

Software

Open Access

GENOMEPOP: A program to simulate genomes in populations

Antonio Carvajal-Rodríguez

Address: Departamento de Bioquímica, Genética e Inmunología. Universidad de Vigo, 36310 Vigo, Spain

Email: Antonio Carvajal-Rodríguez - acraaj@uvigo.es

Published: 30 April 2008

Received: 5 February 2008

BMC Bioinformatics 2008, **9**:223 doi:10.1186/1471-2105-9-223

Accepted: 30 April 2008

This article is available from: <http://www.biomedcentral.com/1471-2105/9/223>

© 2008 Carvajal-Rodríguez; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: There are several situations in population biology research where simulating DNA sequences is useful. Simulation of biological populations under different evolutionary genetic models can be undertaken using backward or forward strategies. Backward simulations, also called coalescent-based simulations, are computationally efficient. The reason is that they are based on the history of lineages with surviving offspring in the current population. On the contrary, forward simulations are less efficient because the entire population is simulated from past to present. However, the coalescent framework imposes some limitations that forward simulation does not. Hence, there is an increasing interest in forward population genetic simulation and efficient new tools have been developed recently. Software tools that allow efficient simulation of large DNA fragments under complex evolutionary models will be very helpful when trying to better understand the trace left on the DNA by the different interacting evolutionary forces. Here I will introduce GenomePop, a forward simulation program that fulfills the above requirements. The use of the program is demonstrated by studying the impact of intracodon recombination on global and site-specific dN/dS estimation.

Results: I have developed algorithms and written software to efficiently simulate, forward in time, different Markovian nucleotide or codon models of DNA mutation. Such models can be combined with recombination, at inter and intra codon levels, fitness-based selection and complex demographic scenarios.

Conclusion: GenomePop has many interesting characteristics for simulating SNPs or DNA sequences under complex evolutionary and demographic models. These features make it unique with respect to other simulation tools. Namely, the possibility of forward simulation under General Time Reversible (GTR) mutation or GTR×MG94 codon models with intra-codon recombination, arbitrary, user-defined, migration patterns, diploid or haploid models, constant or variable population sizes, etc. It also allows simulation of fitness-based selection under different distributions of mutational effects. Under the 2-allele model it allows the simulation of recombination hot-spots, the definition of different frequencies in different populations, etc. GenomePop can also manage large DNA fragments. In addition, it has a scaling option to save computation time when simulating large sequences and population sizes under complex demographic and evolutionary situations. These and many other features are detailed in its web page [1].

Background

There are several situations in population biology research where simulation of DNA sequences is useful. Simulations have been used to for hypothesis testing [2-4], to study the impact of differing demographic scenarios on patterns of human diversity [5], or to simulate the evolution of complex diseases in human populations [6,7]. In addition, population simulation of genetic datasets is also used to estimate population parameters [8-10].

One of the most exciting research areas in the current context of population genetics is the HapMap project. Knowledge about patterns of linkage disequilibrium (LD) in humans is very important from a genomic point of view. The existence of linkage or haplotype blocks [11] or, at least, networks of SNPs in high LD [12], will facilitate the assembly of human genome haplotype maps [13-15] that will enormously improve, among other things, the efficiency of disease gene mapping. It seems that these blocks are mainly defined by recombination hot spots [16,17], but haplotype blocks can also be generated by genetic drift in regions of uniform recombination if rates is low enough [18]. We have now growing empirical knowledge about haplotype block and tagSNP diversity, but less is known about the effect of population demographic history. Though important work has been undertaken in the application of population genetics to LD mapping [19-22] and its relevance to human populations [23-25], we still have an incomplete understanding of how the combined effect of genetic drift, mutation, recombination and migration, affect LD and tagSNP patterns, although it is known that they do [26]. Moreover, recombination is an important evolutionary process to understand how genetic diversity is generated and maintained in populations. Jointly with positive selection, recombination allows for very high rates of evolution [27]. However, the impact of recombination is dependent on other forces, such as selection and demography. Developing tools that allow simultaneous simulation of natural selection, recombination and complex demographic patterns will be of great help in trying to better understand the trace left on the DNA by the different interacting evolutionary forces.

