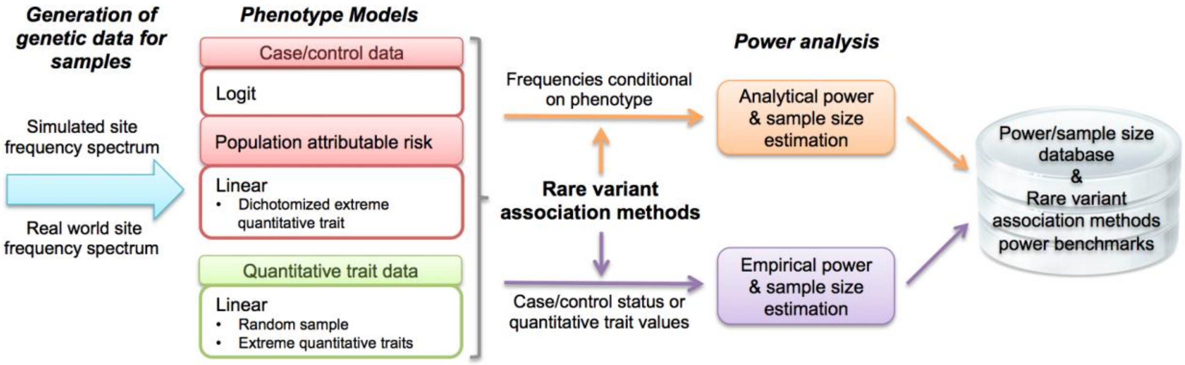
- SUPPLEMENTAL MATERIAL -

Power analysis and sample size estimation for sequence-based association studies

Introduction

Recent evidence suggests that rare variants within the human genome might have a strong impact on the risk for complex diseases. An essential first step in designing genetic association studies is to assess the sample size needed to achieve sufficient statistical power and to choose appropriate statistical methods for association testing. SEQPower employs sophisticated modeling of human genome sequences and complex diseases, and rapidly conducts customized power analysis using recently developed rare variant association methods. This supplemental material provides some basic examples using SEQPower to perform power and sample size evaluation for gene-based rare variant association studies.

Procedures for a typical power analysis in SEQPower consists of four steps: the simulation or sampling of human DNA sequences with rare variants, the simulation of phenotypes conditional on the sequence data, the sampling procedures for the specified study design, and the power and sample size estimation.


Generation of genetic data for samples
Simulated site frequency spectrum
Real world site frequency spectrum
Phenotype Models
Case/control data
Logit
Population attributable risk
Linear
• Dichotomized extreme quantitative trait
Quantitative trait data
Linear
• Random sample
• Extreme quantitative traits
Frequencies conditional on phenotype
Rare variant association methods
Case/control status or quantitative trait values
Power analysis
Analytical power & sample size estimation
Empirical power & sample size estimation
Power/sample size database & Rare variant association methods power benchmarks

Methodology

SEQPower provides a number of study designs and association analysis methods. Power can be evaluated for gene-based association studies under three designs: case-control, quantitative traits, and extreme quantitative traits. For case-control data, for example, phenotypes can be generated conditional on variants within a gene simulated by a European demographic model. Phenotypes can be generated based upon the disease prevalence and odds ratio(s) for specific variants within a gene (the LOGIT model option in SEQPower) or their population attributable risks (the PAR model option in SEQPower). For quantitative traits data a normally distributed phenotype using a linear model can be generated, assuming that the genetic effects are additive. Power calculations can be performed on simulated data using analytical or empirical methods. In this documentation we will perform power analysis using several popular rare variant association methods including the Combined Multivariate Collapsing (CMC) method (Li and Leal, 2008), Weighted Sum Statistic (WSS) (Madsen and Browning, 2009), Kernel Based Adaptive Cluster (KBAC) (Liu and Leal, 2010), Variable Threshold (VT) (Price *et al.*, 2010) and Sequence Kernel Association Test (SKAT) (Wu *et al.*, 2011). For more information on the rare variant association tests, see Table S1. For all tests it is possible to obtain power estimates empirically. Although obtaining power analytically is efficient, it is limited to the CMC method. Power calculations are demonstrated using the LOGIT model to generate case-control data and the linear model to generate quantitative traits.

Calculation of power can be computationally intensive. For the examples provided in the supplemental material relatively small sample sizes are used for demonstration purposes, so the reader can quickly reproduce the results. You may not wish to run all of the examples, but rather examine the input and output to get an idea how power evaluation is performed for rare variant association methods.

There are two versions of SEQPower: a lite and a full version. Full version allows for generation of variant data while the lite version does not. The lite version can be used when it is desired to use variant data from another source which are either simulated or actual variant data. On the SEQPower website the user is provided with data bundles (see section 1.3.1) that contain pre-generated variant data using a variety of population demographic models which can be used for power calculations. These data were generated with the full version of SEQPower using the `spower simulate` command. In this document we will only demonstrate the lite version of SEQPower.

Resources

Data resource

For the examples simulated sequence data can be downloaded from:

`http://bioinformatics.org/spower/download/data/SRV/sfs.tar.gz`.

The data bundle contains files having information on minor allele frequency (MAF) spectrum, purifying selection coefficients and variant positions, which are calculated from simulated data using the following genetic demographic models:

- A two-epoch model (Williamson *et al.*, 2005)
- European population with bottlenecks (Boyko *et al.*, 2008)
- African population with bottlenecks (Boyko *et al.*, 2008)
- European population with bottlenecks and exponential expansion (Kryukov *et al.*, 2009)

The provided data bundle contains simulated variants for genes of length 1,800 bp. Each simulated dataset has 200 haplotype pools. The number (N) of haplotypes within each haplotype pool depends on the population demographic model which was used to generate the data. The number of haplotypes per pool is as follows: two-epoch model, N = 102,680; European population with bottlenecks, N = 105,940; African population with bottlenecks, N = 51,272; and European population with bottlenecks and exponential expansion, N = 180,000. In the examples a haplotype pool that was generated by implementing Kyrukov's European demographic model is used.

