AnTE User Manual

Probabilistic Ancestral Network Reconstruction

for Transposable Elements

version 1.0

Corey Cox and Aaron Wacholder

**Before you run AnTE**

1. Align and trim TE sequences.
   * Sequences can be annotated in the genome using any annotation tool, such as ESP-Clouds or RepeatMasker, and retrieved using associated tools (e.g. bedtools).
   * In most cases, family consensus sequences will need to be aligned to form a master consensus sequence and all sequences will be profile aligned to this master consensus sequence. This alignment can be performed with any appropriate alignment software but the result should be in FASTA format.
   * The final adjustment and trimming currently must be performed manually.
     1. First, we remove all uninformative gaps.
     2. Second, the beginning and end of the sequences must be trimmed so that all sequences have the same length.
   * Optionally, you may provide subfamily assignments from another assignment software (e.g. Coseg) for comparison.
2. Expand the archive
   * Move the archive to the desired destination and expand the archive. The archive will expand into a folder named AnTE.
   * MacOSX/Linux: tar -xzvf AnTE.tar.gz
3. Compile AnTE
   * AnTE is distributed as uncompiled C++ (std 11) source code. It should compile with any standard C++ compiler using the 2011 standard.
   * MacOSX: clang++ -std=c++11 -stdlib=libc++ main.cpp -o AnTE
   * Linux: c++ -std=c++11 main.cpp -o AnTE
4. Adjust the control file (See Appendix C for more detail)
   * At the very least you will need to adjust the file locations and names of your data files and adjust the subfamily information.
     1. ALIGNMENTFILEPATH – Path and file name of the aligned trimmed sequence file (created in Step 1).
     2. SUBFAMILYFILEPATH – Path and file name of a previous subfamily assignment (optional)
     3. FORCEDCANDIDATEFILEPATH – Path and file name of candidate ancestral sequences to be forced into the starting set. (Optional)
     4. SUBFAMILYNAMES – Names of subfamilies supplied from alternate ancestry software.
     5. FORCEEXCLUDE – Force the listed sites to be excluded from the set of discriminatory sites. (optional)
     6. FORCEINCLUDE – Force the listed sites to be included in the set of discriminatory sites. (optional)
   * You may also choose to change the number of generations for running the MCMC but we have provided what we expect to be reasonable defaults.

**Running AnTE**

1. AnTE is run from the command line. We recommend keeping the executable and data files in the same folder and executing the program from there.
   * Run AnTE: ./AnTE
   * You may also run specify a range on the command line when you run AnTE, such as sampling the range 20-30 candidate ancestors.
     1. Specify ancestor range AnTE: ./AnTE 20 30
     2. This command will run the chain for ten different number of allowed ancestral sequences, with a fixed number of ancestors from twenty to thirty. You may also run multiple replicates per dimension.
   * All results generated from the run will be placed in the folder the executable was run from.
2. If you encountered problems with running AnTE please consult Appendix D for known issues. If the problem you are experiencing is not found their please contact the authors at [www.evolutionarygenomics.com](http://www.evolutionarygenomics.com).

# Appendix A – Input file descriptions

**INPUT**

Input file is a fasta-aligned file with all sequences containing large indels removed from the dataset, also trimmed at the beginning and end to create a complete alignment set. All of the subfamilies are merged into one aligned file and the above processing applied to the complete dataset.

*Sequences\_redo\_all\_final.fas* – Actual data file used for the analysis that Aaron presented. Aligned, culled and trimmed as mentioned above. File name/location specified in controls.txt

*Forced\_candidates.txt* – sequences of forced candidates. These do not have sequence name identifiers (no FASTA) and are pre-masked to contain only discriminatory sites.

*prior\_subfam\_assignment.csv* – prior assignment of all sequences to subfamilies as .csv format file (CoSeg)

**CONTROL**

*controls.txt* – constants and filenames for analysis. Must be named controls.txt and located in same folder.

# Appendix B – Output file descriptions

**GENERAL OUTPUT**

*consensus* – Description of the consensus sequence and the masked and unmasked portions on the consensus sequence.

*discrims* – site-by-site information on the discriminatory sites used for analysis.

*mask* – contains the position of all discriminatory sites in the alignment

*mutcounts* – statistics on mutation counts

*numsubs* – number of substitutions from consensus at each position along the alignment.

*priorsubfaminfo* – information about prior subfamily assignments, if provided.

*seqdata* – number of times a nucleotide was observed in each position and calculation of consensus sequence.

*seqinfo* – information on all of the sequences including ID, label, full sequence, unmasked sequence, masked sequence, subfamily, subfamily size, distance from consensus.

*sequence.fasta* – fasta file with label (sequence name) and sequence. This is an output of the sequences used and should match the input sequences fasta file.

*trans* – the final nucleotide substitution matrix.

