Microbial Genomics

SimBac: simulation of whole bacterial genomes with homologous recombination

--Manuscript Draft--

Manuscript Number:	MGEN-D-15-00039R2
Full Title:	SimBac: simulation of whole bacterial genomes with homologous recombination
Article Type:	Methods Paper
Section/Category:	Genomic Methodologies: Novel phylogenetic methods
Corresponding Author:	Nicola De Maio University of Oxford UNITED KINGDOM
First Author:	Thomas Brown
Order of Authors:	Thomas Brown
	Xavier Didelot
	Daniel J. Wilson
	Nicola De Maio
Abstract:	Bacteria can exchange genetic material, or acquire genes found in the environment. This process, generally known as bacterial recombination, can have a strong impact on the evolution and phenotype of bacteria, for example causing the spread of antibiotic resistance across clades and species, but can also disrupt phylogenetic and transmission inferences. With the increasing affordability of whole genome sequencing, the need has emerged for an efficient simulator of bacterial evolution to test and compare methods for phylogenetic and population genetic inference, and for simulation-based estimation. We present SimBac, a whole-genome bacterial evolution simulator that is roughly two orders of magnitude faster than previous software and includes a more general model of bacterial evolution, allowing both within- and between-species homologous recombination. Since methods modeling bacterial recombination generally focus on only one of these two modes of recombination, the possibility to simulate both allows for a general and fair benchmarking. SimBac is available from http://github.com/tbrown91/SimBac and is distributed as open source under the terms of the GNU General Public License.

MICROBIAL GENOMICS

Methods paper template

SimBac: simulation of whole bacterial genomes with homologous recombination

ABSTRACT

Bacteria can exchange genetic material, or acquire genes found in the environment. This process, generally known as bacterial recombination, can have a strong impact on the evolution and phenotype of bacteria, for example causing the spread of antibiotic resistance across clades and species, but can also disrupt phylogenetic and transmission inferences. With the increasing affordability of whole genome sequencing, the need has emerged for an efficient simulator of bacterial evolution to test and compare methods for phylogenetic and population genetic inference, and for simulation-based estimation. We present SimBac, a whole-genome bacterial evolution simulator that is roughly two orders of magnitude faster than previous software and includes a more general model of bacterial evolution, allowing both within- and between-species homologous recombination. Since methods modeling bacterial recombination generally focus on only one of these two modes of recombination, the possibility to simulate both allows for a general and fair benchmarking. SimBac is available from http://github.com/tbrown91/SimBac and is distributed as open source under the terms of the GNU General Public License.

DATA SUMMARY

SimBac, the software we developed to simulate genome-wide bacterial evolution, is distributed as open source under the terms of the GNU General Public License, and is available from GitHub (url - http://github.com/tbrown91/SimBac). A manual and examples of usage of SimBac are provided in the Supplementary Material.

We confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.

IMPACT STATEMENT

Sequencing technologies are revolutionizing microbiology, allowing researchers to investigate with great detail the genetic information in bacteria. This increasingly overwhelming amount of information requires adequate, efficient computer methods to be processed in reasonable time. One of the most important tasks performed by computer methods is simulating data, as this provides a mean for testing hypotheses and checking the performance of other methods in extracting valuable information from data. Previous software specifically developed for simulating bacterial evolution is limited in applicability, having being conceived for limited data and biological phenomena.

We present SimBac, a new simulator of bacterial evolution that can generate data for thousands of bacterial genomes about 100 times faster than previous methods. SimBac also includes a very general model of bacterial evolution that accounts for the fact that bacteria can exchange genetic material with each other, not only within the same population, but also across species boundaries. Thanks to these advancements in SimBac it will be possible to efficiently test hypotheses and estimate parameters comparing real and simulated bacterial data, to test the accuracy of bacterial genomic methods, and to fairly compare methods that make different assumptions regarding bacterial evolution.

INTRODUCTION

Whole-genome bacterial sequencing is rapidly gaining in popularity and replacing multilocus sequence typing (MLST) thanks to its fast and cost-effective provision of higher resolution genetic information (Wilson, 2012, Didelot *et al.*, 2012). Computational algorithms that use genomic data to infer

epidemiological, phylogeographic, phylodynamic, and evolutive patterns are generally hampered by recombination (e.g. Schierup & Hein, 2000, Posada & Crandal, 2002, Hedge & Wilson, 2014), and recent years have seen a surge of methods that measure, identify, and account for bacterial homologous recombination (e.g. Didelot & Falush, 2007, Marttinen *et al*., 2008, Marttinen *et al*., 2012, Croucher *et al*., 2014, Didelot *et al.*, 2010, Didelot & Wilson, 2015). Assessing and comparing the performance of different methods is complicated by the use of different models of recombination, in particular within-species recombination leading to phylogenetically discordant sites (e.g. Didelot *et al.*, 2010), or between-species recombination leading to accumulation of substitutions on specific branches and genomic intervals (e.g. Didelot & Falush, 2007). Simulators of bacterial evolution are routinely used for parameter inference and hypothesis testing (Fearnhead *et al.*, 2005, Fraser *et al.*, 2005) and for method testing and comparison (Falush *et al.*, 2006, Didelot & Falush, 2007, Turner *et al.*, 2007, Buckee *et al.*, 2008, Wilson *et al.*, 2009, Hedge & Wilson, 2014), but simulation software and models used are generally targeted to the specific model of evolution implemented in the methods considered. One of the reasons for this is the lack of general and efficient simulators of bacterial evolution.

Coalescent simulators of eukaryotic evolution usually focus on cross over recombination (see e.g. Arenas & Posada, 2007, 2009, 2014), while bacterial recombination is generally modeled as gene conversion, meaning that in a recombination event only a small fragment of DNA is imported from a donor, whereas most of the genetic material is inherited from the recipient. Many fast and approximate simulation methods (e.g. Marjoram & Wall, 2006, Excoffier & Foll, 2011) cannot be applied to bacterial recombination because the approximations used do not generate the expected long genomic distance correlations in bacterial local trees. Other similar approximate methods are only adequate for low bacterial recombination rates (e.g. Chen *et al*., 2009, Wang *et al.*, 2014). Many forward in time simulation methods (e.g. Chadeau-Hyam *et al.*, 2008, Dalquen *et al.*, 2012) or discrete generation coalescent methods (Excoffier *et al.*, 2000, Laval & Excoffier, 2004) can allow gene conversion, but are generally too slow for simulating whole-genome evolution of large samples or populations.

