ForSim is a simulation and as a consequence its output results are a produce of particular conditions specified by the user in the input file. The purpose is not to estimate parameters of a particular phenogenetic model to specific empirical data, but instead it is to explore the consequences of those input conditions, which are highly flexible.

However, input conditions could be entirely unrealistic, and when that is the case one can conclude that the conditions are unlikely to be an accurate specification of the causal process and history that generated data that are available. This can constrain speculation about evolutionary history, or aspects of genetic architecture.

What counts as 'plausible' outcome results? This is of course a matter of judgment, but there are extensive data on genetic variation in populations, samples, and species. They include nucleotide diversity, allele frequencies, differences between selected and neutral genes, and haplotype structures. In genetic epidemiology there are extensive results concerning haplotype blocks, length of linkage disequilibrium blocks, family relative risks, and so on.

This Supplemental Information provides a sampler of models and ***ForSim*** results to demonstrate that with reasonable input conditions, based on what we know of human variation, simulation results resemble human genetic data.

Plausibility of results, and hence of input parameters can be measured in various ways.

Some typical values of variation in human genes, basically reflecting neutral evolution at the nucleotide level. In running ***ForSim*** you can compute the neutral evolutionary paramter $4N_e\mu$ from your input file parameters, and then compare them to SNP diversity measures in the output results. With directional or purifying selection, genes should have less diversity. Genes evolving neutrally should approximate the expected diversity.

TABLE 6.1: NUCLEOTIDE DIVERSITY, NEUTRAL EXPECTATION OF θ AND EFFECTIVE POPULATION SIZE ESTIMATES FOR HUMANS.

Locus (Length)	π ($\times 10^{-4}$)	θ ($\times 10^{-4}$)	μ ($\times 10^{-9}$)	N_e	Reference
APOE (5.5 kb)	5.3	6.87 (S)	23.5	7300	(Fullerton *et al.*, 2000)
Chr. 1 (10 kb)	5.8	9.51 (S)	14.8	16 000	(Yu *et al.*, 2001)
Chr. 22 (10 kb)	8.8	13.2 (S)	23	14 400	(Zhao *et al.*, 2000)
X chr. (10.2 kb) Xq13.3	3.6	6.8 (S)	18.4	12 300	(Kaessmann *et al.*, 1999;)
X chr. (4.2 kb) *PDHA1*	–	4.41 (ML)	19.2	7700	(Harris and Hey, 1999)
Y chr. (64 kb)	0.74	2.01 (S)	24.8	8100	(Thomson *et al.*, 2000)
mtDNA (15.4 kb) excluding control region	28	28 (π)	340	8200	(Ingman *et al.*, 2000)
Alu insertions	–	–	–	17 500	(Sherry *et al.*, 1997)

N_e is calculated using locus-specific per-generation nucleotide mutation rates (μ). Among the different studies, θ per nucleotide was calculated using estimators based on a variety of sequence characteristics: S (segregating sites), π (pairwise differences) and ML (a maximum likelihood estimator). These sequence-derived estimates are compared with an estimate from *Alu* insertion polymorphisms.

Page 156, Table 6.1, from Jobling et al. *Human Evolutionary Genetics*

Simulation of neutral genes yields values consistent with these expectations as well as of neutral theory. Nucleotide diversity and SNP sojourn times (from mutation to fixation or loss) are also consistent with theory (see next image)

Here are some sample nucleotide diversity (heterozygosity per nucleotide) results for estimates for ***ForSim*** runs of 10 genes, each of 50 Kb, with no selection. Theoretical expectation is approximately $H=\theta/(1+\theta)$, where $\theta = 4N_e\mu$, $N_e=2N$ for a two-sex population, and μ= mutation rate per nucleotide, at mutation-drift equilibrium. These runs are not long enough to achieve equilibrium, which is stochastic in any case, but the results (which are typical) show that the program achieves expected results in these simple conditions. Selected genes have nucleotide diversity correspondingly less, depending on intensity of selection.

ForSimData150149PM053008Run1
mutation rate = 2.5 e -8.0
effective population = 2000
generations = 10000

4nMu = 0.0002
expected diversity = 0.00019996

ABC1 8.066405e-05
ABC2 0.0001040567
ABC3 0.00019958966
ABC4 0.00015603316
ABC5 0.00010092187
DEF1 0.00024475434
DEF2 6.508879e-05
DEF3 9.590798e-05
DEF4 0.00016525363
DEF5 0.00014359048

ForSimData163210PM053008Run1
mutation rate = 8.5 e -8.0
effective population = 2000
generations = 20000

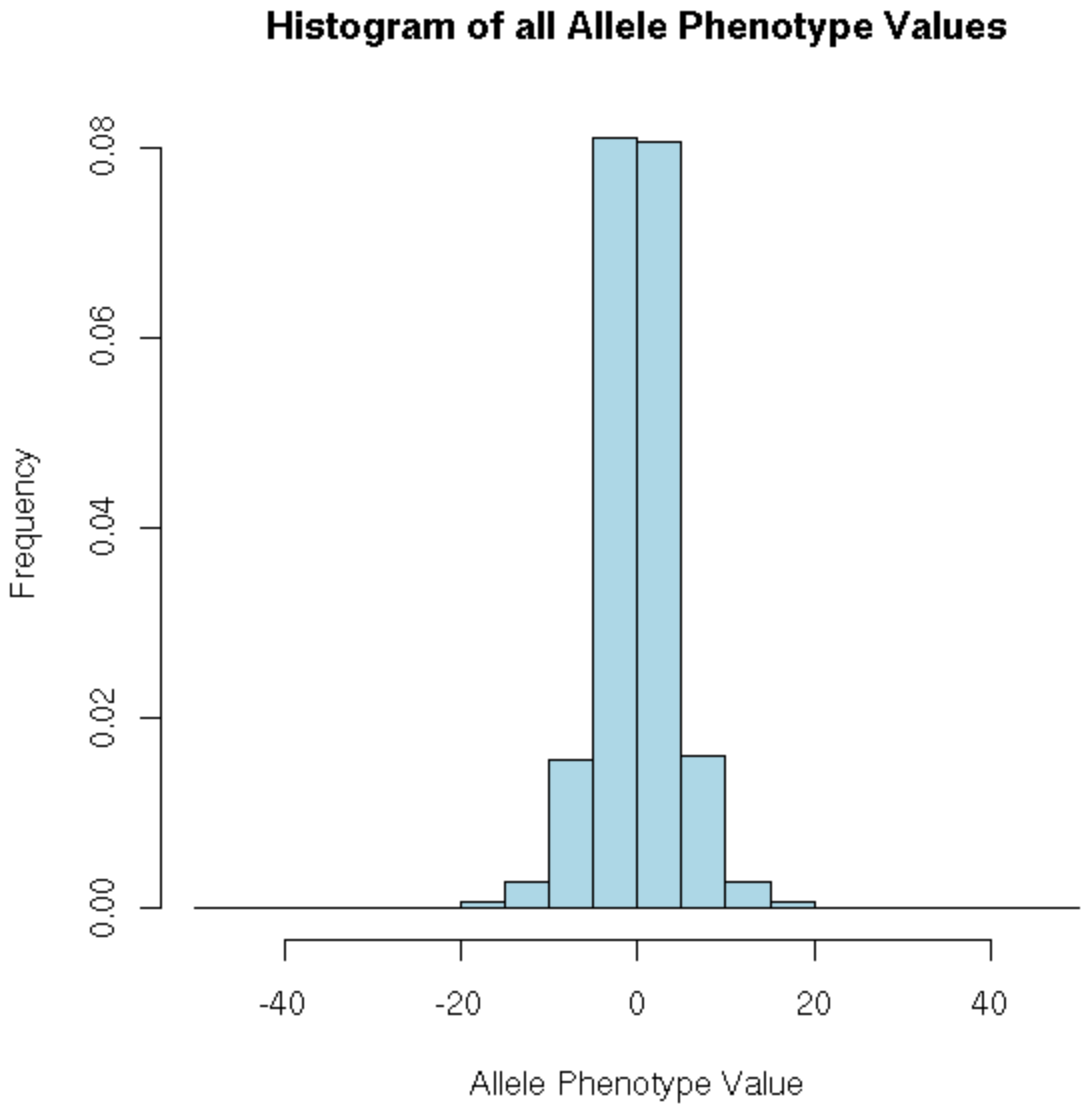
4nMu = 0.00068
expected diversity = 0.000679538

