RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Reproducible Simulations of Realistic Samples for Next-Generation Sequencing Studies Using *Variant Simulation Tools*

Bo Peng*

Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Received 16 June 2014; Revised 14 September 2014; accepted revised manuscript 26 September 2014.
Published online 13 November 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21867

ABSTRACT: Computer simulations have been widely used to validate and evaluate the power of statistical methods for genetic epidemiological studies. Although a large number of simulation methods and software packages have been developed for genome-wide association studies, methodological and bioinformatics challenges have limited their applications in simulating datasets for whole-genome and whole-exome sequencing studies. With the development of more sophisticated statistical methods that make fuller use of available data and our knowledge of the human genome, there is a pressing need for genetic simulators that capture more features of empirical data (e.g., multiallele variants, indels, use of the Variant Call Format) and the human genome (e.g., functional annotations of genetic variants). This article introduces *Variant Simulation Tools* (VST), a module of *Variant Tools* for the simulation of genetic variants for sequencing-based genetic epidemiological studies. Although multiple simulation engines are provided, the core of VST is a novel forward-time simulation engine that simulates real nucleotide sequences of the human genome using DNA mutation models, fine-scale recombination maps, and a selection model based on amino acid changes of translated protein sequences. The design of VST allows users to easily create and distribute simulation methods and simulated datasets for a variety of applications and encourages fair comparison between statistical methods through the use of existing or reproduced simulated datasets.
Genet Epidemiol 39:45–52, 2015. © 2014 Wiley Periodicals, Inc.

KEY WORDS: simulation; genetic variants; next-gen sequencing; rare variant association analysis; reproducible study

Introduction

A large number of computer programs have been developed to simulate pseudo-datasets with known phenotype-genotype associations for the development and applications of statistical methods for genetic epidemiology studies [Hoban et al., 2011; Liu et al., 2008; Peng et al., 2014; Ritchie and Bush, 2010]. These programs have aided in the validation of statistical methods by simulating datasets under model assumptions, and in the evaluation of the performance of these methods in real-world applications by simulating "realistic" datasets that capture more of the complexity of the human genomes and related phenotypes. With continuing advances in genotyping technology, maturation of statistical methods, and our deeper understanding of the genetic causes of human traits and diseases, genetic simulations have become increasingly complex, simulating scenarios from sparse markers for linkage studies [Leal et al., 2005], to high-density common single-nucleotide polymorphisms for genome-wide association (GWA) studies [Li and Li, 2008], to sequences with rare genetic variants for next-generation sequencing analyses [Peng and Liu, 2010].

Because rare genetic variants may be important to the genetic etiology of human diseases [Dickson et al., 2010; Gorlov et al., 2011; Gorlova et al., 2012], and because the effects of single rare variants are too small to be detected using GWA-based approaches, a large number of statistical methods have been developed to detect the combined effect of rare genetic variants in the same exon, gene, or pathway [Cheung et al., 2012; Li and Leal, 2008; Wu et al., 2011; Yi et al., 2011]. A number of simulation methods have been used to simulate data for these studies. For example, a widely used theoretical method simulates genotypes with alleles drawn from a binomial distribution $B(2N, p_i)$, where N is the number of diploid samples and p_i is population allele frequency sampled either from a specified distribution [Jiang and McPeek, 2014] or from a frequency spectrum estimated from empirical data [Auer et al., 2013; Logsdon et al., 2014]. Extensions of this method can model correlations between closely linked loci [Montana, 2005] using linkage disequilibrium (LD) patterns estimated from, for example, haplotypes from the 1000 Genomes Project [Saad and Wijsman, 2014]. Furthermore, resampling-based methods have been used to retain markers and long-range LD of empirical data by sampling haplotypes directly from such data [Wallace, 2013].

Because theoretical and resampling-based methods are limited in their ability to model the effect of historical genetic and demographic features on the genetic composition

*Correspondence to: Bo Peng, Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Unit 1401, Houston, TX 77030, USA. E-mail: bpeng@mdanderson.org

of human populations, coalescent and forward-time methods have been used to simulate haplotype sequences with rare variants. For example, Byrnes et al. [2013] and Lin et al. [2013] used cosi [Schaffner et al., 2005], a popular coalescent simulator, to simulate short haplotype sequences (approximately 1 M base pairs) of the human genome with realistic LD patterns. Cule and de Iorio [2013] used FREGENE [Chadeau-Hyam et al., 2008] to evolve haplotypes with common single-nucleotide polymorphism markers (minor allele frequency >1%) forward-in-time, and Cheung et al. [2012] used SRV [Peng and Liu, 2010] to simulate haplotypes with rare genetic variants. Although coalescent-based simulations are computationally efficient and can be used to simulate large haplotype pools, they are limited in their ability to simulate the effect of natural selection on the human genome, such as the differences in allele frequency spectra in coding and noncoding regions. In comparison, forward-time simulations can simulate haplotypes with more realistic patterns of allele frequency spectrum, at a cost of performance [Peng and Liu, 2010]. These methods are highly flexible and can be used to simulate datasets under complex genetic and demographic assumptions.

