# Supplementary Material

# 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.

* To whom correspondence should be addressed.

**Contact**: merlyescalona@uvigo.es

## Supplementary Material

# 1. NGSPhy workflow

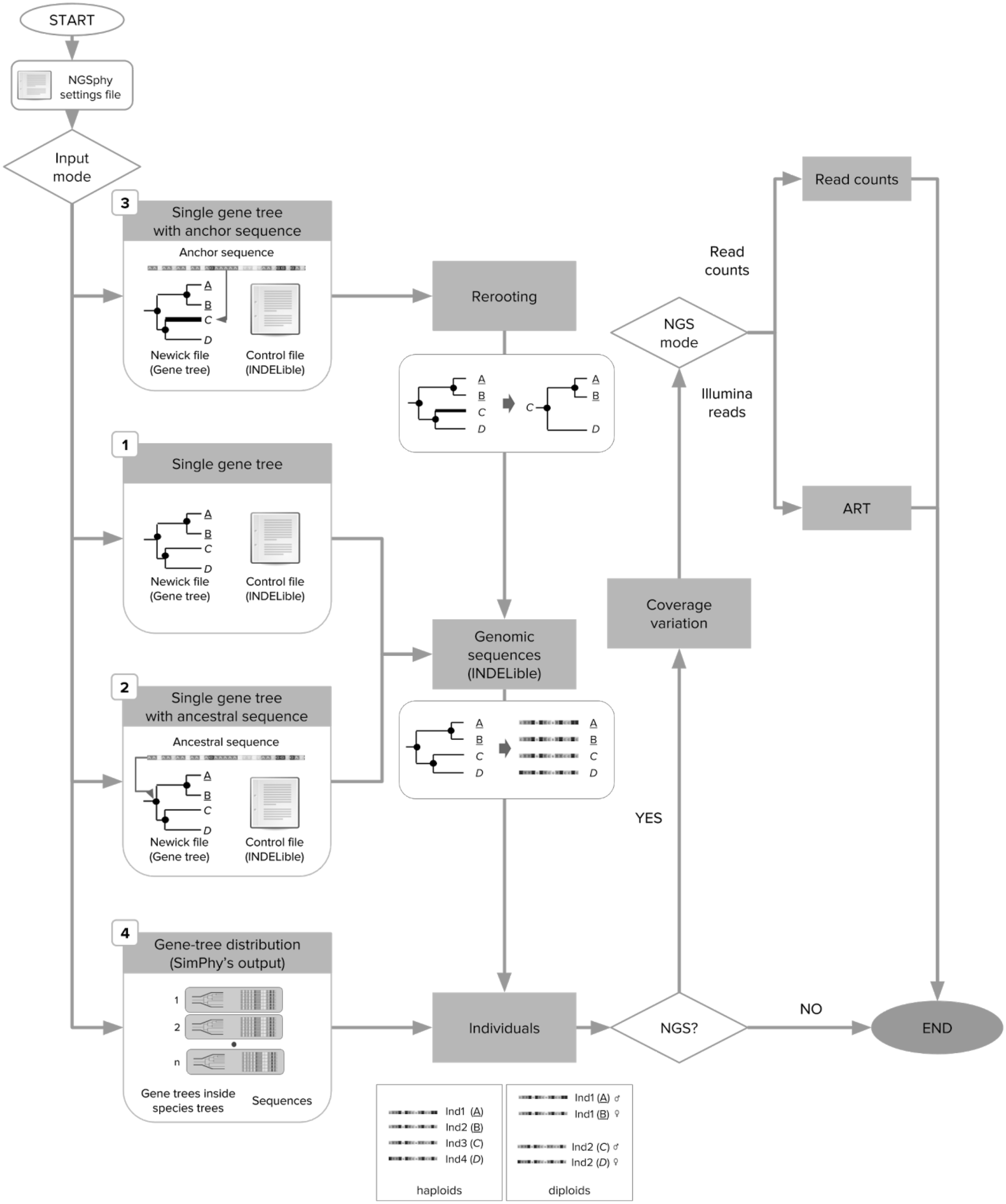

**Figure S1. NGSphy workflow**. At start, NGSphy first verifies the settings files and/or the existence of the corresponding third-party applications. If the input data corresponds to a single gene tree and an user-defined anchor sequence (input mode 3), the tree is first rooted to the selected gene-tree tip. Next, for any single gene tree input mode (input modes 1-3), nucleotide sequences are evolved under the specified substitution model resulting in sequence alignments. Then (for any input mode), haploid or diploid individuals are generated, as desired; for haploid individuals, the resulting sequences are separated into different FASTA files (per genomic fragment); for diploid individuals the sequences are randomly paired from tips within the same gene family and species, and FASTA files are thus generated each comprising both sequences of each fragment of each individual -variation of font emphasis in tip names represent differentiation in species (A, B: species 1; *C*, *D species* 2). Variation of depth of coverage at species, locus and individual level are then generated if desired, and, finally the sequencing data, either Illumina reads (with ART) or read counts (VCF files), is obtained.

## 2. Validation test: phylogenetic reconstruction from simulated alignments

To test whether NGSPhy is working as expected we performed several sanity checks and test runs. Here we describe a particular experiment to check that the simulated alignments have in fact evolved under the user-defined