

Multiple Sequence Alignment

with

PASTA

Michael Nute

Austin, TX

June 17, 2016

Agenda

- Installation recap & Dendropy version issues
- Quick recap of PASTA Algorithm
- Run the GUI
- Explore GUI options and what they do in terms of PASTA
- Run a test alignment
- Explore PASTA outputs and diagnostics
- Run a *different* test alignment
- Compare the PASTA fill-in-the-blank defaults for the two test alignments

PASTA: Installation

We hope everybody has been able to install PASTA based on instructions from our email. If not:

See detailed installation instructions at:

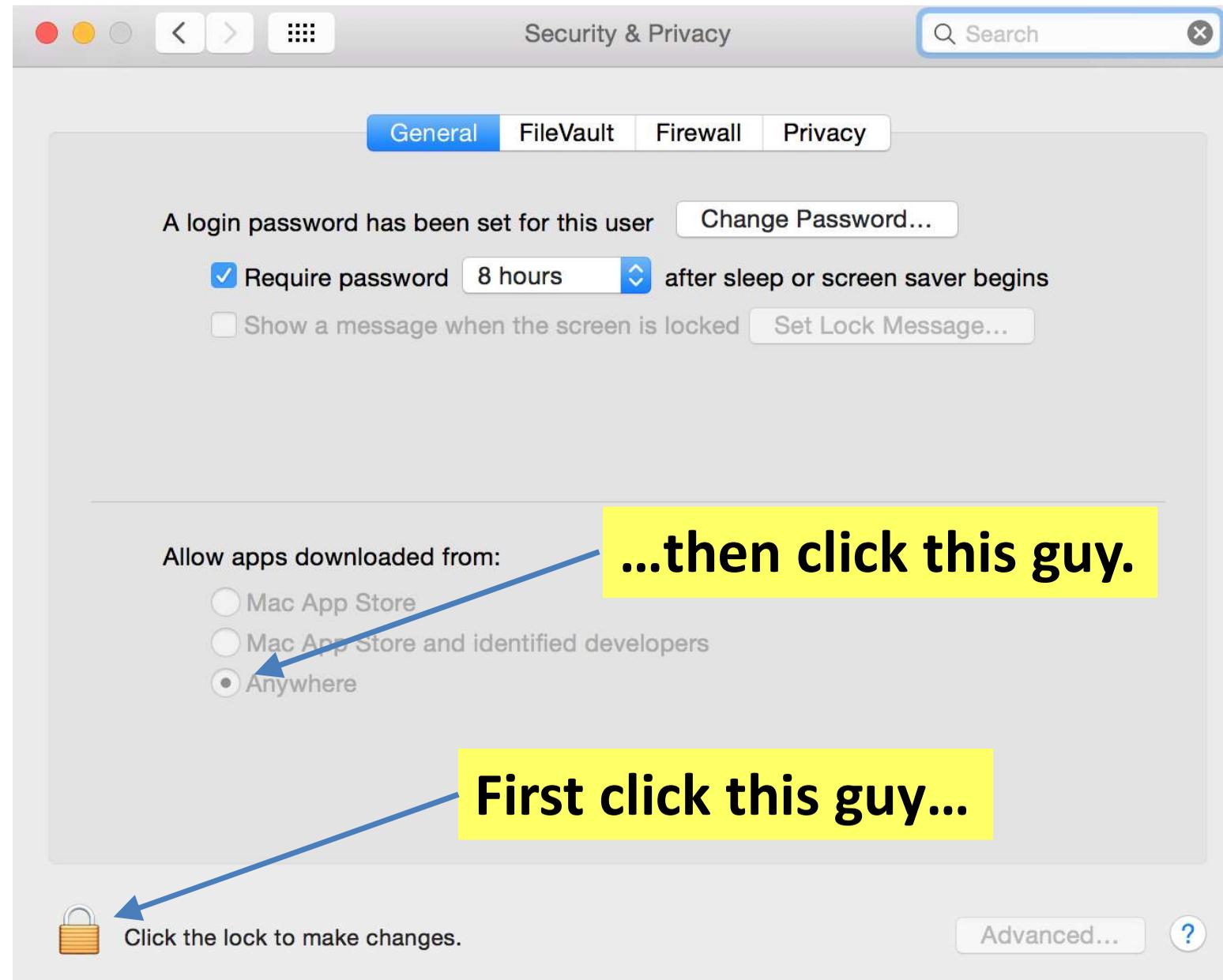
<https://github.com/smirarab/pasta>

Three Options:

- 1) MAC
 - DMG file available at the link above
- 2) Linux
 - Detailed instructions available at the link above
 - Requires JAVA, wxPython,
- 3) Virtual Machine (Recommended: VirtualBox)
 - Virtual appliance available at link above
 - This is the only option for Windows users

PASTA: Installing with a Mac

- If you have a mac, download the .dmg file (link is on the github page).
- Open the dmg and copy the app to a folder you want to use.
- You may need to update settings to allow you to run software downloaded from the internet. Go to **System Preferences** → **Security & Privacy** and you will see the screen on the right:



PASTA: Dealing with DendroPy version issues

- PASTA currently depends on an older version of DendroPy (3.12) and will not run if the newest version (4.0+) is the primary installed version
- To fix this, we need to find a suitable location, download the old code and point Python to that location before running PASTA.
 - Easiest way:
 - Go to your pasta-code folder (the one containing the “pasta” install folder)
 - Download the dendropy 3.12 source & extract it
 - Link that folder to the root pasta folder.
 - Sample commands. See:
http://publish.illinois.edu/michaelnute/files/2014/10/commands_to_fix_dendropy.txt
 1. `wget https://github.com/jeetsukumaran/DendroPy/archive/v3.12.1.zip`
 2. `unzip v3.12.1.zip`
 3. `ls -l DendroPy-3.12.1 <-- VERIFY THAT “dendropy” SUBFOLDER EXISTS`
 4. `dpydir=$(pwd)/DendroPy-3.12.1/dendropy`

PASTA: Dealing with DendroPy version issues

At this point you have two options:

1. Establish a link to the DendroPy 3.12 in the pasta install folder

- This is the easiest and is a permanent solution, but it might cause this to be your default DendroPy installation, which would break anything using DendroPy 4.*

- **Command:**

```
ln -s $dpydir ./pasta/dendropy
```

2. Manually add this to your python path

- This won't break anything, but you'll have to do this every time you want to use PASTA

