Association Analysis of Sequence Data using PLINK/SEQ (PSEQ)

Purpose

PLINK/SEQ (PSEQ) is an open-source C/C++ library for working with human genetic variation data. The specific focus is to provide a platform for analytic tool development for variation data from large-scale resequencing and genotyping projects, particularly whole-exome and whole-genome studies. PSEQ is independent of, but designed to be complementary to, the existing PLINK (Purcell *et al.,* 2007) package. Here we give an overview of analysis of exome sequence data using PSEQ.

Software Resource

This tutorial was completed with PSEQ 0.10, (released on 14-Jul-2014) available from https://atgu.mgh.harvard.edu/plinkseq/download.shtml. Links to PSEQ documentation can also be found on the webpage. Below is an outline of what PSEQ documentation offers:

- Basic Syntax and Conventions
- Project Management
- Data Input
- Attaching Auxiliary Data
- Viewing Data
- Data Output
- Summary Statistics
- Association Analysis
- Locus Database Operations
- Reference Database Operations
- Miscellaneous commands

Exercise Genotype Data

Autosomal exome genotype data was downloaded from the 1000 Genomes pilot data July 2010 release for both the CEU (Utah residents with Northern and Western European ancestry) and YRI (Yoruba in Ibadan, Nigeria) populations. The data sets (CEU.exon.201003.genotypes.vcf.gz and YRI.exon.201003.genotypes.vcf.gz) are available from:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_07/exon/snps
```

The genomic co-ordinate for this data set is hg18 based. To use the PSEQ annotation data source which is hg19 based, you will lift over this data set to use hg19 co-ordinate. Since PSEQ does not provide a liftover feature therefore the data has already been lifted over for you using Variant Association Tools. The resulting data files, **CEU.exon.201003.genotypes.hg19.vcf.gz** and **YRI.exon.201003.genotypes.hg19.vcf.gz**, will be used for this exercise. One data set contains exome data for European-Americans (CEU) from 1000 Genomes while the other for Yoruba (YRI). The liftover feature may also have to be used with your data set as new hg coordinates become available. For additional information see http://varianttools.sourceforge.net/Vtools/Liftover

Phenotype Data

To demonstrate performing an association analysis, we simulated a quantitative trait phenotype (BMI). Please note that these phenotypes are **NOT** from the 1000 genomes project. The phenotype data for the exercise can be found in the text file **phenotype.phe.** This phenotype file contains data for 202 individuals from both the CEU and YRI populations.

Computation Resources

The following tutorial uses a small data set so that the association analysis can be completed in a short period-of-time. Large next-generation sequenced data sets require a reasonably powerful machine with a high-speed internet connection.

Data Cleaning and Variant/Sample Selection

Getting Started

To get a list of PSEQ subcommands use:

```
pseq help
```

Or,

```
pseq help all
```

Create a new project

```
pseq myproj new-project --resources hg19
Creating new project specification file [ myproj.pseq ]
```

The "--resources" flag tells **pseq** where your supporting databases are located. For this exercise the necessary databases have already been created and are within your exercise directory. Instructions on how to create these databases is located at:

`http://atgu.mgh.harvard.edu/plinkseq/resources.shtml`.

Load variant data

Import all vcf files under the current directory:

```
pseq myproj load-vcf --vcf CEU.exon.2010_03.genotypes.hg19.vcf.gz YRI.exon.2010_03.genotypes.hg19.vcf.gz
loading : /home/gmc01/data/pseq/CEU.exon.2010_03.genotypes.hg19.vcf.gz ( 90 individuals )
parsed 3000 rows
loading : /home/gmc01/data/pseq/YRI.exon.2010_03.genotypes.hg19.vcf.gz ( 112 individuals )
parsed 5000 rows
/home/gmc01/data/pseq/CEU.exon.2010_03.genotypes.hg19.vcf.gz : inserted 3489 variants
/home/gmc01/data/pseq/YRI.exon.2010_03.genotypes.hg19.vcf.gz : inserted 5175 variants
```

Note CEU are European-Americans and YRI are Yoruba from Nigeria.

Load phenotype data