Simulation of biological populations under different evolutionary genetic models can be done following backward or forward strategies. Backward simulations, also called coalescent-based simulations, are computationally very efficient because they are based on the history of lineages with surviving offspring in the current population and ignore all individuals that are not ancestral to the present-day population [20]. Hence, coalescent is a sample-based theory relevant to the study of population samples and DNA sequence data. From its beginnings, the basic coalescent has been extended in several useful ways. For exam-

ple, to include structured population models [28-32], changing population size [33-35], recombination [36,37] and selection [38-43].

On the contrary, forward simulations are less efficient because the entire population is simulated from past to present. However, the coalescent framework imposes some limitations that forward simulation does not. The first of these is the same feature that causes its efficiency, namely, the coalescent does not keep track of the complete ancestral information i.e. only takes into account ancestries that survived to form the present-day sample. Thus, if the interest is focused on the evolutionary process itself, rather than on its outcome, forward simulations should be preferred [44]. Second, coalescent simulations are complicated by simple genetic forces such as selection, and although different evolutionary scenarios have been incorporated (see above) it is still difficult to implement models incorporating complex evolutionary situations with selection, variable population size, recombination, complex mating schemes, and so on. In fact, we can only simulate limited forms of recombination and selection under the coalescent. It is known that recombination has a major impact for detecting positive natural selection [45,46]. Shriner *et al* studied the impact of recombination under a neutral model. Anisimova *et al* studied the recombination effect under a coalescent codon-based model i.e. the unit of change was the codon instead of the nucleotide. In the latter case, recombination was not simulated at the intracodon level. Therefore, we still ignore the importance of intracodon recombination under a given codon-based model. Moreover, coalescent methods cannot yet simulate realistic samples of complex human diseases [6]. Indeed, when simulating non-neutral scenarios and/or complex models under the coalescent, much of its computational efficiency is lost (however, see recent work by Marjoram [47] and Liang [48]). Furthermore, the coalescent model is based on specific limiting values and relationships between some important parameters [49]. Hence, there is increasing interest in forward population genetic simulation and new efficient tools have been recently developed [50-52]. Therefore, a program that allows the simulation forward in time, of different Markovian nucleotide or codon models of DNA mutation combined with recombination, at inter and intra codon levels, fitness-based selection and complex demographic scenarios, will be of great interest. Here I will introduce the program GenomePop that fulfills the mentioned requirements.

Implementation

GenomePop uses a simple and efficient algorithm to perform forward simulation of populations and/or genomes. The basic idea considers an individual as the differences (mutations) between this individual and a reference or

consensus genotype. Thus, each individual is no longer represented by its complete sequence or genotype but by the mutations it carries with respect to the consensus. A more detailed explanation of the algorithm is provided at the program web page. Taking advantage of the efficiency of this approach, GenomePop can simulate, forward in time, DNA sequences under specific Markov models. The program allows the simulation of recombination under both nucleotide and codon models of evolution, providing a way to simulate recombination at inter and intracodon levels under codon models. It also permits arbitrary migration models, simulation of SNPs, recombination hot-spots, fitness-based selection and many other features that are detailed in the program web-page. GenomePop has different output formats as GenePop for SNPs and Phylip or Nexus for DNA sequences.

Markov models of DNA mutation

Markov processes are used in molecular evolution to describe the change between nucleotides, aminoacids or codons over evolutionary time. Usually, time is measured as the number of substitutions because molecular sequence data does not allow the separate estimation of the rate and the time, but only of their product [53]. In the context of forward simulation we are not interested in the transition after an arbitrary time *t* (branch length) but just in the transition from a nucleotide or codon to another, given that a mutation occurs. An advantage of this approach is that we need to compute the transition matrix just once at the beginning of the evolutionary process. Therefore, consider a given instantaneous substitution rate matrix *Q*, which allows for a complete definition of any Markovian substitution model [53], the matrix $M = -qQ + I$ is the conditional transition matrix to go from *i* to *j* provided that a substitution occurs, where *q* = diagonal (1/*q_i*) and *I* is the identity matrix [54]. Then, given an instantaneous substitution matrix *Q*, estimated for example using PAUP [55] or Hyphy [56] programs, we can obtain the corresponding transition matrix *M* that can be used to produce the necessary mutation process in a forward in time evolutionary model.

