Variant Simulation Tools

Bo Peng

Sep 25, 2014

Genetic Simulations

Why perform simulations?

To get data that match these (unrealistic) assumptions of our methods	• Validate statistical methods using simulated data based on specific assumptions
To evaluate conditions that could have given rise to current observations	• Simulate data using specific models and compare them to empirical data • Infer parameters from best-match simulations
To get multiple replicates of data	• Calculate empirical statistical power by applying statistical methods to a large number of simulated data • Compare power of multiple methods
To obtain information that are unavailable or too expensive to obtained empirically	• Genotypes of large pedigree, ancestral populations • Samples of very rare disease
To look backward and forward in time	• What are the impact of demographic and genetic features of a population? • Evaluate how changes to a system could change its attributes (cancer intervention)

What to simulate?

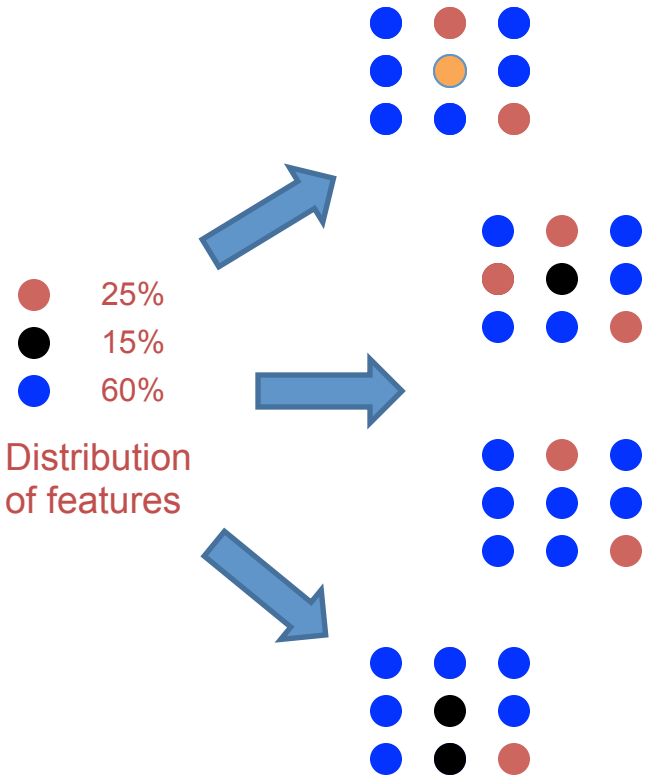
- Haploid and Diploid sequences
- Genetic markers
- Sex chromosomes
- Mitochondrial DNA
- RNA sequence
- Protein sequence

- SNP markers
- Microsatellite markers
- Insertions, deletions, inversion
- Large indels, structural variation
- Copy number variation
- Genotyping error
- Missing data

- Qualitative and quantitative traits
- Random sample
- Extreme traits
- Case control data
- Pedigree data
- Output from genotyping and other platforms

- Impact of bottleneck
- Impact of migration
- Impact of natural selection
- Impact of population expansion
- Impact of recombination

Theoretical simulations



Pros:

- Efficient
- Matching specific assumptions exactly

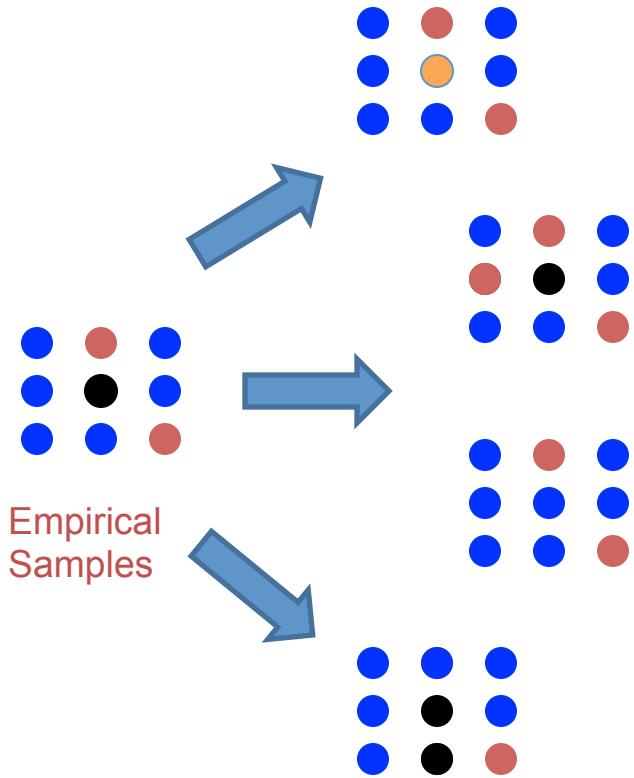
Cons:

- Difficult to handle multiple assumptions
- Difficult to simulate long genomic regions with linked loci

Ideal for:

- Simple data matching specific assumptions

Resampling-based simulations



Pros:

- Efficient
- Realistic samples
- Able to simulate genome-wide samples

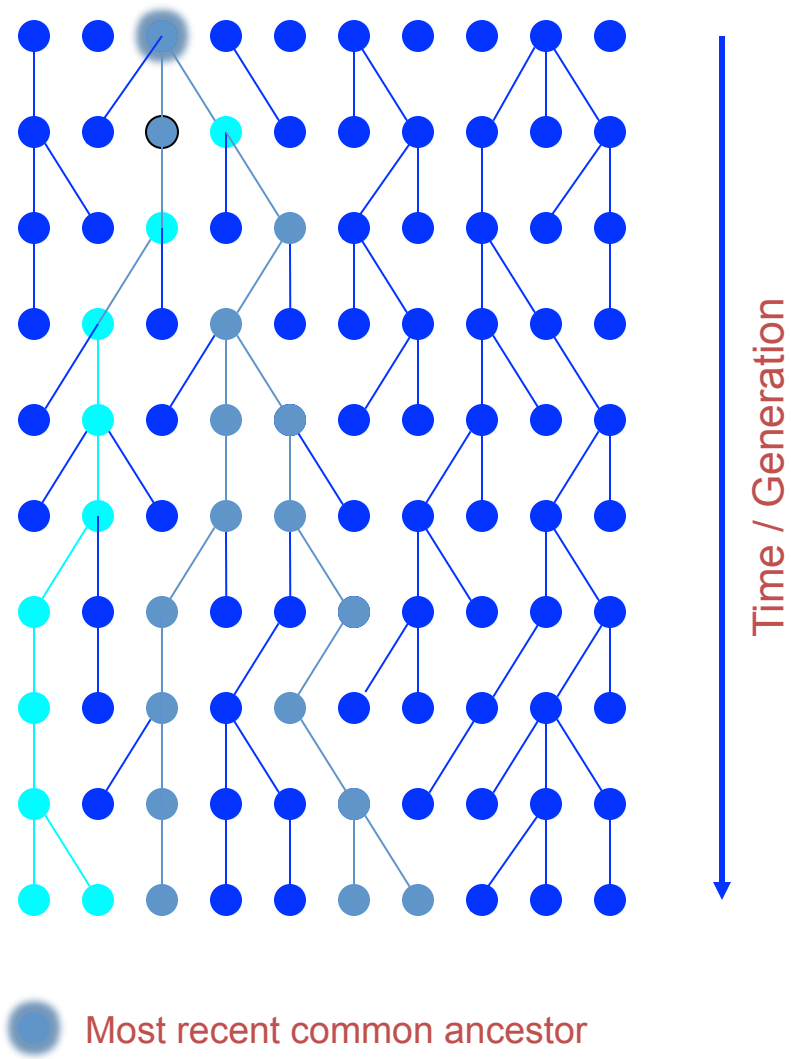
Cons:

- Difficult to match specified conditions
- Source data dependent
- Confounding genomic features
- Difficult to introduce additional genetic variations

Idea for:

- Long genomic regions with realistic features

Coalescent-based Simulation



Pros:

- Very efficient
- Support many mutation and migration models

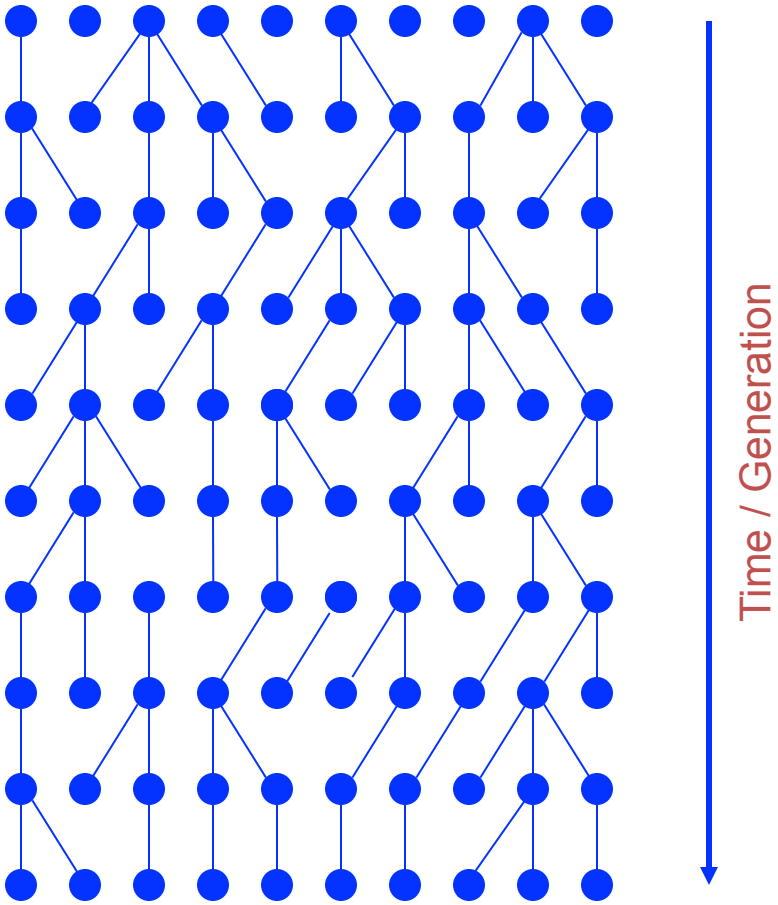
Cons:

- Difficult to simulate genotype-dependent or diploid-specific features e.g. natural selection and penetrance models
- Difficult to simulate long range genomic regions with recombination
- Does not provide good platform for complex disease

Ideal for:

- Large number of short neutral sequences

Forward-time Simulation



Pros:

- Extremely powerful and flexible in modeling natural selection, penetrance, and study designs
- Complete information about ancestral populations

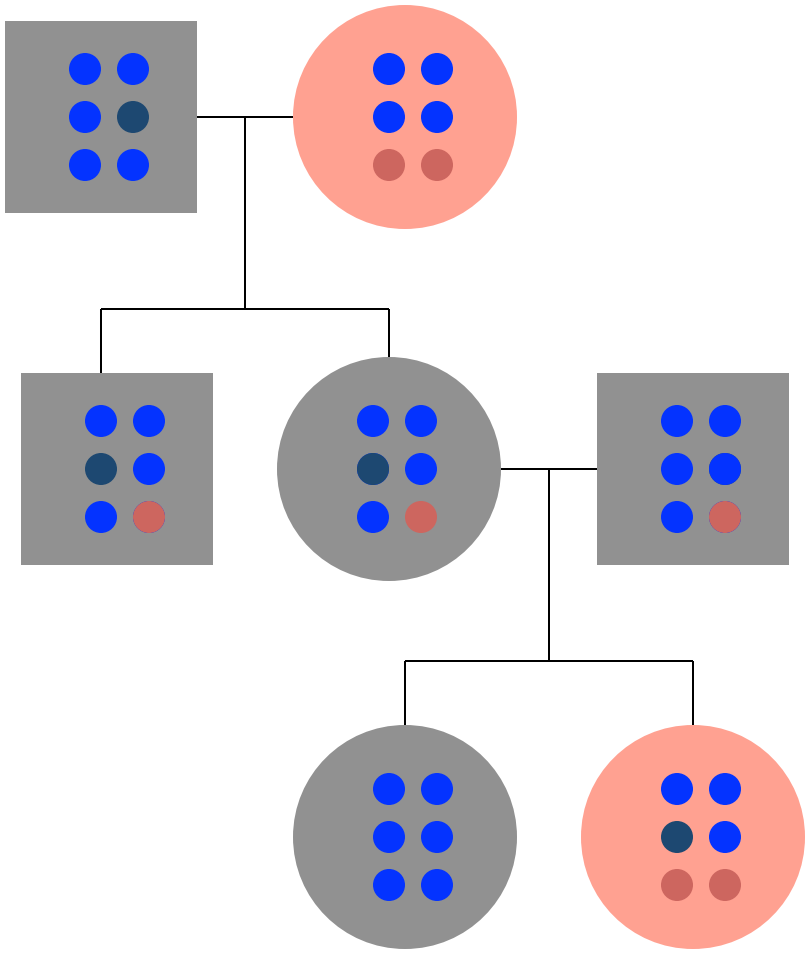
Cons:

- Can be hard to make realistic
- Difficult to simulate long genomic regions and rare phenotype

Ideal for:

- Observational simulations
- Samples under complex evolutionary scenarios and study designs

Gene Dropping



Pros:

- Efficient
- Adapt to arbitrary pedigree structure

Cons:

- Difficult to simulate genotype conditioning on specified traits

Ideal for:

- Simulating pedigree data from existing pedigree structures

Simulation of Genotypes and Phenotypes association – GAW16

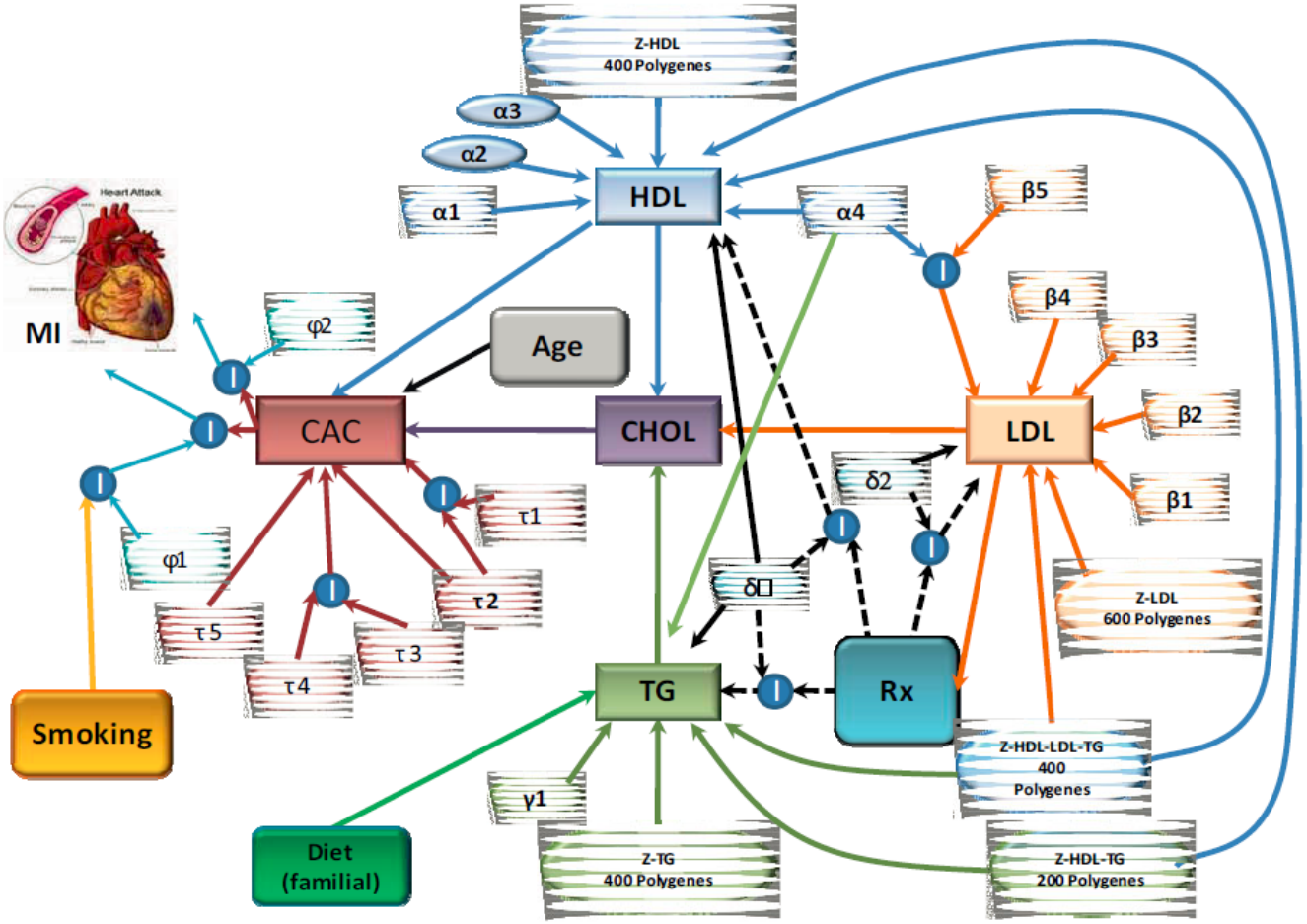
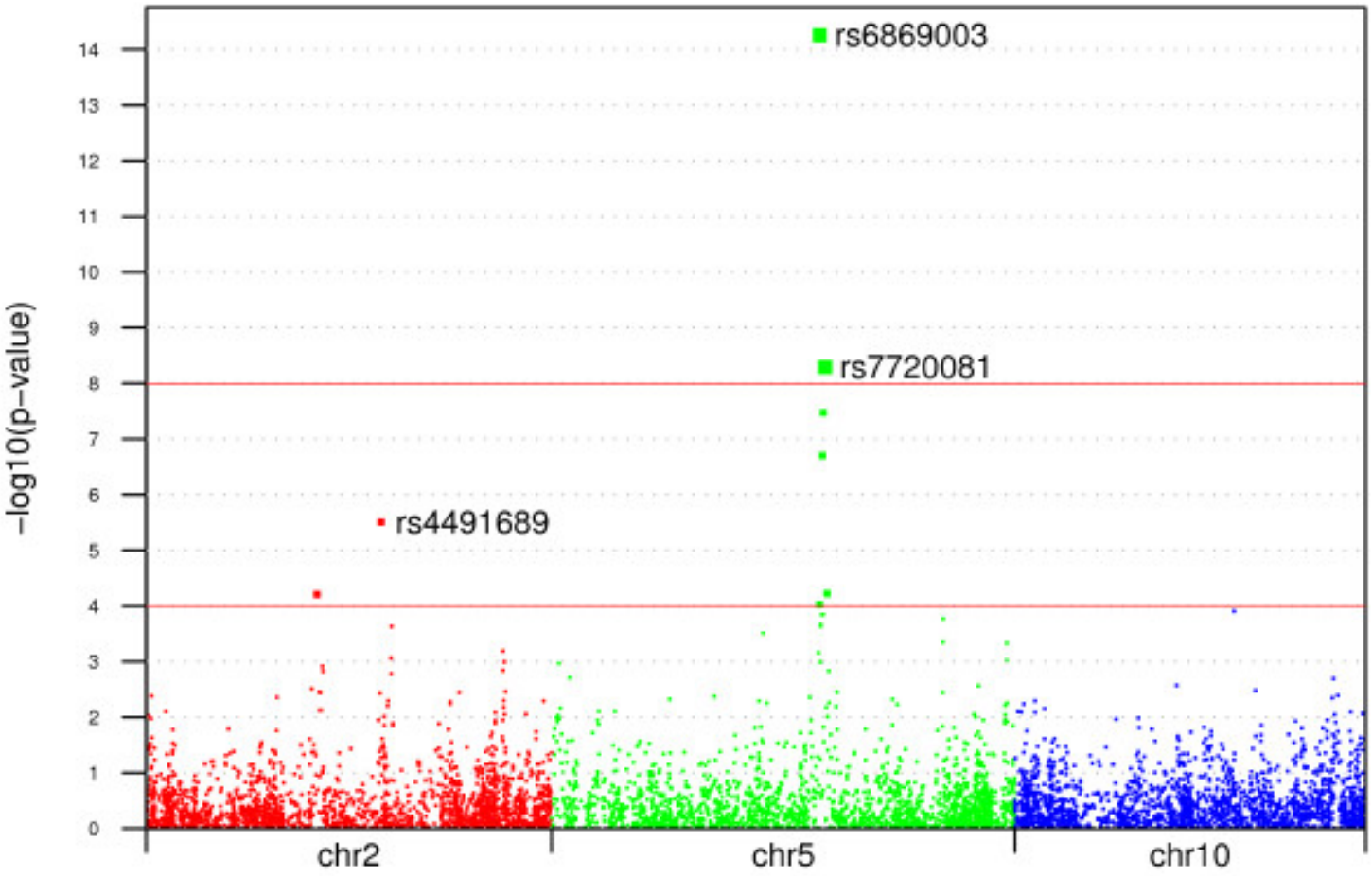


