SUGEN 8.6 Tutorial

Misa Graff, July 2017

Running SUGEN on the server

- Our phenotype data and genetic data (in VCF format) is in the ‘data’ folder. Can go and view it in this folder.

```
$ cd data

$ head phenotypes_sugen.txt
```

```
user1@ubuntu-sugen:~/data$ head phenotypes_sugen.txt
fid     scanID  EV1     EV2     sex     trait   new.trait       disease group   age     sex_f   group_uw
1       p207    0.954701296     -0.085427088    F       273.0572964     282.1944582     0       unc     69      1       0
1       p36     2.074001969     0.481601357     M       89.07522716     97.02145629     0       uw      23      0       1
2       p202    0.929234273     -0.84180236     F       87.78185268     60.7489117      0       unc     19      1       0
2       p370    0.755457507     -0.596455331    F       234.7431294     194.4785816     1       unc     59      1       0
3       p290    -0.709954963    1.270743336     M       140.0262184     146.1398294     0       unc     39      0       0
3       p402    -1.620465422    0.76825257      F       127.2851747     97.88796429     0       unc     34      1       0
3       p421    -0.232063527    -0.961568849    M       239.4223358     214.7733249     0       unc     56      0       0
4       p321    -0.926975536    0.139358715     F       122.2891204     120.4120269     0       unc     29      1       0
4       p379    1.660030242     -0.52371164     M       169.4409833     160.3199493     0       unc     39      0       0
```

Running SUGEN on the server

- VCF file has genetic data in genotype format: GT. (So we won't specify --dosage in SUGEN script)

```
$ zcat SISG_genotype.vcf.gz | head -n 15 | cut -d"     " -f1-11
```

```
user1@ubuntu-sugen:~/data$ zcat SISG_genotype.vcf.gz | head -n 15 | cut -d"     " -f1-11
##fileformat=VCFv4.1
##fileDate=2017-06-22
##source=GWASTools::vcfWrite()
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  p1      p2
1       558390  rs11497407      A       C       .       .       .       GT      1/1     1/1
1       711153  rs12565286      G       C       .       .       .       GT      1/1     1/1
1       713682  rs11804171      C       T       .       .       .       GT      1/1     1/1
1       713754  rs2977670       A       C       .       .       .       GT      1/1     1/1
1       719811  rs2977656       A       T       .       .       .       GT      0/0     0/0
1       740098  rs12138618      G       A       .       .       .       GT      0/0     0/0
1       742429  rs3094315       C       A       .       .       .       GT      0/0     0/0
1       744055  rs3131968       C       G       .       .       .       GT      0/0     0/0
1       755811  rs2519016       G       C       .       .       .       GT      1/1     1/1
1       758311  rs12562034      C       A       .       .       .       GT      1/1     0/1
```

Running SUGEN on the server

- We will makes a directory called ‘results’ to run our analyses from

```
$ cd ../  #go up a directory
$ mkdir results #make a results directory
$ cd results #go in to the results directory
```

Script for a quantitative trait

```
$ ~/SUGEN-master/SUGEN --pheno
../data/phenotypes_sugen.txt --id-col scanID --family-
col fid --vcf ../data/SISG_genotype.vcf.gz --formula
"new.trait=sex+age+EV1+EV2+group" --unweighted --model
linear --out-prefix quan_newtrait
```

- Output:
  - *quan_newtrait.wald.out*
    - Association results
  - *quan_newtrait.log*

**Copy quan_newtrait.wald.out to "C:\easystrata\"

*wald.out description (standard association analyses)

```
$ head quan_newtrait.wald.out
```

- CHROM Chromosome.
- POS Position.
- VCF_ID Variant ID in the VCF file.
- REF Reference allele.
- ALT Alternative allele.
- ALT_AF Alternative allele frequency.
- ALT_AC Alternative allele count.
- N_INFORMATIVE Number of subjects included in the analysis.
- N_REF Number of subjects with two reference alleles.
- N_HET Number of subjects with one reference and one alternative alleles.
- N_ALT Number of subjects with two alternative alleles.
- N_DOSE Number of subjects with genotype dosages.
- BETA Effect estimate.
- SE Standard error estimate of BETA.
- PVALUE p-value.

