SevenBridges

www.sbgenomics.com

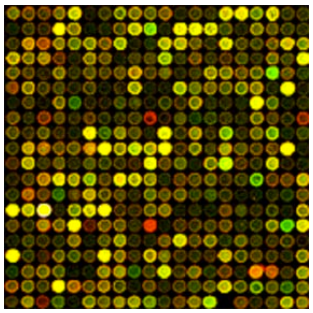
De Novo Transcriptome Assembly and Analysis Pipeline

Lu Zhang, PhD
Director of Research and Development
Seven Bridges Genomics
2014.04.17

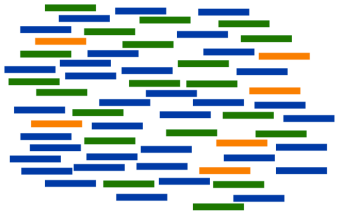
Agenda

Background And Motivation	5 min
Trinity Pipeline	7 min
Example Analysis	8 min

RNA-Seq Is A New Powerful Technology For Studying Transcriptomes



Microarrays (1995)

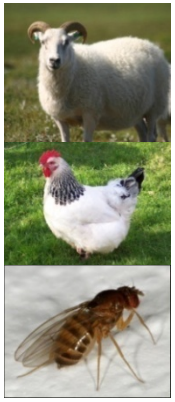


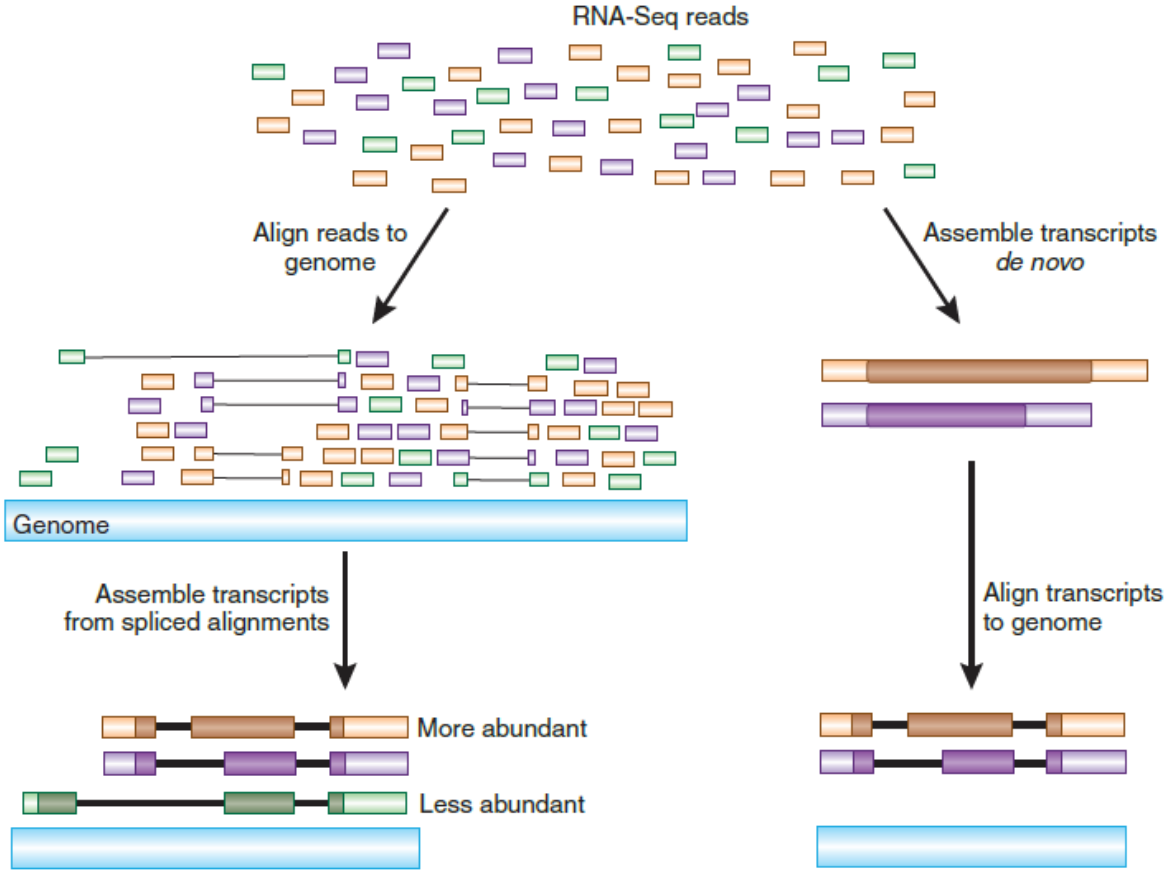
RNA-Seq (2008)

