Imputor User Guide

Matthew Jobin

1. Input

Imputor accepts input from two file types: FASTA files and VCF files.

1.1. Input options

Imputor handles the following input arguments:

- -file (required): The raw input data, in FASTA or VCF format. Note that VCF format files must include the optional FORMAT column with the GT subheader, and include columns where individual samples are marked.
- -ref (required): The reference sequence, in either TXT or OBJ format.
- -tree: The method for importing or constructing a tree, with the following options:
    - *.xml: Input of a file with extension .xml will invoke reading as a phyloxml format tree **(Han & Zmasek 2009)**.
    - PhyML: Invokes the PhyML software **(Guindon et al. 2010)**.
    - RAxML: Invokes the RAxML software **(Stamatakis 2014)**.
- -alpha: The value of the gamma shape parameter used in PhyML. Default is *e*.
- -boot: The number of bootstrap replicates for PhyML. Please see the **PhyML user guide** for detailed instructions . Default is 100.
- -rmodel: The model type for RAxML. Please see the **RAxML user guide** for instructions. Default is GTRCAT.
-

2. Phylogenetic Tree Construction

2.1. PhyML

Unlike methods of maximum parsimony, maximum likelihood methods make use of the branch lengths of a proposed tree .

2.1.1. How to install PhyML for use with Imputor

1. Download and install PhyML for your system as per the **PhyML user guide**.
2. Rename the resulting executable 'phyml' (no quotes).
3. Move your phyml executable to a location that is part of your system's search path. On UNIX-like systems including macOS and Linux, you will usually want to add the line **export PATH="/your path/:$PATH**" to the .bashrc or the .bash profile located in your home directory (where your path is the path to the directory that contains PhyML binary).

2.2. RAxML

2.2.1. How to install RAxML for use with Imputor

1. Download and install RAxML for your system as described in the **RAxML user guide**.
2. Compile the version best suited to your needs. Please see the RaXML docs about parallel computing, makefiles, etc.
3. Rename your compiled executable to 'raxmlHPC' (no quotes).
4. Move your raxmlHPC executable to a location that is part of your system's search path. On UNIX-like systems including macOS and Linux, you will usually want to add the line **export PATH="/your path/:\$PATH**" to the .bashrc or the .bash profile located in your home directory (where your path is the path to the directory that contains RAxML binary).

3. Imputation

Imputation methods have been used to identify mistakenly called sites due to mechanical or laboratory error **(Wang et al. 2012)**. These efforts are often made in order to “fill in” calls marked “missing” by the sequencing process **(Lippold et al. 2014)**. One of the primary methods by which imputation is carried out for Next-Generation Sequencing (NGS) is by comparing NGS sequence reads to that of Sanger sequencing, which is slower but has a lower error rate **(Wall et al. 2014)**.

3.1. Criteria for imputation

IMPUTOR imputes mutations for an input set of data via comparison of variants amongst near neighbors, via the principle of parsimony, wherein neighboring samples on a phylogenetic tree that are identical by descent (IBD) for a derived allele are unlikely to experience a reversion to the ancestral allele amongst one of their members. Under the principle of parsimony, originally introduced as the “principle of minimum evolution”, the course taken in evolutionary history is most likely to match the course that requires the fewest changes **(Edwards & Cavalli-Sforza 1964)**

4. Bibliography

Edwards, A. & Cavalli-Sforza, L., 1964. Reconstruction of evolutionary trees. *Syst. Assoc. Publ.*, No. 6, pp.67–76.

Guindon, S. et al., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3), pp.307–321. Available at: http://sysbio.oxfordjournals.org/content/59/3/307.full.

Han, M.V. & Zmasek, C.M., 2009. phyloXML: XML for evolutionary biology and comparative genomics. *BMC bioinformatics*, 10(1), p.356. Available at: http://www.biomedcentral.com/1471-2105/10/356.

Lippold, S. et al., 2014. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investigative Genetics*, 5(1), p.13. Available at: http://investigativegenetics.biomedcentral.com/articles/10.1186/2041-2223-5-13.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312–1313. Available at: http://bioinformatics.oxfordjournals.org/content/30/9/1312.full.

Wall, J.D. et al., 2014. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Research*, 24(11), pp.1734–1739. Available at: http://genome.cshlp.org/content/24/11/1734.full.

Wang, J.R. et al., 2012. Imputation of Single-Nucleotide Polymorphisms in Inbred Mice Using Local Phylogeny. *Genetics*, 190(2), pp.449–458. Available at: http://www.genetics.org/content/190/2/449.abstract.