AD_____

Award Number: W81XWH-11-1-0261

TITLE: Use of eQTL Analysis for the Discovery of Target Genes Identified by GWAS

PRINCIPAL INVESTIGATOR: Stephen Thibodeau, PhD

CONTRACTING ORGANIZATION: Mayo Clinic, Rochester, MN 55905

REPORT DATE: April 2014

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
April 2014	Final	1 April 2011 – 31 March 2014

4. TITLE AND SUBTITLE
Use of eQTL Analysis for the Discovery of Target Genes Identified by GWAS

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-1-0261

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Stephen Thibodeau

E-Mail: sthibodeau@mayo.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Mayo Clinic

Rochester, MN 55905-0002

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The goal of this grant proposal was to: 1) construct a prostate tissue-specific expression quantitative trait loci (eQTL) dataset; and 2) utilize this dataset to identify candidate genes for existing prostate cancer (PC) risk-single nucleotide polymorphisms (SNPs) that could then be followed up in future studies. To accomplish this goal we performed a genome-wide SNP analysis (Illumina Human Omni 2.5M SNP array) and a genome-wide mRNA expression analysis (RNA sequencing) on a common set of 500 samples of normal prostate tissue sampled from men with PC. Of 500 processed samples, 471 samples passed stringent quality control (QC) and were available for further analysis. Our primary analysis focused on identifying eQTLs for 146 PC risk-SNPs, including all SNPs in linkage disequilibrium with each risk-SNP (r^2 >0.5), resulting in 100 unique risk-intervals. Furthermore, we focused on *cis*-acting associations only where the transcript was located within 2Mb (+/-1Mb) of the risk-SNP interval. Of all SNPs located in the 100 risk-intervals (N=6324 SNPs), 1,718 demonstrated a significant eQTL signal after adjustment for sample histology (% lymphocytes and % epithelial cells) and meeting a Bonferroni-adjusted p-value threshold of 1.96 e-7 (ranged from 1.96 e-7 to 1.52 e-91). Of the 100 PC risk-intervals, 31 (31%) demonstrated a significant eQTL signal and these were associated with 54 genes.

15. SUBJECT TERMS
eQTL dataset

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					USAMRMC
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	94	

Table of Contents

Page

A. INTRODUCTION:

We hypothesized that many of the PC disease-associated SNPs already identified to date will be located in regulatory domains involved in gene transcription. Furthermore, we hypothesized that candidate genes affected by these regulatory elements could be identified by taking advantage of eQTL datasets. Therefore, the objectives of this grant proposal were to: 1) construct a prostate tissue-specific eQTL dataset that could be used to identify candidate genes for any current (or future), predictive (or prognostic) SNP identified for PC; and 2) utilize this dataset to identify candidate genes for existing PC risk SNPs that could then be followed up in future studies. To accomplish this goal, we proposed to perform a genome-wide SNP analysis (using the Illumina Human Omni 2.5M SNP array) and a genome-wide mRNA expression analysis (using RNA sequencing) on a common set of 500 samples of normal prostate tissue sampled from men with PC. The long-term objective of this strategy is to characterize the functional role of the disease-causing SNPs, to identify the genes and biologic pathways affected by these inherited factors, and ultimately to identify targets for disease prediction, risk stratification and identification of treatment targets.

B. BODY:

Statement of work <u>originally</u> proposed for years 1 and 2:

Task 1. Processing of normal prostate tissue for RNA purification (months 1-9)
- **1a.** Cryo-section fresh-frozen tissue from ~500-600 cases (months 1-9)
- **1b.** Create hematoxylin-eosin (H&E) stained slides from each case for review (months 1-9)
- **1c.** Review of sections by a Pathologist. (months 1-9)
- **1d.** Select 500 cases of high-quality samples for RNA extraction (Task 2) (months 10)

Task 2. DNA and RNA Extraction from 500 cases for study (months 11-12)
- **2a.** Use sections from 500 samples selected from Task 1 to purify DNA and total RNA (months 11-12)

Task 3. Genome-wide genotyping of blood DNA from 500 cases for study (months 12-14)
- **3a.** Place blood DNA (already extracted) in 96-well plates for genotyping (months 12)
- **3b.** Genotype samples (months 12-14)
- **3c.** Quality-control checks and data processing – Statistical analyses (months 14)

Task 4. Genome-wide mRNA profiling of tissue RNA from 500 cases for study (months 13-15)
- **4a.** Place RNA in 96-well plates for expression analysis (months 13)
- **4b.** Perform expression analysis (months 13-14)
- **4c.** Quality-control checks and data processing – Statistical analyses (months 15)

Task 5. Create eQTL dataset – Statistical analysis (months 16-24)
- **5a.** Test PC risk-SNPs for their association with transcript level for all mRNAs utilizing data from Tasks 3 and 4 (months 16-18)
- **5b.** Test candidate target gene for association with all other SNPs (months 18-21)
- **5c.** Prepare data for public distribution (months 21-24)

Work performed: Task 1 (Processing of normal prostate tissue for RNA purification)

<u>All of the work proposed for Task 1 has been completed.</u>

In order to achieve our goal of 500 samples of normal prostate tissue, we initially reviewed H&E stained sections from all archived cases available for study; ~4,000. These ~4000 cases were obtained from patients whom had undergone a radical prostatectomy at Mayo Clinic and were available to investigators through the Prostate Cancer SPORE. Typically, one to three pieces of frozen tissue (snap frozen at the time of surgery) were available for each case. At the time each case was initially processed, a representative H&E stained slide was made from each piece of tissue and archived for future investigator review to aid in the process of tissue selection. Although the archived slide allows for an initial evaluation, blocks are used over time and the histology can change. Thus, cutting an additional representative H&E is often necessary to re-evaluate the current state of these blocks.

For this study, the same Pathologist was used throughout the evaluation process to ensure consistency. In our initial pre-screen of the ~4000 normal tissue cases, we first removed all cases where the patient's tumor had a Gleason score greater than 7, cases where tumor was found on the H&E slide and cases where normal

prostate tissue was not available. Following this initial review, 916 pieces of tissue were available for further processing. The archived tissue was then pulled from long-term storage and a fresh representative H&E stained slide was prepared for re-evaluation by a Pathologist. In order to meet the needs of this study, the following criteria were developed for further tissue selection and processing:

1. No tumor present on the new H&E.
2. The section viewed had to be from the posterior region of the prostate – all central and anterior zone tissues were eliminated. The region of interest was determined based on histologic landmarks and Mayo practice processes (posterior region are inked for orientation).
3. No High-Grade Prostatic Intraepithelial Neoplasia (HGPIN).
4. No greater than 1% of the cells on the slide could be lymphocytes.
5. The final percent of epithelial glands present on the slide had to be at least 40%.

Of the 916 cases re-examined, 93 cases met the criteria above, but also contained Benign Prostatic Hyperplasia (BPH), seminal vesicle, urethra, or adjacent central zone. These pieces of tissue were further processed to eliminate the contaminating portion and an additional H&E stained section was prepared to ensure that the block was processed correctly and the unwanted regions were adequately removed.

Following the final review of tissue, 565 cases met the selection criteria noted above. Due to the small number of cases meeting our strict histologic criteria (565 of ~4000 cases reviewed), most of the selected cases did not have blood available for the extraction of DNA (for genotyping). As a result, we chose to take additional sections of the normal prostate tissue, which allowed for the extraction of both RNA (expression) and DNA (genotyping). From past experience, we expected that a degree of histologic change would be present throughout the sectioning process and this would result in an additional ~10% of the cases failing to meet our selection criteria. Thus, we decided to section and evaluate all 565 cases, re-evaluate H&E stained sections once more and then choose the best cases for the final processing.

Work performed: Task 2 (DNA and RNA Extraction from 500 cases for study)

All of the work proposed for Task 2 has been completed.

For the extraction of DNA and RNA, tissue was first sectioned on a cryostat, preparing 10-micron thick sections. Prior to sectioning, however, all of the samples were randomized into cutting groups based on percent epithelium, presence or absence of lymphocytes, the time of original tissue collection, and if the tissue came from prostate cancer patients or from patients having a cysto-prostatectomy due to bladder cancer. The randomization of samples was performed in order to control for any cutting bias that might be introduced as the tissue was processed each day. The 565 cases were sectioned over a period of 26 working days in the following manner: the initial section was taken for an H&E stained slide (to serve as a one-to-one comparison with the initially reviewed H&E section to confirm that no tissue mix-up had occurred), then multiple sections placed in tube 1 for RNA, a 2nd H&E section, multiple sections placed in tube 2 for RNA, 3rd H&E section, multiple sections placed in tube 3 for DNA, 4th H&E section, multiple sections placed in tube 4 for DNA, and the final H&E section. For the RNA-destined tubes, tissue was immediately placed in QIAzol buffer and then snap frozen to ensure high-quality RNA. For the DNA-destined tubes, sections were placed in tubes and initially stored at -80° C. These tubes were then collected the following day, and QIAgen Gentra Puregene cell lysis buffer and proteinase K were added to both DNA tubes and digested overnight at 55° C on a shaking incubator essentially as outlined by the manufacturer. Visual confirmation was done the following day to ensure all of the tissue was digested. The tubes were then considered stable and stored at 4° C pending completion of the DNA extraction.

All five H&Es sections outlined above were evaluated once again by a Pathologist to ensure that no histologic changes had occurred as the tissue was sectioned. Additionally, the 1st H&E was used to compare to the original H&E confirming that no specimen mix-ups had occurred. Upon histologic review of all five H&E slides, roughly 10% of the cases were eliminated due to histologic changes (i.e. the appearance of small cancer foci, change in percent epithelium, appearance of HGPIN, an increase in lymphocytic presence). Following this final review, 505 cases remained that met the initial criteria. Again, because we anticipated that there would be a small number of cases having poor-quality RNA or poor-DNA yield, an additional 19 cases were selected that had 2% infiltrative lymphocytes present for the final process of DNA and RNA extracted. These 524 cases were then split into two batches for RNA extraction and re-randomized again as previously described, but now the randomization scheme also included the day the tissue was processed. This randomization was performed to avoid any batch effects during RNA extraction.

DNA was extracted by first performing a protein precipitation step (Qiagen protein precipitation solution), followed by an isopropanol then Ethanol rinse. The DNA pellet was allowed to dry, then dissolved in TE and allowed to mix overnight. After mixing, DNA was quantified using a nanodrop, and concentrations were standardized. Total RNA was extracted the using the RNeasy Mini Kit (Qiagen) according to the

manufacturer's instructions on the Qiacube. RNA was then assessed for quality using an Agilent chip technology. Cases having a RIN number of 7.0 or greater were considered good quality. Once completed, the optimum set of 500 samples were then selected for the mRNA expression and DNA genotyping studies based on RNA and DNA quality and those samples meeting the most strict selection criteria (i.e., higher percent epithelium, no or fewest lymphocytes present). Following this initial selection, six samples were later omitted because they were found to not meet the original criteria for the grade of tumor (Gleason score of 7 or less).

Work performed: Task 3 (Genome-wide genotyping of blood DNA from 500 cases for study)

All of the work proposed for Task 3 has been completed.

As originally proposed, DNA from 500 tissue samples were selected and randomized to 96-well plates with two CEPH controls on each plate. Samples were then genotyped using the Illumina Human Omni 2.5M SNP array. These studies along with the QC analyses to identify sample and/or SNP quality issues have been completed.

QC analyses included the evaluation of call-rates, minor allele frequencies, and tests of Hardy-Weinberg Equilibrium (HWE) for each of the SNPs. The QC filters that were applied to the genotypic data include excluding SNPs with: 1) call-rate < 95%; 2) MAF < 1%; 3) HWE p-value < 1e-4; 4) concordance in duplicates < 99.5%; and 5) unknown physical position based on current genome build. In addition, we estimated the genotyping error rates by checking for Mendelian consistency and duplicate concordance rates using CEPH controls. Finally, we tested for potential batch effects by testing for allele frequency and call rate differences across plates. Subject level QC included calculation of call-rates, sex determination, as well as calculation of pair wise identity by descent probabilities for all pairs of subjects in order to identify and remove related subjects. See **Appendix 1 and 2** for complete QC report. **Appendix 1** includes information for all SNPs and all samples. **Appendix 2** provides information after excluding problematic SNP and problematic samples and includes additional QC tests.

Overall, the quality of the 2.5M SNP genotyping data is excellent. A total of 17 of 494 samples were flagged for QC reasons; 5 samples had a SNP call rate < 95%, 10 are non-Caucasian (5 African, 5 Asian) and 2 subjects appear to be first cousins. After excluding one of the related pair, we have 478 unrelated, Caucasian samples remaining for analysis. SNP exclusions are summarized below. We have ~1.5M QC-passed SNPs with MAF >= 1% available for analysis.

Sample exclusions:	494 samples
	5 call rate < 95%
	10 non-Caucasian (5 African; 5 Asian)
	1 related pair

Samples remaining:	478
SNP exclusions:	2,372,617 SNPs are on the 2.5M array
	6,409 call rate < 95% (205 failed completely)
	454,736 monomorphic
	902 HWE p-value < 1e-5 (276 with p < 1e-10)

SNPs remaining:	1,910,570
MAF > 1%	1,558,636

Work performed: Task 4 (Genome-wide mRNA profiling of tissue RNA from 500 cases for study)

All of the work proposed for Task 4 has been completed.

In the original statement of work, we had proposed the use of the Illumina humanht-12 BeadChip as the platform to derive the genome-wide mRNA expression dataset. However, the cost of next generation sequencing (NGS) dropped dramatically over the course of our project and, as a result, we explored the option of performing RNA profiling by NGS (RNAseq). The use of RNAseq significantly increased both the quality and value of this dataset. We were able to obtain additional funds to supplement the DOD award to perform these experiments, and following approval by the Scientific Officer, we changed our approach for this task to RNA sequencing. To accomplish the work proposed, we utilized the Agilent SureSelect RNA capture kit for the RNA library preparation and the Illumina HiSeq 2000 for the RNA sequencing. For these experiments, samples were first randomized to library-prep groups. The randomization was performed as previously described, but now the randomization scheme included both the day the tissue was processed and the RNA extraction group. This randomization was performed to avoid any batch effects during sequencing. Samples were indexed such that

five samples were analyzed in a single lane. Our goal was to achieve a minimum of 50 million reads per sample – and this has been accomplished.

The first-phase Bioinformatic analysis was completed using an in-house-developed pipeline, MAP-RSeq. MAP-RSeq is a comprehensive computational pipeline for secondary analysis of RNA-Sequencing data. MAP-RSeq uses a variety of freely available bioinformatics tools along with in-house-developed methods. Alignment and mapping of the reads was performed using Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) and TopHat (http://tophat.cbcb.umd.edu/) software. Gene counts were generated using HTseq software (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html) and gene annotation files were obtained from Illumina (http://cufflinks.cbcb.umd.edu/igenomes.html). For single nucleotide variant (SNV) calling, we used the GATK (http://www.broadinstitute.org/gatk/) software. SNVs were further annotated and filtered for quality, coverage and other criteria using variant quality score recalibration (VQSR) method. MAP-RSeq also provides a list of expressed fusion transcripts using TopHat-Fusion algorithm. All of the bioinformatics analysis using MAP-RSeq has now been completed.

As with the Genotype data, QC assessment of the RNAseq data is also completed. We compared RNA-called genotypes to genotypes from the Illumina Human Omni 2.5M array to test for sample mix-ups. To investigate factors that may influence the number of counts observed, we summarized the log2(gene counts) and the percentage of counts > 0 by subject, lane, flowcells, %GC content per gene and by gene size (counting only the sum of the exons). Data quality was assessed via per-specimen box plots and minus versus average (MVA) plots. The box plots were sorted by various experimental factors, e.g., batch and run order in order to examine global shifts in counts due to these factors. The existence of and functional form of biases between specimens were assessed via residual MVA plots. The modified MVA plot uses a linear model to examine trends in residuals. A detailed description with examples of the QC analyses performed is provided in **Appendix 3**. Overall, the quality of the RNAseq data was excellent.

In addition, a manual review of several Bioinformatically generated sample-specific RNAseq parameters **(Figure 1)** was conducted for each sample. These include the following: junction saturation **(Figure 2 A)**; splice junctions **(Figure 2 B)**; inner distance **(Figure 2 C)**; read duplication; and gene body coverage **(Figure 2 D)**. **Figure 1** shows data for five representative samples, while **Figure 2** shows data for two samples, one with acceptable data (left) and one with unacceptable data (right). From these analyses, eight samples were flagged as potentially problematic.

Figure 1

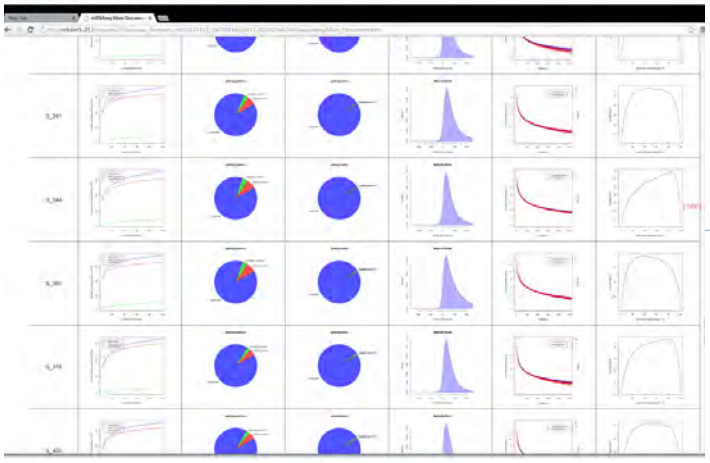
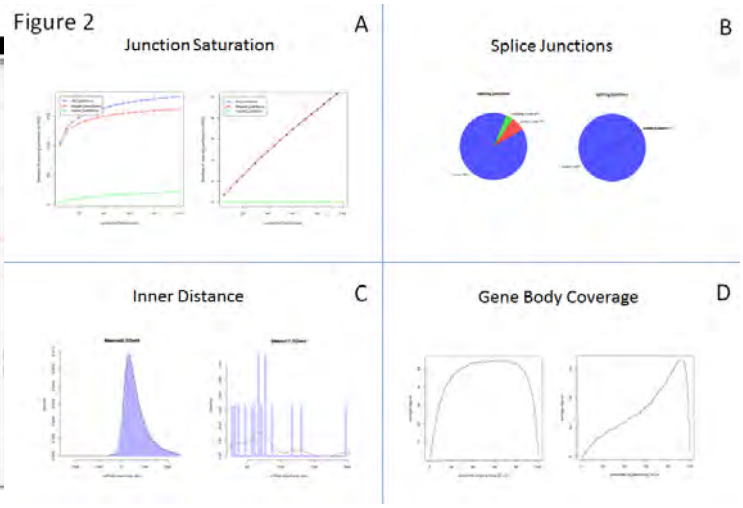


Figure 2



Work performed: Task 5 (Create eQTL dataset)

All of the work proposed for Task 5 has been completed.

For the eQTL dataset, we are interested in both coding (as originally planned) as well as newly described long intergenic non-coding RNA (lincRNA). The standard pipeline described above provides a description of all of the coding transcripts, but not for lincRNAs. As a result, we developed a pipeline to identify, quantify and annotate lincRNA and have applied this to our RNAseq data. These analyses have been completed.

The pipeline consists of several modules:

1) **Candidate transcript assembly module**: this module uses a genome-guided strategy for transcriptome reconstruction. The aligned BAM files (i.e., BAM files from TopHat) were assembled with Cufflinks 2.0.2. The option “Reference Annotation Based Transcript” (RABT) assembly was used

because of its advantage to identify novel transcripts. The GENCODE V16 was used as annotation file to guide the transcript assembly processes.

2) **LincRNA identification module:** this module aimed to identify and report expressed lincRNAs in the RNAseq data. To achieve this, five filtering steps were used as follows.
    a. **Size restriction**: transcripts smaller than 200 nt were removed.
    b. **Removal of known protein-coding regions**: candidate transcripts that overlap with transcripts in the "protein-coding" category in GENCODE V16 were removed.
    c. **Removal of transcript homologous to known proteins**: the blastx program was used to evaluate the similarity between candidate transcripts and known proteins in the RefSeq database (protein with NM_ prefix). The transcripts with E value less than 1e-4 were removed.
    d. **Removal of transcripts predicted to code for proteins**: the candidate transcripts were then assessed for their coding potential by the CPAT tool, an in silico computational model classifying coding and non-coding transcripts. Specifically, a logistic regression model was built based on four sequence features, including open reading frame size, open reading frame coverage, Fickett TESTCODE statistic and hexamer usage bias. A training dataset was constructed containing both known protein-coding (NM_ prefix in RefSeq database) and non-coding transcripts. Compared to other widely used tools such as CPC and PhyloCSF, CPAT has higher sensitivity and specificity (>0.966), and is much faster (i.e., process thousands of transcripts within seconds).
    e. **Known protein domain filter**: the remaining candidate transcripts were then evaluated whether they contain a known protein coding domain. To achieve that, each candidate transcript was translated in all three reading frames and compared against 13,672 known protein family domains documented in the Pfam database Version 26 by the HMMER-3 tool. HMMER-3 uses hidden Markov models (HMMs) to scan each amino acid sequence and classify whether it resembles any of the known domains in the database. Candidate transcripts with a significant Pfam hit (P value less than 1e-5) were excluded.

In total, we identified 72,740 candidate lincRNA transcripts at 38,899 intergenic loci in 494 normal prostate tissue samples. Among these transcripts, significant overlap was observed between them and lincRNAs annotated in GENCODE V17, i.e., 63% of lincRNAs annotated in GENCODE V 17 were also identified in our dataset. These prostate-derived lincRNAs were further examined for evidence of transcriptional activity using the H3K4me3-H3K36mer3 domains generated from nine cell lines in the ENCODE project. Overall, 18,368 lincRNAs (~25%) have evidence of a signature consistent with an actively transcribed gene across the entire locus (both H3K4me3 across the promoter region and H3K36me3 along the transcribed region). Of the remaining transcripts, 7,849 (11%) overlap an H3K4me3 peak alone (promoter region) and 6,856 (9%) overlap an H3K36me3 peak alone (transcribed region). **A manuscript describing the lincRNA work is now in preparation.**

eQTL Analysis. As noted above, genome-wide genotypes and genome-wide mRNA expression levels were obtained with the use of the Illumina Human Omni 2.5M SNP array and by RNA sequencing, respectively. Following extensive QC, our final dataset consisted of: a) 471 normal prostate tissue samples (453 from low Gleason grade PC cases and 18 from Cystoprostatectomy cases); b) 1,542,229 SNPs; and c) 17,252 expressed genes.

For PC, multiple GWAS and confirmatory studies have provided a substantial number of well-validated SNPs (~146) that are associated with an increased risk of developing PC **(Table 1)**. Our primary analysis focused on identifying eQTLs for these 146 PC risk-SNPs, including all SNPs in linkage disequilibrium with each risk-SNP (r^2 >0.5), resulting in a total of 6,324 SNPs to be evaluated in 100 unique risk-intervals. The number of SNPs evaluated for each of the risk regions is shown in **Table 2**.

Furthermore, we focused on *cis*-acting associations only, where the transcript was located within 2Mb (+/-1Mb) of the risk-SNP interval. A total of 3,142 gene transcripts within these intervals were identified. Of these, 867 were not evaluated due to low or no expression, leaving 2,275 for further analysis. The genes localized to each of these regions are shown in **Table 2**.

Of the 6,324 SNPs located in the 100 risk-intervals, 1,718 demonstrated a significant eQTL signal after adjustment for sample histology (percent lymphocytes and percent epithelial cells) and meeting a Bonferroni-adjusted p-value threshold of 1.96e-7 (results ranged from 1.96e-7 to 1.52e-91). Of the 100 PC risk-intervals, 31 (31%) demonstrated a significant eQTL signal and these were associated with 54 genes. Examples for two of the significant eQTL regions of interest are shown in **Appendices 4 and 5**. **Appendix 4** shows data for the risk-SNP region for **rs12653946** on Chromosome 5 (6 Kb region) and the associated gene identified - **IRX4** (all P-value less than e-40). **Appendix 5** shows data for the risk-SNP region for **rs8102476** on Chromosome 19

(30 Kb region) and the associated gene identified - **PPP1R14A** (all P-value less than e-20). **A manuscript describing the eQTL analysis is now in preparation.**

C. KEY RESEARCH ACCOMPLISHMENTS:

- Tissue processing completed.
- Extraction of tissue RNA and DNA completed.
- DNA genotyping of 500 samples using the Illumina Human Omni 2.5M SNP array completed.
- RNA sequencing of 500 samples using the Agilent SureSelect RNA capture kit and the Illumina HiSeq 2000 completed.
- QC assessment of both Genotype and RNAseq data completed.
- Identified, quantified and annotated lincRNA in our RNAseq data (manuscript in preparation).
- eQTL dataset constructed (manuscript in preparation).
- eQTL analysis for 146 reported risk-SNPs completed (manuscript in preparation).
- Identified eQTL signals for 54 riskSNP – gene combinations.

D. REPORTABLE OUTCOMES:

- Three manuscripts now in preparation
- eQTL dataset constructed
- Information from this DOD grant was helpful in our obtaining an NIH award (CA151254)

E. CONCLUSION:

The major goal of this proposal was to construct a prostate tissue-specific expression quantitative trait loci (eQTL) dataset. Tissue processing, RNA and DNA purification, DNA genotyping and RNA expression analysis, and identification of all lincRNA's for the construction of this eQTL dataset has now been completed.

We hypothesized that many of the PC disease-associated SNPs identified to date would be located in regulatory domains involved in gene transcription. Furthermore, we hypothesized that candidate genes affected by these regulatory elements could be identified by taking advantage of an eQTL dataset. The results of this study show convincing data that this is, in fact, the case.

Of the 6,324 SNPs located in the 100 risk-intervals, 1,718 demonstrated a significant eQTL signal after adjustment for sample histology (percent lymphocytes and percent epithelial cells) and meeting a Bonferroni-adjusted p-value threshold of 1.96e-7 (ranged from 1.96e-7 to 1.52e-91). Of the 100 PC intervals containing a PC risk-SNP, 31 (31%) demonstrated a significant eQTL signal and these were associated with 54 genes. Thus, 54 genes have now been identified as candidate risk genes for prostate cancer. <u>This is the largest number of candidate susceptibility genes found to date for prostate cancer</u>.

<u>All aspects of this grant proposal have been completed successfully with very positive and exciting results.</u>

F. REFERENCES: None

G. APPENDICES:

Appendix 1: SNP QC report, for all SNPs and all samples.
Appendix 2: SNP QC report after excluding problematic SNP and problematic samples and includes additional QC tests.
Appendix 3: mRNA QC report
Appendix 4: eQTL analysis for Chromosome 5 region of interest
Appendix 5: eQTL analysis for Chromosome 19 region of interest

H. SUPPORTING DATA:

Table 1: List of PC risk-SNPs used for the study, including chromosome location
Table 2: Number of SNPs and number of genes evaluated for each of the risk regions

Appendix 1

SNP QC report, for all SNPs and all samples

EQTL Test Summary

Inv: SNThibodeau
Statistics Team: McDonnell,Kosel
Bioinformatics Team: Asmann,Middha,Hossain

Mayo Clinic College of Medicine, Health Sciences Research
Rochester MN USA

September 13, 2013

Contents

1 Introduction

This document summarizes GWAS QC analysis performed on the HumanOmni2.5-4v1 chip for Prostate Cancer patients. Data are available for 736 samples from 2,372,617 SNPs including 16 CEPH controls. This summary includes data for 510 samples and 2372617 SNPs including 16 controls.

2 Initial SNP Quality Control

2.1 SNP Call Rates

We first look at how many SNPs drop out using different SNP call rate cutoffs. See Table 1 (p. 6) for the percentage of SNPs retained as the call rate threshold increases. A total of 205 SNPs (0.009%) failed completely. Using a call rate of 98%, 28,443 SNPs (1.2%) will be dropped. Using a call rate of 95%, 6,409 SNPs (0.3%) will be dropped.

2.2 Failed, Monomorphic, and Low Call Rate SNPs by Chromosome

This section describes how many SNPs failed completely, are "monomorphic", or have a call rate $< 95\%$ by chromosome and overall (Table 2, p. 8). First "failed" SNPs are identified, then "Monomorphic", and finally those SNPs with a call rate $< 0.95\%$. The distribution of SNP call rates by chromosome is presented in Figure 1 (p. 4).

2.3 Minor Allele Frequency

The distribution of minor allele frequencies (MAFs) for all SNPs is shown in Figure 2 (p. 5). There are a total of 456,321 (19.23%) monomorphic SNPs and 809,688 (34.13%) SNPs with MAF $< 1\%$.

2.4 Hardy Weinberg P-value

This dataset does not include controls to reliably test for Hardy-Weinberg Equilibrium so the following results should be interpreted with caution. We include only caucasian subjects resulting in 494 independent subjects. Chromosomes X, Y, XY, and MT markers

Figure 1: SNP Call Rates by Chromosome

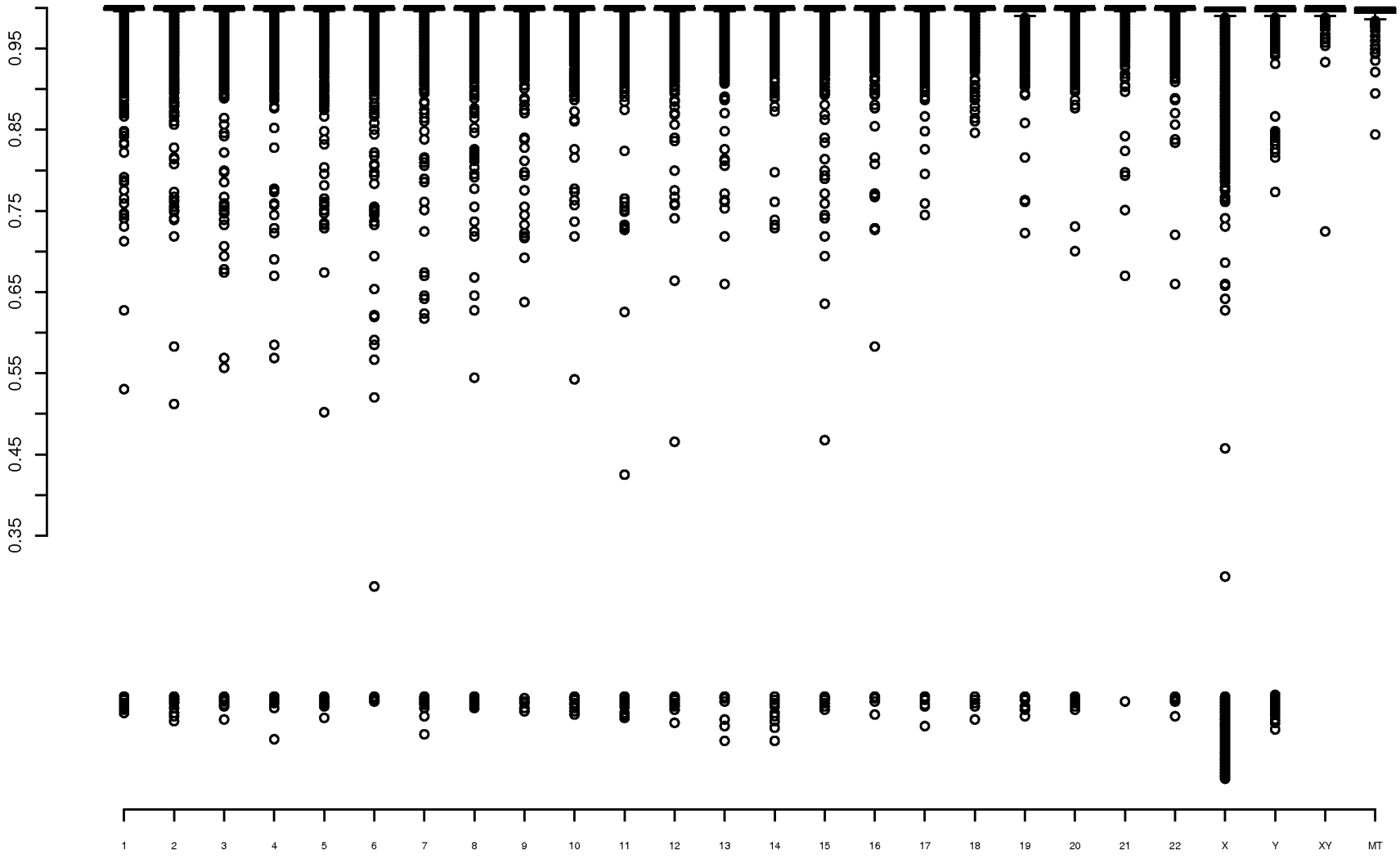


Figure 2: Histogram of Minor Allele Frequencies

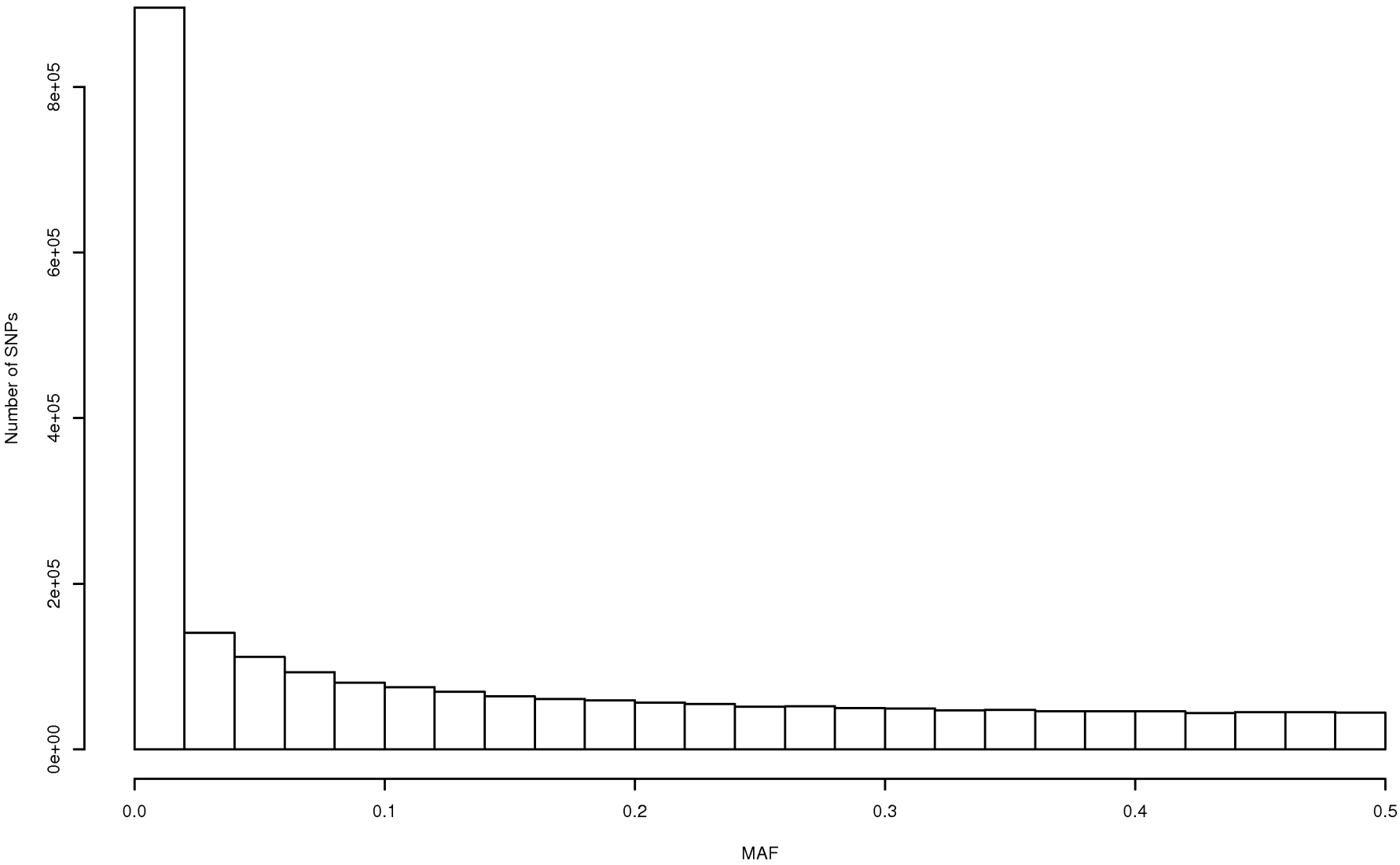


Table 1: SNP Call Rates

CallRate	NumSNPsBelow	%Below	NumSNPsAbove	%Above
0.000	205	0.000	2372412	100.000
0.800	2200	0.100	2370417	99.900
0.850	2458	0.100	2370159	99.900
0.900	2906	0.100	2369711	99.900
0.910	3111	0.100	2369506	99.900
0.920	3424	0.100	2369193	99.900
0.930	3968	0.200	2368649	99.800
0.940	4877	0.200	2367740	99.800
0.950	6409	0.300	2366208	99.700
0.960	9328	0.400	2363289	99.600
0.970	14625	0.600	2357992	99.400
0.980	28443	1.200	2344174	98.800
0.990	159173	6.700	2213444	93.300
1.000	901479	38.000	1471138	62.000

are excluded from this summary as are SNPs that failed on all samples and SNPs with MAF < 0.05. There are 1,242 SNPs have a HWE p-value $<$ 10e-05 (see Figure 3, p. 7).

