SimPEL User's Manual r1.0.0

December 7th, 2017

SimPEL is short for **Sim**ulation-based **P**ower **E**stimation for sequencing studies of **L**ow-prevalence conditions. SimPEL addresses the need for power estimation in low-prevalence condition studies, taking into account their inherently small sample sizes and analytical procedures. SimPEL integrates customizable parameters to realistically model study design outcomes and provide applicable information towards further refinement of experimental procedure. SimPEL is implemented as a function of the established JAWAMix5 tool (Long et al., 2013), an HDF5-based Java implementation for association mapping.

Installation

A tarball will be available for download at https://owncloud.westgrid.ca/index.php/s/s4d86JjmRXycLKZ. The archive will include all mandatory input files. MD5 hash values for each of the files are also included in the download. As Java is platform independent and the software comes "batteries-included," there is no specific installation required. The JAWAMix5 jar file is ready for immediate use, provided that Java has been installed in the system. Through this, users can easily verify that they indeed have the correct files and that no errors have occurred during the download and extraction process.

Please note that the files in the tarball are for user convenience. This allows users to start testing the program without the burden of locating and downloading multiple files from different websites. Once users have confirmed that they are able to run the program without any issues, it is strongly encouraged to download up to date files from the original websites and cite the updated publications for each file. Should users choose to use the included files in the tarball, please cite SimPEL in the publication as well as the following papers for the tarball files:

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., … Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, *48*(12), 1581–1586. https://doi.org/10.1038/ng.3703

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Usage

All parameters are implemented as:

java -Xmx4g -jar /filepath/jawamix5.jar simpel <parameters>.

An example command line is as follows:

java -Xmx4g -jar jawamix5.jar simpel -population_genotypes g1k_all.hdf5 -out example_output.txt -causal_gene_pool HLA_ErbB.txt -population_pedigree integrated_call_samples_v2.20130502.ALL.ped -all_genes gencode.v12.genenames.gtf -mafs ExAC.r0.3.1.sites.AC.txt -pathogenicity mcap_v1_0.txt -tmp_folder tmp_chrs/ -num_cases 10 -parents 8 -compound_het

In this example, all the files are present in current working directory. The number of case-control pairings is designated as 10, of which 8 of the controls are parents of the cases. Instances of compound heterozygosity will also be considered by the program, as indicated by the inclusion of the -compound_het flag.

For all file names, the file path must be specified if the file in question is not present in the current working directory.

Important Note to Users

While SimPEL allows users to supply their own files in place of the default input files, it is important that the format of the files are identical to those of the default input files. Another very important thing to note is that the files used must all align to the same version of the human genome. All default SimPEL files are based on **Build 37;** therefore if any of the default files are used, all custom files must also align to Build 37. (Alternatively, the users may want to upgrade to Build 38. Should this be the case, users must ensure that all files comply to the same coordinate system.)

Mandatory Parameters

- -population_genotypes <hdf5 file>

The input HDF5 file for population genotypes utilized by the simulation process. For default usage, the whole genomes provided by the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) are recommended. The compressed CSV file has been included within the SimPEL package. Through a single "import" command, one can easily convert the CSV file to an HDF5 file (please refer to the JAWAMix5 Users Manual). The HDF5-based data format allows the users to store large genotype files in the disk while accessing them as though they are in the memory. The runtime of disk-based solution is similar to approaches using memory (Long et al., 2013). This ensures that the software is not memory-intensive and that the programs run at a rapid pace when working with large files.

File format: The first row is a header specifying the IDs of the individuals. Each subsequent row represents a the presence or absence of a genetic variant. The columns are separated by commas. The first column represents the chromosome number (integer); the second column represents the location coordinate. All following columns represents the genotype calls for each individual at each variant. 0, 1, and 2 are each assumed to represent homozygous reference allele, heterozygous allele, and homozygous mutational allele(s), respectively.

- -out <txt file>

The output text file. Input a file path and the file name and the program will generate the output file accordingly. The output will show the simulated causal gene and mutations within the samples as well as the rank of the simulated causal gene in the rank success list for each round of simulation. The overall power of the simulated study is displayed at the end of the file.

- -causal_gene_pool <txt file>

The input text file for the causal gene pool. By default, a list of genes is expected and a text file containing gene pathogenicity scores is used in conjunction. Should the specific casual gene pool be unknown, a list of all human genes can be used. SimPEL also offers the ability to specify different inheritance models. This can be done by supplying a list of genes reported by existing studies. To this end, six files containing genes associated with six

different disease models have been included with the SimPEL package. For example, if the condition in question is assumed to be autosomal-recessive, the Autosomal Recessive Genes file included within the package can be used as the input for this flag.

File format: Each row represents a separate gene. Sequentially, the columns represent the gene name, the chromosome it is located on, the start position, and the end position. The columns are separated by tab spaces.

- -population_pedigree <ped file>

The input pedigree file for pool of samples to draw from for simulations. This file is included as part of the SimPEL package and serves as the pedigree information for the genomes from the 1000 Genomes Project. The two files are used in conjunction to form a pool of samples to draw from when generating cases/controls for simulation rounds.

File format: The format follows the 1000 Genome Project sample information PED file. SimPEL only reads the following information: sample ID, gender, ethnic groups, parental IDs, and sibling IDs. Should a user wish to swap in and use a custom PED file in the place of the default file, please ensure that the information used by the program are in the correct columns.