- **Command:**

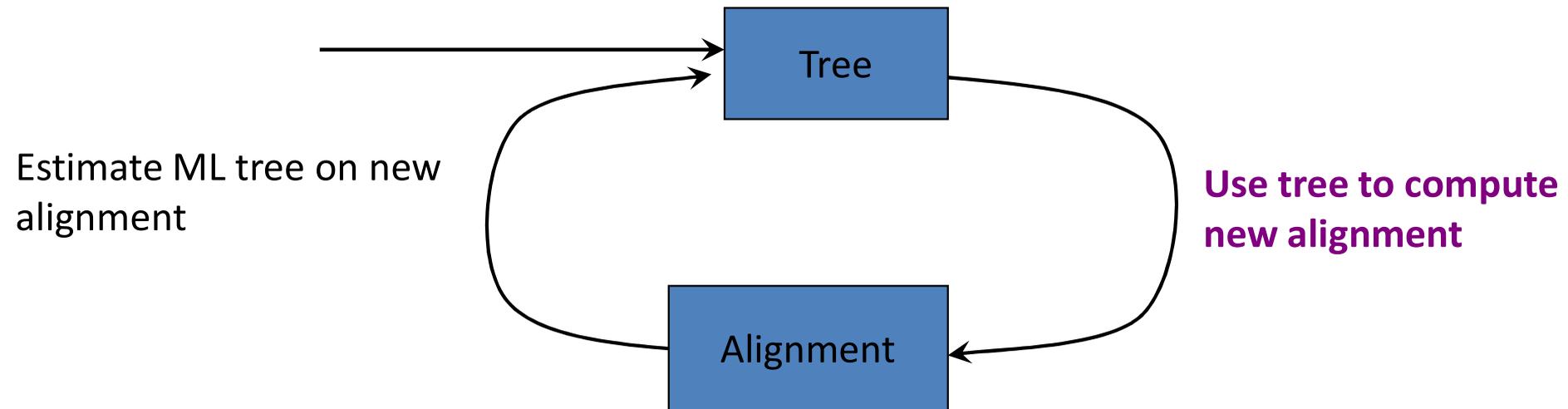
```
export PYTHONPATH=$dpydir:$PYTHONPATH
```

Test your installation: `python ./pasta/run_pasta.py -h`

I promise that a permanent, easy fix for this is coming soon.

SATé and PASTA Algorithms

Obtain initial alignment and
estimated ML tree

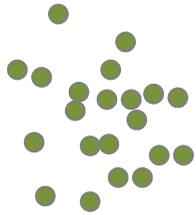


Repeat until termination condition, and
return the alignment/tree pair with the best ML score