Figure 3: Q-Q plot of HWE p-values (573 p-values have been truncated at 10e-10)

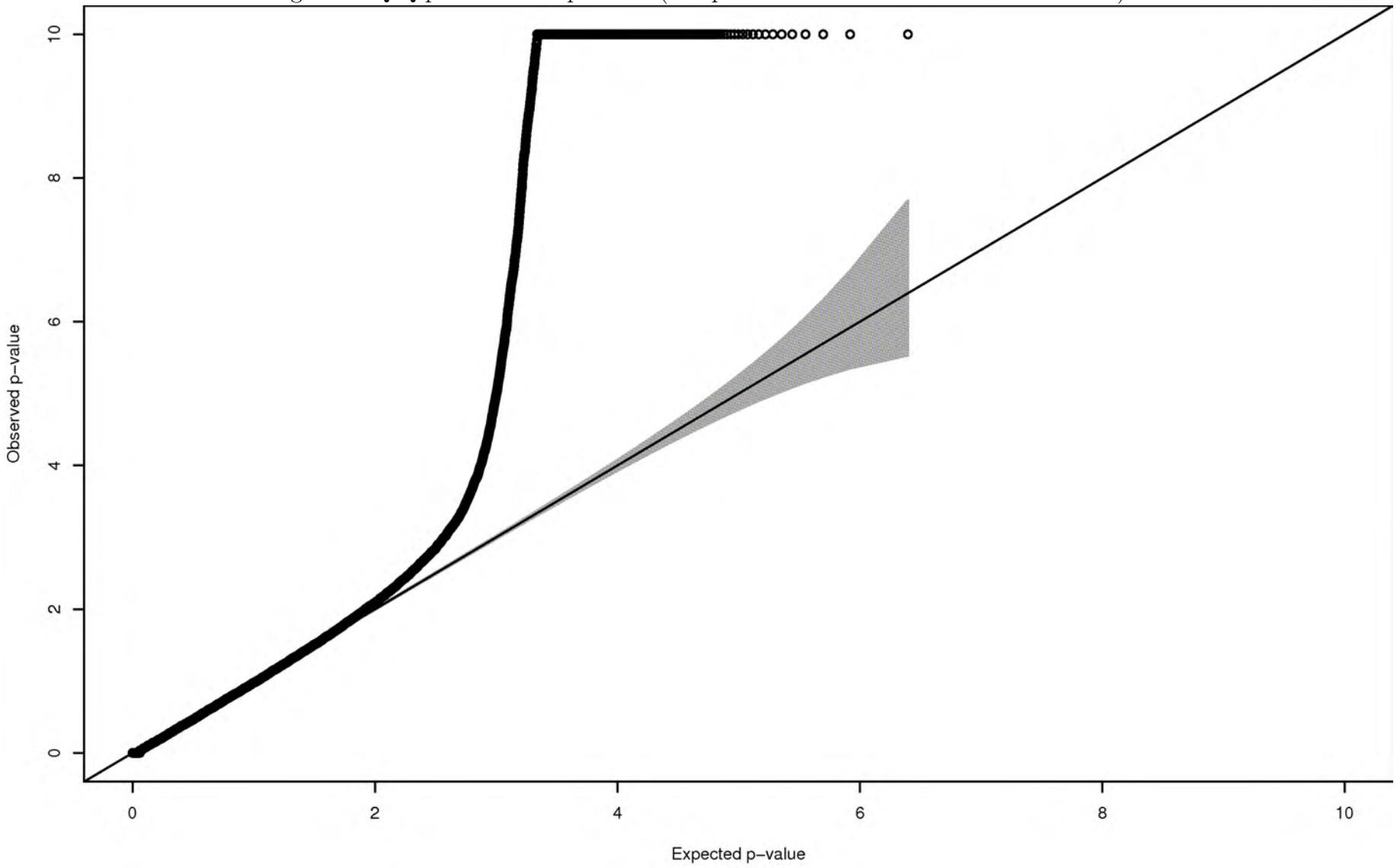


Table 2: SNP QC Summary by Chromosome - CEPH samples excluded

		Failed		Monomorphic		Callrate<0.95		Remaining	
Chrom	TotalSNPs	N	%	N	%	N	%	N	%
1	184072	10	0.01	37394	20.31	267	0.15	146401	79.53
2	194126	8	0.00	39033	20.11	245	0.13	154840	79.76
3	163672	16	0.01	31653	19.34	193	0.12	131810	80.53
4	152846	7	0.00	28989	18.97	193	0.13	123657	80.90
5	145453	4	0.00	29638	20.38	170	0.12	115641	79.50
6	154686	7	0.00	28652	18.52	259	0.17	125768	81.31
7	129072	5	0.00	24646	19.09	209	0.16	104212	80.74
8	125515	6	0.00	23393	18.64	189	0.15	101927	81.21
9	103011	6	0.01	19384	18.82	140	0.14	83481	81.04
10	119408	8	0.01	22824	19.11	163	0.14	96413	80.74
11	116095	4	0.00	23212	19.99	192	0.17	92687	79.84
12	112722	3	0.00	22343	19.82	158	0.14	90218	80.04
13	83483	4	0.00	14950	17.91	102	0.12	68427	81.97
14	76510	6	0.01	14566	19.04	105	0.14	61833	80.82
15	72294	3	0.00	13249	18.33	104	0.14	58938	81.53
16	76610	5	0.01	13546	17.68	139	0.18	62920	82.13
17	66387	4	0.01	12459	18.77	152	0.23	53772	81.00
18	68552	5	0.01	12196	17.79	90	0.13	56261	82.07
19	47733	3	0.01	8787	18.41	131	0.27	38812	81.31
20	56542	4	0.01	10103	17.87	94	0.17	46341	81.96
21	32075	4	0.01	5604	17.47	32	0.10	26435	82.42
22	33310	3	0.01	4993	14.99	105	0.32	28209	84.69
X	55208	34	0.06	12690	22.99	1165	2.11	41319	74.84
Y	2561	46	1.80	1887	73.68	14	0.55	614	23.98
XY	418	0	0.00	49	11.72	2	0.48	367	87.80
MT	256	0	0.00	81	31.64	6	2.34	169	66.02
Overall	2372617	205	0.01	456321	19.23	4619	0.19	1911472	80.56

Table 3: Minor Allele Frequency - CEPH samples and failed SNPs excluded

MAFcutoff	Ndrop	%Drop	Nkeep	%Keep
0.001	456321	19.200	1916091	80.800
0.010	809688	34.100	1562724	65.900
0.050	1095145	46.200	1277267	53.800
0.100	1321988	55.700	1050424	44.300

3 Initial Sample Quality Control

3.1 Sample Call Rates

Figure 4 (p. 11) shows the call rates for all samples, all samples minus CEPH controls, and CEPH controls using all SNPs (excluding chromosome Y). Table 4 (p. 10) shows the number of samples that exceed various call rate exclusion thresholds. Similarily Table 5 (p. 10) shows call rates for all non-CEPH samples, and Table 6 (p. 12) shows call rates for CEPH samples only. For example using a call rate of 95%, 5 samples (1%) will be dropped and using a call rate of 98%, 6 samples (1.2%) will be dropped.

Table 4: Number of Samples Dropped by Call Rate Threshold (Y chromosome excluded) All Samples

cutoff	Ndrop	%Drop	Nkeep	%Keep
0.950	5	1.000	505	99.000
0.980	6	1.200	504	98.800
0.990	8	1.600	502	98.400
0.995	13	2.500	497	97.500
1.000	510	100.000	0	0.000

Table 5: Number of Samples Dropped by Call Rate Threshold (Y chromosome excluded) No CEPH

cutoff	Ndrop	%Drop	Nkeep	%Keep
0.950	5	1.000	489	99.000
0.980	6	1.200	488	98.800
0.990	8	1.600	486	98.400
0.995	13	2.600	481	97.400
1.000	494	100.000	0	0.000

Figure 4: Histogram of Sample Call Rates (Y chromosome excluded)

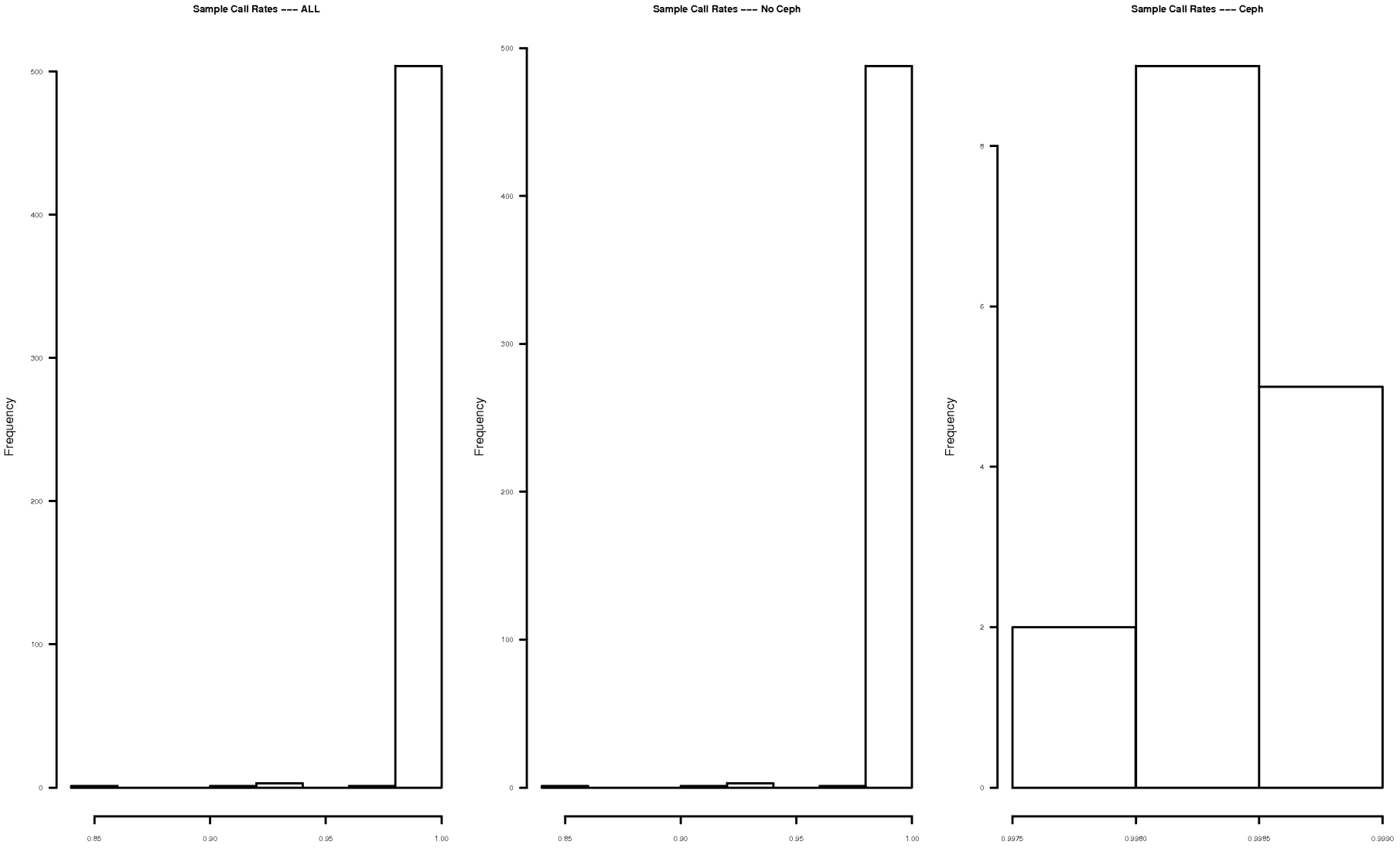


Table 6: Number of Samples Dropped by Call Rate Threshold (Y chromosome excluded) CEPH Only

cutoff	Ndrop	%Drop	Nkeep	%Keep
0.950	0	0.000	16	100.000
0.980	0	0.000	16	100.000
0.990	0	0.000	16	100.000
0.995	0	0.000	16	100.000
1.000	16	100.000	0	0.000

3.2 Sample Sex Check

In this section, information from Chromosomes X and Y is used to estimate sex. Subjects whose reported sex does not match the estimated sex using SNP data are presented in Table 7 (p. 13) with all subjects displayed in Figure 5 (p. 14). Table 7 column descriptions are shown below.

- **PEDSEX**: Recorded sex for this sample (1=Male, 2=Female)
- **SNPSEX**: Sex esimated from Chromosome X variants
- **STATUS**: Displays "PROBLEM" or "OK" for each individual
- **F**: Plink chromosome X inbreeding (homozygosity) estimate
- **No.Ygeno**: Number of SNVs on Chromosome Y
- **cr.chry**: Chromosome Y call rate
- **No.Xgeno**: Number of SNVs on Chromosome X

The expectation is that F is more than 0.8 for Males and less than 0.20 for Females. We would expect cr.chry to be near 1 for Males and near *0 for Females (given the pseudo-autosomal region of Chromosome Y).*

Table 7: Sex Check

IID	FID	PEDSEX	SNPSEX	STATUS	F	No.Ygeno	cr.chry	het.chrx	No.Xgeno

3.3 Sample Heterozygosity

A histogram of the overall heterozygosity per sample is shown in Figure 6. We also analyzed the per-sample heterozygosity by chromosome. In Figure 7 (p. 16), the horizontal dotted red line is the median heterozygosity for all samples.

4 Duplicate Concordance

Table 8: Duplicated Samples

Sample	Number of Replicates	Matched	Mismatch (missing)	Mismatch (called)	Missing (all replicates)	Total SNPs	Concordance
QC1025302437	6	2356459	14102	150	1906	2372617	0.99994
QC1025302436	5	2356085	16002	184	346	2372617	0.99992
QC1025302407	5	2357152	13313	139	2013	2372617	0.99994

This study included 3 samples which were each run multiple times. In Table 8 (p. 13) we look at the number of SNPs whose genotypes:

- matched across all replicates,
- did not match due to missingness in one or more replicates,
- were called differently in the replicates, or
- were missing for all replicates.

Figure 5: Sex assignment verification from Plink. Samples shown in red were flagged as errors by Plink.

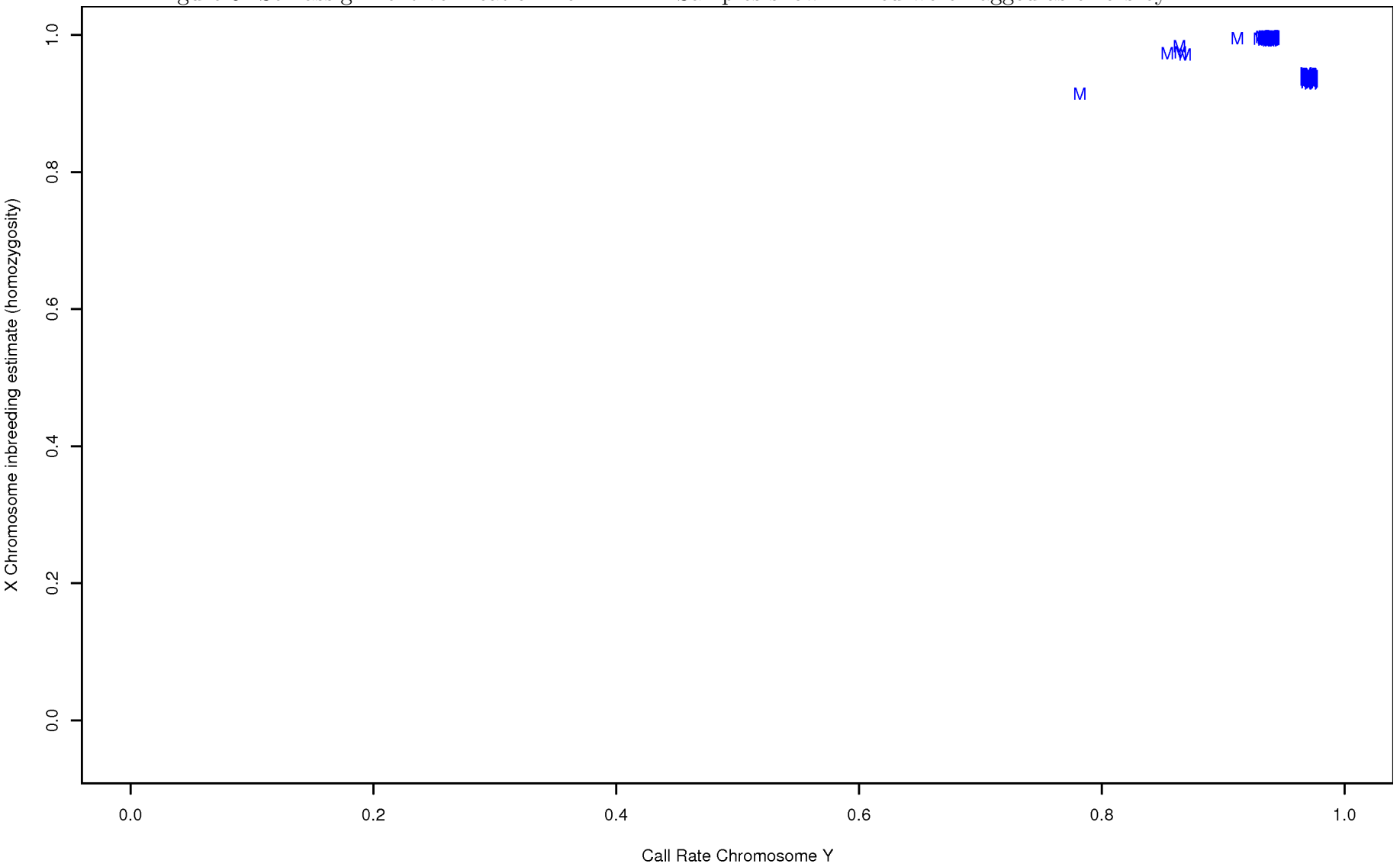


Figure 6: Sample Heterozygosity

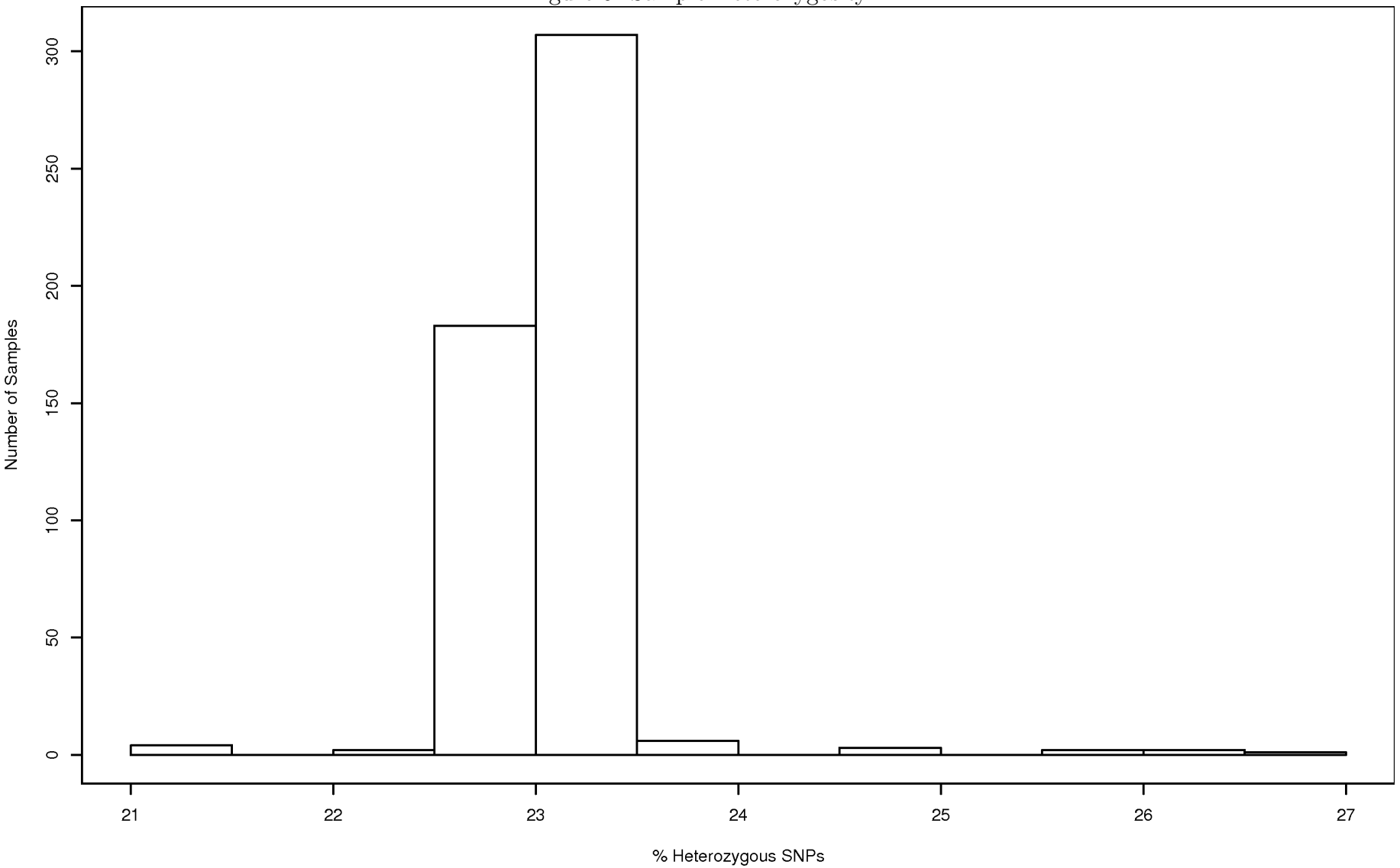
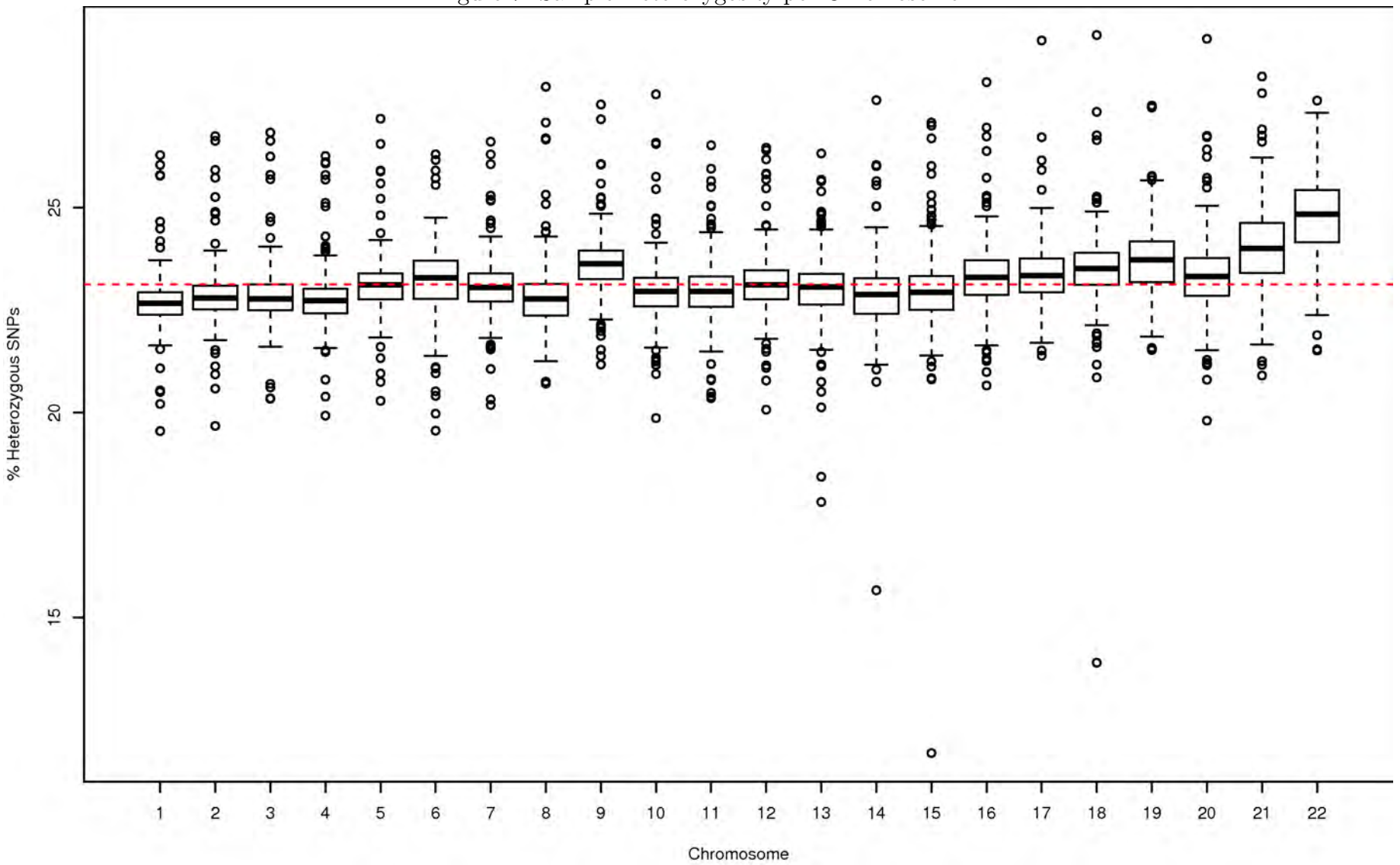


Figure 7: Sample Heterozygosity per Chromosome



Appendix 2

SNP QC report after excluding problematic SNP and problematic samples and includes additional QC tests

EQTL Test Summary

Inv: SNThibodeau
Statistics Team: McDonnell,Kosel
Bioinformatics Team: Asmann,Middha,Hossain

Mayo Clinic College of Medicine, Health Sciences Research
Rochester MN USA

September 14, 2013

Contents

1 Introduction

This document summarizes GWAS QC analysis performed on the HumanOmni2.5-4v1 chip for Prostate Cancer patients. Data are available for 736 samples from 2,372,617 SNPs including 16 CEPH controls. This summary includes data for 510 samples and 2366208 SNPs including 16 controls.

2 Initial SNP Quality Control

2.1 SNP Call Rates

We first look at how many SNPs drop out using different SNP call rate cutoffs. See Table 1 (p. 6) for the percentage of SNPs retained as the call rate threshold increases. Using a call rate of 98%, 22,034 SNPs (0.9%) will be dropped. Using a call rate of 95%, 0 SNPs (0%) will be dropped.

2.2 Failed, Monomorphic, and Low Call Rate SNPs by Chromosome

This section describes how many SNPs failed completely, are "monomorphic", or have a call rate $< 95\%$ by chromosome and overall (Table 2, p. 8). First "failed" SNPs are identified, then "Monomorphic", and finally those SNPs with a call rate $< 0.95\%$. The distribution of SNP call rates by chromosome is presented in Figure 1 (p. 4).

2.3 Minor Allele Frequency

The distribution of minor allele frequencies (MAFs) for all SNPs is shown in Figure 2 (p. 5). There are a total of 454,736 (19.22%) monomorphic SNPs and 807,572 (34.13%) SNPs with MAF $< 1\%$.

2.4 Hardy Weinberg P-value

This dataset does not include controls to reliably test for Hardy-Weinberg Equilibrium so the following results should be interpreted with caution. We include only caucasian subjects resulting in 494 independent subjects. Chromosomes X, Y, XY, and MT markers

Figure 1: SNP Call Rates by Chromosome

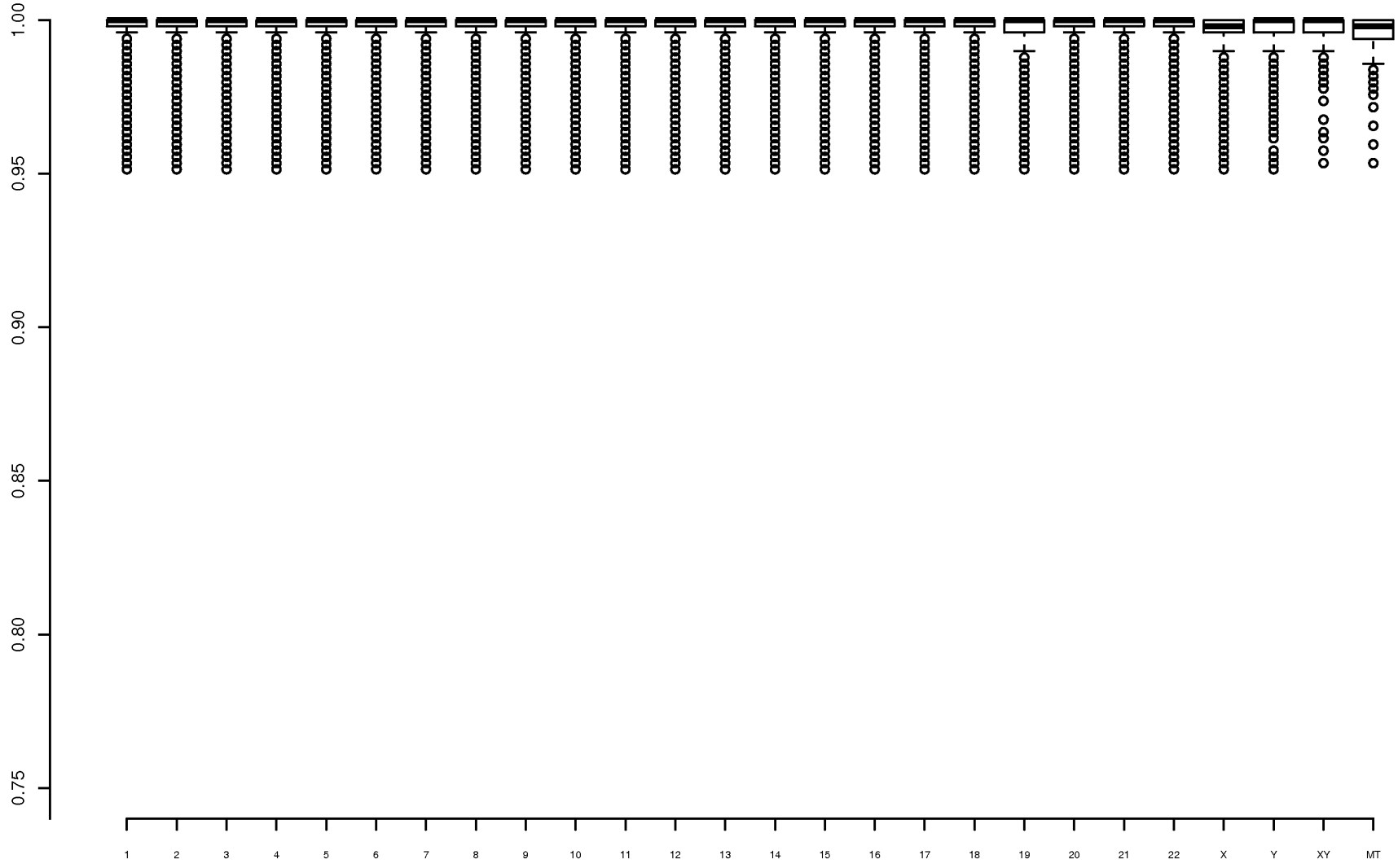


Figure 2: Histogram of Minor Allele Frequencies

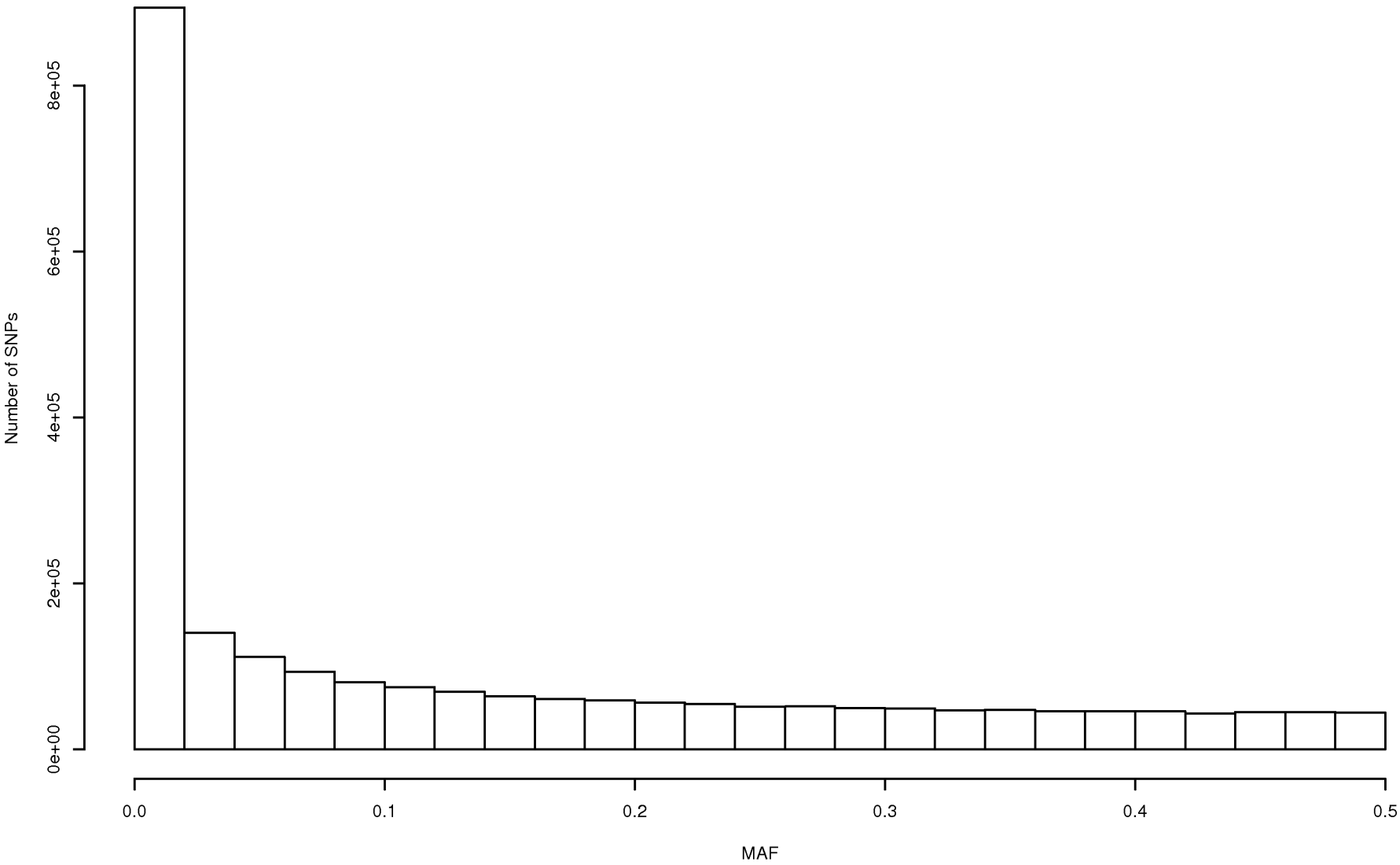


Table 1: SNP Call Rates

CallRate	NumSNPsBelow	%Below	NumSNPsAbove	%Above
0.000	0	0.000	2366208	100.000
0.800	0	0.000	2366208	100.000
0.850	0	0.000	2366208	100.000
0.900	0	0.000	2366208	100.000
0.910	0	0.000	2366208	100.000
0.920	0	0.000	2366208	100.000
0.930	0	0.000	2366208	100.000
0.940	0	0.000	2366208	100.000
0.950	0	0.000	2366208	100.000
0.960	2919	0.100	2363289	99.900
0.970	8216	0.300	2357992	99.700
0.980	22034	0.900	2344174	99.100
0.990	152764	6.500	2213444	93.500
1.000	895070	37.800	1471138	62.200

are excluded from this summary as are SNPs that failed on all samples and SNPs with MAF < 0.05. There are 902 SNPs have a HWE p-value $<$ 10e-05 (see Figure 3, p. 7).