With the development of more sophisticated rare variant association methods that make fuller use of available data and our knowledge of the human genome, there is a pressing need for genetic simulators that capture more features of empirical data and the human genome. Features that are not well supported by existing simulators include, but are not limited to, the simulation of multiallele variants, indels, and support for the commonly used Variant Call Format (VCF). In addition, because a large proportion of rare genetic variants are likely to be neutral or only weakly deleterious [Boyko et al., 2008; Kryukov et al., 2007; Yampolsky et al., 2005], weighting schemes that explore the structure of genes and functional annotation of genetic variants of the human genome would be more effective than those that do not utilize such information [Byrnes et al., 2013]. Simulation methods that model distributions and functional effect of mutations in coding and noncoding regions are therefore needed.

We developed *Variant Simulation Tools* (VST), a simulator that simulates genetic samples for next-generation sequencing studies. Although multiple simulation engines are provided, the core of VST is a novel forward-time simulation engine that simulates real nucleotide sequences of the human genome using DNA mutation models, fine-scale recombination maps, and a selection model based on amino acid changes of translated protein sequences. The design of VST allows users to easily create and distribute simulation methods and simulated datasets for a variety of applications and to reproduce simulation studies performed by VST.

Methods

Design and User-Interface

VST is implemented and distributed as a module of *Variant Tools* [San Lucas et al., 2012]. It uses *Variant Tools* to manage, analyze, and export sample genotypes and uses its pipeline feature to perform simulations. It also uses *Variant Tools*' online repository to store and distribute simulation models and simulated datasets.

VST utilizes a two-tier design. The core of VST provides a set of functions that perform tasks such as extracting and importing genotypes from the 1000 Genomes Project [Abecasis et al., 2012], evolving a population forward in time, applying a disease model, drawing samples from simulated populations, and calling external programs to analyze simulated datasets. These functions follow a standard interface and can be "connected" to perform simulations of varying complexity. On top of the VST core, simulation models are described in simulation specification files, which essentially specify how to call VST functions to perform particular types of simulations. Whereas the core functions of VST are distributed with *Variant Tools* and are relatively stable, simulation specification files are stored separately at the *Variant Tools* resource repository. A new simulation model will be available to all VST users as soon as its specification is uploaded to the repository.

VST uses the command line interface of *Variant Tools* and is provided as one of its subcommands. A simulation is performed by command

```
% vtools simulate SPECFILE [MODEL] [—seed
    SEED] [options].
```

This command locates a local or online simulation specification file and executes one of the models with model-specific options. Option `MODEL` can be ignored if `SPECFILE` defines only one simulation model. VST by default uses a randomly selected seed for each simulation but a previous simulation could be repeated by specifying the seed of that simulation using parameter `—seed`. Similar to other components of *Variant Tools*, users can use command

```
% vtools show simulations
```

to get a list of all available simulation models, and command

```
% vtools show simulation SPECFILE
```

to learn the details of simulation models defined in SPECFILE. Output of the last command includes descriptions of the simulation method, output, typical applications, steps of the simulation, and model parameters. Advanced users who would like to create new simulation models or further customize existing models can write new simulation specifications or modify existing ones.

Core Functions and Simulation Engines

VST provides an increasing number of core functions for different types of simulations. Under the hood of a uniform interface and file formats, VST includes simulation engines for backward-time, forward-time, theoretical, and resampling-based simulations. The backward-time coalescent simulation engine is currently implemented by importing datasets simulated by ms [Hudson, 2002]. The theoretical simulation engine simulates genotypes either from random frequency drawn from a specified allele frequency spectrum or from allele frequencies of an empirical dataset. Resampling-based simulation is implemented as drawing from a large haplotype pool expanded from provided haplotypes, subject to mutation and recombination. The introduction of new mutants to a resampling-based simulation can be controlled by natural selection.

