NGSphy: phylogenomic simulation of next-generation sequencing data

Merly Escalona [1], Sara Rocha [1] and David Posada [1,2,3]

[1] Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain
[2] Biomedical Research Center (CINBIO), University of Vigo, 36310 Vigo, Spain.
[3] Galicia Sur Health Research Institute, 36310 Vigo, Spain.

Abstract

Advances in sequencing technologies have made feasible the obtention of massive datasets for phylogenomic inference, often consisting of large numbers of loci from multiple species and individuals. The analysis of next-generation sequencing (NGS) in phylogenomics implies a complex computational pipeline where multiple technical and methodological decisions are necessary that might influence the final tree obtained, from coverage to assembly, mapping, variant calling and/or phasing. In order to assess the influence of these variables, we introduce NGSphy, an open-source tool for the simulation of Illumina reads/read counts obtained from haploid/diploid individual genomes with thousands of independent gene families evolving under a common species tree. In order to resemble real NGS experiments, NGSphy includes multiple options to model sequencing coverage (depth) heterogeneity across species, individuals and loci, including off-target or uncaptured loci. For comprehensive simulations covering multiple evolutionary scenarios, parameter values for the different replicates can be sampled from user-defined statistical distributions.

Availability and Implementation: Source code and documentation are available at http://github.com/merlyescalona/ngsphy

Contact: merlyescalona@uvigo.es

1. Introduction

Next-generation sequencing (NGS) technologies facilitate nowadays the obtention of large "phylogenomic" datasets with hundreds or thousands of loci with several individuals from multiple species (McCormack et al 2013). At the same time, genome-wide data have brought a renewed interest in the discordance between gene trees and species trees and in methods of simulation

and phylogenetic reconstruction that deal with large amounts of loci and potential sources of phylogenetic incongruence (Mallo and Posada 2016). Importantly, the assembly of multiple sequence alignments from NGS reads is not free from errors and biases. There are many variables that might interfere with the accuracy of the final gene trees and species trees inferred from NGS data, from experimental-design aspects such as number of samples, sequencing technology or coverage, to methodological decisions along the processes of assembly or mapping, orthology inference, variant and genotype calling, or phasing. In this context, the simulation of NGS reads can be quite useful to properly understand and improve the NGS phylogenomic pipeline. However, at the phylogenomic scale we need to be able to simulate NGS data from multiple gene families or loci with potentially discordant phylogenies, and represented by multiple individuals. To the best of our knowledge, the only tool which generates NGS data from phylogenies is TreeToReads (McTavish et al. 2017), yet this tool cannot easily simulate data along multiple gene trees, diploid individuals, or consider NGS coverage heterogeneity, and therefore is not very suitable for the phylogenomic scale. Here we introduce NGSphy, an easy-to-use pipeline to generate NGS data (read counts or Illumina reads) from multiple loci belonging to multiple haploid/diploid individuals and species under the gene tree / species tree paradigm, and with different options to control coverage variation across species, loci and individuals.

2. Description and implementation

NGSphy is written in Python and is open-source. Its workflow is depicted in Supplementary Figure 1. Parameter values and options for the simulations are specified in a settings file. Arguments and conditions for the different replicates can be sampled from user-defined statistical distributions, or set as fixed. In the simplest scenario, the user just needs to specify a single gene tree, a substitution model, and the sequencing design. In addition, the user can provide an ancestral or current (anchor) nucleotide sequence that will be used at the root of the tree to start the simulation (otherwise the root sequence is simulated according to the stationary frequencies of the specified substitution model). Nucleotide sequence alignments are evolved using INDELible (Fletcher and Yang 2009) or a version of it modified by us in order to force the sequence at the root. For more complex scenarios, NGSphy is able to read directly the output of SimPhy (Mallo et al 2015), a program that simulates multiple gene trees evolving within a species tree under incomplete lineage sorting, gene duplication and loss, horizontal gene transfer or gene conversion, plus the corresponding multilocus alignments obtained with INDELible. Before

producing the NGS reads, the genomic sequences at the tip of each gene tree are assigned to haploid (directly) or diploid (by random sampling within species) individuals.