Figure 3: Q-Q plot of HWE p-values (276 p-values have been truncated at 10e-10)

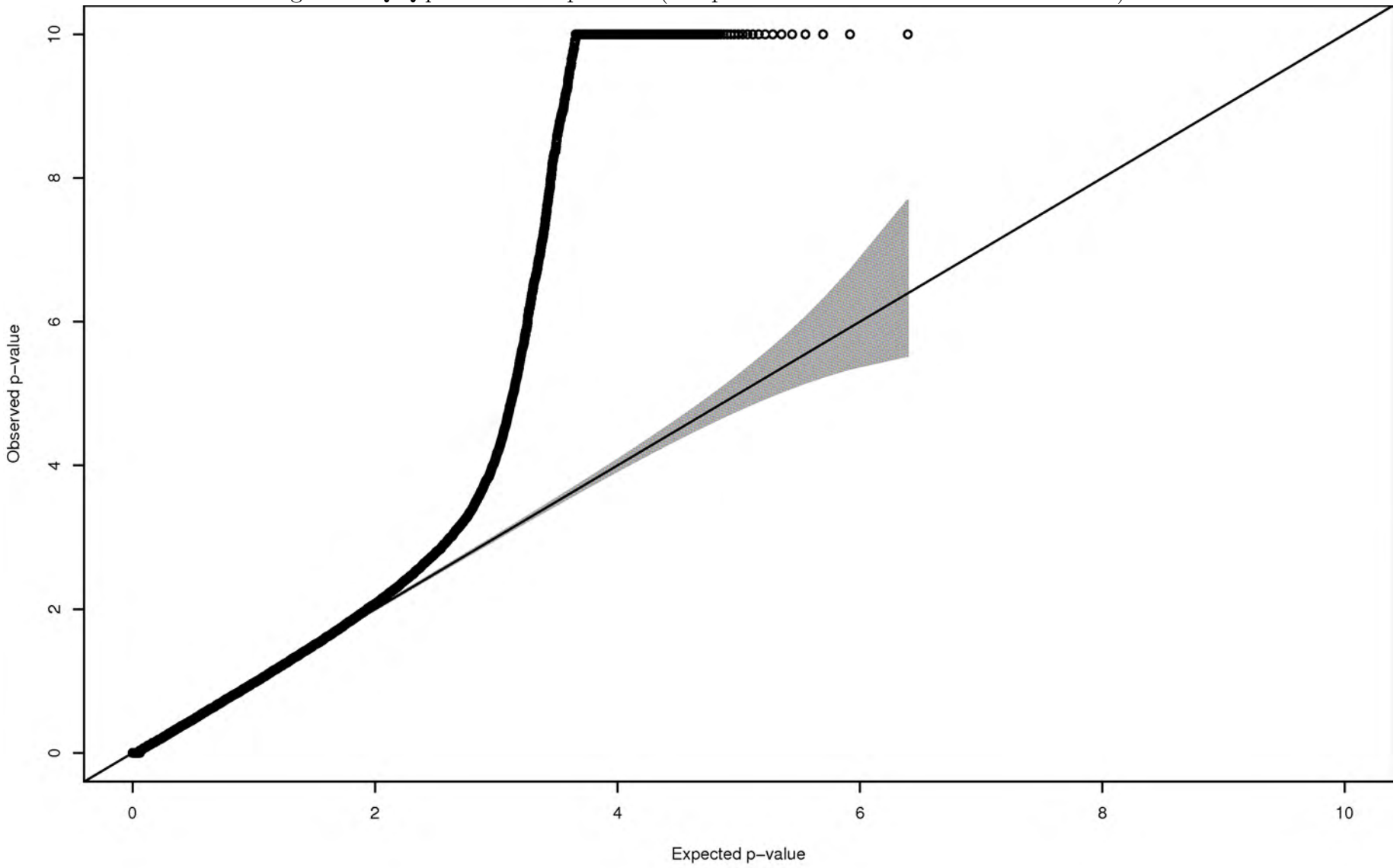


Table 2: SNP QC Summary by Chromosome - CEPH samples excluded

		Failed		Monomorphic		Callrate<0.95		Remaining	
Chrom	TotalSNPs	N	%	N	%	N	%	N	%
1	183728	0	0.00	37327	20.32	0	0.00	146401	79.68
2	193824	0	0.00	38984	20.11	0	0.00	154840	79.89
3	163427	0	0.00	31617	19.35	0	0.00	131810	80.65
4	152609	0	0.00	28952	18.97	0	0.00	123657	81.03
5	145233	0	0.00	29592	20.38	0	0.00	115641	79.62
6	154374	0	0.00	28606	18.53	0	0.00	125768	81.47
7	128819	0	0.00	24607	19.10	0	0.00	104212	80.90
8	125280	0	0.00	23353	18.64	0	0.00	101927	81.36
9	102842	0	0.00	19361	18.83	0	0.00	83481	81.17
10	119219	0	0.00	22806	19.13	0	0.00	96413	80.87
11	115865	0	0.00	23178	20.00	0	0.00	92687	80.00
12	112532	0	0.00	22314	19.83	0	0.00	90218	80.17
13	83353	0	0.00	14926	17.91	0	0.00	68427	82.09
14	76390	0	0.00	14557	19.06	0	0.00	61833	80.94
15	72174	0	0.00	13236	18.34	0	0.00	58938	81.66
16	76447	0	0.00	13527	17.69	0	0.00	62920	82.31
17	66220	0	0.00	12448	18.80	0	0.00	53772	81.20
18	68440	0	0.00	12179	17.80	0	0.00	56261	82.20
19	47589	0	0.00	8777	18.44	0	0.00	38812	81.56
20	56429	0	0.00	10088	17.88	0	0.00	46341	82.12
21	32030	0	0.00	5595	17.47	0	0.00	26435	82.53
22	33196	0	0.00	4987	15.02	0	0.00	28209	84.98
X	53137	0	0.00	11818	22.24	0	0.00	41319	77.76
Y	2386	0	0.00	1772	74.27	0	0.00	614	25.73
XY	416	0	0.00	49	11.78	0	0.00	367	88.22
MT	249	0	0.00	80	32.13	0	0.00	169	67.87
Overall	2366208	0	0.00	454736	19.22	0	0.00	1911472	80.78

Table 3: Minor Allele Frequency - CEPH samples and failed SNPs excluded

MAFcutoff	Ndrop	%Drop	Nkeep	%Keep
0.001	454736	19.200	1911472	80.800
0.010	807572	34.100	1558636	65.900
0.050	1092475	46.200	1273733	53.800
0.100	1318885	55.700	1047323	44.300

3 Initial Sample Quality Control

3.1 Sample Call Rates

Figure 4 (p. 11) shows the call rates for all samples using all SNPs (excluding chromosome Y). Table 4 (p. 10) shows the number of samples that exceed various call rate exclusion thresholds. Similarily Table 5 (p. 10) shows call rates for all non-CEPH samples, and Table 6 (p. 12) shows call rates for CEPH samples only. For example using a call rate of 95%, 5 samples (1%) will be dropped and using a call rate of 98%, 6 samples (1.2%) will be dropped.

Table 4: Number of Samples Dropped by Call Rate Threshold (Y chromosome excluded) All Samples

cutoff	Ndrop	%Drop	Nkeep	%Keep
0.950	5	1.000	489	99.000
0.980	6	1.200	488	98.800
0.990	7	1.400	487	98.600
0.995	12	2.400	482	97.600
1.000	494	100.000	0	0.000

Table 5: Number of Samples Dropped by Call Rate Threshold (Y chromosome excluded) No CEPH

cutoff	Ndrop	%Drop	Nkeep	%Keep
0.950	5	1.000	489	99.000
0.980	6	1.200	488	98.800
0.990	7	1.400	487	98.600
0.995	12	2.400	482	97.600
1.000	494	100.000	0	0.000

Figure 4: Histogram of Sample Call Rates (Y chromosome excluded)

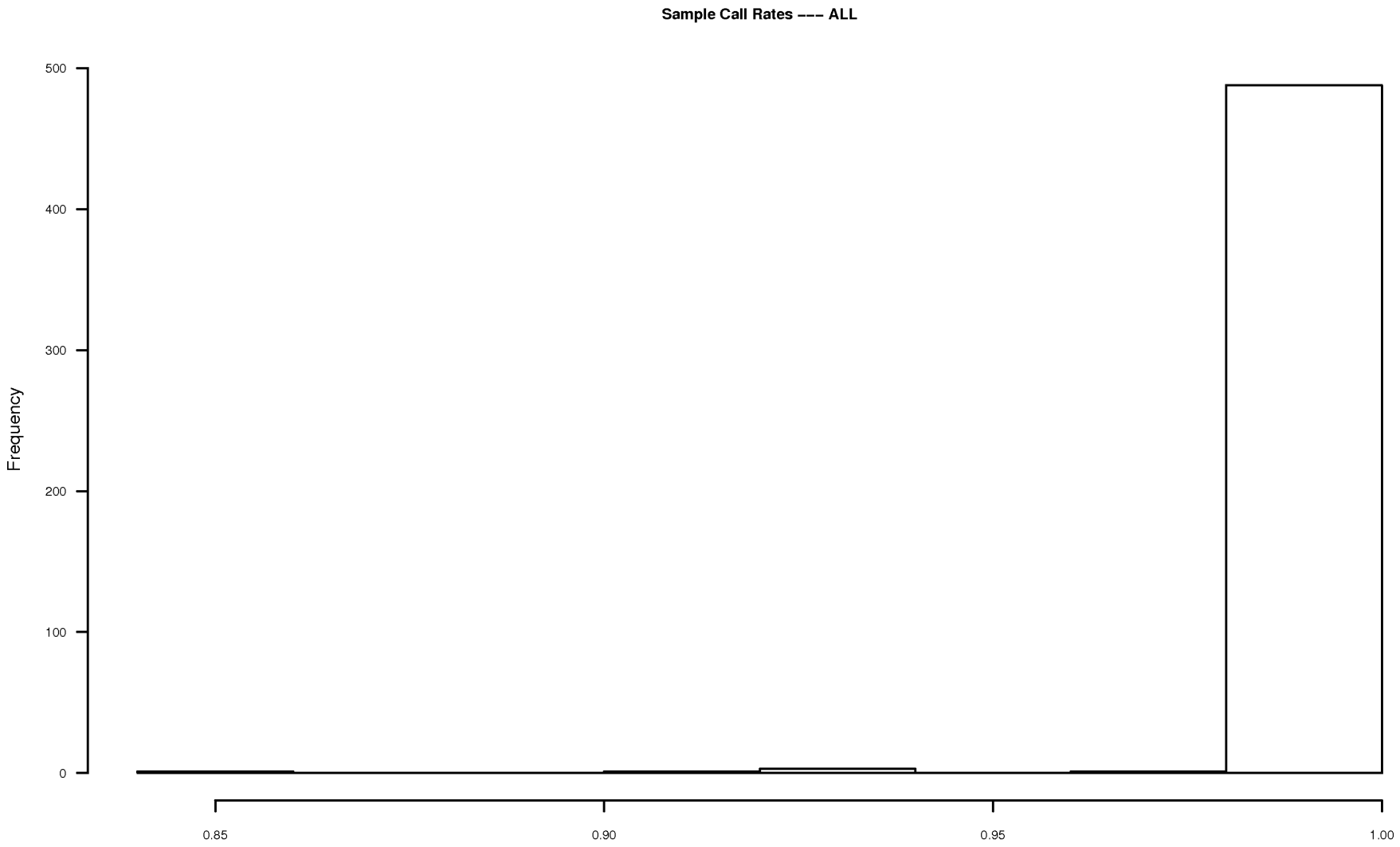


Table 6: Number of Samples Dropped by Call Rate Threshold (Y chromosome excluded) CEPH Only

cutoff	Ndrop	%Drop	Nkeep	%Keep
0.950	0		0	
0.980	0		0	
0.990	0		0	
0.995	0		0	
1.000	0		0	

3.2 Sample Sex Check

In this section, information from Chromosomes X and Y is used to estimate sex. Subjects whose reported sex does not match the estimated sex using SNP data are presented in Table 7 (p. 13) with all subjects displayed in Figure 5 (p. 14). Table 7 column descriptions are shown below.

- **PEDSEX**: Recorded sex for this sample (1=Male, 2=Female)
- **SNPSEX**: Sex esimated from Chromosome X variants
- **STATUS**: Displays "PROBLEM" or "OK" for each individual
- **F**: Plink chromosome X inbreeding (homozygosity) estimate
- **No.Ygeno**: Number of SNVs on Chromosome Y
- **cr.chry**: Chromosome Y call rate
- **No.Xgeno**: Number of SNVs on Chromosome X

The expectation is that F is more than 0.8 for Males and less than 0.20 for Females. We would expect cr.chry to be near 1 for Males and near *0 for Females (given the pseudo-autosomal region of Chromosome Y).*

Table 7: Sex Check

IID	FID	PEDSEX	SNPSEX	STATUS	F	No.Ygeno	cr.chry	het.chrx	No.Xgeno

3.3 Sample Heterozygosity

A histogram of the overall heterozygosity per sample is shown in Figure 6. We also analyzed the per-sample heterozygosity by chromosome. In Figure 7 (p. 16), the horizontal dotted red line is the median heterozygosity for all samples.

4 Batch Effects

Table 8: Plate Mapping

WG0232831-DNA	1
WG0232832-DNA	2
WG0232833-DNA	3
WG0232834-DNA	4
WG0232835-DNA	5
WG0232836-DNA	6
WG0232837-DNA	7
WG0232838-DNA	8

Table 8 (p. 13) will act as map for the following batch effect plots regarding Plate. To test for Plate effects in variant calling, we performed a chi-squared test for each SNP comparing the allele frequency estimated using samples on one Plate to the allele frequency estimated from the remaining Plates. We then took the mean of the chi-squared statistics for each Plate across all SNPs. The numbers in the plot (Figure 8) (p. 17) indicates Plate. Figure 9 (p. 18) shows boxplots of the sample call rate for each Plate. The dashed horizontal line is drawn at the 98% percentile of missingness rates for the SNPs used in the figure. Figure 10 (p. 19) shows boxplots of the sample heterozygosity rate for each Plate. The dashed horizontal line is drawn at the median heterozygosity rate across samples.

Figure 5: Sex assignment verification from Plink. Samples shown in red were flagged as errors by Plink.

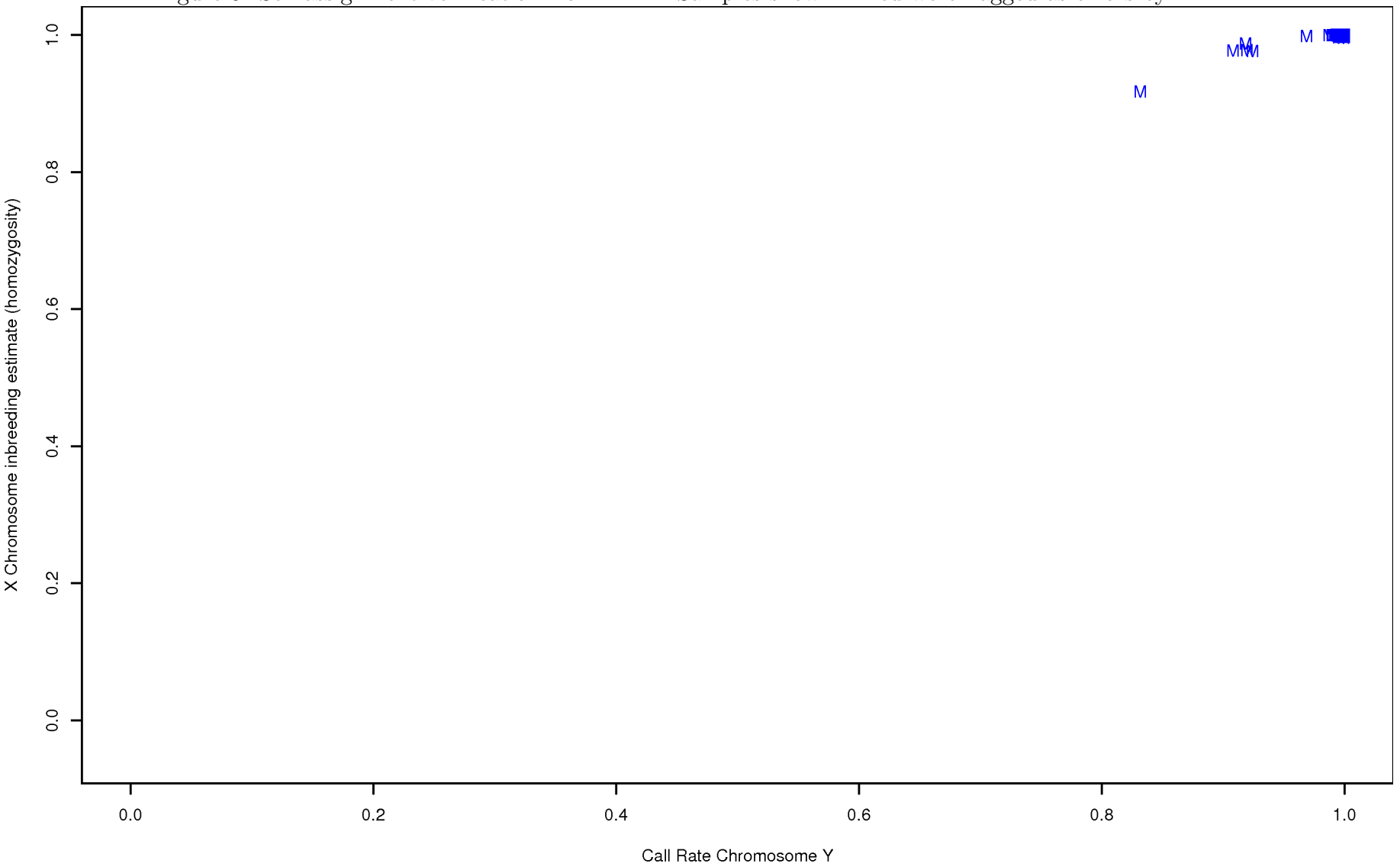


Figure 6: Sample Heterozygosity

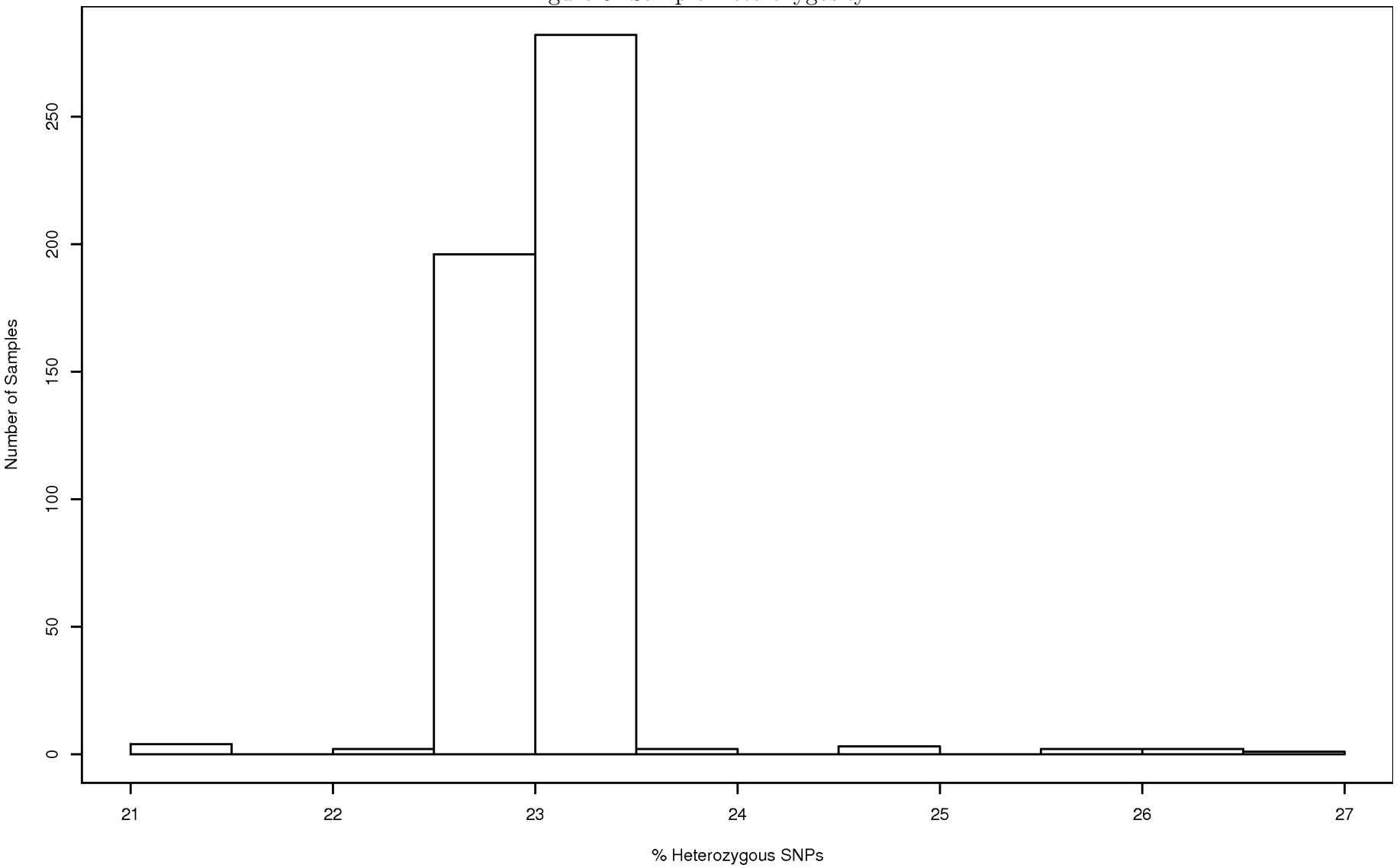


Figure 7: Sample Heterozygosity per Chromosome

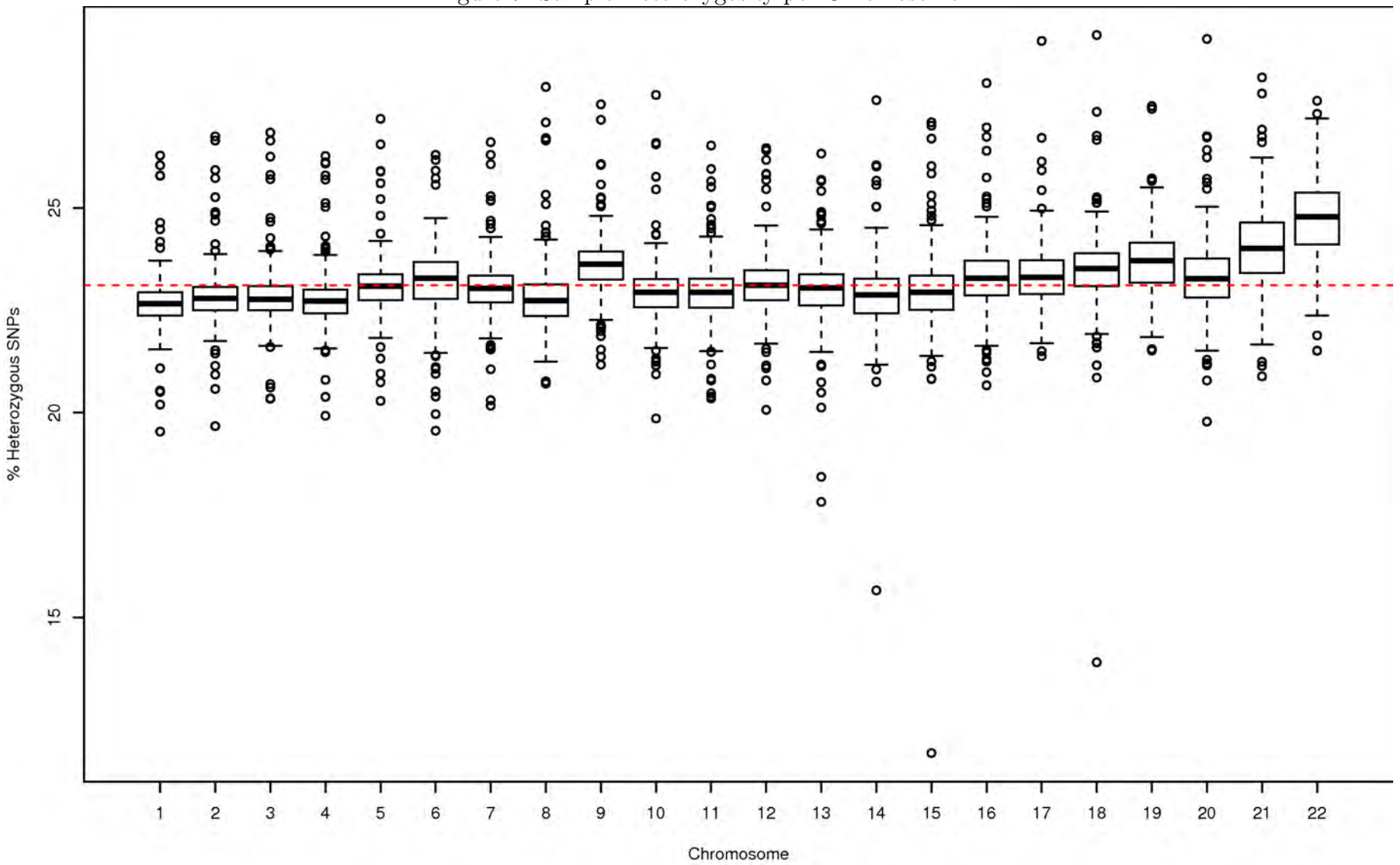


Figure 8: Test for Batch Effects

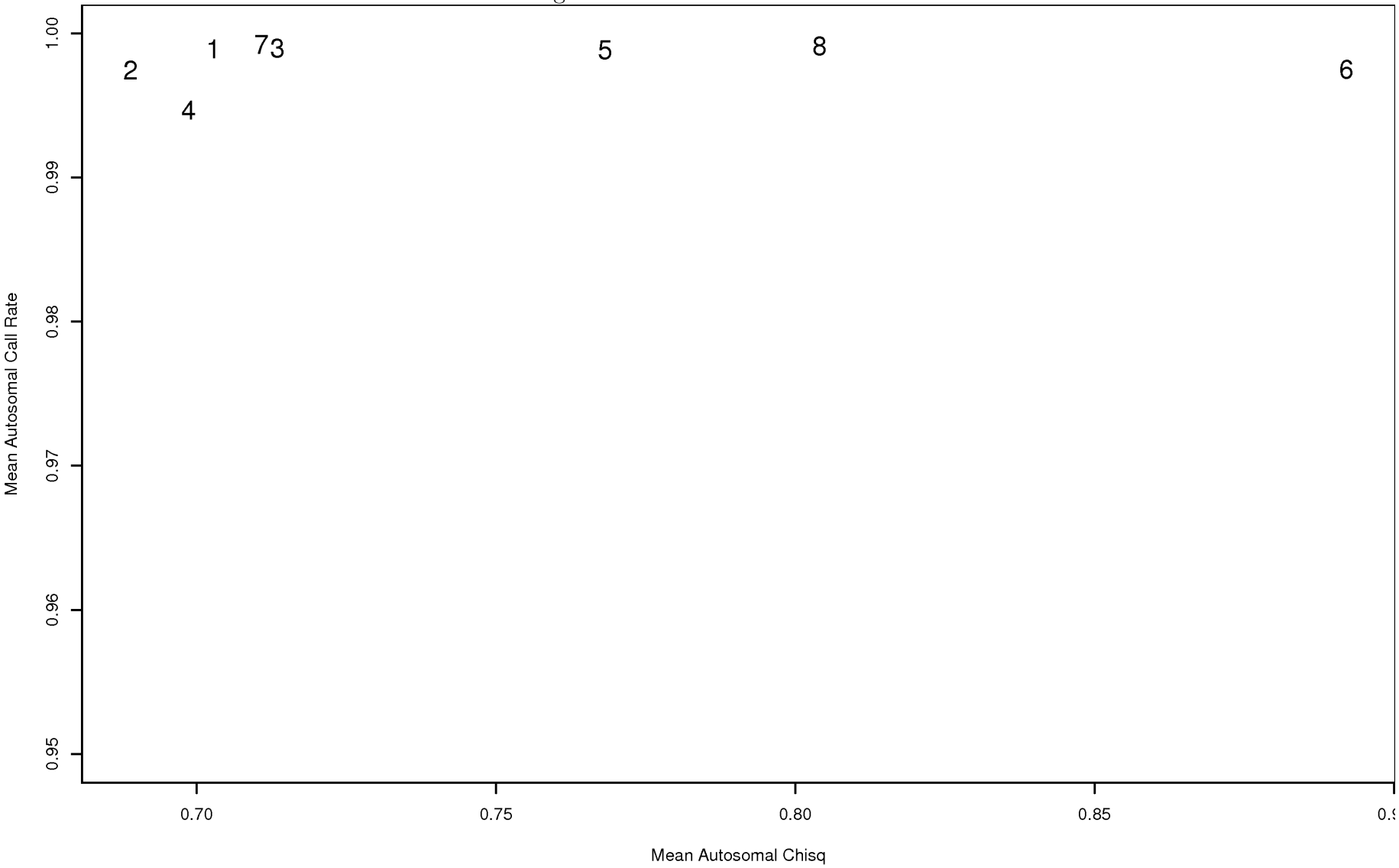


Figure 9: Sample Call Rate by Plate

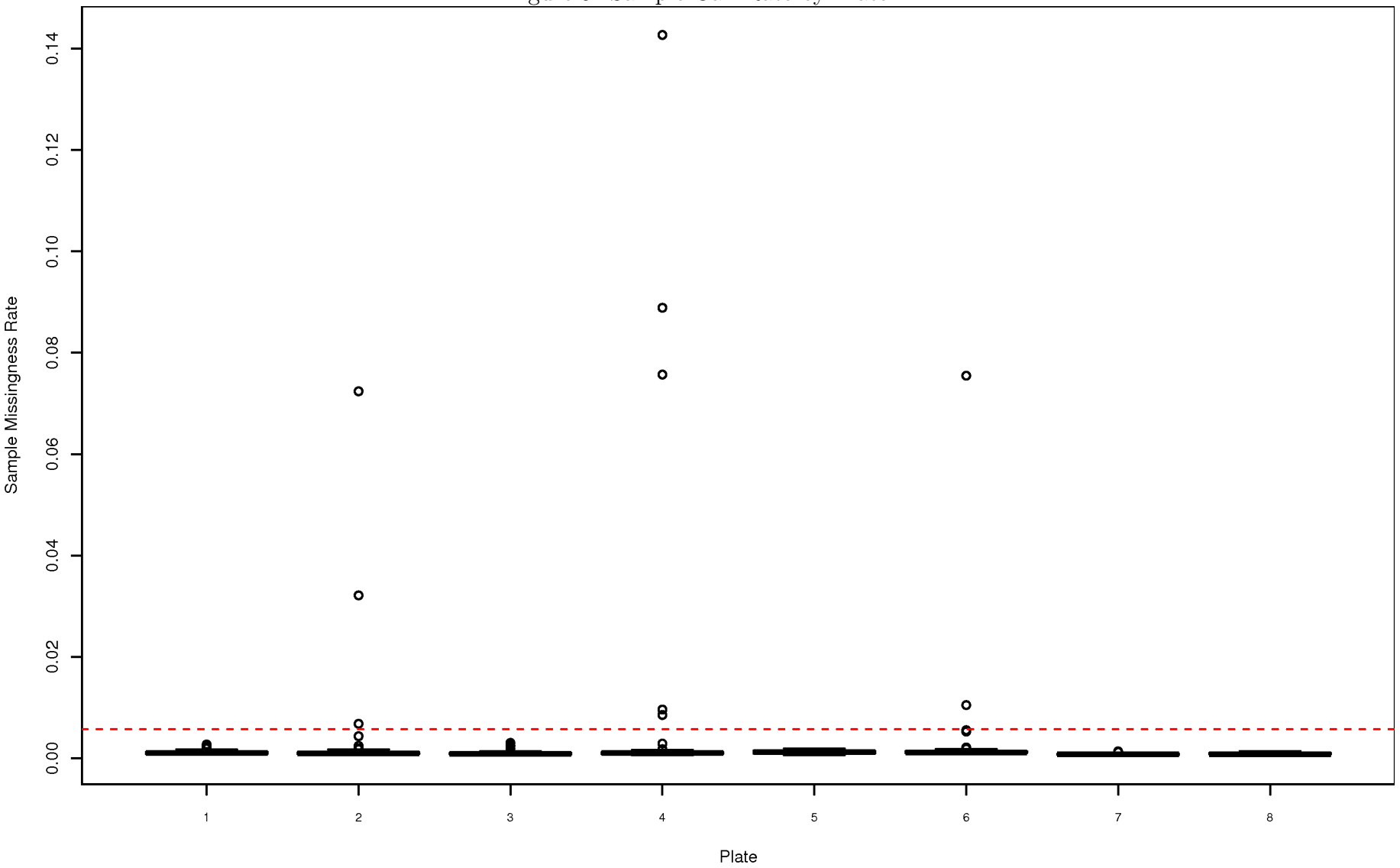
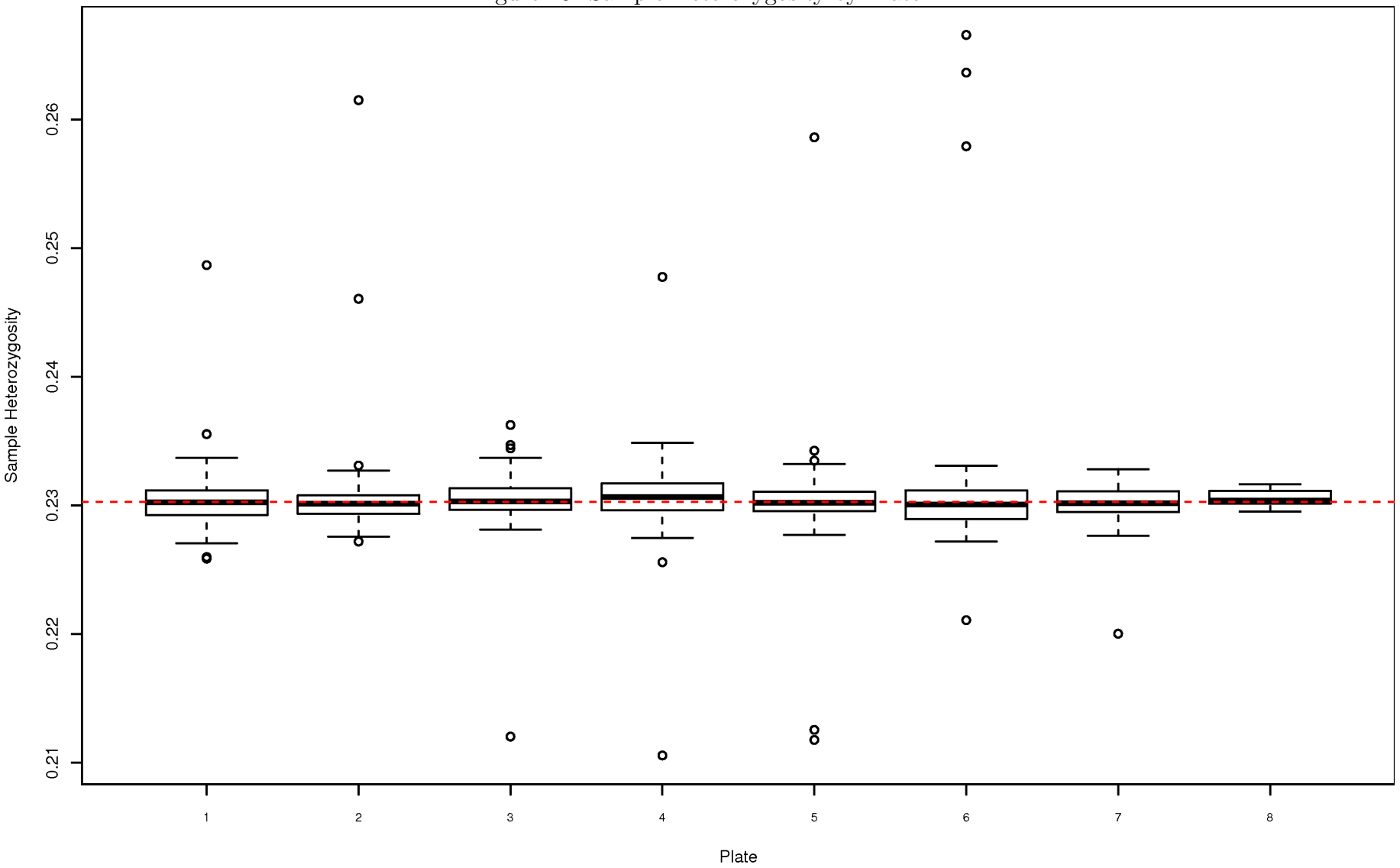