An exact and fast method to simulate gene conversion is the coalescent model of Wiuf & Hein (2000) included in ms (Hudson, 2002) and its extensions (Mailund *et al.*, 2005, Hellenthal & Stephens, 2007, Ramos-Onsins & Mitchell-Olds 2007). Recently, this model has been implemented in simulation software specific for bacterial evolution, SimMLST (Didelot *et al.*, 2009).

SimMLST is optimized for MLST data which requires to simulate several short distant loci, and, similarly to ms, only simulates within-species bacterial recombination. For these reasons, these methods are not generally suited for large genome-wide bacterial simulation studies or for testing different models and assumptions of recombination.

Here we present SimBac, a new method for simulating bacterial evolution. SimBac implements an efficient coalescent-based algorithm for simulating genome-wide bacterial evolution, and includes a new and more general model of bacterial recombination that extends the classical within-species recombination (Didelot *et al.*, 2009) by allowing the user to specify any degree of recombination between species.

THEORY AND IMPLEMENTATION

We simulate evolution backward in time under the standard coalescent model with gene conversion, and generate an ancestral recombination graph (ARG, see Wiuf & Hein, 2000). Within-species recombination events are modelled as a copy-pasting of a small fragment of DNA from the donor lineage sequence into the recipient.

The computational efficiency of SimBac derives from algorithmic improvements over previous software. First, instead of rejection sampling of recombination events as in Didelot *et al.*, 2009, we developed an analytical solution that only samples recombination events effectively altering ancestral material of lineages (details of the methods are given in the Supplementary Material). Second, we represent ancestral material with a more efficient data structure. These new features allow about 100-fold faster simulation of

bacterial genome-wide evolution compared to SimMLST (see Fig. 1). Also, our method generally outperforms ms (Hudson, 2002) when many recombination (or equivalently gene conversion) events are expected.

Our software also provides the possibility to simulate a circular or linear genome, and entire or fragmented bacterial genome, and offers a recombination model that allows a mixture of between- and within-species recombination. Within-species recombination is modelled as the coalescent with gene conversion (Wiuf & Hein, 2000, Didelot *et al.,* 2009) with fragments lengths distributed geometrically with mean δ, and with all sites having the same per-site recombination initiation rate *R* (scaled by the effective population size). As the coalescent process is simulated backward in time, any extant lineage can be the recipient of a recombining interval from a donor lineage, which is then added to the other extant lineages. In such a case, the recombining interval becomes part of the genome of the new donor lineage (see Fig. 2(b)). Every site of the genome of every extant lineage becomes the start of a recombining interval at the same rate *R*.

Between-species recombination is modelled as a separate process backward in time with a specific scaled per-site recombination initiation rate R_e and a specific distribution of imported fragments lengths (geometric with mean δ_e). When a between-species recombination event occurs at a recipient lineage and interval, the donor lineage is not tracked back in time as for within-species recombination, but instead substitutions are introduced into the recombining interval, similar to the model in ClonalFrame (Didelot & Falush, 2007). Therefore, we do not simulate species evolution as in Arenas & Posada (2014), but rather assume that each recombining segment is donated by a different lineage within a given divergence range.

However, differently from ClonalFrame, the donor sequence is obtained adding a random amount of divergence (uniformly sampled within the interval $[D_1, D_2]$, specified by the user) into the corresponding homologous sequence from the root of the ARG. This model accounts for the excess of substitutions caused by between-species recombination as in ClonalFrame, but at the same time also generates the homoplasies that are expected if the recipient lineage does not lead to the root of the local tree. More details on the methods of

simulation and a summary of the algorithm are provided in the Supplementary Material.

To showcase the possible applications of our software, we extend the investigation of phylogenetic inference accuracy by Hedge & Wilson (2014). The authors investigated the effect of low bacterial recombination rates (up to a scaled per-site rate of R=0.01) on the inference of clonal frame. Using SimBac, we are able to simulate higher recombination rates (up to R=0.1) in reasonable time, and we show that for highly recombining bacteria, and in particular for older phylogenetic branches, the probability of reconstructing the phylogenetic topology is reduced further to around 91% (Fig. 3).

CONCLUSION

Simulation of genome evolution is important as it allows inference of parameters from data and testing of evolutionary hypothesis, and because it is routinely used to benchmark and compare different microbial genomic analysis methods. We present SimBac, a new method for simulating genome-wide bacterial evolution implemented and distributed as open source software (http://github.com/tbrown91/SimBac). Our model of bacterial recombination is more general than those used by most methods in the field, in that it can describe any mixture of within-species and between-species recombination, and as such, it can fit the assumptions of most methods, or it can provide a more realistic background for comparing methods with different hypothesis. Also, our efficient implementation achieves an approximately 100-fold increase in computational efficiency over previous similar effort, allowing inference and benchmarking over considerably larger datasets. For example, a thousand 1Mbp genomes with *R=0.01* can be generated in about 6 minutes. SimBac can generate a wide range of possible outputs: sequence alignments, ARGs graphics (see Fig. 2), clonal frames, local genealogies, and lists of recombination events. Although only a JC substitution model (Jukes & Cantor 1969) is presently included in SimBac, in practice this is not a restriction because the local genealogies can be used to generate alignments under a vast choice of nucleotide and codon substitution models using for example SeqGen

(Rambaut & Grassly 1997) or INDELible (Fletcher & Yang, 2009) (see Arenas, 2013).

Although SimBac generalizes the applicability of SimMLST, it currently lacks the wide set of options of some simulators of evolution, in particular of forward simulators that allow very general demographic, speciation, selection, migration, and rate variation patterns (e.g. Chadeau-Hyam *et al.*, 2008, Dalquen *et al.*, 2012). In fact, many of these features present considerable methodological hurdles in being incorporated in computationally efficient coalescent simulators.

Yet, future extensions of our method could consist of the inclusion of distributive conjugal transfer (Gray *et al.*, 2013), of non-homogenous genomic rates of recombination (see e.g. Everitt *et al.*, 2013, Arenas & Posada, 2014), or of demographic events and population structure (Arenas & Posada, 2007, Arenas & Posada, 2014).

ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council [EP/F500394/1 to TB]; the Biotechnology and Biological Sciences Research Council [BB/L023458/1 to XD]; the National Institute for Health Research [HPRU-2012-10080 to XD]; the Wellcome Trust to DJW; the Royal Society [101237/Z/13/Z to DJW]; and the Oxford Martin School to NDM.

We thank Jessica Hedge for comments on the project.

ABBREVIATIONS

Multilocus sequence typing (MLST); ancestral recombination graph (ARG).