Figure 1

The Genetic Analysis Workshop 16 Problem 3 diagram. Figure 1 shows simulated phenotypes emulating the lipid domain (HDL, LDL, TG, and CHOL) and its contribution to cardiovascular disease risk (CAC and MI). Simulated major genes are symbolized with Greek letters. There are 1,000 polygenes for each trait HDL, LDL, and TG, several of them with pleiotropic effects. Continued lines and arrows show causality/interaction (I); dashed lines show pharmacogenetic effects only for subjects treated with medication, where response was dependent on the subjects' genotypes. Environmental factors such as diet, smoking, and medication were modeled in the simulation.

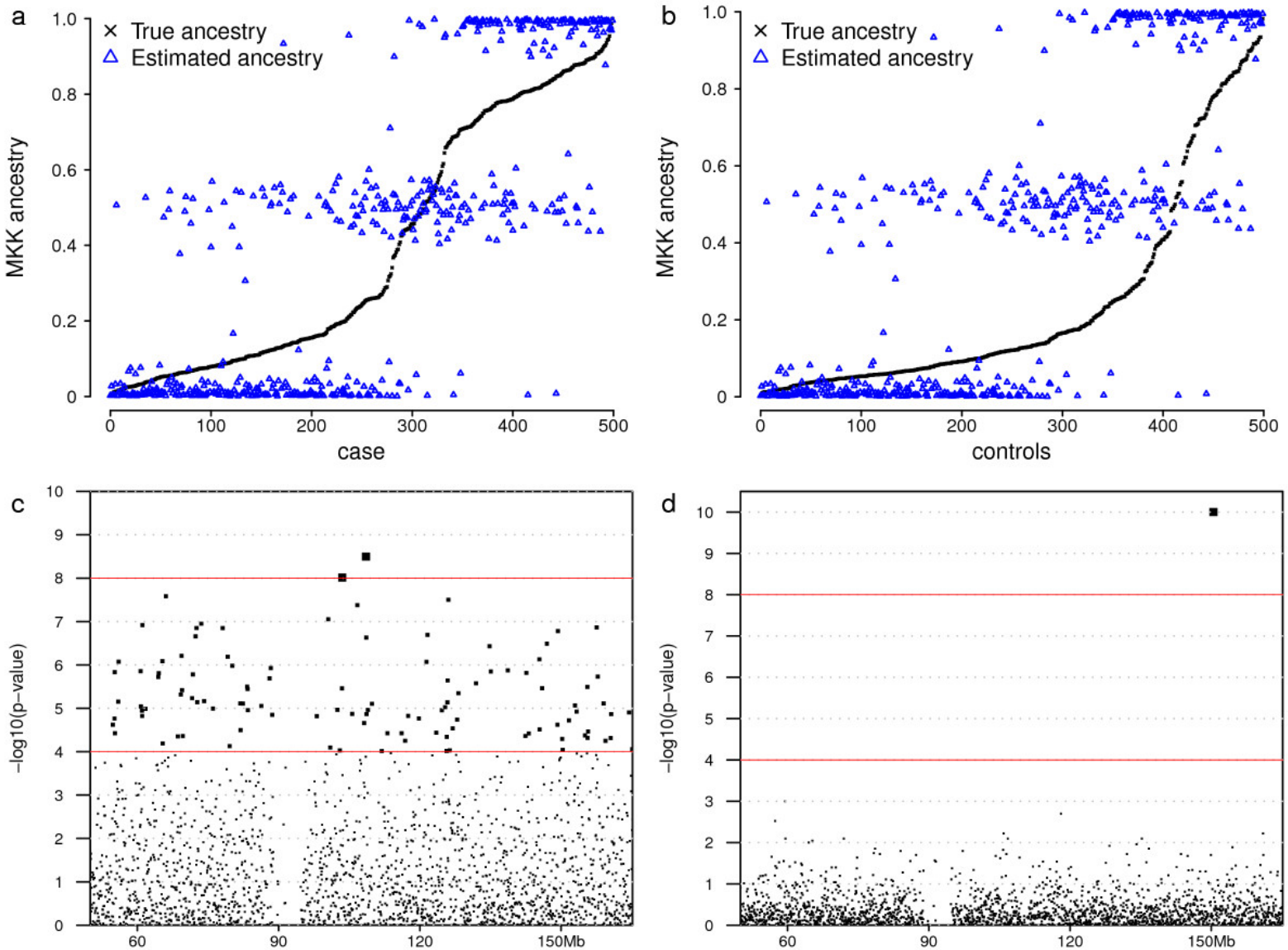
Sample Applications

Simulation of GWA Studies

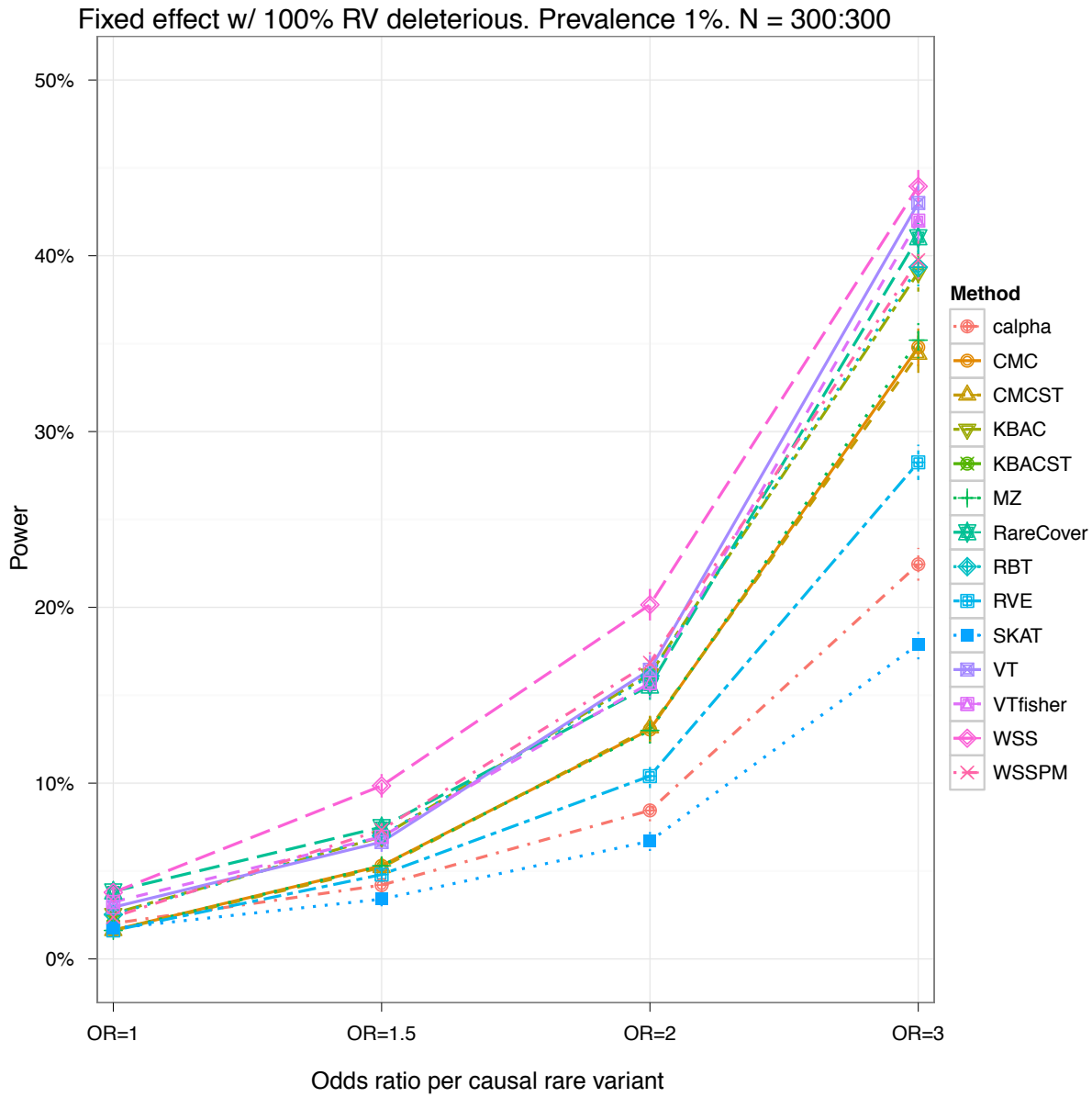


Peng & Amos, BMC Bioinformatics, 2010

Impact of population structure



Rare variant association analysis



Design and implementation

Why yet another simulator?

