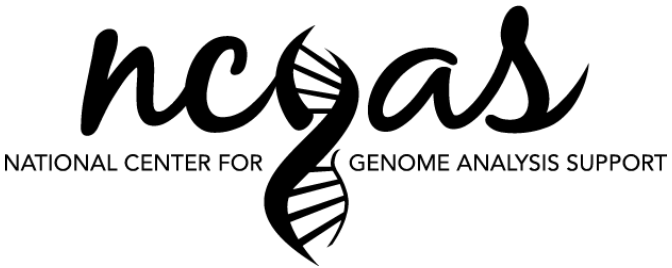
NCGAS makes Robust Assembly Even Easier with Added Features to our Accessible *de novo* Transcriptome Assembly Workflow

Sheri Sanders
ss93@iu.edu



Problem

Many users new to *de novo* assemblies gravitate toward Trinity for it's ease, but…

…Trinity can give large numbers of false positives – which is fine if you have a good idea on how to filter/curate.

Degree of ease should not dictate best practice for a project!

However, picking a workflow can be daunting…

Assembler Biases

Any one assembler is going to be biased in one way or another...

	Pros	Cons
Trinity	easy to use, good isoform handling, great downstream pipelines and annotation	high redundancy, highest memory needs
TransABySS	Low redundancy, has found novel things in our experience, better run time and memory needs	generally not as robust as others
Velvet/Oases	Low noise sensitivity	produces more chimeras
SOAPdenovo	error handling of trinity, graph method of oases, fast, high contiguity, low redundancy	sensitive to coverage

Assembler

CDTA – Combined *de novo* Transcriptome Assembly

It is unlikely that different assembly algorithms will experience the same biases/errors in assembly.

- Use multiple assemblers with multiple parameters (kmers)
- Get as much data as possible and look for concordance between the different assemblers.

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

See paper on CDTA – He et al. 2016

Why we started doing this…

In several projects (particularly in large or polyploid systems), we were not recovering genes we knew were expressed – we had qPCR to back them up! No one assembler got all the target genes – the CDTA did!

We've seen quality increases in the transcriptome when we run this pipeline.

It has been published in best practices for RNA-seq to use multiple parameters at least.

It's easier to defend in publication!

Workflow Overview

Assemble: Submit all the assemblies in parallel

- Assemblies run for k=25 (trinity) and k={35,45,55,65,75,85} for Velvet/Oases, SOAPdenovo, TransABySS.
- Running in parallel decreases time to run all 19 assemblies!
- Not entirely pushbutton – you are should be LOOKING AT WHAT YOU ARE RUNNING.

Combine: Submit all the combiner scripts in parallel (there are three)

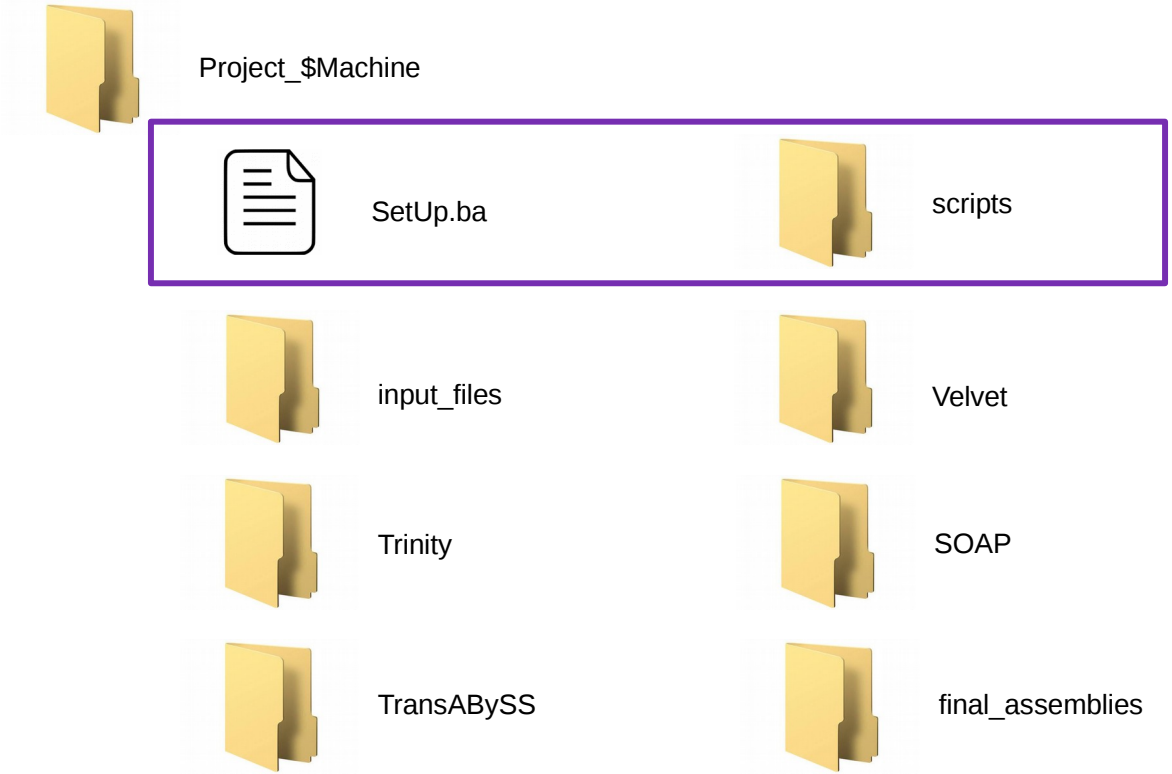
- These combine all the kmers from SOAP, Velvet, and TransABySS
- Also label each contig with kmer and assembler for evaluation if interested
- Moves all the final groups to the final_assemblies folder