gene tree. The simulation process started from the gene tree in Figure S2, using the tip 1_0_0 as anchor (i.e., providing a known sequence corresponding to that tip). We ran 100 replicates of NGSphy in inputmode 3 (single gene tree with user-defined anchor sequence). The sequence alignments were simulated under a JC69 model (Jukes and Cantor 1969), equal base frequencies and a length of 1000 bp. The simulated alignments were used to reconstruct maximum likelihood (ML) trees with raxml-ng (Kozlov 2017, April 4), using the (known) JC69 model. Ten heuristic searches were performed per alignment, starting on maximum parsimony trees. The Robinson-Foulds (RF) (Robinson and Foulds 1981) and Branch Score distances (BSD) (Kuhner and Felsenstein 1994) were used to compare the input gene tree and the estimated ML trees respect to topology and branch lengths, respectively. All RF scores were always zero, while the BSD were always minimal (mean = 0.0555, standard deviation = 0.0175), suggesting that at least the alignment simulation is correct.

In the input mode used for this test, the anchor tip is used to re-root the tree, and then used by indelible-ngsphy to generate the locus alignment. This process involves the generation of a zero-branch-length between the anchor tip and what is considered the *root* node by indelible. To show that this re-rooting process was not leading to any error and that the generated *anchor* sequence is identical to the one defined by the user as *anchor*, we measured the p-distance between them. In all cases, this distance was zero.

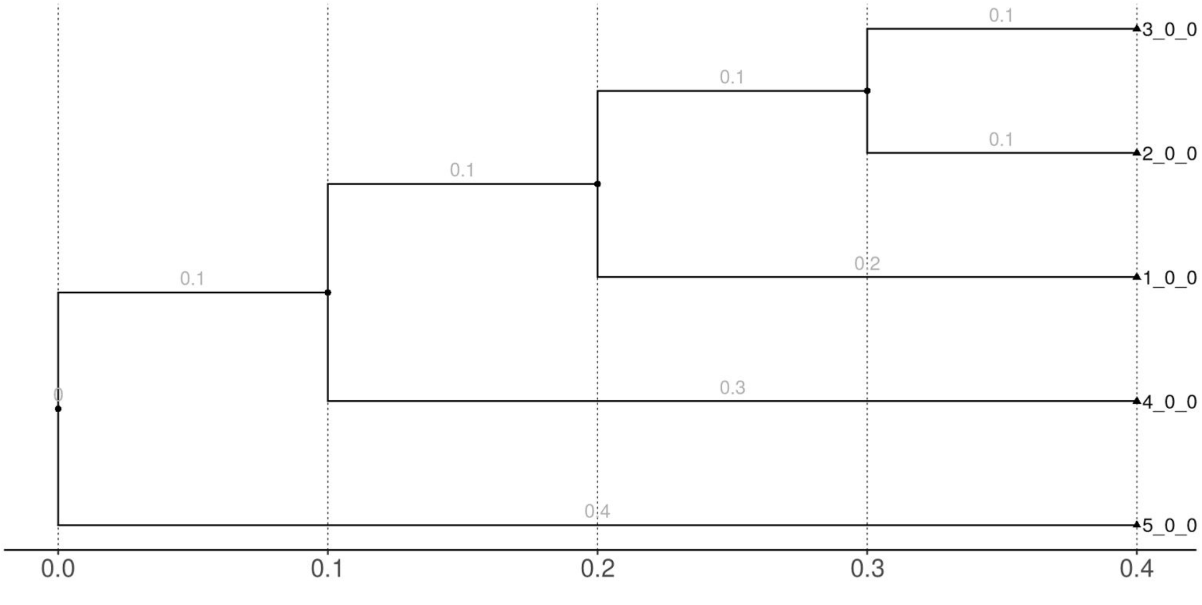

**Figure S2:** Gene-tree with five tips used for the validation. Numbers above the branches represent branch lengths in expected number of substitutions.