Clean and plot results in EasyStrata

- Open EasyStrata script (we copied this to you laptop C drive yesterday): "C:\easystrata\Easystrata_quan_newtrait.ecf"
  - *Make any necessary changes to paths or file name*
  - *We will go through the script to QC and plot results for the quantitative trait (other scripts are similar)*

```
##PLEASE add path to where results, plots, etc will be output
DEFINE      --pathOut C:\easystrata

            ##define column names and column classes; don't need to change these
            --acolIn CHROM;POS;VCF_ID;REF;ALT;ALT_AF;N_INFORMATIVE; BETA;SE;PVALUE
            --acolInClasses character;numeric;character;character;character;numeric;numeric;numeric;numeric;numeric
            --strMissing NA
            --strSeparator TAB

##PLEASE add Define path to the results:
#optional: tag name to be used in any merging where 2 columns have the same name; ShortName will be used in plots, reports, etc
EASYIN      --fileIn C:\easystrata\quan_newtrait.wald.out
                        --fileInTag quan_newtrait
                        --fileInShortName quan_newtrait
```

Clean and plot results in EasyStrata

```
###################################
## EASYSTRATA Scripting interface:
START EASYSTRATA

#########################################################################
## Merge in info, type information --colRefMarker is the ID of the variant in the list to extract --colInMarker is the the ID of the variant in your input file
#########################################################################
MERGE --colInMarker VCF_ID
                        --fileRef C:\easystrata\snpAnnot.txt
                        --acolIn rsID;info;type;oevar
                        --acolInClasses character;numeric;numeric;numeric
                        --colRefMarker rsID
```

Clean and plot results in EasyStrata

```
####################################
## EASYSTRATA Scripting interface:
START EASYSTRATA

####################################
## Filter, cleaning SNPs
####################################
## Remove monomorphic SNPs:
CLEAN --rcdClean (ALT_AF==0)|(ALT_AF==1) --strCleanName numDrop_Monomorph

## Missings:
CLEAN --rcdClean is.na(PVALUE) --strCleanName numDrop_Missing_P
CLEAN --rcdClean is.na(BETA) --strCleanName numDrop_Missing_BETA
CLEAN --rcdClean is.na(SE) --strCleanName numDrop_Missing_SE
CLEAN --rcdClean is.na(ALT_AF) --strCleanName numDrop_Missing_EAF
CLEAN --rcdClean is.na(N_INFORMATIVE) --strCleanName numDrop_Missing_N
CLEAN --rcdClean info<0.8 --strCleanName numDrop_lowinfo

## Sanity checks:
CLEAN --rcdClean PVALUE<0|PVALUE>1 --strCleanName numDrop_invalid_P
CLEAN --rcdClean ALT_AF<0|ALT_AF>1 --strCleanName numDrop_invalid_ALT_AF
```

Clean and plot results in EasyStrata

```
##get number of rare SNPs with <0.1% ALT_AF
GETNUM     --rcdGetNum (ALT_AF<0.01 | ALT_AF>0.99) --strGetNumName numSNPs_lowmaf

##create minor allele count and count SNPs with MAF<5; can filter on this if desired
ADDCOL --rcdAddCol 2*pmin(ALT_AF*N_INFORMATIVE,(1-ALT_AF)*N_INFORMATIVE,na.rm=TRUE) --colOut MAC
GETNUM     --rcdGetNum MAC<5 --strGetNumName numSNPs_MACle5

##get the number of SNPs with PVALUE<min (7e-04 suggestive; 7e-06 bonf. significant)
GETNUM     --rcdGetNum PVALUE<7e-04 --strGetNumName numSNPs_7e04
GETNUM     --rcdGetNum PVALUE<7-06 --strGetNumName numSNPs_7e06

##get the number of SNPs that are genotyped ‘type=2’
GETNUM     --rcdGetNum type==2 --strGetNumName numSNPs_genotyped

#####################################
##annotate known loci; list of known loci for your trait if available; includes 3 columns "Chr" "Pos" "Colour",
#ANNOTATE --fileAnnot /path2/known_loci.txt
#    --numPosLim 500000
#    --colInChr CHROM
#    --colInPos POS
#    --colOutAnnot Annot_knownloci
#    --strAnnotTag KNOWN
```

Clean and plot results in EasyStrata

```
#####################################
## Qqplot: pvalue ## with and without known loci
#####################################

QQPLOT
        --acolQQPlot PVALUE
        --astrColour black
        ##path to list of known loci, with at least 2 columns "Chr" "Pos"
        #--fileRemove /path2/known_loci.txt
        #--numRemovePosLim 500000
        #--strRemovedColour orangered
        #--colInChr CHROM
        #--colInPos POS


#calculate lambda; suppress GC correction for now
GC      --colPval PVALUE
        --blnSuppressCorrection 1
```

