## simple and multiple regression of the phenotype

Load the genotype data.

```
if(!file.access(EXOME.GENO.FN)==0 | !file.access(SEQ.GENO.FN)==0) {
  ## genotype data has not yet been processed
  source('plink_genotype_preprocessing.R')
} else {
  ## load processed genotype data
  load(EXOME.GENO.FN)
  load(SEQ.GENO.FN)
}
```

```
## Reading map from file '../data/qc_report/consensus/consensus.nomiss.map' ...
## ... done.  Read positions of 8732 markers from file '../data/qc_report/consensus/consensu
## Reading genotypes from file '../data/qc_report/consensus/consensus.nomiss.ped' ...
## ...done.  Read information for 226 people from file '../data/qc_report/consensus/consensu
## Analysing marker information ...
## Writing to file '../data/qc_report/consensus/consensus.genabel' ...
## ... done.
## Reading map from file '../data/exomechip_data/innocenti_082613.nomiss.map' ...
## ... done.  Read positions of 242895 markers from file '../data/exomechip_data/innocenti_0
## Reading genotypes from file '../data/exomechip_data/innocenti_082613.nomiss.ped' ...
##   ... read 10201590 genotypes ...
##   ... read 20160285 genotypes ...
##   ... read 30118980 genotypes ...
##   ... read 40077675 genotypes ...
##   ... read 50036370 genotypes ...
## ...done.  Read information for 226 people from file '../data/exomechip_data/innocenti_082
## Analysing marker information ...
##   ... analysed 10000048 genotypes ...
##   ... analysed 20000096 genotypes ...
##   ... analysed 30000144 genotypes ...
##   ... analysed 40000192 genotypes ...
##   ... analysed 50000014 genotypes ...
## Writing to file '../data/exomechip_data/innocenti_082613.genabel' ...
## ... done.
## ids loaded...
## marker names loaded...
## chromosome data loaded...
## map data loaded...
## allele coding data loaded...
## strand data loaded...
## genotype data loaded...
## snp.data object created...
```

```
## assignment of gwaa.data object FORCED; X-errors were not checked!
## Excluding people/markers with extremely low call rate...
## 8732 markers and 226 people in total
## 0 people excluded because of call rate < 0.1
## 207 markers excluded because of call rate < 0.1
## Passed: 8525 markers and 226 people
##
## RUN 1
## 8525 markers and 226 people in total
## 6093 (71.47%) markers excluded as having low (<5%) minor allele frequency
## 275 (3.226%) markers excluded because of low (<95%) call rate
## 213 (2.499%) markers excluded because they are out of HWE (FDR <0.2)
## 0 (0%) people excluded because of low (<95%) call rate
## Mean autosomal HET is 0.319 (s.e. 0.02167)
## 0 people excluded because too high autosomal heterozygosity (FDR <1%)
## Mean IBS is 0.7364 (s.e. 0.01504), as based on 2000 autosomal markers
## 0 (0%) people excluded because of too high IBS (>=0.95)
## In total, 2135 (25.04%) markers passed all criteria
## In total, 226 (100%) people passed all criteria
##
## RUN 2
## 2135 markers and 226 people in total
## 0 (0%) markers excluded as having low (<5%) minor allele frequency
## 0 (0%) markers excluded because of low (<95%) call rate
## 0 (0%) markers excluded because they are out of HWE (FDR <0.2)
## 0 (0%) people excluded because of low (<95%) call rate
## Mean autosomal HET is 0.319 (s.e. 0.02167)
## 0 people excluded because too high autosomal heterozygosity (FDR <1%)
## Mean IBS is 0.736 (s.e. 0.01508), as based on 2000 autosomal markers
## 0 (0%) people excluded because of too high IBS (>=0.95)
## In total, 2135 (100%) markers passed all criteria
## In total, 226 (100%) people passed all criteria
## ids loaded...
## marker names loaded...
## chromosome data loaded...
## map data loaded...
## allele coding data loaded...
## strand data loaded...
## genotype data loaded...
## snp.data object created...
## assignment of gwaa.data object FORCED; X-errors were not checked!
## Excluding people/markers with extremely low call rate...
## 242895 markers and 165 people in total
## 0 people excluded because of call rate < 0.1
## 33 markers excluded because of call rate < 0.1
## Passed: 242862 markers and 165 people
```

```
##
## RUN 1
## 242862 markers and 165 people in total
## 215649 (88.79%) markers excluded as having low (<5%) minor allele frequency
## 162 (0.0667%) markers excluded because of low (<95%) call rate
## 775 (0.3191%) markers excluded because they are out of HWE (FDR <0.2)
## 1 (0.6061%) people excluded because of low (<95%) call rate
## Mean autosomal HET is 0.3545 (s.e. 0.007301)
## 1 (0.6061%) people excluded because too high autosomal heterozygosity (FDR <1%)
## Excluded people had HET >= 0.3953
## Mean IBS is 0.7134 (s.e. 0.007848), as based on 2000 autosomal markers
## 0 (0%) people excluded because of too high IBS (>=0.95)
## In total, 26578 (10.94%) markers passed all criteria
## In total, 164 (99.39%) people passed all criteria
##
## RUN 2
## 26578 markers and 164 people in total
## 47 (0.1768%) markers excluded as having low (<5%) minor allele frequency
## 0 (0%) markers excluded because of low (<95%) call rate
## 0 (0%) markers excluded because they are out of HWE (FDR <0.2)
## 0 (0%) people excluded because of low (<95%) call rate
## Mean autosomal HET is 0.3547 (s.e. 0.006596)
## 0 people excluded because too high autosomal heterozygosity (FDR <1%)
## Mean IBS is 0.7122 (s.e. 0.007808), as based on 2000 autosomal markers
## 0 (0%) people excluded because of too high IBS (>=0.95)
## In total, 26531 (99.82%) markers passed all criteria
## In total, 164 (100%) people passed all criteria
##
## RUN 3
## 26531 markers and 164 people in total
## 0 (0%) markers excluded as having low (<5%) minor allele frequency
## 0 (0%) markers excluded because of low (<95%) call rate
## 0 (0%) markers excluded because they are out of HWE (FDR <0.2)
## 0 (0%) people excluded because of low (<95%) call rate
## Mean autosomal HET is 0.3547 (s.e. 0.006596)
## 0 people excluded because too high autosomal heterozygosity (FDR <1%)
## Mean IBS is 0.716 (s.e. 0.007817), as based on 2000 autosomal markers
## 0 (0%) people excluded because of too high IBS (>=0.95)
## In total, 26531 (100%) markers passed all criteria
## In total, 164 (100%) people passed all criteria

if(!file.access(PHENO.FN)==0) {
  source(purl('phenotype_preprocessing.R'))
} else {
  load(PHENO.FN)
}
```