Biological models

There are two basic biological models implemented in GenomePop, namely "viral" and "non-viral". The only

difference that distinguishes them is just that in the viral model the initial sequences are different in each population, as the different viruses infect different individuals. Thus, the user can define a viral model indicating the percentage of sequence identity (0–100) between the sequences of the distinct populations. By default the sequence identity is zero i.e. the sequences at each population are randomly settled. In the non-viral model the initial sequence is the same for every population (identity of 100%).

DNA models, recombination and selection

There are different DNA models implemented in GenomePop (Table 1). In any of them, the user can decide to allow recurrent mutation, i.e. multiple site hits or not. Models can be haploid or diploid. Population size can be constant or variable. In the four-allele models, the sequences can be generated by the program or provided by the user. In the case of the 2-allele model (SNPs) just one or several chromosomes can be considered. In this same model, recombination can be constant or a hot spot recombination model can be defined. In the latter, the recombination rate *r* is per haploid region and generation. If no hot spots are defined, the expected number of recombination events between any two sites *i* and *j* will be $2rd_{ij}/(L-1)$ where *d_{ij}* is the implied region length and *L* is the chromosome length. The number of recombination events between the two chromosome extremes 0 and *L* - 1 will be $2rd_{ij}/(L-1) = 2r$. In GenomePop, the effect of natural selection can be modelled in two different ways: 1) by its effects on the *dN/dS* ratio i.e. by defining a codon model, and 2) via the fitness effect of mutation on specific loci. The user can run either of two models. The codon model option runs a MG94 codon model [57] with a given *dN/dS* combined with any defined nucleotide model. This model of codon evolution will be implemented by the instantaneous rate matrix to go from codon *i* to *j*. That is, $Q_{ij} = \theta_{mn}k\pi_n$ where θ_{mn} accounts for biased nucleotide, *m* to *n* substitutions; *k* = 1 or ω for synonymous or nonsynonymous mutation rates respectively and π_n is the equilibrium frequency of the target nucleotide. This corresponds to the MG94 model [57] with the restriction of $\alpha = 1$. Nucleotide equilibrium frequencies are used instead of codon frequencies. To simulate a given *dN/dS* we simply set $\omega = dN/dS$. Alternatively, the user can set the

Table 1: GenomePop DNA models

| DNA Model | GenomePop Notation | Output format | Recombination | Selective sites |
|-----------------|--------------------|---------------|---------------|-----------------|
| 2 allele | JC2 | Genepop | Hot spots | Yes |
| Jukes Cantor | JC4 | Phylip/Nexus | Constant | Yes |
| GTR | GTR | Phylip/Nexus | Constant | Yes |
| MG94 × (JC/GTR) | Codon true | Phylip/Nexus | Constant | Yes |

GTR: General Time Reversible Model [63]. MG94: Muse and Gaut [57] codon model.

codon model option to false (default option) and define specific sites under directional selection with a given selective coefficient which will apply when a mutation occurs at such site. The user can also force all sites to undergo selection. The selection coefficient, s , can be constant or sampled from a gamma distribution with user-defined shape parameter β and scale parameter β/s . The β parameter allows for modelling of the fitness effects distribution, e.g. a low value of β (0.1) will sample many mutations with low effect and few with high. A β parameter of 1 corresponds to the exponential distribution. If we set β to 0 then a constant effect model is applied. Moreover, GenomePop permits the combination of both kinds of models of selection, codon and fitness-based, though the biological meaning of such a mixture is not clear.