Clean and plot results in EasyStrata

```
#####################################
## Manhattan plot
## - can display known loci using physical distance criterion d<500kB to either side, if we use --fileAnnot
#####################################
MHPLOT --colMHPlot PVALUE
            --colInChr CHROM
            --colInPos POS
            --numWidth 1200
            --numHeight 800
            --anumAddPvalLine 7e-4;7e-6
            --anumAddPvalLineLty 6;6
            --astrAddPvalLineCol blue;orangered
            --numDefaultSymbol 20
            --arcdSymbolCrit PVALUE<7e-4;PVALUE<7e-6
            --anumSymbol 8;8
            --arcdColourCrit PVALUE<7e-4;PVALUE<7e-6
            --astrColour blue;orangered
            --arcdCexCrit PVALUE<7e-4;PVALUE<7e-6
            --anumCex 1.2;1.3
            --astrDefaultColourChr gray51;gray66
            --blnYAxisBreak 1
            ##Path to list of known loci for your trait if available; includes 3 columns "Chr" "Pos" "Colour",
            #--fileAnnot /path2/known_loci.txt
            #--numAnnotPosLim 500000
```

Clean and plot results in EasyStrata

```
#####################################
## Extract SNPs from list; --colRefMarker is snpID in the list to extract --colInMarker is the snpID in your input file
#EXTRACTSNPS --colInMarker VCF_ID
#                           --fileRef /path2/snp_list_to_extract.txt
#                           --colRefMarker rsid
#                           --strTag known

#####################################
## Clump results by 'suggestive' significance (P<7e-4) using physical distance criterion distance<500kB to either side
INDEP       --rcdCriterion PVALUE<7e-4
                           --colIndep PVALUE
                           --colInChr CHROM
                           --colInPos POS
                           --numPosLim 500000
## Function INDEP creates two files labeled _indep.txt and _indepX.txt
## _indep.txt contains ALL SNPs that passed P<7e-4
##       Column aLociTag can be used to distinguish between independent loci (each number represents an indep locus)
##             Column aTopHit can be used to identify the TopHit for the respective locus (all other SNPs are set to NA)
## _indepX.txt only contains the Top Hits
```

Clean and plot results in EasyStrata

```
####################################
## this will output cleaned results after dropping based on QC
WRITE --strMode txt
        --strPrefix CLEANED.
        --strSep TAB
        --strMissing .



STOP EASYSTRATA
#########################################################################################################
```

Run EasyStrata script for quantitative trait

- Run in R:

```
> library(EasyStrata)
> EasyStrata("C:/easystrata/Easystrata_quan_newtrait.ecf")
```

- Output plots and files:

Easystrata_quan_newtrait.ecf.out	Easystrata log file
Easystrata_quan_newtrait.rep	Report file with file names, snps in and out, cleaning steps, lambda, requested counts, etc
CLEANED.quan_newtrait.txt	Cleaned output after removing SNPs based on QC protocol
quan_newtrait.mh.png	Manhattan plot
quan_newtrait.qq.png	QQ plot
quan_newtrait.indep.txt	All SNPs with PVALUE<7e-04
quan_newtrait.indepX.txt	Most significant SNPs within +/-500Kb with 7e-04

Questions:

Looks in the **Easystrata_quan_newtrait.rep** file and at the plots **quan_newtrait.mh.png**, **quan_newtrait.qq.png**.

1. What is the lambda? Is there inflation?
2. How many SNPs were dropped for low info?
3. How many monomorphic SNPs were dropped?
4. How many reach the different PVALUE thresholds (i.e. 7e-04 and 7e-06)?

Script for a binary trait

```
~/SUGEN-master/SUGEN --pheno
../data/phenotypes_sugen.txt --id-col scanID --family-
col fid --vcf ../data/SISG_genotype.vcf.gz --formula
"disease=sex+age+EV1+EV2+group" --score --unweighted --
model logistic --out-prefix bin_trait
```

- Output:
  - *bin_trait.score.snp.out*
    - Single variant association results
  - *bin_trait.mass.out*
    - gene-based summary statistics are stored in MASS 7.0 format; or converted for use in RAREMETAL, seqMETA.
  - *bin_trait.log*

**Copy bin_trait.score.snp.out to "C:\easystrata\"

*score.snp.out description

```
$ head bin_trait.score.snp.out
```

- GENE_ID In single-variant analysis (i.e., [--group group_file] is not specified), GENE_ID equals CHROM:POS.
- CHROM Chromosome.
- POS Position.
- VCF_ID Variant ID in the VCF file.
- REF Reference allele.
- ALT Alternative allele.
- ALT_AF Alternative allele frequency.
- ALT_AC Alternative allele count.
- N_INFORMATIVE Number of subjects included in the analysis.
- N_REF Number of subjects with two reference alleles.
- N_HET Number of subjects with one reference and one alternative alleles.
- N_ALT Number of subjects with two alternative alleles.
- N_DOSE Number of subjects with genotype dosages.
- U Score statistic.
- V Variance estimate of U.
- BETA Effect estimate.
- SE Standard error estimate of BETA.
- PVALUE p-value.