Of particular interest is a forward-time simulation engine with related recombination, mutation, selection, penetrance, and quantitative trait models based on simuPOP [Peng and Kimmel, 2005]. In comparison to previous forward-time simulations that simulate hypothetical genomic regions with two-state mutations (wild-type and mutant alleles), uniform recombination, and selection models based on the combined effect of individual mutations with random fitness effect [Chadeau-Hyam et al., 2008; Hernandez, 2008; Peng and Liu, 2010], this simulation engine simulates the evolution of DNA sequences of the human genome with fitness effects evaluated from amino acid changes of the translated protein sequences. More specifically, before starting a simulation, this simulation engine identifies all genes (isoforms) from the NCBI reference sequences database [Pruitt et al., 2007] that locate within or overlap with the user-provided regions, locates exons and coding regions, and records positions of codons on the forward and reverse strands. It then determines recombination rates at each locus according to a fine-scale genetic map with recombination rates and hotspots provided by the HapMap project [Myers et al., 2005]. Starting from an initial population with only reference sequences at these regions, the simulation engine evolves the population forward in time, following a user-specified demographic model that models the evolution of one or more human populations. During evolution, a nucleotide mutation model introduces new mutants to the population. If one or more mutations happen within the coding regions, the DNA sequences of all affected genes are transcribed and translated to protein sequences. Changes of amino acids are categorized as missense (regular amino acid changes), stoploss, and stopgain mutations and are used to evaluate the fitness and/or traits of individuals. There can be different nucleotide changes at a locus in the population (multiple alternative alleles), and the same mutation can have different fitness effects if other mutations change nucleotides on the same codon. The most damaging effect is considered if a single mutation causes codon changes in more than one gene. This evolutionary model ignores mutations in noncoding regions and synonymous mutations in coding regions and leads to lower allele frequencies for more damaging (nonsynonymous) mutations, which is consistent with theoretical estimates and empirical observations [Gorlov et al., 2008; Gorlova et al., 2012]. Most simulation models can take advantage of a scaling approach to reduce the computational burden of forward-time simulations [Hoggart et al., 2007; Peng and Amos, 2010]. Such simulations apply magnified (multiplied by λ) genetic forces (mutation, recombination, and selection) to smaller populations of size N/λ for t/λ generations and are usually good approximations to the full-scale simulations [Hernandez, 2008].

In addition to simulation engines, VST provides functions for data acquisition, file format conversion, and generation of random samples, case-control samples, and pedigrees with fixed or random structure. Because simulated populations might not have enough individuals with a rare disease or an extreme trait, samples are not drawn directly from the simulated population. Instead, VST repeatedly produces offspring from parents chosen randomly from the simulated population, keeping or discarding offspring according to their affection status or trait value, until enough samples are collected. This method essentially draws samples from an infinite-sized offspring population of the simulated population. It reduces but does not eliminate the need to simulate large populations with rare diseases [Peng, 2010] because the diversity of samples, especially affected individuals, is still determined by the size of parental populations.

Distribution of Simulation Models and Simulated Datasets

Depending on the workflow of particular simulation studies, simulated datasets can be exported in standard VCF format or in delimiter-separated text formats (for phenotype) or saved as *Variant Tools* snapshots. Whereas genotypes in VCF format can be readily analyzed by other programs, the snapshots can be easily downloaded and loaded by *Variant Tools* and analyzed by more than 20 rare variant association methods implemented in *Variant Association Tools* [Wang et al., 2014]. All simulation models and simulated datasets are stored online in the variant tools repository. Whereas casual users can download simulated datasets and use them without running VST, other users can simulate data using download simulation models or models created by them.

Results

As a demonstration of the simulation engines of VST, we simulated the evolution of a sequence of 63,000 base pairs on chromosome 17 (chr17:41,200,001–41,263,000) using four simulation methods: a neutral coalescent simulation using ms [Hudson, 2002] (wrapped by VST), neutral forward-time simulation, simulation with natural selection, and resampling-based simulation. This region overlaps with five isoforms (NM_007294, NM_007297, NM_007298, NM_007299, NM_007300) of BRCA1, with 5,337 nucleotides (8.47%) within the coding regions of one of the isoforms.