## Model specification

Here we define two models: a simple regression of SNP on phenotype, and a full model including available covariates. Both sample sex and site of ascertainment are natural categorical covariates. The encoding of the dosing regimen deserves some careful consideration. This can be treated as a continuous covariate, a categorical covariate, or perhaps some different altogether. Currently, I am treating it as a continuous variable. I am able to include it in fitting the SKAT-O model as an added benefit. We simply do not have the degrees of freedom necessary (i.e. number of samples) to include 33 dummy variables.

```
trans.fun <- my.invnorm
basic.model <- trans.fun(ANC) ~ 1
full.model <- trans.fun(ANC) ~ sex + site + as.numeric(dose)
pca.model <- trans.fun(ANC) ~ sex + site + as.numeric(dose) + eigenvect.1 + eigenvect.2 + ei
```

## PGRNseq GWAS

### Simple Regression

```
seq.simplereg.results <- mlreg(basic.model, seq.geno, trait="gaussian")
qqunif(seq.simplereg.results[,"P1df"])
title('PGRNseq Simple Linear Regression GWAS')
```

### Multiple Regression

```
seq.multiplereg.results <- mlreg(full.model, seq.geno, trait="gaussian")
qqunif(seq.multiplereg.results[,"P1df"])
title('PGRNseq Multiple Linear Regression GWAS')
```

### All Samples with PCA Adjustment

```
r seq.pcareg.results <- mlreg(pca.model, seq.geno, trait="gaussian")
qqunif(seq.pcareg.results[,"P1df"]) title('PGRNseq All Samples
with Covariates and top 5 PCs')
```

## Exome chip GWAS

### Simple Regression

```
exome.reg.results <- mlreg(basic.model, exome.geno, trait="gaussian")
qqunif(exome.reg.results[,"P1df"])
title('Exome chip Simple Linear Regression GWAS')
```