THE Truth

Truth that we think we know

Truth that we can model

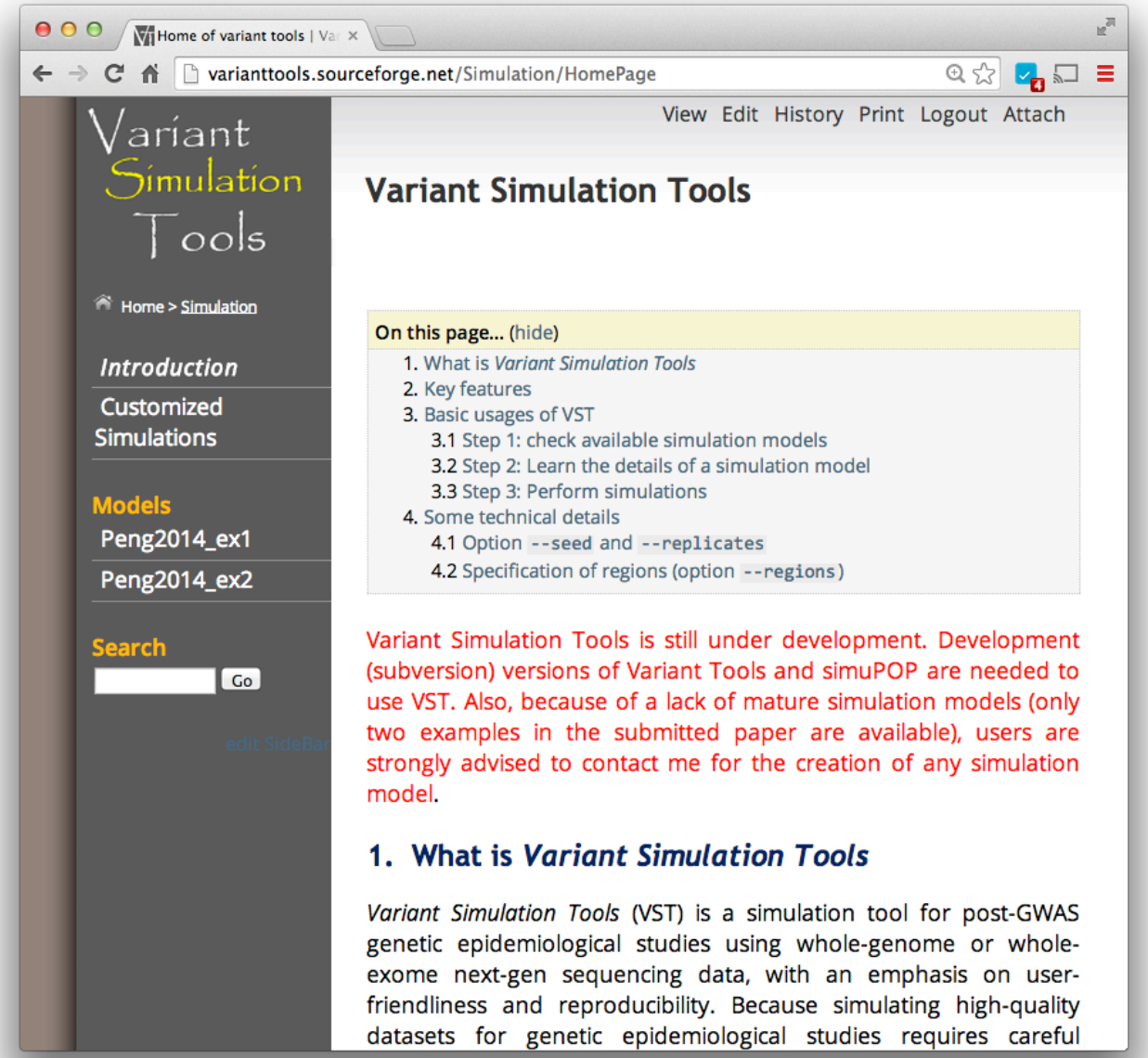
Truth that we can simulate

Many new methods are using prior biological knowledge in some way, for filtering, for pathway analysis, and for Bayesian priors

GAP to be filled

Most of our models follow standard genetic models (dominant, recessive, additive etc) and simple phenotype models.

Variant Simulation Tools

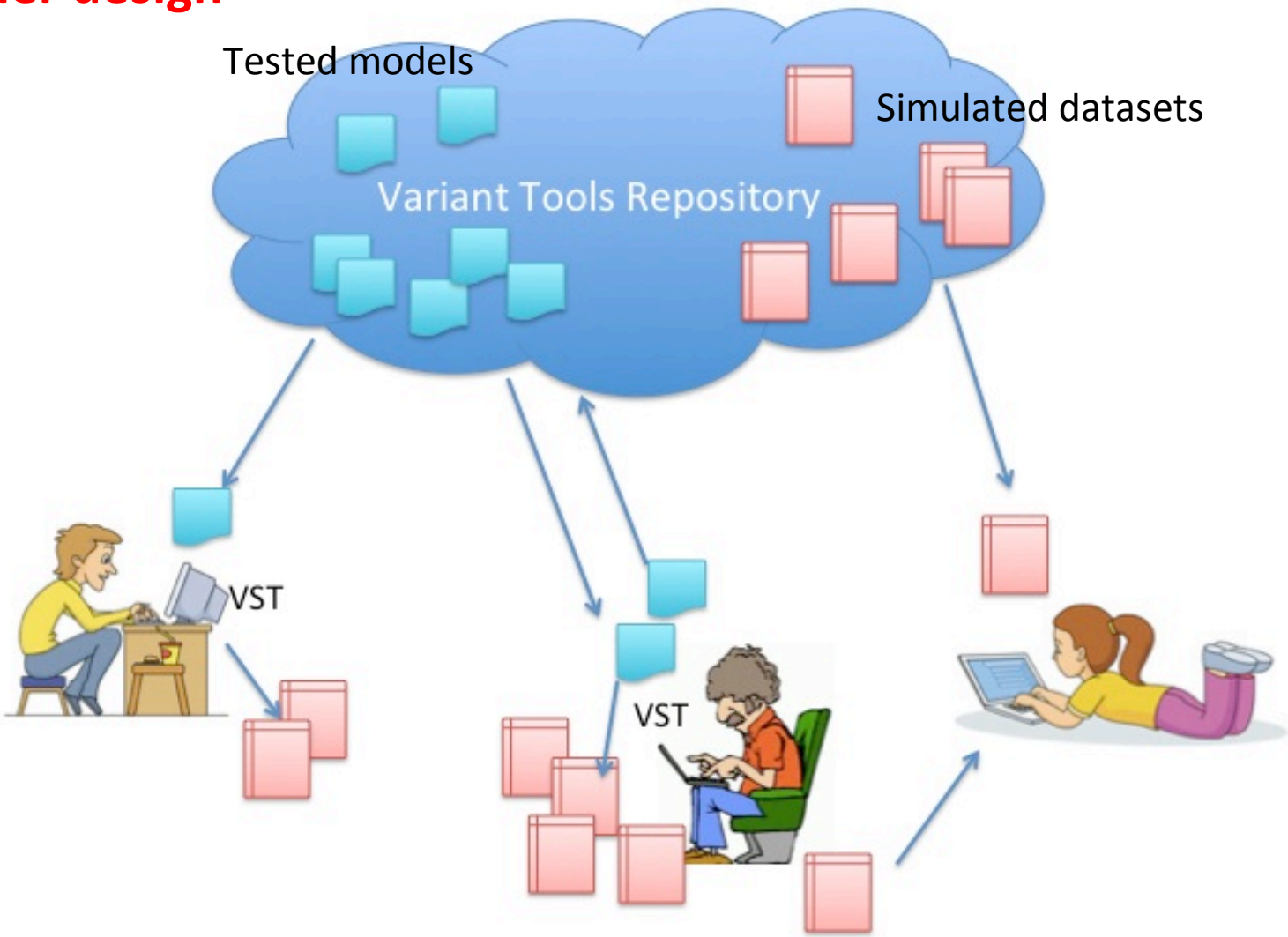


VST is a simulation tools for post-GWA genetic epidemiological studies, with emphases on realism, user-friendliness and reproducibility.

