 2012. Fine mapping diabetes related traits, including insulin resistance, in heterogeneous stock rats. Physiol Genomics. 44:1013-1026.

[23] Solberg Woods L, He H, Holl K, Litrell J, Prokop JW, Zagloul M, Keele G, Xie Y, Baur B, Fox J, Robinson M, Levy S, Valdar W. Genetic fine-mapping and identification of a likely causal variant for adiposity traits in outbred rats. *Obesity*. December 2014. *Under review.*

[24] Valdar W, Holmes CC, Mott R, Flint J. Mapping in structured popu- lations by resample model averaging. *Genetics* 182: 1263–1277, 2009.

[25] Variant Visualizer, Rat Genome Database Web Site, Medical College of Wisconsin, Milwaukee, Wisconsin. World Wide Web (URL: http://rgd.mcw.edu/). [January 2015].

[26] Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segre AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Bostrom K, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jor- gensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M,

Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proenca C, Prokopenko I, Rath- mann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparso T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllen- sten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pram- staller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI; MAGIC investigators; GIANT Consortium. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42: 579–589, 2010.