Genes using other models and different lengths can be conveniently simulated using the spower simulate command which is found in the full version of SEQPower or using other software such as SFSCODE. For additional information on using the full version of SEQPower to generate variant data please see the online documentation `http:// bioinformatics.org/spower/srvbatch`

Purifying selection coefficients in the simulated data can be used to determine the "functionality" of each variant. Variants with selection coefficients $> 1\times10^{-5}$ or $< -1\times10^{-5}$ are deemed to be "functional" while those with selection coefficients between -1×10^{-5} and 1×10^{-5} are "neutral". As will be seen later, although a variant is functional it is not necessarily causal and the simulations are performed with various proportions of the variants being causal. Using the above selection coefficients, approximately 37% of the variants are neutral, mimicking the proportion of synonymous variants within the exome. Neutral variants are not analyzed, which is also usually the case for synonymous variants in real world analyses. Other selection coefficient cut-offs can be used to determine which variants are functional and neutral. Additionally annotation other than selection coefficients can also be used to determine functionality of the generated variants.

Online support

Documentation and *tutorial* can be found at `http://bioinformatics.org/spower`. The *documentation* provides two manuals; 1) User guide: with information on commands and recommendations on choice of parameters; and 2) Technical notes: with details on the implementation of the simulations and statistical analysis. The *tutorial* contains additional examples on power and sample size estimation.

Getting Started

Command options to implement simulations and power calculation can be found in Table S2. From the command terminal, the `-h` option can be used to obtain help and useful information, for example,

```
spower -h
usage: spower [-h] [--version] {LOGIT,PAR,BLNR,LNR,ELNR,show,execute} ...
```

To view information for a specific model or option, type `spower <name> -h`, for example:

```
spower LOGIT -h
```

Input of a typical SEQPower command is composed of:

- Modeling variant-phenotype associations;
- Information on samples collected;
- Sequencing and genotyping artifacts, *i.e.*, missing data and sequencing errors;

- For empirical power calculations, options for rare variant association test to be performed. For analytical power analysis and sample size estimation the CMC method is used. For a complete listing of rare variant association methods which can be used for power and sample size estimation please see Table S1.

The command line options for each model in SEQPower consist of one required positional argument, *i.e.*, the input data. Additionally optional arguments can be used. For some of the optional arguments either short or long syntax can be used (see Table S2), which are equivalent. For clarity, hereafter we will use long version syntax for those crucial optional arguments in the provided examples.

Below is an example command and screen output for running the LOGIT model to generate the phenotype data and using CMC method to perform the association analysis. This example aims to demonstrate the SEQPower interface and we will examine in detail the usage of each parameter option in the next section:

```
spower LOGIT Kryukov2009European1800.sfs --sample_size 1000 --OR_rare_detrimental  1.5 --method \
      "CFisher --name CMC" -r 1000 -j 4 -l 1 -o exercise
```

```
INFO: Loading data from [Kryukov2009European1800.sfs] ...
INFO: 200 units found
INFO: 1 units will be analyzed
R1 : 100% |=====================================================| Time: 0:01:13
INFO: Tuning [exercise.SEQPowerDB]  ...
INFO: Result saved to [exercise.csv/.loci.csv/.SEQPowerDB]
```

View and interpret the output

The lines starting with INFO are information lines printed on screen. The first three INFO lines provide a summary of input data. For the simulated data an analysis "unit" refers to sequence of a gene from one "haplotype pool". The output data are saved to plain text files `*.csv` as well as a database file `*.SEQPowerDB`. You can either browse the text file with a text editor, or more conveniently, use `spower show options`. To show all available fields in text output,

```
spower show exercise.csv
```

```
title
name
method
power
power_median power_std
case_cmaf
case_cmaf_median
case_cmaf_std
cmaf
cmaf_detrimental
cmaf_neutral
cmaf_protective
ctrl_cmaf
ctrl_cmaf_median
ctrl_cmaf_std
...
```

To extract a particular field, for example, the power estimates:

```
spower show exercise.csv  power*
```

```
+-------+---------------+
| power |   power_std   |
+-------+---------------+
| 0.289 | 0.01433453870  |
+-------+---------------+
```

Logging and Summary

Optional argument `-v` controls verbosity levels. The default verbosity level of the program is “2”, which will output all INFO and WARNING messages on the screen. `-v 1` will only show an overall progress bar for all test units and `-v 0` will suppress all screen output.

The results of the power analysis are stored in a database. A text file containing the power analysis results is generated. Two additional files are generated: the `*.log` file records the command line history for SEQPower and all INFO, WARNING and DEBUG messages generated during runtime, while `*.loci.csv` contains summary information of each locus of the unit being analyzed. Below are a few examples of the types of information provided in the `*.loci.csv` file:

- Odds ratios per variant site
- Population attributable risk per variant site
- Functionality of variant sites: neutral or functional
- Causality of variant sites: non-causal, increased or decreased disease risk, or quantitative trait values
- Genotype frequency per variant site

`spower show` command can be applied to the summary text file to view all column names in the text file or the values of a particular column in the file, e.g. “maf”,

```
spower show exercise.loci.csv
spower show exercise.loci.csv maf
```

Details on association analysis methods

Use `spower show tests` to list all available tests, and `spower show test <name>` to list options for a specific association test.

```
spower show tests
spower show test SKAT
```

Details on technical procedures for association test methods can be found in Table S1.