RNA-Seq Advantages Over Microarray:

- Not limited to existing genomic sequence
- Able to detect new genes
- Detects structural variants; alter. splicing, gene fusion
- Low background signal
- Large dynamic range of expression level
- Single-base resolution
- Highly accurate
- High level of reproducibility
- Price becoming comparable to microarrays

Strategies For Reconstructing Transcripts Using RNA-Seq

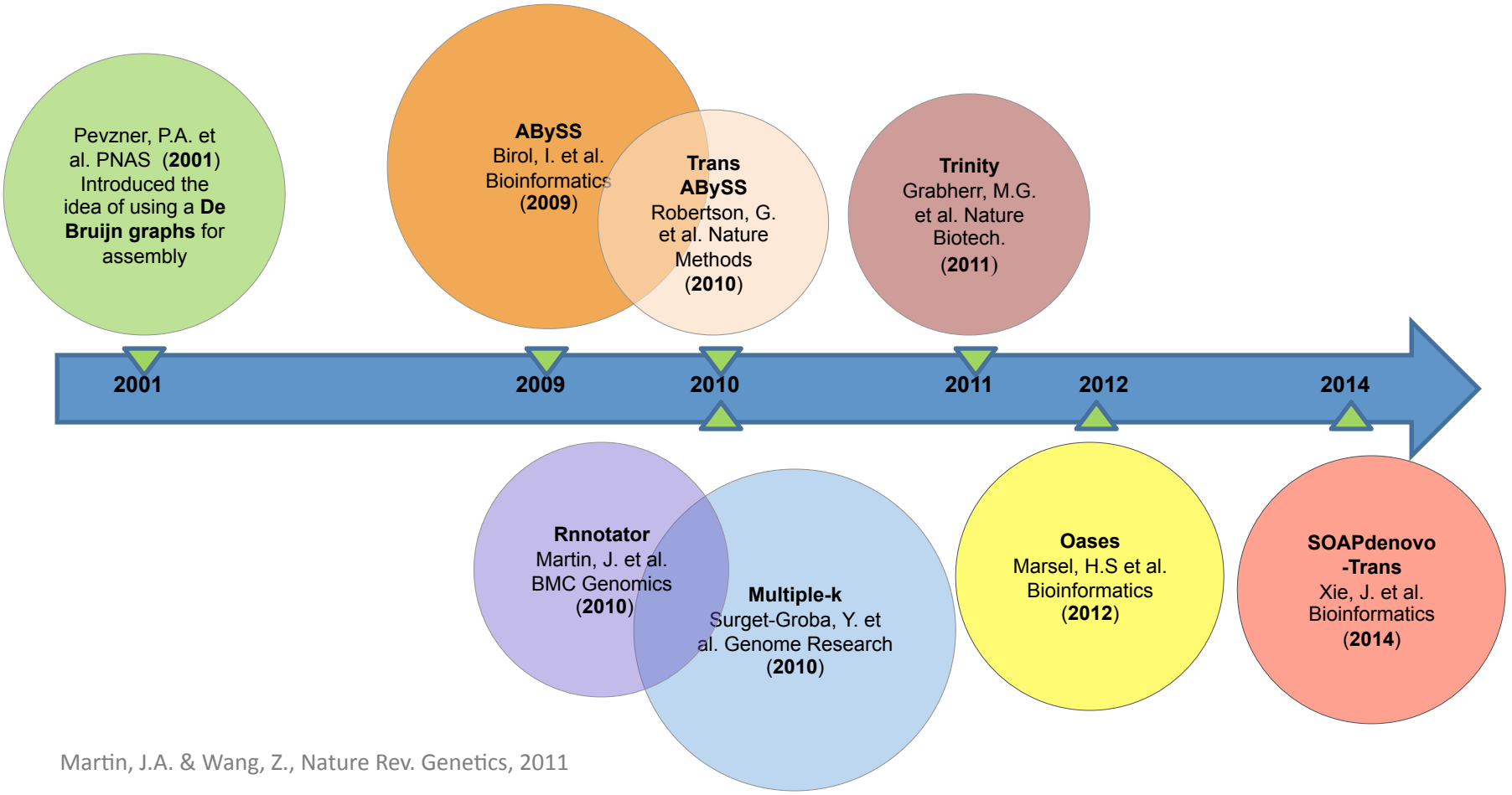