Migration models

Two basic migration schemes, island model and one-dimensional stepping stone, are pre-defined in GenomePop. However, the user can define any migration model of interest (Figure 1). To do this, set the flow model to 'user' in the standard input file and then just introduce a scheme similar to that of Figure 1 in a file called Migration-Model.txt. In this file, the lines beginning with '#' are comments. To indicate how individuals will migrate from a given population just begin the line with the word "pop". The order of appearance of each population in the file will correspond with its index i.e. the first population that appear is the population number one, etc. The number below "pop" refers to the migration level, i.e. the number of different migration rates defined from this population. The next line should begin with a migration rate (between 0 and 1) followed, in the same line, by the target popula-

tion(s). We should have as many of these kinds of lines as the migration level indicates, i.e. if the migration level is 2 we should have two lines beginning with a migration rate. More detailed explanation and specific examples are given in the program web page.

Scaling

Clearly, the more complex the model defined, the slower the simulation. To avoid high computation times, GenomePop incorporates a scaling option based on the fact that, under neutral models, we can scale the population size N and the time t , provided the consequent correction to the mutation (μ), migration (m) and recombination (r) rates holds the corresponding compound products $N\mu$, Nr , Nm , etc., constant.

Results

Input file

The input file should be called GenomePopInput.txt. In this file, lines beginning with '#' are comments and will be ignored. In Figure 2 we can see an example of an input file. Note that the input is flexible, i.e. the minimum input for GenomePop to work appropriately corresponds to the first line and the values below it. This line must begin with the identifier 'chromsize' and the line below with the corresponding desired values. Note that, in lines with identifiers, only the first word matters for the program.

Thus, the input in Figure 2 generates 100 datasets under a GTR model with substitution rates typical for HIV [58]. Both recurrent and retromutation are allowed. The system will evolve 1 chromosome of 1 Kb under the given model over 20,000 generations. As can be seen in Figure 2, a scal-

```
# Population sending emigrants
Pop
# the number of different migration rates from this population
2
# migration rate 0.01 from this population to population 2
0.01 2
# migration rate 0.001 from this population to populations 3 and 4
0.001 3 4

# The second population sending emigrants to population 1
Pop
1
0.01 1
```

Figure 1
Example of a user-defined migration model.

ing of 10 was used, which implies that both, population size and the number of generations, was divided by 10 and mutation was multiplied by the same factor. A more exhaustive explanation of the input facilities of GenomePop is provided at the program web page.

Example and validation of the Markov mutation method

For each obtained dataset from the input in Figure 2, the best-fit model of nucleotide substitution under the Akaike information criteria (AIC) was estimated with Modeltest v3.6 [59], using maximum likelihood (ML) estimates from PAUP* [55]. The percentage of correct model estimation (GTR) was 97% although some datasets, about 29%, were also assigned invariable sites or rate heterogeneity among sites. The substitution pattern and equilibrium frequencies were correctly estimated.

Examples and validation of other general features

As GenomePop has many different features and models it is difficult to validate every possibility or circumstance. However, strong effort has been made to validate the program as thoroughly as possible. For example, both unscaled and scaled simulations were performed under a Jukes-Cantor model with diversity $\theta = 4N\mu = 0.004$ over 10^4 generations and then θ was estimated using the finite-sites correction of Watterson θ [60]. The accuracy was quite good, obtaining estimates of 0.0043 ± 0.00015 and 0.0037 ± 0.00016 for the unscaled and scaled cases

respectively. Recombination was also tested by evolving datasets for $6N$ generations under a Jukes-Cantor 4-allele model with different values for the parameter $\rho = 4NrL$, where N is population size, r is recombination rate per site and L is the DNA sequence length (the corresponding parameter in GenomePop is 'Rec' = $r \times L$). Namely, we ran cases with ρ equal to 0, 50 and 100. Recombination was then accurately estimated using the program Kpairwise [58]. GenomePop allows also studying 2-allele SNPs at different frequencies in different populations. In Figure 3 we define a 2-allele model (JC2) with different initial composition at each population (viral model) and 10 independent SNPs (recombination 'Rec' = $10 \times 0.5 = 5$). The populations have different sizes (100 and 120) and migration occurs under the island model. Note that when defining different population sizes, the original population size provided in the 'chromsize' line under the 'pop-sizeKmax' identifier is overwritten.