Figure 10: Sample Heterozygosity by Plate



5 PLINK Relationship Checking

This study consists of 494 presumed unrelated individuals. Relationship checking was performed by estimating the proportion of alleles shared identical by descent (IBD) for all pairs of subjects. PLINK was used to estimate IBD. Independent SNPs were selected for analysis by first excluding all SNPs with callrate < 0.95%, MAF < 0.05%, and HWE pvalue < 1e-06. Remaining SNPs were pruned using Plink such that pairwise correlation between SNPs (r2) is less than 0.01. A total of 21395 were used for this analysis. Figure 11 (p. 22) shows the IBD plot for all study samples. If this study includes both related and unrelated samples, then panel A shows the unrelated samples and panel B shows related samples. Relationship codes shown in Figure 11 along with their expected IBD sharing are shown below.

```
CODE     RELATIONSHIP          E(IBD0)   E(IBD1)  E(IBD2)
-----------------------------------------------------------
 PO   : Parent-Offspring       0         1.00     0
 FS   : Full-Sibling           0.25      0.50     0.25
 HS   : Half-Sibling           0.50      0.50     0
 AV   : Avuncular              0.50      0.50     0
 GPC  : Grandparent-grandchild 0.50      0.50     0
 FC   : First-Cousin           0.75      0.25     0
 HA   : Half-Avuncular         0.75      0.25     0
 HFC  : Half-First-Cousin      0.875     0.125    0
 HSFC: Half-Sib+First-Cousin   0.375     0.50     0.125
 U    : Unrelated              1.00      0        0
```

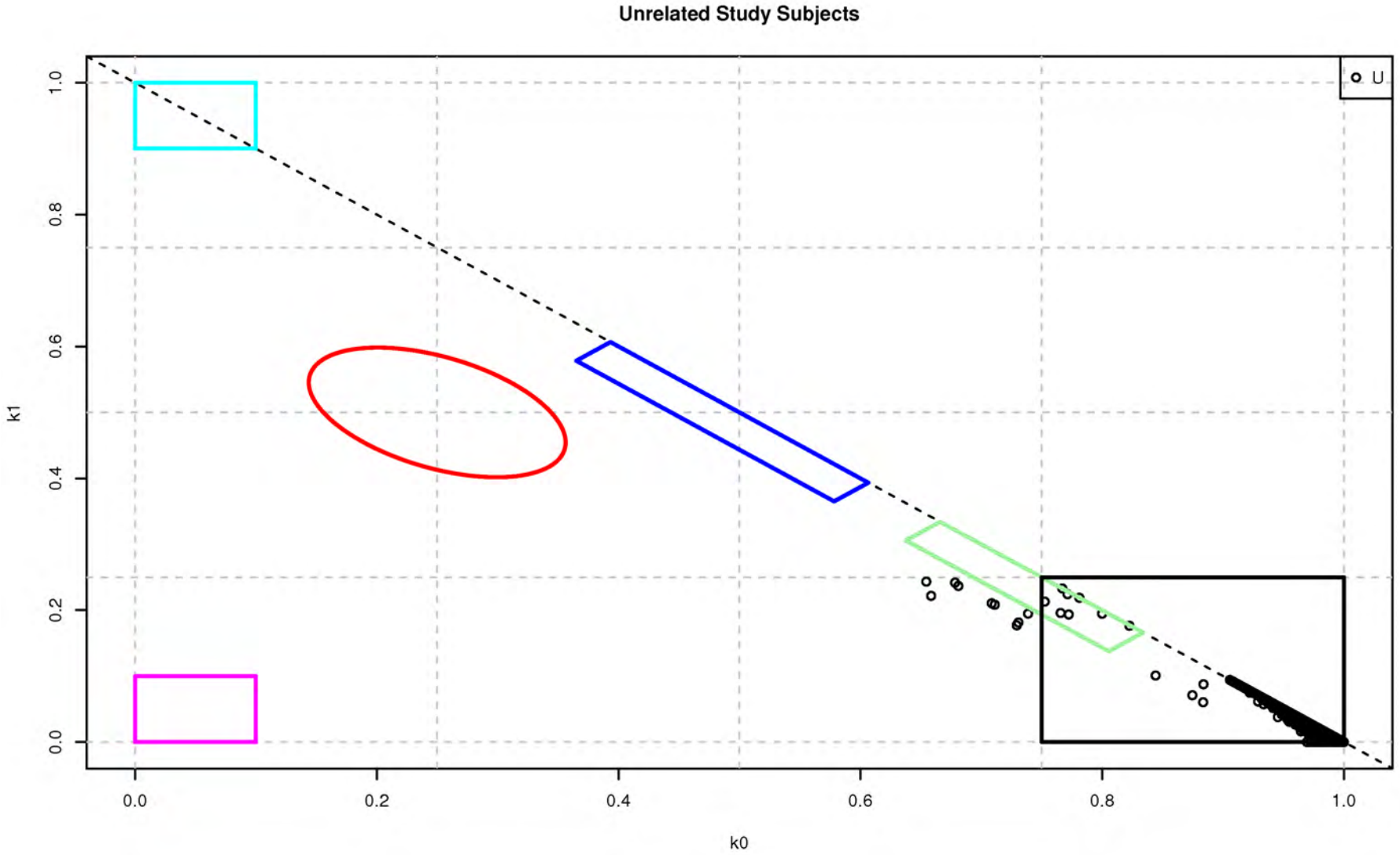
Table 9: Check for Cryptic relatedness: Unrelated pairs

FID1	IID1	FID2	IID2	Z0	Z1	Z2	PI_HAT	RT	Obs.RT
1213802311	1213802311	1211702138	1211702138	0.7714	0.2243	0.0044	0.1165	U	FC
1213802311	1213802311	1211001831	1211001831	0.7812	0.2188	0.0000	0.1094	U	FC
1213802218	1213802218	1211702092	1211702092	0.7671	0.2329	0.0000	0.1164	U	FC
1211800763	1211800763	1211702138	1211702138	0.7087	0.2105	0.0808	0.1861	U	Q

1211800763	1211800763	1213802245	1213802245	0.7112	0.2083	0.0805	0.1846	U	Q
1211800763	1211800763	1211001831	1211001831	0.7308	0.1815	0.0876	0.1784	U	Q
1211702138	1211702138	1213802245	1213802245	0.7294	0.1771	0.0935	0.1820	U	Q
1211702138	1211702138	1211001831	1211001831	0.6546	0.2433	0.1021	0.2237	U	Q
1211001818	1211001818	1211800765	1211800765	0.6586	0.2218	0.1196	0.2305	U	Q
1211001818	1211001818	1211702155	1211702155	0.6811	0.2368	0.0821	0.2005	U	Q
1211001818	1211001818	1213103091	1213103091	0.7526	0.2130	0.0345	0.1409	U	FC
1213802245	1213802245	1211001831	1211001831	0.7388	0.1948	0.0665	0.1639	U	Q
1211800765	1211800765	1211702155	1211702155	0.6784	0.2418	0.0799	0.2007	U	Q
1211800765	1211800765	1213103091	1213103091	0.7724	0.1935	0.0340	0.1308	U	FC
1211702155	1211702155	1213103091	1213103091	0.7657	0.1958	0.0385	0.1364	U	FC

All pairs of unrelated subjects with the probability of sharing 0 alleles IBD < 0.80 are shown in Table 9 (p. 21). There are 15 pairs of unrelated subjects who have higher than expected IBD sharing. Related pairs whose IBD sharing does not match expected are shown in Table **??** (p. **??**). All relative pairs where the absolute value of expected minus observed sharing is greater than 0.25 for any of the IBD sharing probabilities is included. These tables includes both the expected relationship type (column labelled 'RT') and the observed relationship type based on estimated IBD probabilities (column labelled '$Obs.RT$'). There are 0 pairs of related subjects whose relationships appear to be different than expected. Relationship codes shown in these tables are described on page 20.

Figure 11: Estimated IBD sharing between all pairs of subjects. If study includes pedigrees, then the IBD sharing is split into two panels: Panel A includes all unrelated pairs of subjects and Panel B includes all related pairs within pedigrees. Each relationship is displayed in a different symbol and color. Relationship codes are described on page 20.



Thibodeau eQTL mRNA NGS QC

Inv: Thibodeau
Statistical Team: Schaid, McDonnell, Riska, Fogarty

September 17, 2013

Contents

1 Introduction

This document describes the mRNA-seq quality control checks and initial analysis performed for the "Thibodeau eQTL mRNA NGS QC" project. A total of 493 subjects contributed 493 samples consisting of N=19 cystoprostatectomy samples, N=474 low gleason samples. 493 subject(s) gave 1 samples. There are 0 repeated samples (). Samples were run up to 5 per lane, with the groupings listed in Table 1.

There were 23,398 Genes presented in the original data (46 Genes mapped to 2 different chromosomes and 3 Genes mapped to 3 different chromosomes). Of all the genes, 780 (3.3%) had no counts for all samples and were removed from further analysis (genes deemed undetectable/noise). The remaining genes were distributed across all the chromosomes (Table 2). For genes that mapped to both chromosome X and Y, only the chromosome X version was retained. After filtering, there was only 3 gene (FAM45B, MIR1256, TTL) mapped to more than 1 location (chr10, chrX, chr10, chrX, chr13, chr2). Additionally, there were still 37 Genes that mapped to chromosome Y (AMELY, BCORP1, CD24, CSPG4P1Y, DDX3Y, EIF1AY, GYG2P1, KDM5D, LINC00230A, NCRNA00185, NLGN4Y, PCDH11Y, PRKY, RBMY1A3P, RBMY2EP, RBMY2FP, RPS4Y1, RPS4Y2, SRY, TBL1Y, TMSB4Y, TSPY1, TSPY2, TTTY10, TTTY12, TTTY13, TTTY14, TTTY15, TTTY16, TTTY18, TTTY19, TTTY22, TTTY5, TXLNG2P, USP9Y, UTY, ZFY).

Flowcell	Run.Name	Subjects	N
1	121112_SN7001166_0111_BD1KD4ACXX	s_10,s_114,s_142,s_202,s_21,s_23,s_280,s_313 s_341,s_344,s_360,s_378,s_435,s_449,s_452,s_459 s_471,s_501,s_511,s_547,s_61	21
2	121112_SN7001166_0111_BD1KD4ACXX_2	s_549,s_87	2
3	121116_SN725_0269_BD1KC5ACXX	s_104,s_141,s_172,s_176,s_224,s_354,s_375,s_392 s_398,s_405,s_410,s_414,s_42,s_432,s_450,s_453 s_504,s_506,s_516,s_539,s_65,s_80	22
4	121120_SN414_0250_AC1F36ACXX	s_11,s_110,s_12,s_173,s_196,s_238,s_35,s_394 s_404,s_422,s_423,s_438,s_444,s_451,s_472,s_479 s_532,s_536	18
5	121120_SN414_0251_BD1KDGACXX	s_106,s_160,s_165,s_169,s_217,s_218,s_239,s_24 s_246,s_249,s_258,s_301,s_339,s_355,s_36,s_370 s_400,s_419,s_443,s_478,s_486,s_497,s_510,s_527	24
6	121128_SN7001166_0114_AD1K24ACXX	s_133,s_163,s_166,s_187,s_198,s_226,s_27,s_270 s_274,s_276,s_286,s_304,s_307,s_314,s_324,s_383 s_41,s_437,s_474,s_492,s_509,s_541,s_546,s_77 s_9,s_95,s_96,s_98	28
7	121129_SN616_0231_AC1GC0ACXX	s_126,s_145,s_155,s_182,s_194,s_260,s_272,s_275 s_279,s_285,s_288,s_321,s_34,s_372,s_441,s_446 s_447,s_477,s_483,s_507,s_553,s_556,s_70	23
8	121129_SN616_0232_BD1K1UACXX	s_167,s_241,s_338,s_365,s_476,s_498,s_62,s_86	8
9	121130_SN414_0256_AD1M44ACXX	s_1,s_119,s_153,s_156,s_157,s_266,s_268,s_31 s_343,s_348,s_367,s_4,s_402,s_408,s_465,s_484 s_519,s_525,s_551,s_558,s_60,s_76,s_78,s_82	24
10	121205_SN725_0272_AC1H54ACXX	s_105,s_118,s_137,s_140,s_147,s_168,s_181,s_183 s_191,s_2,s_232,s_264,s_294,s_333,s_352,s_387 s_388,s_393,s_417,s_448,s_488,s_49,s_496,s_50 s_512,s_562	26
11	121205_SN725_0273_BD1M9VACXX	s_152,s_171,s_178,s_210,s_25,s_269,s_287,s_337 s_347,s_366,s_377,s_440,s_467,s_482,s_490,s_534 s_538,s_542,s_59,s_84,s_89,s_91,s_94	23

Flowcell	Run.Name	Subjects	N
12	121213_SN725_0275_BC1GGBACXX	s_100,s_101,s_109,s_113,s_121,s_125,s_13,s_134 s_144,s_17,s_185,s_195,s_243,s_326,s_340,s_380 s_409,s_413,s_43,s_458,s_466,s_475,s_480,s_5 s_505,s_522,s_530,s_79,s_81,s_97	30
13	121214_SN7001166_0118_AD1LW9ACXX	s_131,s_15,s_158,s_177,s_19,s_193,s_253,s_259 s_319,s_32,s_33,s_373,s_382,s_397,s_407,s_421 s_425,s_461,s_513,s_550,s_7,s_75	22
14	121214_SN7001166_0119_BD1M77ACXX	s_123,s_129,s_235,s_282,s_316,s_346,s_357,s_386 s_390,s_395,s_468,s_52,s_535,s_555,s_63	15
15	121218_SN616_0237_AD1M5BACXX	s_115,s_116,s_151,s_18,s_180,s_205,s_255,s_257 s_290,s_293,s_317,s_318,s_359,s_368,s_412,s_415 s_427,s_442,s_45,s_469,s_47,s_515,s_526,s_548 s_56,s_68,s_85	27
16	130104_SN7001166_0126_AC1MU4ACXX	s_111,s_135,s_149,s_174,s_209,s_215,s_221,s_229 s_278,s_30,s_308,s_310,s_315,s_363,s_364,s_385 s_396,s_406,s_481,s_489,s_491,s_493,s_495,s_514 s_518,s_528,s_537,s_543,s_545,s_57,s_64,s_69 s_92	33
17	130104_SN7001166_0127_BC1N0KACXX	s_102,s_112,s_122,s_124,s_132,s_138,s_143,s_199 s_22,s_234,s_320,s_327,s_329,s_369,s_381,s_39 s_403,s_416,s_44,s_46	20
18	130104_SN7001166_0127_BC1N0KACXX_2	s_533	1
19	130111_SN7001166_0128_AD1NCWACXX	s_161,s_291,s_349,s_433,s_434,s_456,s_503,s_53	8
20	130125_SN316_0280_BC1KPWACXX	s_148,s_162,s_170,s_201,s_216,s_263,s_38,s_384 s_40,s_430,s_485,s_6,s_72,s_74,s_93	15
21	MERGE_3_28_2013-1	s_108,s_117,s_127,s_128,s_136,s_16,s_184,s_186 s_188,s_189,s_203,s_206,s_212,s_213,s_227,s_233 s_247,s_254,s_261,s_265,s_267	21
22	MERGE_3_28_2013-2	s_28,s_281,s_306,s_311,s_312,s_323,s_325,s_328 s_330,s_336,s_345,s_350,s_351,s_361,s_362,s_374 s_376,s_391,s_401,s_424,s_426,s_428,s_439,s_460	33

Flowcell	Run.Name	Subjects	N
		s_464,s_499,s_517,s_554,s_565,s_71,s_8,s_83	
		s_99	
23	MERGE_3_28_2013-3	s_120,s_150,s_164,s_190,s_192,s_197,s_200,s_208	10
		s_214,s_228	
24	MERGE_3_28_2013-4	s_231,s_237,s_242,s_245,s_248,s_250,s_252,s_256	38
		s_26,s_271,s_273,s_277,s_283,s_289,s_295,s_297	
		s_298,s_322,s_332,s_342,s_358,s_389,s_411,s_418	
		s_420,s_431,s_445,s_455,s_457,s_463,s_470,s_473	
		s_523,s_524,s_55,s_557,s_58,s_88	
25	MERGE_3_28_2013-5	s_3	1

Table 1: Samples in each Flowcell

chr01	chr02	chr03	chr04	chr05	chr06	chr07	chr08	chr09
2279	1447	1226	854	993	1184	1086	780	925
chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18
881	1405	1152	385	759	770	916	1326	319
chr19	chr20	chr21	chr22	chrX	chrY			
1535	644	286	528	881	37			

Table 2: Chromosome distribution of Genes

Summaries of the log_2 (counts) and $\%counts > 0$ by subject, by flowcell, by group, by $\%GCcontent$, and by gene size (counting only the sum of the exons) are included in the following sections. These factors can influence then number of counts observed

2 Assessing log_2 (Gene Counts)

2.1 By Subject and Lane

Figure 1 shows the distribution of Gene Counts separately for each subject via boxplots. The plots are color-coded to indicate tumor type. Because the values are presented on a log_2 scale, the Gene Counts is actually the Gene Counts + 1 so that those genes with a count of zero are also included in the figure. Figure 2 and 3 to 27 shows the same subjects, but this time the boxes are color-coded by RunID. The hope is that the boxplots are relatively consistent across all the subjects. Figure 28 to 52 shows the distribution of gene counts via line graph. Figure 53 shows, for each subject, the sum of all the Gene Counts. Lines are used to separate subjects by RunID. The red line in the middle of the dots is the median of each RunID.

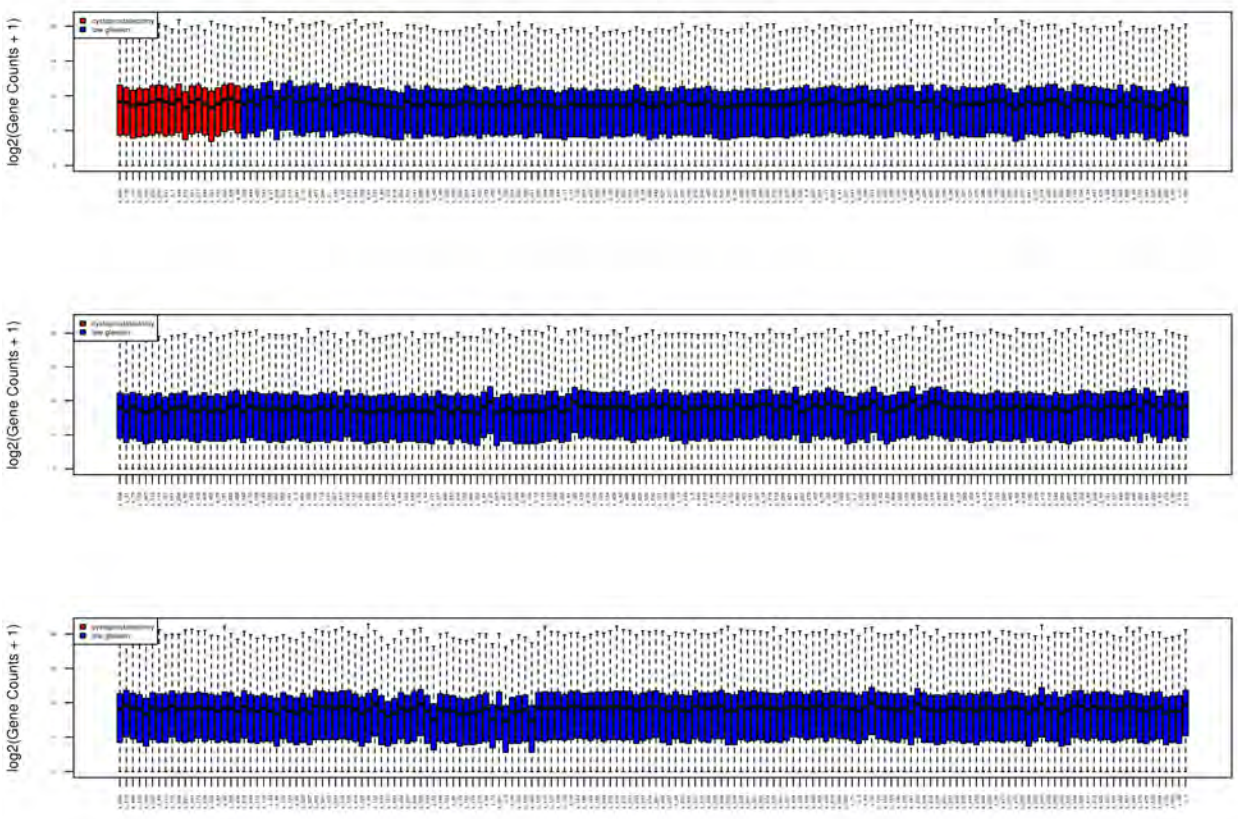


Figure 1: Distribution of log2(Gene Counts) for each Subject color -coded by Group

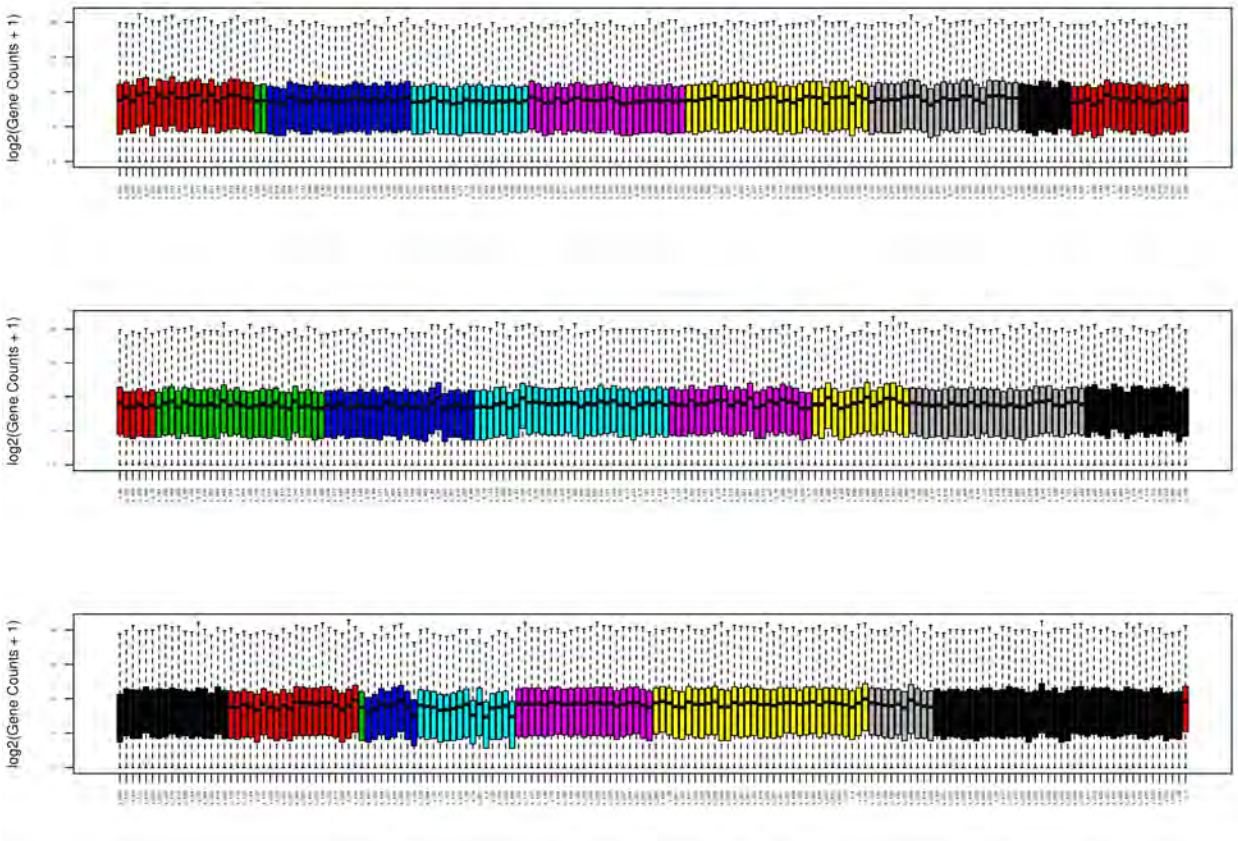


Figure 2: Distribution of log2(Gene Counts) for each Subject color -coded by RunID

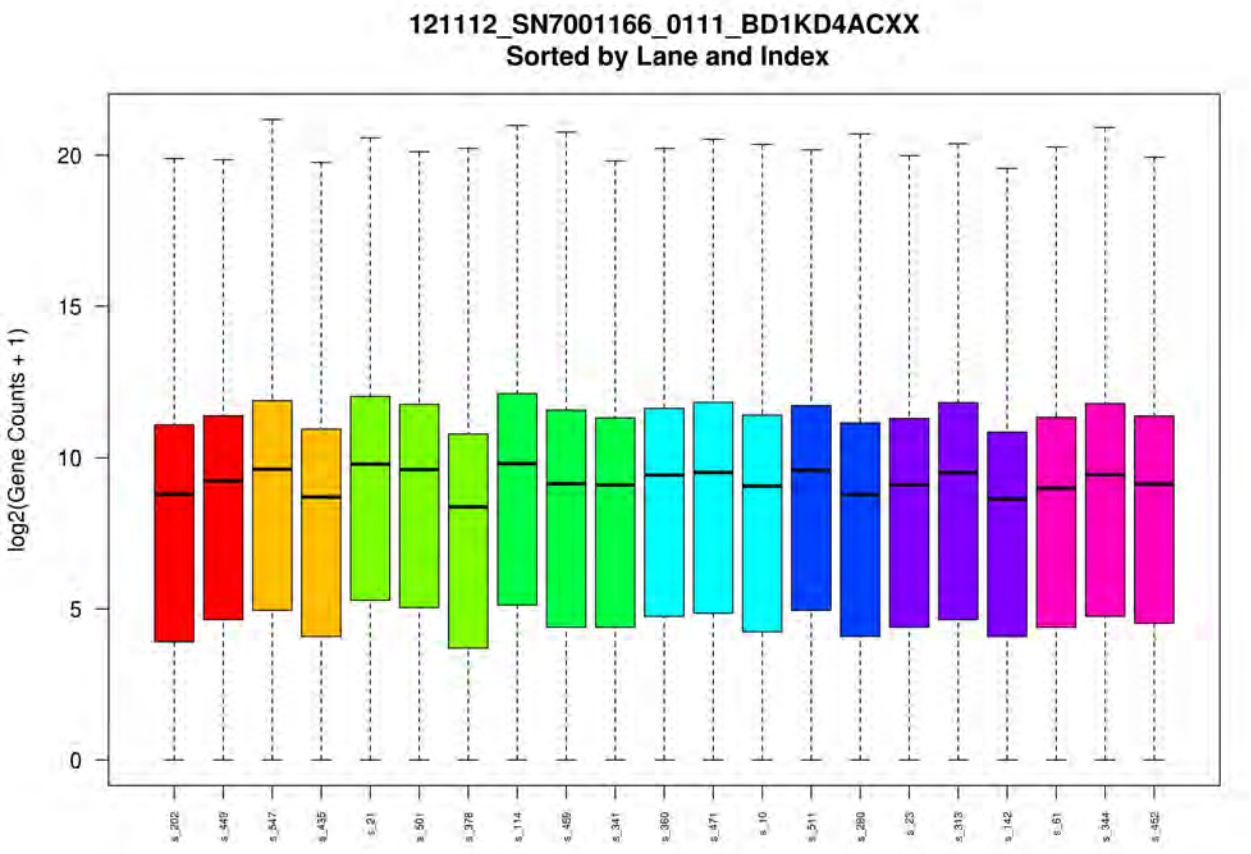


Figure 3: Distribution of log2(Total Gene Counts) for each Subject by RunID

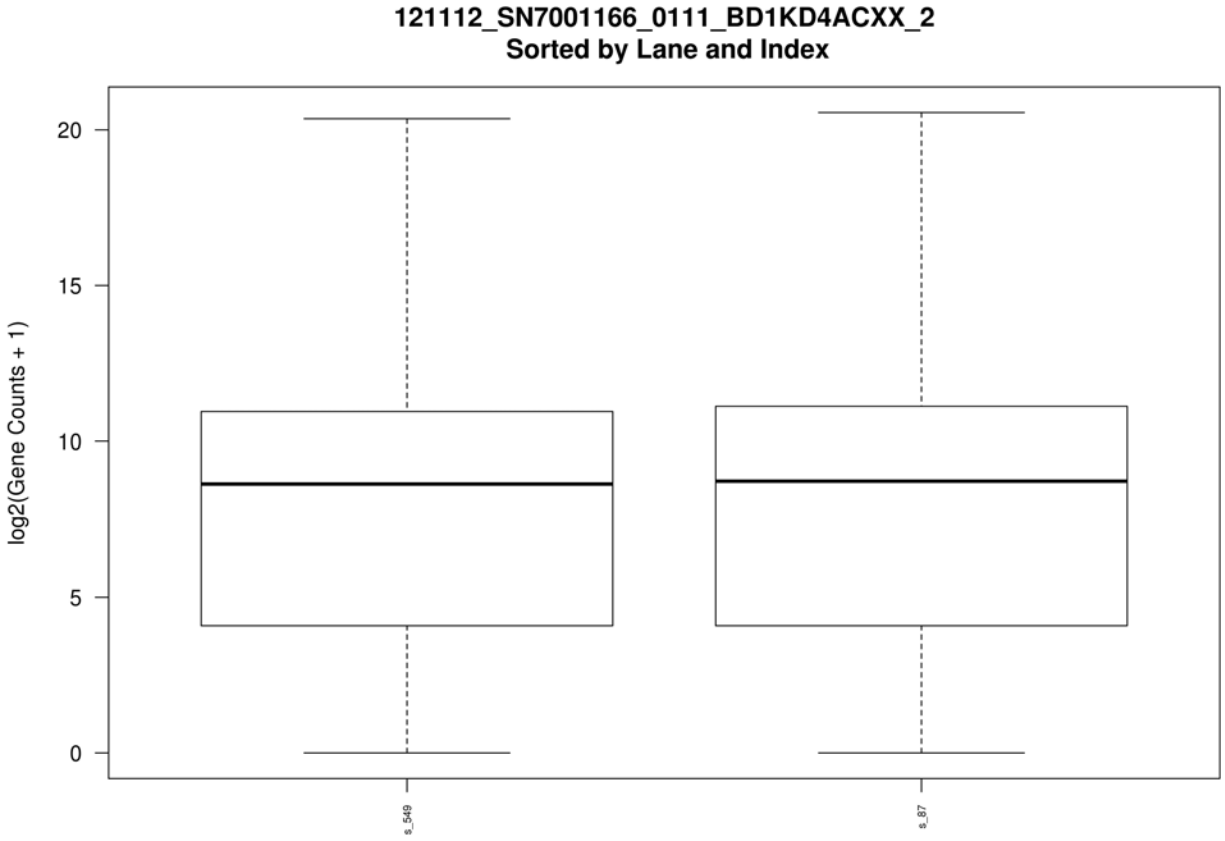


Figure 4: Distribution of log2(Total Gene Counts) for each Subject by RunID

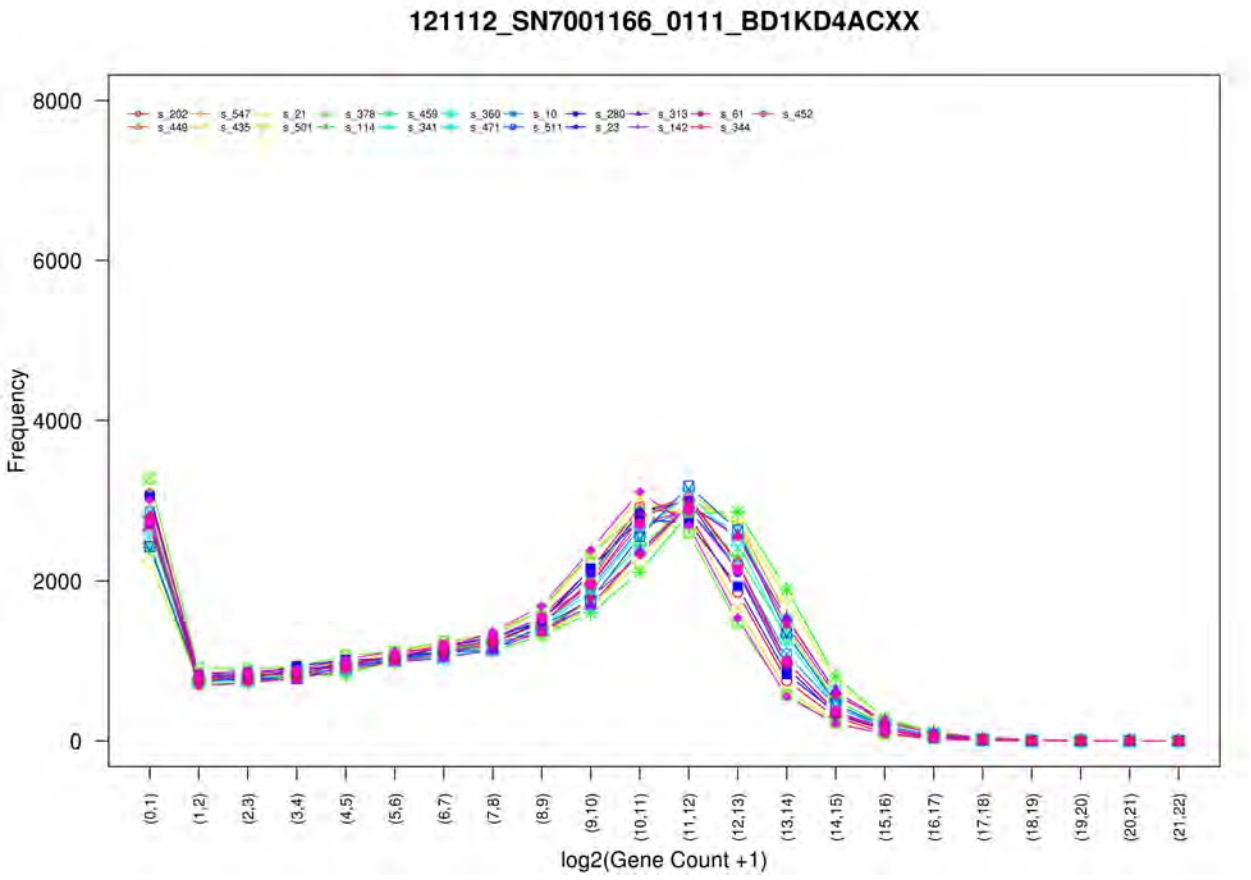


Figure 28: Distribution of log2(Total Gene Counts) for each Subject by RunID

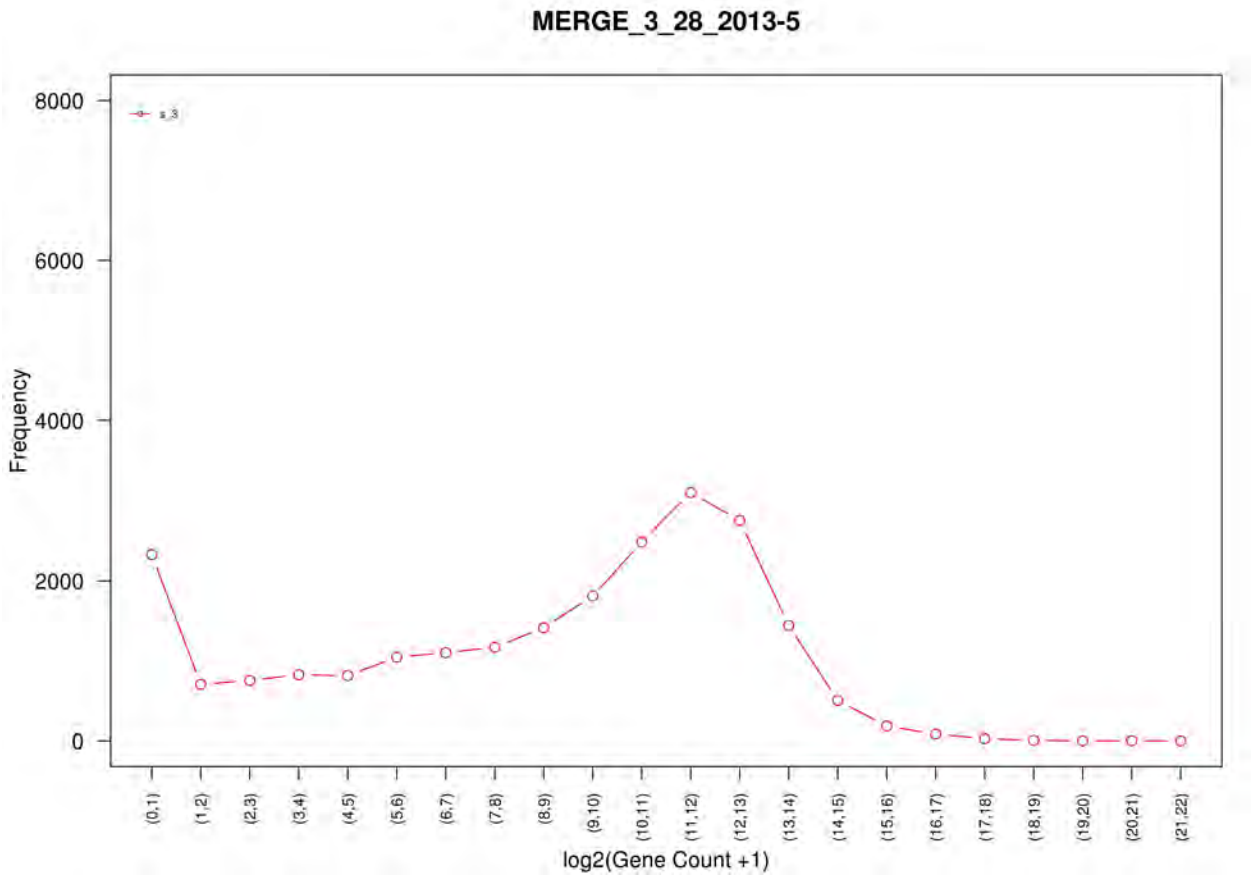


Figure 52: Distribution of log2(Total Gene Counts) for each Subject by RunID

2.2 By GC Content

Because GC Content is known to impact expression levels and can be impacted by PCR, it is important to evaluate whether there are individual subjects that show overall Gene Count levels that vary by %GC. Figure 54 shows a smoothed color density representation of the scatterplot with %GC on the x-axis and $log_2(GeneCount)$ on the y-axis. A loess smoother line is shown indicating the general pattern of all the Gene Count values for this particular subject. Similarly, Figure 55 to 79 shows the loess smoother line for each subject. Based on this plot, it appears that the overall pattern is similar for all samples. Figure 80 shows the distribution of log2(Gene Count+1) by deciles of %GC by flowcell. Again, there is clearly a lower Gene Count when the %GC is higher, but the patterns are similar for most samples.