ABC1 0.0003133212
ABC2 0.00031387487
ABC3 0.00066747639
ABC4 0.00057207781
ABC5 0.00034591112
DEF1 0.00072690955
DEF2 0.00035551785
DEF3 0.00035355036
DEF4 0.00042386894
DEF5 0.00027785088

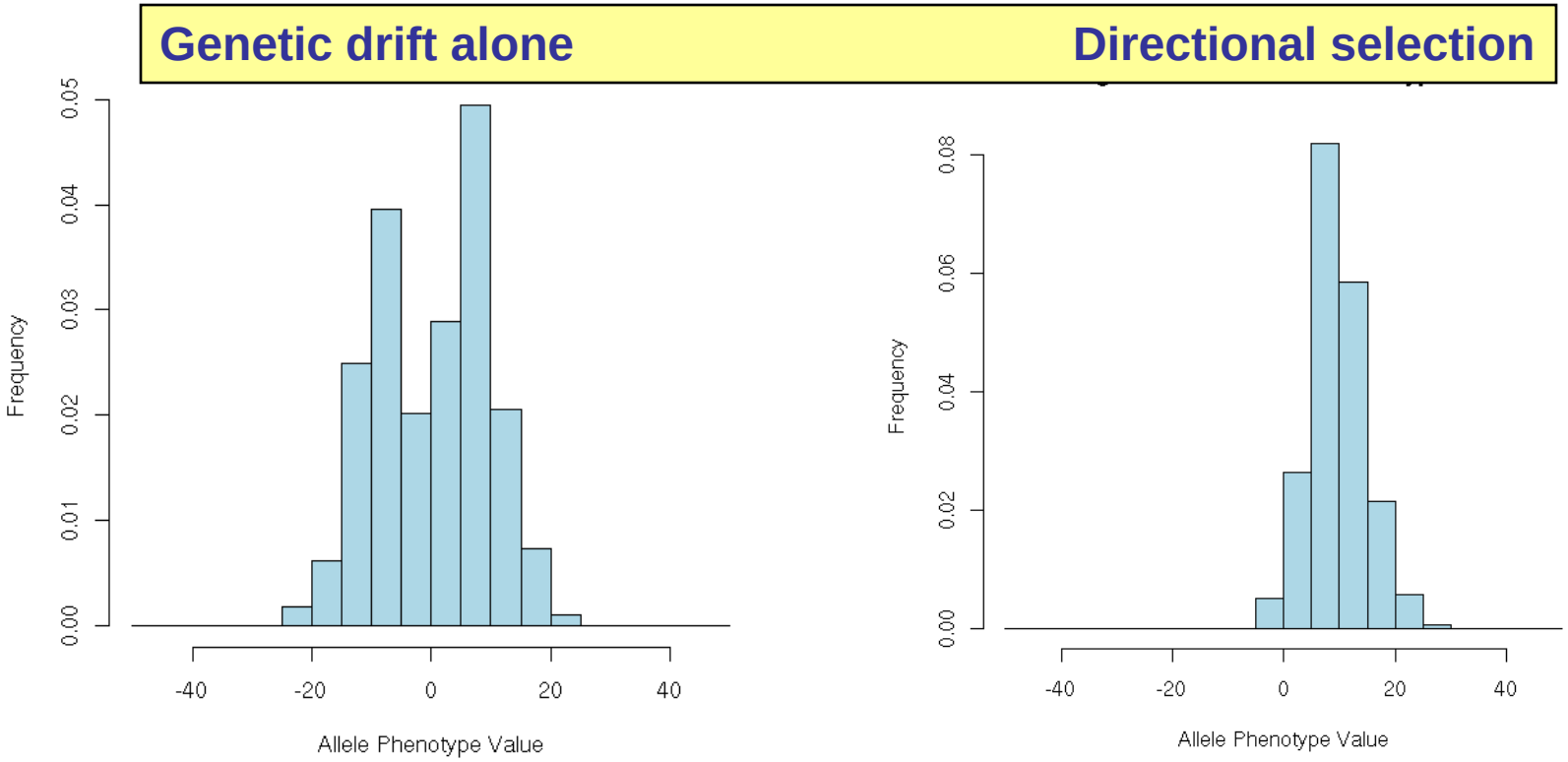
A SAMPLE: CURRENT *ForSim* TEST-RUN PARAMETERS

Populations: 1
Individuals: 10,000 (like human species effective population size)
Time: 10,000 generations (=250,000 human years)
Population growth pattern: stationary
Chromosomes: 1, 1Mb long
Genes: 2 per chromosome, each 1 kb long
Mutation rate: 5×10^{-8} (no hot spots)
Recombination: 1 per Morgan (no hot spots)
Baseline sequence: A,C,G,T random, all 25% at the beginning
Traits: 1 (quantitative)
Genes affecting traits: Eight genes, each affecting the trait
Allelic effects distribution: Gamma(1,3)
Selection: Balancing, truncation-directional, none
Environmental or stochastic trait and selection variance: none