**CHAIN-SPECIFIC OUTPUT**

For each fixed number of ancestors in the range specified at the command line one of the following files will be generated in the format *filename\_X\_Y*. Where X is the fixed number of ancestors and Y is the iteration number of the current run.

*data\_50\_0* – Data summary file calculating mean frequency of replication and probability of being replicative of all candidate ancestors.

*fullpost\_50\_0* - posterior probabilities each element was replicated from each candidate ancestral sequence

*loglike\_50\_0* – step, beta, and log likelihood for thermodynamic integration. Thermondynamic integration is disabled in default control file, so this file will be empty unless enabled.

*mcmc\_steps\_50\_0* – Record file of log-likelihood and ancesetral frequency for each candidate ancestor for each sampling of the MCMC chain.

*meleminform\_50\_0* – ID, frequency, and masked sequence for each unique masked element

*newcands\_50\_0* – Candidates from search performed during burn-in. If no search during burn-in this file will be blank.

*post\_50\_0* – Stepwise posterior ancestral probabilities.

*seqinform\_50\_0* – Sequence information file. Contains id, label, sequence, masked sequence, unmasked sequence, length, prior subfamily assignment, new subfamily assignment, master element id, age, deviation from ancestor.

*trans\_50\_0* – Log transition probability of each candidate versus each discriminatory site sequence.

*tree\_50\_0* - Position of all sequences in the tree, -2 if the sequence is not in the tree.

# Appendix C – Control File Description

**CONTROL PARAMETER DESCRIPTIONS**

Parameter Name Default Value

SDEVBOUND 3.0

Discriminatory sites are assigned if observed variation from consensus at a site exceeds this value.

MAGNIFICATION 2.0

Mutation rate magnification factor on off-diagonal entries of the nucleotide transition matrix.

MINSUBSPROB 0.01

Minimum substitution probability

GENERATIONS 1000000

Number of generations to run the chain

REPORTFREQ 1000

Number of generations between sampling

BURNIN 100000

Number of generations to run the chain before we begin sampling

STARTCANDIDATESEARCH 4200000

Time to start candidate search, if > burnin no candidate search

SEARCHFREQ 1000

Search Frequency

MELTTIME 1000000

Time to begin melting integration; if exceeds number of generations, thermodynamic integration will not occur

INCREMENTBETAFREQ 1000

Frequency for incrementing beta in thermodynamic integration

RANDOMSEED 42

Starting random seed (change for unique runs)

DATALESS 0

If 1, run the chain with no data

CONSTDIM 60

Number of dimensions (number of allowed ancestral sequences), can be set on command line

ALIGNMENTFILEPATH ./sequence\_all\_070313\_ag.fas

Path and name of the sequence data file

SUBFAMILYFILEPATH ./prior\_subfam\_assignment.csv

Path and name of the subfamily identification from other software

FORCEDCANDIDATEFILEPATH ./forced\_candidates.txt

File and path of candidates to force into starting candidate set

SUBFAMILYNAMES sub\_0,sub\_1,sub\_2,sub\_3,sub\_5,sub\_6,sub\_7,sub\_8, sub\_9,sub10,sub11,sub12,sub13,sub14,sub15

Names of initial subfamilies

FORCEEXCLUDE 4,9,22,60,88,192,199,265,313,360

Sites excluded from consideration for being discriminatory sites

FORCEINCLUDE 7,24,35,56,62,67,88,94,97,98,108,122,124,182,239,240,293,300,301,302,303,370

Ancestral candidates that must be included among discriminatory sites.

**DEFAULT CONTROL FILE**

SDEVBOUND 3.0

MAGNIFICATION 2.0

MINSUBSPROB 0.01

GENERATIONS 1000000

REPORTFREQ 1000

BURNIN 100000

STARTCANDIDATESEARCH 4200000

SEARCHFREQ 1000

MELTTIME 1000000

INCREMENTBETAFREQ 1000

RANDOMSEED 42

DATALESS 0

CONSTDIM 60

ALIGNMENTFILEPATH ./sequence\_all\_070313\_ag.fas

SUBFAMILYFILEPATH ./prior\_subfam\_assignment.csv

FORCEDCANDIDATEFILEPATH ./forced\_candidates.txt

SUBFAMILYNAMES sub\_0,sub\_1,sub\_2,sub\_3,sub\_5,sub\_6,sub\_7,sub\_8,sub\_9,sub10,sub11,sub12,sub13,sub14,sub15

FORCEEXCLUDE 4,9,22,60,88,192,199,265,313,360

FORCEINCLUDE 7,24,35,56,62,67,88,94,97,98,108,122,124,182,239,240,293,300,301,302,303,370

# Appendix D – Known Issues

1. On MacOSX, if the control file does not have a space at the end of each line, all lines after the first without a space will not be imported into the software and the AnTE will run with the hard-coded defaults.