Peng, Genetic Epi, 2014

Variant Tools Repository

Two-tier design



Simulation Specification Files

```
[*_0]
action=CheckVariantToolsVersion('2.3.1')
comment=Check the version of variant tools. Version 2.3.1 or higher is required
    for the execution of this simulation.

[*_1]
action=ImportModules(['simuPOP.demography', 'VST_srv.py'])
comment=Import required models

[*_10]
input_emitter=EmitInput(select=${:not glob.glob('*.proj')})
action=RunCommand('vtools init Peng2014_ex1')
comment=Create a new project if there is no existing project under the current
    directory.

[ex1_neutral_20]
action=RunCommand('vtools use refGene')
comment=Link the refGene database to the project. This database is required
    to parse the regions for gene structure.

[ex1_neutral_30]
action=CreatePopulation(
    size=1000,
    regions='%(regions)s',
    output='cache/ex1_neutral_init_${seed}.pop')
output='cache/ex1_neutral_init_${seed}.pop'
comment=Create an empty simuPOP population for specified regions.

[ex1_neutral_40]
action=EvolvePopulation(
    output='ex1_neutral_evolved_${seed}.pop',
    mutator=sim.SNPMutator(u=1.8e-8 * %(scale)s, v=1.8e-8 * %(scale)s),
    demoModel = MultiStageModel([
        InstantChangeModel(T=81000 / %(scale)s, N0=8100 / %(scale)s,
            G=[70000 / %(scale)s, 71000 / %(scale)s], NG=[7900 / %(scale)s, 8100 / %(scale)s]),
        ExponentialGrowthModel(T=370 / %(scale)s, NT=900000 / %(scale)s)
     ]))
comment=Evolve the population with a SNP mutator, without recombination and natural selection.

[ex1_neutral_50]
Peng2011_srv.pipeline                                                    104,90        42%
```

Variant tools pipeline

Multiple models in one spec file

Can execute arbitrary commands

Allow additional pipeline steps and functions in Python

NOT user-friendly

Variant Tools + simuPOP

- Storage, annotation, and manipulation of variants
- Pipeline mechanism
- User interface
- Gene annotation
- Variant tools repository
- Integration with Variant Association Tools

- Mutant-based storage model for the simulation of rare variants
- Fine-scale recombination with hotspot
- Flexible natural selection models
- Demographic models

Simple command line interface

- Commands to show all simulation models

  `vtools show simulations`

- Clear documentation

  `vtools show simulation SPECFILE`

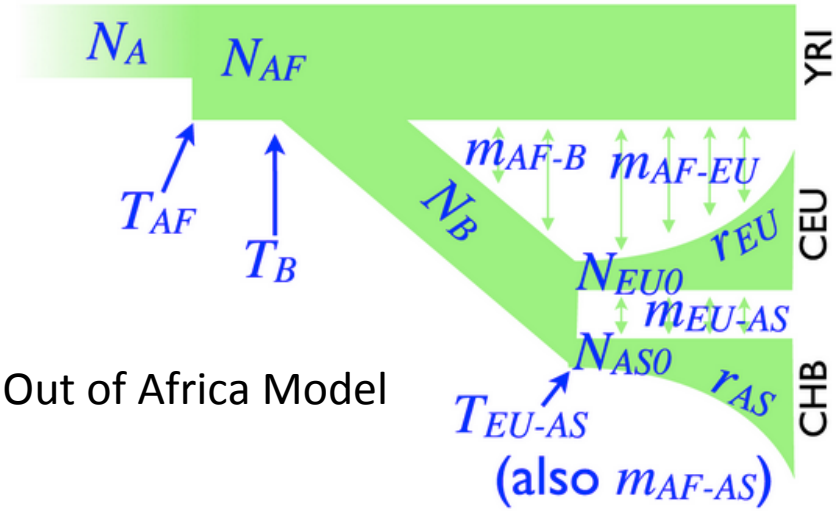
- Simple interface with no or few parameters