With the trees, multiple sequence alignments and individuals in place, NGSphy can directly produce read counts for single-nucleotide variants (SNVs) or generate Illumina reads using ART (Huang et al. 2012). Read counts are produced taking into account an user-defined error rate. The sequencing coverage (depth) at each position is sampled from a Negative Binomial distribution whose mean is the sampled coverage for the corresponding species, locus and individual. For diploid individuals, coverage is further splitted among homologous chromosomes with equal probability. If desired, genotype likelihoods for every site can be computed as in GATK (McKenna et al 2010). The simulated read counts are written to a set of VCF files, one per locus. NGSphy can also call ART with the parameter values indicated in the settings file and produce Illumina reads for each individual in ALN, BAM and/or FASTQ format. ART is a very flexible program that uses customized read error model parameters and quality profiles.

NGSphy can simulate the variation in depth of coverage (herein coverage) that may occur in NGS due to differences in quantity or quality of DNA samples, technical problems when generating libraries or genomic changes in GC content. Coverage variation is implemented hierarchically. For each replicate, an experiment-wide coverage is assigned according to the user –fixed or sampled from a distribution of choice. Coverage variation among loci and individuals is then introduced independently according to Gamma distribution with a mean of one and and a shape parameter sampled from an user-defined distribution. In addition, coverage variation across species/taxa can be directly specified by the user. For targeted-sequencing experiments NGSphy can simulate off-target loci –with a percentage of the expected coverage–, uncaptured loci, or a coverage decay related to the phylogenetic distance to a selected reference sequence (Bragg et al. 2016). Finally, NGSphy allows for multi-threaded, parallel execution. In addition, it can generate job templates for execution in computational clusters.

