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 0.7362 (s.e. 0.01513), 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 0.7355 (s.e. 0.01539), 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 168 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 168 people
```

```
##
## RUN 1
## 242862 markers and 168 people in total
## 215542 (88.75%) markers excluded as having low (<5%) minor allele frequency
## 163 (0.06712%) markers excluded because of low (<95%) call rate
## 777 (0.3199%) markers excluded because they are out of HWE (FDR <0.2)
## 1 (0.5952%) people excluded because of low (<95%) call rate
## Mean autosomal HET is 0.3535 (s.e. 0.007208)
## 1 (0.5952%) people excluded because too high autosomal heterozygosity (FDR <1%)
## Excluded people had HET >= 0.3939
## Mean IBS is 0.7184 (s.e. 0.007884), as based on 2000 autosomal markers
## 0 (0%) people excluded because of too high IBS (>=0.95)
## In total, 26681 (10.99%) markers passed all criteria
## In total, 167 (99.4%) people passed all criteria
##
## RUN 2
## 26681 markers and 167 people in total
## 30 (0.1124%) 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 0.3535 (s.e. 0.006513)
## 0 people excluded because too high autosomal heterozygosity (FDR <1%)
## Mean IBS is 0.7179 (s.e. 0.007716), as based on 2000 autosomal markers
## 0 (0%) people excluded because of too high IBS (>=0.95)
## In total, 26651 (99.89%) markers passed all criteria
## In total, 167 (100%) people passed all criteria
##
## RUN 3
## 26651 markers and 167 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 0.3535 (s.e. 0.006513)
## 0 people excluded because too high autosomal heterozygosity (FDR <1%)
## Mean IBS is 0.7163 (s.e. 0.007353), as based on 2000 autosomal markers
## 0 (0%) people excluded because of too high IBS (>=0.95)
## In total, 26651 (100%) markers passed all criteria
## In total, 167 (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 can be treated as a continuous covariate, a categorical covariate, or perhaps some different altogether. Currently, I am treating it as a continuous variable in order for this to be considered

```
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')
```