Corroborate: Run EviGene in final_assemblies folder

- Cleans up the over-assembly

Pending queue load and size of data… this can take ~2 days - ~weeks

Running the Workflow



0. Run set up script

Running the Workflow: Set Up Script - NEW

Initial set up is now stream-lined with a set up script.

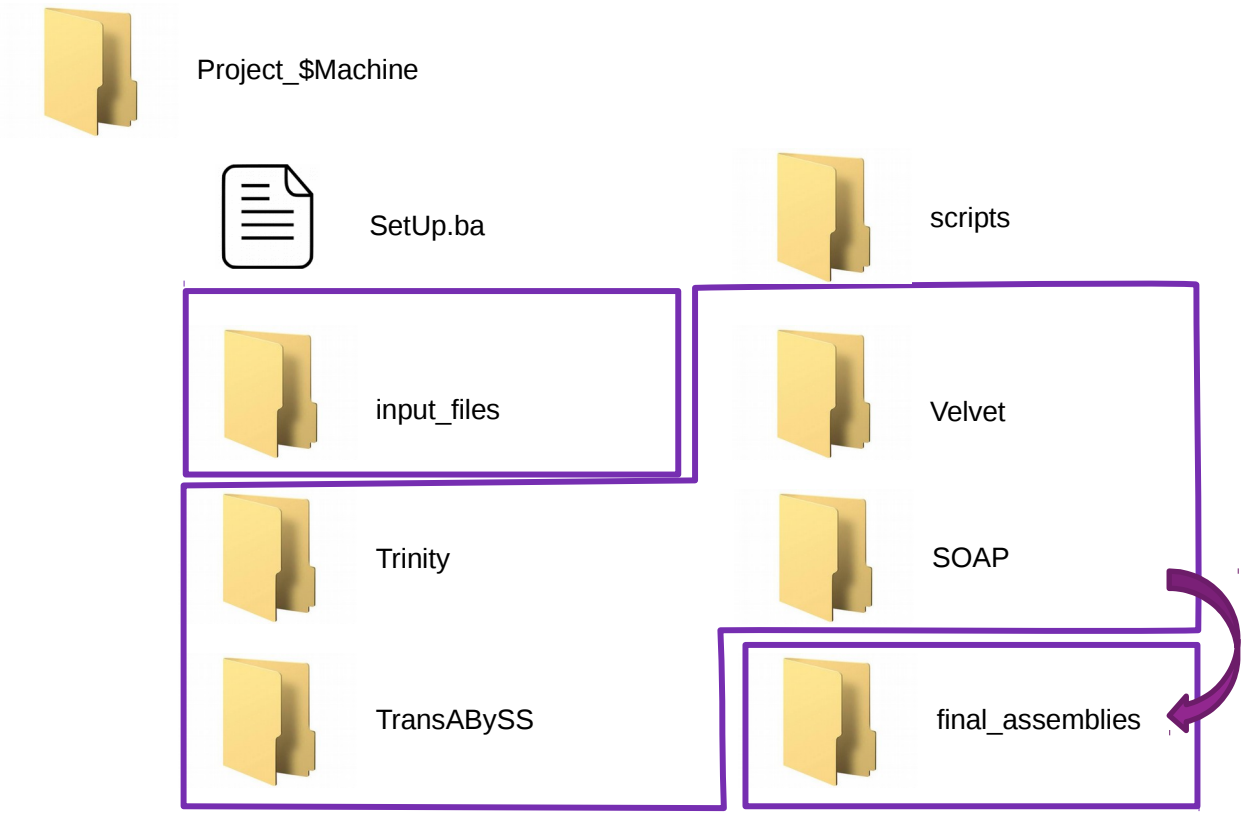
This script does the following:

1) Sets email for job files
2) Sets working directory for all job files (automatically)
3) Sets pipeline to double or single strand (default: double) – ALSO NEW
4) Automatically adjusts insert length (double stranded only) and read length – ALSO NEW
   -Default 400bp insert, 150bp read length

./Setup.sh –e email@university.edu –s double –i 400 –r 150
./Setup.sh –e email@university.edu –s single –r 150

This replaces manual commands to change these scripts that were confusing for many users.