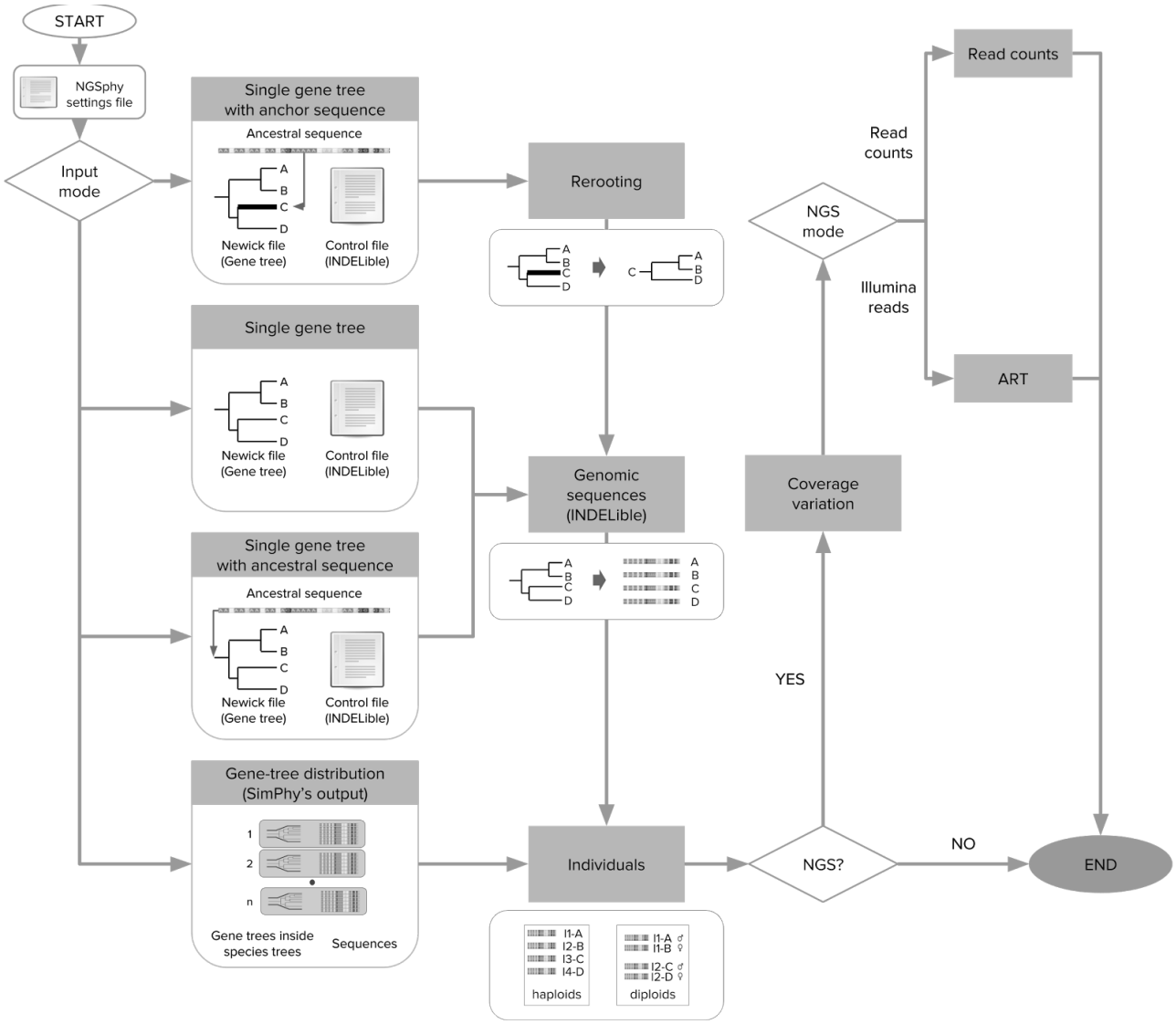
Funding

This work was supported by the Spanish Government (research grants BFU2012-33038 and BFU-BFU2015-63774-P to D.P.; FPI graduate fellowship BES-2013-067181 to M.E. and Juan de la Cierva postdoctoral fellowship FPDI-2013-17503 to S.R.) and the European Research Council (ERC-617457-PHYLOCANCER to D.P.)

Conflict of Interest: none declared.

Acknowledgements

We thank Diego Mallo for help regarding SimPhy. We also thank the members of the Phylogenomics lab (http://darwin.uvigo.es), and Rute R. da Fonseca and Anders Albrechtsen (http://popgen.dk) for their critical comments along the development of the software.



SUPPLEMENTARY FIGURE 1: NGSphy workflow.

References

- Bragg, J. G., Potter, S., Bi, K., and Moritz, C. (2016). Exon capture phylogenomics: efficacy across scales of divergence. Molecular ecology resources 16(5), 1059-1068.
- Fletcher, W. and Yang, Z. (2009) INDELible: A flexible simulator of biological sequence evolution. Molecular Biology and Evolution. 26 (8): 1879-1888.
- Huang, W., Li,L. , Myers J.R., and Marth, G.T. (2012) ART: a next-generation sequencing read simulator. Bioinformatics. 28 (4): 593-594.
- Mallo, D., de Oliveira Martins, L. and Posada, D. (2015). SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees. Systematic Biology. 65(2): 334-344.
- Mallo, D., and Posada, D. (2016). Multilocus inference of species trees and DNA barcoding. Phil. Trans. R. Soc. B., 371(1702), 20150335.
- McCormack, J. E, Hird, S. M., Zellmer, A. J, Carstens, B.C. and Brumfield, R. T. (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution. 66 526-538.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 20(9), 1297-1303.
- McTavish, E. J., Pettengill, J., Davis, S., Rand, H., Strain, E., Allard, M. and Timme, R. E. (2017) TreeToReads - a pipeline for simulating raw reads from phylogenies. BMC Bioinformatics 18:178.