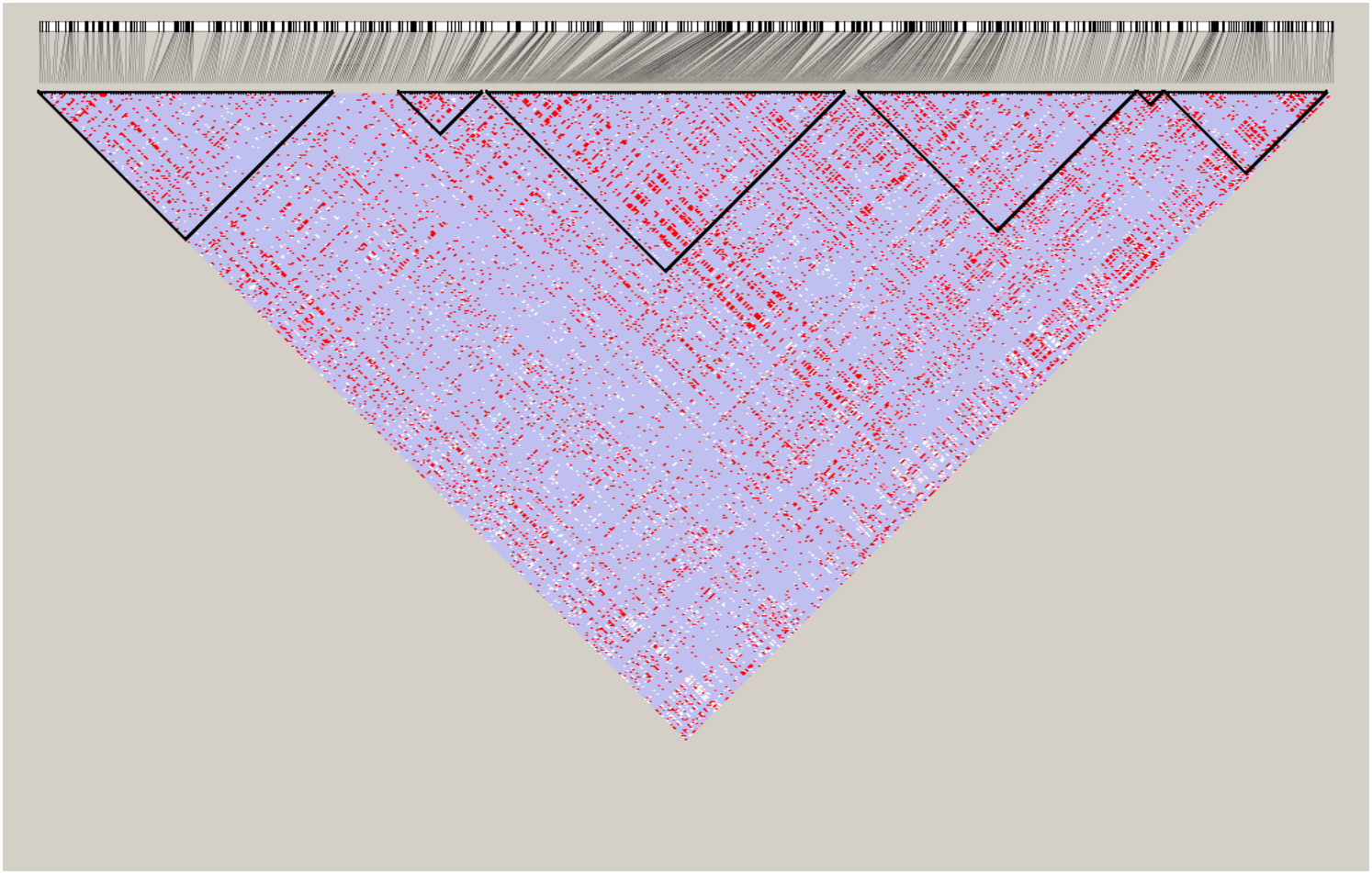
Distribution of phenotypic effects of new mutations (results of Γ(1,3)



Natural selection moves the mean, reduces the genetic variance. One neutral gene, one gene with fitness=0 for individuals <1σ from the current generation population phenotype mean.



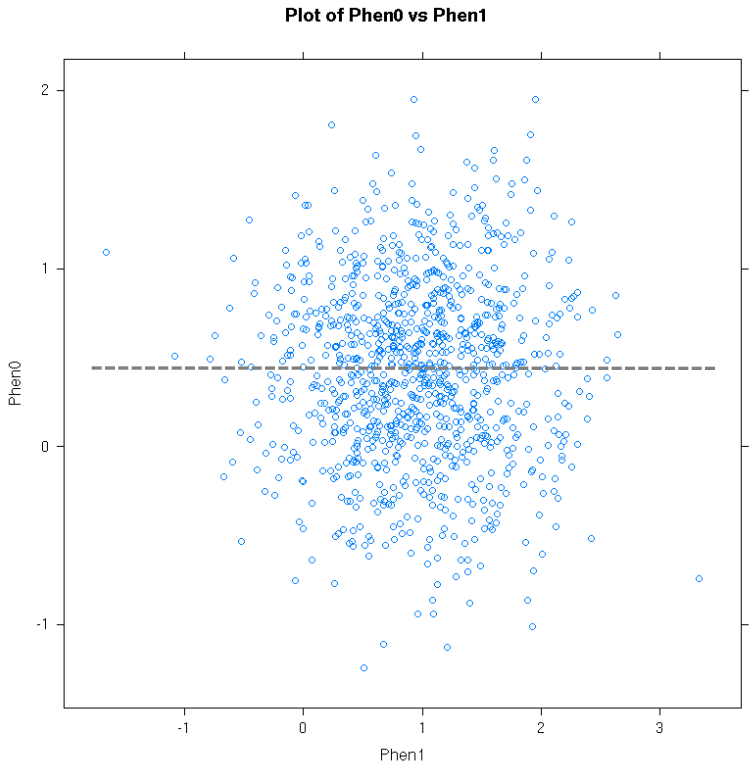
LD distribution along a simulated chromosome; no selection
(the black triangles are LD haplotype blocks as identified by Haploview*)