```
pseq myproj load-pheno --file phenotype.phe
Processed 202 rows
```

The "phenotype.phe" file contains phenotypes for SEX, BMI and RACE (BMI is body mass index, males are denoted by a 1 and females by 2). Instruction on formatting .phe file can be found at https://atgu.mgh.harvard.edu/plinkseq/input.shtml#phe.

View variants and samples

To view variant sites info:

pseq myproj v-view | head

```
chr1:1115461    .               C/T     .       1       PASS
chr1:1115503    .               T/C     .       1       SBFilter
chr1:1115510    .               C/T     .       1       PASS
chr1:1115548    .               G/A     .       1       PASS
chr1:1115604    .               C/A     .       1       PASS
chr1:1118275    rs61733845      C/T     .       2       PASS
chr1:1119399    .               C/T     .       1       PASS
chr1:1119434    .               C/A     .       1       PASS
chr1:1120370    .               C/G     .       1       PASS
chr1:1120377    .               T/A     .       1       PASS
```

v-view command outputs a per-variant level view of a project, with the above fields: chromosome (base-position); variant-ID (or '.' If novel); ref/alt alleles; a sample/file identifier (or '.' If consensus variant); # of samples the variant observed in; filter values for samples (here 'PASS' means that the variant site passes all filter and 'SBFilter' means that the variant site fails to pass the strand bias (SB) filter). More details about v-view command can be found at https://atgu.mgh.harvard.edu/plinkseq/view.shtml#var

To view samples and phenotypes:

i-view command writes to standard output to view individuals' phenotype information

pseq myproj i-view | head

```
#BMI (Float) "BMI"
#RACE (String) "RACE"
#SEX (Integer) "SEX"
#PHE    .
#STRATA .
#ID     FID     IID     MISS    SEX     PAT     MAT     META
NA06984 .       .       0       0       .       .       BMI=36.353;RACE=CEU;SEX=1
NA06985 .       .       0       0       .       .       BMI=21.415;RACE=CEU;SEX=2
NA06986 .       .       0       0       .       .       BMI=26.898;RACE=CEU;SEX=1
NA06989 .       .       0       0       .       .       BMI=25.015;RACE=CEU;SEX=2
```

There are 3 fields, BMI, RACE and SEX contained in the input phenotype file, phenotype.phe. The headers are #ID – main unique individual ID; FID – optional family ID; IID: optional individual ID; MISS – a flag to indicate missing data; SEX – sex; PAT – paternal ID; MAT – maternal ID; META – meta information of fields from input phenotype file. More details about i-view command outputs can be found at https://atgu.mgh.harvard.edu/plinkseq/view.shtml#ind.

Summary

To view a summary of the complete project

pseq myproj summary

Command above will generate a long list of output. To view summaries of portions of the project, i.e., variant data, phenotype data, locus data, reference data, sequence data, input files and meta data:

pseq myproj var-summary

```
---Variant DB summary---

6987 unique variants
File tag : 1 (3489 variants, 90 individuals)
File tag : 2 (5175 variants, 112 individuals)
```

pseq myproj ind-summary

```
---Individual DB summary---

202 unique individuals
Phenotype : BMI (Float) "BMI"
Phenotype : RACE (String) "RACE"
Phenotype : SEX (Integer) "SEX"
```

pseq myproj loc-summary

pseq myproj ref-summary

pseq myproj seq-summary

pseq myproj file-summary

pseq myproj meta-summary

More details about viewing summary information for project databases can be found at https://atgu.mgh.harvard.edu/plinkseq/proj.shtml#summ

Based on the “pseq myproj var-summary” command there are 6987 unique variant sites for CEU and YRI, with the CEU sample having 3489 variant sites and the YRI sample 5175 variant sites. .

For an overview of variant summary statistics:

```
pseq myproj v-stats
NVAR    6987
RATE    0.568384
MAC     19.8557
MAF     0.0691347
SING    2064
MONO    30
TITV    3.57264
TITV_S  3.77778
DP      8426.74
QUAL    NA
PASS    0.999857
FILTER|PASS     0.999857
FILTER|SBFilter 0.000143123
PASS_S  1
```

v-stats command obtains summary statistics across variants. Output statistics are NVAR – total number of variants; RATE – average call rate; MAC – mean minor allele count; MAF – mean minor allele frequency; SING – number of singletons; MONO – number of monomorphic sites; TITV – transition/transversion (Ti/Tv) ratio; TITV_S – Ti/Tv ratio for singletons; DP – mean variant read depth; QUAL – mean QUAL score from VCF; PASS – proportion of variants that PASS all FILTERS; FILTER|PASS – proportion of variants that pass all filters; FILTER|SBFilter – proportion of variants that fail to pass SB filter. More details about v-stats command outputs can be found at https://atgu.mgh.harvard.edu/plinkseq/stats.shtml#var

For individual level summary statistics:

```
pseq myproj i-stats | head
```

ID	NALT	NMIN	NHET	NVAR	RATE	SING	TITV	PASS	PASS_S	QUAL	DP
NA06984	719	568	480	3162	0.452555	8	3.61789	568	8	NA	13489
NA06985	655	531	420	3144	0.449979	10	3.5	531	10	NA	13530.3
NA06986	773	643	503	3437	0.491914	22	3.69343	643	22	NA	12535.8
NA06989	699	532	469	3130	0.447975	8	3.22222	532	8	NA	13549.7
NA06994	591	464	377	3002	0.429655	3	3.59406	464	3	NA	13923.8
NA07000	802	613	517	3388	0.484901	10	3.67939	613	10	NA	12292.6
NA07037	800	631	512	3374	0.482897	4	3.60584	631	4	NA	12357.4
NA07048	817	675	607	3373	0.482754	15	3.29936	675	15	NA	12909.5
NA07051	825	637	507	3451	0.493917	13	3.05732	637	13	NA	11929

i-stats command obtains a matrix of summary statistics for every individual in a project. Output statistics are ID – individual ID; NALT – number of non-reference genotypes; NMIN – number of genotypes with a minor allele; NHET – number of heterozygous genotypes for individual; NVAR – total number of called variants for individual; RATE – genotyping rate for individual; SING – number of singletons individuals has; TITV – mean Ti/Tv for variants for which individual has a nonreference genotype; PASS – number of variants passing for which individual has a nonreference genotype; PASS_S - number of singletons passing for which individual has a (singleton) nonreference genotype; QUAL - mean QUAL for variants for which individual has a nonreference genotype; DP - mean variant DP for variants for which individual has a nonreference genotype. More details about i-stats command output can be found at https://atgu.mgh.harvard.edu/plinkseq/stats.shtml#ind

The file tags (listed at the top of the "pseq myproj var-summary" results as "1" for the CEU imported VCF file and "2" for YRI imported VCF file) can be changed to more identifiable names using the commands:

pseq myproj tag-file --id 1 --name CEU

pseq myproj tag-file --id 2 --name YRI

To view changes use the command:

pseq myproj var-summary