We ran this example over 200 generations and then analyze the output with the GenePop 4.0 program [61]. As expected the SNPs were detected as independent. We then changed the value of recombination to 0 ('Rec' = 0) and then GenePop 4.0 tell us that the 10 SNPs are linked, as expected. Note the many possibilities that the program provides in the context of studying SNPs under complex evolutionary situations. We can define any number of populations under any user-defined migration model. We

```

chromsize  numcr      popsize  Npops   gen      HaploidGenomeMutRate  Rec
1000      1            1000    1       20000    0.1                    0.0

scale
10

recurrentmut  retromutation
true          true

runs  diploid  constantMetapopSize
100   true      true

sample size
50

model
GTR
rates
3.0 5.0 0.9 1.3 5.3 1.0

# the first three equilibrium freqs: A C G
freqs
0.35 0.15 0.25
    
```

Figure 2
Input file to generate 100 datasets under a GTR model.

```

# Two populations with 10 independent SNPs fixed for different alleles each.
# NOTE that the initially given population size are overwritten to be distinct and of size 100 and 120

Biological Model
viral

chromsize   numcroms   popsizeKmax  Numpops   maxgen   HaploidGenomemutRate  Rec
10          1          1000        2         200     0.000001              5

# the next line is unnecessary, with 2 pops island = 1dim-stepping stone (default model)
flowmodel
ISLAND

migration
0.01

recurrentmut
true true

runs          diploid   constantMetapopSize
1            true      false

# the following line overwrites the popsizeKmax value for each population
# if, for example we change the pop2 size to be 12000 we will see the invasion of allele 2 in the pop 1

different pop sizes
100 120
sample size
20
model
jc2
# next line fixes pop 1 with all 1's and pop2 with all 2's
SNP freqs
1.0 0.0

```

Figure 3
Input file to generate 10 independent SNPs at different frequencies in different populations.

can set any number of SNPs with the desired linkage relationships. The SNPs can be set at distinct initial frequencies in the different populations, for example, 'SNPfreqs' at 1.0 and 0.0 defines the first population with allele 1 fixed and the second with allele 2 fixed.

Impact of recombination on estimation of positive selection

We performed a simple experiment to test the impact of recombination on dN/dS estimation. We ran 50 replicates, with and without population recombination per gene, $4Nr = 40$ and 0 , respectively. The runs were performed under a MG94 \times JC model both with $dN/dS = 1$ and $dN/dS = 2.5$ evolving 333 codons for $10N$ generations with an effective population size of $N = 10^3$ to get samples of 20 sequences. The dN/dS ratio was estimated with the FEL (Fixed effects Likelihood) model of Hyphy [62] which computes global and site by site dN/dS ratio. A p value of 0.1 was used to infer sites under positive selection. As can be seen in Table 2 a dN/dS of 2.5 provokes the detection of some sites under positive selection (1 or 2, not shown)

in only 30% of the replicates (NSS = 0.3 in Table 2). Furthermore in the strictly neutral case ($dN/dS = 1$), one positive selected site was assigned in 10% of the replicates as expected given the p value used. If we correct by this 10% of false positive tests then positive selected sites were detected only in 20% of the replicates under a dN/dS value of 2.5 and no recombination. This is in agreement with the conservative nature of the FEL method [62]. Also noteworthy is that recombination had no impact on global dN/dS estimation but had important effects on the number of sites detected under positive selection as is evident upon inspecting Table 2. It seems also that the effect of intracodon recombination is negligible. Interestingly, it appears that the effect of recombination is somewhat higher under non-neutral dN/dS than in the neutral case. The impact of recombination on positive selection detection has already been studied [45,46]. However, as far as we know, the comparison of the impact of recombination under neutral or positive dN/dS jointly with the effect of intracodon recombination has never been studied before. The significance of this effect should be studied with more

Table 2: Impact of recombination on dN/dS estimation under a Jukes Cantor model.

| 4Nr | Expected ω | Estimated ω | NPSS |
|--------|-------------------|--------------------|-------------|
| 0 | 1 | 1.02 ± 0.03 | 0.1 ± 0.05 |
| 40 | 1 | 1.06 ± 0.04 | 9.9 ± 0.56 |
| 40 ncb | 1 | 1.01 ± 0.03 | 8.8 ± 0.49 |
| 0 | 2.5 | 2.62 ± 0.12 | 0.3 ± 0.07 |
| 40 | 2.5 | 2.57 ± 0.11 | 13.1 ± 0.77 |
| 40 ncb | 2.5 | 2.58 ± 0.13 | 12.7 ± 0.65 |

N: Population size. r = Recombination rate per gene. $\omega = dN/dS$.
NPSS: Average number of positive selection sites. ncb: no codon break allowed.

replicates and cases, which is out of the scope of the present work.