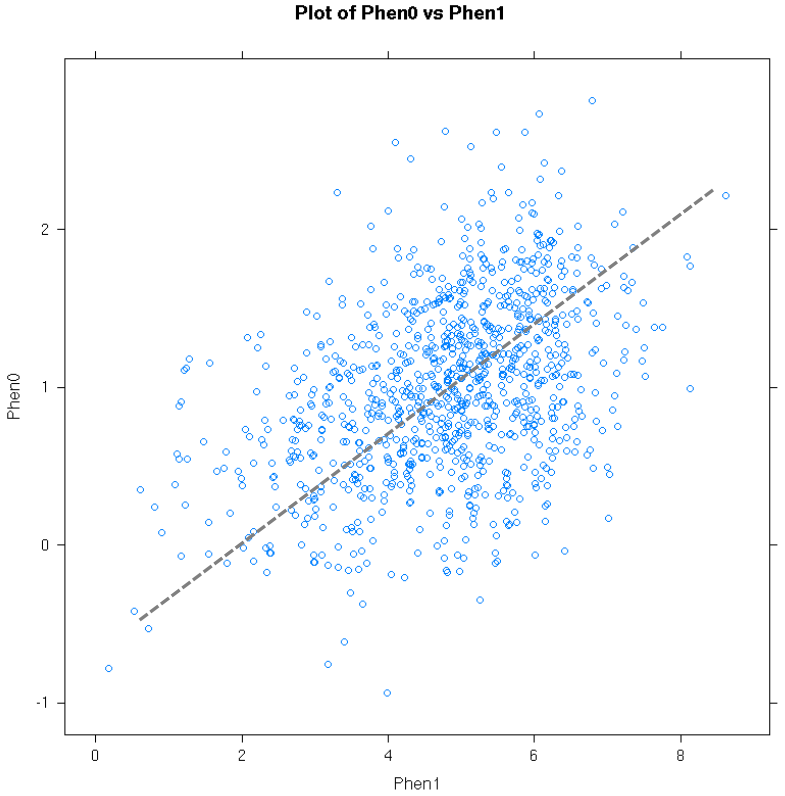


*http://www.broad.mit.edu/mpg/haploview/

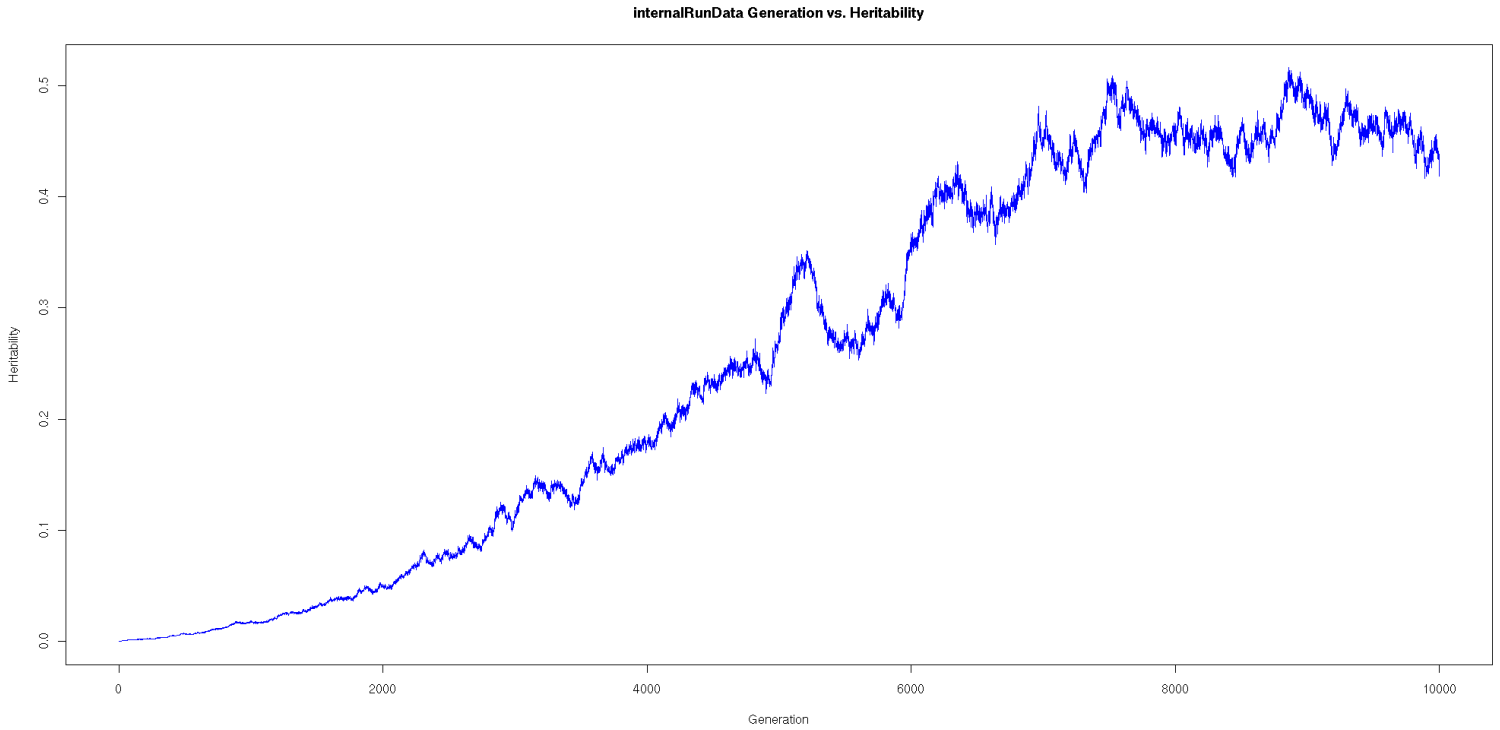
Generation by selection of correlated phenotypes

Two phenotypes, A and B, evolving neutrally. Left, variation in GeneA affects Phen0, while variation in GeneB affects Phen1; the traits are not correlated. Right, same as left except that GeneB affects both Phen0 and Phen1, resulting in correlation between the traits.





Heritability (h^2) plotted per generation, showing the build-up from generation 0 in which there is no genetic variation that affects the simulated trait. Over subsequent generations mutations arise that have phenotypic effects, while environmental contributions remain constant (stochastically imposed on each individual from *Nor*(0,1)). The values plateau (except for stochastic variation) at 45-50%, typical of most complex traits in nature. Short run of 6K generations for a single trait, in a population of 1K. Values depend on parameteric conditions, of course, especially selection, phenogenetic effects of mutations, etc.