```
---Variant DB summary---

6987 unique variants
File tag : CEU (3489 variants, 90 individuals)
File tag : YRI (5175 variants, 112 individuals)
```

This will help us later for viewing population specific data as well as filtering and analyzing data based on population.

Variant statistics

Variant statistics such as Hardy-Weinberg equilibrium, minor allele count, and minor allele frequency can be output using the "v-freq" command:

pseq myproj v-freq | head

VAR	CHR	POS	REF	ALT	FILTER	QUAL	TI	GENO	MAC	MAF	REFMIN	HWE	HET	NSNP
chr1:1115461	1	1115461	C	T	PASS	.	1	0.311881	4	0.031746	0	1	0.0634921	3
chr1:1115503	1	1115503	T	C	SBFilter	.	1	0.282178	4	0.0350877	0	1	0.0701754	2
chr1:1115510	1	1115510	C	T	PASS	.	1	0.331683	2	0.0149254	0	1	0.0298507	2
chr1:1115548	1	1115548	G	A	PASS	.	1	0.262376	1	0.00943396	0	1	0.0188679	1
chr1:1115604	1	1115604	C	A	PASS	.	0	0.287129	3	0.0258621	0	1	0.0517241	0
chr1:1118275	1	1118275	C	T	PASS	.	1	0.579208	45	0.192308	0	0.367544	0.282051	0
chr1:1119399	1	1119399	C	T	PASS	.	1	0.49505	3	0.015	0	1	0.03	1
chr1:1119434	1	1119434	C	A	PASS	.	0	0.49505	1	0.005	0	1	0.01	0
chr1:1120370	1	1120370	C	G	PASS	.	0	0.49505	16	0.08	0	0.478564	0.14	2

Please note that it is not valid to filter for deviation from HWE using the entire project since there are two populations, instead the HWE much be examined for each individual project.
For population specific variant statistics use the "--mask" flag with the "file" option:

pseq myproj v-freq --mask file=CEU | head

VAR	CHR	POS	REF	ALT	FILTER	QUAL	TI	GENO	MAC	MAF	REFMIN	HWE	HET	NSNP
chr1:1115503	1	1115503	T	C	SBFilter	0	1	0.633333	4	0.0350877	0	1	0.0701754	1
chr1:1115548	1	1115548	G	A	PASS	0	1	0.588889	1	0.00943396	0	1	0.0188679	0
chr1:1118275	1	1118275	C	T	PASS	0	1	0.677778	3	0.0245902	0	1	0.0491803	0
chr1:1120377	1	1120377	T	A	PASS	0	0	0.988889	1	0.00561798	0	1	0.011236	1
chr1:1120431	1	1120431	G	A	PASS	0	1	0.855556	6	0.038961	0	1	0.0779221	0
chr1:3548136	1	3548136	T	C	PASS	0	1	0.811111	18	0.123288	1	1	0.219178	0
chr1:3548832	1	3548832	G	C	PASS	0	0	0.988889	13	0.0730337	0	1	0.146067	0
chr1:3551737	1	3551737	C	T	PASS	0	1	0.988889	1	0.00561798	0	1	0.011236	1
chr1:3551792	1	3551792	G	A	PASS	0	1	1	8	0.0444444	0	1	0.0888889	0

pseq myproj v-freq --mask file=YRI | head

VAR	CHR	POS	REF	ALT	FILTER	QUAL	TI	GENO	MAC	MAF	REFMIN	HWE	HET	NSNP
chr1:1115461	1	1115461	C	T	PASS	0	1	0.5625	4	0.031746	0	1	0.0634921	1
chr1:1115510	1	1115510	C	T	PASS	0	1	0.598214	2	0.0149254	0	1	0.0298507	1
chr1:1115604	1	1115604	C	A	PASS	0	0	0.517857	3	0.0258621	0	1	0.0517241	0
chr1:1118275	1	1118275	C	T	PASS	0	1	0.5	42	0.375	0	0.395585	0.535714	0
chr1:1119399	1	1119399	C	T	PASS	0	1	0.892857	3	0.015	0	1	0.03	1
chr1:1119434	1	1119434	C	A	PASS	0	0	0.892857	1	0.005	0	1	0.01	0
chr1:1120370	1	1120370	C	G	PASS	0	0	0.892857	16	0.08	0	0.478564	0.14	1
chr1:1120431	1	1120431	G	A	PASS	0	1	0.741071	67	0.403614	0	0.360868	0.542169	4
chr1:1120488	1	1120488	A	C	PASS	0	0	0.857143	10	0.0520833	0	1	0.104167	3

As you see, the "--mask" flag is used to set conditions for the viewing or filtering variants or individuals.
More details about "v-freq" command can be found at
https://atgu.mgh.harvard.edu/plinkseq/tutorial.shtml

Data Cleaning

Removal of low quality variants

To view the number of variants that passed all quality filters:

```
pseq myproj v-view --mask any.filter.ex | head
```

```
chr1:1115461    .               C/T     .       1       PASS
chr1:1115510    .               C/T     .       1       PASS
chr1:1115548    .               G/A     .       1       PASS
chr1:1115604    .               C/A     .       1       PASS
chr1:1118275    rs61733845      C/T     .       2       PASS
chr1:1119399    .               C/T     .       1       PASS
chr1:1119434    .               C/A     .       1       PASS
chr1:1120370    .               C/G     .       1       PASS
chr1:1120377    .               T/A     .       1       PASS
chr1:1120431    rs1320571       G/A     .       2       PASS
```

```
pseq myproj v-view --mask any.filter.ex | wc -l
```

There are 6986 unique variant sites that have passed the quality filters. The "--mask" flag gives the condition(s) that must be met for the variant to be listed. Here "any.filter.ex" tells **pseq** to remove any variants that failed 1 or more quality filters. Only variants that have a 'PASS' value in the FILTER field of the vcf file will be selected. More details about filtering variants on FILTER field can be found at https://atgu.mgh.harvard.edu/plinkseq/masks.shtml#filter

To view the number of variants that failed any quality filter:

```
pseq myproj v-view --mask any.filter | wc -l
```

One variant failed the filter. To select only variants that passed all quality filters:

```
pseq myproj var-set --group pass --mask any.filter.ex
```

```
pseq myproj var-summary
```

```
---Variant DB summary---

6987 unique variants
File tag : CEU (3489 variants, 90 individuals)
File tag : YRI (5175 variants, 112 individuals)

Set pass containing 8663 variants
```

The "var-set" option tells **pseq** that we will be creating a new set of variants, the input following the "--group" flag gives the name of the new variant set, and the input following the "--mask" flag gives the condition(s) that must be met for the variant to be included in the new variant set.

If we consider variant sites with a read depth < 15 as low quality variant sites and we want to remove variants that did not meet this threshold. Note that 'DP', which denotes total read depth of a variant site, is contained in the INFO field of vcf file.

```
pseq myproj var-set --group pass_DP15 --mask include="DP>14" var=pass
```

```
pseq myproj var-summary
```

```
---Variant DB summary---

6987 unique variants
File tag : CEU (3489 variants, 90 individuals)
File tag : YRI (5175 variants, 112 individuals)

Set pass containing 8663 variants
Set pass_DP15 containing 8662 variants
```

Only one variant site is removed. The "var=allpass" option allows us to use a previously defined variant set as a reference for additional filtering of a previously filtered variant set. By using various "--mask" commands you can filter out variants that are not useful for your particular study.

Filter data by genotype read depth 10

```
pseq myproj var-set --group pass_DP15_DPgeno10 --mask geno=DP:ge:11 var=pass_DP15
```

```
pseq myproj var-summary
---Variant DB summary---

6987 unique variants
File tag : CEU (3489 variants, 90 individuals)
File tag : YRI (5175 variants, 112 individuals)

Set pass containing 8663 variants
Set pass_DP15 containing 8662 variants
Set pass_DP15_DPgeno10 containing 8662 variants
```

This command sets all genotypes with a sequencing depth (DP) < 11 to null using the option "geno=DP:ge:11". In the vcf file, genotype level DP information is contained in the genotype columns, present under each individual ID and is specific to every individual's genotype. Available genotype level information is denoted by FORMAT column in the vcf file.

Association Tests for a Quantitative Trait

NOTE: From this step forward the association tests will be performed for the CEU population only. The "file=YRI" tag can be used to perform the same tests on the YRI data.

Select CEU variant sites

```
pseq myproj var-set --group pass_DP15_DPgeno10_CEU --mask file=CEU var=pass_DP15_DPgeno10
```

```
pseq myproj var-summary
```

```
---Variant DB summary---
…
Set pass_DP15_DPgeno10_CEU containing 3488 variants
```

There are 3488 variant sites that can be found in CEU population dataset after QC.

Exclude variant sites with HWE p-value < 5.7e-7

```
pseq myproj var-set --group pass_DP15_DPgeno10_CEU_HWE --mask hwe=5.7e-7:1 var=pass_DP15_DPgeno10_CEU
```