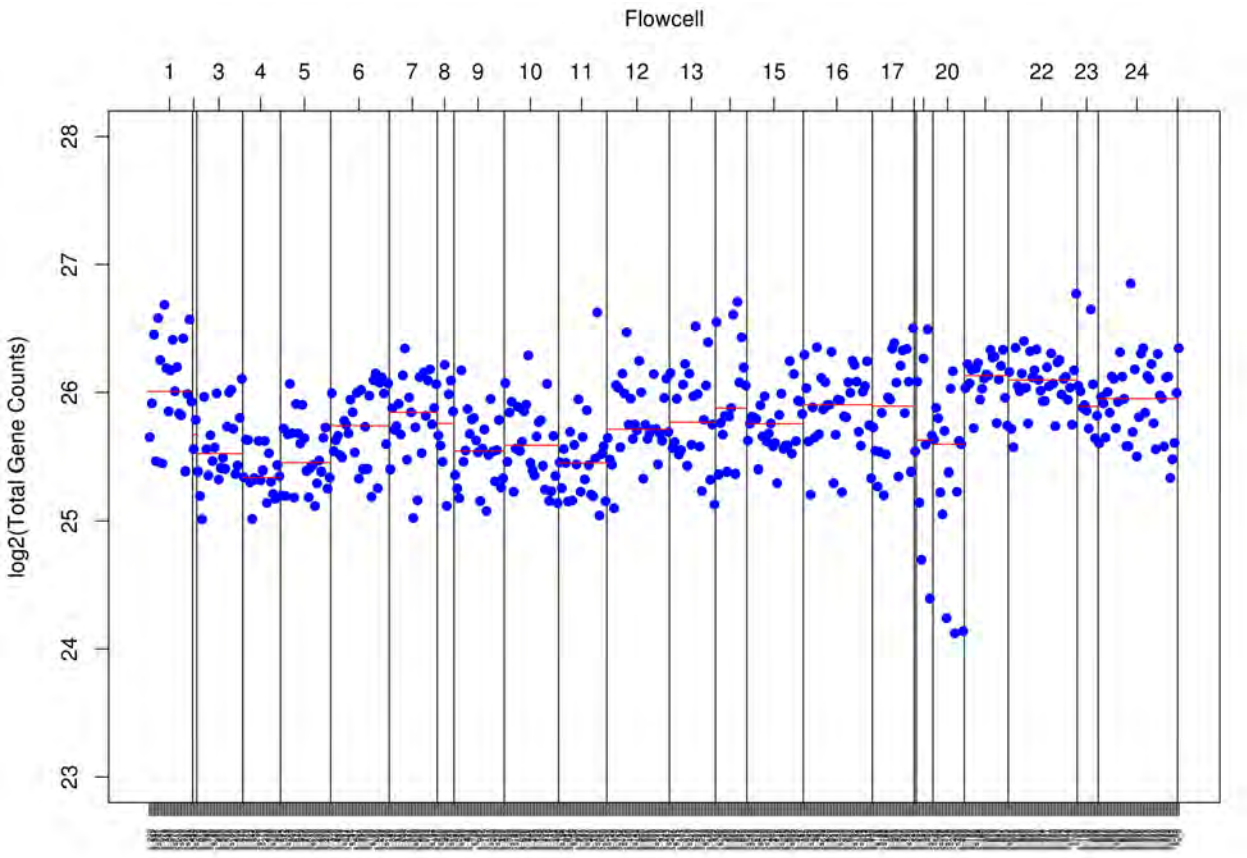


Figure 53: Distribution of Total Gene Counts) for each Subject by RunID

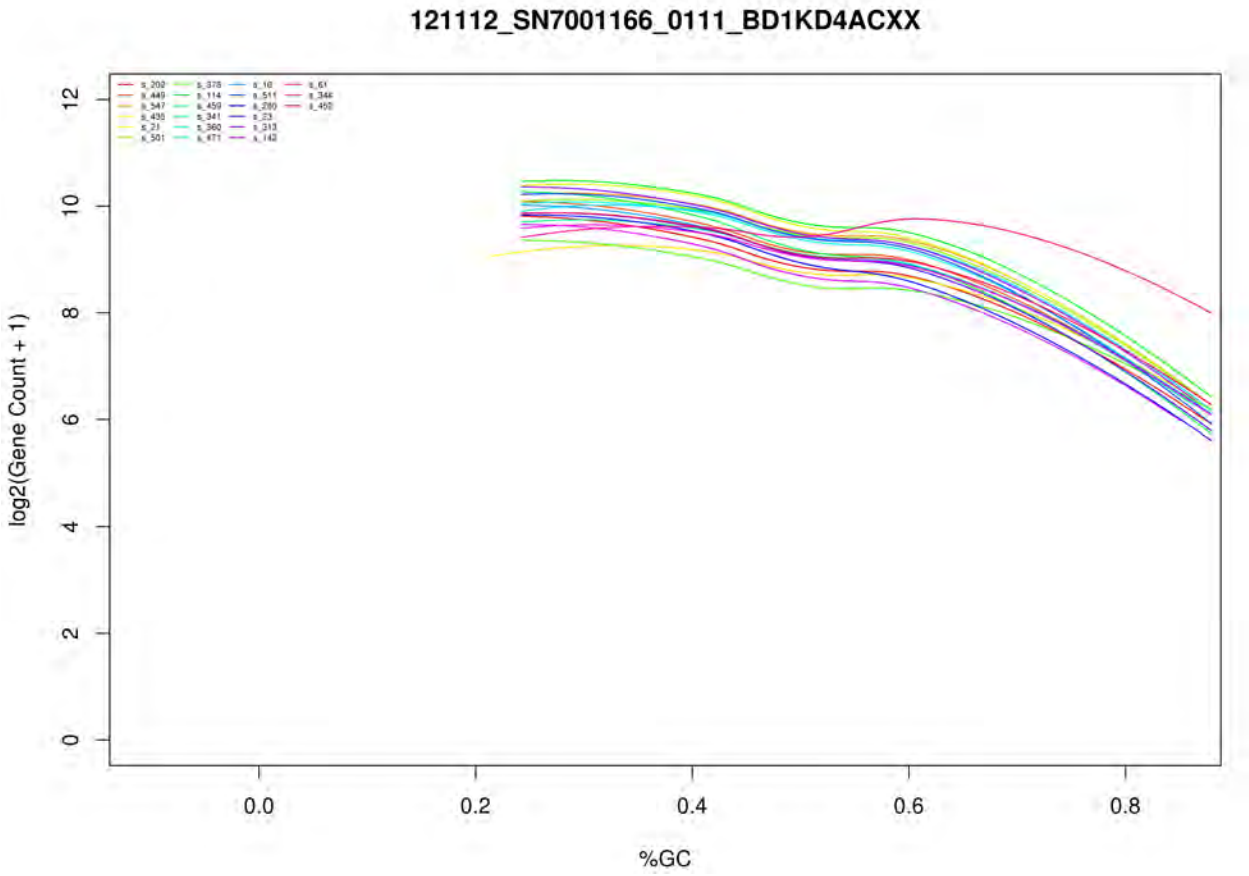


Figure 55: Distribution of Percent GC versus log2(Gene Count + 1) with a loess smoother for each subject by flowcell

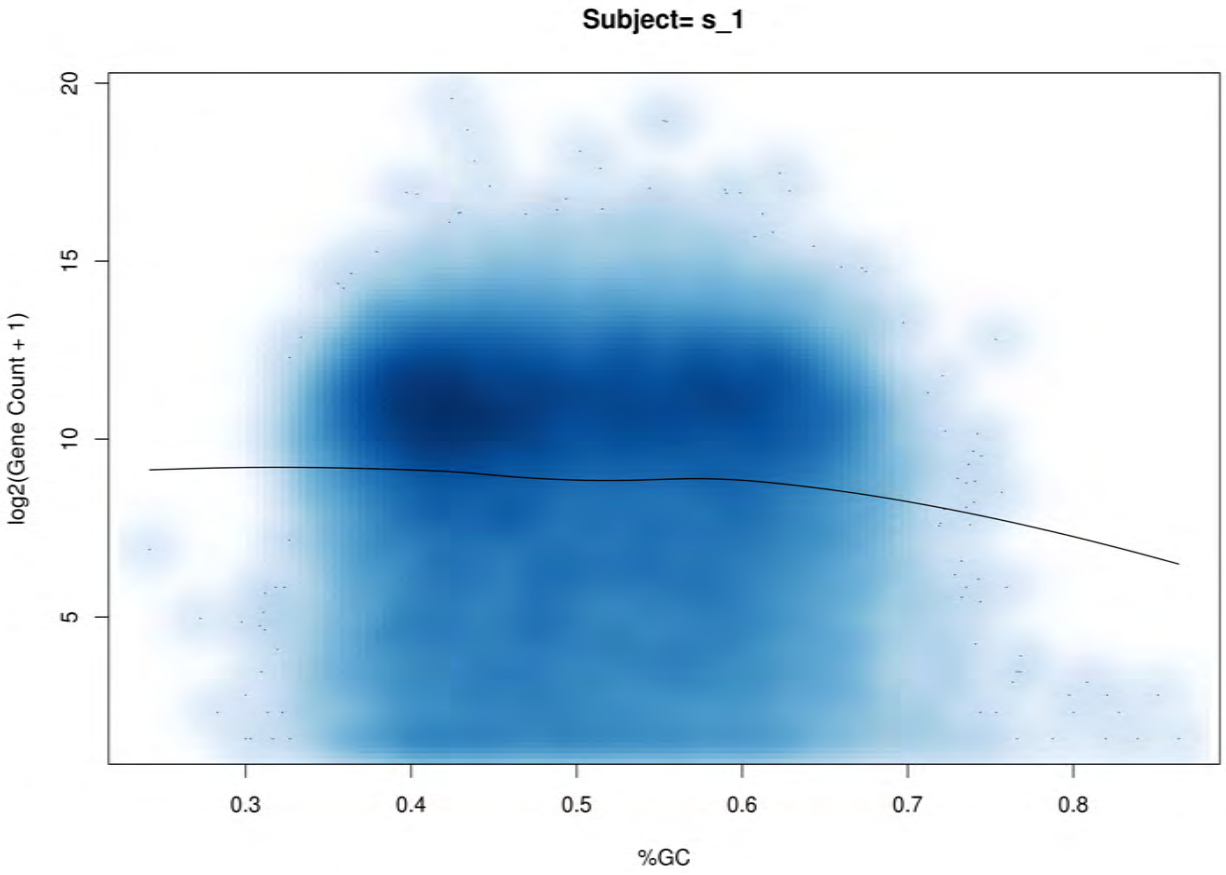


Figure 54: Distribution of %GC versus log2(Gene Count + 1) for subject S1 with a loess smoo

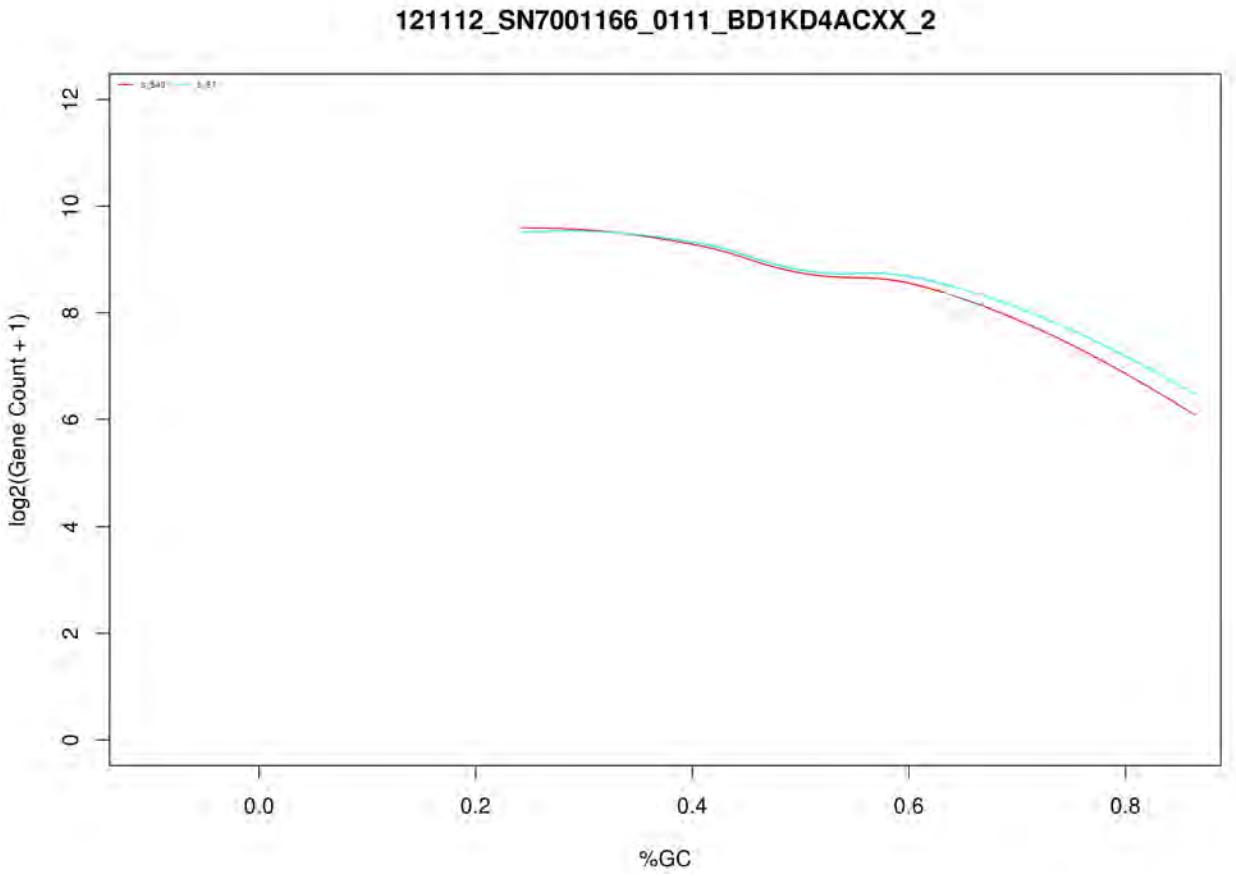


Figure 56: Distribution of Percent GC versus log2(Gene Count + 1) with a loess smoother for each subject by flowcell

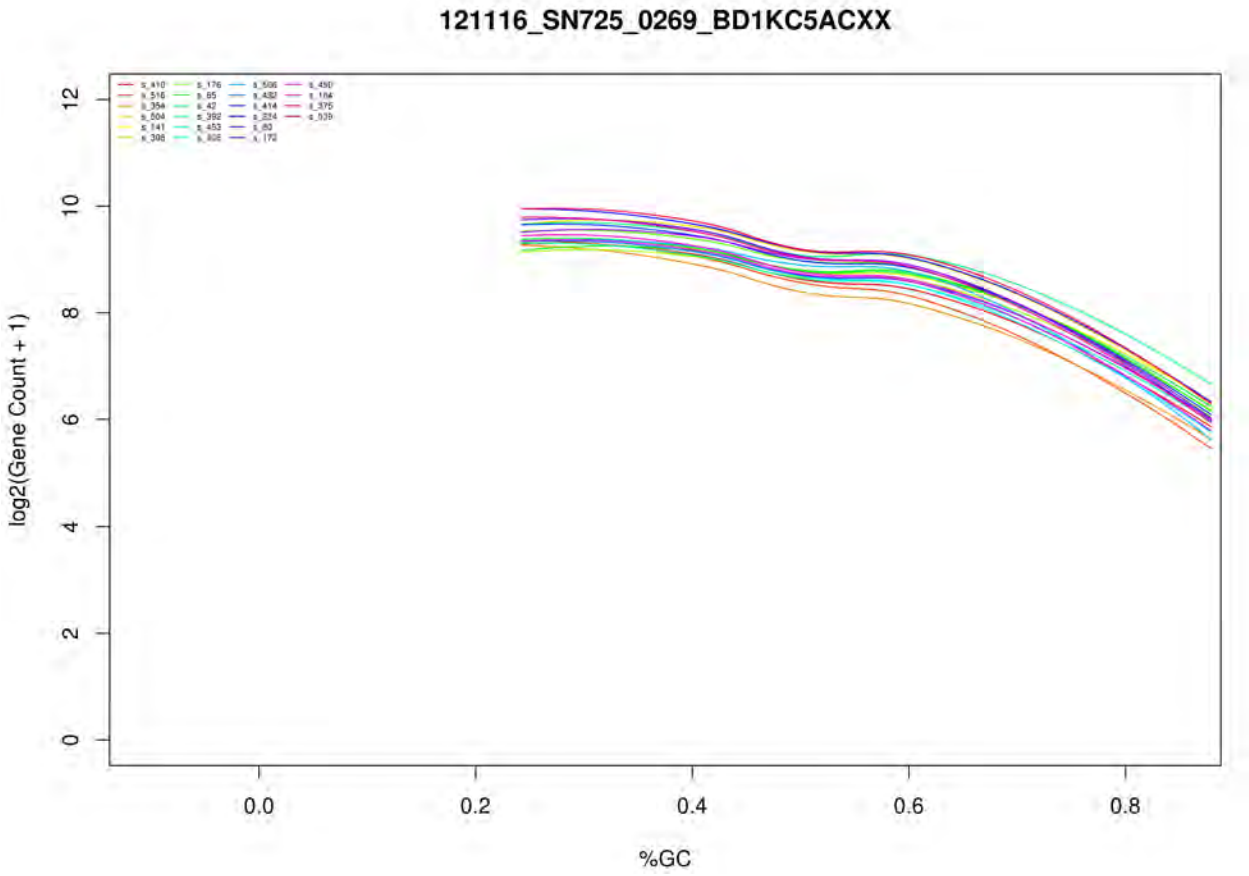


Figure 57: Distribution of Percent GC versus log2(Gene Count + 1) with a loess smoother for each subject by flowcell

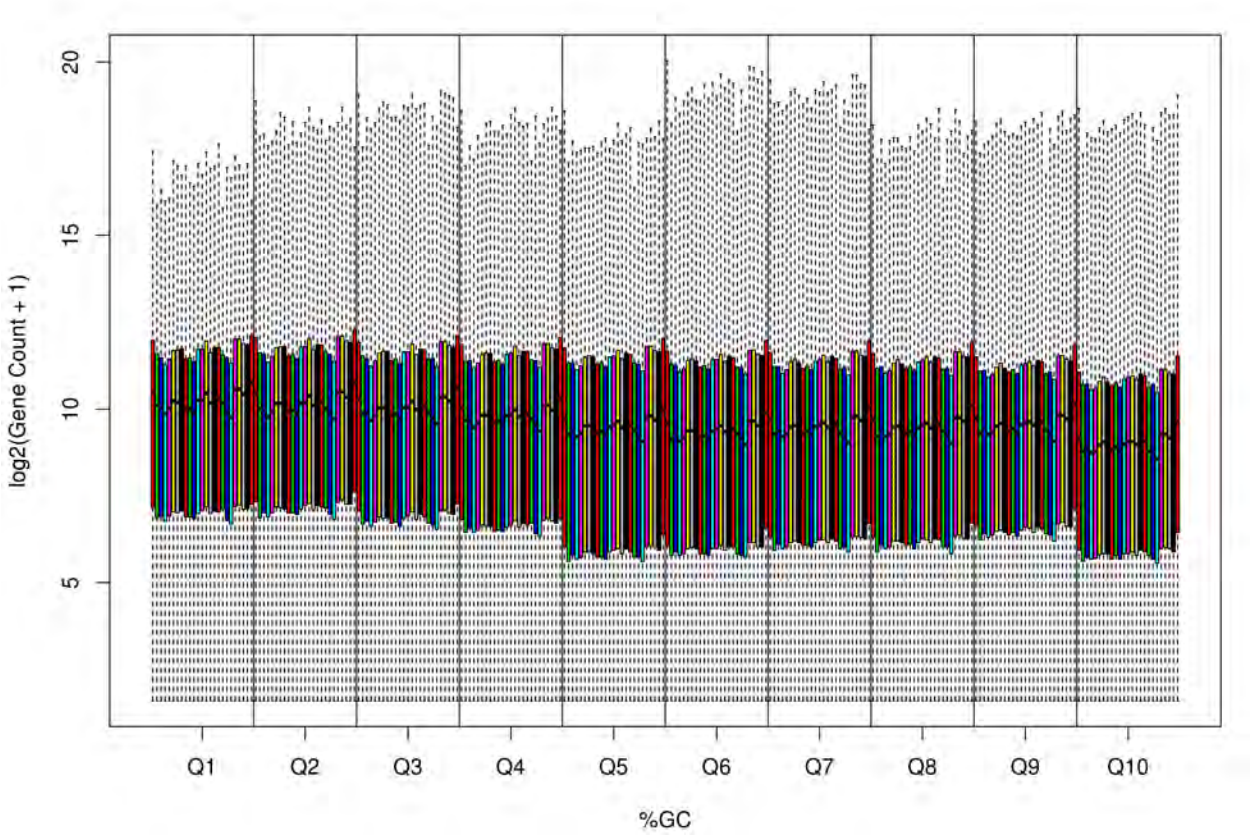


Figure 80: Distribution of log2(Gene Count+1) by deciles of %GC and flowcell

2.3 By Gene Size

Gene Size is known to impact expression levels and hence it is important to assess overall Gene Count levels by Gene size. Figure 81 shows boxplots of Gene Counts by quintiles of Gene size, Figure 82 shows boxplots of Gene Counts by quintiles of Gene size and flowcells, and Figure 83 shows the distribution of log2(Gene Count+1) with smoothed lines for each subject. Patterns differ by size but there is no extreme ouliers.

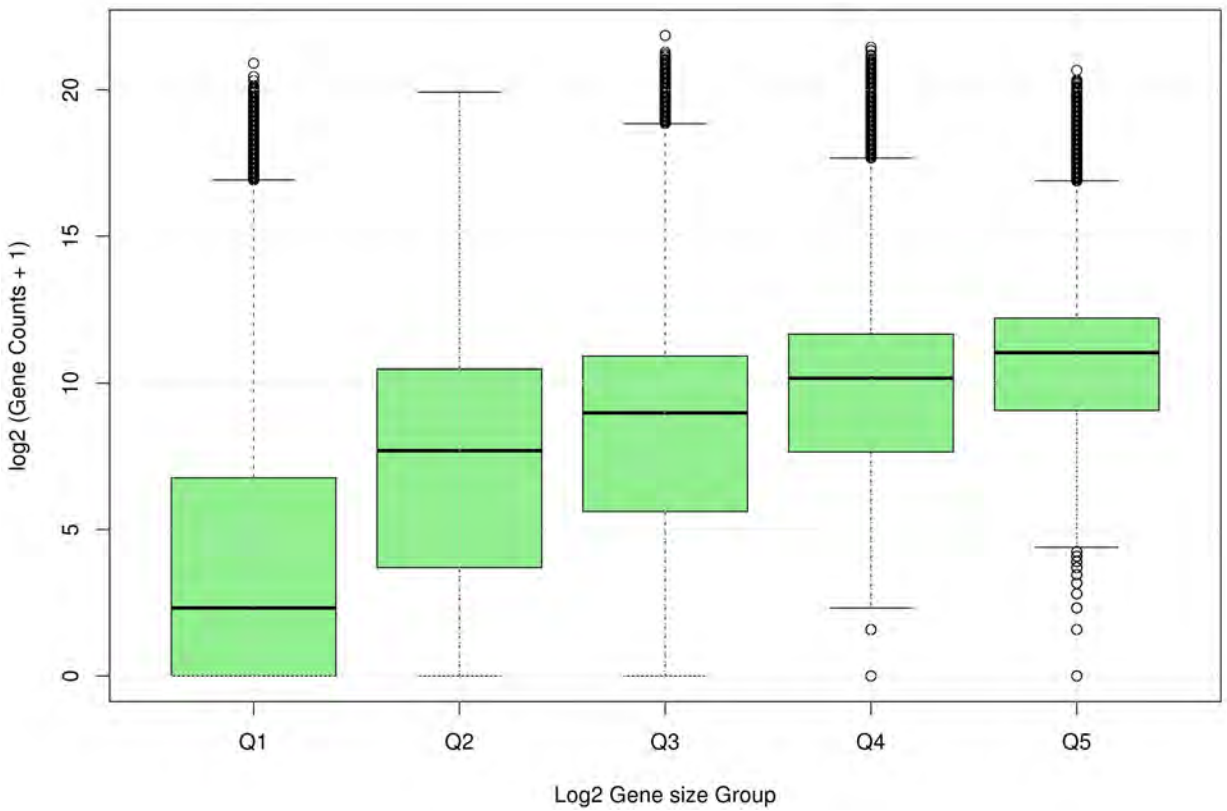


Figure 81: Distribution of log2(Gene Count+1) by Gene Size (5 groups)

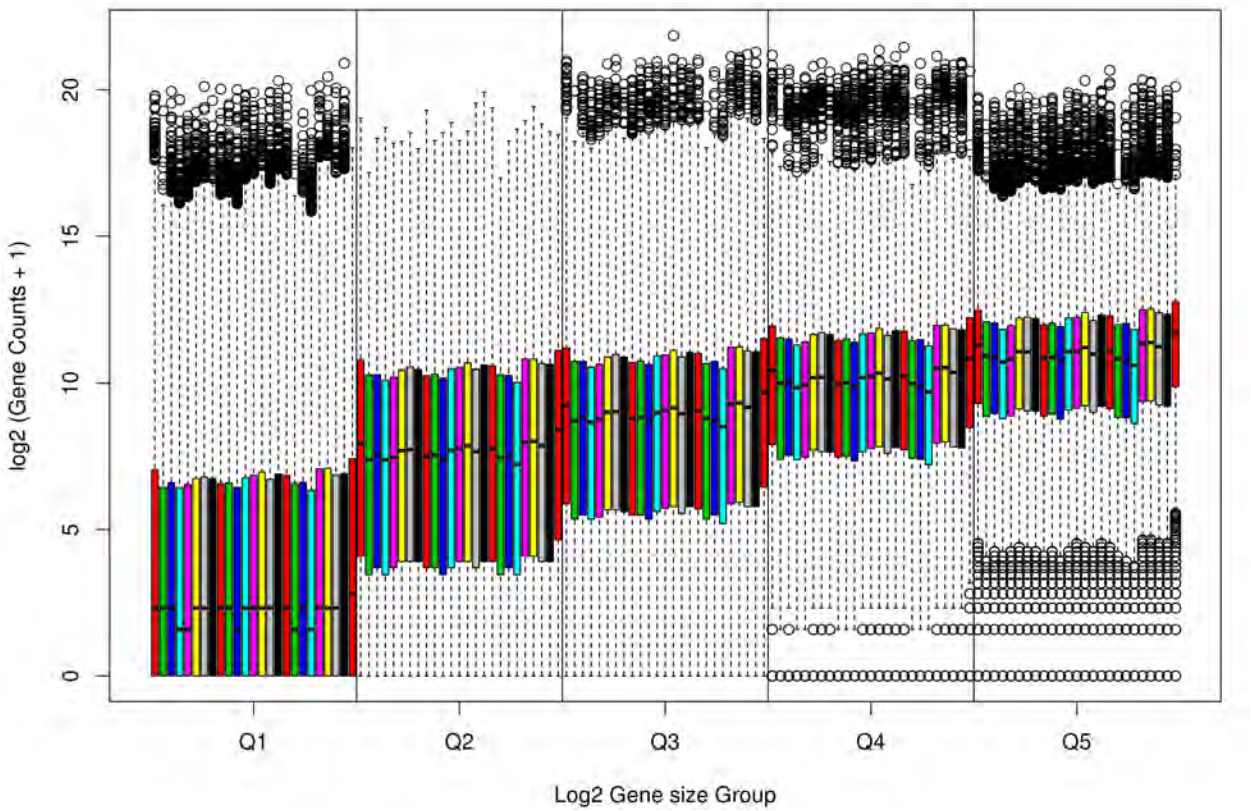


Figure 82: Distribution of log2(Gene Count+1) by flowcell and Gene Size (5 groups) color-coded by flowcell

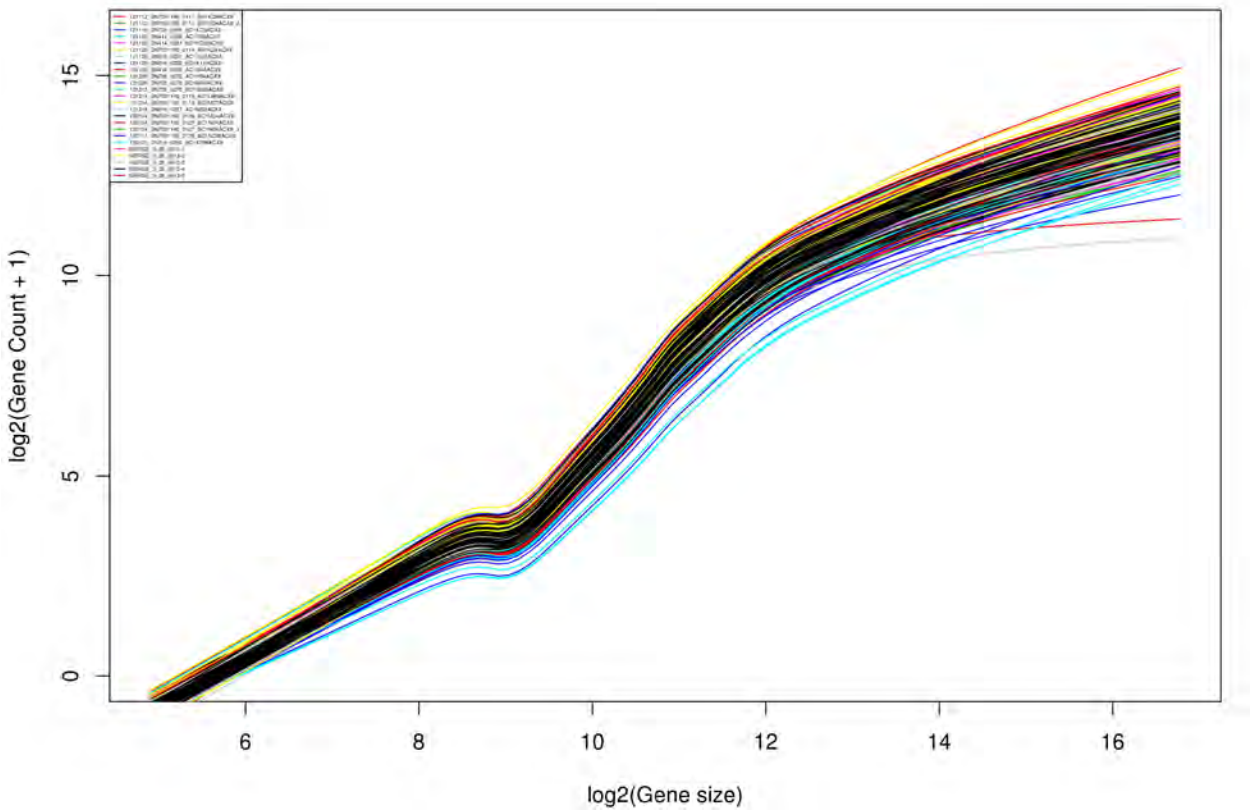


Figure 83: Distribution of log2(Gene Count+1) by Gene Size. Lowess smoothed lines are shown for each subject

2.4 Individual Gene Counts versus the average Gene Count

Finally, it is useful to look at how individual Gene Counts differ from the average (Figure 84).