REFERENCES

Arenas, M. & Posada, D. (2007). Recodon: Coalescent simulation of coding DNA sequences with recombination, migration and demography. *BMC Bioinformatics* **8**, 458.

Arenas, M. & Posada, D. (2010). Coalescent simulation of intracodon recombination. *Genetics* **184.2**, 429-437.

Arenas, M. (2013). Computer programs and methodologies for the simulation of DNA sequence data with recombination. *Frontiers in genetics*, ***4*.**

Arenas, M. & Posada, D. (2014). Simulation of Genome-Wide Evolution under Heterogeneous Substitution Models and Complex Multispecies Coalescent Histories. *Molecular Biology and Evolution* **31(5)**, 1295-1301.

Buckee, C., Jolley, K., Recker, M., Penman, B., Kriz, P., Gupta, S. & Maiden, M. (2008). Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitides*. *PNAS* **105**, 15082-15087.

Chadeau-Hyam, M., Hoggart, C. J., O'Reilly, P. F., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *Bmc Bioinformatics*, ***9*(1),** 364.

Chen, G. K., Marjoram, P., & Wall, J. D. (2009). Fast and flexible simulation of DNA sequence data. *Genome research*, ***19*(1),** 136-142.

Croucher, N., Page, A., Connor, T., Delaney, A., Keane, J., Bentley, S., Parkhill, J. & Harris, S. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43**, e15.

Dalquen, D. A., Anisimova, M., Gonnet, G. H., & Dessimoz, C. (2012). ALF—a simulation framework for genome evolution. *Molecular biology and evolution*, ***29*(4),** 1115-1123.

Didelot, X. & Falush, D. (2007). Inference of Bacterial Microevolution Using Multilocus Sequence Data. *Genetics* **175**, 1251-1266.

Didelot, X., Lawson, D. & Falush, D. (2009). SimMLST: simulation of mulit-locus sequence typing data under a neutral model. *Bioinformatics* **25**, 1442-1444.

Didelot, X., Lawson, D., Darling, A. & Falush, D. (2010). Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences. *Genetics* **186**, 1435-1449.

Didelox, X., Bowden, R., Wilson, D., Peto, T. & Crook, D. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature Review Genetics* **13**, 601-612.

Didelot, X. & Wilson, D. (2015). ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Computational Biology* **11**, e1004041.

Drummond, A. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 214.

Everitt, R. G., Didelot, X., Batty, E. M., Miller, R. R., Knox, K., Young, B. C., ... & Wilson, D. J. (2014). Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nature communications*, ***5***.

Excoffier, L., Novembre, J., & Schneider, S. (2000). SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *Journal of Heredity*, ***91*****(6),** 506-509.

Excoffier, L. & Foll, M. (2011). Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27.9**, 1332-1334.

Falush, D., Torpdahl, M., Didelot, X., Conrad, D., Wilson, D. & Achtman, M. (2006). Mismatch induced speciation in *Salmonella*: model and data. *Philos Trans R Soc Lond B Biol Sci* **361**, 2045-2053.

Fearnhead, P., Smith, N., Barrigas, M., Fox, A. & French, N. (2005). Analysis of Recombination in *Campylobacter jejuni* from MLST Population Data. *Journal of Molecular Evolution* **61**, 333-340.

Fletcher, W., & Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, ***26*****(8),** 1879-1888.

Fraser, C., Hanage, W. & Spratt, B. (2005). Neutral microepidemic evoluation of bacterial pathogens. *PNAS* **102**, 1968-1973.

Gansner, E., Koutsofios, E., North, S. & Vo, K. (1993). A Technique for Drawing Directed Graphs. *IEEE Trans. Softw. Eng.* **19**, 1968-1973.

Gray, T. A., Krywy, J. A., Harold, J., Palumbo, M. J., & Derbyshire, K. M. (2013). Distributive conjugal transfer in mycobacteria generates progeny with meiotic-like genome-wide mosaicism, allowing mapping of a mating identity locus. *PLoS Biol*, **11(7)**, e1001602.

Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W. & Gascue, O. (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321.

Hedge, J. & Wilson, D. (2014). Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not. *mBio* **5**, e02158-14.

Hellenthal, G., & Stephens, M. (2007). msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, ***23*****(4),** 520-521.

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

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, ***3***, 21-132.

Laval, G., & Excoffier, L. (2004). SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, ***20*****(15),** 2485-2487.

Mailund, T., Schierup, M. H., Pedersen, C. N., Mechlenborg, P. J., Madsen, J. N., & Schauser, L. (2005). CoaSim: a flexible environment for simulating genetic data under coalescent models. *BMC bioinformatics*, ***6*(1),** 252.

Marjoram, P. & Wall, J.D. (2006). Fast Coalescent Simulation. *BMC genetics* **7.1**, 16.

Marttinen, P., Baldwin, A., Hanage, W. P., Dowson, C., Mahenthiralingam, E., & Corander, J. (2008). Bayesian modeling of recombination events in bacterial populations. *BMC bioinformatics*, **9(1)**, 421.

Marttinen, P., Hanage, W. P., Croucher, N. J., Connor, T. R., Harris, S. R., Bentley, S. D., & Corander, J. (2012). Detection of recombination events in bacterial genomes from large population samples. *Nucleic acids research*, **40(1)**, e6-e6.

Posada, D. & Crandall, K. (2002). The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.* **54**, 396-402.

Rambaut, A., & Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences: CABIOS*, ***13*(3),** 235-238.

Ramos-Onsins, S. E., & Mitchell-Olds, T. (2007). Mlcoalsim: multilocus coalescent simulations. *Evolutionary bioinformatics online*, **3**, 41.

Robinson, D., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, ***53*(1),** 131-147.

Schierup, M. & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879-891.

Turner, K., Hanage, W., Fraser, C., Connor, T. & Spratt, B. (2007). Assessing the reliability of eBURST using simulated populations with known ancestry. *BMC Microbiol* **7**, 30.

Wang, Y., Zhou, Y., Li, L., Chen, X., Liu, Y., Ma, Z. M., & Xu, S. (2014). A new method for modeling coalescent processes with recombination. *BMC bioinformatics*, ***15*(1),** 273.

Wilson, D. (2012). Insights from Genomics into Bacterial Pathogen Populations. *PLoS Pathog* **8**, e1002874.

Wilson, D., Gabriel, E., Leatherbarrow, A., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Hart, C., Diggle, P. & Fearnhead, P. (2009). Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Mol Biol Evol* **26**, 385-397.

Wiuf, C. & Hein, J. (2000). The coalescent with gene conversion. *Genetics* **155**, 451-462.