```
pseq myproj var-summary
```

```
---Variant DB summary---
…
Set pass_DP15_DPgeno10_CEU containing 3479 variants
```

There are 3479 variant sites that are in HWE (Hardy-Weinberg equilibrium) in CEU population. Details about tests for deviation from HWE can be found at http://en.wikipedia.org/wiki/Hardy–Weinberg_principle. Here we use a p-value cutoff of 5.7e-7 to exclude variant sites, for more details see reference http://www.nature.com/nature/journal/v447/n7145/full/nature05911.html

Filter variants by minor allele frequency (MAF)

We wish to analyze variant sites with different allele frequencies. In order to obtain the different data sets the following commands are used.

To extract variant sites with MAF ≥ 0.05:

```
pseq myproj var-set --group pass_DP15_DPgeno10_CEU_HWE_MAFgt05 --mask maf=0.05:0.5
var=pass_DP15_DPgeno10_CEU_HWE
```

```
pseq myproj var-summary
```

```
---Variant DB summary---
…
Set pass_DP15_DPgeno10_CEU_HWE_MAFgt05 containing 1429 variants
```

There are 1429 variant sites in the CEU data set that pass QC with a MAF ≥ 0.05. These variant sites are saved to the variant table; `pass_DP15_DPgeno10_CEU_HWE_MAFgt05.`

To extract variant sites with MAF ≤ 0.01:

```
pseq myproj var-set --group pass_DP15_DPgeno10_CEU_HWE_MAFlt01 --mask "mac=1 maf=0.01"
var=pass_DP15_DPgeno10_CEU_HWE
```

```
pseq myproj var-summary
```

```
---Variant DB summary---
Set pass_DP15_DPgeno10_CEU_HWE_MAFlt01 containing 1083 variants
```

There are 1083 variant sites in the CEU dataset which pass QC with a MAF ≤ 0.01. The variant sites are saved to the variant table; pass_DP15_DPgeno10_CEU_HWE_MAFlt01. Note that condition “mac=1” excludes monomorphic sites.

More details about --mask options on filtering variants on sample polymorphism can be found at `https://atgu.mgh.harvard.edu/plinkseq/masks.shtml#maf`

Analysis of common variants (MAF ≥ 0.05)

To run a linear or logistic regression on each single variant, use the glm command. The type of test will depend on the phenotype (quantitative trait or dichotomous disease trait).

To detect single variant association between quantitative phenotype BMI, controlling for sex and a group of variants, contained in variant table pass_DP15_DPgeno10_CEU_HWE_MAFgt05, filtered using each of the previous filtering conditions:

```
pseq myproj glm --phenotype BMI --covar SEX --mask var=pass_DP15_DPgeno10_CEU_HWE_MAFgt05 > SNV_CEU.result
```

```
head SNV_CEU.result
```

VAR	REF	ALT	N	F	BETA	SE	STAT	P
chr1:3548136	T	C	73	0.876712	-1.53374	1.85033	-0.828897	0.40998
chr1:3548832	G	C	89	0.0730337	1.13049	2.26738	0.49859	0.619341
chr1:6524501	T	C	86	0.0697674	0.433904	2.49357	0.174009	0.862282
chr1:6524688	T	C	88	0.0511364	-1.86795	2.70494	-0.690568	0.491718
chr1:11710561	T	G	47	0.117021	-0.347495	1.92692	-0.180337	0.857716
chr1:17914057	G	A	86	0.0755814	-1.59486	2.34734	-0.679432	0.498754
chr1:17914122	G	A	85	0.0823529	2.61561	2.1748	1.20269	0.232558
chr1:17961345	C	T	68	0.110294	2.99054	2.00047	1.49492	0.139775
chr1:17981184	A	C	80	0.15	-1.83108	1.63531	-1.11972	0.266315

The output statistics are VAR – variant identifier; REF – reference allele; ALT – alternate allele(s); N – number of individuals included in analysis; F – frequency of the alternate allele(s); BETA – regression coefficient; SE – standard error of estimate; STAT – test statistic; P – asymptotic p-value. More details about linear and logistic regression models can be found at https://atgu.mgh.harvard.edu/plinkseq/assoc.shtml#glm

To view the results sorted by p-value:

```
cat SNV_CEU.result | awk '{if(FNR==1) print $0; if(NR>1) print $0 | "sort -k9"}' | grep -v "NA\s\+NA\s\+NA" | head
```

VAR	REF	ALT	N	F	BETA	SE	STAT	P
chr11:108383676	A	G	90	0.138889	6.36308	1.60942	3.95365	0.000156342
chr19:16008388	A	C	53	0.122642	6.88317	1.73915	3.95778	0.000239339
chr19:16006413	G	A	80	0.1	6.31788	1.78167	3.54604	0.000669193
chr14:39901157	C	A	36	0.0555556	10.8531	3.12283	3.47542	0.00144933
chr16:57735900	G	C	80	0.29375	-4.18114	1.43663	-2.91039	0.004718
chr2:49189921	C	T	90	0.588889	-3.345	1.17772	-2.84025	0.0056123
chr7:156742501	C	G	9	0.277778	-12.1592	2.89402	-4.20149	0.00567644
chr2:49191041	C	T	89	0.58427	-3.36254	1.19515	-2.81348	0.00607226
chr15:25926204	C	G	83	0.0783133	5.79532	2.13611	2.71302	0.00816109

Analysis of rare variants (MAF ≤0.01)

PSEQ has a collection of gene-based tests, see https://atgu.mgh.harvard.edu/plinkseq/assoc.shtml#genic for details.

However, Currently only the SKAT and SKAT-O can be used to analyze quantitative traits so the SKAT test will be used in the following rare variant burden analysis (if we choose to use other tests, e.g. WSS – frequency-weighted test, VT – variable threshold test, etc., the following error will be returned.

```
pseq myproj assoc --tests fw vt --phenotype BMI
pseq error : only SKAT/SKAT-O can handle quantitative traits
```

To perform SKAT, where rare variants aggregated across a gene region, a group-by mask is required. Here we use loc.group=refseq, where refseq denotes NCBI Reference Sequence Database. More details about

grouping variants can be found at https://atgu.mgh.harvard.edu/plinkseq/masks.shtml#groups. More details about refseq can be found at http://www.ncbi.nlm.nih.gov/refseq/

When performing single variant analysis data QC can be performed and then variant table containing selected variants can be analyzed. If a rare variant aggregate association test is being performed it is not possible using PSEQ to specify the name of the variant table, instead all of the QC parameters must be included in the command line in addition to the association test parameters.

Running the SKAT test using the variant table results in an error:

```
pseq myproj assoc --tests skat --phenotype BMI --covar SEX --mask var=pass_DP15_DPgeno10_CEU_HWE_MAFlt01 loc.group=refseq > SKAT_CEU.result

pseq error : you cannot specify other includes in the mask with loc.group
```

Additional details can be found at https://atgu.mgh.harvard.edu/plinkseq/whatisnew.shtml),

Although we use the most recent version pseq-0.10 in this exercise (for which there is no updated documentation), the error still remains unresolved. Therefore, we have to redo cleaning on original data by re-specifying each filtering condition and run SKAT using one command as below:

```
pseq myproj assoc --tests skat --phenotype BMI --covar SEX --mask include="DP>14" geno=DP:ge:11 file=CEU hwe=5.7e-7:1 "mac=1 maf=0.01" loc.group=refseq > SKAT_CEU.result