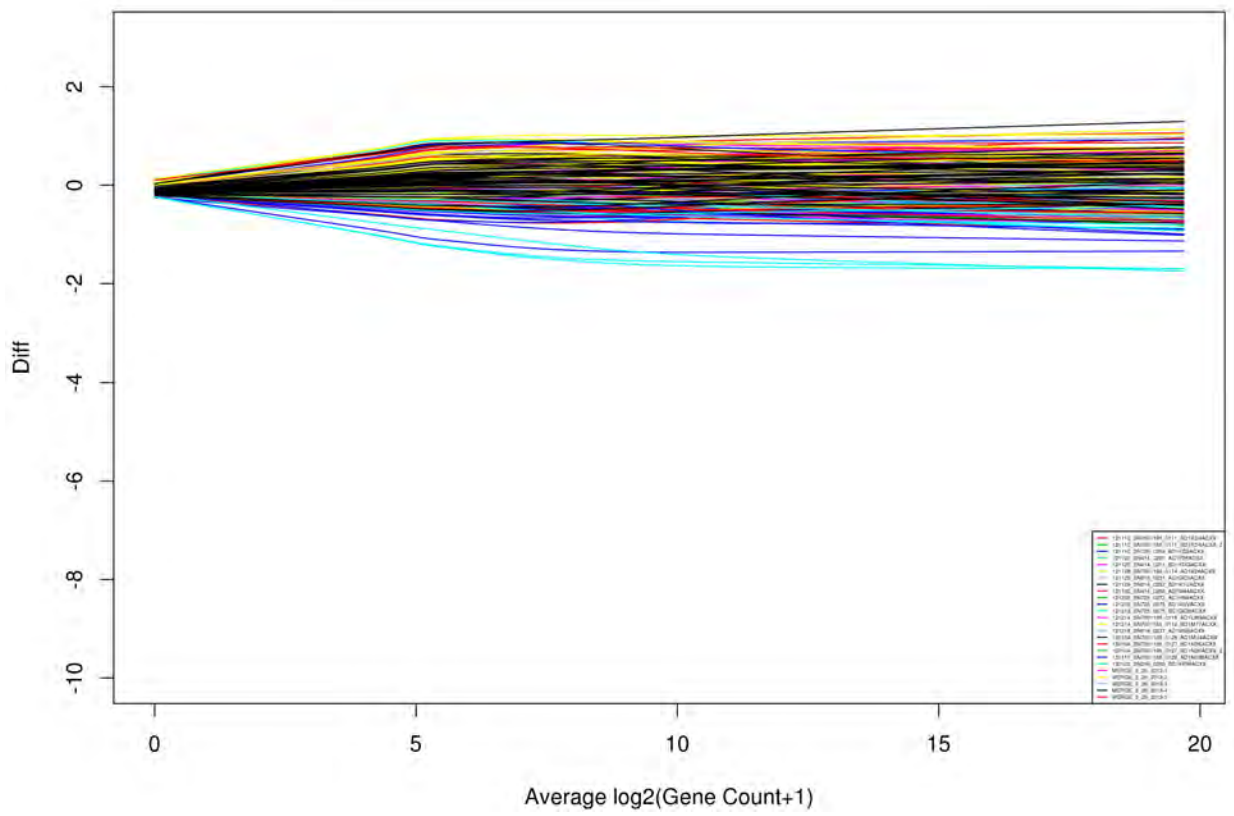


Figure 84: MA Plot showing the difference of log2(Gene Count+1) - mean(log2(Gene Count+1)) versus mean(log2(Gene Count+1)). Lowess smoothed lines are shown for each subject and color coded by flowcell.

3 Normalizing Data

In much of the literature RPKM (reads per kilobase per million) has been used to normalize the mRNA-seq count data. The objective is to take into account the fact that some runs, because of the applification step, are going to produce higher counts. Additionally, this approach takes into account the fact that some genes are larger than others and therefore will have larger counts. Count data typically is analyzed assuming either a Poisson or Negative Binomial distribution. Unfortunately, RPKM changes the underlying structure of the data and renders the distributional assumptions invalid when directly adjusting the ratio. The preferred approach is to model the original gene counts and adjust for additional factors by means of an offset in a Negative Binomial model.
The RPKM for a given sample (subject) is as follows:

C = Number of reads mapped/assigned to a gene for that sample

L = exon length in base-pairs for a gene

N = Total mapped reads for the sample

These are combined in the equation for $RPKM = (10^9 * C)/(N * L)$

3.1 CQN normalization

Recent publications have shown that %GC content can have a large impact on Gene Counts and may need to be accounted for in the analysis. The CQN approach uses the %GC Content in addition to total mapped reads and Gene Length to create an appropriate offset variable for each subject-gene combination.

The CQN package in R was used to estimate an offset for each subject and gene combination, taking into account exon length (gene size) for each gene, %GC content, and total mapped reads for each subject. This offset was then used in the edgeR package in R to run the analysis testing for group differences. Figures 85, 86, 87, and 88 show QC plots after normalization (per subject, by GC Content, by Gene size, and Mean vs Average).

3.2 Sample Filters

A total of 493 passed sample QC filters. 0 sample did not pass QC filters and will be removed from further analysis. Table 3 shows the excluded sample and the reason for exclusion.

SampleID	Use.Status	Eexclude.Reason

Table 3: List of Excluded Samples

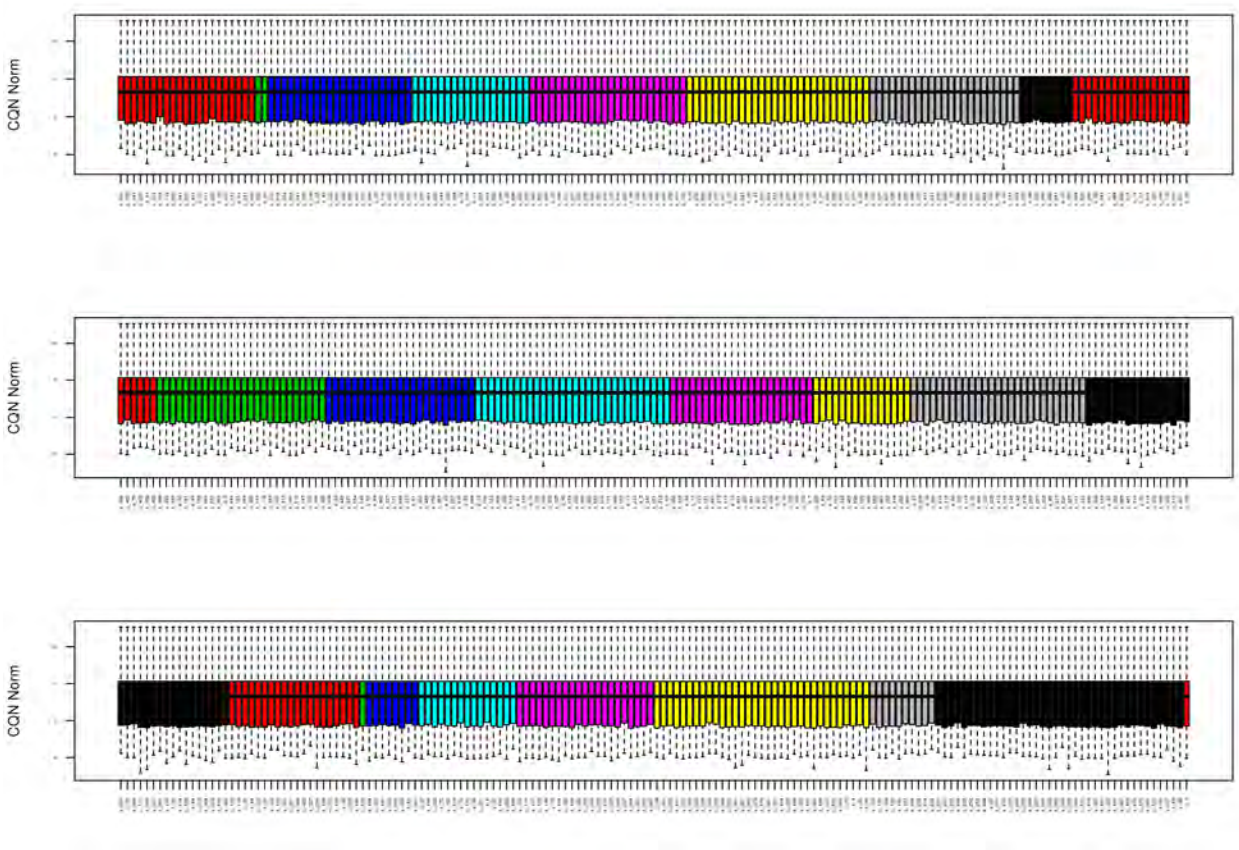


Figure 85: Distribution of normalized Gene Counts/million (on log2 scale) for each subject.

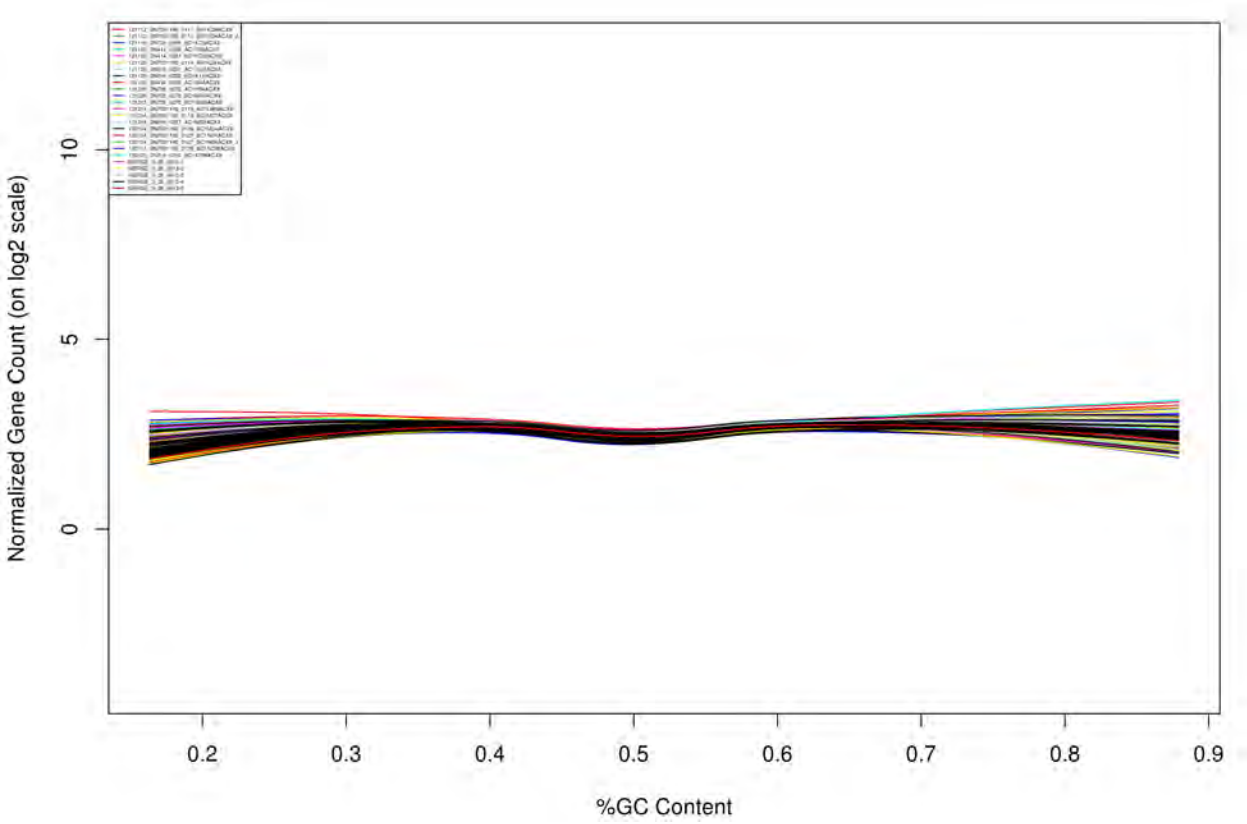


Figure 86: Distribution of normalized Gene Count (on log2 scale) by GC Content. Lowess smoothed lines are shown for each subject

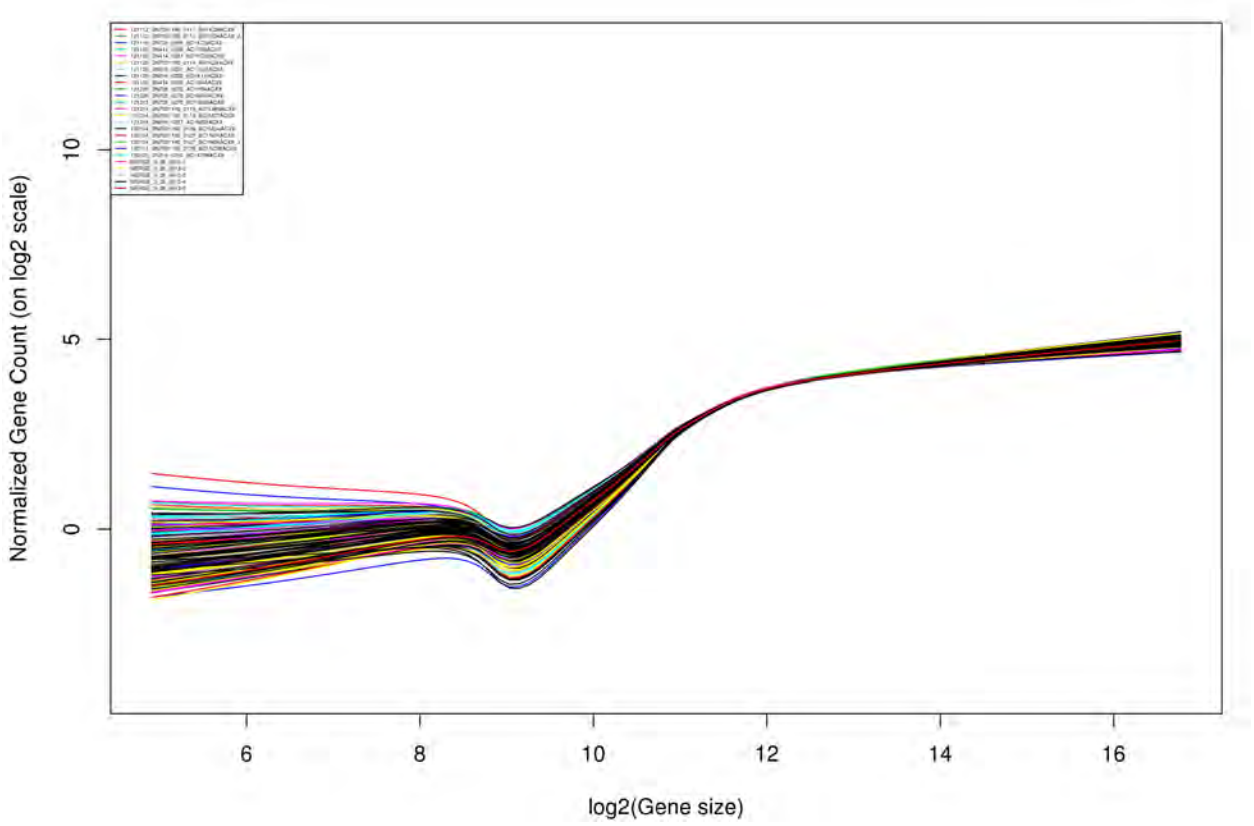


Figure 87: Distribution of normalized Gene Count (on log2 scale) by Gene Size. Lowess smoothed lines are shown for each subject

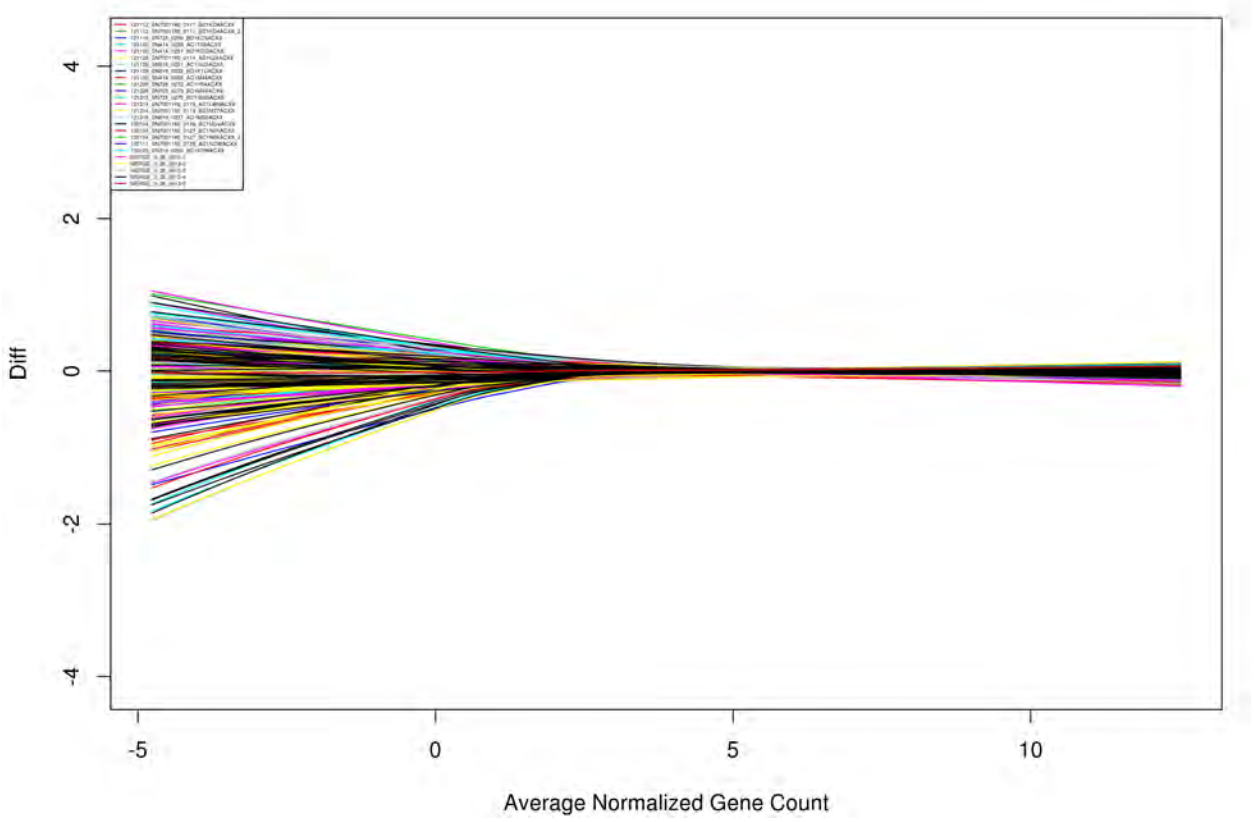


Figure 88: MA Plot showing the difference of the normalized Gene Count - mean(normalized Gene Count) versus mean(normalized Gene Count). Lowess smoothed lines are shown for each subject.

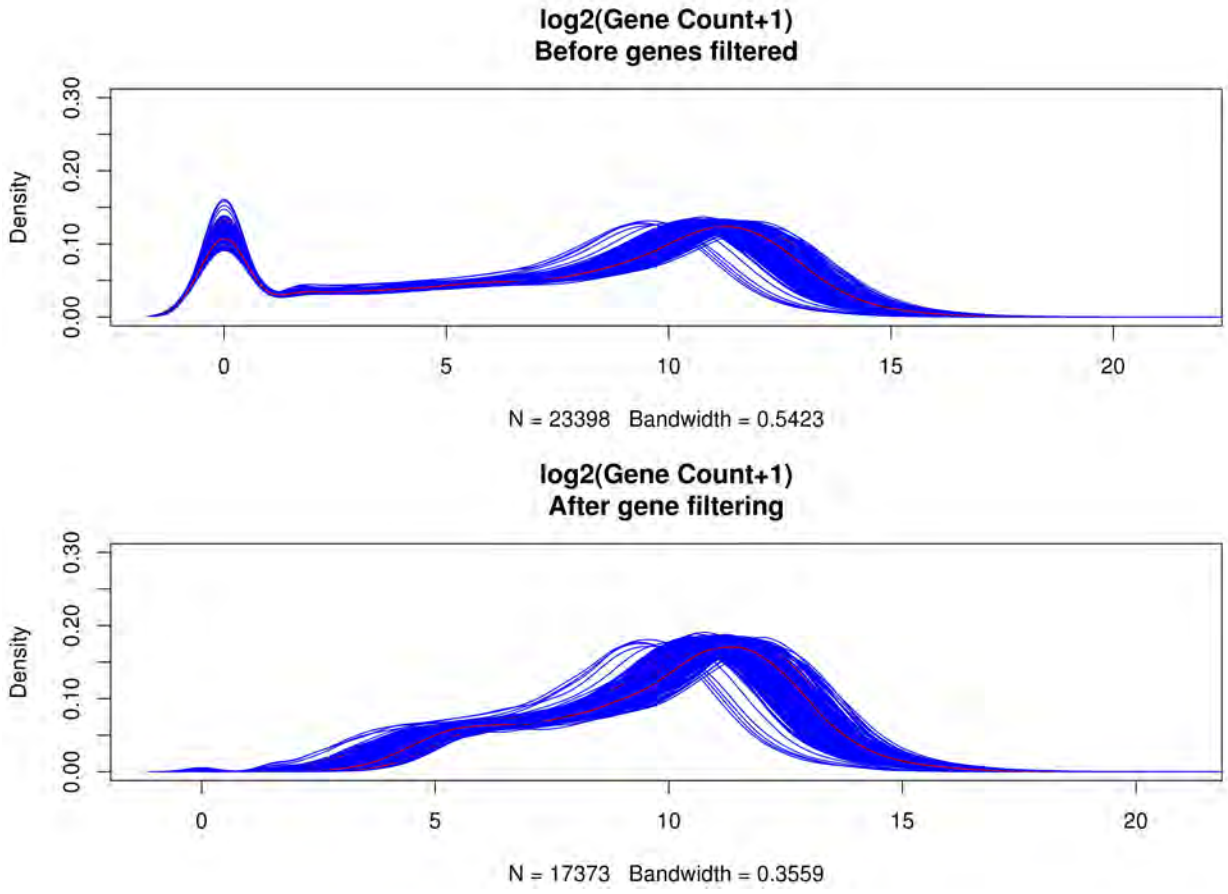
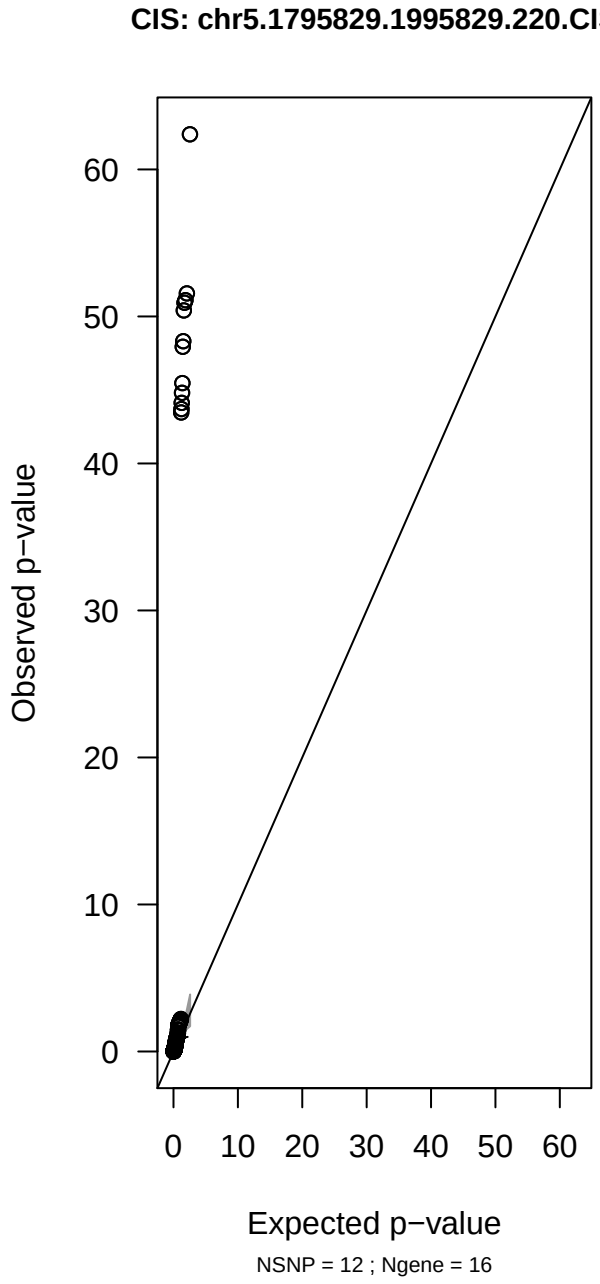


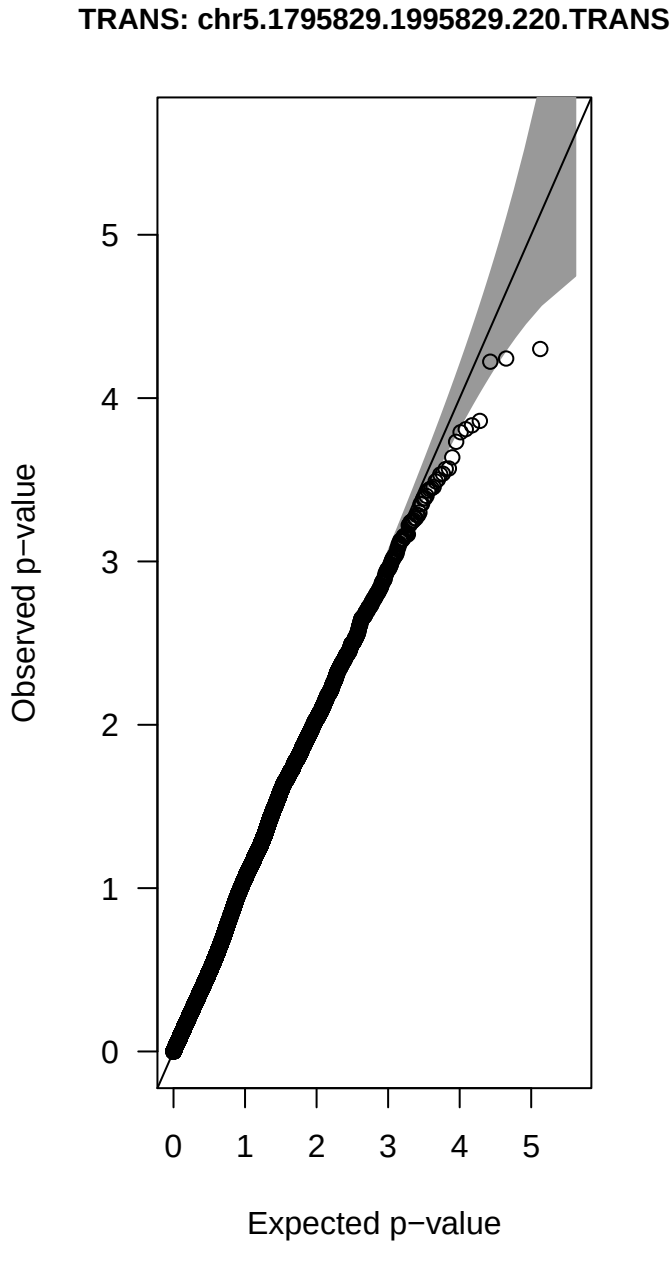
Figure 89: Distribution of log2(Gene Counts + 1) for each Subject by filtering

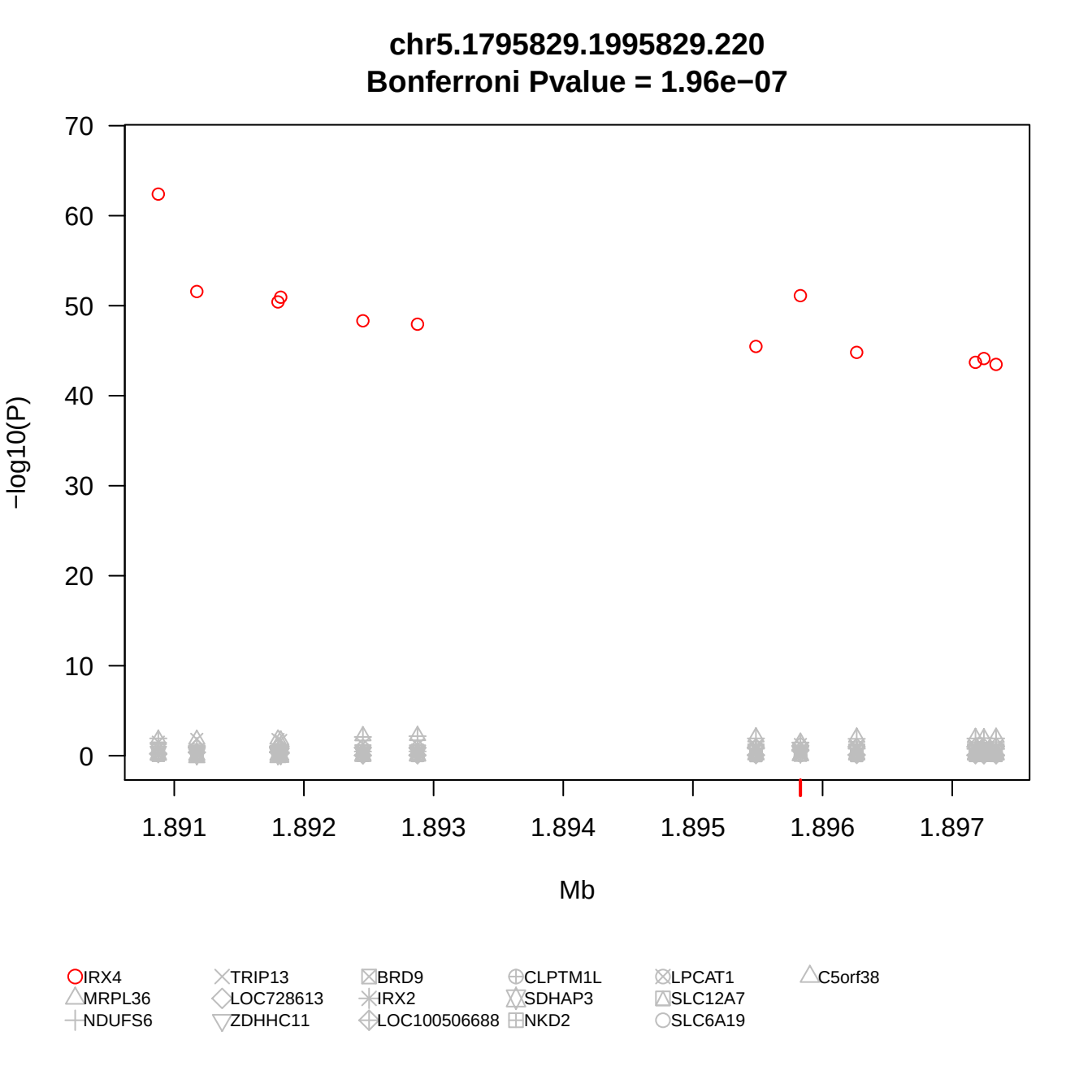
3.3 Gene Filters

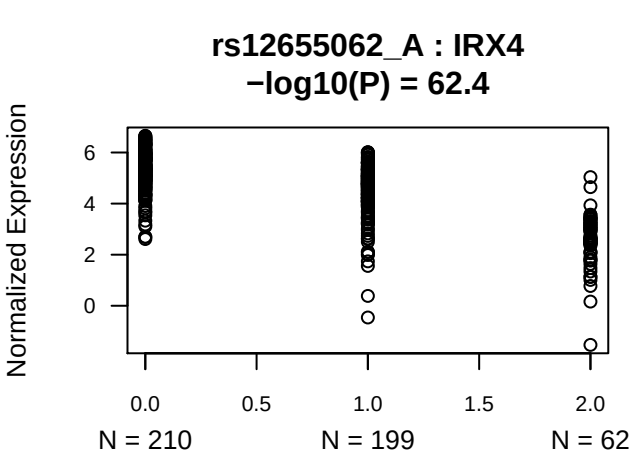
Of the remaining genes with at least 1 count, 5,225 (23.1%) had a median count of less than 16 in the analysis groups and were removed from further analysis (genes deemed undetectable/noise). This filter was applied on the raw count data. The normalized count data will not to done again, we will simply remove the filtered out genes prior to analysis. Figure 89 shows the distribution of the Log2(Gene Count + 1) for each subject before and after filtering for low gene count.