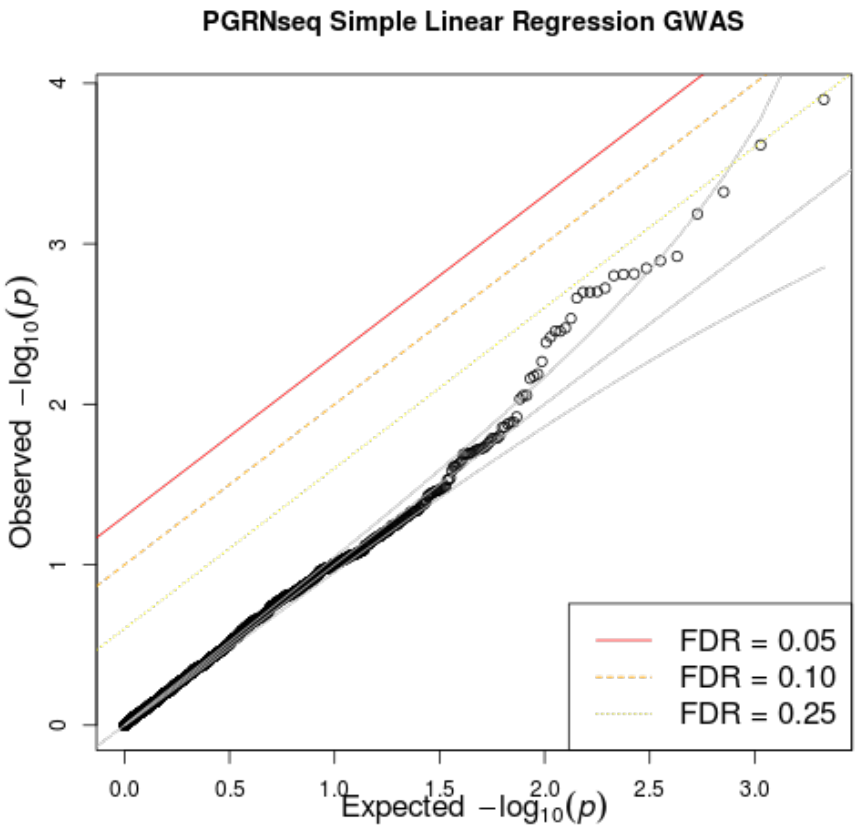


Figure 1: plot of chunk seq_simple_regression

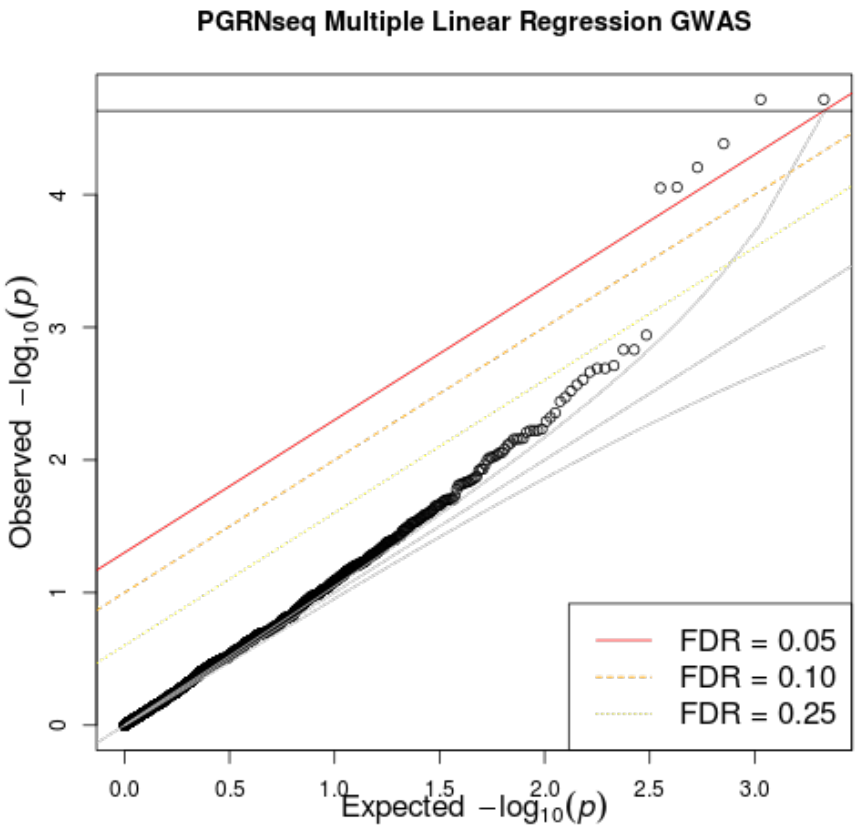


Figure 2: plot of chunk seq_multiple_regression

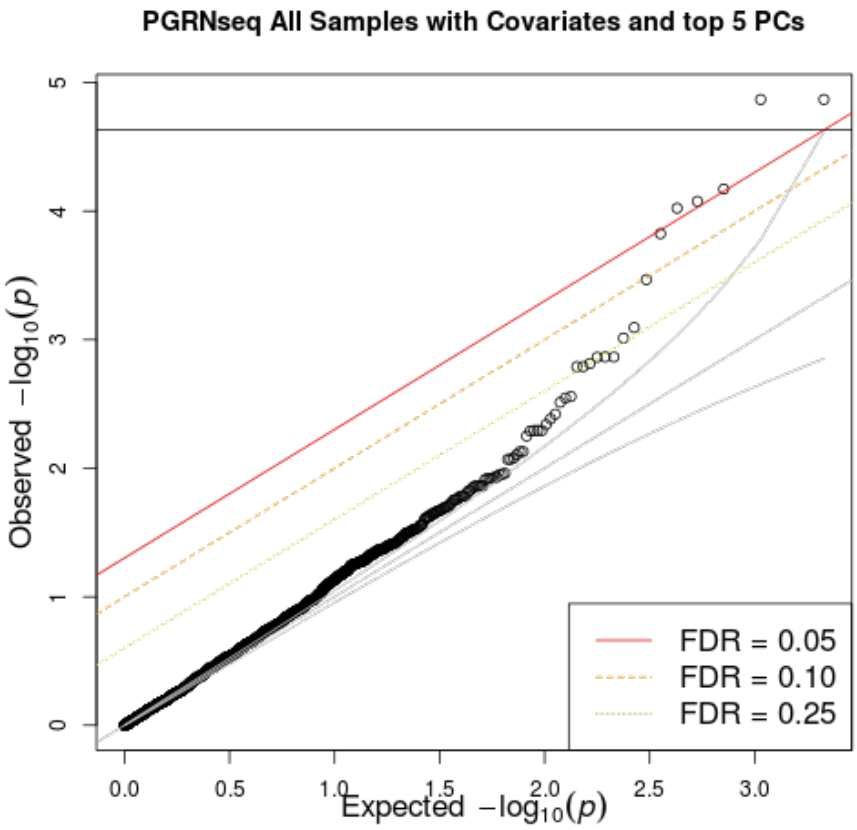


Figure 3: plot of chunk seq_allsamplepca_regression

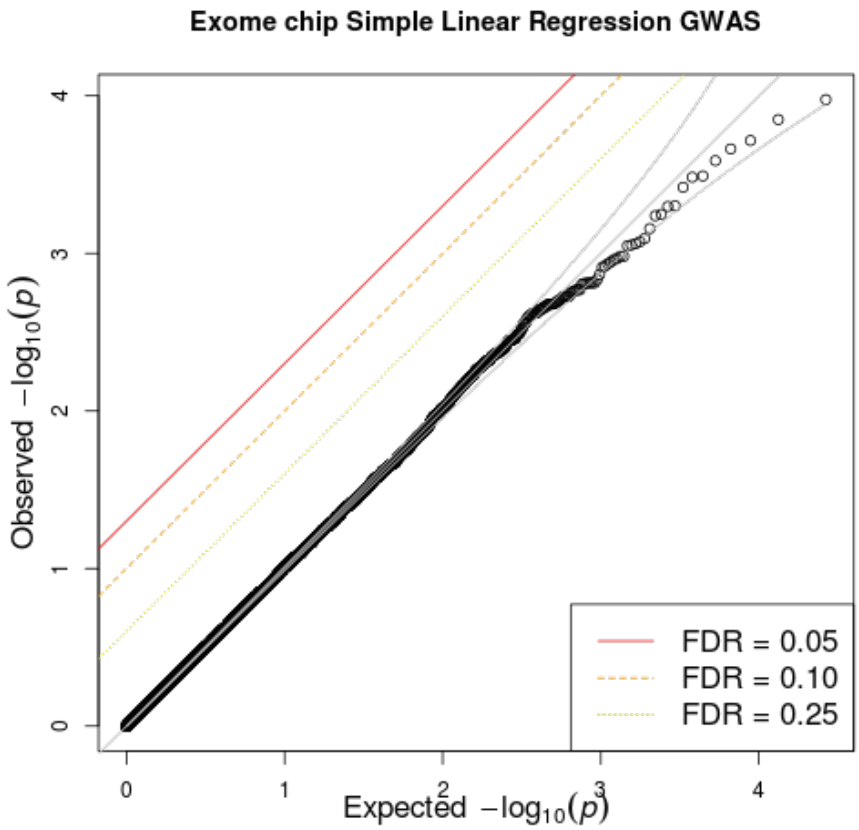