Clean and plot results in EasyStrata

- Open EasyStrata script (we copied this to you laptop C drive yesterday): “C:\easystrata\Easystrata_bin_trait.ecf”
  - *Make any necessary changes to paths or file name*

```
##PLEASE add path to where results, plots, etc will be output
DEFINE      --pathOut C:\easystrata

            ##define column names and column classes; don't need to change these
            --acolIn CHROM;POS;VCF_ID;REF;ALT;ALT_AF;N_INFORMATIVE; ALT_AF_CASE;N_CASE;BETA;SE;PVALUE
            --acolInClasses character;numeric;character;character;character;numeric;numeric; numeric;numeric;numeric;numeric;numeric
            --strMissing NA
            --strSeparator TAB

##PLEASE add Define path to the results:
#optional: tag name to be used in any merging where 2 columns have the same name; ShortName will be used in plots, reports, etc
EASYIN      --fileIn C:\easystrata\bin_trait.score.snp.out
                        --fileInTag bin_trait
                        --fileInShortName bin_trait
```

Run EasyStrata script for binary trait

- Run in R:

```
> library(EasyStrata)
> EasyStrata("C:/easystrata/Easystrata_bin_trait.ecf")
```

- Output plots and files:

Easystrata_bin_trait.ecf.out	Easystrata log file
Easystrata_bin_trait.rep	Report file with file names, snps in and out, cleaning steps, lambda, requested counts, etc
CLEANED.bin_trait.txt	Cleaned output after removing SNPs based on QC protocol
bin_trait.mh.png	Manhattan plot
bin_trait.qq.png	QQ plot
bin_trait.indep.txt	All SNPs with PVALUE<5e-05
bin_trait.indepX.txt	Most significant SNPs within +/-500Kb with PVALUE<5e-05

- Questions:

Looks in the **Easystrata_bin_trait.rep** file and at the plots **bin_trait.mh.png, bin_trait.qq.png**.

1. What is the lambda? Is there inflation?
2. How many reach the different PVALUE thresholds (i.e. 5e-05 and 6.7e-06)?

Looks in the **bin_trait.indep.txt and bin_trait.indepX.txt** files

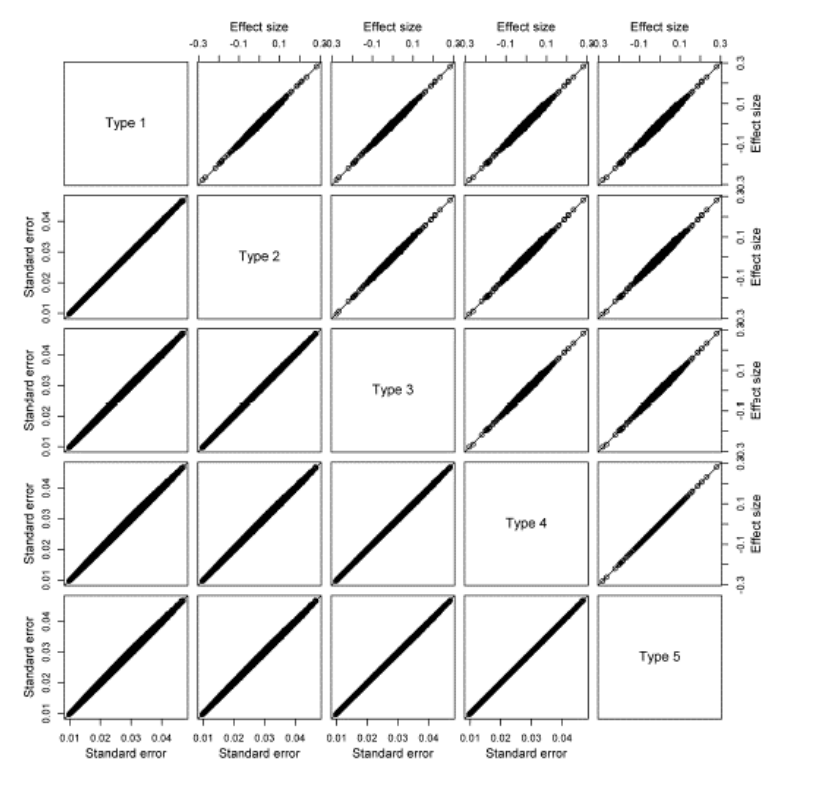
3. How many SNPs met the PVALUE thresholds?
4. What are their allele frequencies and MAC?