All coalescent and forward-time simulations assumed a demographic model that mimics the evolution of the European population. This model evolves an initial population of 8,100 individuals for 81,000 generations and expands it to 900,000 individuals in 370 generations after a short bottleneck of 7,900 individuals [Kryukov et al., 2007]. A mutation rate of 1.8×10^{-8} was used for ms. A Jukes-Cantor DNA evolution model with mutation rate 2.4×10^{-8} was used for the forward-time simulations because the Jukes-Cantor model with mutation rate μ mutates a nucleotide to any of the four nucleotides at equal probability and has an effective mutation rate of 0.75 μ. For VST simulations with natural selection, we used constant fitness values 0.005, 0.02, and 0.1 for missense, stoploss, and stopgain mutations, respectively. A multiplicative model was used to combine fitness values if an individual carried more than one nonsynonymous mutation. This model applies strong purifying selection to stopgain mutations and practically disallows the existence of such mutations in the population. The resampling-based simulation extracts genotypes at this region from the 1000 Genomes Project and expands the population to 10,000 individuals in 10 generations, subject to the same mutation, recombination, and selection forces as the forward-time simulation.

With 10,000 replicates for each simulation method, we drew a sample of 700 individuals from the simulated population. We counted the number of mutants at each locus and then the number of loci in each mutation frequency class. The results from two neutral models obtained from ms and VST differed in that the VST simulations yielded less singleton variants. This difference may be partly due to the different mutation models used by the two programs. Whereas ms uses an infinite-site model in which new mutations always lead to a new segregating site, VST uses a finite-site mutation model where mutations can happen at existing segregating sites (Fig. 1A). The allele frequency spectra of neutral and selection simulations differ only slightly because only 8.47% of the simulated region is under the influence of natural selection. We can observe a more significant difference between allele frequency spectra if we limit the loci to coding regions (Fig. 1B).

Except for resampling-based simulations, the allele frequency spectra of the simulated datasets are, however, different from what we observed from 700 random samples of the 1000 Genomes Project [Abecasis et al., 2012] (Fig. 1, blue bars). The observed allele frequency spectrum has more singletons than most simulated datasets, especially for coding regions of the VST simulations with natural selection. The reasons for this phenomenon can be multifold but we suspect that most rare variants were introduced during the rapid expansion of the European population, which had a much stronger effect than natural selection in coding regions.