ForSim Parameter values for a sample run

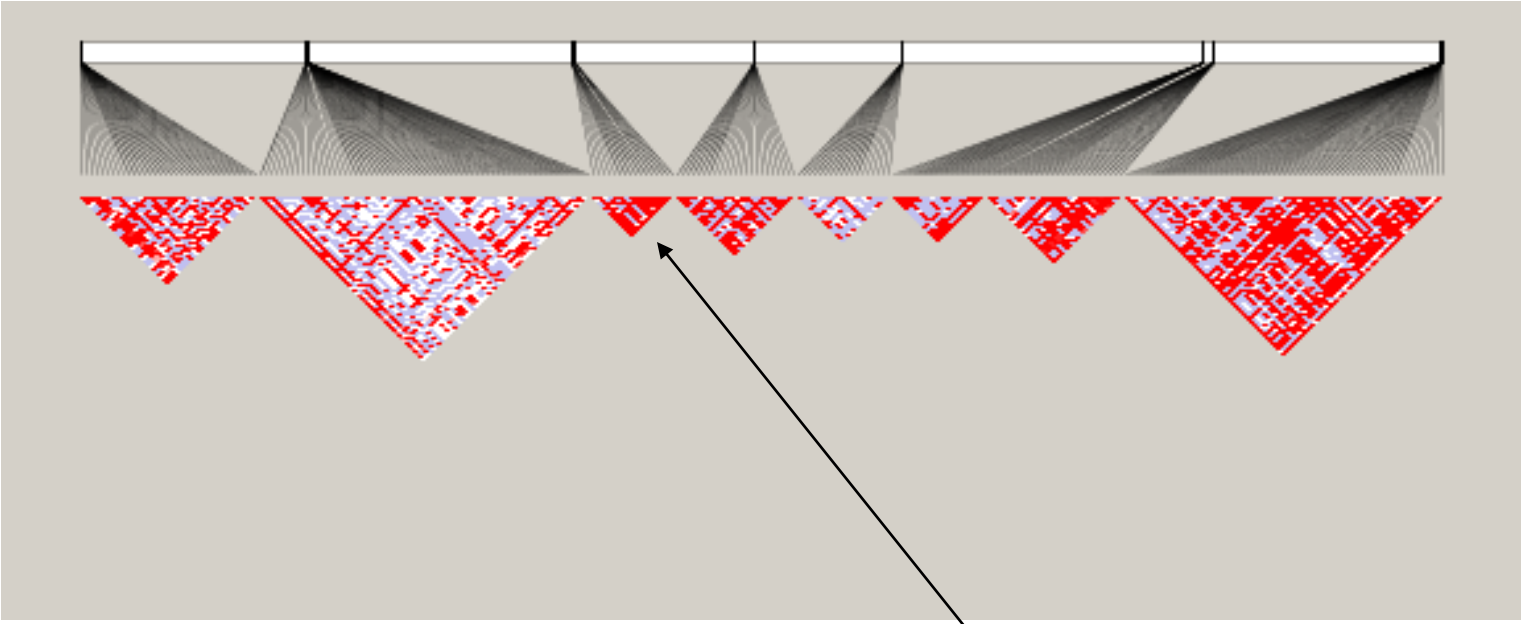
One population (no subdivision)
1 Chromosome of 40 Mb
Simulated for N_e=10,000 (stochastic) for 10,000 generations (250,000 years)
8 Genes: 4 are 20kb, 2 are 40kb long (6 do, 2 don't affect the trait)
Standard human recombination 1%/Mb, mutation 2.5×10^{-8}/base
Stabilizing natural selection that is effectively neutral (Selection against trait value greater than ±3.6 SD from the mean)
Prevalence of 'affecteds': about 7.7%
Sibling relative risk: about 4.3
tagSNPs identified from 30 random trios, using *Haploview*, parameters minor allele freq $\geq$ 0.05; pairwise $r^2 \geq 0.6$, LOD$\geq$3.3 (the typically used parameters)

Risk an Relative Risk from a ***ForSim*** simulation run.
This pattern of prevalence and sibling risk/relative risk values are common in human disease (this is a standard output file relativeRisk.txt)

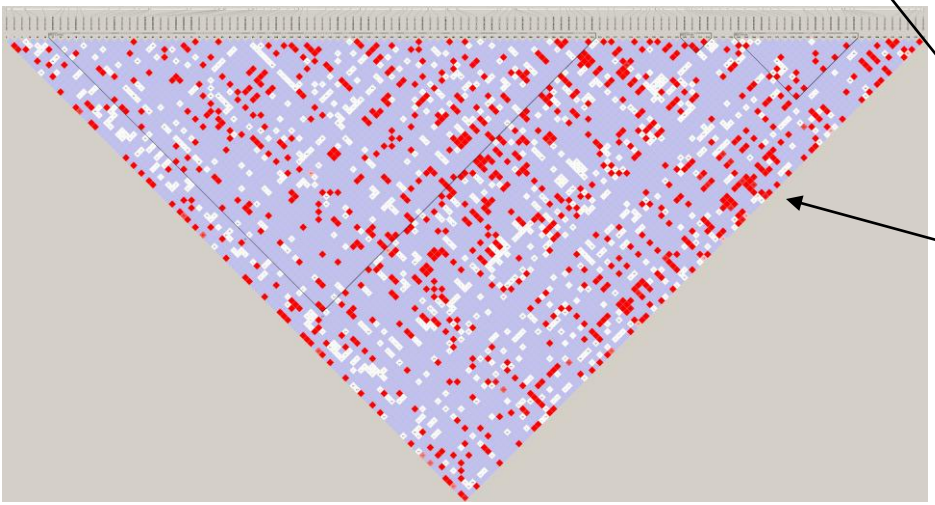
9294 of 101384 offspring in final generation are affected
631 first siblings have affected second siblings in final generation
13362 of 14771 first two siblings have concordant affectation status
first siblings == 14855
first siblings affected == 1448
sibling pairs == 14771
sibling pairs, both affected == 631
overall risk == 0.0916713 (pop prevalence=7.7%; inflated here because relative risk applies to single ascertainment pedigrees)
sibling risk == 0.435773
sibling relative risk == 4.75365

ForSim Identifying LD and taggable LD blocks in output data using HapMap-like simulated data, 30 random parent-offspring trios. Top: Results from Haploview and simulation in previous images. Bottom: from a different run, a gene with weak LD.