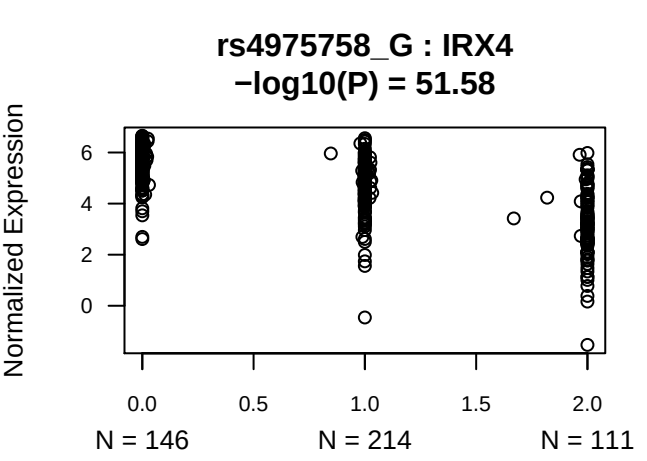
Appendix 4: eQTL analysis for Chromosome 5 region of interest

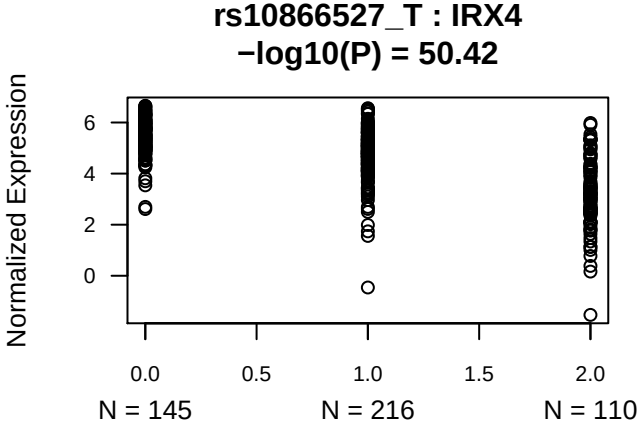


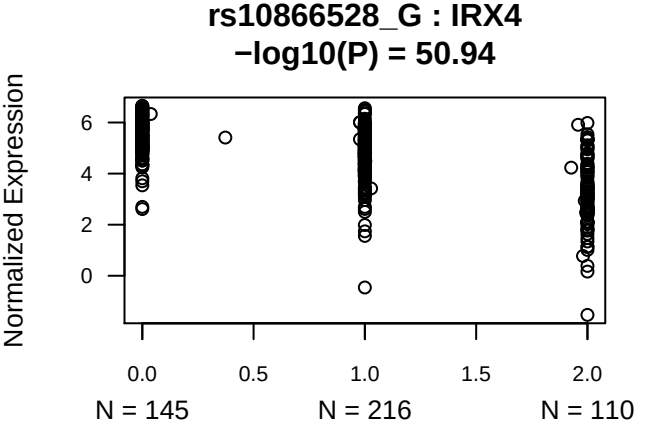


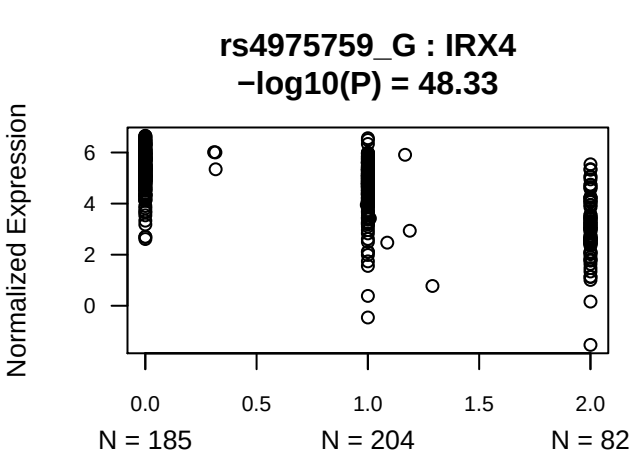

chr5.1795829.1995829.220
Bonferroni Pvalue = 1.96e−07
−log10(P)
70
60
50
40
30
20
10
0
1.891
1.892
1.893
1.894
1.895
1.896
1.897
Mb
IRX4
MRPL36
NDUFS6
TRIP13
LOC728613
ZDHHC11
BRD9
IRX2
LOC100506688
CLPTM1L
SDHAP3
NKD2
LPCAT1
SLC12A7
SLC6A19
C5orf38

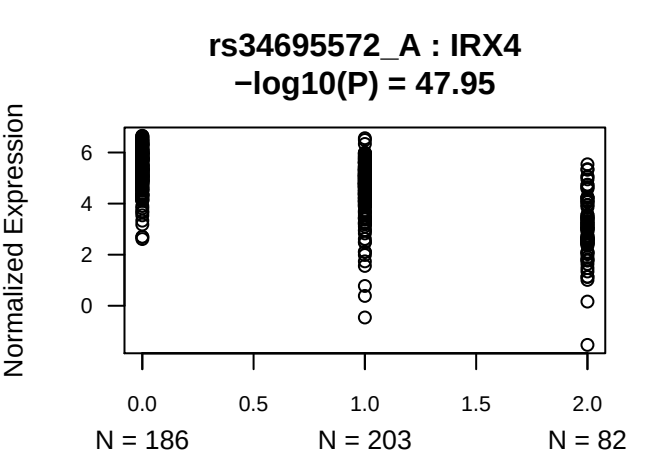

rs12655062_A : IRX4
−log10(P) = 62.4
Normalized Expression
6
4
2
0
0.0
0.5
1.0
1.5
2.0
N = 210
N = 199
N = 62

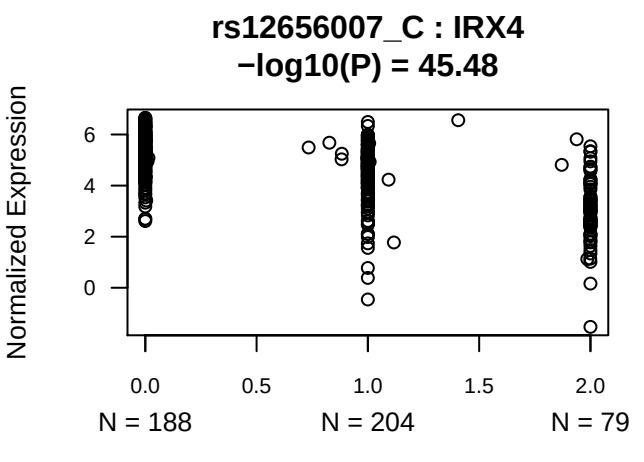

rs4975758_G : IRX4
−log10(P) = 51.58
Normalized Expression
6
4
2
0
0.0
0.5
1.0
1.5
2.0
N = 146
N = 214
N = 111

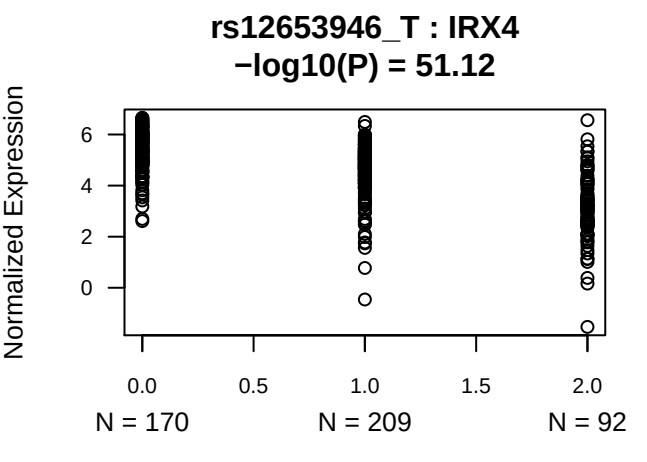

rs10866527_T : IRX4
−log10(P) = 50.42
Normalized Expression
6
4
2
0
0.0
0.5
1.0
1.5
2.0
N = 145
N = 216
N = 110

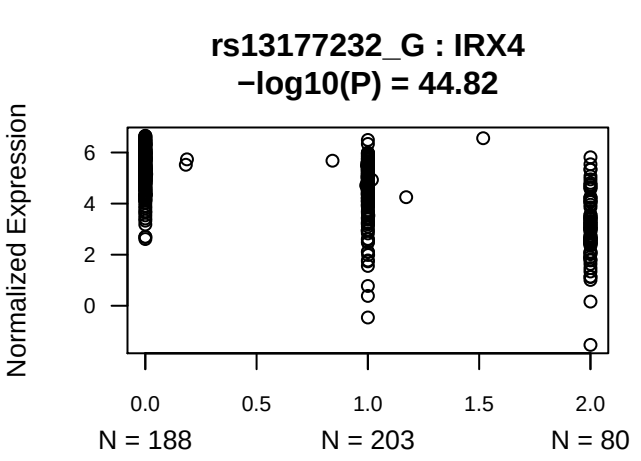

rs10866528_G : IRX4
−log10(P) = 50.94
Normalized Expression
6
4
2
0
0.0
0.5
1.0
1.5
2.0
N = 145
N = 216
N = 110

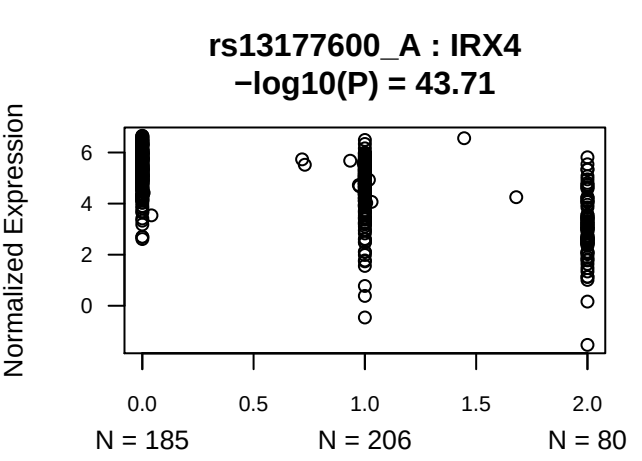

rs4975759_G : IRX4
−log10(P) = 48.33
Normalized Expression
0
2
4
6
0.0
0.5
1.0
1.5
2.0
N = 185
N = 204
N = 82

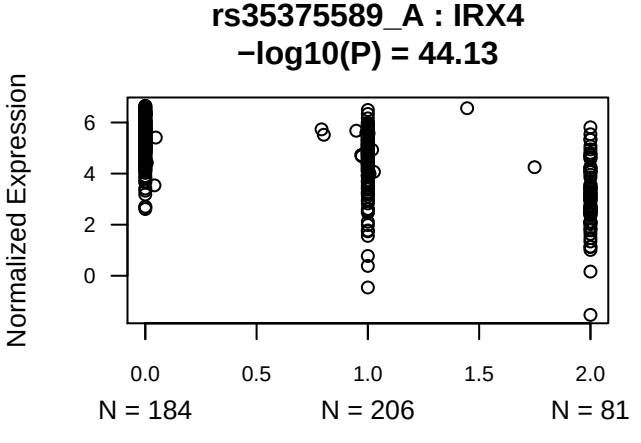

rs34695572_A : IRX4
−log10(P) = 47.95
Normalized Expression
0
2
4
6
0.0
0.5
1.0
1.5
2.0
N = 186
N = 203
N = 82

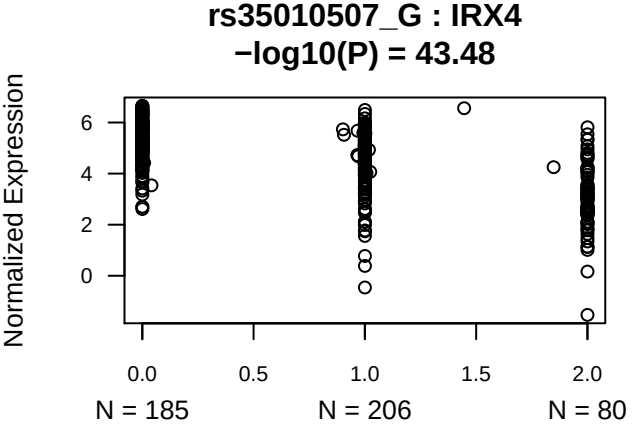

rs12656007_C : IRX4
−log10(P) = 45.48
Normalized Expression
0
2
4
6
0.0
0.5
1.0
1.5
2.0
N = 188
N = 204
N = 79


rs12653946_T : IRX4
−log10(P) = 51.12
Normalized Expression
0
2
4
6
0.0
0.5
1.0
1.5
2.0
N = 170
N = 209
N = 92


rs13177232_G : IRX4
−log10(P) = 44.82
Normalized Expression
0
2
4
6
0.0
0.5
1.0
1.5
2.0
N = 188
N = 203
N = 80

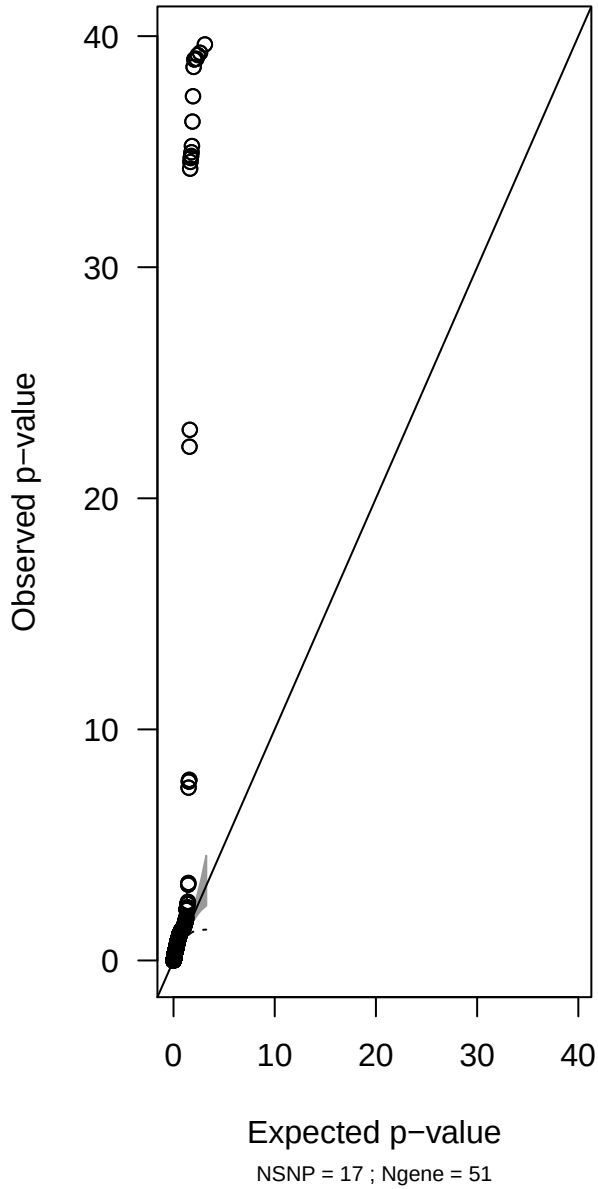

rs13177600_A : IRX4
−log10(P) = 43.71
Normalized Expression
0
2
4
6
0.0
0.5
1.0
1.5
2.0
N = 185
N = 206
N = 80

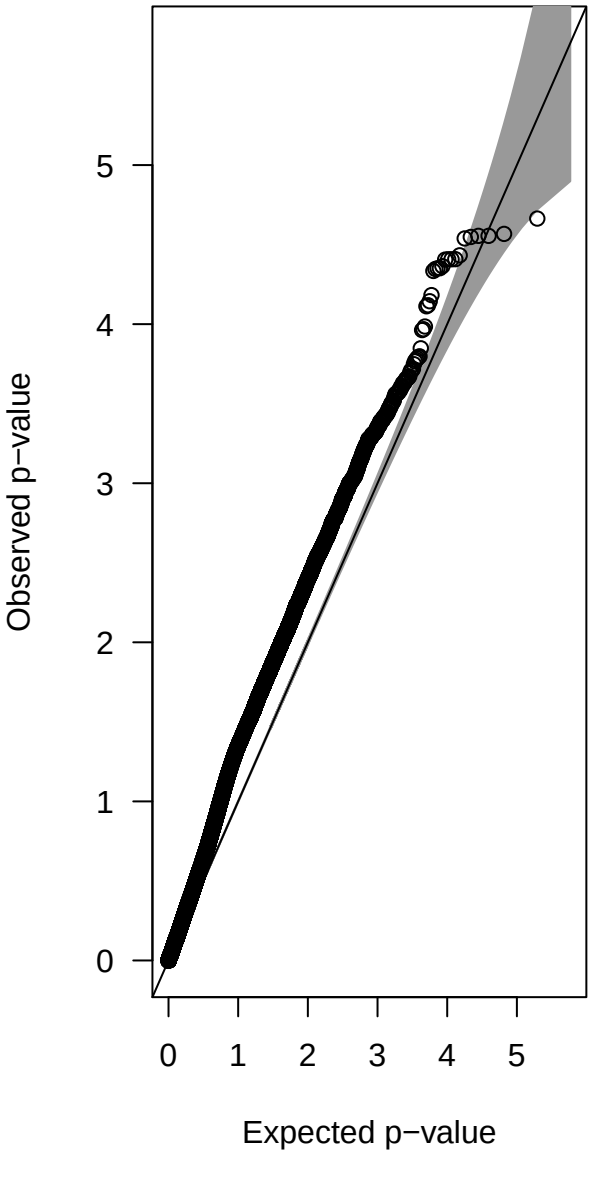

rs35375589_A : IRX4
−log10(P) = 44.13
Normalized Expression
0
2
4
6
0.0
0.5
1.0
1.5
2.0
N = 184
N = 206
N = 81


rs35010507_G : IRX4
−log10(P) = 43.48
Normalized Expression
0
2
4
6
0.0
0.5
1.0
1.5
2.0
N = 185
N = 206
N = 80

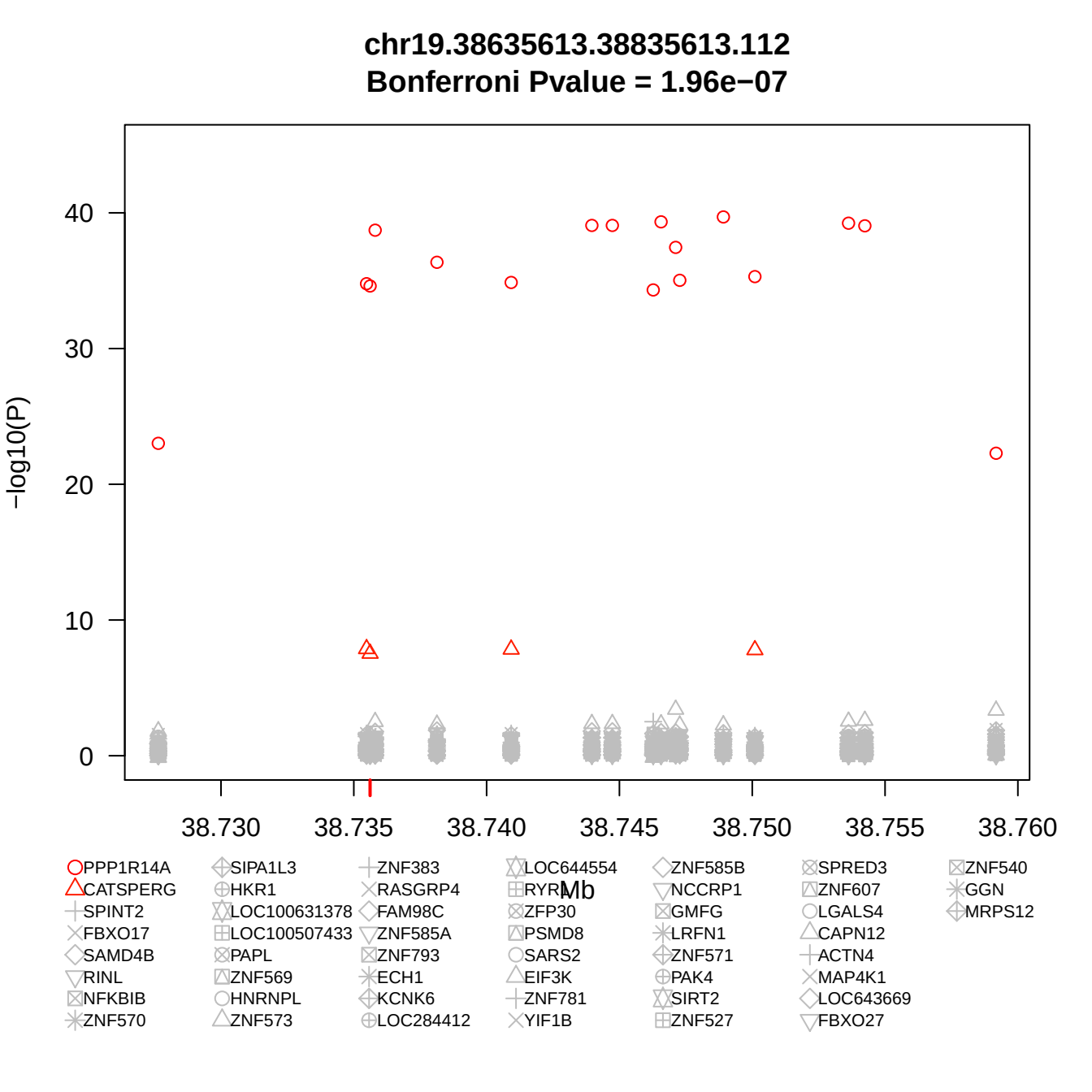
Appendix 5: eQTL analysis for Chromosome 19 region of interest

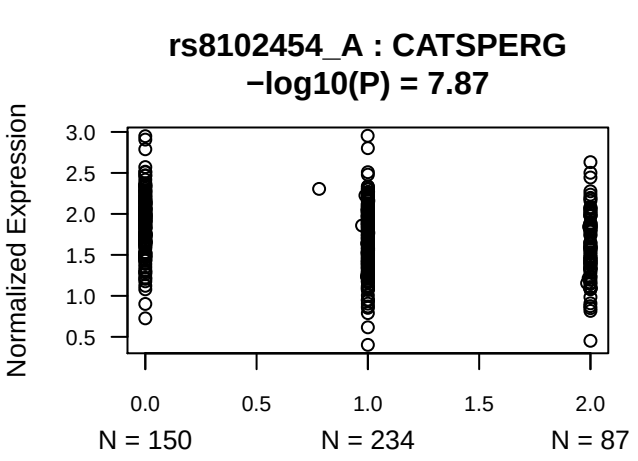
CIS: chr19.38635613.38835613.112.CIS **TRANS: chr19.38635613.38835613.112.TRAN**

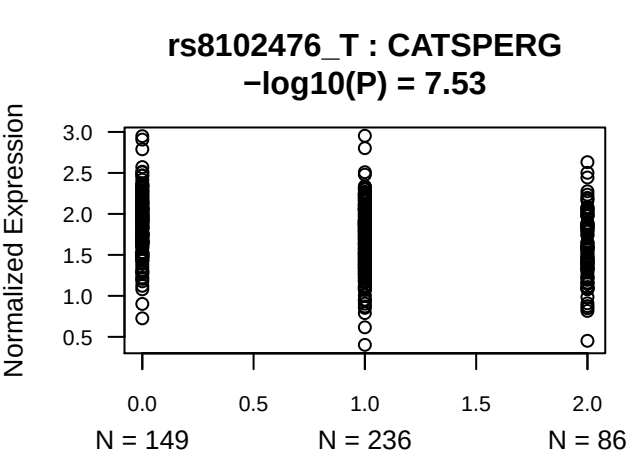


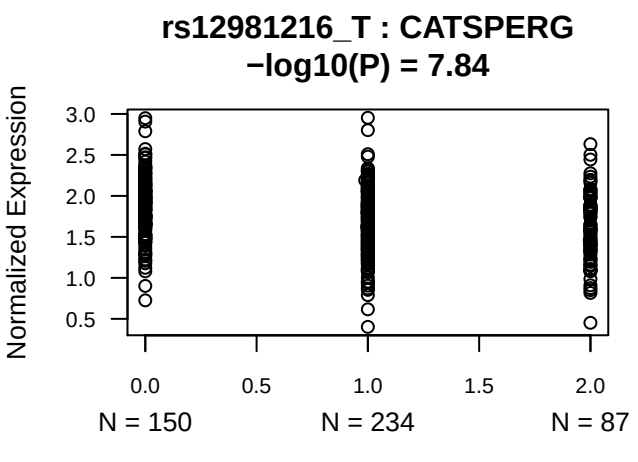


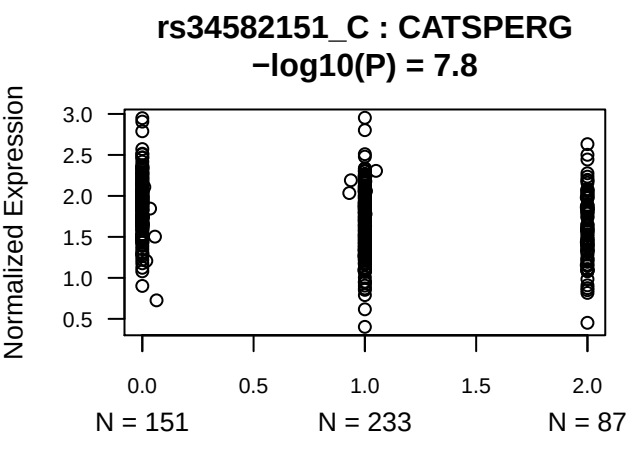
NSNP = 17 ; Ngene = 51

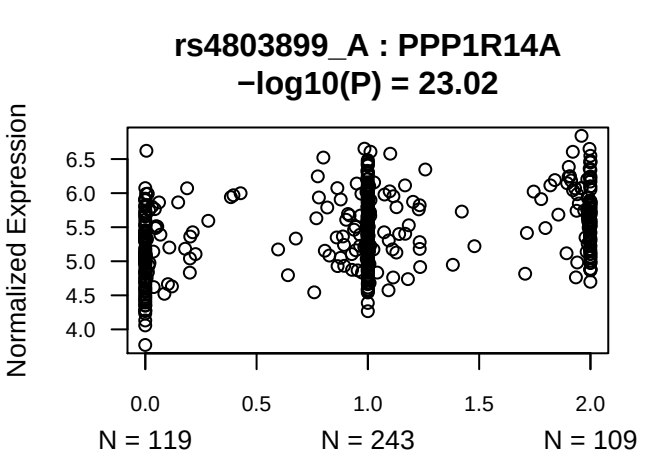

chr19.38635613.38835613.112
Bonferroni Pvalue = 1.96e−07
−log10(P)
40
30
20
10
0
38.730
38.735
38.740
38.745
38.750
38.755
38.760
Mb
PPP1R14A
CATSPERG
SPINT2
FBXO17
SAMD4B
RINL
NFKBIB
ZNF570
SIPA1L3
HKR1
LOC100631378
LOC100507433
PAPL
ZNF569
HNRNPL
ZNF573
ZNF383
RASGRP4
FAM98C
ZNF585A
ZNF793
ECH1
KCNK6
LOC284412
LOC644554
RYR1
ZFP30
PSMD8
SARS2
EIF3K
ZNF781
YIF1B
ZNF585B
NCCRP1
GMFG
LRFN1
ZNF571
PAK4
SIRT2
ZNF527
SPRED3
ZNF607
LGALS4
CAPN12
ACTN4
MAP4K1
LOC643669
FBXO27
ZNF540
GGN
MRPS12

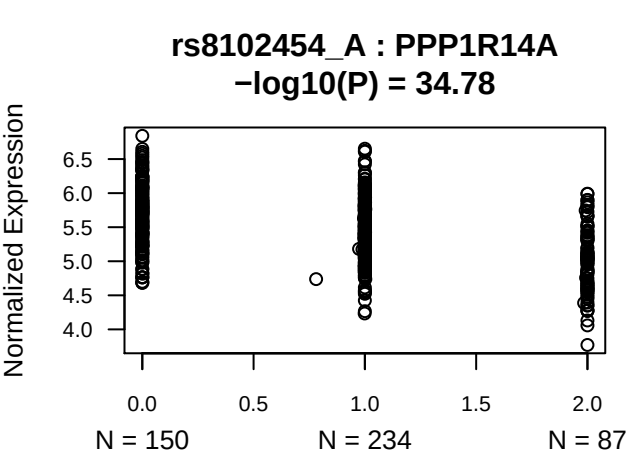

rs8102454_A : CATSPERG
−log10(P) = 7.87
Normalized Expression
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
N = 150
N = 234
N = 87

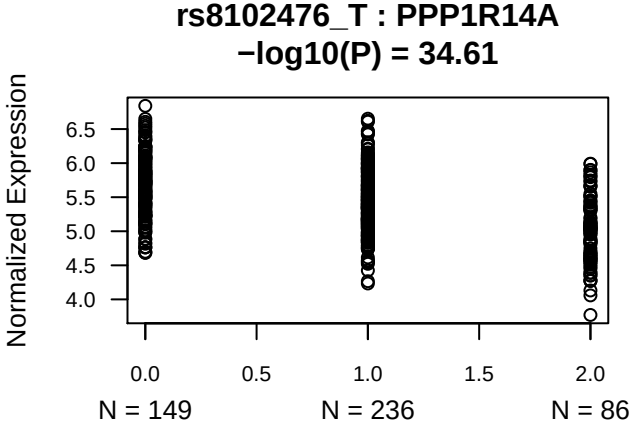

rs8102476_T : CATSPERG
−log10(P) = 7.53
Normalized Expression
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
N = 149
N = 236
N = 86

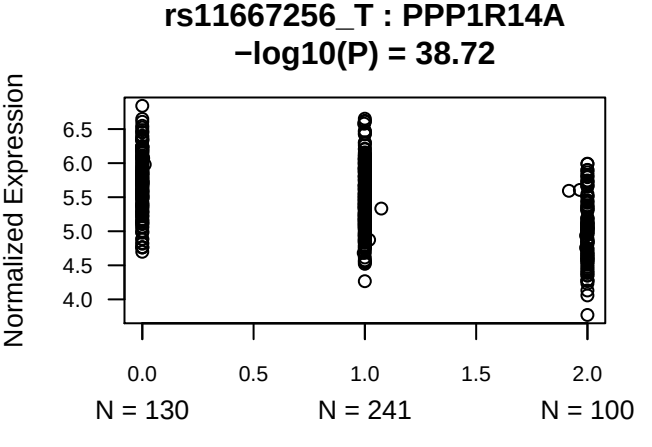

rs12981216_T : CATSPERG
−log10(P) = 7.84
Normalized Expression
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
N = 150
N = 234
N = 87

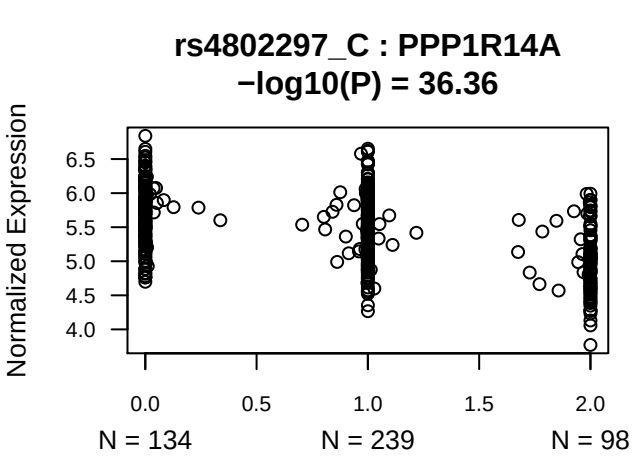

rs34582151_C : CATSPERG
−log10(P) = 7.8
Normalized Expression
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
N = 151
N = 233
N = 87

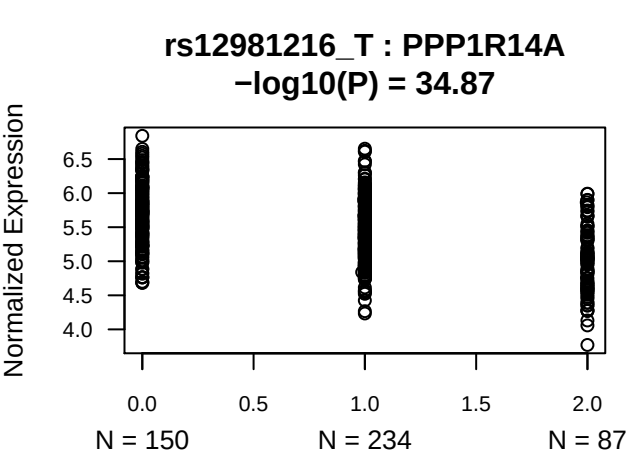

rs4803899_A : PPP1R14A
−log10(P) = 23.02
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 119
N = 243
N = 109

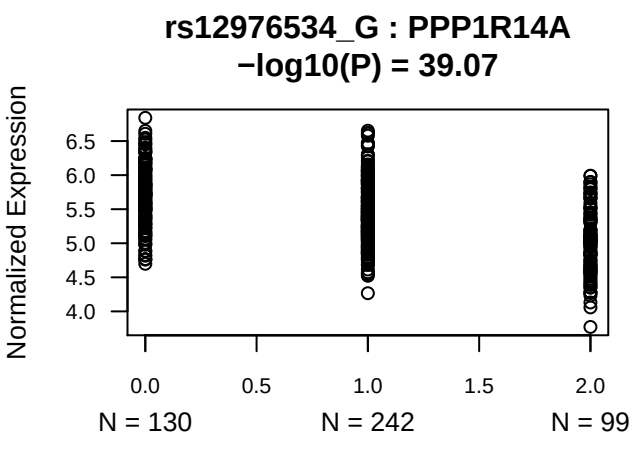

rs8102454_A : PPP1R14A
−log10(P) = 34.78
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 150
N = 234
N = 87

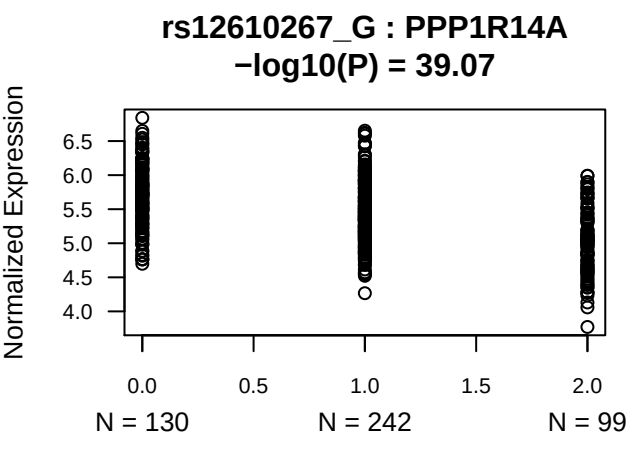

rs8102476_T : PPP1R14A
−log10(P) = 34.61
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 149
N = 236
N = 86

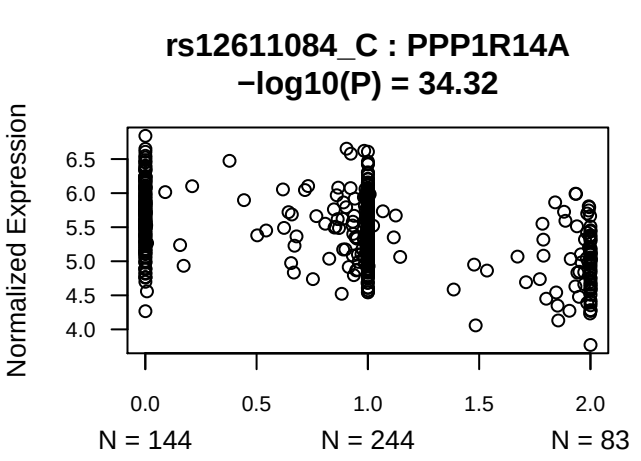

rs11667256_T : PPP1R14A
−log10(P) = 38.72
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 130
N = 241
N = 100

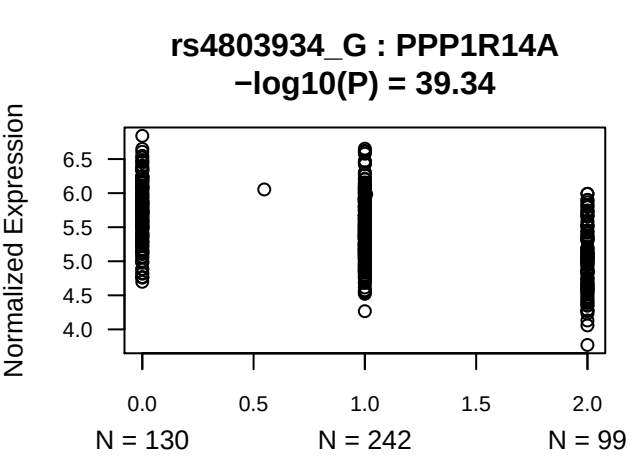

rs4802297_C : PPP1R14A
−log10(P) = 36.36
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 134
N = 239
N = 98

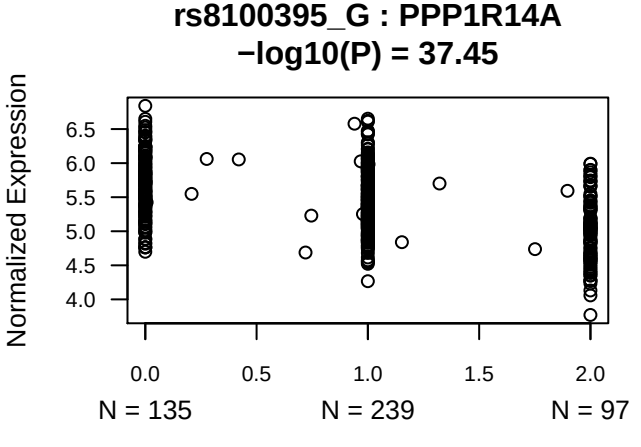

rs12981216_T : PPP1R14A
−log10(P) = 34.87
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 150
N = 234
N = 87

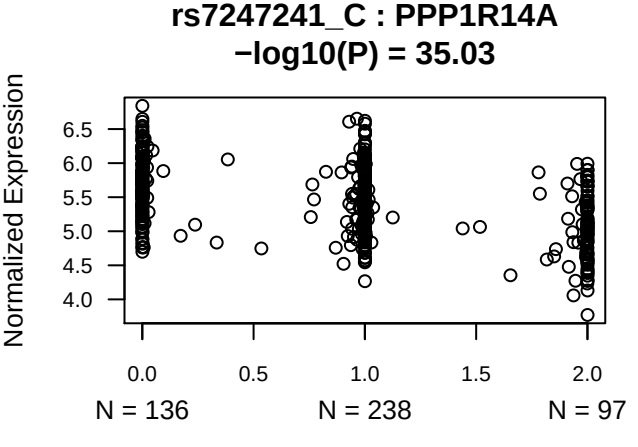

rs12976534_G : PPP1R14A
−log10(P) = 39.07
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 130
N = 242
N = 99

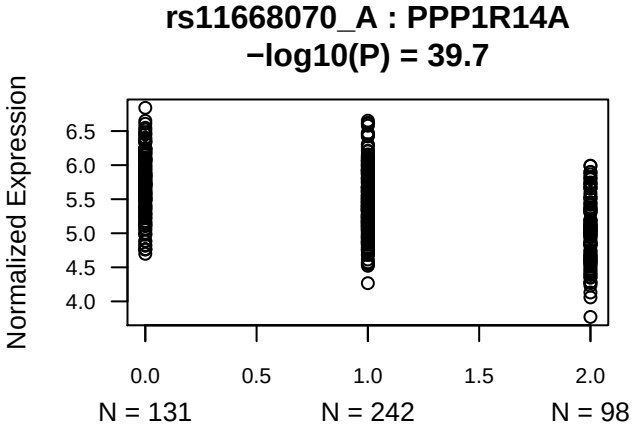

rs12610267_G : PPP1R14A
−log10(P) = 39.07
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 130
N = 242
N = 99