Example 1: Power Analysis for Case-control Design

In this example binary phenotype-genotype associations will be modeled using disease prevalence and variant-specific odds ratios. This model is named `LOGIT` in SEQPower. The example in the previous section is revisited, but this time we use the KBAC association method to analyze the simulated data. For demonstration purposes most arguments are written out including some that use default values.

```
spower  LOGIT  Kryukov2009European1800.sfs  \
--def_rare  0.01  --def_neutral  -0.00001  0.00001  --moi  A  \
--proportion_detrimental  1  --proportion_protective  0  \
--OR_rare_detrimental  1.5  --OR_common_detrimental  1  --baseline_effect  0.01  \
--sample_size  1000  --p1  0.5  --limit  1  \
--alpha  0.05  \
--method "KBAC  --name  K1  --mafupper  0.01  --maflower  0  --alternative  1  \
--moi  additive  --permutations  1000  --adaptive  0.1"  \
--replicates  1000  \
--jobs  4  -o  exercise
```

A few key parameters are explained below. It should be noted that some parameters are not specified but instead their default values are used (see Table S2).

- `--def_rare` The definition of a “rare” variant. A rare variant is often specified as having a MAF ≤ 0.01. An alternative definition for rare variants may be used e.g. MAF<0.5% or it may be desired to analyze both rare and low frequency variants and a higher MAF can be used e.g. < 5%.
- `--def_neutral` Variants determined to be neutral by their functional annotation score. In this simulated dataset selection coefficients are used to annotate functionality. For this example, variants having selection coefficient between -0.00001 and 0.00001 are considered to be neutral.
- `--moi` “MOI” stands for “mode of inheritance”. This option controls the underlying MOI for simulation of phenotypes conditional on functional/causal variant sites. The default MOI is “additive”; other options include “dominant”, “recessive” and “multiplicative”. For some of the association methods, e.g. WSS, different coding can be used which takes into account different underlying genetic models. If it is desired to apply MOI for association tests, the --moi argument which is nested inside the --method option can be used.
- `--proportion_detrimental` This parameter allows us to model situations where not all functionally deleterious variants are causal. Here we assume all deleterious variants are causal for the trait of interest, thus `--proportion_detrimental 1.0`. We can use a value less than 1.0 if we want to model the impact of deleterious variants which are non-causal.
- `--proportion_protective` This is the proportion of causal variants which are protective.
- `--OR_rare_detrimental` The odds ratio per “rare” (MAF < `def_rare`) detrimental variant.
- `--OR_common_detrimental` The odds ratio per “common” (MAF ≥ `def_rare`) detrimental variant.
- `--baseline_effect` The baseline odds ratio in population. For common complex traits involving rare variants, this is approximately the same as prevalence of disease in population.

- `--sample_size` Total sample size.
- `--p1` Proportion of cases. For the case-control design, 50% cases and 50% controls will yield maximum power.
- `--limit` In the input data `Kryukov2009European1800.sfs` there are many haplotype pools which were generated using the same forward time simulation setting. The number of haplotype pools used in the analysis can be set using the `--limit` argument. In this demonstration only one haplotype pool is used (`--limit 1`).
- `--alpha` Significance level for which power will be evaluated.
- `--method` Name of rare variant association method to be applied, see Table S1 and `spower show tests`

  * `--name` A unique name assigned to a specific analysis. For the same rare variant association method you might wish to have runs using different analysis parameters, in which case you may want to specify a name in order to distinguish between different settings. These labels are particularly useful when you have multiple configurations of the same method in the same command, for example `--method “KBAC --name K1 ... --mafupper 0.01” “KBAC --name K5 ... --mafupper 0.05”`
  * `--mafupper 0.01/--maflower 0.0` We define rare variants within a population haplotype pool using `--def_rare` option in order to generate phenotypic data. For rare variant association methods we also want to define frequencies of the variants to be analyzed based upon the observed frequencies within the sample. This can be done using the `--mafupper` and `--maflower` commands.
  * `--alternative` To indicate if a one-sided or two-sided test should be performed. Use 1 for one-sided test, 2 for two-sided test. When only testing for detrimental variants, it is more powerful to apply a one-sided test.
  * `--permutations/--adaptive` In the previous section we used Fisher’s method for CMC test implementation, which does not require permutations, since p-values can be obtained analytically instead of empirically. For rare variant tests, for which p-values must be obtained empirically, permutation is used. The number of permutations needed to estimate the p-value, depends on the α level. For $\alpha = 0.05$, it is sufficient to use 1,000 permutations. For exome studies with $\alpha = 2.5\times10^{-6}$ (Bonferroni-corrected p-value for testing 20,000 genes) greater than 10^7 permutations should be used. Although SEQPower can perform “adaptive” permutation (via argument `--adaptive`), when calculating power a large number of permutations would still be required because the majority of replicates are falling under the alternative. Therefore it is usually not feasible to use permutation-based tests to evaluate power for very small α levels.
- `--replicates` Number of replicates to be used for power / sample size estimation.
- `--jobs` Number of CPUs to be used to run the command. The input of this parameter depends on your computational environment.

Result of the above analysis is as follows:

```
+-------+---------------+
| power |   power_std   |
+-------+---------------+
| 0.347 | 0.01505293991 |
+-------+---------------+
```

Variable effect sizes of rare variants

The previous example uses a fixed odds ratio =1.5 for detrimental variants. To use variable odds ratios, *e.g.* in range [1.2, 3.0], the following is input:

```
spower LOGIT Kryukov2009European1800.sfs --sample_size 1000 \
--OR_rare_detrimental 1.2 --ORmax_rare_detrimental 3.0 \
  --method CFisher -r 1000 -j 4 -l 1 -o exercise
```

Instead of using a fixed odds ratio, now the odds ratio will be generated based on the underlying MAFs. You can observe the generated odds ratios in the resulting `*.loci.csv` file:

```
spower show exercise.loci.csv effect*
```

```
+---------------+
|  effect_size  |
+---------------+
...
|      1.0      |
...
| 1.42859361773 |
...
|      3.0      |
| 2.99121719826 |
| 2.99820319113 |
| 2.99940345946 |
|      1.2      |
...
```

The effect size for detrimental variants will range from 1.2 to 3.0; effect size of neutral variants is 1.0.