DATA BIBLIOGRAPHY

Brown, T., Didelot, X., Wilson, D.J. & De Maio, N., GitHub https://github.com/tbrown91/SimBac (2015).

FIGURES AND TABLES

Figure 1

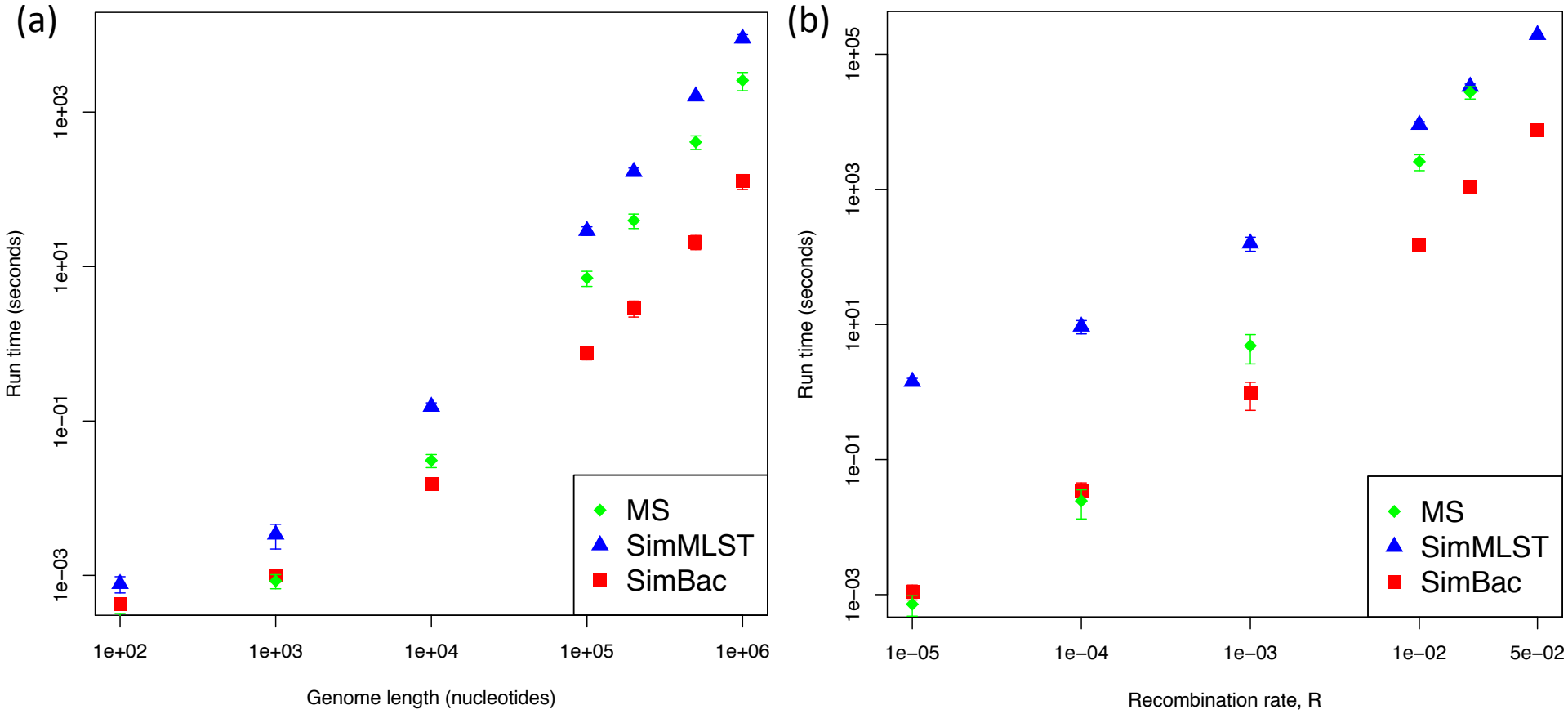
Comparison of run-time of SimMLST, ms and SimBac. Only gene conversion (no cross-over) is simulated in ms, to model bacterial evolution. (a) Average time to simulate the ARG for a fixed recombination rate R=0.01 and genome length from 100bp to 1Mbp. (b) Average time to simulate the ARG for a fixed genome length of 1Mbp and recombination rate increasing from R=0 to R=0.05. 100 Simulations were performed for each dot, except for SimMLST at R=0.02 and R=0.05, and ms at R=0.02, where 10 simulations were performed due to the elevated computational demand. ms was not run at R=0.05 because a single run required >4 days. Error bars show ±1 standard deviation.