LD along a simulated chromosome



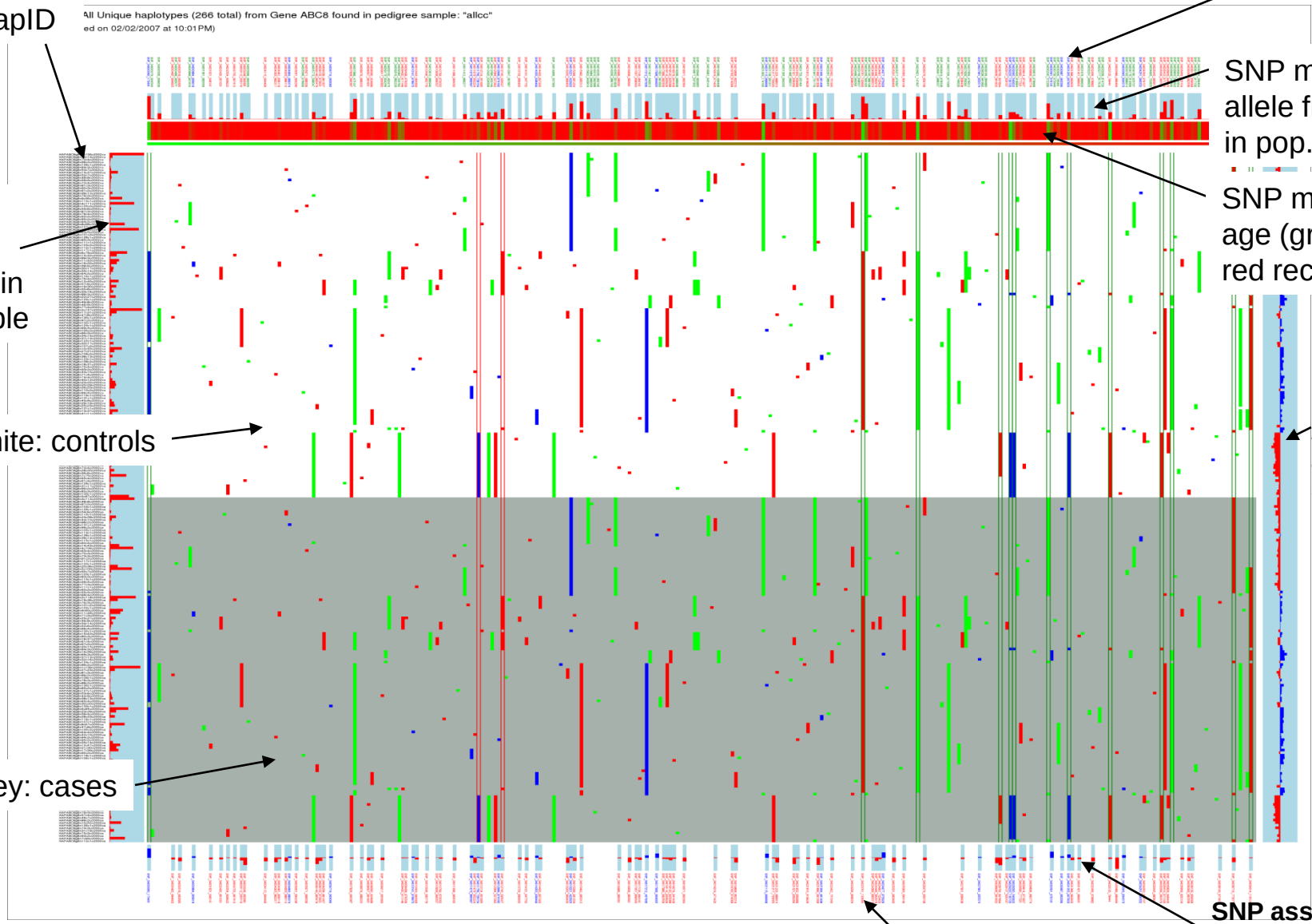
LD within a large simulated gene



Tagging would be useful in this gene

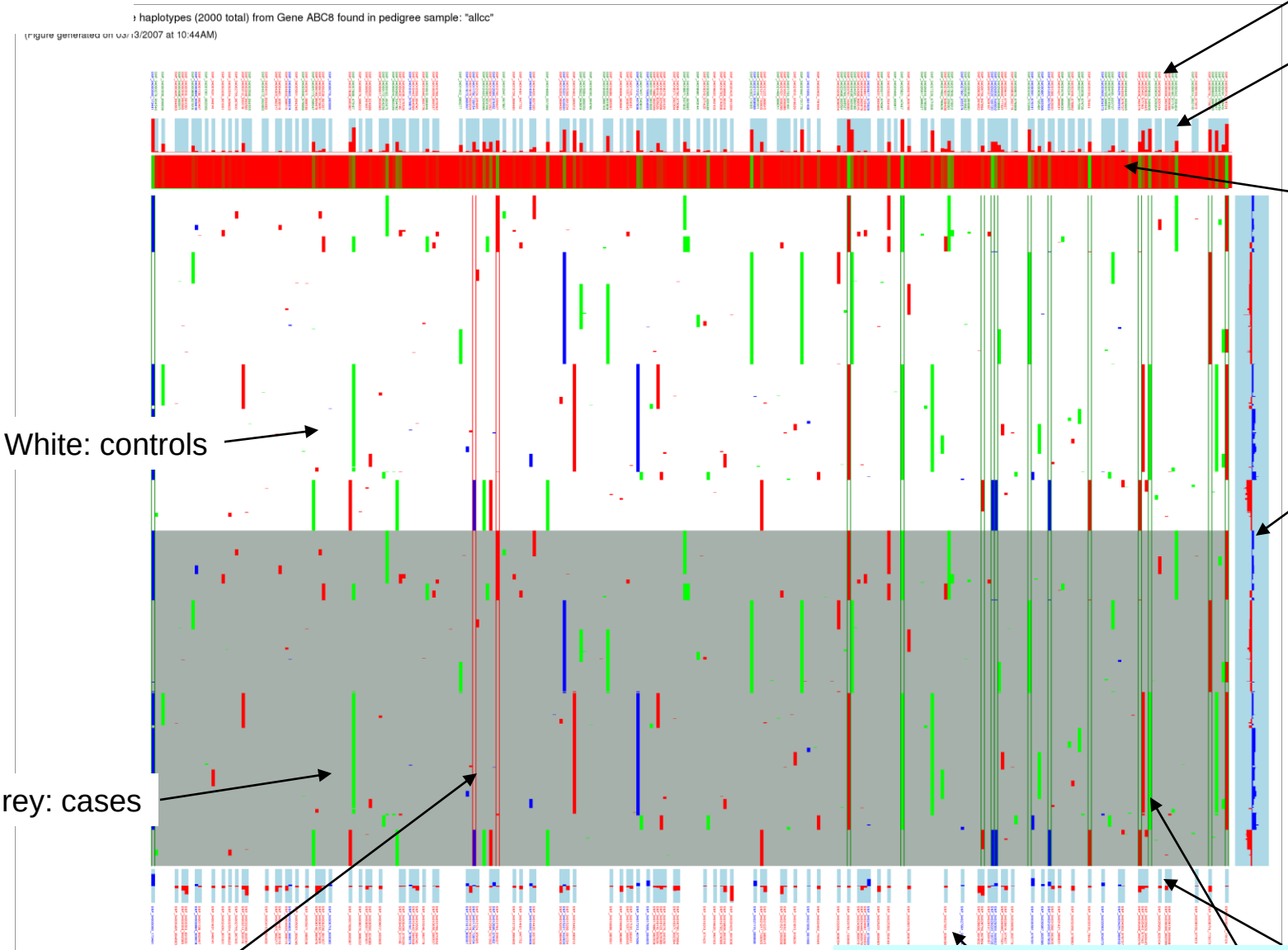
but not this one

Unique haplotypes in cases and controls from this simulation (see user Manual)