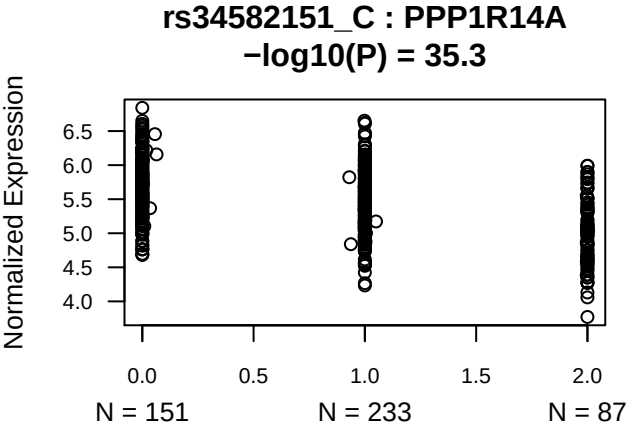

rs12611084_C : PPP1R14A
−log10(P) = 34.32
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 144
N = 244
N = 83

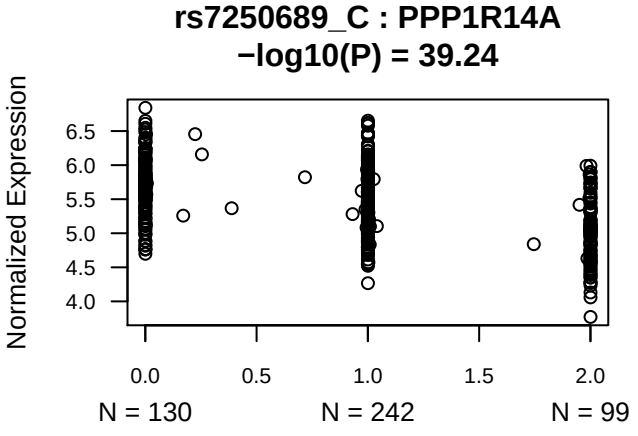

rs4803934_G : PPP1R14A
−log10(P) = 39.34
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 130
N = 242
N = 99

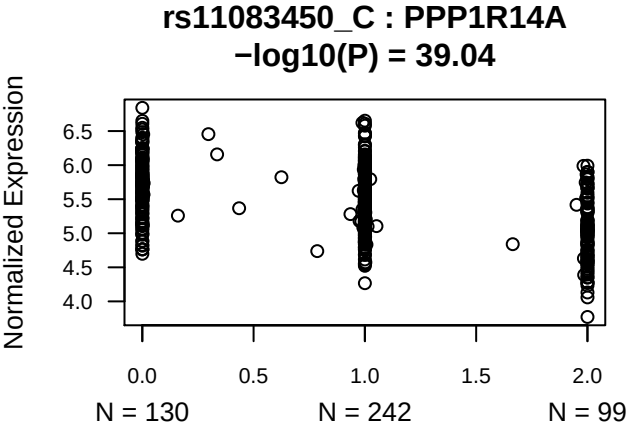

rs8100395_G : PPP1R14A
−log10(P) = 37.45
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 135
N = 239
N = 97


rs7247241_C : PPP1R14A
−log10(P) = 35.03
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 136
N = 238
N = 97


rs11668070_A : PPP1R14A
−log10(P) = 39.7
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 131
N = 242
N = 98


rs34582151_C : PPP1R14A
−log10(P) = 35.3
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 151
N = 233
N = 87


rs7250689_C : PPP1R14A
−log10(P) = 39.24
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 130
N = 242
N = 99


rs11083450_C : PPP1R14A
−log10(P) = 39.04
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 130
N = 242
N = 99

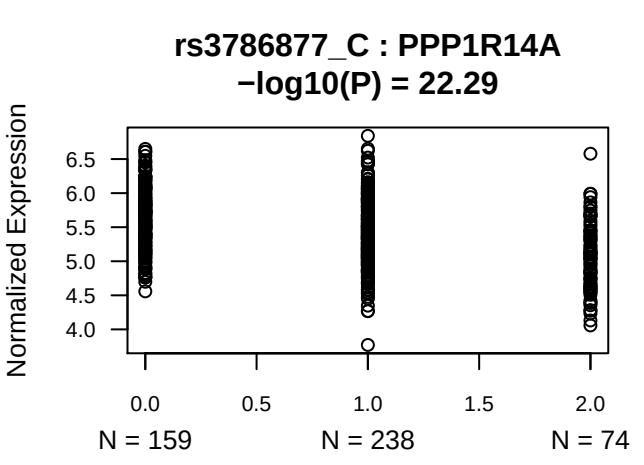

rs3786877_C : PPP1R14A
−log10(P) = 22.29
Normalized Expression
6.5
6.0
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0
N = 159
N = 238
N = 74

Table 1: List of PC risk-SNPs used for the study, including Chromosome location

SNP ID	Chr.	GRCh37/hg19 Position	Alleles (risk)	Intra/intergenic	Gene	LD region	
rs636291	1p35	10,556,097	A/G			chr1.10456097.10656097.112	unpublished meta analysis
rs17599629	1q21	150,658,287	G/A	intronic	GOLPH3L	chr1.150558287.150758287.49	unpublished meta analysis
rs1218582	1q21.3	154,834,183	A/G	Intragenic-intron	*KCNN3*	chr1.154734183.154934183.172	Eeles 2013
rs4245739	1q32.1	204,518,842	A/C	Intragenic-intron	*MDM4*	chr1.204418842.204618842.68	Eeles 2013
rs1775148	1q32	205,757,824	C/T			chr1.205657824.205857824.106	unpublished meta analysis
rs11902236	2p25.1	10,117,868	G/A	Intragenic-intron	GRHL1	chr2.10017868.10217868.182	Eeles 2013
rs9287719	2p25	10,710,730	C/T	161 bp 3'	NOL10	chr2.10610730.10810730.147	unpublished meta analysis
rs13385191	2p24.1	20,888,265	A/G	Intragenic-intron	C2orf43	chr2.20788265.20988265.160	Takata 2010
rs1465618	2p21	43,553,949	A/G	Intragenic-intron	THADA	chr2.43453949.43653949.82	Eeles 2009
rs721048	2p15	63,131,731	A/G	Intragenic-intron	EHBP1	chr2.63031731.63401164.88	Gudmundsson 2008
rs6545977	2p15	63,301,164	A/G	Intergenic		chr2.63031731.63401164.88	Eeles 2009
rs2028898	2p11.2	85,777,270	C/T	Intragenic-intron	*GGCX*	chr2.85677270.85894297.100	Akamatsu 2012
rs10187424	2p11.2	85,794,297	A/C	Intergenic		chr2.85677270.85894297.100	Kote-Jarai 2011
rs12621278	2q31.1	173,311,553	A/G	Intragenic-intron	ITGA6	chr2.173211553.173411553.175	Eeles 2009
rs7584330	2q37.3	238,387,228	A/G/C	Intergenic		chr2.238287228.238543226.217	Kote-Jarai 2011
rs2292884	2q37.3	238,443,226	A/G	intragenic-intron	MLPH	chr2.238287228.238543226.217	Schumacher 2011
rs3771570	2q37.3	242,382,864	G/A	Intragenic-intron	*FARP2*	chr2.242282864.242482864.97.	Eeles 2013
rs9311171	3p22.2	37,996,477	G/T	Intragenic-intron	CTDSPL	chr3.37896477.38096477.85	Murabito 2007
rs2660753	3p12.1	87,110,674	C/T	Intergenic		chr3.87010674.87341497.136	Eeles 2008
rs9284813	3p12.1	87,152,169	A/G	Intergenic		chr3.87010674.87341497.136	Takata 2010
rs17181170	3p12.1	87,173,324	A/G	Intergenic		chr3.87010674.87341497.136	Eeles 2009
rs7629490	3p11.2	87,241,497	C/T	Intergenic		chr3.87010674.87341497.136	Schumacher 2011
rs2055109	3p11.2	87,467,332	C/T	Intergenic		chr3.87367332.87567332.101.	Akamatsu 2012
rs7611694	3q13.2	113,275,624	A/C	Intragenic-intron	*SIDT1*	chr3.113175624.113375624.105	Eeles 2013
rs10934853	3q21.3	128,038,373	A/C	Intragenic-intron	EEFSEC	chr3.127938373.128138373.49	Gudmundsson 2009
rs6763931	3q23	141,102,833	G/A/T	intragenic-intron	ZBTB38	chr3.141002833.141202833.103	Kote-Jarai 2011
rs345013	3q24	145,173,788	C/T	Intergenic		chr3.145073788.145273788.65.	Murabito 2007
rs10936632	3q26.2	170,130,102	A/C	Intergenic		chr3.170030102.170230102.97	Kote-Jarai 2011
rs10009409	4q13	73,855,253	T/C	65 kb 3'	COX18	chr4.73755253.73955253.78	unpublished meta analysis
rs1894292	4q13.3	74,349,158	G/A	Intragenic-intron	AFM	chr4.74249158.74449158.75	Eeles 2013
rs12500426	4q22.3	95,514,609	A/C	Intragenic-intron	PDLIM5	chr4.95414609.95662877.135.	Eeles 2009
rs17021918	4q22.3	95,562,877	C/T	Intragenic-intron	PDLIM5	chr4.95414609.95662877.135.	Eeles 2009
rs7679673	4q24	106,061,534	A/C	Intergenic		chr4.105961534.106161534.83	Eeles 2009
rs2242652	5p15.33	1,280,028	G/T	intragenic-intron	TERT	chr5.1180028.1398733.193	Kote-Jarai 2011, Nam 2012
rs7725218	5p15.33	1,282,414	A/G	intragenic-intron	TERT	chr5.1180028.1398733.193	Kote Jarai 2013
rs2853676	5p15.33	1,288,547	A/G	intragenic-intron	TERT	chr5.1180028.1398733.193	Kote Jarai 2013
rs13190087	5p15.33	1,298,733	A/C	intergenic		chr5.1180028.1398733.193	Kote Jarai 2013
rs12653946	5p15.33	1,895,829	C/T	Intergenic		chr5.1795829.1995829.220	Takata 2010, Cheng 2012
rs2121875	5p12	44,365,545	G/T	intragenic-intron	FGF10	chr5.44265545.44465545.72.	Kote-Jarai 2011
rs4466137	5q14.3	82,985,739	G/T	Intragenic-intron	HAPLN1	chr5.82885739.83085739.124	Murabito 2007
rs37181	5q23.1	115630004	A/C	intergenic		chr5.115530004.115730004.117	Kote-Jarai 2011
rs6869841	5q35.2	172,939,426	G/A	Intergenic		chr5.172839426.173039426.176	Eeles 2013
rs4713266	6p24	11,219,030	C/T	intronic	NEDD9	chr6.11119030.11319030.200	unpublished meta analysis
rs115457135 (rs7767188)	6p22	30,073,776	A/G	intronic	TRIM31	chr6.29973776.30173776.600	unpublished meta analysis
rs130067	6p21.33	31,118,511	A/G	missense	CCHCR1	chr6.31018511.31218511.772	Kote-Jarai 2011
rs3096702 (rs114376585)	6p21.32	32,192,331	G/A	Intergenic		chr6.32092331.32292331.467	Eeles 2013
rs115306967	6p21	32,400,939	G/C	5 kb 5'	HLA-DRB6	chr6.32300939.32500939.715	unpublished meta analysis
rs1983891	6p21.1	41,536,427	C/T	Intragenic-intron	FOXP4	chr6.41436427.41636427.190	Takata 2010
rs10498792	6p12.3	51,666,631	C/T	Intragenic-intron	PKHD1	chr6.51566631.51766631.91	Murabito 2007
rs9443189	6q14	76,495,882	G/A			chr6.76395882.76595882.52.	unpublished meta analysis
rs2273669	6q21	109,285,189	A/G	Intragenic-intron	ARMC2	chr6.109185189.109385189.70	Eeles 2013
rs339331	6q22.1	117,210,052	C/T	Intragenic-intron	RFX6	chr6.117110052.117310052.82	Takata 2010
rs1933488	6q25.2	153,441,079	A/G	Intragenic-intron	*RGS17*	chr6.153341079.153541079.117	Eeles 2013
rs651164	6q25.3	160,581,374	A/G	Intergenic		chr6.160481374.160681374.169	Eeles 2009, Schumacher 2011
rs9364554	6q25.3	160,833,664	C/T	Intragenic-intron	SLC22A3	chr6.160,733,664.160933664.151	Eeles 2008
rs12155172	7p15.3	20,994,491	A/G	Intergenic		chr7.20894491.21094491.113	Eeles 2009, 2013
rs10486567	7p15.2	27,976,563	A/G	Intragenic-intron	JAZF1	chr7.27876563.28076563.143	Thomas 2008, Zheng 2009
rs56232506	7p12	47,437,244	A/G	intronic	TNS3	chr7.47337244.47537244.142	unpublished meta analysis
rs6465657	7q21.3	97,816,327	C/T	Intragenic-intron	LMTK2	chr7.97716327.97916327.72	Eeles 2008, 2009
rs2928679	8p21.2	23,438,975	C/T	Intergenic		chr8.23338975.23626463.307	none?
rs1512268	8p21.2	23,526,463	A/G/T	Intergenic		chr8.23338975.23626463.307	Eeles 2009, Takata 2010, Cheng 2012
rs11135910	8p21.2	25,892,142	G/A	Intragenic-intron	*EBF2*	chr8.25792142.25992142.187.	Eeles 2013
rs979200	8q24.21	127,923,720	A/G	Intergenic		chr8.127823720.128723639.822	Salinas 2008
rs12543663	8q24.21	127,924,659	A/C	Intergenic		chr8.127823720.128723639.822	Al Olama 2009
rs10086908	8q24.21	128,011,937	T/C	Intergenic		chr8.127823720.128723639.822	Al Olama 2009
rs1016343	8q24.21	128,093,297	C/T	Intergenic		chr8.127823720.128723639.822	Eeles 2008, Schumacher 2011
rs13252298	8q24.21	128,095,156	A/G	Intergenic		chr8.127823720.128723639.822	Schumacher 2011
rs1456315	8q24.21	128,103,937	A/G	Intergenic		chr8.127823720.128723639.822	Takata 2010
rs13254738	8q24.21	128,104,343	A/C	Intergenic		chr8.127823720.128723639.822	Salinas2008, Haiman 2007, Cheng 2012
rs6983561	8q24.21	128,106,880	A/C	Intergenic		chr8.127823720.128723639.822	Salinas2008, Haiman 2007, Cheng 2012
rs16901979	8q24.21	128,124,916	A/C	Intergenic		chr8.127823720.128723639.822	Gudmundsson 2007A, 2009, Zheng 2009
rs10505483	8q24.21	128,125,195	A/T	Intergenic		chr8.127823720.128723639.822	Cheng 2012
rs16902094	8q24.21	128,320,346	A/G	Intergenic		chr8.127823720.128723639.822	Gudmundsson 2009
rs445114	8q24.21	128,323,181	C/T	Intergenic		chr8.127823720.128723639.822	Gudmundsson 2009, Schumacher 2011
rs620861	8q24.21	128,335,673	C/T	Intergenic		chr8.127823720.128723639.822	Al Olama 2009
rs6983267	8q24.21	128,413,305	G/T	Intergenic		chr8.127823720.128723639.822	Eeles 2008, Thomas 2008, Yeager 2007, Zheng 2009, Schumacher 2011
rs7837328	8q24.21	128,423,127	A/G	Intergenic		chr8.127823720.128723639.822	Yeager 2007
rs7000448	8q24.21	128,441,170	C/T	Intergenic		chr8.127823720.128723639.822	Salinas2008, Haiman 2007
rs1447295	8q24.21	128,485,038	A/C	Intergenic		chr8.127823720.128723639.822	Gudmundsson 2007A, 2009, Yeager 2007
rs4242382	8q24.21	128,517,573	A/G	Intergenic		chr8.127823720.128723639.822	Thomas 2008
rs4242384	8q24.21	128,518,554	A/C	Intergenic		chr8.127823720.128723639.822	Eeles 2008, 2009, Schumacher 2011
rs10090154	8q24.21	128,532,137	C/T	Intergenic		chr8.127823720.128723639.822	Salinas 2008, Haiman 2007, Cheng 2012
rs7837688	8q24.21	128,539,360	G/T	Intergenic		chr8.127823720.128723639.822	Takata 2010
rs7005795	8q24.21	128,623,639	G/T	Intergenic		chr8.127823720.128723639.822	none?
rs17694493	9p21	22,041,998	G/C	intronic	CDKN2B-AS1	chr9.21941998.22141998.99	unpublished meta analysis
rs817826	9q31.2	110,156,300	T/C	Intergenic		chr9.110056300.110256300.186	Xu 2012
rs1571801	9q33.2	124,427,373	A/C	Intragenic-intron	DAB2IP	chr9.124327373.124527373.113	Duggan 2007
rs76934034	10q11	46,082,985	T/C	intronic	MARCH8	chr10.45982985.46182985.48	unpublished meta analysis
rs3123078	10q11.23	51,524,971	C/T	Intergenic		chr10.51424971.51649496.56.	Eeles 2009
rs10993994	10q11.23	51,549,496	C/T	Intergenic		chr10.51424971.51649496.56.	Eeles 2008, Thomas 2008, Takata 2010, Schumacher 2011
rs3850699	10q24.32	104,414,221	A/G	Intragenic-intron	*TRIM8*	chr10.104314221.104514221.67	Eeles 2013
rs2252004	10q26.12	122,844,709	G/T	Intergenic		chr10.122744709.123132519.253.	Akamatsu 2012
rs11199874	10q26.12	123,032,519	A/G	Intergenic		chr10.122744709.123132519.253.	Nam 2012
rs4962416	10q26.13	126,696,872	C/T	Intragenic-intron	CTBP2	chr10.126596872.126796872.265	Thomas 2008
rs7127900	11p15.5	2,233,574	A/G	Intergenic		chr11.2133574.2333574.160.	Eeles 2009
rs1938781	11q12	58,915,110	T/C	Intragenic-intron	*FAM111A*	chr11.58815110.59015110.73	Akamatsu 2012
rs12418451	11q13.2	68935419	A/G	intergenic	LNCRNA RP11-554A11.8	chr11.68835419.69095958.206	Zheng 2009
rs11228565	11q13.2	68,978,580	A/G	Intergenic		chr11.68835419.69095958.206	Gudmundsson 2009
rs7931342	11q13.2	68,994,497	G/T	Intergenic		chr11.68835419.69095958.206	Eeles 2008
rs10896449	11q13.2	68,994,667	A/G	Intergenic		chr11.68835419.69095958.206	Thomas 2008, Zheng 2009
rs7130881	11q13.3	68,995,958	A/G	Intergenic		chr11.68835419.69095958.206	Eeles 2009, Schumacher 2011
rs11568818	11q22.2	102,401,661	A/G	Intergenic		chr11.102301661.102501661.177	Eeles 2013
rs11214775	11q23	113,807,181	G/A	intronic	HTR3B	chr11.113707181.113907181.105	unpublished meta analysis
rs731236	12q13.12	48238757	C/T	synonymous	VDR	chr12.48138757.48519618.295.	Bonilla 2011
rs80130819	12q13	48,419,618	A/C	17 kb 3'	SENP1	chr12.48138757.48519618.295.	unpublished meta analysis
rs10875943	12q13.12	49,676,010	C/T	Intergenic		chr12.49576010.49776010.86.	Kote-Jarai 2011
rs902774	12q13.13	53,273,904	A/G	Intergenic		chr12.53173904.53373904.144	Schumacher 2011

rs12827748	12q21.31	80088578	C/T	intergenic		chr12.79988578.80188578.44	Bonilla 2011
rs1270884	12q24.21	114,685,571	G/A	Intergenic		chr12.114585571.114785571.183.	Eeles 2013
rs9600079	13q22.1	73,728,139	G/T	Intergenic		chr13.73628139.73828139.181	Takata 2010
rs1529276	13q33.1	103,928,007	A/T	Intergenic		chr13.103828007.104028007.161	Murabito 2007
rs8008270	14q22.1	53,372,330	G/A	Intragenic-intron	*FERMT2*	chr14.53272330.53472330.72.	Eeles 2013
rs7153648	14q23	61,122,526	C/G			chr14.61022526.61222526.42.	unpublished meta analysis
rs7141529	14q24.1	69,126,744	A/G	Intragenic-intron	*RAD51L1*	chr14.69026744.69226744.245	Eeles 2013
rs8014671	14q24	71,092,256	G/A	16 kb 5'	TTC9	chr14.70992256.71192256.161	unpublished meta analysis
rs4775302	15q21.1	46,639,808	A/G	Intergenic		chr15.46539808.46739808.116.	Nam 2013
rs12051443	16q22	71,691,329	A/G			chr16.71591329.71791329.67.	unpublished meta analysis
rs684232	17p13.3	618,965	A/G	Intergenic		chr17.518965.718965.117.	Eeles 2013
rs11649743	17q12	36,074,979	A/G	Intragenic-intron	HNF1B	chr17.35974979.36201156.186	Sun 2008, Levin 2008
rs4430796	17q12	36,098,040	**A**/G	Intragenic-intron	HNF1B	chr17.35974979.36201156.186	Thomas 2008, Gudmundsson 2007, Levin 2008, Eeles 2009, Gudmundsson 2009
rs7501939	17q12	36,101,156	**C**/T	Intragenic-intron	HNF1B	chr17.35974979.36201156.186	Eeles 2008, Sun 2008, Levin 2008, Takata 2010, Schumacher 2011
rs11650494	17q21.32	47,345,186	G/A	Intergenic		chr17.47245186.47536749.149.	Eeles 2013
rs1859962	17q24.3	69,108,753	**G**/T	Intergenic		chr17.69008753.69208753.130	Eeles 2008, Gudmundsson 2007, Levin 2008, Eeles 2009, Schumacher 2011,
rs7241993	18q23	76,773,973	G/A	Intergenic		chr18.76673973.76873973.112	Eeles 2013
rs8102476	19q13.2	38,735,613	C/T	Intergenic		chr19.38635613.38835613.112	Gudmundsson 2009
rs11672691	19q13.2	41,985,587	G/A	Intragenic-intron	LOC100505495	chr19.41885587.42085624.89.	Al Olama 2012, 2013
rs887391	19q13.2	41,985,624	C/T	Intergenic		chr19.41885587.42085624.89.	Hsu 2009
rs2735839	19q13.33	51,364,623	A/**G**	Intergenic		chr19.51264623.51464623.230	Eeles 2008
rs103294	19q13.4	54,797,848	T/C	Intragenic-intron	*LILRA3*	chr19.54697848.54897848.146	Xu 2012
rs12480328	20q13	49,527,922	T/C			chr20.49427922.49627922.124	unpublished meta analysis
rs2427345	20q13.33	61,015,611	G/A	Intergenic		chr20.60915611.61115611.153	Eeles 2013
rs6062509	20q13.33	62,362,563	A/C	Intragenic-intron	*ZGPAT*	chr20.62262563.62462563.84.	Eeles 2013
rs1041449	21q22	42,901,421	G/A			chr21.42801421.43001421.189	unpublished meta analysis
rs2238776	22q11	19,757,892	G/A			chr22.19657892.19857892.132	unpublished meta analysis
rs11704416	22q13.1	40,436,973	G/C	Intergenic		chr22.40336973.40552119.71	Al Olama 2012, 2013
rs9623117	22q13.1	40,452,119	C/T	Intragenic-intron	TNRC6B	chr22.40336973.40552119.71	Sun 2009
rs5759167	22q13.2	43,500,212	G/T	Intergenic		chr22.43400212.43618275.191	Eeles 2009
rs742134	22q13.2	43,518,275	A/G	intragenic-intron	BIK	chr22.43400212.43618275.191	Schumacher 2011
rs2405942	Xp22.2	9,814,135	A/G	Intragenic-intron	*SHROOM2*	chr23.9714135.9914135.117	Eeles 2013
rs1327301	Xp11.22	51,210,057	C/T	Intergenic		chr23.51110057.51341672.32	Eeles 2009
rs5945572	Xp11.22	51,229,683	**A**/G	Intergenic		chr23.51110057.51341672.32	Gudmundsson 2008
rs5945619	Xp11.22	51,241,672	**C**/T	Intergenic		chr23.51110057.51341672.32	Eeles 2008
rs2807031	Xp11	52,896,949	C/T	intronic	XAGE3	chr23.52796949.52996949.17	unpublished meta analysis
rs5919432	Xq12	67,021,550	C/A	Intergenic		chr23.66921550.67121550.26	Kote-Jarai 2011
rs6625711	Xq13	70,139,850	A/T	36 kb 3'	SLC7A	chr23.70039850.70239850.62	unpublished meta analysis
rs4844289	Xq13	70,407,983	G/A	16kb 3'	NLGN3	chr23.70307983.70507983.74	unpublished meta analysis

Table 2: Number of SNPs and number of genes evaluated for each of the risk regions

Risk SNP ID	LD region for SNP evaluation	2 Mb ROI for gene evaluation	# SNPs evaluated (nSNPs)	# genes in ROI (total)	# genes evaluated (ngene)	# tests (Nfreq)
rs636291	chr1.10456097.10656097.112	chr1.9456097.11656097.112	71	27	23	1662
rs17599629	chr1.150558287.150758287.49	chr1.149558287.151758287.49	11	75	61	670
rs1218582	chr1.154734183.154934183.172	chr1.153734183.155934183.172	87	76	63	5411
rs4245739	chr1.204418842.204618842.68	chr1.203418842.205618842.68	131	36	28	3683
rs1775148	chr1.205657824.205857824.106	chr1.204657824.206857824.106	40	31	27	1100
rs11902236	chr2.10017868.10217868.182	chr2.9017868.11217868.182	11	21	18	198
rs9287719	chr2.10610730.10810730.147	chr2.9610730.11810730.147	269	26	19	4667
rs13385191	chr2.20788265.20988265.160	chr2.19788265.21988265.160	11	13	12	132
rs1465618	chr2.43453949.43653949.82	chr2.42453949.44653949.82	8	19	17	136
rs721048, rs6545977	chr2.63031731.63401164.88	chr2.62031731.64401164.88	41	14	13	460
rs2028898, rs10187424	chr2.85677270.85894297.100	chr2.84677270.86894297.100	111	37	34	3727
rs12621278	chr2.173211553.173411553.175	chr2.172211553.174411553.175	106	16	13	1354
rs7584330, rs2292884	chr2.238287228.238543226.217	chr2.237287228.239543226.217	258	23	19	4902
rs3771570	chr2.242282864.242482864.97.	chr2.241282864.243482864.97	69	35	27	1863
rs9311171	chr3.37896477.38096477.85	chr3.36896477.39096477.85	27	25	21	544
rs2660753, rs9284813, rs17181170, rs762949(	chr3.87010674.87341497.136	chr3.86010674.88341497.136	165	9	7	1126
rs2055109	chr3.87367332.87567332.101.	chr3.86367332.88567332.101	87	8	6	522
rs7611694	chr3.113175624.113375624.105	chr3.112175624.114375624.105	24	26	23	552
rs10934853	chr3.127938373.128138373.49	chr3.126938373.129138373.49	44	33	27	1192
rs6763931	chr3.141002833.141202833.103	chr3.140002833.142202833.103	51	14	11	559
rs345013	chr3.145073788.145273788.65.	chr3.144073788.146273788.65	89	4	4	324
rs10936632	chr3.170030102.170230102.97	chr3.169030102.171230102.97	30	21	16	480
rs10009409	chr4.73755253.73955253.78	chr4.72755253.74955253.78	1	18	10	10
rs1894292	chr4.74249158.74449158.75	chr4.73249158.75449158.75	2	22	14	28
rs12500426, rs1702191(	chr4.95414609.95662877.135.	chr4.94414609.96662877.135	170	7	5	850
rs7679673	chr4.105961534.106161534.83	chr4.104961534.107161534.83	21	8	8	177
rs2242652, rs7725218, rs2853676, rs1319008(	chr5.1180028.1398733.193	chr5.180028.2398733.193	22	33	25	545
rs12653946	chr5.1795829.1995829.220	chr5.795829.2995829.220	12	22	16	192
rs2121875	chr5.44265545.44465545.72.	chr5.43265545.45465545.72	124	10	9	1114
rs4466137	chr5.82885739.83085739.124	chr5.81885739.84085739.124	26	7	5	130
rs37181	chr5.115530004.115730004.117	chr5.114530004.116730004.117	71	13	10	733
rs6869841	chr5.172839426.173039426.176	chr5.171839426.174039426.176	19	17	13	247
rs4713266	chr6.11119030.11319030.200	chr6.10119030.12319030.200	2	20	18	36
rs115457135 (rs7767188)	chr6.29973776.30173776.600	chr6.28973776.31173776.600	178	79	47	8327
rs130067	chr6.31018511.31218511.772	chr6.30018511.32218511.772	21	129	95	1972
rs3096702 (rs114376585)	chr6.32092331.32292331.467	chr6.31092331.33292331.467	11	129	103	1133
rs115306967	chr6.32300939.32500939.715	chr6.31300939.33500939.715	108	126	101	10908
rs1983891	chr6.41436427.41636427.190	chr6.40436427.42636427.190	56	34	25	1408
rs10498792	chr6.51566631.51766631.91	chr6.50566631.52766631.91	94	18	9	846
rs9443189	chr6.76395882.76595882.52.	chr6.75395882.77595882.52	66	8	8	528
rs2273669	chr6.109185189.109385189.70	chr6.108185189.110385189.70	183	19	15	2823
rs339331	chr6.117110052.117310052.82	chr6.116110052.118310052.82	100	24	17	1691
rs1933488	chr6.153341079.153541079.117	chr6.152341079.154541079.117	76	9	8	608
rs651164	chr6.160481374.160681374.169	chr6.159481374.161681374.169	1	22	17	17
rs9364554	chr6.160733664.160933664.151	chr6.159733664.161933664.151	23	22	17	391
rs12155172	chr7.20894491.21094491.113	chr7.19894491.22094491.113	12	8	8	103
rs10486567	chr7.27876563.28076563.143	chr7.26876563.29076563.143	28	27	26	725
rs56232506	chr7.47337244.47537244.142	chr7.46337244.48537244.142	19	10	5	95
rs6465657	chr7.97716327.97916327.72	chr7.96716327.98916327.72	43	17	13	593
rs2928679, rs1512268	chr8.23338975.23626463.307	chr8.22338975.24626463.307	128	29	25	3147
rs11135910	chr8.25792142.25992142.187.	chr8.24792142.26992142.187	39	11	11	403
rs979200, rs12543663, rs10086908, rs1016343, rs13252298, rs1456315, rs13254738, rs6983561, rs16901979, rs10505483, rs16902094, rs445114, rs620861, rs6983267, rs7837328, rs7000448, rs1447295, rs4242382, rs4242384, rs10090154, rs7837688, rs7005795	chr8.127823720.128723639.822	chr8.126823720.129723639.822	462	12	6	2365
rs17694493	chr9.21941998.22141998.99	chr9.20941998.23141998.99	33	29	10	329
rs817826	chr9.110056300.110256300.186	chr9.109056300.111256300.186	2	4	3	6
rs1571801	chr9.124327373.124527373.113	chr9.123327373.125527373.113	11	35	22	242
rs76934034	chr10.45982985.46182985.48	chr10.44982985.47182985.48	2	30	14	28
rs3123078, rs10993994	chr10.51424971.51649496.56.	chr10.50424971.52649496.56	135	23	14	1912
rs3850699	chr10.104314221.104514221.67	chr10.103314221.105514221.67	46	52	39	1724
rs2252004	chr10.122744709.123132519.253.	chr10.121744709.124132519.253	157	8	7	1129
rs4962416	chr10.126596872.126796872.265	chr10.125596872.127796872.265	25	23	18	450
rs7127900	chr11.2133574.2333574.160.	chr11.1133574.3333574.160	84	53	32	2688
rs1938781	chr11.58815110.59015110.73	chr11.57815110.60015110.73	134	44	16	2135
rs12418451, rs11228565, rs7931342, rs10896449, rs7130881	chr11.68835419.69095958.206	chr11.67835419.70095958.206	99	21	16	1555
rs11568818	chr11.102301661.102501661.177	chr11.101301661.103501661.177	3	21	13	39
rs11214775	chr11.113707181.113907181.105	chr11.112707181.114907181.105	12	20	15	180
rs731236, rs80130819	chr12.48138757.48519618.295.	chr12.47138757.49519618.295	51	47	33	1436
rs10875943	chr12.49576010.49776010.86.	chr12.48576010.50776010.86	54	63	45	2418
rs902774	chr12.53173904.53373904.144	chr12.52173904.54373904.144	52	68	41	2157
rs12827748	chr12.79988578.80188578.44	chr12.78988578.81188578.44	6	8	5	30
rs1270884	chr12.114585571.114785571.183.	chr12.113585571.115785571.183	35	14	13	469
rs9600079	chr13.73628139.73828139.181	chr13.72628139.74828139.181	18	6	6	108
rs1529276	chr13.103828007.104028007.161	chr13.102828007.105028007.161	33	13	7	460
rs8008270	chr14.53272330.53472330.72.	chr14.52272330.54472330.72	44	14	14	308
rs7153648	chr14.61022526.61222526.42.	chr14.60022526.62222526.42	87	17	14	1270
rs7141529	chr14.69026744.69226744.245	chr14.68026744.70226744.245	6	18	16	101
rs8014671	chr14.70992256.71192256.161	chr14.69992256.72192256.161	13	22	13	171
rs4775302	chr15.46539808.46739808.116.	chr15.45539808.47739808.116	53	11	10	530
rs12051443	chr16.71591329.71791329.67.	chr16.70591329.72791329.67	112	27	23	2624
rs684232	chr17.518965.718965.117.	chr17.-481035.1718965.117	75	34	30	2168
rs11649743, rs4430796, rs750193(	chr17.35974979.36201156.186	chr17.34974979.37201156.186	30	36	25	740
rs11650494	chr17.47245186.47536749.149.	05705.47905705.122, chr17.46245186.48	75	57	42	3160
rs1859962	chr17.69008753.69208753.130	chr17.68008753.70208753.130	129	4	3	387
rs7241993	chr18.76673973.76873973.112	chr18.75673973.77873973.112	13	10	9	113
rs8102476	chr19.38635613.38835613.112	chr19.37635613.39835613.112	17	60	51	866
rs11672691	chr19.41885587.42085624.89.	chr19.40885587.43085624.89	20	70	56	1112
rs2735839	chr19.51264623.51464623.230	chr19.50264623.52464623.230	4	109	73	292
rs103294	chr19.54697848.54897848.146	chr19.53697848.55897848.146	35	136	56	1953
rs12480328	chr20.49427922.49627922.124	chr20.48427922.50627922.124	52	22	19	988
rs2427345	chr20.60915611.61115611.153	chr20.59915611.62115611.153	16	49	28	448
rs6062509	chr20.62262563.62462563.84.	chr20.61262563.63462563.84	117	70	46	5382
rs1041449	chr21.42801421.43001421.189	chr21.41801421.44001421.189	14	27	19	266

rs2238776	chr22.19657892.19857892.132	chr22.18657892.20857892.132	17	52	37	629
rs11704416, rs9623117	chr22.40336973.40552119.71	chr22.39336973.41552119.71	71	39	33	2296
rs5759167, rs742134	chr22.43400212.43618275.191	chr22.42400212.44618275.191	23	36	32	731
rs2405942	chr23.9714135.9914135.117	chr23.8714135.10914135.117	40	9	6	250
rs1327301, rs5945572, rs594561	chr23.51110057.51341672.32	chr23.50110057.52341672.32	130	20	5	663
rs2807031	chr23.52796949.52996949.17	chr23.51796949.53996949.17	42	31	10	419
rs5919432	chr23.66921550.67121550.26	chr23.65921550.68121550.26	179	5	5	898
rs6625711	chr23.70039850.70239850.62	chr23.69039850.71239850.62	19	37	25	475
rs4844289	chr23.70307983.70507983.74	chr23.69307983.71507983.74	40	39	28	1132