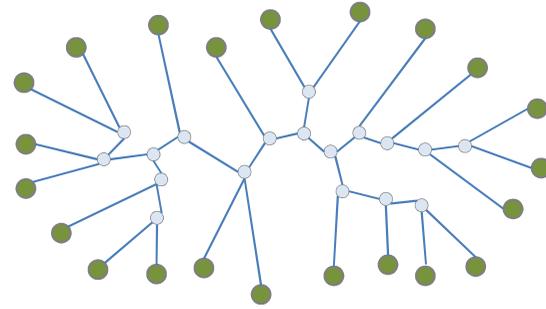
PASTA Algorithm

Input: unaligned sequences

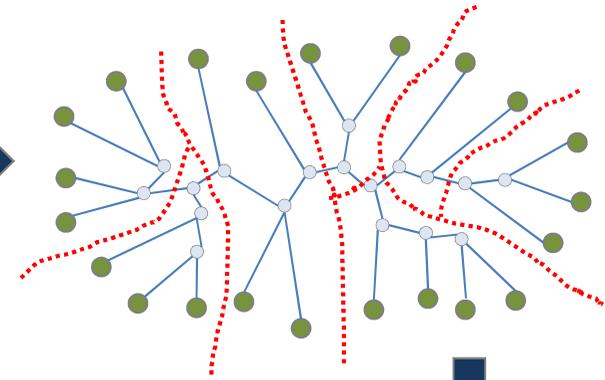
1) Get initial alignment



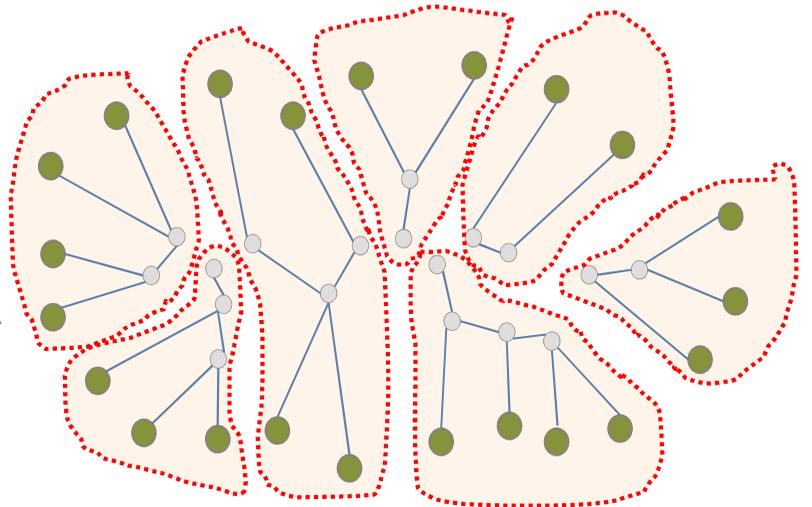
2) Estimate tree on current alignment



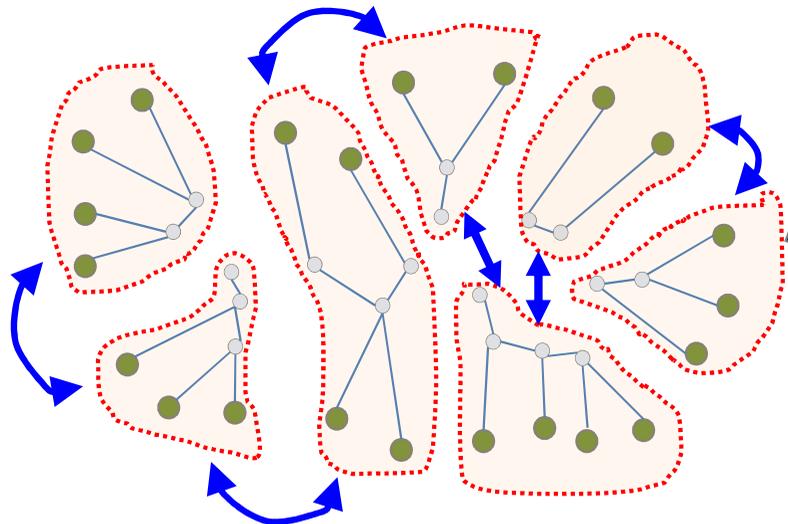
3) Break into subsets according to tree



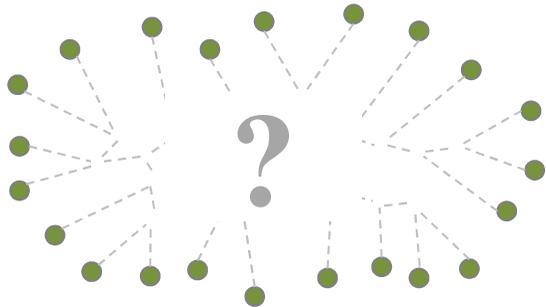
4) Use external aligner to align subsets



5) Use external profile aligner to merge subset alignments



6) Use transitivity to merge subset pairs into a full alignment, scrap the old tree



(repeat)

PASTA GUI (Linux version, but Mac looks the same)

PASTA - Practical Alignment using SATe and TraAnsitvity

External Tools

- Aligner: MAFFT
- Merger: MUSCLE
- Tree Estimator: FASTTREE
- Model: GTR+G20

Job Settings

- Job Name: pastajob
- Output Dir.:
- CPU(s) Available: 1
- Max. Memory (MB): 1024

Sequences and Tree

- Sequence file ...
- Multi-Locus Data
- Data Type: DNA
- Initial Alignment Use for initial tree
- Tree file (optional) ...

Workflow Settings

- Algorithm Two-Phase (not PASTA)
- Post-Processing Extra RAXML Search

PASTA Settings

- Max. Subproblem: Percentage (50) Size (200)
- Decomposition: Centroid
- Time Limit (hr) (24)
- Iteration Limit (100)
- Return: Final

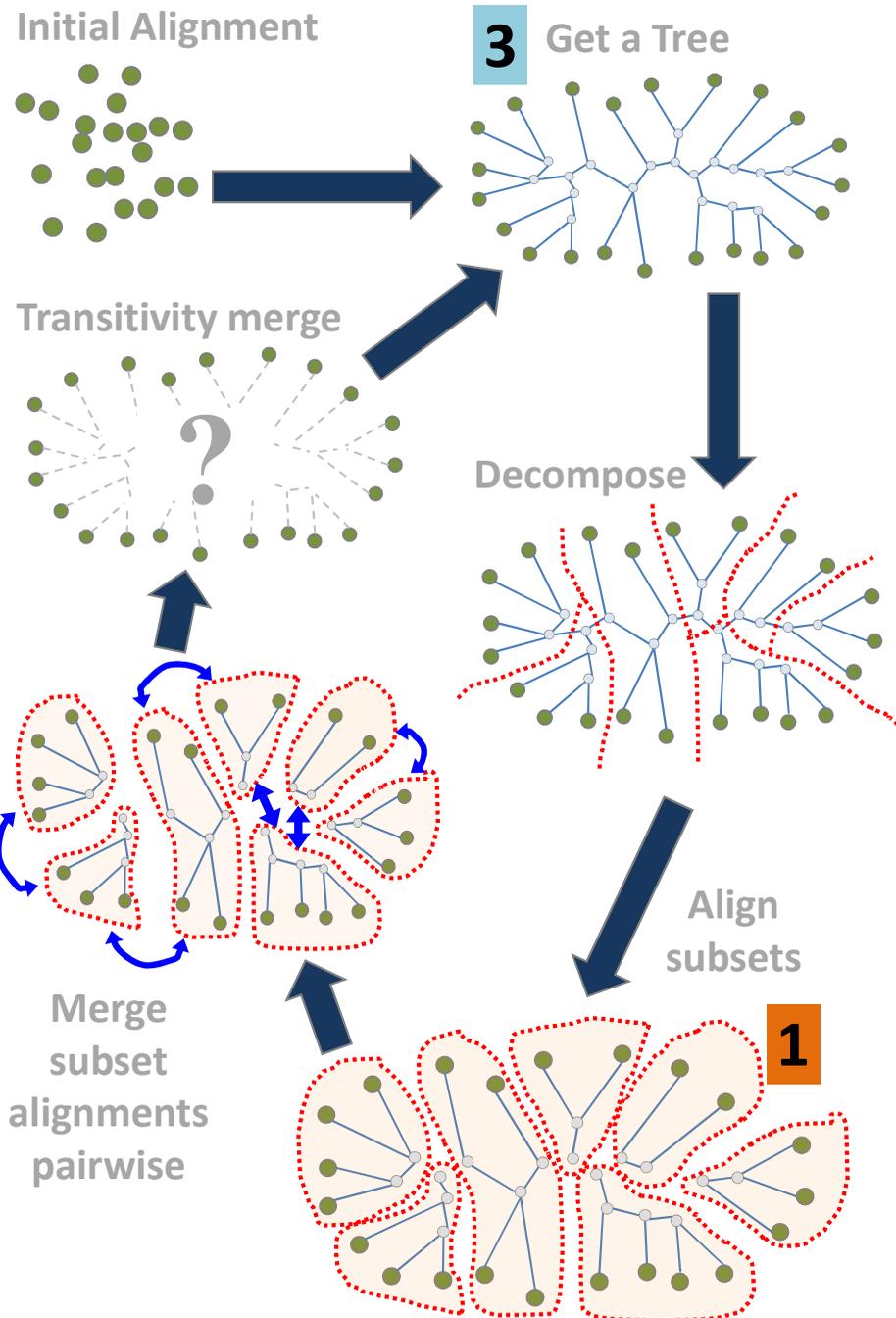
Start

PASTA 1.6.4, 2013-2016

Running Log (2016-06-15 22:37:41 CST)

PASTA Ready!