Conclusion

GenomePop has interesting characteristics for simulating SNPs or DNA sequences under complex models of evolution and demography. These features make it unique with respect to other simulation tools. Namely, the possibility of forward simulation under GTR mutation or GTR × MG94 codon models with intra-codon recombination, simulation of any user-defined migration pattern, diploid or haploid models, constant or variable population sizes, fitness-based selection, etc. Under the 2-allele model it allows the simulation of recombination hot-spots, the definition of different frequencies in different populations, etc. GenomePop can also manage large DNA fragments and has a scaling option to save computation time when simulating large sequences or population sizes under complex demographic and evolutionary situations. It has many other features that are detailed in the web page [1].

Availability and requirements

Project name: GenomePop v. 1.0

Project home page: <http://webs.uvigo.es/acraaj/GenomePop.htm>

Operating system(s): Windows and Linux (the source will be provided to compile for Mac)

Programming language: C++

License: GNU GPL.

Authors' contributions

AC-R had the original idea for the work, designed and implemented the algorithms and wrote the manuscript.

Acknowledgements

I am grateful to A. Caballero, H. Quesada, S.T. Rodríguez-Ramilo and two anonymous reviewers for discussion and comments on the manuscript. I also want to thank Sergei L Kosakovsky Pond for his help with HYPHY. This work was supported by grant CPE03-004-C2 from Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA) and from Dirección Xeral de Investigación e Desenvolvemento from Xunta de Galicia. AC-R is currently funded by an Isidro Parga Pondal research fellowship from Xunta de Galicia (Spain).

References

- Carvajal-Rodríguez A: **GenomePop: software to simulate the evolution of genomes and populations.** [<http://webs.uvigo.es/acraaj/GenomePop.htm>].
- Liu Y, Nickle DC, Shriner D, Jensen MA, Gerald H, Learn J, Mittler JE, Mullins JI: **Molecular clock-like evolution of human immunodeficiency virus type 1.** *Virology* 2004, **329**:101-108.
- Liu Y, Mullins JI, Mittler JE: **Waiting times for the appearance of cytotoxic T-lymphocyte escape mutants in chronic HIV-1 infection.** *Virology* 2006, **347**(1):140-146.
- Caballero A, Cusi E, Garcia C, Garcia-Dorado A: **Accumulation of deleterious mutations: Additional Drosophila melanogaster estimates and a simulation of the effects of selection.** *Evolution* 2002, **56**(6):1150-1159.
- Carvajal-Rodríguez A, Rolan-Alvarez E, Caballero A: **Quantitative variation as a tool for detecting human-induced impacts on genetic diversity.** *Biological Conservation* 2005, **124**(1):1-13.
- Peng B, Amos CI, Kimmel M: **Forward-Time Simulations of Human Populations with Complex Diseases.** *PLoS Genet* 2007, **3**(3):e47.
- Peng B, Kimmel M: **Simulations provide support for the common disease-common variant hypothesis.** *Genetics* 2007, **175**(2):763-776.
- Keightley PD: **Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: a simulation study.** *Genetics* 1998, **150**(3):1283-1293.
- Wakeley J: **Nonequilibrium migration in human history.** *Genetics* 1999, **153**(4):1863-1871.
- Wakeley J: **The coalescent in an island model of population subdivision with variation among demes.** *Theor Popul Biol* 2001, **59**(2):133-144.
- Goldstein DB: **Islands of linkage disequilibrium.** *Nat Genet* 2001, **29**:109-111.
- Nothnagel M, Rohde K: **The effect of single-nucleotide polymorphism marker selection on patterns of haplotype blocks and haplotype frequency estimates.** *Am J Hum Genet* 2005, **77**(6):988-998.
- International-HapMap-Consortium: **The International HapMap Project.** *Nature* 2003, **426**(6968):789-796.
- International-HapMap-Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437**(7063):1299-1320.
- International-HapMap-Consortium: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**(7164):851-861.
- Jeffreys AJ, Holloway JK, Kauppi L, May CA, Neumann R, Slingsby MT, Webb AJ: **Meiotic recombination hot spots and human DNA diversity.** *Philos Trans R Soc Lond B Biol Sci* 2004, **359**(1441):141-152.
- Greenawalt DM, Cui X, Wu Y, Lin Y, Wang HY, Luo M, Tereshchenko IV, Hu G, Li JY, Chu Y, Azaro MA, Decoste CJ, Ching NO, Gao R, Shen L, Shih WJ, Lange K, Li H: **Strong correlation between meiotic crossovers and haplotype structure in a 2.5-Mb region on the long arm of chromosome 21.** *Genome Res* 2006, **16**(2):208-214.
- Liu N, Sawyer SL, Mukherjee N, Pakstis AJ, Kidd JR, Kidd KK, Brookes AJ, Zhao H: **Haplotype block structures show significant variation among populations.** *Genet Epidemiol* 2004, **27**(4):385-400.
- Nordborg M, Tavaré S: **Linkage disequilibrium: what history has to tell us.** *Trends Genet* 2002, **18**(2):83-90.
- Rosenberg NA, Nordborg M: **Genealogical trees, coalescent theory and the analysis of genetic polymorphisms.** *Nat Rev Genet* 2002, **3**(5):380-390.