  `vtools simulate SPECFILE [model] [opt]`

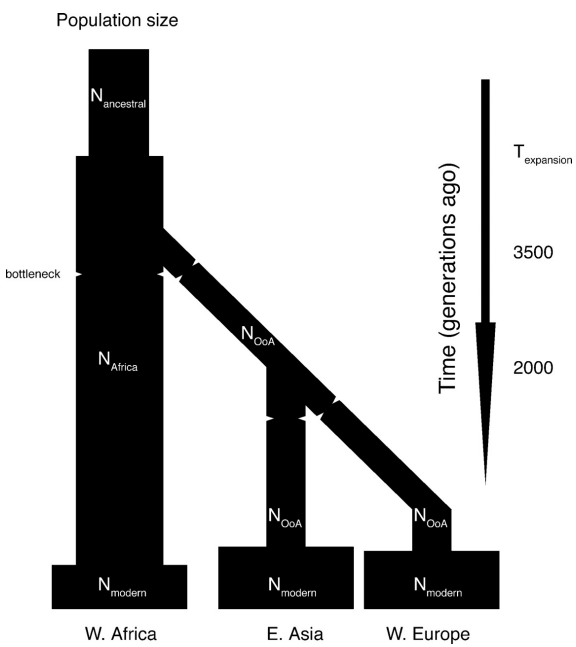
- Downloadable simulated datasets

  `vtools show snapshots`

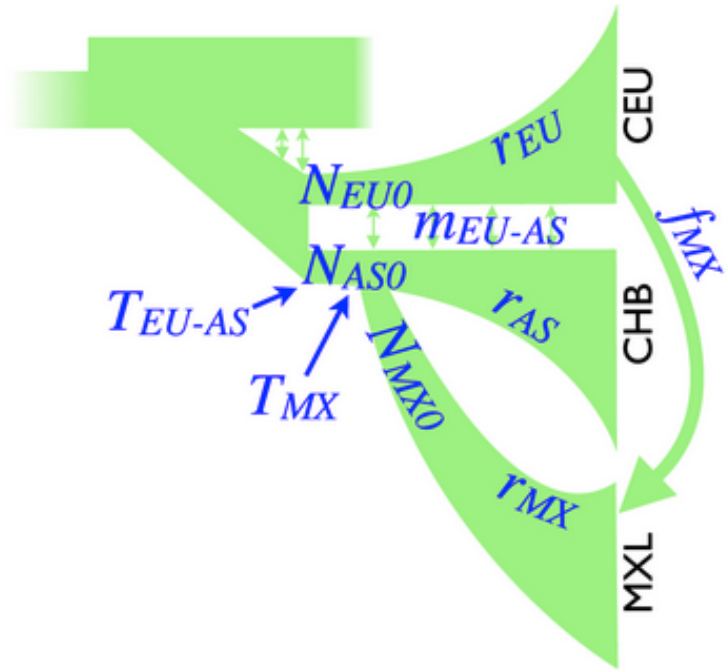
Realistic Demographic Models



Out of Africa Model



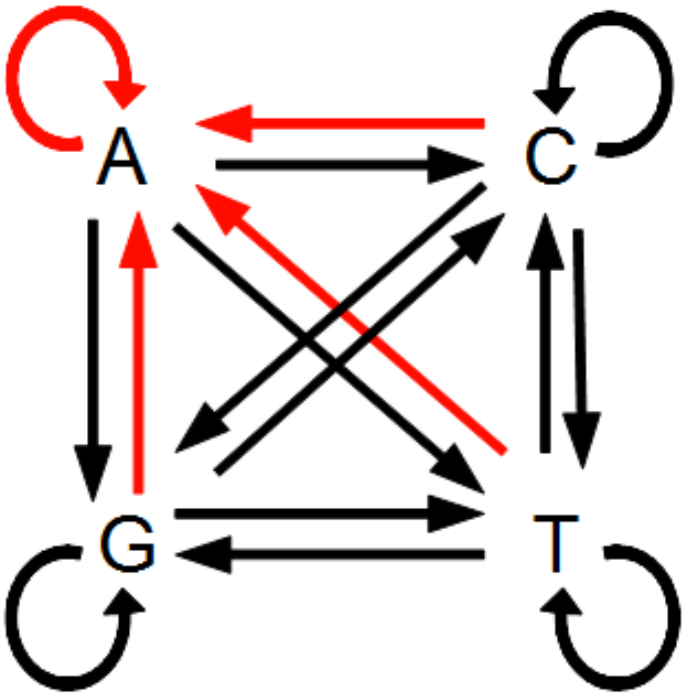
Cosi model



Settlement of New World Model

Schaffner et al, genome research, 2005
Gutenkunst, PLoS Genetics, 2009

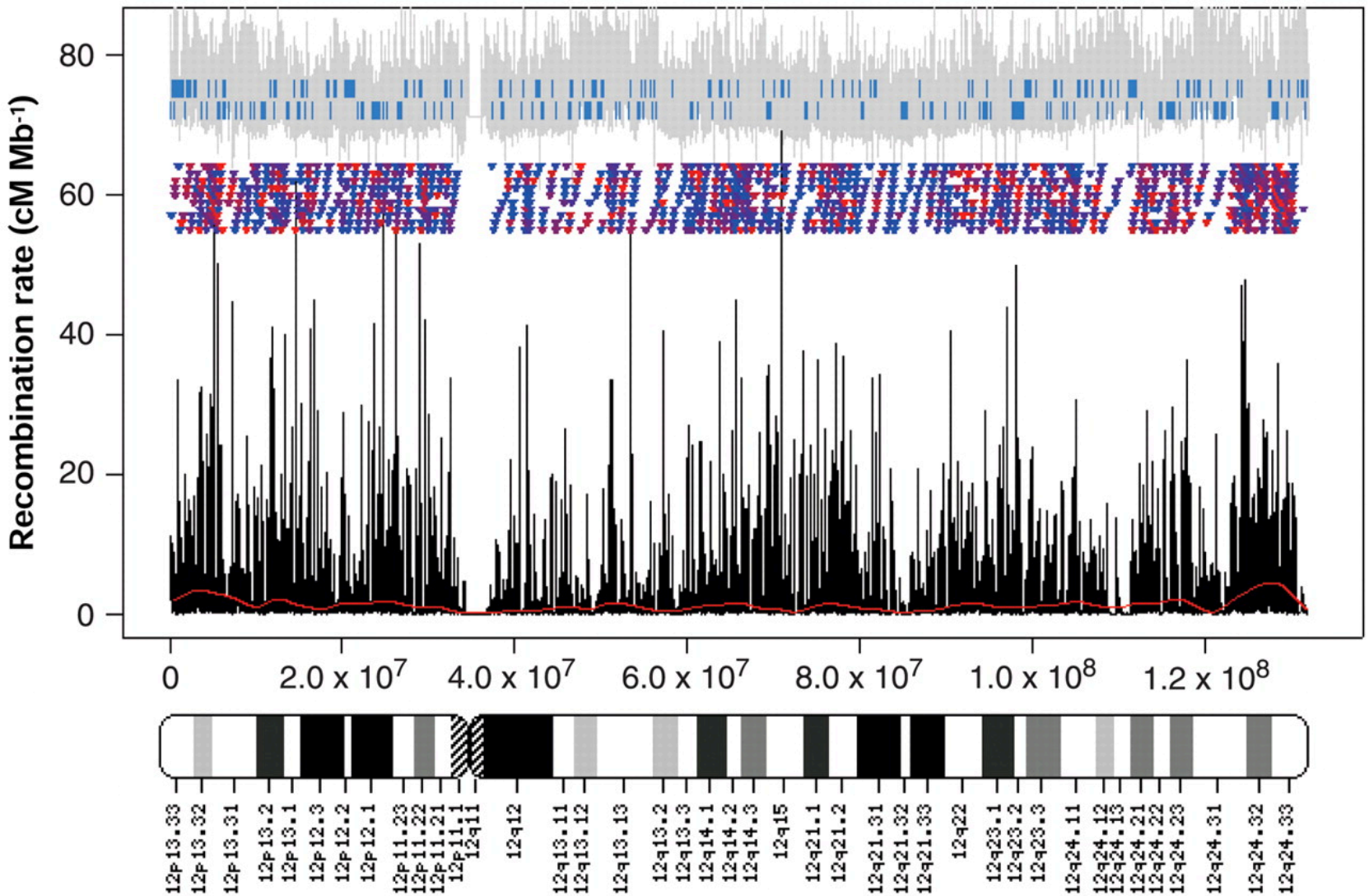
Nucleotide mutation models



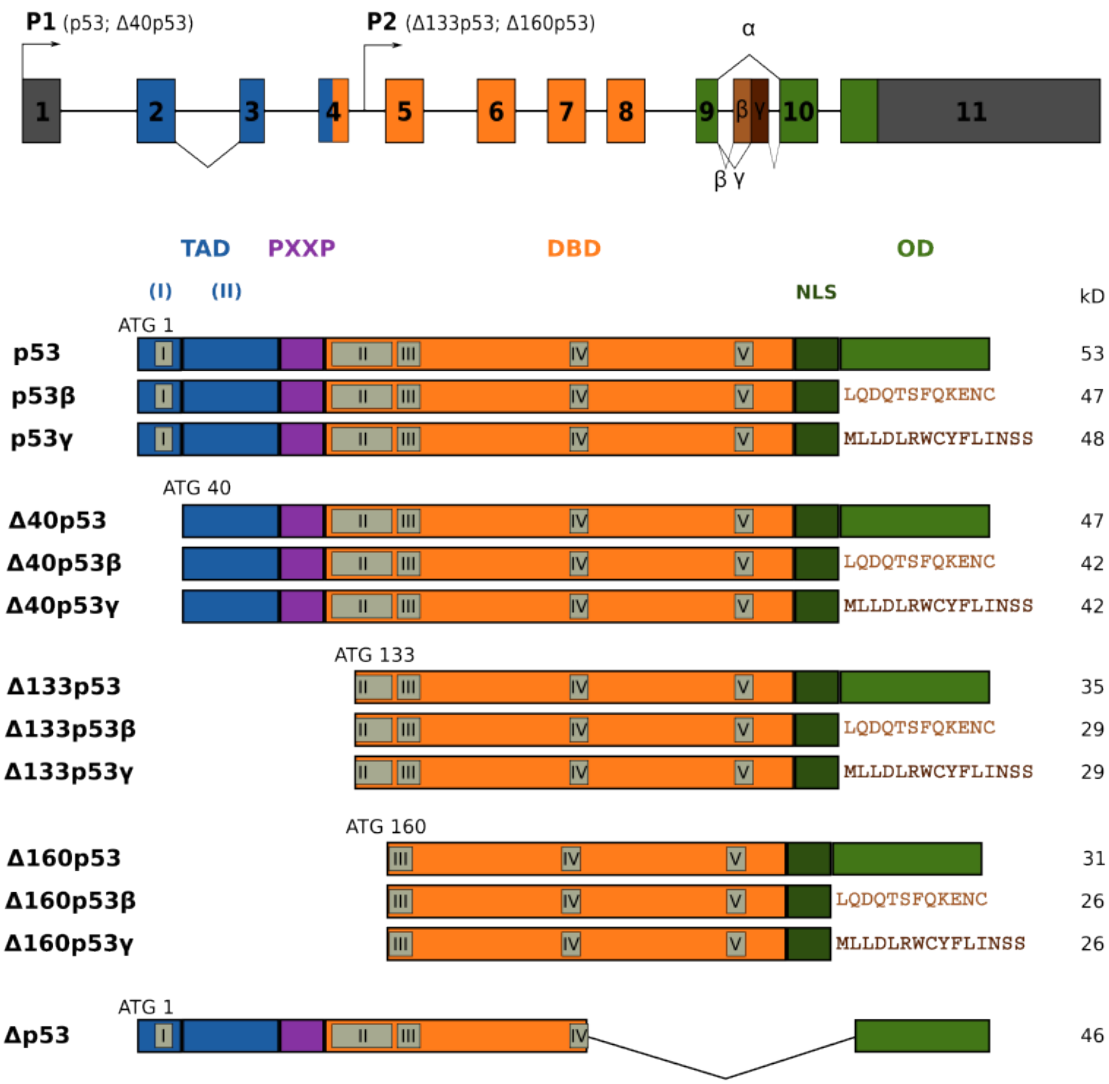
- Mutating real nucleotide sequences of specified regions of the human genome
- Allow multiple-alternative alleles
- Can model difference in transition and transversion rates

- Jukes Cantor 1969 model
- Kimura 1980 model
- Felsenstein 1981 model
- HKY 1985 model
- …