PASTA Algorithm



PASTA GUI

The screenshot shows the PASTA GUI interface with the following settings:

- External Tools:**
 - Aligner: MAFFT (labeled with a yellow '1')
 - Merger: MUSCLE (labeled with a yellow '2')
 - Tree Estimator: FASTTREE (labeled with a yellow '3')
 - Model: GTR+G20
- Job Settings:**
 - Job Name: pastajob
 - Output Dir.:
 - CPU(s) Available: 1
- Sequences and Tree:**
 - Sequence file ...
 - Multi-Locus Data:
 - Data Type: DNA
 - Initial Alignment: Use for initial tree
 - Tree file (optional) ...
- Workflow Settings:**
 - Algorithm: Two-Phase (not PASTA)
 - Post-Processing: Extra RAXML Search
- PASTA Settings:**
 - Max. Subproblem: Percentage (50) / Size (200)
 - Decomposition: Centroid
 - Time Limit (hr): 24
 - Iteration Limit: 100
 - Return: Final

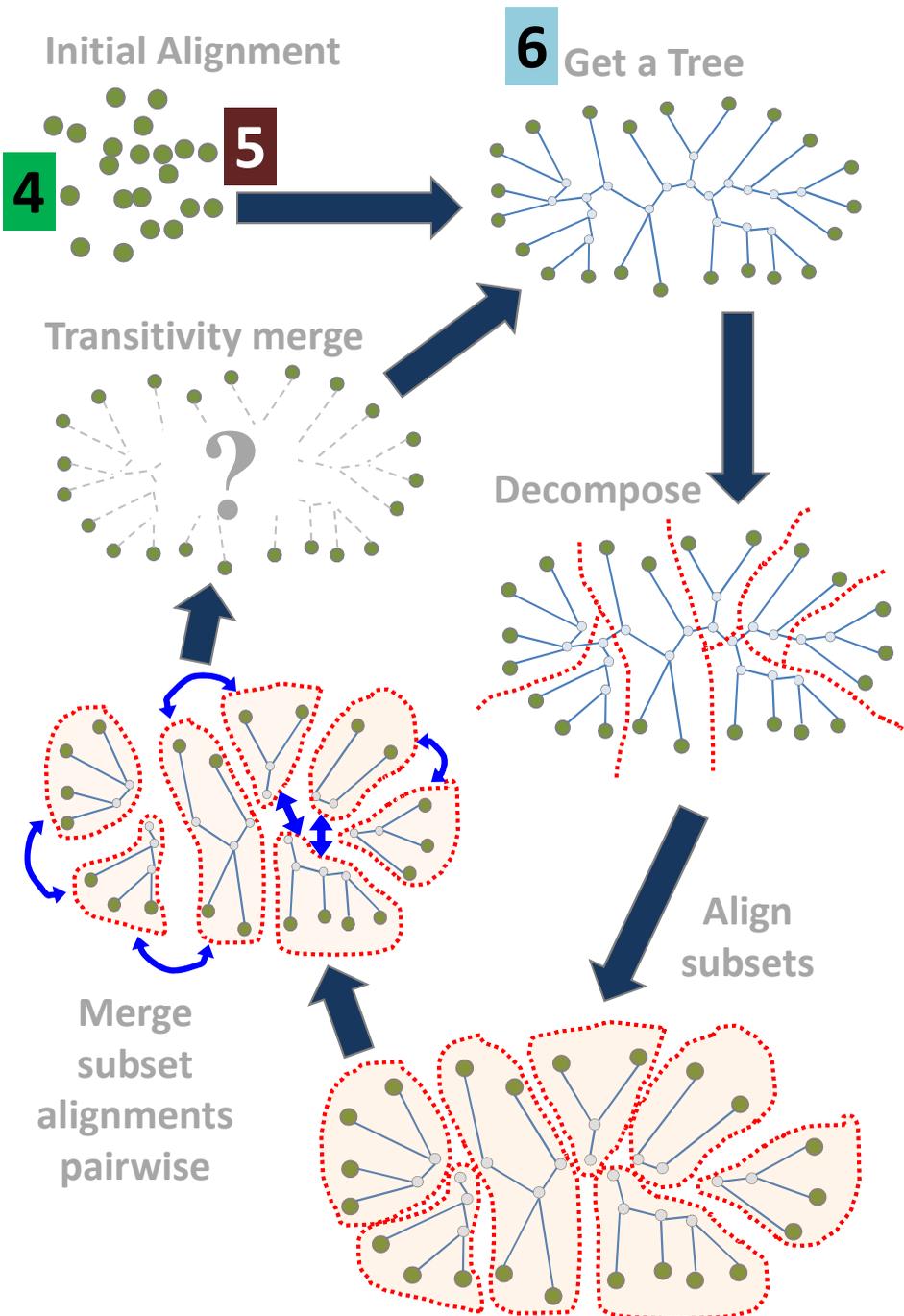
A yellow callout bubble points to the Tree Estimator dropdown, stating: "This applies to the Tree Estimator in particular".

At the bottom of the GUI, the following text is visible:

```
PASTA 1.6.4, 2013-2016
Running Log (2016-06-15 22:37:41 CS
PASTA Ready!
```

- 1) This is the alignment tool used to align the subsets (several options).
- 2) Tool for merging two subset alignments. (OPAL or MUSCLE)
- 3) Tool to estimate a maximum likelihood tree (FastTree or RAXML)

PASTA Algorithm



PASTA - Practical Alignment using SATe and TraAnsitivty

This should be checked if the sequence file (4) should be treated as aligned. If not checked, PASTA will generate a fast progressive alignment to start.

The basic input to the problem: FASTA file with sequences in need of alignment

Data type (DNA, RNA or Protein)

The user can provide a starting tree that will cause the algorithm to skip the initial alignment step.