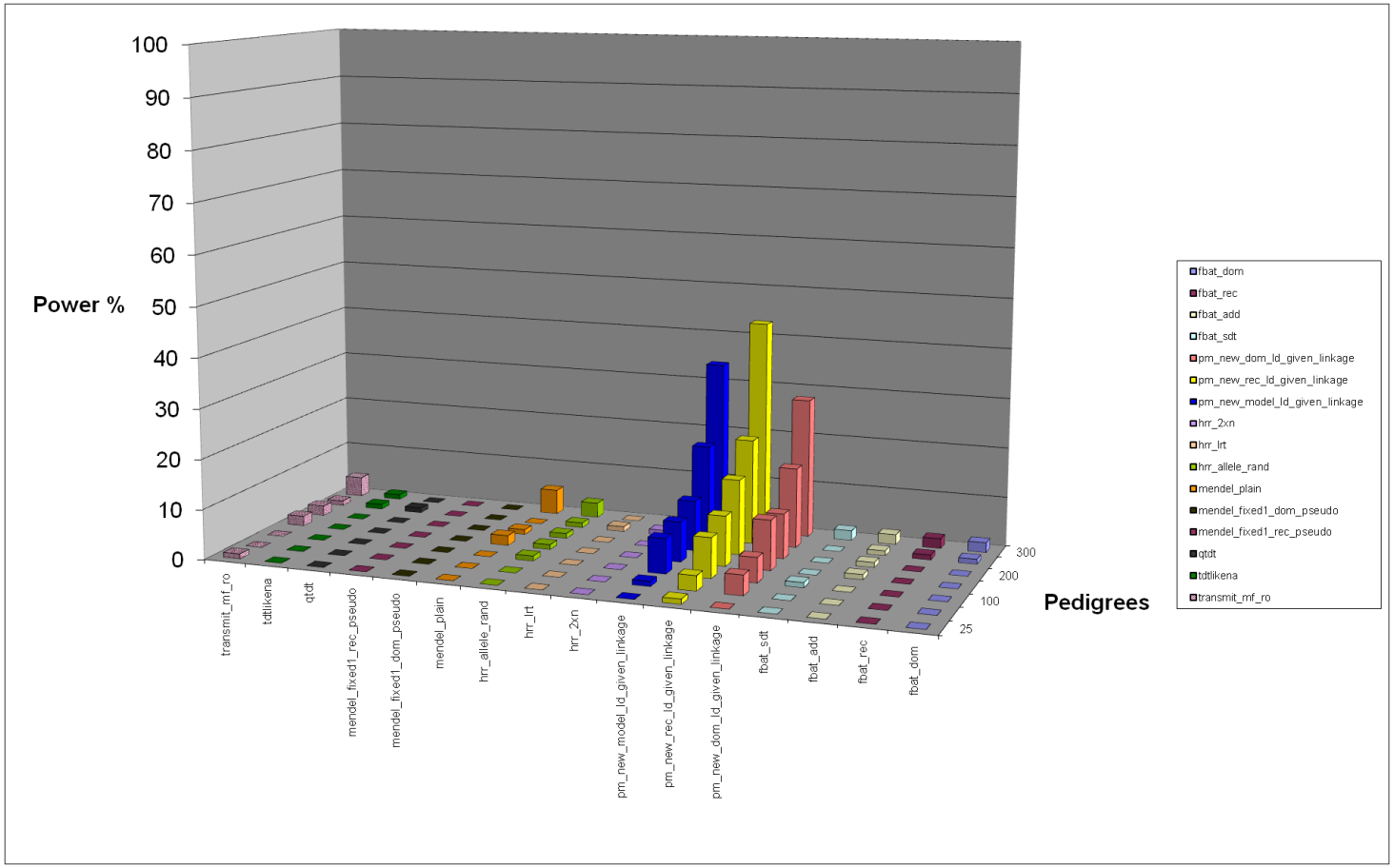


Each row one haplotype. Green: SNP allele no effect
Red: SNP allele negative effect; Blue: SNP allele pos effect