Haas, B.J & Zody, M.C.
Nature Biotech. 2010

Current Methods In RNA-Seq Assembly

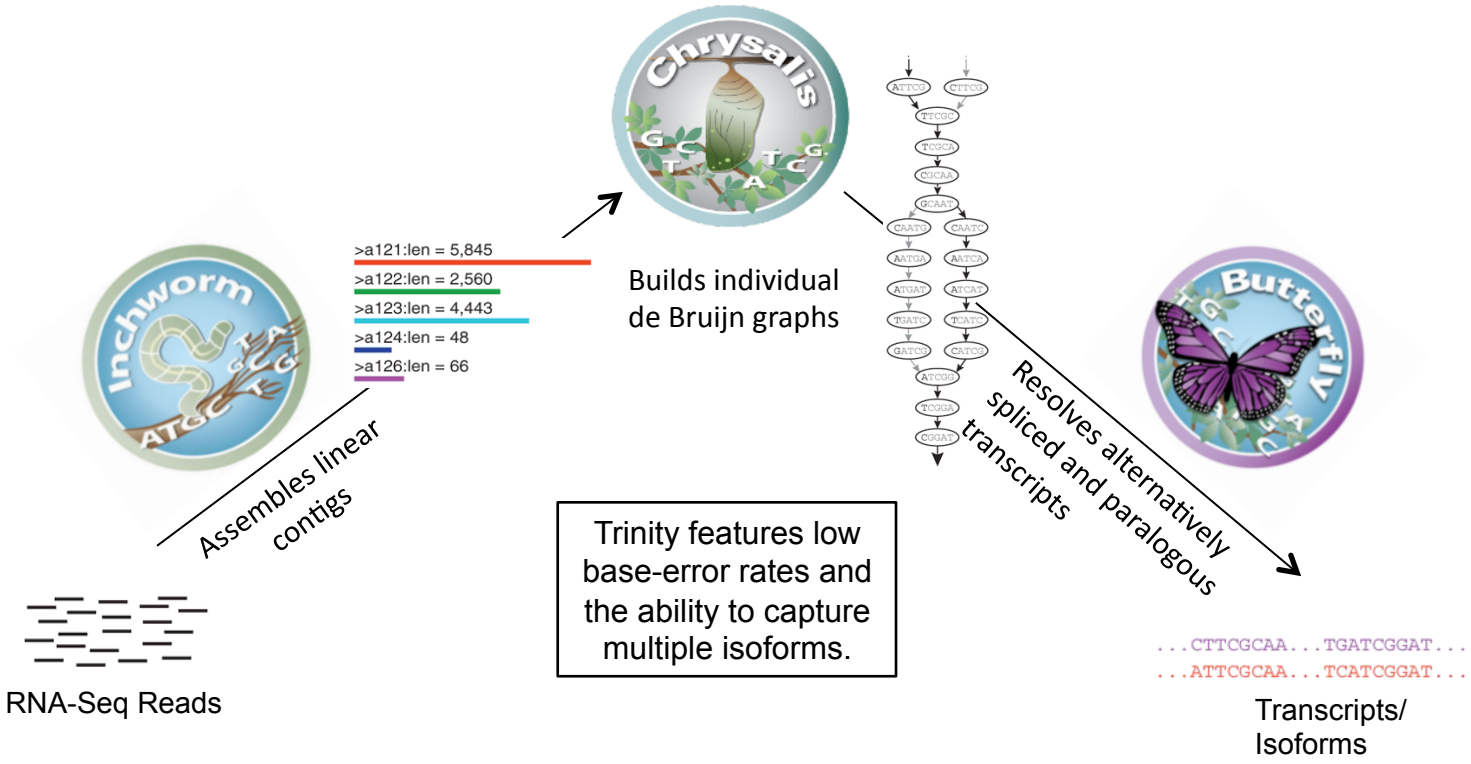
Development Timeline



Martin, J.A. & Wang, Z., Nature Rev. Genetics, 2011

Trinity Algorithm
Step-wise Assemble Transcripts *De Novo*

- Developed by Broad Institute and Hebrew University of Jerusalem



Trinity features low base-error rates and the ability to capture multiple isoforms.