Fine-scale recombination map



Protein-based selection and trait models

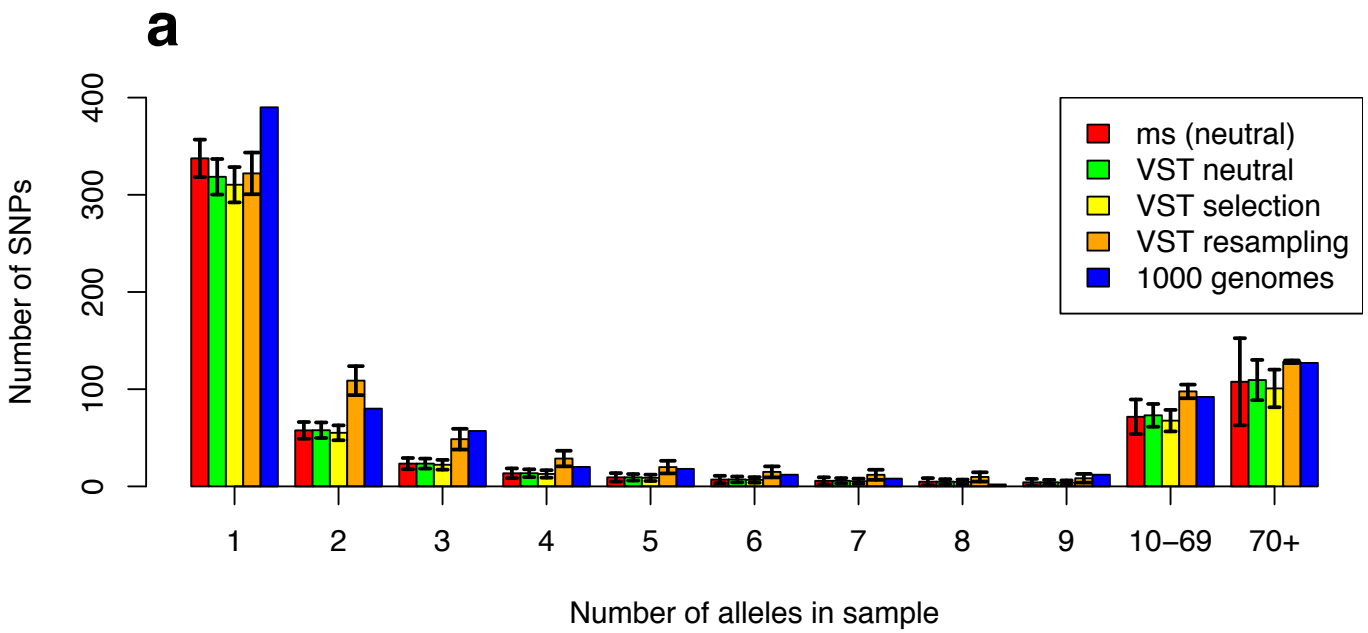


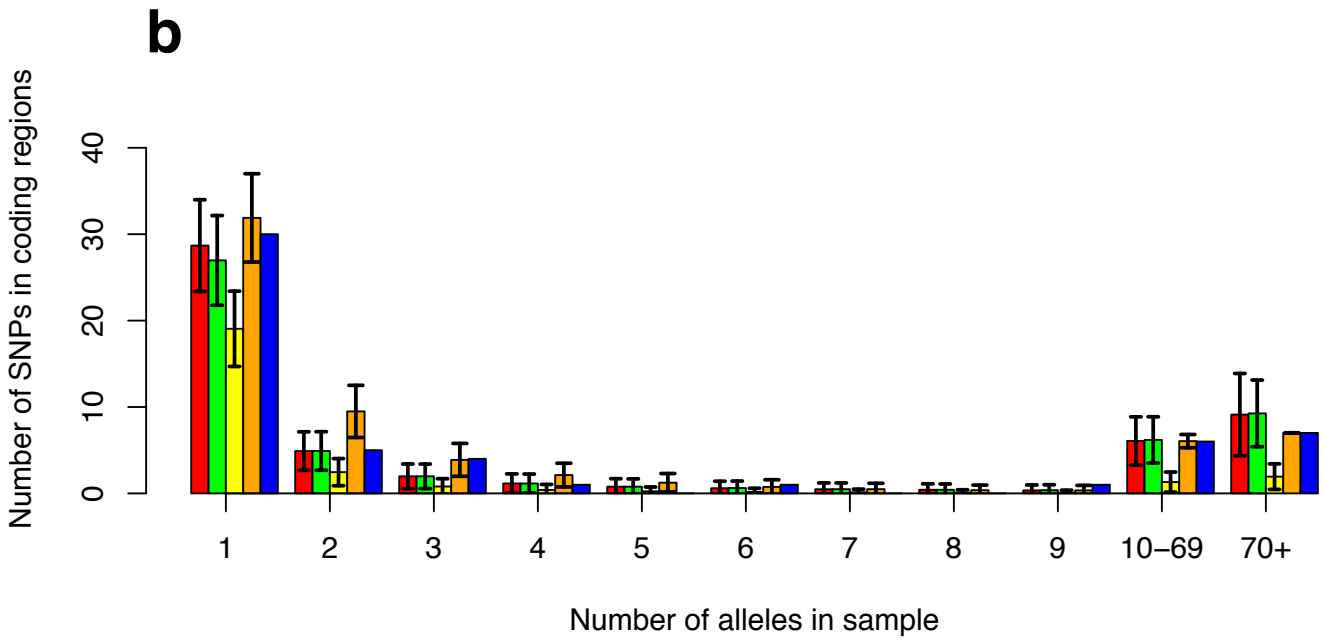
- Mutations in introns are silent
- One mutation can cause different fitness effects for multiple isoforms of a gene
- One mutation can have different fitness effect due to the occurrence of another mutation
- Different mutations can happen at the same location
- Fitness effect for regular non-synonymous, stop-gain and stop-loss mutations.

Example 1

Model Details

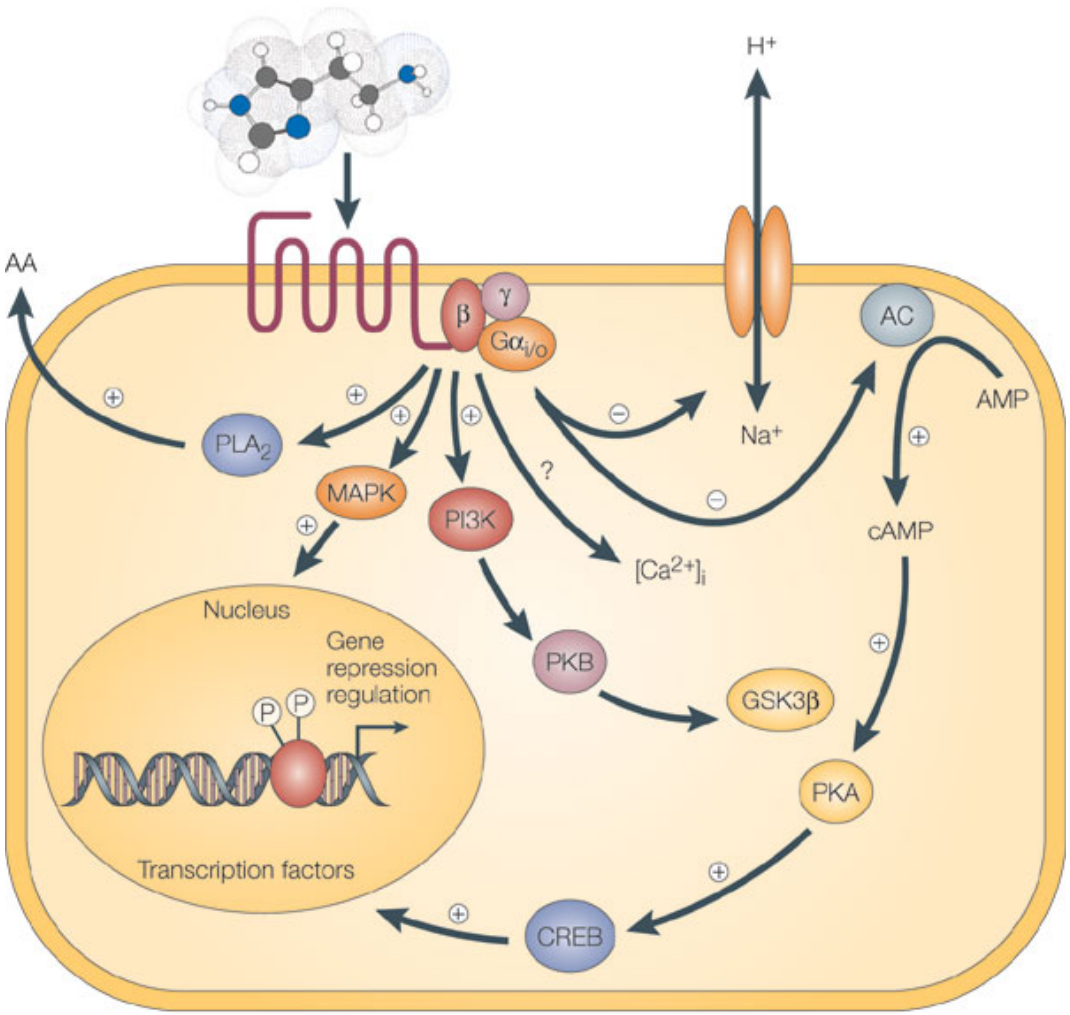
- chr17:41,200,001-41,263,000 (63,000 bp)
- NM_007294, NM_007297, NM_007298, NM_007299, NM_007300 (BRCA1)
- 5337 (8.47%) in coding regions of one of the isoforms
- Demographic model of European populations (Kryukov et al, 2007)
- Mutation rate 1.8×10^{-8} using a Jukes-Cantor model
- Constant fitness values 0.005, 0.02, and 0.1 for missense, stoploss, and stopgain mutations


a
Number of SNPs
0
100
200
300
400
ms (neutral)
VST neutral
VST selection
VST resampling
1000 genomes
1
2
3
4
5
6
7
8
9
10–69
70+
Number of alleles in sample


b
Number of SNPs in coding regions
0
10
20
30
40
1
2
3
4
5
6
7
8
9
10–69
70+
Number of alleles in sample

Example 2

G Protein Coupled Receptors signaling pathway



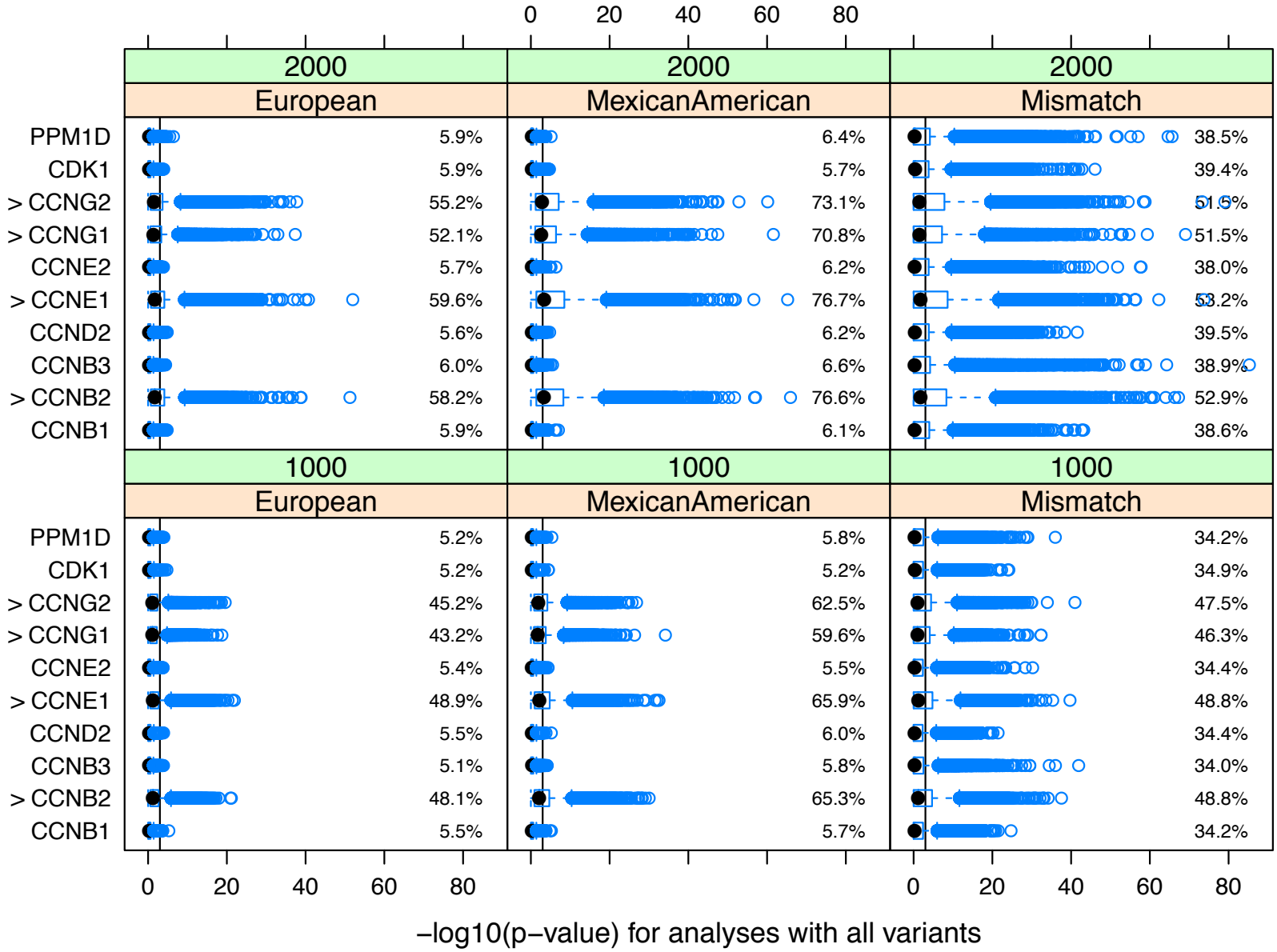
Nature Reviews | Drug Discovery

- 20 genes in the GPCR pathway on chromosomes 6, 8, and 10 and X
- Overlap with 27 isoforms of 15 genes
- Coding regions of these genes range from 563 to 1818 base pairs and represents 16.2% of the total simulated region (17,841 of 110,387 bp)
- Five causal genes