head -20 SKAT_CEU.result
```

LOCUS	POS	ALIAS	NVAR	TEST	P	I	DESC
NM_000055	chr3:165548187	G/A	W=1	0:0			
NM_000055	chr3:165548187..165548187	BCHE	1	SKAT	0.237374	.	.
NM_000112	chr5:149359938	C/G	W=1	0:0			
NM_000112	chr5:149360143	T/C	W=1	0:0			
NM_000112	chr5:149360212	A/G	W=1	0:0			
NM_000112	chr5:149360215	T/C	W=1	0:0			
NM_000112	chr5:149361245	G/A	W=1	0:0			
NM_000112	chr5:149359938..149361245	SLC26A2	5	SKAT	0.293096	.	.
NM_000119	chr15:43498537	C/T	W=1	0:0			
NM_000119	chr15:43499436	G/A	W=1	0:0			
NM_000119	chr15:43500478	C/T	W=1	0:0			
NM_000119	chr15:43498537..43500478	EPB42	3	SKAT	0.422114	.	.
NM_000122	chr2:128016983	C/T	W=1	0:0			
NM_000122	chr2:128038204	T/C	W=1	0:0			
NM_000122	chr2:128016983..128038204	ERCC3	2	SKAT	0.386466	.	.
NM_000124	chr10:50732644	G/C	W=1	0:0			
NM_000124	chr10:50738781	T/C	W=1	0:0			
NM_000124	chr10:50740844	G/A	W=1	0:0			
NM_000124	chr10:50740861	C/T	W=1	0:0			

For each gene region the list of the variants within the gene are listed, followed by gene-based association results. The I field is only available for case control data and provides the smallest possible empirical p-value which can be obtained for the variant sites and the DESC field which is also only available for case control data and it provides the number of case and control alternative alleles. Since we are analyzing quantitative trait data these fields are blank. Detailed explanation about each output field can be found at https://atgu.mgh.harvard.edu/plinkseq/assoc.shtml#genic

To view the smallest p-values for each SKAT test:

```
cat SKAT_CEU.result | grep SKAT | grep -v "P=NA" | sort -k6 | head -15
NM_024837       chr15:50152449..50264848        ATP8B4  5       SKAT    0.00405073      .       .
NM_001055       chr16:28617413..28617413        SULT1A1 1       SKAT    0.00418122      .       .
NM_177529       chr16:28617413..28617413        .       1       SKAT    0.00418122      .       .
NM_177530       chr16:28617413..28617413        .       1       SKAT    0.00418122      .       .
NM_177534       chr16:28617413..28617413        .       1       SKAT    0.00418122      .       .
NM_177536       chr16:28617413..28617413        .       1       SKAT    0.00418122      .       .
NM_001137559    chr12:121746337..121764935      ANAPC5  3       SKAT    0.00621198      .       .
NM_016237       chr12:121746337..121764935      .       3       SKAT    0.00621198      .       .
NM_006371       chr3:33174163..33174163         CRTAP   1       SKAT    0.00748816      .       .
NM_006944       chr2:234959642..234967570       SPP2    3       SKAT    0.00753125      .       .
NM_018328       chr2:149221327..149241000       MBD5    4       SKAT    0.00755692      .       .
NM_000782       chr20:52779338..52779338        CYP24A1 1       SKAT    0.00794735      .       .
NM_001128915    chr20:52779338..52779338        .       1       SKAT    0.00794735      .       .
NM_001018088    chr15:62204043..62302757        .       3       SKAT    0.0221564       .       .
NM_017684       chr15:62204043..62302757        VPS13C  3       SKAT    0.0221564       .       .
```

Note that each test has been performed on each alternative transcript (NM_*) of each gene, e.g. transcripts NM_001055, NM_177529, NM_177530, NM_177534 and NM_177536 all belong to gene SULT1A1.

Questions

Repeat the above analysis but using the data from the Yoruba (YRI) population and answer the following questions.

Question 1

List the four smallest p-values for the single variant tests for the common variants i.e. MAF ≥0.05:

1.) _____

2.) _____

3.) _____

4.) _____

List the four smallest p-values for the SKAT rare variant test:

1.) _____

2.) _____

3.) _____

4.) _____

Answers

Question 1

Single variant test

1.) ___ chr21:26979752_____ 0.00084882_____

2.) ___ chr17:3445901 _____0.000956475_____

3.) ___ chr17:9729445_____ 0.0010022_____

4.) ___ chr19:15303225_____0.0011692_____

SKAT aggregate burden test

1.) ___ NM_207317_____ 0.0210752_____

2.) ___ NM_032048_____ 0.0238947_____

3.) ___ NM_002738_____ 0.0255961_____

4.) ___ NM_212535_____ 0.0255961_____