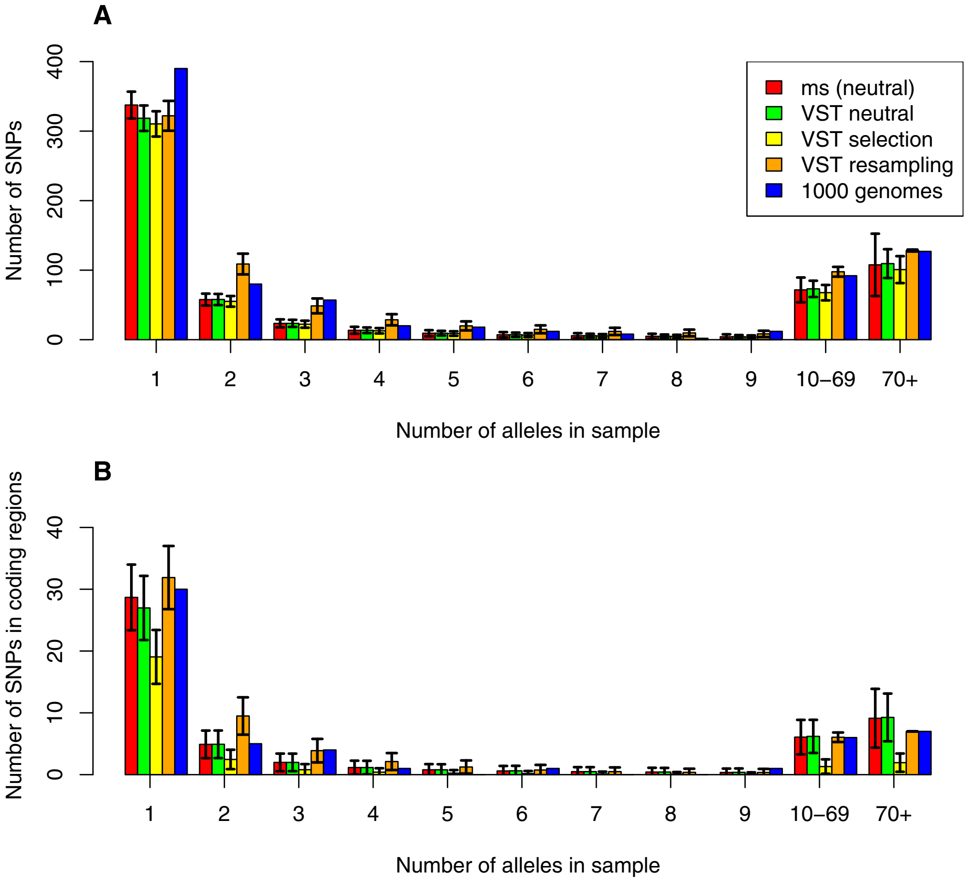

Figure 1. Comparison of four simulation results for the first nine mutant frequency classes $1 \leq i \leq 9$ and the number of single-nucleotide polymorphisms (SNPs) with 10 to 69 mutants, and at least 70 (5%) mutants, (A) in chr17:41,200,001–41,263,000 (63,000 bp) or (B) in coding regions of BRCA1 (5,337 bp). The *x*-axes show the number of alleles in a sample of 700 individuals; the *y*-axes show the number of SNP markers in each mutant frequency class. Means ± SD are shown, which were obtained from the four sets of simulations, each with 10,000 replicates.

To demonstrate the application of VST in rare variant association analysis, we simulated a miniature exome sequencing study with variants in the exon regions of 20 genes (isoforms) in the G Protein Coupled Receptors signaling pathway. These genes reside on chromosomes 6, 8, and 10 and chromosome X. The simulated regions overlap with 27 isoforms of 15 genes in the NCBI reference sequences database [Pruitt et al., 2007]. The coding regions of these genes range from 563 to 1818 bp and represent 16.2% of the total simulated region (17,841 of 110,387 bp).

We started with an ancestral population of 7,300 individuals and evolved it for 100,000 generations, following a Settlement of New World model that models the evolution of Africa, Asian, Mexican, and European populations and the formation of the Mexican American population [Gutenkunst et al., 2009]. The evolutionary process was subject to the influence of a K80 nucleotide mutation model [Kimura, 1980] with mutation rate of 1.8×10^{-8} and a transition transversion ratio of 2; a fine-scale recombination with hotspot model with average recombination rate (per region) ranging from 6.14×10^{-9} to 6.23×10^{-6}; and a natural selection model with selection coefficients of 0.0001, 0.0001, and 0.001 for missense, stoploss, and stopgain mutations, respectively. Part of the evolutionary process (80,000 of the 91,200 generation burn-in stage) was shared among replicate simulations by starting the simulations from a saved population. A scaling factor of 4 was used to speed up the simulations.

We simulated a genetic disease caused by nonsynonymous mutations in four of the 15 simulated genes. Under this penetrance model, an individual who carries a missense, stoploss, and stopgain mutation has probabilities of 0.001, 0.001, and 0.01, respectively, to be affected. Individuals will have higher probabilities to be affected if they carry more than one mutation (a multiplicative model is used) and a probability of 0.0001 if they carry no nonsynonymous mutation. We drew 1,000 and 2,000 cases and matching numbers of controls from the simulated populations of size 151,521 using three models. The first model drew cases and controls from the same European population; the second model drew cases and controls from the Mexican American population, which is admixed from the Mexican and European populations; and the third model drew cases from the Asian population and controls from the European population. We applied a burden test proposed by Morris and Zeggini [2010] to each of the simulated datasets, first to all variants, then to only nonsynonymous, stopgain, and stoploss variants identified by snpEff [Cingolani et al., 2012]. Five genes were excluded from the latter analysis because of small numbers of nonsynonymous mutations in these genes.

Figure 2 plots the box-and-whisker plots of negative $\log_{10}$ *P*-values for association analyses between disease status and all variants (Fig. 2A) or all nonsynonymous mutations (Fig. 2B) using 1,000 (bottom) or 2,000 (top) cases and matching numbers of controls. In contrast to the European and Mexican American simulations, which had reasonable false-positive rates (approximately 5% for significance level 0.05) for noncausing genes, analyses of the mismatch simulations yielded high proportions of spurious associations between all genes and disease status. This signifies the importance of using ethnicity-matched samples for case-control association analysis and the danger of using public controls (e.g., data from the 1000 Genomes Project) for association analysis. Because the disease is caused by nonsynonymous mutations in the causal genes, limiting association tests to such variants is predictably more powerful than tests based on all variants. It is interesting, however, that association tests using an admixed population are more powerful than tests using all European samples. The reasons behind this phenomenon require further investigation of the distribution of causal variants in the simulated populations.