Presence of non-causal variants

Now we assume only 80% of the deleterious variants are causal, *i.e.*, although functionally not neutral, 20% of deleterious variants do not contribute to the particular phenotype of interest. We perform analysis under this assumption as follows:

```
spower LOGIT Kryukov2009European1800.sfs --sample_size 1000 \
--OR_rare_detrimental 1.2 --ORmax_rare_detrimental  3.0 \
--proportion_detrimental 0.8 \
--method CFisher -r 1000 -j 4 -l 1 -o exercise
```

With this option the program will still simulate all the deleterious variant sites, regardless of them being causal or not. However since they are not associated with the phenotype, they are essentially “noise” in data and can reduce the power.

```
+-------+------------------+
| power |    power_std     |
+-------+------------------+
| 0.084 | 0.00877177291088  |
+-------+------------------+
```

Exclusion of causal variants

Different from presence of non-causal variants, this option models the situation when variants are presented in the population and may or may not have contributed to the phenotype, but were removed from sample dataset for association analysis, due to lack of coverage, low mapping / variant calling quality, *etc.* (exclusion of true positive signals in analysis). For example we set 5% sites missing due to such artifacts:

```
spower  LOGIT  Kryukov2009European1800.sfs  --sample_size  1000  \
--OR_rare_detrimental  1.5  --missing_sites  0.05  \
--method  CFisher  -r 1000  -j 4  -l 1  -o  exercise
```

```
+-------+----------------+
| power |    power_std    |
+-------+----------------+
| 0.325 | 0.0148113132436 |
+-------+----------------+
```

It is also possible to model the proportion of missing data by MAF, using the `--missing_low_maf` option to set variants below certain population MAF as missing, irrespective of functionality. The example below mimics the "exome chip" design, where all variants on the chip have population MAFs greater than 0.000125, *i.e.* appear at least 3 times in 12,000 sequence samples that were used to design the exome chip (G. Abecasis, personal communication).

```
spower  LOGIT  Kryukov2009European1800.sfs  --sample_size  1000  \
--OR_rare_detrimental  1.5  --missing_low_maf  0.000125  \
--method  CFisher  -r 1000  -j 4  -l 1  -o  exercise
```

Using multiple association methods

It is possible to apply multiple association tests in one command, in order to compare methods. In previous sections we have introduced CMC and KBAC. Here we also use three additional rare variant association tests:

- WSS. The WSS method up-weights rarer variants which can amplify association signals if they are causal, and is powerful particularly when rarer variants are functional and causal. The weights are based upon the frequency of the variants in the controls. The WSS method can be applied using a semi-permutation or full permutation algorithm. The full permutation option can be triggered using `--permutation,` with the default being semi-permutation. Although the full permutation version provides a more accurate estimate of power it is computationally intensive.

- VT. The VT method maximizes the test statistic over all possible MAF frequency cutoffs in the range specified by `--maflower/--mafupper`, and will correct for multiple testing within a permutation framework.
- SKAT. Due to the computational intensity of the SKAT method, p-values are obtained analytically.

In the last section of this supplemental material we provide a discussion on the choice of various association methods for power estimates. For now we will complete this example by showing the command to run five tests simultaneously and output the analysis results:

```
spower  LOGIT  Kryukov2009European1800.sfs  --sample_size  1000  \
--OR_rare_detrimental  1.5    \
--method  "CFisher  --alternative  1 --name CMC" "KBAC  --permutations  1000  --alternative  1" \
"WSSRankTest  --alternative  1 –name WSS" "VTtest  --alternative  1  --permutations  1000" "SKAT  disease"  \
-r 1000  -j 4  -l 1  -o  exercise
```

```
INFO:  Loading  data  from  [Kryukov2009European1800.sfs]  ...
INFO:  200  units  found
INFO:  1  units  will  be  analyzed
R1  : 100%  |===================================================================================|  Time:  0:22:17
INFO:  Tuning  [exercise.SEQPowerDB]  ...
INFO:  Result  saved to  [exercise.csv/.loci.csv/.SEQPowerDB]
```

```
spower show exercise.csv method power
```

```
+-------------+-------+
|    method   | power |
+-------------+-------+
|     WSS    | 0.374 |
|    VTtest   | 0.318 |
|     CMC     | 0.316 |
|     KBAC    | 0.352 |
|     SKAT    | 0.094 |
+-------------+-------+
```

Analytical power and sample size estimation

Command interface to perform analytical power analysis differs from empirical power analysis only in the `--method` argument: without using this argument, analytical power calculation will be performed using CMC.

```
spower LOGIT Kryukov2009European1800.sfs --sample_size 1000 --OR_rare_detrimental  1.5  -j 4 –l 1 –o exercise
```

If the `--power` option is used instead of `--sample_size`, analytical sample size estimation will be performed. For example to estimate sample size required to achieve power of 80%,

```
spower LOGIT Kryukov2009European1800.sfs –power 0.8 --OR_rare_detrimental  1.5  -j 4 –l 1 –o exercise
```