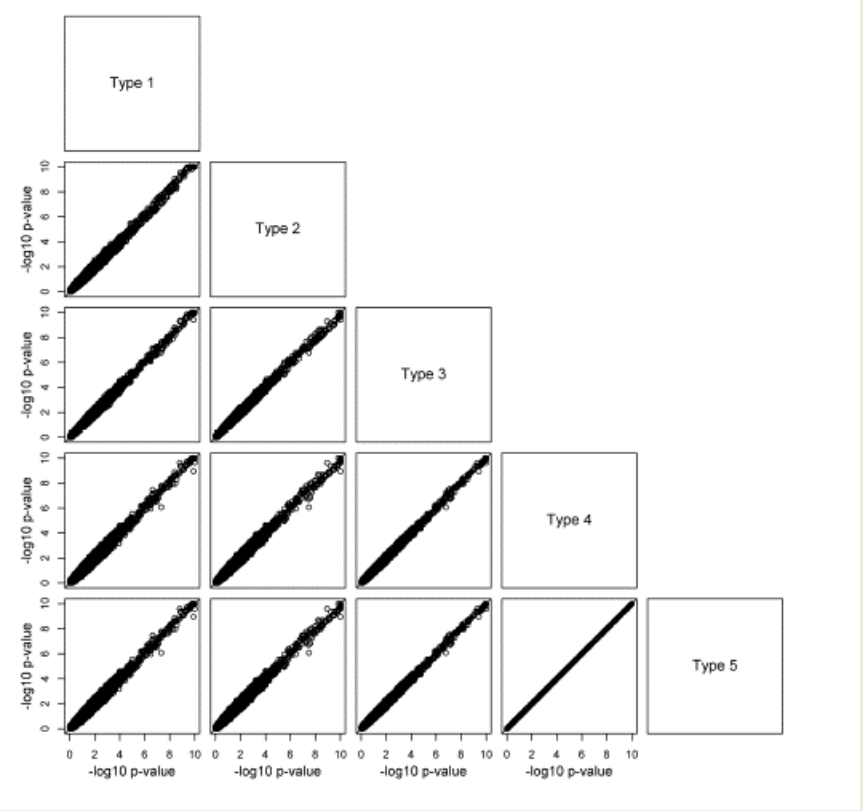
Heterogeneous variances

- Models that combine heterogeneous race/ethnic groups from the PAGE consortium in different ways.
  1. *Use GEE in pooled analysis of all populations*
  2. *Use GEE only for SOL, pooled analysis of all the other data using OLS, and then meta-analyze the two sets of results*
  3. *Use GEE for all Hispanics, pooled analysis of all the other ethnicities using OLS, and then meta-analyze the two sets of results*
  4. *Use GEE for all self-reported racial/ethnic groups separately, then meta-analyze results from all groups*
  5. *Use GEE for all Hispanics, OLS for other racial/ethnic groups separately, and then meta-analyze (METAL) results for separate analyses*

Heterogeneous variances

Comparing results for the 5 models in a *trait that has similar variances between race ethnic groups*.
Not much difference between models results.