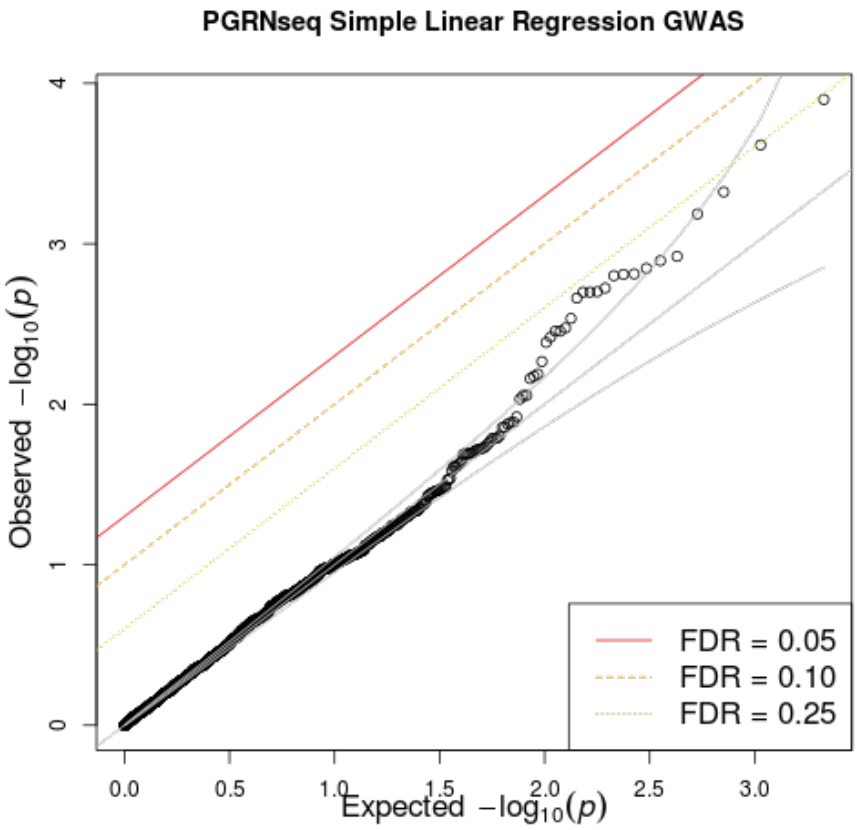

Figure 1: plot of chunk seq_simple_regression

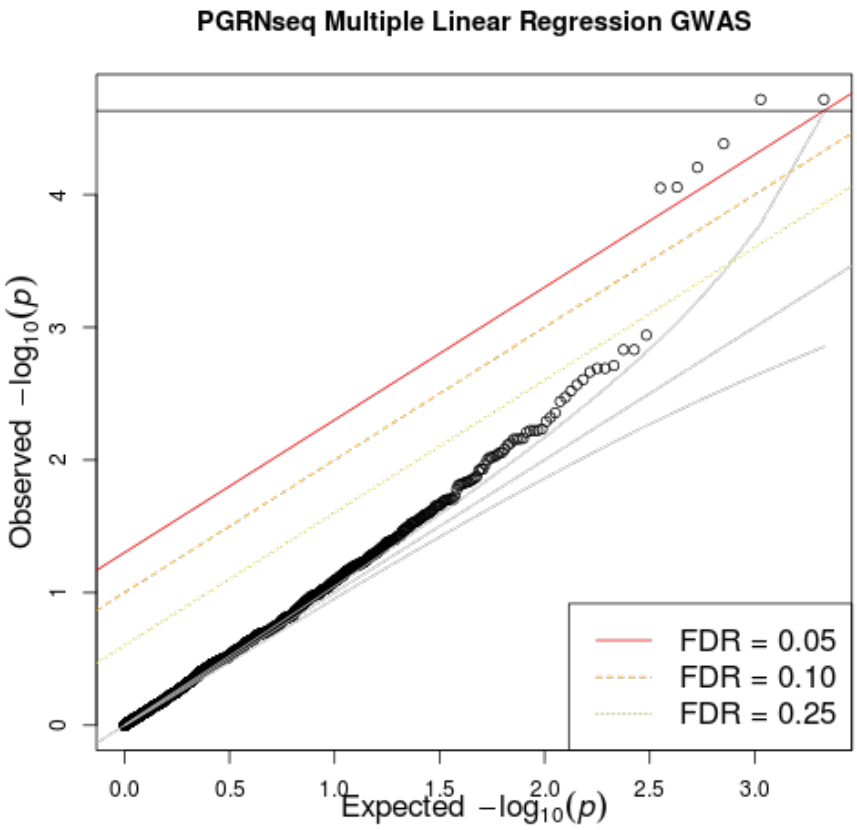

Figure 2: plot of chunk seq_multiple_regression

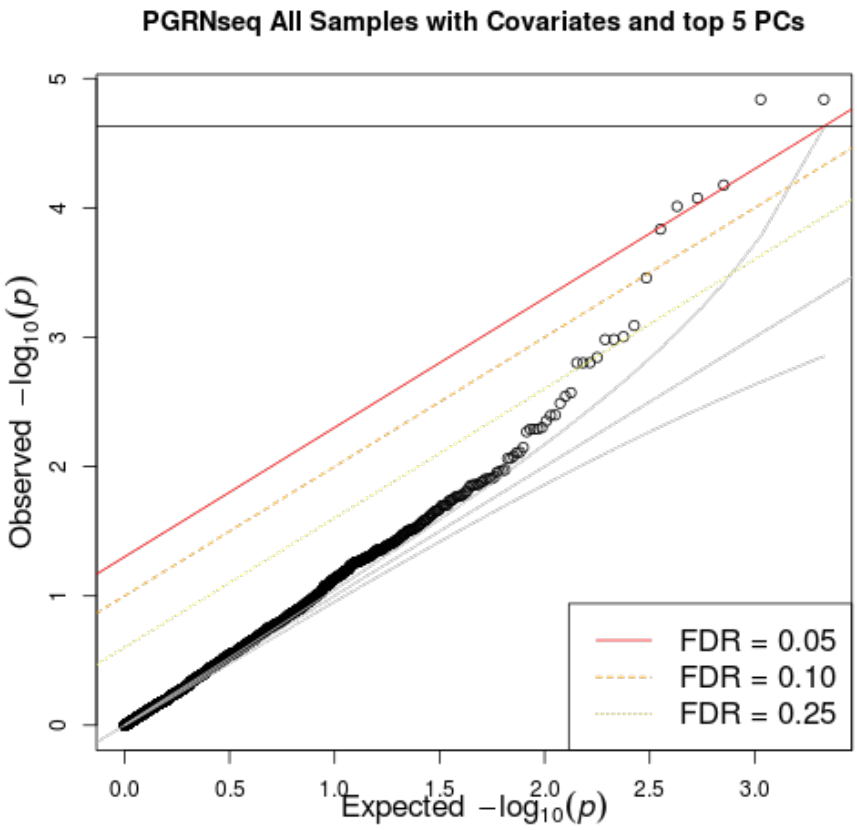

Figure 3: plot of chunk seq_allsamplepca_regression

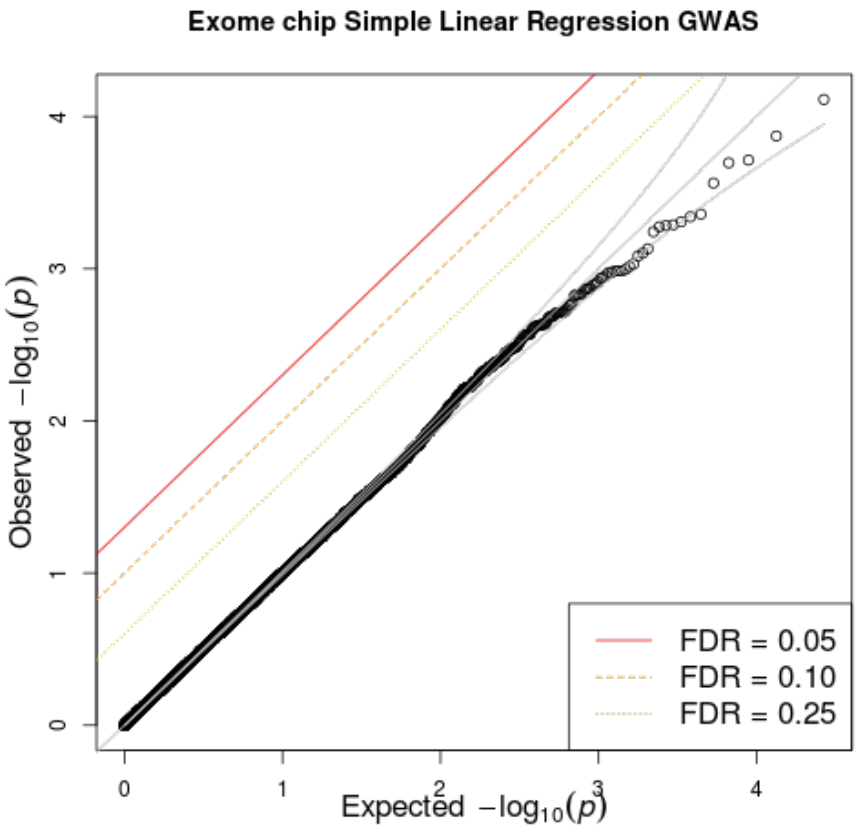

Figure 4: plot of chunk exome_simple_regression

Multiple Regression

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

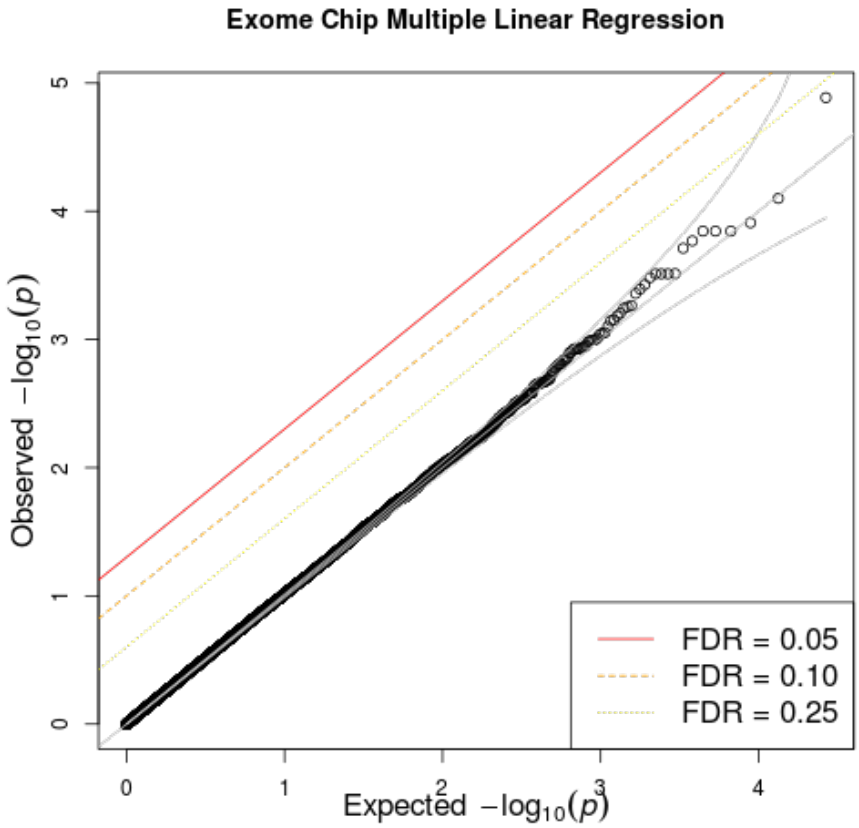

Figure 5: plot of chunk exome_multiple_regression

Known signals

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

A: 27%
G: 73%

```
rs10929302.res <- results(seq.pcareg.results)['chr2:234665782:G:A', c('A1', 'A2','N', 'effB'
print(xtable(rs10929302.res, digits=6), include.rownames=FALSE)
```

% latex table generated in R 3.1.1 by xtable 1.7-3 package % Tue Aug 26 17:08:15 2014

A1	A2	N	effB	se_effB	P1df
T	G	206.000000	-0.291090	0.101094	0.003984