Discussion

This article introduces VST, a simulation package that simulates samples for genetic epidemiological studies using next-generation sequencing data. VST aims to simulate realistic samples for such studies with a forward-time simulation engine that models the distributions and functional effect of synonymous and nonsynonymous mutations in coding and noncoding regions. Because of the use of DNA nucleotide mutation models and protein-based selection and penetrance models, datasets simulated by VST consist of real mutations on the human genome and can be further annotated and analyzed by other programs. Additional features such as the simulation of indels and structural variations, effects of frame-shift and splice site mutations, genotyping error, and missing data could be incorporated to capture more complexity of empirical studies. To avoid making the simulations unnecessarily complex and difficult to interpret, these features will be introduced gradually to new simulation models that consider these effects.

The realism of simulations is, however, limited by our knowledge about reality and the models to model it. Because of our incomplete understanding about the functional effect of mutations on simulated genes and the complex evolutionary history of human populations that shape the genomic composition of the current population, it is not possible to simulate datasets that match empirical data exactly. Datasets simulated by the forward-time engine of VST therefore reflect random outcomes of the underlying models and should not be compared site by site to empirical data. Users who need to simulate data with real variants observed in human populations can use a resampling-based approach (also provided by VST), although such simulations have their own problems (see Xu et al. [2013] for a comparison of multiple methods).

The forward-time simulation engine of VST is based on a highly optimized simulation environment simuPOP [Peng and Kimmel, 2005]. Using a scaling factor of 4, a forward-time simulation of example 2 could be completed in 25 min on a computer with 2.4G Hz Xeon CPU and 8G of RAM. Datasets simulated using a scaling factor of 4 are comparable to those from full-scale simulations, which would take 5.5 hr to complete. A larger scaling factor could be used but