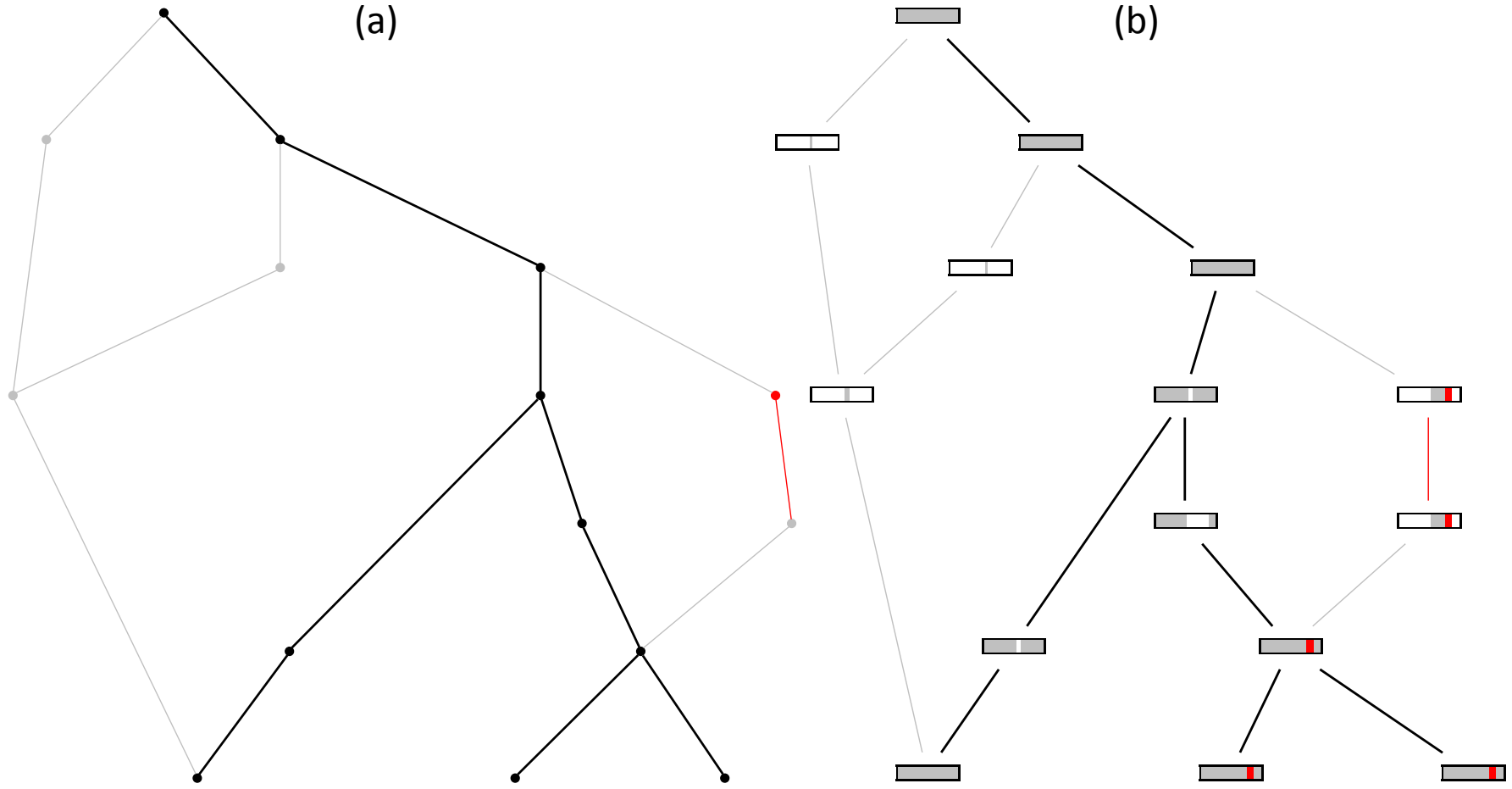
Figure 2

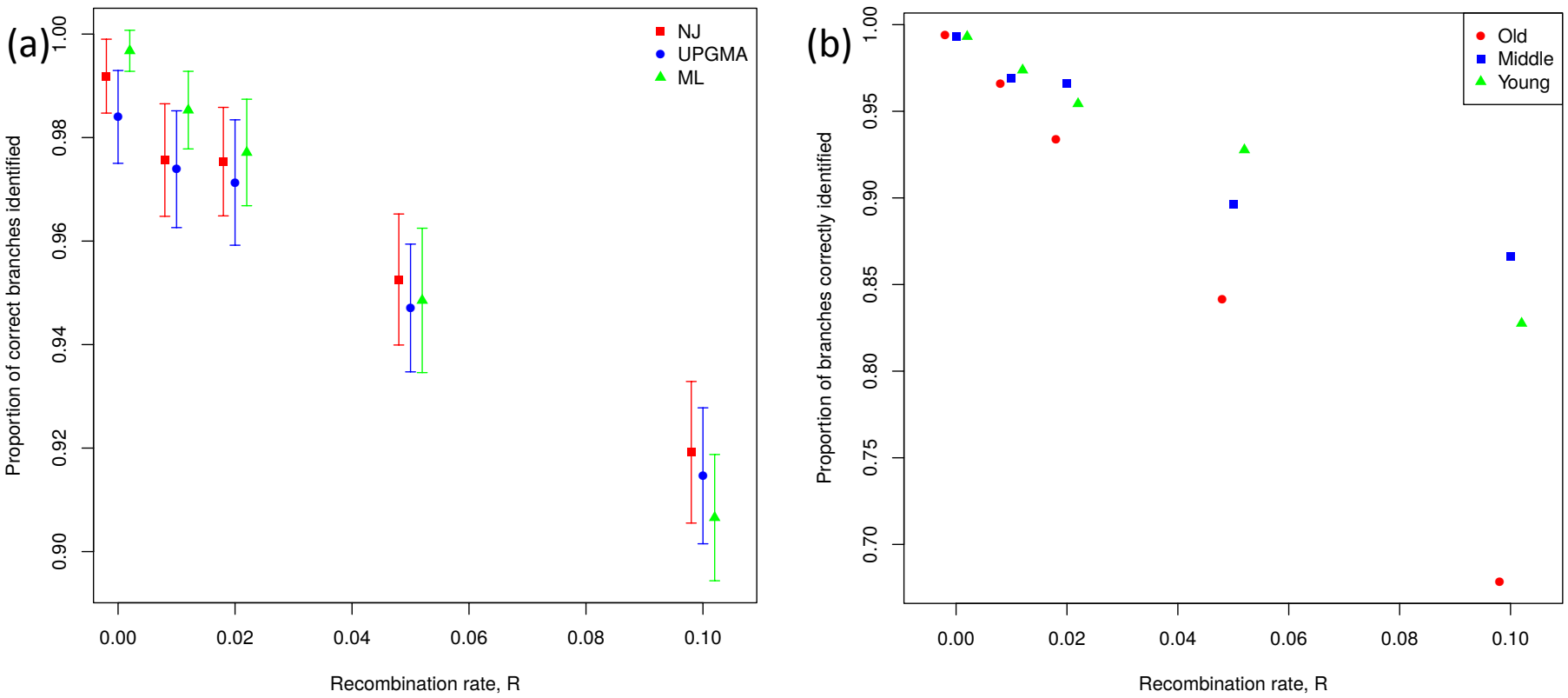
Examples of Ancestral Recombination Graphs (ARGs) generated and plotted by SimBac. Branches represent ARG lineages, and time is considered from to go backward from the bottom to the top of the tree. Branch merges (from bottom to top) represent coalescent events, while branch splits represent recombination events. (a) Example ARG with the clonal frame lineages marked in black, the non-clonal lineages in grey, and a recombination event involving an external species marked in red. (b) Same ARG as before, but with ancestral material of each lineage represented as a rectangle in the corresponding node. Each colored vertical bar inside each rectangle represent a genomic segment. Genomic segments that are present in the ancestral material are colored in grey, those absent are in white, and those imported from an external species are in red.

Figure 3

Accuracy of clonal frame estimation from recombining bacterial genomes.
The X axis shows the recombination rate R under which simulations are performed.
The Y axis shows the accuracy of inference, as the proportion of branches correctly estimated using the Robinson-Foulds metric (Robinson & Foulds, 1981). Ten independent replicates are used for R=0.1 and a hundred in all other cases. Genomes are 1Mbp long and the scaled mutation rate is fixed at 0.01. (a) Accuracy of three phylogenetic methods: Neighbour Joining (NJ), Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Maximum Likelihood (ML). Error bars represent ±1 standard deviations. (b) Clonal frame branches were separated into three age categories: young, middle-aged, and old (respectively with a distance between the branch mid-point and the root of more than 2.09, between 1.32 and 2.09, and less than 1.32 N_e generations). The ML accuracy for each age category is plotted separately in different colors.