Sequence file ... **4**

~~Multiple sequence data~~ <-- not implemented yet

Data Type DNA **5**

Initial Alignment Use for initial tree

Tree file (optional) ... **6**

Workflow Settings

Algorithm Two-Phase (not PASTA)

Post-Processing Extra RAXML Search

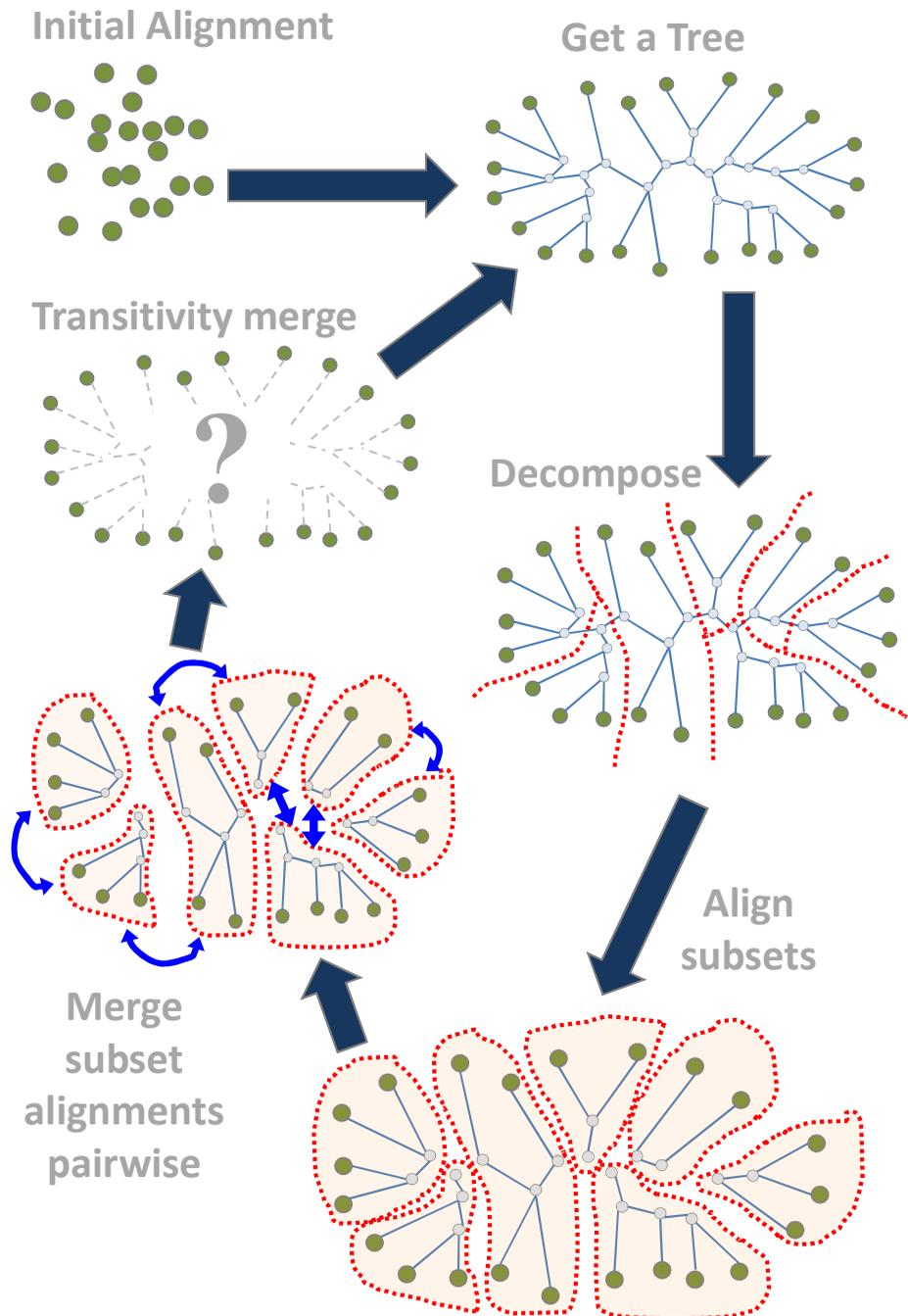
Start

PASTA 1.6.4, 2013-2016

Running Log (2016-06-15 22:37:41 CST)

PASTA Ready!

PASTA Algorithm



Basic administrative settings:
Job Name – all output files will start with this name.
Output Dir – folder where output files will go.
CPUs – number of processors
Max. Memory (MB) – only applies to Java when OPAL is called.

External Tools
Aligner: MAFFT
Merger: MUSCLE
Tree Estimator: FASTTREE
Model: GTR+G20

Job Settings
Job Name: pastajob
Output Dir.:
CPU(s) Available: 1
Max. Memory (MB): 1024

Sequences and T...
Sequence fi...
Da...
Initial Ali...
Tree file (optiona...)

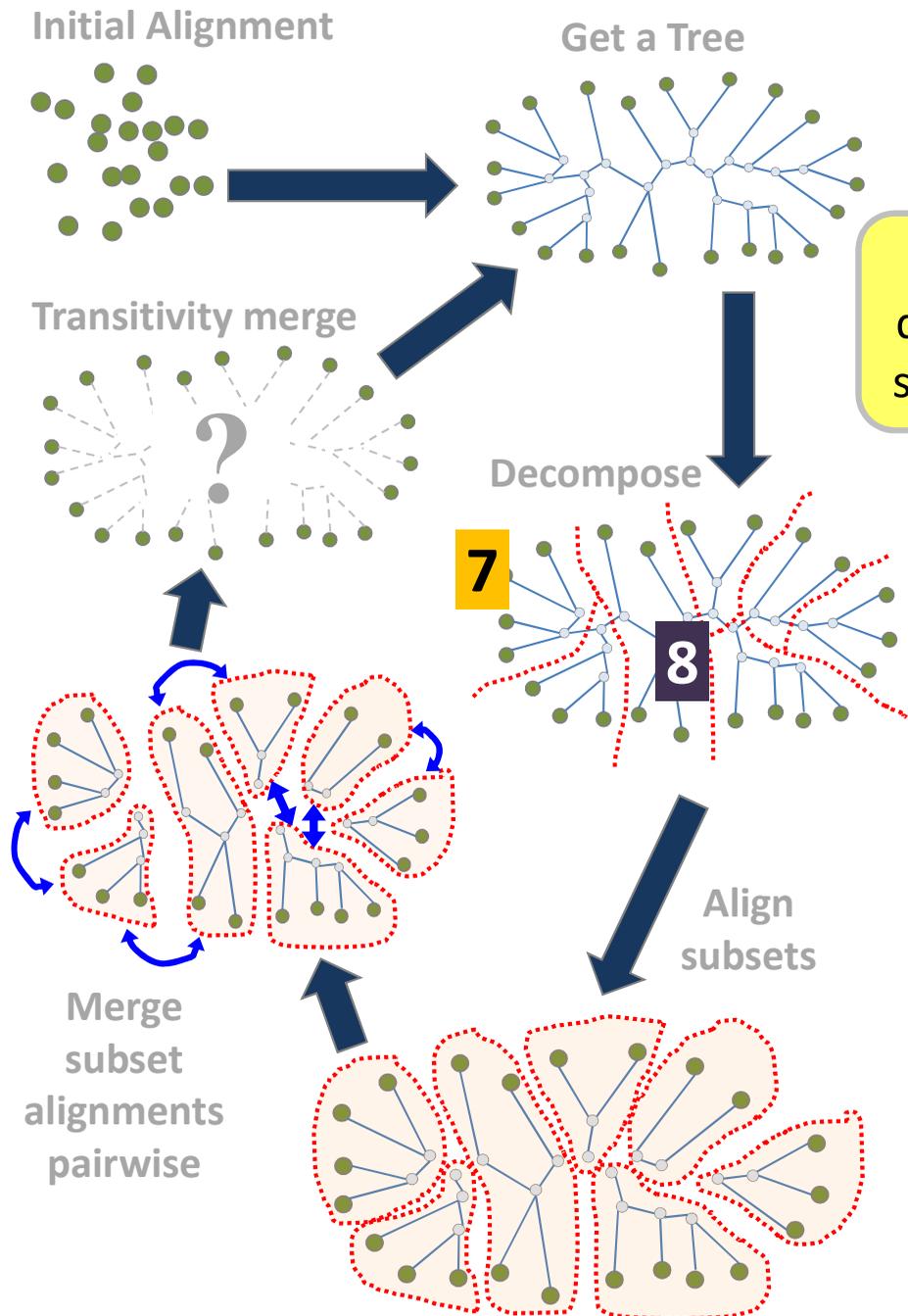
Workflow Settings
Algorithm: Two-Phase (not PASTA)
Post-Processing: Extra RAXML Search