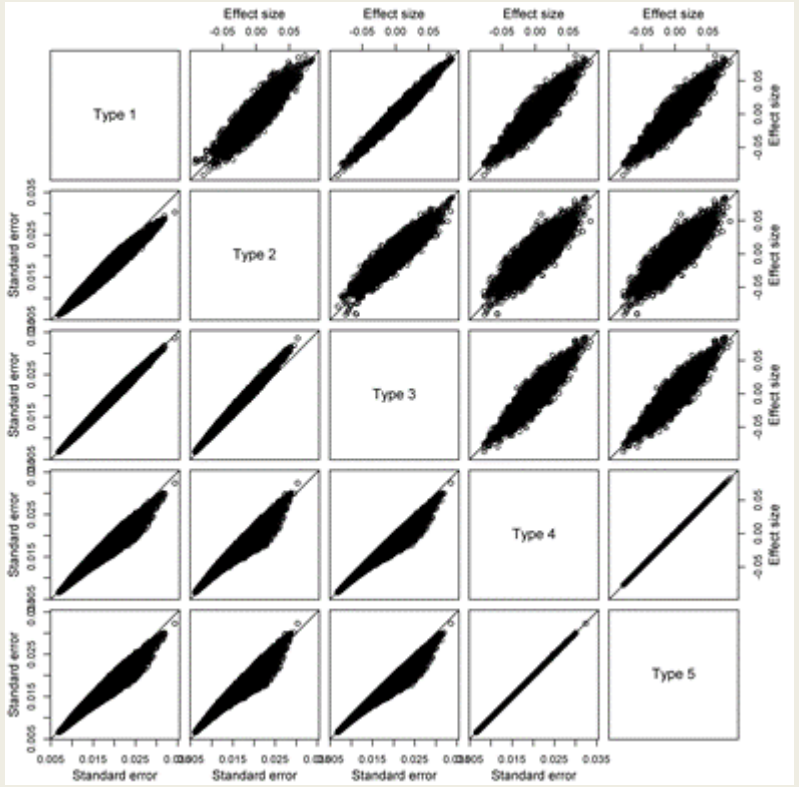


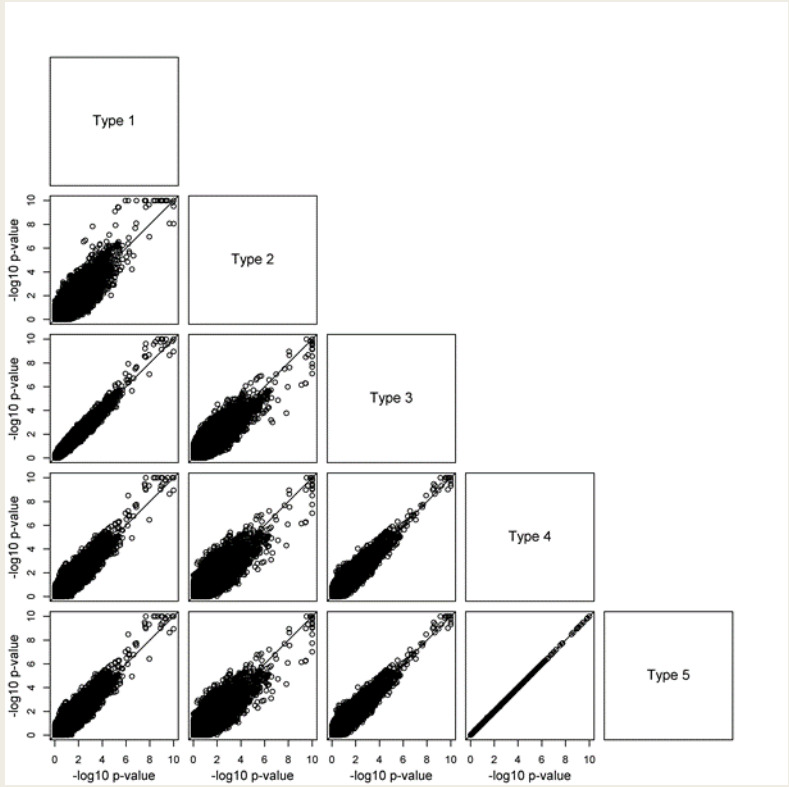


C-Reactive Protein

Heterogeneous variances

Comparing results for the 5 models in a *trait that has different variances between race ethnic groups*.
Large differences in results between models 1-3 with each other and Models 4 and 5.
Models 4 and 5 (meta-analyses that combined race/ethnic groups) are similar.