21. Stumpf MPH, McVean GAT: **Estimating recombination rates from population-genetic data.** *Nature Reviews Genetics* 2003, **4**:959-968.
22. Hein J, Wiuf C, Schierup MH: **Gene genealogies, variation and evolution : a primer in coalescent theory.** Oxford , Oxford University Press; 2005:XIII, 276 s..
23. Kruglyak L: **Prospects for whole-genome linkage disequilibrium mapping of common disease genes.** *Nat Genet* 1999, **22**:139-144.
24. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: models and data.** *Am J Hum Genet* 2001, **69**(1):1-14.
25. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome.** *Science* 2004, **304**(5670):581-584.
26. Gu S, Pakstis AJ, Li H, Speed WC, Kidd JR, Kidd KK: **Significant variation in haplotype block structure but conservation in tag-SNP patterns among global populations.** *Eur J Hum Genet* 2007, **15**(3):302-312.
27. Marais G, Charlesworth B: **Genome evolution: recombination speeds up adaptive evolution.** *Curr Biol* 2003, **13**(2):R68-70.
28. Bahlo M, Griffiths RC: **Coalescence time for two genes from a subdivided population.** *J Math Biol* 2001, **43**(5):397-410.
29. Bahlo M, Griffiths RC: **Inference from gene trees in a subdivided population.** *Theor Popul Biol* 2000, **57**(2):79-95.
30. Beerli P, Felsenstein J: **Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach.** *Proceedings of the National Academy of Sciences, USA* 2001, **98**(8):4563-4568.
31. Notohara M: **The coalescent and the genealogical process in geographically structured population.** *J Math Biol* 1990, **29**:59-75.
32. Wilkinson-Herbots HM: **Genealogy and subpopulation differentiation under various models of population structure.** *J Math Biol* 1998, **37**(6):535-585.
33. Griffiths RC, Tavaré S: **Sampling theory for neutral alleles in a varying environment.** *Philosophical Transactions of the Royal Society of London, Series B* 1994, **344**:403-410.
34. Mohle M, Sagitov S: **A classification of coalescent processes for haploid exchangeable population models.** *Annals of Probability* 2001, **29**(4):1547-1562.
35. Tajima F: **The effect of change in population size on DNA polymorphism.** *Genetics* 1989, **123**:597-601.
36. Hey J, Wakeley J: **A coalescent estimator of the population recombination rate.** *Genetics* 1997, **145**:833-846.
37. Hudson RR, Kaplan NL: **The coalescent process in models with selection and recombination.** *Genetics* 1988, **120**:831-840.
38. Kaplan NL, Darden T, Hudson RR: **The coalescent process in models with selection.** *Genetics* 1988, **120**:819-829.
39. Krone SM, Neuhauser C: **Ancestral processes with selection.** *Theor Popul Biol* 1997, **51**(3):210-237.
40. Neuhauser C, Krone SM: **The genealogy of samples in models with selection.** *Genetics* 1997, **145**:519-534.
41. Donnelly P, Nordborg M, Joyce P: **Likelihoods and simulation methods for a class of nonneutral population genetics models.** *Genetics* 2001, **159**(2):853-867.
42. Barton NH, Etheridge AM, Sturm AK: **Coalescence in a random background.** *Annals of Applied Probability* 2004, **14**(2):754-785.