## 3. Use case: effect of the variation of depth of coverage on SNP recovery

One of the possible uses of NGSphy might be the optimization of depth of coverage for a given purpose. In this case we designed a small experiment to visualize the potential trade-off between NGS coverage and SNP discovery. In this case we used NGSphy to simulate a single sequence alignment from a given gene-tree (Figure S3) and from it we generated 100 NGS datasets at different depths of coverage. The sequence alignments were simulated under a JC69 substitution model, with equal base frequencies and a length of 10000 bp. The Illumina runs generated 150 bp paired-end reads for all individuals at a coverage of 2X, 10X, 50X, 100X and 200X (100 replicates for each level). The detailed settings are shown in Table S1.

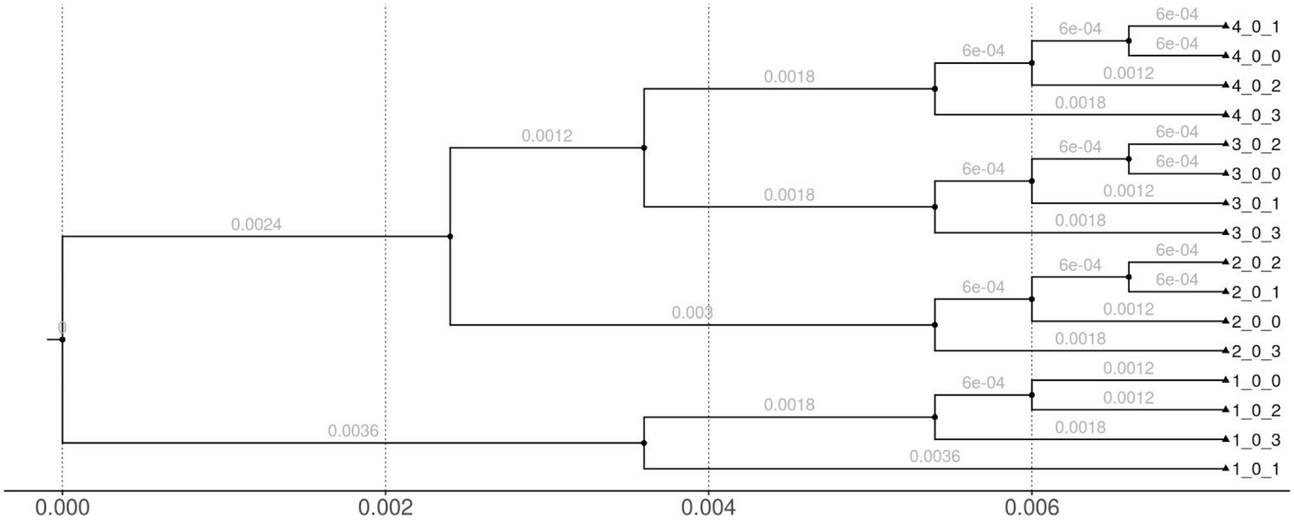

**Figure S3:** Gene-tree used for the use case simulation. It represents four species with two individuals per species. Numbers above the branches represent branch lengths, in expected number of substitutions.

Mapping was carried out using the MEM algorithm of BWA Version 0.7.7-r441 (Li and Durbin 2009), against a randomly chosen reference (sequence 1_0_2 in all cases). Following a standardized best-practices pipeline (Van der Auwera and Carneiro 2013) mapped reads from all replicates were independently processed, performing local realignment around indels and removing PCR duplicates. Variant calling was made with GATK (Mckenna et al. 2010), using the single-sample variant-calling joint-genotyping framework using the HaplotypeCaller and GenotypeVCF modules. SNP calls from each replicate were compared to the true variant sites, showing that SNP recovery increased very rapidly until 10X, when almost all true variants were called (Figure S4). All raw data and code is available at https://www.github.com/merlyescalona/ngsphy/manuscript/supp.material/scripts.