Iteration Limit: 100
Return: Final

Start

PASTA 1.6.4, 2013-2016
Running Log (2016-06-15 22:37:41 CST)
PASTA Ready!

PASTA Algorithm



Decomposition Steps:

- Start by choosing a branch according to the **Decomposition** option (Centroid or Longest Branch).
- For each of the two subsets created, if the number of taxa is greater than **Max. Subproblem**, then repeat on that subset.

Stopping criteria for the decomposition. Can be either a fixed size or a percentage of the total taxa.

How to decide where to bisect the tree, (either Centroid Edge or the Longest Branch).

The screenshot shows the PASTA software interface with the following settings:

- CPU(s) Available:** 1
- Max. Memory (MB):** 1024
- Max. Subproblem:** Percentage (7) with a value of 50.
- Decomposition:** Centroid (8).
- Iteration Limit:** 100 (checked).
- Time Limit (hr):** 24
- Final:** Final

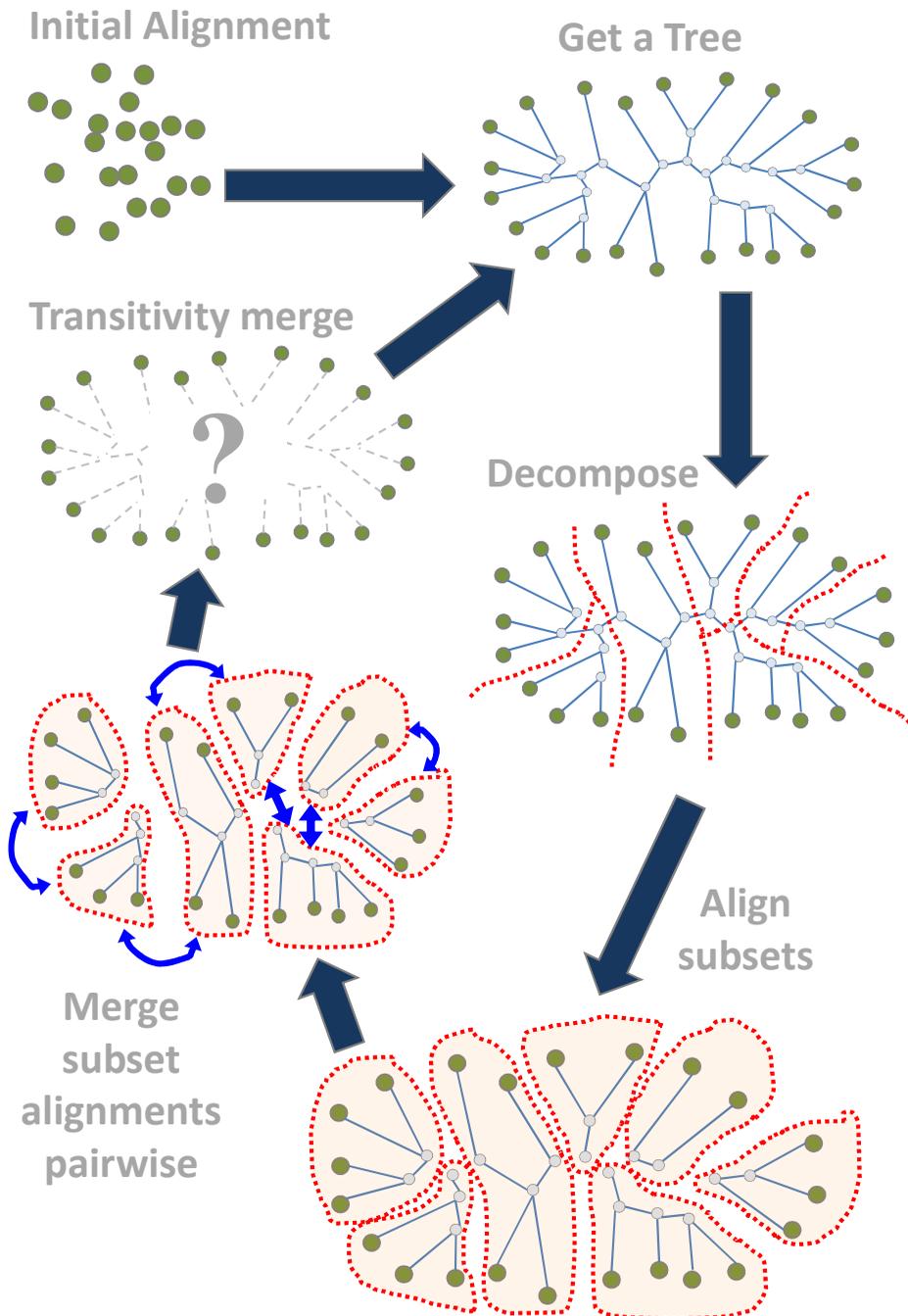
Additional interface elements include: Data Type: DNA; Multi-Locus Data: unchecked; Extra RAXML Search: unchecked; and a Start button.