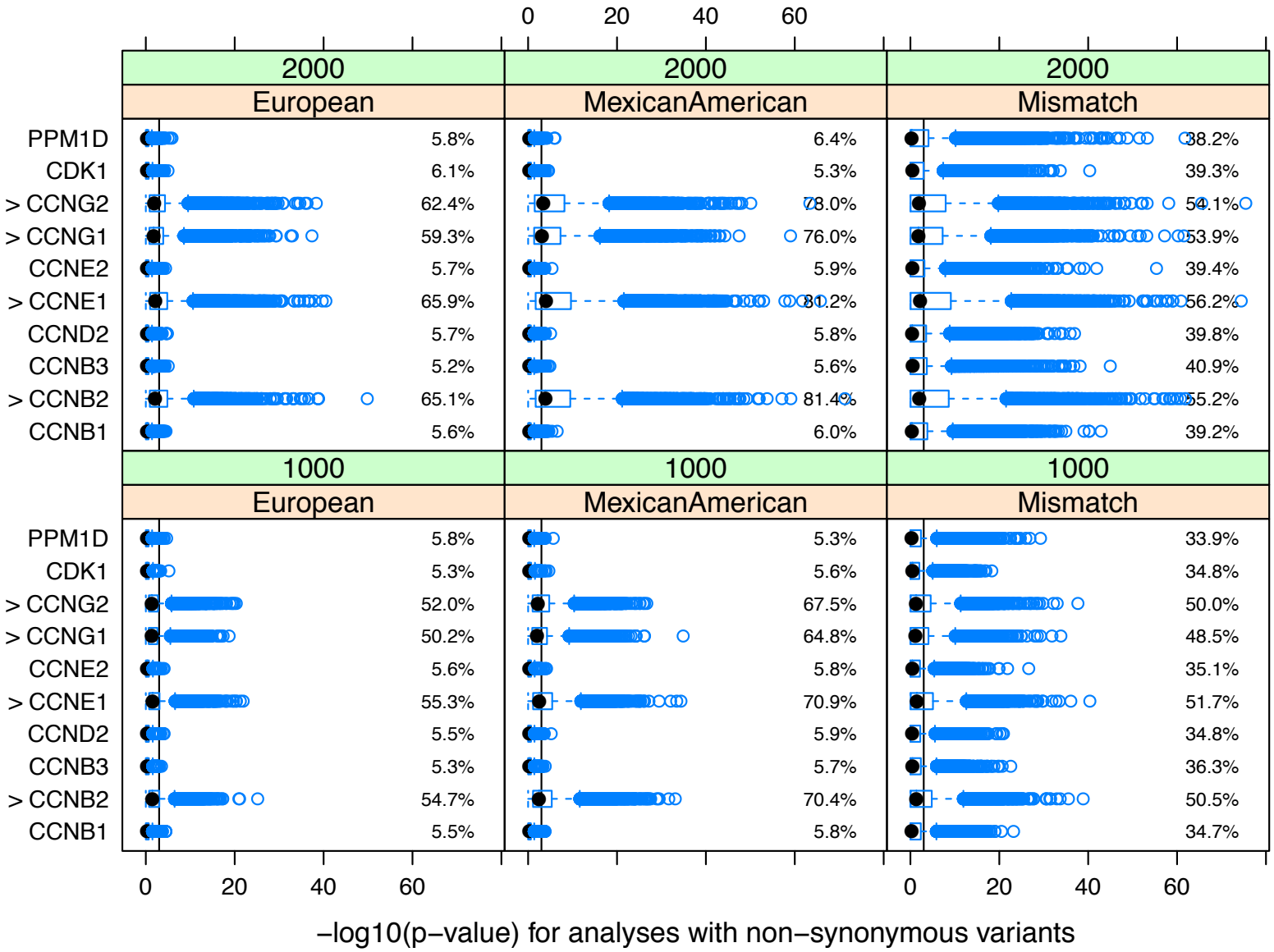
Model details

- Settlement of New World model with AF, AS, EU, MX, and MXL populations.
- K80 mutation model with an ti/tv ratio of 2
- Recombination rates from 6.14×10^{-9} to 6.23×10^{-6}
- Fitness effect of 0.0001, 0.0001, and 0.001 for missense, stoploss, and stopgain mutations
- Draw case controls samples from EU, MXL or EU/AS (mismatch) populations
- Analyze all variants or non-synnonymous mutations (annotated by snpEff)

All variants

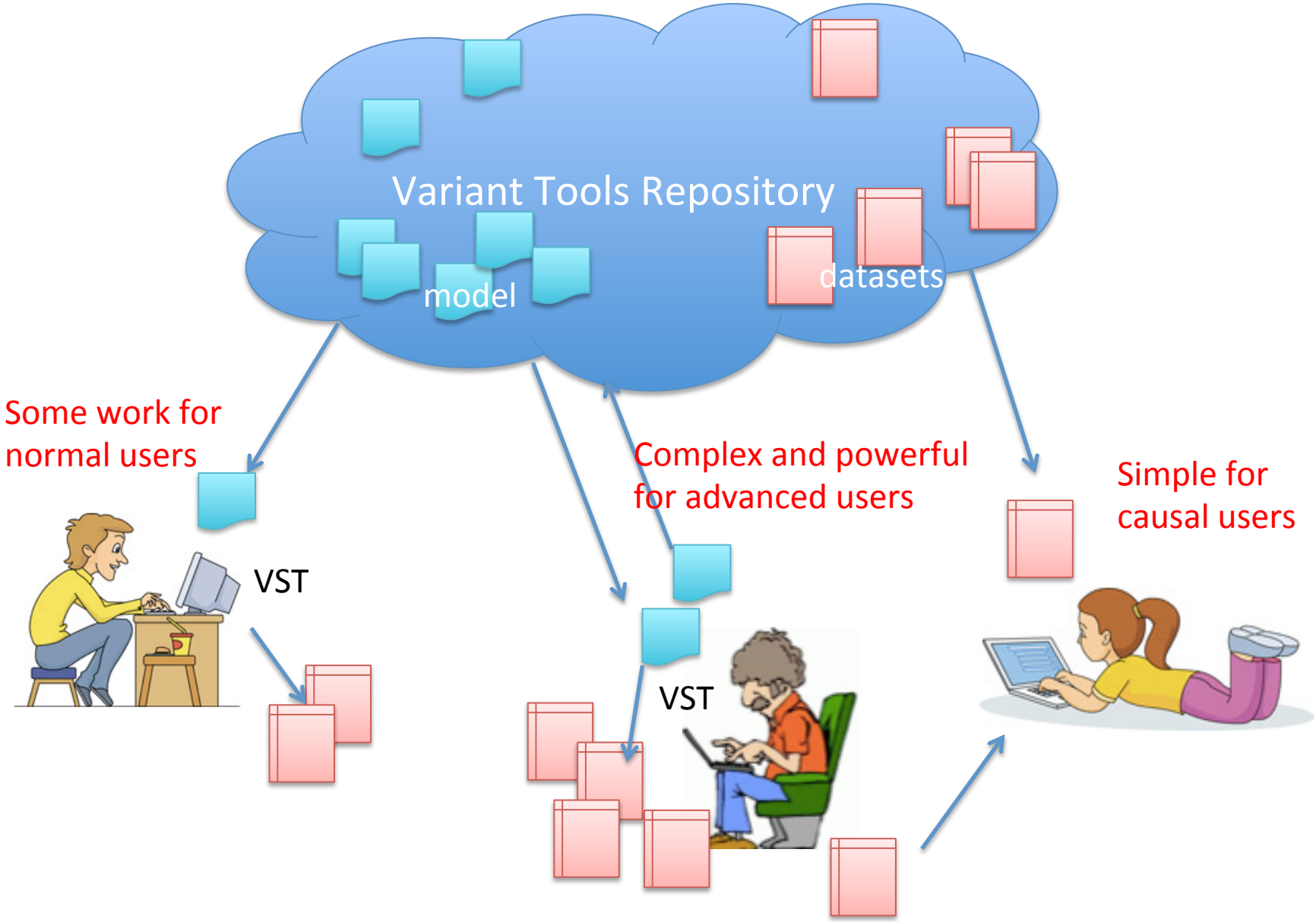


Limit to non-synonymous variants



Discussions

Simple and powerful?



Reproducibility

- Variant Tools repository encourages the sharing of simulation models
- Option --seed to reproduce simulations
- Less option means easier reproducibility
- Available simulated datasets

Limitations

- Limited to models that are provided by authors and power users (but the existing models are already comparable to other single-application tools)
- Performance of the forward-time simulation engine
  - Scaling approach
  - Hybrid simulations

Acknowledgements

- Dr. Paul Scheet
- Gao Wang
- Dr. Biao Li
- Dr. Xiaoming Liu
- Dr. John Weinstein





- Grant 1R01HG005859 (Dr. Paul Scheet)
- The Prevent Cancer Foundation
- The Michael and Susan Dell Foundation
- The Chapman Foundation
- MD Anderson High Performance Computing Cluster