(a)
Run time (seconds)
1e−03
1e−01
1e+01
1e+03
1e+02
1e+03
1e+04
1e+05
1e+06
Genome length (nucleotides)
MS
SimMLST
SimBac
(b)
Run time (seconds)
1e−03
1e−01
1e+01
1e+03
1e+05
1e−05
1e−04
1e−03
1e−02
5e−02
Recombination rate, R
MS
SimMLST
SimBac


(a)
(b)


(a)
NJ
UPGMA
ML
Proportion of correct branches identified
1.00
0.98
0.96
0.94
0.92
0.90
0.00
0.02
0.04
0.06
0.08
0.10
Recombination rate, R
(b)
Old
Middle
Young
Proportion of branches correctly identified
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.00
0.02
0.04
0.06
0.08
0.10
Recombination rate, R

SimBac: simulation of whole bacterial genomes with homologous recombination

Supplementary Information

Thomas Brown [1], Xavier Didelot [2], Daniel J. Wilson [3,4,5] and Nicola De Maio [3,4,*]
1 Doctoral Training Centre, University of Oxford, Oxford, United Kingdom
2 Department of Infectious Disease Epidemiology, Imperial College, London, United Kingdom
3 Institute for Emerging Infections, Oxford Martin School, Oxford, United Kingdom
4 Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
5 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom
*** E-mail: nicola.demaio@ndm.ox.ac.uk**

User Manual

SimBac jointly simulates bacterial genomes with the clonal genealogy under a coalescent model with recombination. Such simulations can be used to test phylogenetic analysis of real data sets.

Running from the command line

SimBac can be run from the command line using a combination of the following options. Passing no arguments to SimBac will display the possible arguments.

-N NUM Sets the number of isolates (default is 100)
-T NUM Sets the value of θ, the site-specific mutation rate, between 0 and 1 (default is 0.01)
-m NUM Sets the lower bound of site-mutation (divergence) in a region of external recombination, between 0 and 1 (default is 0)
-M NUM Sets the upper bound of site-mutation (divergence) in a region of external recombination, between 0 and 1 (default is 0)
-R NUM Sets the per-site rate of internal (within species) recombination, R_i, (default is 0.01)
-r NUM Sets the per-site rate of external (between species) recombination, R_e, (default is 0)
-D NUM Sets the average length of an internal recombinant interval, δ_i (default is 500)
-e NUM Sets the average length of an external recombinant interval, δ_e (default is 500)
-B NUM,...,NUM Sets the number and lengths of fragments of genetic material (default is 10000)
-G NUM,...,NUM Sets the size of gaps between each fragment, must be the same number of gaps as there are numbers of genetic fragments (default is 0,...,0)
-s NUM Use given seed to initiate random number generation
-o FILE Name of file to write generated sequences (FASTA format)
-c FILE Name of file to write clonal genealogy (Newick format)
-l FILE Name of file to write local trees (Newick format)
-b FILE Name of file to write log of internal recombination breaks
-f FILE Name of file to write log of external recombination breaks
-d FILE Name of file to export ancestral recombination graph (DOT file)
-a Include ancestral material in the DOT graph

Output format

SimBac produces the following output files:

- FASTA file of simulated sequences. If more than one fragment of genetic information is specified, the output is in the eXtended Multi-Fasta Alignment (XMFA) format. In this situation the simulated gene fragments are separated with an '=' sign.
- The clonal genealogy in Newick format
- The local trees contained in the simulated data. This file is a list of Newick trees each of which is preceded by the number of sites that share the current local tree.
- A full description of the graph representing the ancestry of the sample in the DOT language (Fig. 2 in the Main Text). This can be used in conjunction with the graphviz and the DOT program to produce figures illustrating the ancestry. The examples show the ancestry with and without the ancestral material included at each node. The clonal genealogy is shown in bold and external recombination events are shown in red. In the graph showing the ancestral material, the ancestral material remaining at each node is shown in grey and any external genetic material is shown in red.

Examples

To simulate 100 genomes each 1Mbp long with an internal recombination rate $R_i = 0.01$ and mutation rate $\theta = 0.01$ run:
./SimBac -N 100 -B 1000000 -R 0.01 -T 0.01 -o sequences.fasta -c clonal.nwk -l local.nwk
This produces the simulated sequences and the clonal genealogy in the files 'sequences.fasta' and 'clonal.nwk'. The local trees are written to 'local.nwk'

To simulate 100 genomes with internal and external recombination rate $R_e = 0.01$ and average break length of 500bp run:
./SimBac -N 100 -B 1000000 -R 0.01 -D 500 -r 0.01 -e 500 -b internal.log -f external.log
This produces two log files with the start- and end-points of all internal and external recombination events.

To simulate sequences undergoing internal and external recombination with mutation in an external recombinant interval occurring with probability in the interval $[0.5, 1]$, run:
./SimBac -N 100 -B 100000 -R 0.01 -D 500 -r 0.01 -e 500 -m 0.5 -M 1 -o sequences.fasta -c clonal.nwk
This produces the sequences and clonal genealogy.

To produce a DOT file with the ancestral information included in the graph run:
./SimBac -n 10 -B 1000 -R 0.01 -D 50 -r 10 -e 50 -d graph.dot -a

To simulate a linear genome, add a large gap to the end of the genome to prevent any recombinant intervals including both the first and last elements of the genome. For example to simulate a linear genome of length 100kbp run:
./SimBac -N 100 -B 100000 -G 1000000 -o sequences.fasta -c clonal.nwk
This places a gap of 1Mbp at the end of the genome.

Supplementary Methods

Hereby we will use the notation of [2], except that we will assume that there is a circular genome of length G (with sites $1 \dots G$), and that ancestral material of each lineage is a subset of this genome. Ancestral material for a sample (an ARG tip) need not be the whole genome, but might be a subset of the genome made of different loci, for example in the case of MLST data, so that we can simulate both genome data or MLST data. Ancestral material for a node consists of b non-overlapping ordered intervals, $I_1 \dots I_b$, of lengths respectively $L_1 \dots L_b$, and with $I_i = [s_i, e_i]$ (implying $e_i - s_i = L_i - 1$). Also for easiness of presentation (due to genome circularity) we will set $e_0 = e_b - G$, which is intendedly negative. The recombination rate per site per genome will be $R/2$. It should be noted that $\rho = 2R$, where ρ is the rate of recombination initiation or termination in LDhat [3]. Lastly, recombining intervals have a geometric distribution with mean δ.