Example 2: Power Analysis for Quantitative Traits

Quantitative traits are generated under a linear model (`LNR` command in SEQPower). Simulation and analysis of quantitative traits share some options with the previous example for case-control data. This example will focus on its unique options for quantitative traits. Take the following command for example:

```
spower LNR Kryukov2009European1800.sfs --sample_size  1000 \
--meanshift_rare_detrimental 0.2 \
--method  "CollapseQt  --name CMC --alternative  2" \
-r 1000 -j 4 -l 1 -o exercise
```

Notice the new options

- `--meanshift_rare_detrimental` The genetic effect is now modeled by mean-shift in phenotype value due to the genetic factor.
- `--method CollapseQt` This is the quantitative trait version of the CMC method implemented as a linear regression score statistic with regressors being the collapsed genotype scores. We use a two-sided test here, a fair assumption for quantitative traits, since it is not known *a priori* if associated variants will increase or decrease quantitative trait values, although in the simulation only variants which increase quantitative trait values are modeled.

```
+-------+-----------------+
| power |    power_std    |
+-------+-----------------+
| 0.228 | 0.0132671021704  |
+-------+-----------------+
```

Variable effect sizes model

Variable effect sizes model can also be applied to the generation of quantitative trait data, for example:

```
spower LNR Kryukov2009European1800.sfs --sample_size  1000 \
--meanshift_rare_detrimental 0.2 --meanshiftmax_rare_detrimental 0.5 \
--method  "CollapseQt  --alternative  2" \
-r 1000 -j 4 -l 1 -o exercise
```

```
+-------+-----------------+
| power |    power_std    |
+-------+-----------------+
| 0.541 | 0.0157581407533  |
+-------+-----------------+
```

Example 3: Extreme Quantitative Traits Design

Samples with extreme quantitative traits can be obtained in two ways:

- From existing cohorts individuals with extreme quantitative traits are sampled and analyzed.
- Individuals with quantitative trait values above or below a certain value are sampled from the population and analyzed.

The default extreme quantitative trait model (`ELNR` command option in SEQPower) implements the first theme, with samples obtained from existing cohorts:

```
spower ELNR Kryukov2009European1800.sfs --sample_size 1000 \
--meanshift_rare_detrimental 0.2 --QT_thresholds 0.4 0.6 \
--method "CollapseQt --alternative 2" \
-r 1000 -j 4 -l 1 -o exercise
```

For the power analysis of quantitative traits there are additional options, some of which are shown below:

- `--QT_thresholds` Lower and upper cutoffs for extreme quantitative traits. For this power calculation, a sample of size 1000 is generated but only those individuals with QT values less than 40th percentile or greater than 60th percentile will be analyzed. Therefore a sample size of 800 individuals is used in the power analysis. This is a model for sampling the extremes of quantitative traits from a finite population.

```
spower show exercise.csv sample* power
```

```
+-------+-----------------+----------------------+
| power |    power_std    | sample_size_analyzed |
+-------+-----------------+----------------------+
| 0.207 | 0.0128121426779 |        800.0         |
+-------+-----------------+----------------------+
```

The ELNR with `--p1` option implements the second theme, with samples obtained from an infinite population:

```
spower ELNR Kryukov2009European1800.sfs --sample_size 1000 --p1 0.5 \
--meanshift_rare_detrimental 0.5 --QT_thresholds 0.4 0.6 \
--method "CollapseQt --alternative 2" \
-r 1000 -j 4 -l 1 -o exercise
```

- `--p1` With this option, quantitative trait values are generated conditional on the simulated genotype data using a cumulative density function for the standard normal distribution and those individuals with simulated trait values in between $\Phi^{-1}(0.4)$ and $\Phi^{-1}(0.6)$ (Φ is the cumulative density function for standard normal distribution) are sampled, as though they come from an infinite population. Sampling continues until a sample size of 1,000 individuals is obtained.

Power analysis result is as follows:

```
+-------+-----------------+----------------------+
| power |    power_std    | sample_size_analyzed |
+-------+-----------------+----------------------+
| 0.877 | 0.0103860964756 |       1000.0         |
+-------+-----------------+----------------------+
```

Example 4: Simulation-only mode for case-control and quantitative trait data

It is sometimes desirable to simulate datasets for purposes other than power and sample size calculations. The `GroupWrite` method in SEQPower will output data into bundles containing the following files:

- *Genotype file* One variant per row
  first column: variant id; subsequent columns: sample haplotypes at each variant site.
- *Phenotype file* One subject per row
  first column: subject id; second column: quantitative / binary phenotypes.

Phenotypes are generated conditional on the variant data, but the data are saved to files instead of analyzing the generated replicates to perform power analysis. Please note variant data is not simulated but instead the pre-generated haplotype pools are used.

```
spower LOGIT Kryukov2009European1800.sfs --sample_size 1000 --OR_rare_detrimental 1.5 \
--method "GroupWrite ExerciseSimulation" -j 4 -o exercise -v1
```

```
INFO: Loading data from [Kryukov2009European1800.sfs] ...
INFO: 200 units found
scanning: unit 200 - 100% |>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>| Time: 0:00:53
$ ls ExerciseSimulation
 R100_geno.txt   R130_mapping.txt R160_pheno.txt   R191_geno.txt    R3_mapping.txt  R6_pheno.txt
 R100_mapping.txtR130_pheno.txt   R161_geno.txt    R191_mapping.txt R3_pheno.txt    R70_geno.txt
 R100_pheno.txt  R131_geno.txt    R161_mapping.txt R191_pheno.txt   R40_geno.txt    R70_mapping.txt
 R101_geno.txt   R131_mapping.txt R161_pheno.txt   R192_geno.txt    R40_mapping.txt R70_pheno.txt
 R101_mapping.txtR131_pheno.txt   R162_geno.txt    R192_mapping.txt R40_pheno.txt   R71_geno.txt
 R101_pheno.txt  R132_geno.txt    R162_mapping.txt R192_pheno.txt   R41_geno.txt    R71_mapping.txt
 R102_geno.txt   R132_mapping.txt R162_pheno.txt   R193_geno.txt    R41_mapping.txt R71_pheno.txt
 ...
```

Example 5: Power Calculation for a Range of Input Parameters

Browse result output from multiple SEQPower commands

So far we have covered a number of SEQPower examples. With output option `-o exercise` in action for all examples, we have saved all results from every command in this tutorial to a database named `exercise.SEQPowerDB`.