43. Fearnhead P: **Perfect simulation from nonneutral population genetic models: Variable population size and population subdivision.** *Genetics* 2006, **174**(3):1397-1406.
44. Calafell F, Grigorenko EL, Chiknian AA, Kidd KK: **Haplotype evolution and linkage disequilibrium: A simulation study.** *Hum Hered* 2001, **51**(1-2):85-96.
45. Anisimova M, Nielsen R, Yang Z: **Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites.** *Genetics* 2003, **164**(3):1229-1236.
46. Shriner D, Nickle DC, Jensen MA, Mullins JJ: **Potential impact of recombination on sitewise approaches for detecting positive natural selection.** *Genet Res* 2003, **81**:115-121.
47. Marjoram P, Wall JD: **Fast "coalescent" simulation.** *BMC Genet* 2006, **7**:16.
48. Liang L, Zollner S, Abecasis GR: **GENOME: a rapid coalescent-based whole genome simulator.** *Bioinformatics* 2007, **23**(12):1565-1567.
49. Wakeley J: **The limits of theoretical population genetics.** *Genetics* 2005, **169**(1):1-7.
50. Balloux F: **EASYPop (version 1.7): a computer program for population genetics simulations.** *J Hered* 2001, **92**(3):301-302.
51. Peng B, Kimmel M: **simuPOP: a forward-time population genetics simulation environment.** *Bioinformatics* 2005, **21**(18):3686-3687.
52. Guillaume F, Rougemont J: **Nemo: an evolutionary and population genetics programming framework.** *Bioinformatics* 2006, **22**(20):2556-2557.
53. Yang Z, Balding D, Bishop M, Cannings: **Adaptive Molecular Evolution.** In *Handbook of Statistical Genetics* Wiley J. and Sons Ltd.; 2003.
54. Karlin S, Taylor HM: **A second course in stochastic processes.** New York , Academic Press; 1981:XVIII, 542 s..
55. Swofford DL: **PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).** 4th edition. Sunderland, Massachusetts , Sinauer Associates; 2002.
56. Kosakovsky Pond SL, Frost SDW, Muse SV: **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics* 2005, **21**(5):676-679.
57. Muse SV, Gaut BS: **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.** *Mol Biol Evol* 1994, **11**(5):715-724.
58. Carvajal-Rodríguez A, Crandall KA, Posada D: **Recombination Estimation under Complex Evolutionary Models with the Coalescent Composite Likelihood Method.** *Mol Biol Evol* 2006, **23**(4):817-827.
59. Posada D, Crandall KA: **Modeltest: testing the model of DNA substitution.** *Bioinformatics* 1998, **14**(9):817-818.
60. McVean GAT, Awadalla P, Fearnhead P: **A coalescent based-method for detecting and estimating recombination from gene sequences.** *Genetics* 2002, **160**:1231-1241.
61. Raymond M, Rousset F: **GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism.** *J Heredity* 1995, **86**:248-249.
62. Kosakovsky Pond SL, Frost SD: **Not so different after all: a comparison of methods for detecting amino acid sites under selection.** *Mol Biol Evol* 2005, **22**(5):1208-1222.
63. Rodríguez F, Oliver JF, Marin A, Medina JR: **The general stochastic model of nucleotide substitution.** *J Theor Biol* 1990, **142**:485-501.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