All haplotypes in cases and controls from this simulation

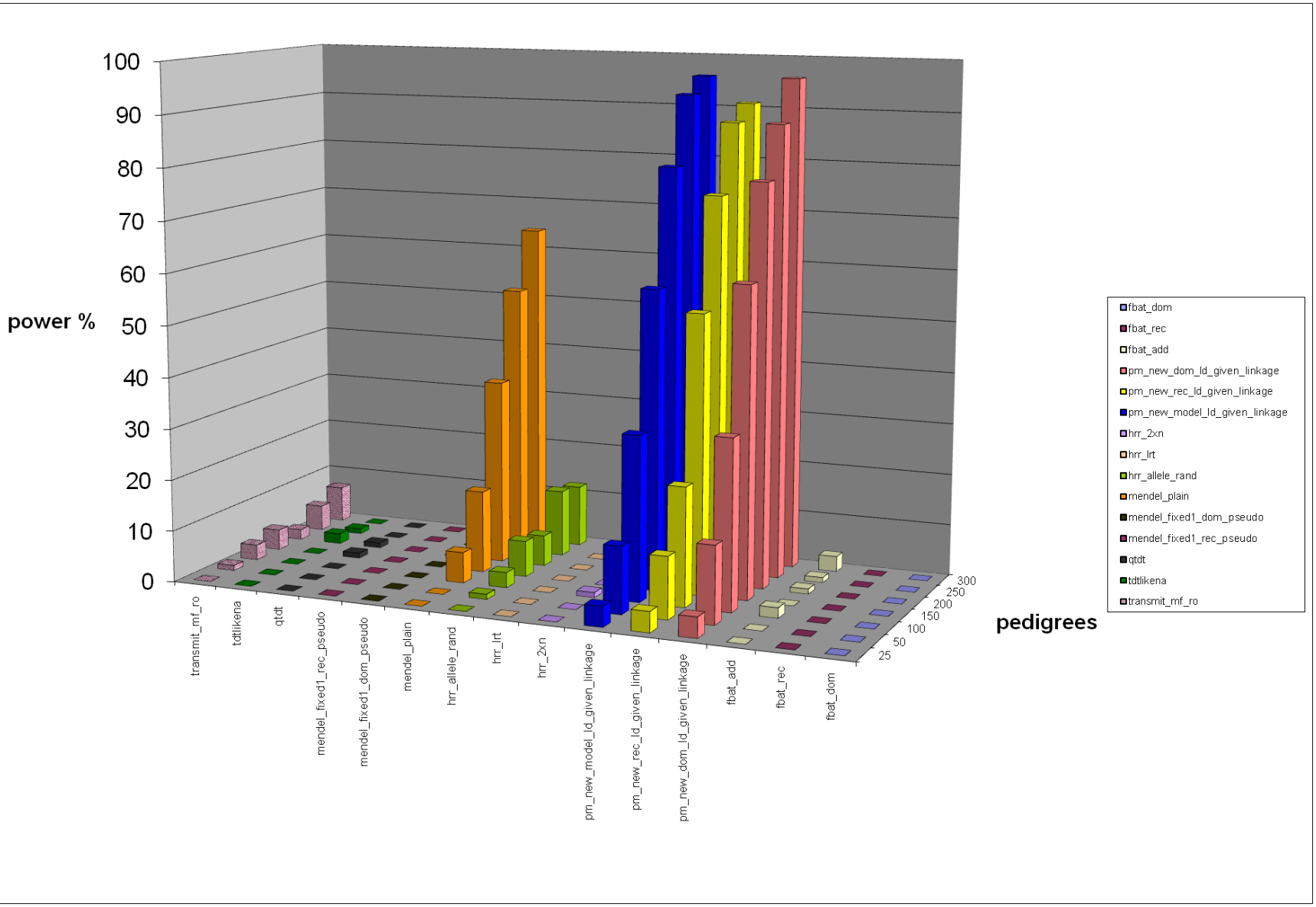


For this simulation, the relative power of one of the tagSNPs to detect phenotypic effects, comparing Terwilliger's Pseudomarker and other standard, available inferential programs, for various causal models, as a function of number of tested 3-generation single-ascertainment pedigrees. Details are unimportant, but show the use of ***ForSim*** output in genetic epidemiological software. TagSNPs were identified by the default Haploview parameter criteria using all SNPs available at the end of the run.

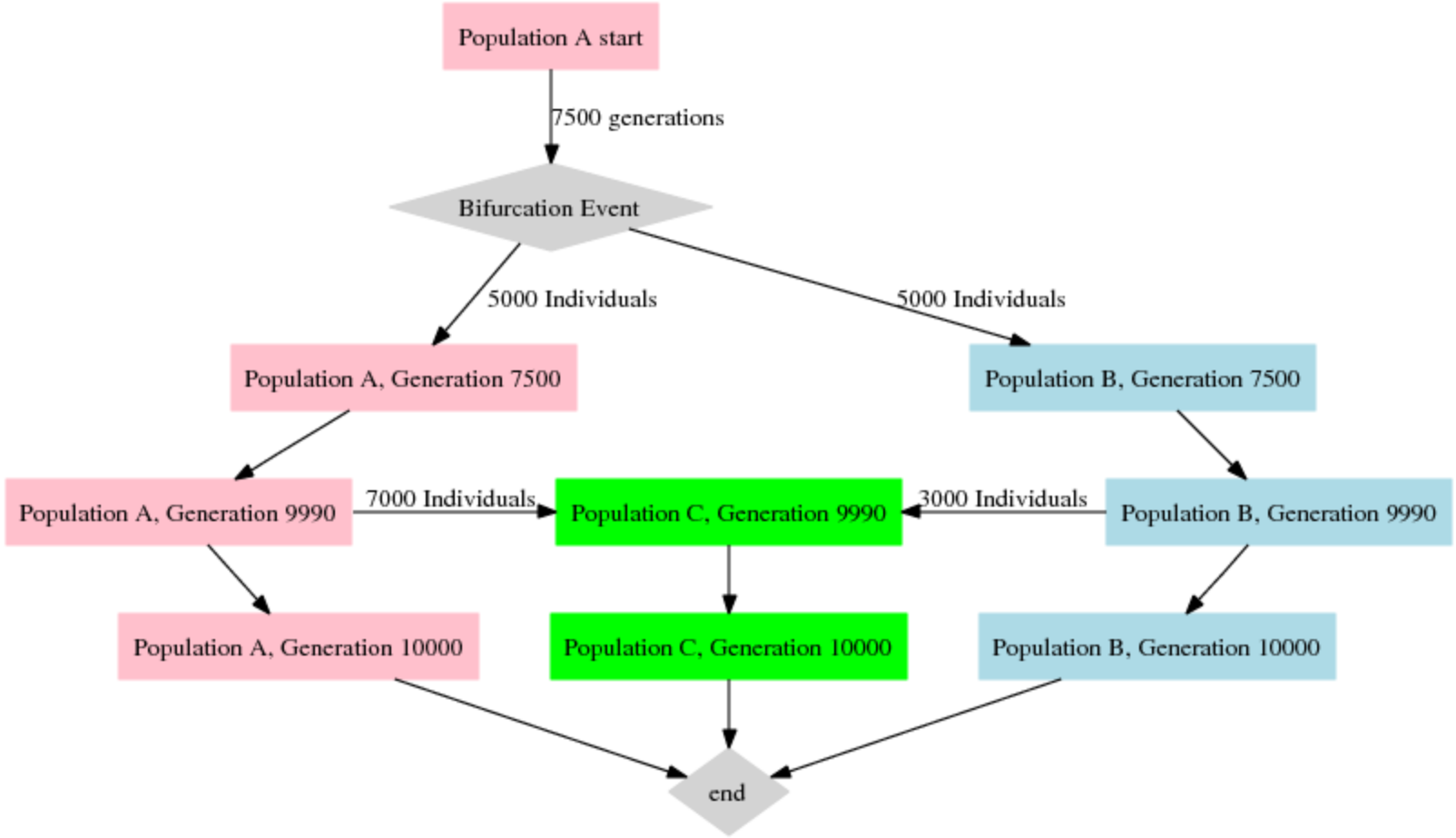


Other programs were: LAMP, Genehunter (TDT implementation), Transmit, QTDT, FBAT, MENDEL, HHRR from ANALYZE, and Pseudomarker (www.helsinki.fi/~tsjuntun/pseudomarker/ tested for dominant, recessive, and general model)

For this run, the relative power of a second example tagSNP to detect phenotypic effects, as in previous image. These two images show that programs have different relative efficacy, which depends on the (uncontrollable) underlying LD and the effectiveness of SNPtagging. Pseudomarker is always more powerful because of its design.



This is the flow diagram for a sample run evolving human continental populations and then creating an admixed population at the end, much as African-American and Hispanic-Americans have been produced. This diagram and the corresponding input file are in the ***ForSim*** user's Manual.



Haploview plot of 30 parent-offspring trios from run described in previous image. Population A and B evolved independently for thousands of generations after common ancestry (much as human Africans and Europeans did) and then Pop C was formed by admixture (30% from Pop B, 70% from A) 10 generations before the end of the run. Pops A, B accumulated largely different SNPs during 2500 gens after split.

Population C shows increased LD due to admixture. Arrows identify genes affecting a simulated trait that was under weak balancing selection, showing their increased LD, which would be useful in SNPtagged association mapping.