The database can be browsed using the command `spower show`.

```
spower show exercise.SEQPowerDB
spower show exercise.SEQPowerDB LOGIT
```

To select a power analysis result of interest, for example

```
spower show exercise.SEQPowerDB LOGIT method power title --condition "where power between 0.25 and 0.95"
```

```
+---------+-------+
|  method | power |
+---------+-------+
|   CMC   | 0.289 |
|   CMC   | 0.31  |
|    K1   | 0.347 |
|   CMC   | 0.772 |
|   CMC   | 0.795 |
+---------+-------+
```

Although results of interest are selected from SEQPower database, the output is confusing since we cannot determine from the output under which scenarios the power estimates are obtained. To resolve this problem we need to assign a unique ID for each power calculation task when running them in batches, as will be demonstrated in the next section.

Run SEQPower commands in shell with unique IDs

We utilize simple programming in shell languages to run power calculation in batch mode, assigning unique identifiers to each calculation (the `--title` option), and extract output from the result database. We now save everything to a new database called `exercise2.SEQPowerDB`. Below is an example under the Linux `Bash` shell:

```
for i in 1 1.5 2 2.5 3 3.5 4; do
spower LOGIT Kryukov2009European1800.sfs --sample_size 1000 --OR_rare_detrimental $i \
--method "CFisher --name CMC$i" --title FixedOR$i \
-r 1000 -j 4 -l 1 -o exercise2
done
```

A range of fixed odds ratios from 1 to 4 is evaluated. To view the output:

```
spower show exercise2.SEQPowerDB  LOGIT method power title
```

```
+--------+-------+------------+
| method | power |   title    |
+--------+-------+------------+
|  CMC1  |  0.05 |  FixedOR1  |
| CMC1.5 |  0.28 | FixedOR1.5 |
|  CMC2  |  0.74 |  FixedOR2  |
| CMC2.5 |  0.84 | FixedOR2.5 |
|  CMC3  |  0.9  |  FixedOR3  |
| CMC3.5 |   1   | FixedOR3.5 |
|  CMC4  |   1   |  FixedOR4  |
+--------+-------+------------+
```

Discussion

Choice of association methods

There are many rare variant association methods available in SEQPower. Since both the relative and absolute power of each rare variant association method depend heavily on the assumptions made in simulations, there is no one method that outperforms other methods in a variety of circumstances, *i,e.*

there is no single “best” method. However under some general settings, we can make recommendations on choice of association methods with consideration of the power and sample size. For example, the inclusion of non-causal variants (`-P` option in SEQPower) and exclusion of causal variants (`--missing_sites` option) will drastically reduce power for all methods. Methods involving weighting are sometimes more powerful than plain counting methods in the presence of non-causal variants. Methods running multiple comparisons such as VT and RareCover may seem less powerful than other “fixed variant set” methods if the underlying causal rare variants MAF is exactly the MAF range being analyzed. In real world data analyses it is also often observed that these methods are less powerful, as the advantage of maximization over candidate sets is often too small to cancel out the penalty required for multiple comparison. Finally, when deleterious variants are predominant in the genetic region of interest, methods analyzing both protective and deleterious variants such as SKAT and C-alpha are usually underpowered compared to other methods which focus on testing for unidirectional effects. When testing for unidirectional effects, the one-sided versions of association methods (if applicable) are usually more powerful than the two-sided ones.

Table S1 SEQPower Rare Variant Association Methods

Method	Feature	Phenotype	p-value	Power	SEQPower Command	Reference
aSum	Two-stage test for bi-directional effects	Binary	Empirical	Empirical	`aSum`	(Han and Pan, 2010)
c-alpha	Variance component test for bi-directional effects	Binary	Empirical / Analytical	Empirical	`cAlpha`	(Neale *et al.*, 2011)
BRV [1]	Variant counting method	Binary / Quantitative	Empirical / Analytical	Empirical	`BurdenQt / BurdenBt`	(Auer *et al.*, 2013)
CMC	Variant collapsing method	Binary / Quantitative	Empirical / Analytical	Empirical / Analytical [2]	`CFisher / CollapseQt / CollapseBt`	(Li and Leal, 2008)
KBAC	Genotype weighted counting method	Binary / Quantitative	Empirical	Empirical	`KBAC`	(Liu and Leal, 2010)
RareCover	Maximization approach over variants	Binary	Empirical	Empirical	`RareCover`	(Bhatia *et al.*, 2010)
RBT	Weighted counting method for bi-directional effects	Binary / Quantitative	Empirical	Empirical	`RBT`	(Ionita-Laza *et al.*, 2011)
SKAT	Weighted variance component test for bi-directional effects	Binary / Quantitative	Empirical / Analytical	Empirical	`SKAT disease / SKAT quantitative`	(Wu *et al.*, 2011), (Lee *et al.*, 2012)
WSS	Variant weighted counting method	Binary / Quantitative	Empirical	Empirical	`WSSRankTest / WeightedBurdenQt / WeightedBurdenBt`	(Madsen and Browning, 2009)
VT	Maximization approach-over allele frequencies	Binary / Quantitative	Empirical	Empirical	`VTtest / VariableThresholdsQt`	(Price *et al.*, 2010)

aSum: Adaptive Sum test; BRV: Burden of Rare Variants; CMC: Combined Multivariate and Collapsing; KBAC: Kernel Based Adaptive Clustering; RBT: Replication Based Test; SKAT: Sequencing Kernel Association Test; WSS: Weighted Sum Statistic; VT: Variable Threshold test.