Running the Workflow



0. Run set up script
1. Place input files and normalize
2. Run all assemblies
   - READMEs in all
   - Combine script in each to merge/tag kmer assemblies
3. Corroborate

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Running the Workflow: Cleaning it up with EviGene

Evidential Gene

- Removes perfect redundancy (fastanrdb)
- Removes perfect fragments (CD-HIT)
- Clusters similar nucleotide sequences (CD-HIT-EST)
- Uses BLASTn to find 98% identity, exon sized aligments (BLAST)

Uses this information to identify highest quality cds regions and classifies transcripts as main (okay), alternative (okalt), and dropped (drop) sets.

See eugenes website or our blog on how it works for more details!
NOTE: This software only looks for quality metrics “internally” – no external data is used

Benefits

Pretty much filters for you – usually I end up with the expected 20-30k transcripts in the main set.

You get a separate fasta with all the alternatives (tagged with which main transcript they are associated with), as well as a table of main and alternative forms.

Automatically gives you cds, aa, and fa formats

Replicability is high for a filtering paradigm

You start with working scripts that you can easily change, with documentation.

Quality Control - NEW

Automated script will perform the following quality control metrics for each of the 19 assemblies, as well as the post-EviGene clean up:

Quality (via quast)

- N_{50}
- Gene/isoform ratio – what are your expectations?
- Length metrics - > 1000bp, >5000 bp, >25,000bps
- GC – does this match expected?

Completeness

- BUSCO

There are other options to consider for QC:

- qPCR
- BLAST to a similar organism*

Annotation - NEW

We've recently added an annotation workflow as well – once you complete EviGene, you can run the RunAnnotation.sh and RunTrinotate.sh scripts.

Output:

- Completed Trinotate database of your loci
  - this includes BLAST to UniProt/Swiss-Prot, KEGG and GO terms.
  - option to include custom BLAST for closely related organism
- Final naming and clean up of fasta files for amino acid and nucleotides.
- A table listing all the loci and the purposed isoforms

Takes ~2 days for most assemblies. Resulting table can be subset and merged into DE tables as well for easier interpretation.

Differential Expression - NEW

We've added documentation and an example on how to run Differential Expression on data coming out of the pipeline. This is not fully automated, but helpful scripts are included, such as:

README – walks through steps but is also a script to run an example

subset_fasta.pl – subsets fasta given list of desired and original fasta

We also offer resources on choosing a DE program!

User Projects

Environmental toxicology/disease

- Birds as indicators of contaminant exposure in Great Lakes, using Tree Swallows*
- Chestnut Tree response to bleeding canker in Europe

Aquatic studies

- Skin and blood microbiomes in sharks to evaluate health in non-invasive ways**
- Killer whale transcriptome analyses to determine mechanisms of toxic injury to reproductive health**
- Soft octocoral, Corky sea finger, from southwest continental shelf of Puerto Rico

Polyploid systems

- Allotetraploid catostomid fishes to study genome duplication
- Allopolyploid salamanders to study genome duplication in disease response
- Identifying stilbenoid pathways in allotetraploid peanut***

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

*poster 113 **microbiome methodology test ***published

User Reception

Great for us – NCGAS does a LOT of transcriptome assemblies

- Sweet potato, coffee, peanut, fly, frog, daphnia, salamander…

Working well for our users

- “The use of this pipeline has **saved me tons of time** from having to figure out the script for each assembly program and it is VERY easy to use, especially for a **person like myself who barely understands Linux!”**

- “If this pipeline was not available, I would have **most likely used only one package and at one kmer size** for my assembly, and it would have have **probably taken me just as long** to figure out and run.”

How do you get all this?

Github – Torque and SLURM versions available at github.com/ncgas

Go to our website – ncgas.org, and click training

Contact us (help@ncgas.org)!

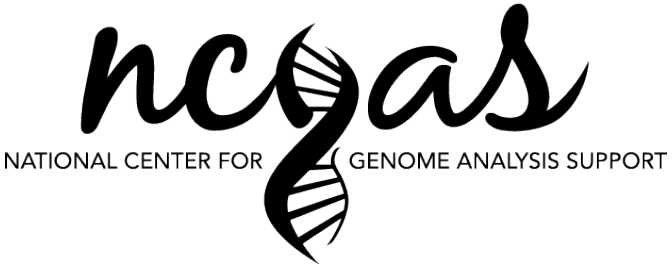
de novo Transcriptome Assembly Workshop

NCGAS runs a two and a half day workshop on this pipeline! Next one is April 29-May 1, 2019.

Topics:

- HPC resources freely available
- Managing and moving data
- Using this pipeline
- Downstream analyses with hands-on demos
- Manipulating jobs to be more efficient

See our workshop listings here for up-to-date information:
www.ncgas.org/Workshops.php



/ncgasiu @ncgas ncgas.org