**Table S1. NGSphy settings parameters**. <coverage> varies for 2, 10, 20, 100 and 200x.

| | | |
|---|---|---|
| general | path | ./output/<coverage> |
| | output_folder_path | NGSphy_test2_<coverage> |
| | ploidy | 1 |
| data | inputmode | 1 |
| | gene_tree_file | files/supp.test2.tree |
| | indelible_control_file | files/control.supp.test2.txt |
| coverage | experiment | <coverage> |
| ngs-reads-art | fcov | true |
| | l | 150 |
| | m | 250 |
| | p | true |
| | q | true |
| | s | 50 |
| | ss | HS25 |
| execution | environment | bash |
| | runART | on |
| | threads | 2 |

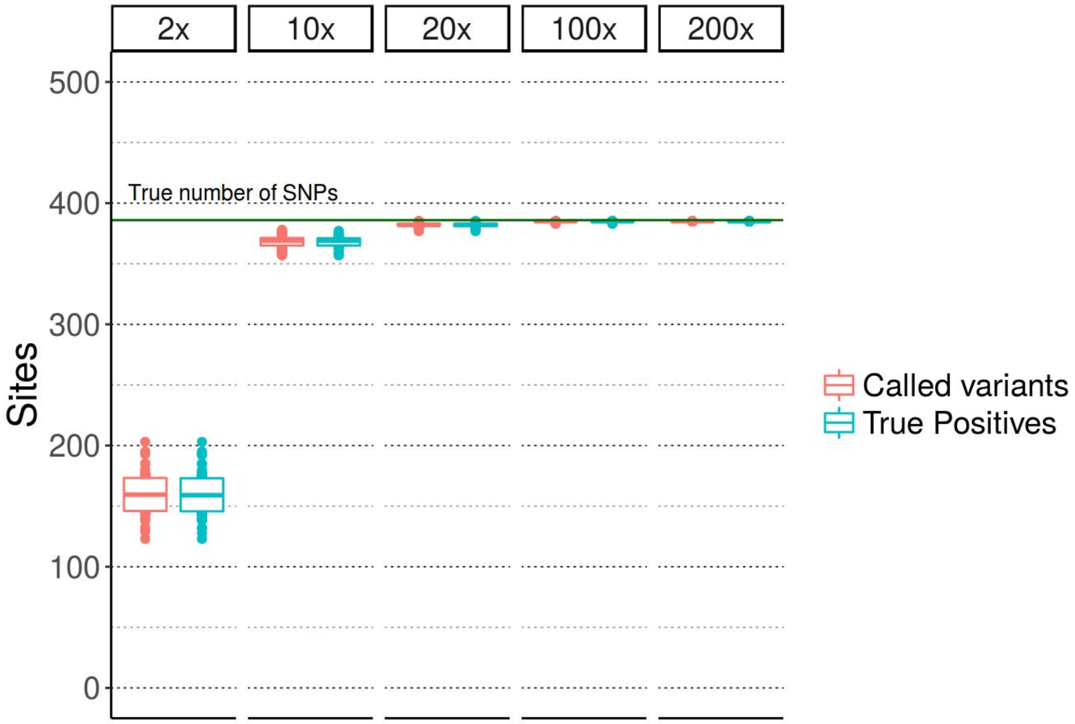

**Figure S4:** Called variants and true positives (mean and Q1/Q3) at different depths of coverage. The true number of SNPs is 386. At 100x and 200x only 1 site (average) is not recovered.

## References

Jukes T. H. and Cantor C. R. (1969) Evolution of Protein Molecules. In *Mammalian Protein Metabolism*, 21–132.

Kozlov, A (2017) *raxml-ng* (version RAxML-NG v0.2.0 BETA; April 4.). http://doi.org/10.5281/zenodo.492245.

Kuhner M. K. and Felsenstein J.( 1994) A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Molecular Biology and Evolution* 11 (3):459–68.

Li H. and Durbin R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14):1754–60.

Mckenna A. *et al*. (2010) The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. *Genome Research* 20:1297–1303.

Robinson D. F. and Foulds L. R. (1981) Comparison of Phylogenetic Trees. *Mathematical Biosciences* 53 (1):131–47.

Van der Auwera G. A. and Carneiro M. O. (2013)From FastQ Data to High-confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*. Wiley Online Library. http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1110s43/full.