[1] BRV and GRANVIL (Morris and Zeggini, 2010) are roughly equivalent but BRV does not have a denominator with the number of variant sites which is not robust to type I error when there are missing data. [2] Analytical sample size calculation can also be performed.

Table S2: SEQPower Command Options

Short syntax	Long syntax	Default	Description
Power and sample size calculation options:			
-	--moi	A	mode of inheritance: "A", additive (default); "D", dominant; "R", recessive; "M", multiplicative (does not apply to quantitative traits model)
-	--resampling	FALSE	directly draw sample genotypes from given haplotype pools (sample genotypes will be simulated on the fly if haplotype pools are not available)
-	--def_rare	0.01	definition of rare variants: variant having "MAF <= frequency" will be considered a "rare" variant; the opposite set is considered "common"
-	--def_neutral	None	annotation value cut-offs that defines a variant to be "neutral" (e.g. synonymous, non-coding etc. that will not contribute to any phenotype); any variant with "function_score" X falling in this range will be considered neutral
-	--def_protective	None	annotation value cut-offs that defines a variant to be "protective" (i.e., decrease disease risk or decrease quantitative traits value); any variant with "function_score" X falling in this range will be considered protective
-P	--proportion_detrimental	None	proportion of deleterious variants associated with the trait of interest, i.e., the random set of the rest (1 - p) x 100% deleterious variants are non-causal: they do not contribute to the phenotype in simulations yet will present as noise in analysis
-Q	--proportion_protective	None	proportion of protective variants associated with the trait of interest, i.e., the random set of the rest (1 - p) x 100% protective variants are non-causal: they do not contribute to the phenotype in simulations yet will present as noise in analysis
-	--sample_size	None	total sample size
-	--p1	None	proportion of affected individuals , or individuals with high extreme QT values sampled from infinite population (default set to None, meaning to sample from finite population specified by --sample_size option).
-	--def_valid_locus	None	upper and lower bounds of variant counts that defines if a locus is "valid", i.e., locus having number of variants falling out of this range will be ignored from power calculation
-	--rare_only	FALSE	remove from analysis common variant sites in the population, i.e., those in the haplotype pool having MAF > $def_rare
-	--missing_as_wt	FALSE	label missing genotype calls as wild type genotypes
-	--missing_low_maf	None	variant sites having population MAF < P are set to missing
-	--missing_sites	None	proportion of missing variant sites
-	--missing_sites_deleterious	None	proportion of missing deleterious sites
-	--missing_sites_protective	None	proportion of missing protective sites
-	--missing_sites_neutral	None	proportion of missing neutral sites
-	--missing_sites_synonymous	None	proportion of missing synonymous sites
-	--missing_calls	None	proportion of missing genotype calls
-	--missing_calls_deleterious	None	proportion of missing genotype calls at deleterious sites
-	--missing_calls_protective	None	proportion of missing genotype calls at protective sites
-	--missing_calls_neutral	None	proportion of missing genotype calls at neutral sites
-	--missing_calls_synonymous	None	proportion of missing genotype calls at synonymous sites
-	--error_calls	None	proportion of error genotype calls
-	--error_calls_deleterious	None	proportion of error genotype calls at deleterious sites
-	--error_calls_protective	None	proportion of error genotype calls at protective sites
-	--error_calls_neutral	None	proportion of error genotype calls at neutral sites
-	--error_calls_synonymous	None	proportion of error genotype calls at synonymous sites
-	--power	None	power for which total sample size is calculated (this option is mutually exclusive with option '--sample_size')
-r	--replicates	1	number of replicates for power evaluation
-	--alpha	0.05	significance level at which power will be evaluated