Figure 4: plot of chunk exome_simple_regression

### Multiple Regression

```
exome.reg.results <- mlreg(full.model, exome.geno, trait="gaussian")
qqunif(exome.reg.results[,"P1df"])
title('Exome Chip Multiple Linear Regression')
```

## Known signals

UGT1A1*93: rs10929302
hg19 chr2:234,665,782 G/A
1000 Genomes allele frequencies:

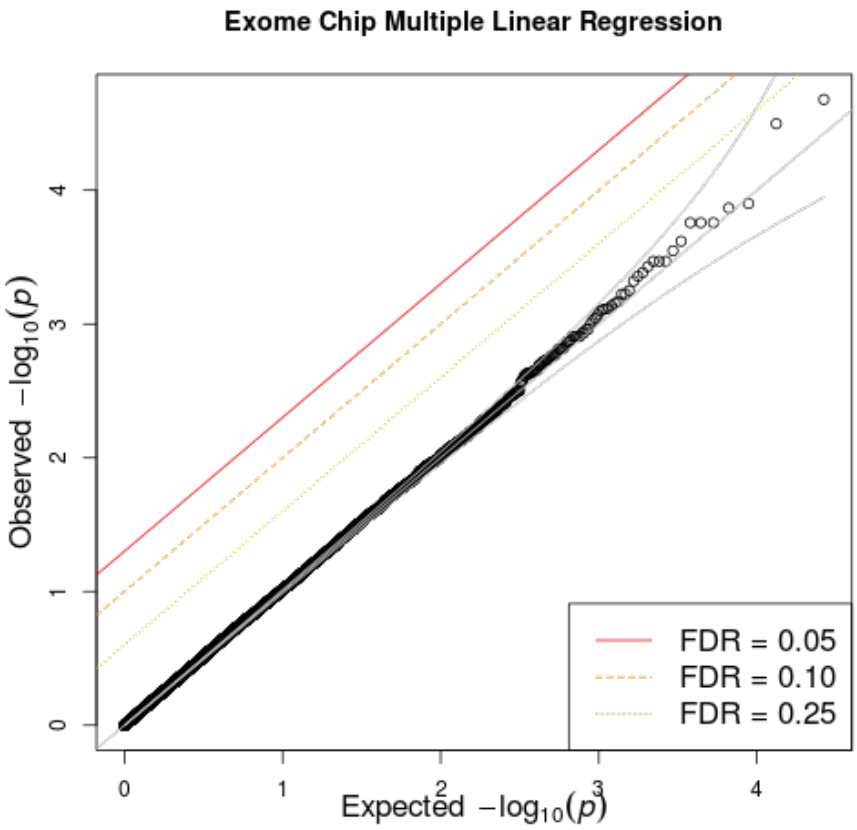


Figure 5: plot of chunk exome_multiple_regression

A: 27%
G: 73%

```
rs10929302.res <- results(seq.pcareg.results)['chr2:234665782:G:A', c('A1', 'A2','N', 'effB'
kable( rs10929302.res )
```

| |A1 |A2 | N| effB| se_effB| P1df|
|:----------------|:-|:-|---:|------:|------:|---:|
|chr2:234665782:G:A |T |G | 206| -0.2923| 0.101| 0.0038|