Running Log (2016-06-15 22:37:41 CST):

PASTA 1.6.4, 2013-2016

PASTA Ready!

PASTA Algorithm



The screenshot shows the PASTA software interface with the following settings and annotations:

- External Tools:**
 - Aligner: MAFFT
 - Merger: MUSCLE
 - Tree Estimator: FASTTREE
 - Model: GTR+G20
- Job Settings:**
 - Job Name: pastajob
 - Output Dir.:
 - CPU(s) Available: 1
 - Max. Memory (MB): 1024
- Sequences and Tree:**
 - Sequence file ...
 - Multi-Locus Data
 - Data Type: DNA
 - Initial Alignment: Use for initial tree
 - Tree file (optional) ...
- Workflow Settings:**
 - Algorithm: Two-Phase (not PASTA)
 - Post-Processing: Extra RAXML Search
- PASTA Settings:**
 - Max. Subproblem: Percentage (50) / Size (200)
 - Decomposition: Centroid
 - Time Limit (hr): 24
 - Iteration Limit: 100
 - Return: Final

Annotations (yellow boxes):

- "(see below)" points to the "Tree file (optional) ..." field.
- "When to Stop Running?" points to the "Iteration Limit" setting.
- "Should final tree be RAXML?" points to the "Post-Processing" section.
- "Which iteration to return? (Final or Highest Likelihood)" points to the "Return" dropdown.

Running Log (2016-06-15 22:37:41 CST):

PASTA Ready!

Two-Phase search is simply 1) run an alignment, 2) get a tree from it. This is completely different than PASTA and *if this is checked, PASTA (formally) will not be run.*

Example 1: small.fasta

Step 1: Read in the data.
Located at <pasta-
folder>/data/small.
fasta

The screenshot shows the PASTA software interface. The main window has several settings: Aligner (MAFFT), Merger (OPAL), Tree Estimator (FASTTREE), and Model (GTR+G20). A 'Choose sequences...' dialog box is open, showing a file browser with 'phylolab' selected in the breadcrumb path. The file list shows 'small.fasta' selected. A 'Read input data now?' dialog box is also present, with 'OK' selected. The bottom panel shows the execution log.

This is the PASTA install folder on the Virtual Machine

Name	Size	Modified
figwasps		06/15/2014
hummingbirds		06/15/2014
16S.E.ALL.referene.fasta	1.9 MB	06/15/2014
16S.E.ALL.unaligned.fasta	544.0 KB	06/15/2014
anolis.fasta	8.8 KB	06/15/2014
BBA0067-half.input.fasta	93.0 KB	06/15/2014
large.fasta	1.3 MB	06/15/2014
pythonidae.fasta	86.2 KB	06/15/2014
small.fasta	52.6 KB	06/15/2014

Read input data now?
Do you want PASTA to read the data now? (this causes PASTA to customize some of the settings for your data).

PASTA 1.6.4, 2013-2016
Running Log (2016-06-17 03:40:01 CST)
Read 1 file(s) with aligned DNA data. Total of 32 taxa found.
Parsing of the file "/home/phylolab/tools/pasta/data/small.fasta" returned 32 sequences of length = 1650
PASTA Ready!

*Reads in the DATA
and sets Type,
prints some stats:*

Example 1: small.fasta

Importing the data caused the GUI to automatically set several settings based on the size, data type, etc...

External Tools
Aligner: MAFFT
Merger: OPAL
Tree Estimator: FASTTREE
Model: [empty]

Sequences and T
Sequence file ...: /home/phy
Multi-Locus ...: [empty]
Data Type: DNA
Initial Alignment: Use for initial tree
Tree file (optional) ...: [empty]

Job Settings
Job Name: test_small
Output Dir.: /home/phy/phylo/phylo/pastajob
CPU(s) Available: 2
Max. Memory (MB): 1024

PASTA Settings
Max. Subproblem: Percentage (50) Size (16)
Decomposition: Centroid
 Time Limit (hr) (24)
 Iteration Limit (3)
Return: Final

Log
PASTA 1.6.4, 2013-2016
Running Log (2016-06-17 03:40:01 CST)
Read 1 file(s) with aligned DNA data. Total of 32 taxa found.
Parsing of the file "/home/phy/phylo/phylo/pasta/data/small.fasta" returned 32 sequences of length = 1650
PASTA Ready!

It noticed that the data type was DNA

It also noticed that this fasta file contains aligned sequences.

Example 1: small.fasta

Step 2: name the job & set the output folder:

Recommended: Use the create folder dialog to create a specific folder for these outputs.

Job Settings

Job Name	test_small
Output Dir.	/home/phylo/phylo/pastatest/pastajob
CPU(s) Available	2
Max. Memory (MB)	1024

PASTA Settings

Max. Subproblem	<input type="radio"/> Percentage	50
	<input checked="" type="radio"/> Size	16
Decomposition	Centroid	
<input type="checkbox"/> Time Limit (hr)		24
<input checked="" type="checkbox"/> Iteration Limit		3
Return	Final	

PASTA 1.6.4, 2013-2016

Running Log (2016-06-17 03:40:01 CST)

Read 1 file(s) with aligned DNA data. Total of 32 taxa found.
Parsing of the file "/home/phylo/phylo/tools/pasta/data/small.fasta" returned 32 sequences of length = 1650

PASTA Ready!

Example 1: small.fasta

Step 3: Say "GO"

The screenshot shows the PASTA software interface with the following settings:

- External Tools:** Tree Estimator: FASTTREE, Model: GTR+G20
- Job Settings:** Job Name: test_small, Output Dir.: /home/phylo/phylo/pastatest/pastajob, CPU(s) Available: 2, Max. Memory (MB): 1024
- PASTA Settings:** Max. Subproblem: Size (16), Decomposition: Centroid, Iterative: checked, Time Limit (hr): 24, Iterations: 3
- Sequences and Tree:** Sequence file: /home/phy, Data Type: DNA, Initial Alignment: Use for initial tree (checked)
- Workflow Settings:** Algorithm: Two-Phase (PASTA), Post-Processing: Extra R...