-l	--limit	None	if specified, will limit calculations to the first N groups in data .
-o	--output	None	output filename
-t	--title	None	unique identifier of a single command run
-v	--verbosity	2	verbosity level: 0 for absolutely quiet, 1 for less verbose, 2 for verbose, 3 for more debug information
-s	--seed	0	seed for random number generator, 0 for random seed
-j	--jobs	2	number of CPUs to use when multiple replicates are required via "-r" option .
-m	--methods	None	Method of one or more association tests. Parameters for each method should be specified together as a quoted long argument (e.g. --methods "m --alternative 2" "m1 --permute 1000"), although the common method parameters can be specified separately, as long as they do not conflict with command arguments. (e.g. --methods m1 m2 -p 1000 is equivalent to --methods "m1 -p 1000" "m2 -p 1000".). You can use command 'spower show tests' for a list of association tests, and 'spower show test TST' for details about a test.
-	--discard_samples	None	Discard samples that match specified conditions within each test group. Currently only expressions in the form of "%(NA)>p" is provided to remove samples that have more 100*p percent of missing values.
-	--discard_variants	None	Discard variant sites based on specified conditions within each test group. Currently only expressions in the form of '%(NA)>p' is provided to remove variant sites that have more than 100*p percent of missing genotypes. Note that this filter will be applied after "--discard_samples" is applied, if the latter also is specified.
LOGIT ***command options for binary phenotype simulation under LOGIT model:***			
-a	--OR_rare_detrimental	1	odds ratio for detrimental rare variants
-b	--OR_rare_protective	1	odds ratio for protective rare variants
-A	--ORmax_rare_detrimental	None	maximum odds ratio for detrimental rare variants, applicable to variable effects model
-B	--ORmin_rare_protective	None	minimum odds ratio for protective rare variants, applicable to variable effects model
-c	--OR_common_detrimental	1	odds ratio for detrimental common variants
-d	--OR_common_protective	1	odds ratio for protective common variants
-f	--baseline_effect	0.01	penetrance of wild type genotypes
PAR ***command options for binary phenotype simulation under population attributable risk model:***			
-a	--PAR_rare_detrimental	0	Population attributable risk for detrimental rare variants
-b	--PAR_rare_protective	0	Population attributable risk for protective rare variants
-c	--PAR_common_detrimental	0	Population attributable risk for detrimental common variants
-d	--PAR_common_protective	0	Population attributable risk for protective common variants
-	--PAR_variable	FALSE	use variable population attributable risks: the smaller the MAF the larger the PAR
-f	--baseline_effect	0.01	penetrance of wild type genotypes
LNR ***command options for quantitative phenotype simulation under linear model:***			
-a	--meanshift_rare_detrimental	0	mean shift in quantitative value w.r.t standard deviation due to detrimental rare variants i.e., by "MULTIPLIER * sigma"
-b	--meanshift_rare_protective	0	mean shift in quantitative value w.r.t. standard deviation due to protective rare variants i.e., by "MULTIPLIER * sigma"
-A	--meanshiftmax_rare_detrimental	None	maximum mean shift in quantitative value w.r.t standard deviation due to detrimental rare variants i.e., by "MULTIPLIER * sigma", applicable to variable effects model
-B	--meanshiftmax_rare_protective	None	maximum mean shift in quantitative value w.r.t standard deviation due to protective rare variants i.e., by "MULTIPLIER * sigma", applicable to variable effects model
-c	--meanshift_common_detrimental	0	mean shift in quantitative value w.r.t standard deviation due to detrimental common variants i.e., by "MULTIPLIER * sigma"
-d	--meanshift_common_protective	0	mean shift in quantitative value w.r.t standard deviation due to protective common variants i.e., by "MULTIPLIER * sigma"
BLNR and ELNR ***command options for binary or extreme quantitative phenotype simulation under linear model:***			
-a	--meanshift_rare_detrimental	0	mean shift in quantitative value w.r.t standard deviation due to detrimental rare variants i.e., by "MULTIPLIER * sigma"
-b	--meanshift_rare_protective	0	mean shift in quantitative value w.r.t. standard deviation due to protective rare variants i.e., by "MULTIPLIER * sigma"

-A	--meanshiftmax_rare_detrimental	None	maximum mean shift in quantitative value w.r.t standard deviation due to detrimental rare variants i.e., by "MULTIPLIER * sigma", applicable to variable effects model
-B	--meanshiftmax_rare_protective	None	maximum mean shift in quantitative value w.r.t standard deviation due to protective rare variants i.e., by "MULTIPLIER * sigma", applicable to variable effects model
-c	--meanshift_common_detrimental	0	mean shift in quantitative value w.r.t standard deviation due to detrimental common variants i.e., by "MULTIPLIER * sigma"
-d	--meanshift_common_protective	0	mean shift in quantitative value w.r.t standard deviation due to protective common variants i.e., by "MULTIPLIER * sigma"
-	--QT_thresholds	[0.5, 0.5]	lower/upper percentile cutoffs for quantitative traits in extreme QT sampling
The `show` command options:			
Association test name, power analysis result text file name or SEQPower database file name		None	type of information to display, which can be 'tests' for a list of all association tests, 'test TST' for details of an association test TST, 'FILENAME.csv' for all column names in a csv file, 'FILENAME.csv [colnames]' for values of columns in a csv file; 'FILENAME.SEQPowerDB' for all table names in a SEQPower database file, 'FILENAME.SEQPowerDB TABLE' for all column names in a table, 'FILENAME.SEQPowerDB TABLE [colnames]' for values of specified columns in a table, and 'FILENAME.SEQPowerDB TABLE [colnames] --condition QUERY' for filtered/formatted values of columns in a table. Wildcard symbol '*' for colnames is allowed.
The `execute` command options:			
-s	--sliding	None	specify variable parameters
-f	--fixed	None	specify fixed parameters
-	--plot	FALSE	generate plot instead of running simulations
-	--dry_run	FALSE	print generated commands to screen instead of executing them

References

Auer,P.L. *et al.* (2013) Testing for rare variant associations in the presence of missing data. *Genet. Epidemiol.*, **37**, 529–538.

Bhatia,G. *et al.* (2010) A Covering Method for Detecting Genetic Associations between Rare Variants and Common Phenotypes. *PLoS Comput Biol*, **6**, e1000954.

Boyko,A.R. *et al.* (2008) Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLoS Genet*, **4**, e1000083.

Han,F. and Pan,W. (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.*, **70**, 42–54.

Ionita-Laza,I. *et al.* (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.*, **7**, e1001289.

Kryukov,G.V. *et al.* (2009) Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl. Acad. Sci.*, **106**, 3871 –3876.

Lee,S. *et al.* (2012) Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *Am. J. Hum. Genet.*, **91**, 224–237.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Liu,D.J. and Leal,S.M. (2010) A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genet*, **6**, e1001156.

Madsen,B.E. and Browning,S.R. (2009) A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet*, **5**, e1000384.

Morris,A.P. and Zeggini,E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, **34**, 188–193.

Neale,B.M. *et al.* (2011) Testing for an Unusual Distribution of Rare Variants. *PLoS Genet*, **7**, e1001322.

Price,A.L. *et al.* (2010) Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am. J. Hum. Genet.*, **86**, 832–838.

Williamson,S.H. *et al.* (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 7882 –7887.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.