Grabherr, M.G., Nature Biotech., 2011

Beyond assembly

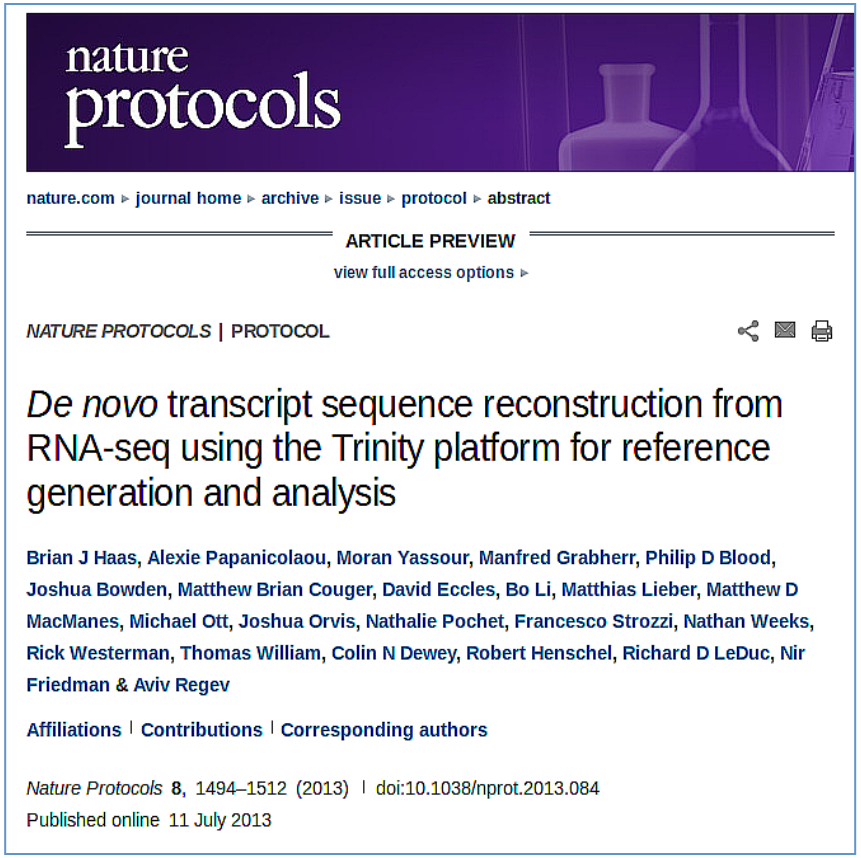
What's 

?

Trinity Pipeline Protocol
From Raw Data To Insights

Brian Haas, Broad Institute

What happened?
finding biological significance



nature protocols

nature.com ▸ journal home ▸ archive ▸ issue ▸ protocol ▸ abstract

ARTICLE PREVIEW

view full access options ▸

NATURE PROTOCOLS | PROTOCOL

De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

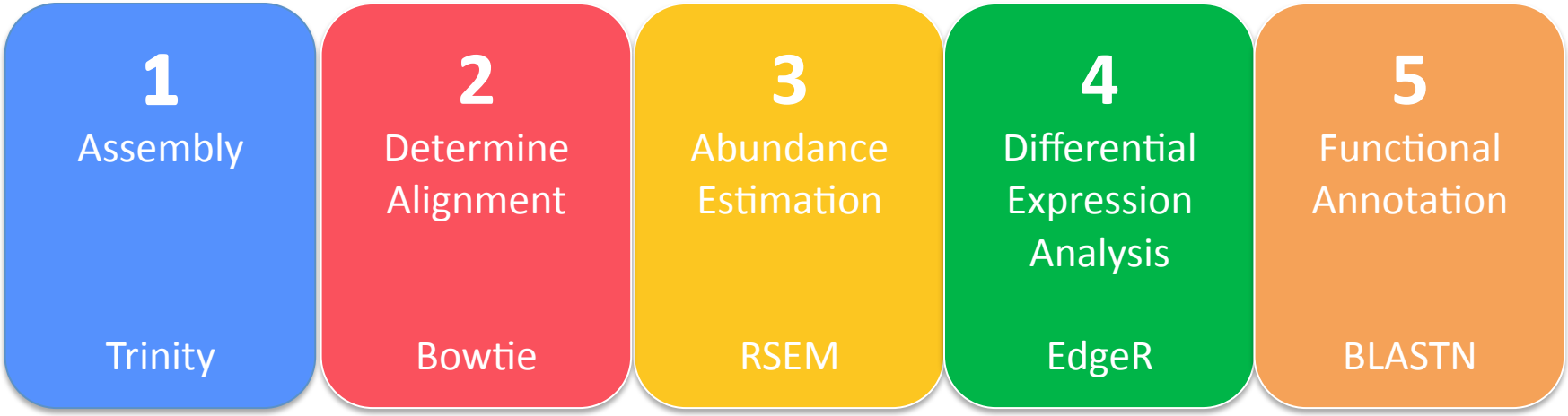
Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