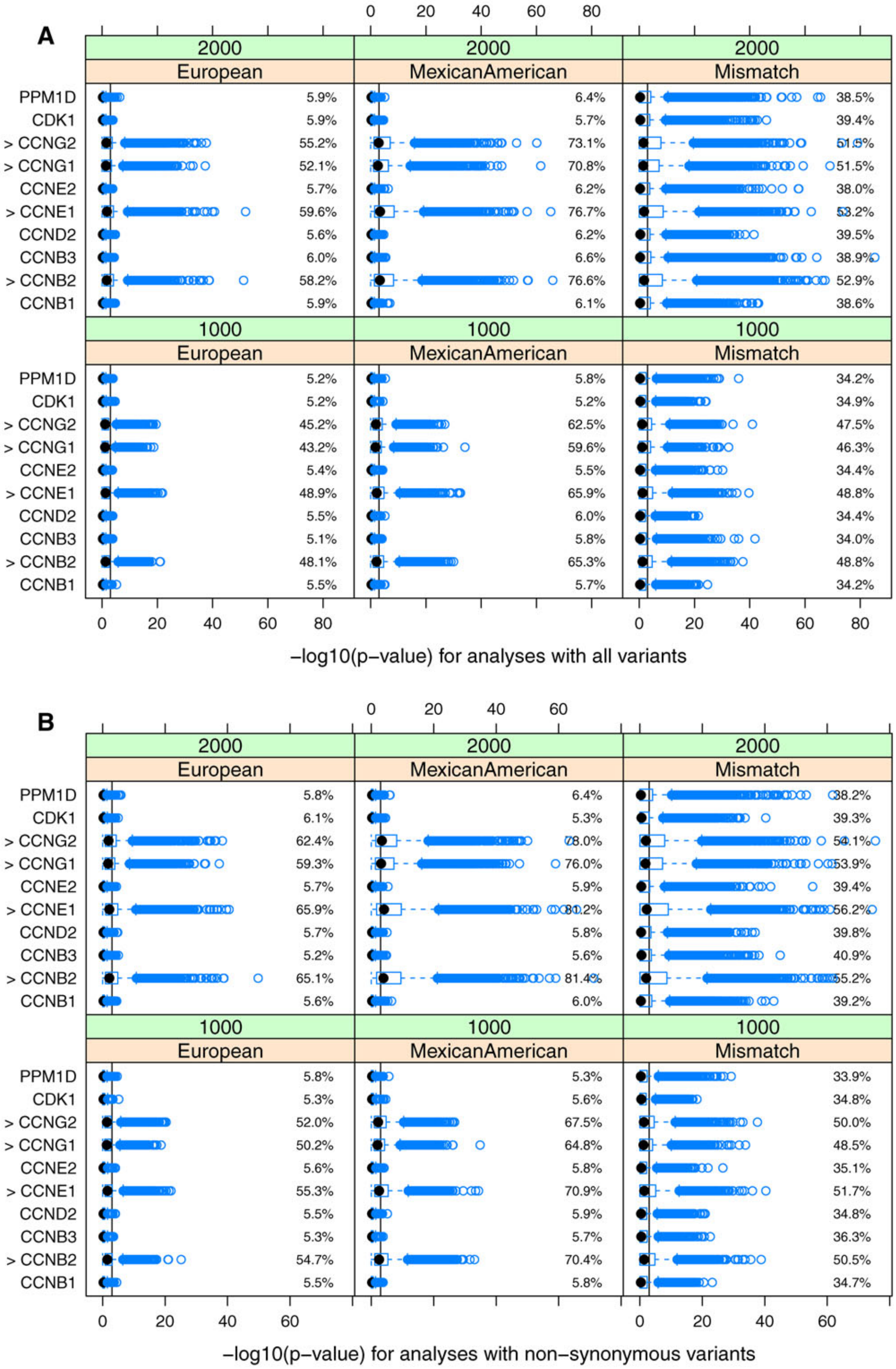

Figure 2. Box-and-whisker plots of negative log10 *P*-values (*x*-axis) of association analyses of 10 genes (*y*-axis) using 1,000 cases (bottom) or 2,000 cases (top) and matching numbers of controls, for cases and controls drawn from European (left), Mexican American (middle), and separately from Asian and European populations (right). Each box-and-whisker plot represents *P*-values of 10,000 replicate simulations, although some tests do not yield valid *P*-values due to insufficient number of variants. Genes that are causal are marked by a leading ">" before their names. The vertical lines represent *P*-values of 0.05, and the numbers after box-and-whisker plots are the power of tests estimated by percentage of replicates with *P*-values less than or equal to 0.05.

excessive scaling might lead to skewed patterns of diversity in the simulated sequences [Uricchio and Hernandez, 2014], and result in related samples if large samples are drawn from a relatively small simulated population. Although other simulation techniques (e.g., using coalescent-based simulation for the burn-in stage of the simulations) could be used to improve the performance of forward-time simulations, a cluster system is generally recommended for large-scale simulations using VST.

Because a VST simulation is executed as a *Variant Tools* pipeline, it can be used to analyze simulated datasets using arbitrary external commands. For example, simulation models defined in Peng2014_ex2 contain steps to draw samples for all three models, call snpEff to annotate simulated datasets, and select and analyze variants using *Variant Association Tools* [Wang et al., 2014]. Such flexibility allows the specification and distribution of a complete simulation study in a simulation specification file. We encourage users of VST to send us their simulation models for published simulation studies and share them through the *Variant Tools* repository. This will allow interested users to reproduce the simulations using VST and apply their own statistical methods. The sharing of simulation models and simulated datasets facilitates reproducible simulation studies and allows fair comparison between statistical methods, which is essential for identifying the strengths and weaknesses of these methods [Cheng et al., 2014].

Web Resources

Variant Simulation Tools: http://varianttools.sourceforge.net/Simulation
Variant Tools: http://varianttools.sourceforge.net
simuPOP: http://simupop.sourceforge.net

Acknowledgments

This work was supported in part by grant 1R01HG005859 from the National Human Genome Research Institute, the Chapman Foundation, the Michael and Susan Dell Foundation (honoring Lorraine Dell), the Prevent Cancer Foundation, and MD Anderson Cancer Center Support grant P30 CA016672. All simulations were performed using the MD Anderson High Performance Computing Cluster. The author has declared no conflicts of interest.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.

Auer PL, Wang G, Project NES, Leal SM. 2013. Testing for rare variant associations in the presence of missing data. *Genet Epidemiol* 37(6):529–538.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR and others. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 4(5):e1000083.

Byrnes AE, Wu MC, Wright FA, Li M, Li Y. 2013. The value of statistical or bioinformatics annotation for rare variant association with quantitative trait. *Genet Epidemiol* 37(7):666–674.

Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, de Iorio M, Balding DJ. 2008. Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics* 9:364.

Chen HS, Hutter CM, Mechanic LE, Amos CI, Bafna V, Hauser E, Hernandez RD, Li C, Liberles DA, McAllister K, and others. 2014. Genetic simulation tools for post-genome wide association studies of complex diseases. *Genet Epidemiol* 341(2):617–631.