Effective recombination rate for a lineage

In [2], the recombination rate per site is $R/2$. We call a the ancestral material of this lineage. If a recombination event happens on the considered lineage, then a recombining interval r is picked at random from the genome, and if $r \cap a \neq \emptyset$ (and $a - r \neq \emptyset$ for lineages not in the clonal frame) then the two new recombining lineages are created, otherwise the recombination event is rejected.

Here we propose to sample recombination events and recombining intervals conditional on $r \cap a \neq \emptyset, a - r \neq \emptyset$, or just on $r \cap a \neq \emptyset$ for lineages in the clonal frame, such that no rejection ever occurs while simulating. To do this, we first define a lineage-specific effective recombination rate. This is the rate at which recombination events occur satisfying $r \cap a \neq \emptyset, a - r \neq \emptyset$ (or just $r \cap a \neq \emptyset$ for clonal frame lineages). As in [2], we assume that the rate of initiation of a recombination event is the same for each site of the genome. Under these assumptions, and assuming as in [1] a geometric distribution with mean δ for recombination interval lengths, the rate at which a recombination event is started between e_0 and s_1, and includes s_1, is:

$$\begin{aligned}
\frac{R_{s_1-e_0}}{2} &= \frac{R}{2} \sum_{i=0}^{s_1-e_0-1} (1-\delta^{-1})^i = \\
&= \frac{R}{2} \left[\sum_{i=0}^{\infty} (1-\delta^{-1})^i - \sum_{i=s_1-e_0}^{\infty} (1-\delta^{-1})^i \right] = \\
&= \frac{R}{2} \left[\delta - \delta(1-\delta^{-1})^{(s_1-e_0)} \right] = \\
&= \frac{R}{2} \delta (1 - (1-\delta^{-1})^{(s_1-e_0)}).
\end{aligned}$$

Where $\left(1-\delta^{-1}\right)^i$ is the probability of a recombinant break having length greater than i. Now, let us assume we have a lineage with ancestral material $a = \cup_{i=1}^{b} [s_i, e_i]$ union of non-empty, ordered, disjoint intervals. As mentioned before, $e_0 = e_b - G$. The amount of ancestral material in a lineage is defined as: $L = \sum_{i=1}^{b} L_i$ The rate of recombination events satisfying $r \cap a \neq \emptyset$ for that lineage is then:

$$\frac{R_a}{2} = \left(\sum_{i=1}^{b} \frac{R_{s_i - e_{i-1}}}{2} \right) + \frac{R}{2}(L - b).$$

Finally, the lineage-specific recombination rate satisfying $r \cap a \neq \emptyset$, and $a - r \neq \emptyset$ is:

$$\frac{R'_a}{2} = \frac{R_a}{2} - \left(\sum_{i=1}^{b} \frac{R_{s_i - e_{i-1}}}{2} (1-\delta^{-1})^{G-(s_i-e_{i-1})} \right) - \frac{R}{2}(1-\delta^{-1})^{G-1}(L-b).$$

Additionally, for a clonal lineage without ancestral material the recombination rates will be 0.

Probability of recombination initiating sites

Conditional on an effective recombination event on a non clonal frame lineage (that is, satisfying $r \cap a \neq \emptyset, a - r \neq \emptyset$) occurring on ancestral material a, the probability that the first ancestral site affected by r is s_i is:

$$P'_{s_i} = \frac{R_{s_i - e_{i-1}}(1 - (1 - \delta^{-1})^{G-(s_i - e_{i-1})})}{R'_a},$$

and the probability that it is any other site in a is

$$\frac{R(1 - (1 - \delta^{-1})^{G-1})}{R'_a}.$$

If the considered recombining lineage is in the clonal frame instead (with recombination satisfying only $r \cap a \neq \emptyset$), the probabilities are

$$P_{s_i} = \frac{R_{s_i - e_{i-1}}}{R_a},$$

and

$$\frac{R}{R_a}$$

respectively.

After the starting site of $r \cap a$ has been picked, the ending site of r is chosen according to a geometric distribution with mean δ for a lineage in the clonal frame. In a non-clonal lineage, the ending site of r is chosen according to the same geometric distribution, but conditional on $|r| \leq G - (s_i - e_{i-1})$ if the starting site of r is s_i, or $|r| \leq G - 1$ otherwise.

External recombination events

Simulation of external recombination events follows the same protocol as that of a clonal recombination event, with $R_e/2$ and δ_e replacing $R/2$ and δ, respectively. As we are only interested in the imported fragment from the external species, the recombinant interval need only satisfy the condition $r \cap a \neq \emptyset$, as in a clonal lineage.