- -all_genes <gtf file>

The input GTF file for all human genes. This file is included as part of the SimPEL package and serves as a compiled list of all genes within the human genome. It is recommended that users use the file included with the package and not substitute it with a separate file.

- -mafs <txt file>

The input text file for minor allele counts in healthy populations. By default, we have included data from the Exome Aggregation Consortium (ExAC) as part of the SimPEL package (Lek et al., 2016). For more information, please refer to the ExAC flagship publication or visit the ExAC website.

File format: The first row is a header specifying the information provided by each column. The columns, in order represent the chromosome, the position, the reference base, the altered base, and the allele count. Every subsequent row represents a minor allele in healthy populations.

- -pathogenicity <txt file>

The input text file for gene pathogenicity scores. By default, a gene pathogenicity file scored as per the M-CAP guidelines (Jagadeesh et al., 2016) is included as part of the SimPEL package and provides applicable pathogenicity scores for rare missense variants in the human genome. Pathogenicity score cutoffs can also be specified as an optional parameter, details below.

File format: The first row is a header specifying the information provided by each column. The columns, in sequential order, represent the chromosome the variant is located, the position, the reference base, the altered base, and its M-CAP pathogenicity score. Every subsequent row represents a different variant. Please note that if the user wishes to use another tool to produce pathogenicity scores, the input file must follow this format and the score cutoff below must be adjusted accordingly.

- -tmp_folder <folder filepath>

The input file path to a folder to store temporary files generated by the tool. The folder name does need to be specified but does not need to be created beforehand as SimPEL will generate the folder if it is not currently present.

Optional Parameters

Some of the assignable parameters below are used in calculations. Please refer to the **Supplementary Materials and Methods** for complete descriptions of their roles within the simulation and analytical framework. Should the following parameters be not specified or not included, default values will be used.

- -num_cases <integer>

Sets the sample size used for simulations. Default value is 4, indicating 4 controls and 4 cases.

- -sim_rounds <integer>

Sets the number of simulations to be performed. Default value is 100.

- -rank_success <integer>

Sets the size of the list of top ranking candidates based on the simulation score. Indicates a successful study if the causal gene appears on this list. Default value is 10. This option reflects the false positive rates of the study. If the user believes the sample size is insufficient to guarantee a satisfactory power, this number may be increased to obtain a higher power at the cost of increasing the false positive rate. The expectation is that the investigator will discover a large number of genes for experimental follow-up upon completion of the proposed study, and will be able to prioritize based on existing biological insight.

- -num_causal_genes <integer>

Sets the number of causal genes for the condition in the simulations. Default value is 1. This value reflects the heterogeneity of the condition. Generally for small sample sizes, it is required to assume that there is only one causal gene for the condition to ensure acceptable power. Should there be sufficient number of samples (i.e. ~10 with prior knowledge on the potential causal gene pool), multiple causal genes can be considered.

- -prevalence <double>

Sets the prevalence of the low-prevalence condition in the population in the simulations. Default value is 0.01.

- -penetrance <double>

Sets the penetrance of the low-prevalence condition in the simulations. Default value is 0.9.

- -min_score <double>

Sets the pathogenicity score cutoff for the provided list of causal genes. Any scores falling below the cutoff will not contribute to the causal ranking. Default value is 0.025, as recommended by the M-CAP guidelines.

- -conf_score <double>

Sets the confidence of pathogenicity scores in the simulations. Default value is 0.95. This value reflects the confidence in the correctness of the pathogenicity scores. Please refer to the **Supplementary Materials and Methods** for a description of the confidence score's effect on power estimations.

- -weight _score <double>

Sets the weight multiple for pathogenicity for each individual case in the simulations. Default value is 1.0. 0 means that the pathogenicity score is ignored.

- -pop_freq <double>

Sets the population frequency cutoff. Variants with a higher population frequency than this score will not be deemed as functional. Default value is 0.001.

- -parents <integer>

Sets the number of generated controls within the simulations as parents of the cases. The sum of parents and siblings must not be greater than the designated sample size. Any controls not parents are siblings are deemed unrelated to the cases. Default value is 0.

- -siblings <integer>

Sets the number of generated controls within the simulations as siblings of the cases. The sum of parents and siblings must not be greater than the designated sample size. Any controls not parents are siblings are deemed unrelated to the cases. Default value is 0.

- -include_all_genes

Inclusion of this flag indicates that any human gene is considered in association mapping, not just those within the input causal gene pool file.

- -compound_het

Upon the inclusion of this flag, the simulation will generate two random heterozygous mutations in the same gene. The analysis will be adjusted accordingly to account for the likelihood of prioritizing one or two of the case-specific functional mutations in each of the genes.

Contacts

Quan Long, quan.long@ucalgary.ca

Copyright License (MIT Open Source)

Acknowledgements

Funding. This work is supported by the Rare Disease Foundation, University of Calgary Startup and URGC Seed Grants, Canada Foundation for Innovation, NSERC Discovery Grant (Q.L.) and Alberta Children's Hospital Research Institute graduate student fellowship (L.M.).

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., … Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, *48*(12), 1581–1586. https://doi.org/10.1038/ng.3703

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Long, Q., Zhang, Q., Vilhjalmsson, B. J., Forai, P., Seren, Ü., & Nordborg, M. (2013). JAWAMix5: an out-of-core HDF5-based java implementation of whole-genome association studies using mixed models. *Bioinformatics*, *29*(9), 1220–1222. https://doi.org/10.1093/bioinformatics/btt122