Cheung YH, Wang G, Leal SM, Wang S. 2012. A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. *Genet Epidemiol* 36(7):675–685.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80–92.

Cule E, de Iorio M. 2013. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet Epidemiol* 37(7):704–714.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. *PLoS Biol* 8(1):e1000294.

Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI. 2011. Evolutionary evidence of the effect of rare variants on disease etiology. *Clin Genet* 79(3):199–206.

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* 82(1):100–112.

Gorlova OY, Ying J, Amos CI, Spitz MR, Peng B, Gorlov IP. 2012. Derived SNP alleles are used more frequently than ancestral alleles as risk-associated variants in common human diseases. *J Bioinform Comput Biol* 10(2):1241008.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5(10):e1000695.

Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* 24(23):2786–7278.

Hoban S, Bertorelle G, Gaggiotti OE. 2011. Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet* 13(2):110–122.

Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, de Iorio M, Balding DJ. 2007. Sequence-level population simulations over large genomic regions. *Genetics* 177(3):1725–1731.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.

Jiang D, McPeek MS. 2014. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet Epidemiol* 38(1):10–20.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16(2):111–120.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 80(4):727–739.

Leal SM, Yan K, Muller-Myhsok B. 2005. SimPed: a simulation program to generate haplotype and genotype data for pedigree structures. *Hum Hered* 60(2):119–122.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83(3):311–321.

Li C, Li M. 2008. GWAsimulator: a rapid whole-genome simulation program. *Bioinformatics* 24(1):140–142.

Lin W-Y, Yi N, Lou X-Y, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N. 2013. Haplotype kernel association test as a powerful method to identify chromosomal regions harboring uncommon causal variants. *Genet Epidemiol* 37(6):560–570.

Liu Y, Athanasiadis G, Weale ME. 2008. A survey of genetic simulation software for population and epidemiological studies. *Hum Genomics* 3(1):79–86.

Logsdon BA, Dai JY, Auer PL, Johnsen JM, Ganesh SK, Smith NL, Wilson JG, Tracy RP, Lange LA, Jiao S and others. 2014. A variational bayes discrete mixture test for rare variant association. *Genet Epidemiol* 38(1):21–30.

Montana G. 2005. HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics* 21(23): 4309–4311.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34(2):188–193.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310(5746):321–324.

Peng B. 2010. Simulating gene-environment interactions in complex human diseases. *Genome Med* 2(3):21.

Peng B, Amos C. 2010. Forward-time simulation of realistic samples for genome-wide association studies. *BMC Bioinformatics* 11(1):442.

Peng B, Chen HS, Mechanic LE, Racine B, Clarke J, Gillanders E, Feuer EJ. 2014. Genetic data simulators and their applications: an overview. *Genet Epidemiol* 39(1):2–10.

Peng B, Kimmel M. 2005. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* 21(18):3686–3687.

Peng B, Liu X. 2010. Simulating sequences of the human genome with rare variants. *Hum Hered* 70(4):287–291.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue):D61–D65.

Ritchie MD, Bush WS. 2010. Genome simulation approaches for synthesizing in silico datasets for human genomics. *Adv Genet* 72:1–24.

Saad M, Wijsman EM. 2014. Power of family-based association designs to detect rare variants in large pedigrees using imputed genotypes. *Genet Epidemiol* 38(1): 1–9.

San Lucas FA, Wang G, Scheet P, Peng B. 2012. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* 28(3):421–422.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15(11):1576–1583.

Uricchio LH, Hernandez RD. 2014. Robust forward simulations of recurrent hitchhiking. *Genetics* 197(1):221–236.

Wallace C. 2013. Statistical Testing of shared genetic control for potentially related traits. *Genet Epidemiol* 37(8):802–813.

Wang GT, Peng B, Leal Suzanne M. 2014. Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. *Am J Hum Genet* 94(5):770–783.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89(1):82–93.

Xu Y, Wu Y, Song C, Zhang H. 2013. Simulating realistic genomic data with rare variants. *Genet Epidemiol* 37(2):163–172.

Yampolsky LY, Kondrashov FA, Kondrashov AS. 2005. Distribution of the strength of selection against amino acid replacements in human proteins. *Hum Mol Genet* 14(21):3191–3201.

Yi N, Liu N, Zhi D, Li J. 2011. Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. *PLoS Genet* 7(12):e1002382.