Fasting Glucose

Heterogeneous variances

Residual Standard Error $\hat{\sigma}^2$

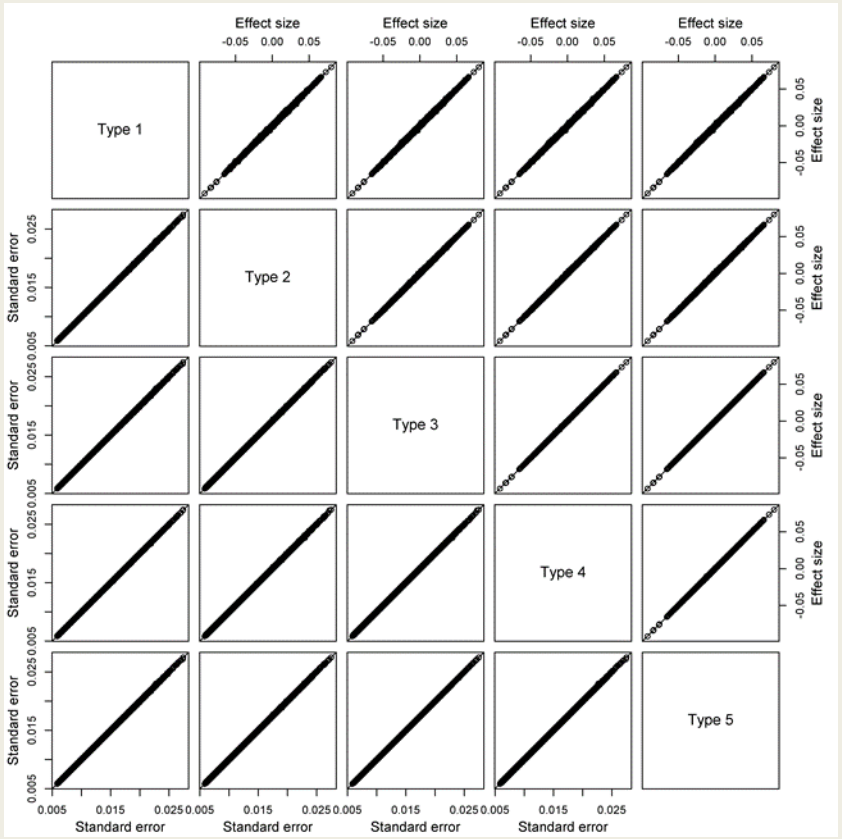
Trait	Pooled	SOL only	HA	AA	AS	HI	NA	All races, no HA	Other
C-reactive protein	1.1	1.07	1.07	1.17	1.1	1.16	1.02	1.15	1.13
Fasting glucose	0.85	0.56	0.78	0.97	0.74	0.35	0.8	0.94	1.05

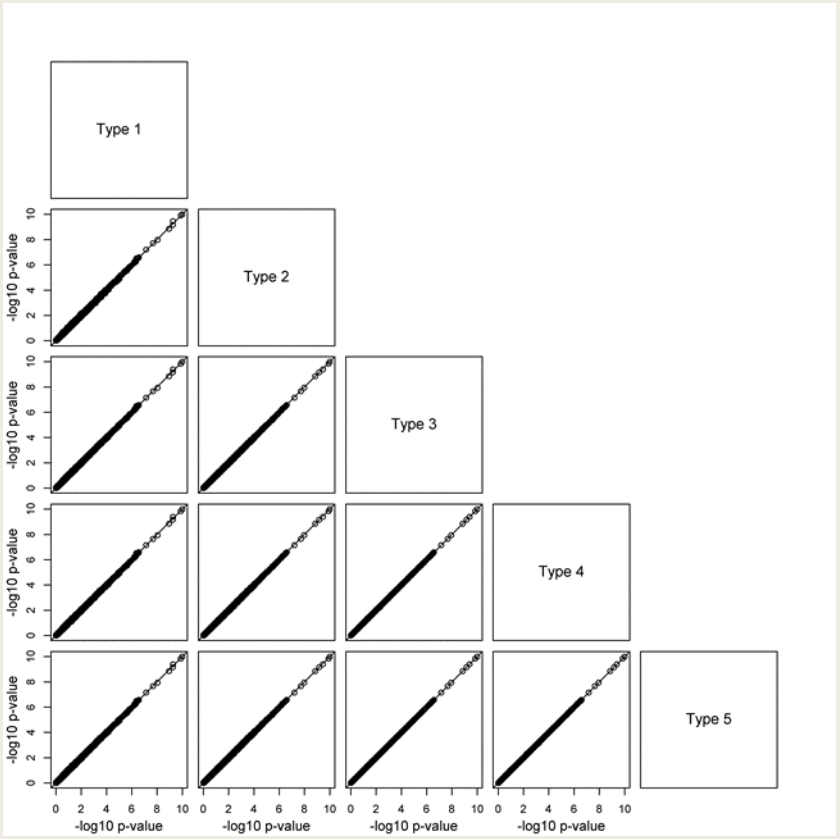
NOTE: HS = Hispanic, AA = African, AS = Asian, HI = Hawaiian, NA = Native American.

- **Solution:** pooled analysis including race x covariate interactions and allowing for heterogeneous trait variances

Heterogeneous variances

Comparing results for the 5 models in a *trait that has different variances between race ethnic groups* **after accounting** for differences in pooled analyses. Little difference in results between *ALL* models.