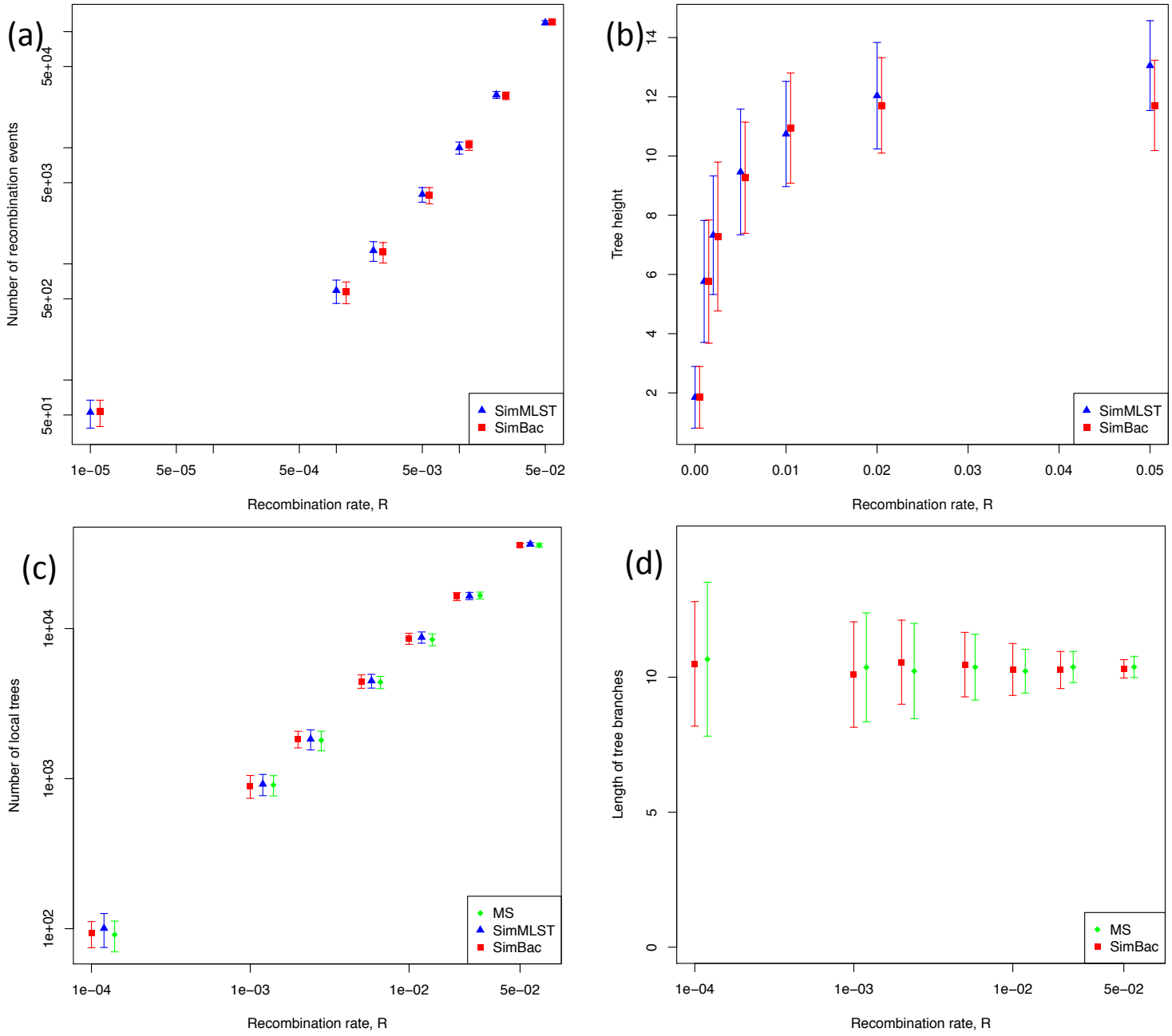
Technical description of SimBac

Here, n denotes the number of isolates for which the data is to be simulated. For a genome of length G, simulate the ARG with internal recombination rate $R_i/2$ and external recombination rate $R_e/2$. The average length of internal and external recombining segments are given by δ_i and δ_e, respectively.

Input: $n,\ R_i/2,\ R_e/2,\ \delta_i,\ \delta e,\ G$
Output: Simulated ancestral recombination graph with clonal genealogy
Set number of lineages $k = n$
for $x = 1 \ldots n$ **do**
  Calculate internal and external recombination rates, $\frac{R_{i,x}}{2}$, $\frac{R_{e,x}}{2}$
  Ancestral material a_x is given by $[0, G]$
  Lineage x is clonal
end for
while $k > 1$ **do**
  Calculate the rates of internal and external recombination given by:
  $\frac{R_i}{2} = \sum_{x=1}^{k} \frac{R_{i,x}}{2}$ and $\frac{R_e}{2} = \sum_{x=1}^{k} \frac{R_{e,x}}{2}$
  Increment current time by an amount distributed exponentially with parameter $\binom{k}{2} + \frac{R_i}{2} + \frac{R_e}{2}$
  Let $u \sim U(0, 1)$
  if $u < k(k-1)/(k(k-1) + R_i + R_e)$ **then**
    Coalescent event
    Choose two lineages x and y at random from the k remaining nodes and replace with the new lineage z
    The lineage z is clonal if x or y is clonal
    The ancestral material of z is $a_z = a_x \cup a_y$
    Update total number of lineages containing each element of the genome
    for all Nucleotides **do**
      if Number of lineages containing given nucleotide is one **then**
        Remove nucleotide from a_z
      end if
    end for
    Calculate the internal and external rate of recombination for the new lineage, $\frac{R_{i,z}}{2}$ and $\frac{R_{e,z}}{2}$
    The number of lineages, k is decreased by one
  else if $u < (k(k-1) + R_i)/(k(k-1) + R_i + R_e)$ **then**
    Internal recombination event
    Choose one lineage x weighted by $\frac{R_{i,x}}{2}$
    Determine a recombining interval, r, distributed geometrically with parameter δ_i.
    if x is clonal **then**
      Choose r such that $r \cap a_x \neq \emptyset$
    else
      Choose r such that $r \cap a_x \neq \emptyset$ and $a - r \neq \emptyset$
    end if
    Create two new lineages, y and z.
    Lineage z is clonal if x is clonal
    $a_y = a_x \cap r$ and $a_z = a_x - r$
    Lineage y is not clonal
    Calculate the new recombination rates for lineages y and z, $\frac{R_{i,y}}{2}$, $\frac{R_{e,y}}{2}$, $\frac{R_{i,z}}{2}$ and $\frac{R_{e,z}}{2}$
    The number of lineages, k is increased by one
  else
    External recombination event
    Choose one lineage, x weighted by $\frac{R_{e,x}}{2}$
    Choose a recombinant interval, r, distributed geometrically with parameter δ_e conditioned on $r \cap a_x \neq \emptyset$
    The material satisfying $a_x \cap r$ will be simulated as genetic material from an external species
  end if
end while

Methods Validation

Fig. S 1: Validation of SimBac. To validate SimBac we compared summary statistics of its simulated data to those of simMLST and ms. Not all statistics are available for every software. Genome length is 10^5bp. On X axis is always shown the scaled, per-site recombination rate R and error bars represent $\pm$ 1 standard deviations. 10 simulations were performed for R=0.02 and 0.05 in simMLST in the top two plots, 100 simulations in all other cases. (a) Total number of recombination events. (b) Height of the ARG. (c) Number of local trees (identical neighbouring local trees were merged). (d) Average sum of branch lengths for local trees (ms values are scaled by a factor of 2 as it assumes diploidy, while SimBac and SimMLST assume haploidy).



References

[1] Didelot,X., Falush,D. (2006) Inference of Bacterial Microevolution Using Multilocus Sequence Data. *Genetics*, **175**, 1251-1266.

[2] Didelot, X. Lawson, D, Falush, D. (2009) SimMLST: simulation of multi-locus sequence typing data under a neutral model *Bioinformatics*, **25**(11): 1442–1444.

[3] McVean, G. Awadalla, P. Fearnhead, P. (2002) A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences *Genetics*, **160**: 1231–1241.