Red arrows point to the **Start** button, which is highlighted with a red box. The bottom of the window shows a running log with the text: "PASTA 1.6.4, 2013-2016", "Running Log (2016-06-17 03:40:01 CST)", "Read 1 file(s) with aligned DNA data. Total of 32 taxa...", "Parsing of the file "/home/phylo/phylo/pasta/d.../small.fasta" returned 32 sequences of length = 1650", and "PASTA Ready!"

Example 1: Examining the Output Folder

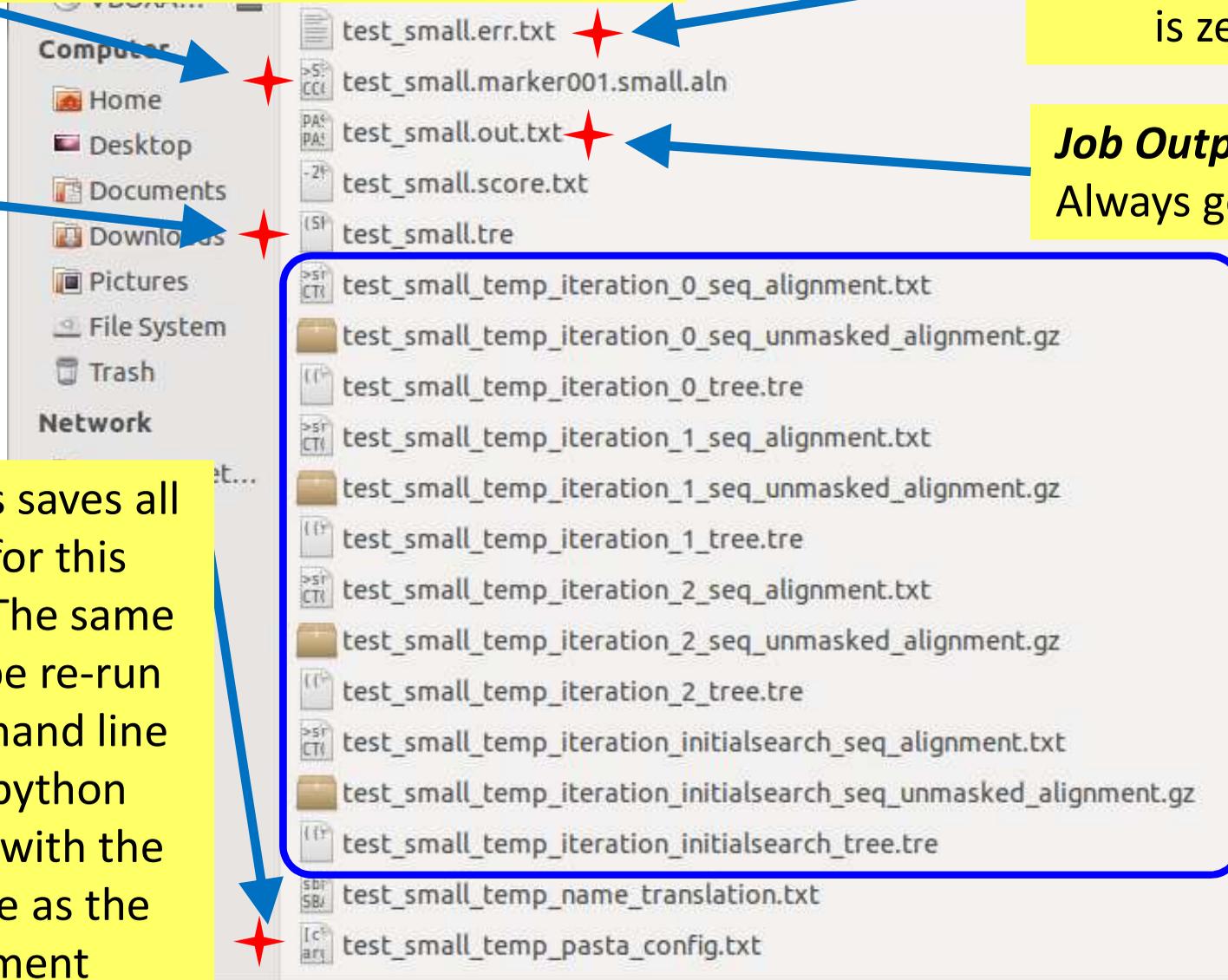
★ = Important File

Final Alignment: always in this name format:
<jobname>.marker001.<original-fasta-name>.aln

Job Output (Errors): contains PASTA console output when errors are reported. If this file is zero bytes, that is a good thing.

Final Tree

Job Output: contains PASTA console output. Always good to examine this file after a run.



Config File: This saves all the settings for this particular job. The same exact job can be re-run from the command line by running “python run_pasta.py” with the path to this file as the ONLY argument

Intermediate alignments and trees after the initial search and after each iteration. Useful mainly for diagnostics and debugging

Example 2: BBA0067 (time permitting)

- (protein data)

Final Tips & Best Practices

- After running an alignment, it is always a good idea to look at the console outputs generated to verify that PASTA did what it was expected to do. If the error file is non-zero size, read that too.
- The PASTA default settings are appropriate and well-chosen for most applications. Unless you have a good reason to use something else, this is a good starting point.
- PASTA scales with the number of cores available, so giving it as many processors as possible is a good idea.
- There are more settings available than what is in the GUI. Check the config file output for any pasta job to see the full list. Also can type “python run_pasta.py -h” from the pasta folder to see a thorough help menu
- Approximate running time benchmarks (length=1500 base pairs):
 - 100 Sequences: <10 minutes on a laptop
 - 1000 Sequences: About 1-3 hours on a 16-core server
 - 10000 Sequences: About 8-15 hours on a 16-core server
 - (Should scale about linearly after this, but will depend on settings...)

Resources

- PASTA User Group:

<https://groups.google.com/forum/#!forum/pasta-users>

- Link to these slides:

<http://publish.illinois.edu/michaelnute/useful-files/>

- Github Repository (which has more documentation, including full install instructions):

<http://github.com/smirarab/pasta>

My Email: nute2@illinois.edu