Fasting Glucose

Accounting for Heterogeneous variances

```
$ ~/SUGEN-master/SUGEN --pheno
../data/phenotypes_sugen.txt --id-col scanID --family-
col fid --vcf ../data/SISG_genotype.vcf.gz --formula
"new.trait=sex+age+EV1+EV2+group" --unweighted --model
linear --hetero-variance group --out-prefix
quan_newtrait_het
```

- Note flag for `--hetero-variance group`
- Output:
  - *quan_newtrait_het.wald.out*
    - Association results
  - *quan_newtrait_het.log*

**Copy quan_newtrait_het.wald.out to "C:\easystrata\"

Clean and plot results in EasyStrata

- Open EasyStrata script (we copied this to your laptop C drive yesterday):

"C:\easystrata\Easystrata_quan_newtrait_het.ecf"

  - *Make any necessary changes to paths or file name*

```
##PLEASE add path to where results, plots, etc will be output
DEFINE      --pathOut C:\easystrata

            ##define column names and column classes; don't need to change these
            --acolIn CHROM;POS;VCF_ID;REF;ALT;ALT_AF;N_INFORMATIVE; BETA;SE;PVALUE
            --acolInClasses character;numeric;character;character;character;numeric;numeric;numeric;numeric;numeric
            --strMissing NA
            --strSeparator TAB

##PLEASE add Define path to the results ;we will merge the results before and after accounting for heterogeneous variances
#optional: tag name to be used in any merging where 2 columns have the same name; ShortName will be used in plots, reports, etc
EASYIN      --fileIn C:\easystrata\quan_newtrait_het.wald.out
                        --fileInTag quan_newtrait_het
                        --fileInShortName quan_newtrait_het
……….
##towards the end of the file
MERGE --colInMarker VCF_ID
                        --fileRef C:\easystrata\CLEANED.quan_newtrait.txt
```

Run EasyStrata script for quantitative trait results (which accounted for heterogeneous variances)

- Run in R:

```
> library(EasyStrata)
> EasyStrata("C:/easystrata/Easystrata_quan_newtrait_het.ecf")
```

- Output plots and files:

Easystrata_quan_newtrait_het.ecf.out Easystrata log file
Easystrata_quan_newtrait_het.rep Report file with file names, snps in and out, cleaning steps, lambda, requested counts, etc
CLEANED.quan_newtrait_het.txt Cleaned output after removing SNPs based on QC protocol
quan_newtrait_het.mh.png Manhattan plot
quan_newtrait_het.qq.png QQ plot
quan_newtrait_het.png Scatter plot comparing p-values of results before and after accounting for heterogeneous variances
quan_newtrait_het.indep.txt All SNPs with PVALUE<5e-05
quan_newtrait_het.indepX.txt Most significant SNPs within +/-500Kb with PVALUE<5e-05
quan_newtrait_het.BETA.png Compare BETA between results with and without using the hetero-variance flag
quan_newtrait_het.SE.png Compare SE between results with and without using the hetero-variance flag
quan_newtrait_het.PVALUE.png Compare PVALUE between results with and without using the hetero-variance flag

Question: How do the results compare between the analyses run with and without the hetero-variance option?

Running just a subset of the data

```
$ ~/SUGEN-master/SUGEN --pheno
../data/phenotypes_sugen.txt --id-col scanID --family-
col fid --vcf ../data/SISG_genotype.vcf.gz --formula
"new.trait=sex+age+EV1+EV2" --unweighted --model linear
--subset group="uw" --out-prefix quan_newtrait_group_uw
```

- Note flag for `--subset group="uw"`

Conditional analyses

```
$ ~/SUGEN-master/SUGEN --pheno
../data/phenotypes_sugen.txt --id-col scanID --family-
col fid --vcf ../data/SISG_genotype.vcf.gz --formula
"new.trait=sex+age+EV1+EV2+group" --unweighted --model
linear --cond ../data/cond.list --out-prefix
quan_newtrait_cond
```

- Note flag for `--cond ../data/cond.list`
- Output:
  - *quan_newtrait_cond.wald.out*
    - Association results
  - *quan_newtrait_cond.log*

Question: what is the P-value for rs6656541 before and after conditioning?

Exact only a subset of SNPs

```
$ ~/SUGEN-master/SUGEN --pheno
../data/phenotypes_sugen.txt --id-col scanID --family-
col fid --vcf ../data/SISG_genotype.vcf.gz --formula
"new.trait=sex+age+EV1+EV2+group" --unweighted --model
linear --extract-file ../data/extract.list --cond
../data/cond.list --out-prefix
quan_newtrait_extract_cond
```

- Note flag for `–extract-file ../data/extract.list`
- Output:
  - *quan_newtrait_extract_cond.wald.out*
    - Association results
  - *quan_newtrait_extract_cond.log*