Affiliations | Contributions | Corresponding authors

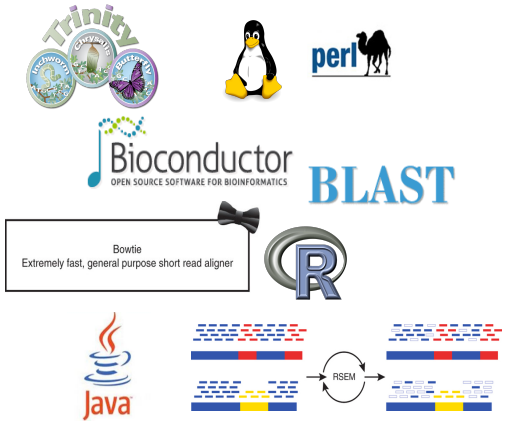
Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084
Published online 11 July 2013

Case study: 20+ processing steps to analyze *S. pombe* data

Trinity Discovery Workflow Main Components



Prepare For Analysis – Traditional Approach



Big Data

Plants and large size eukaryotes need millions to billions of RNA-Seq reads to assemble a transcriptome.

Processing Framework Commands

Powerful Computers

~1G of RAM per 1M of pairs of Illumina reads.
At least 8 cores per sample.
Fast storage system.

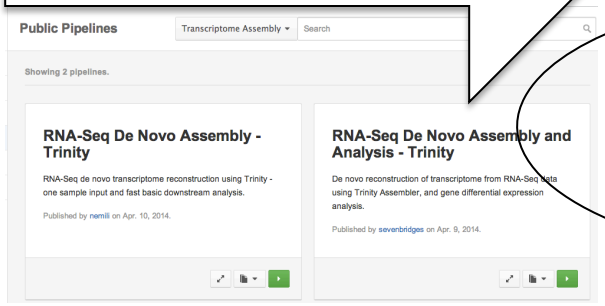
Typical challenges of large data problems:
How to handle distributed synchronization?
How to optimize memory and CPU usage?
How to handle intermediate results?
Where to store large files?
What happened when tasks fail?

Prepare For Analysis – Seven Bridges Approach

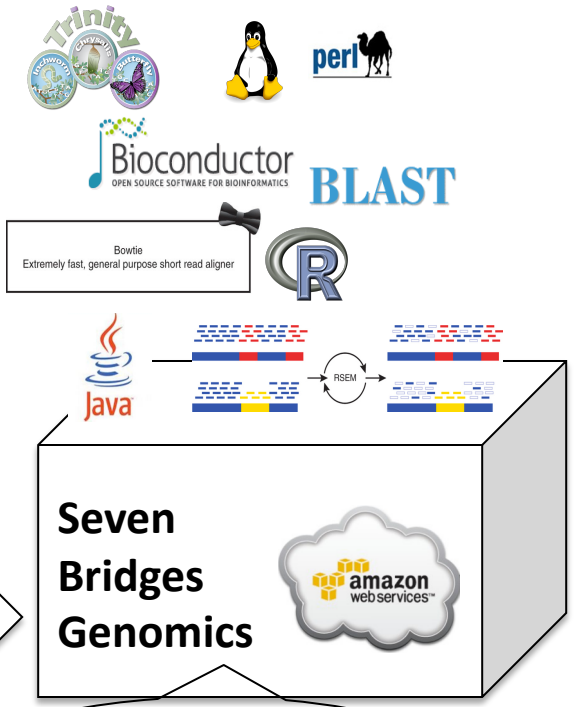
Big Data

Plants and large size eukaryotes need millions to billions of RNA-Seq reads to assemble a transcriptome.

Processing Framework
Web Interface



Automated workflow
Leverage advantages of cloud
Parallelism
Scalability



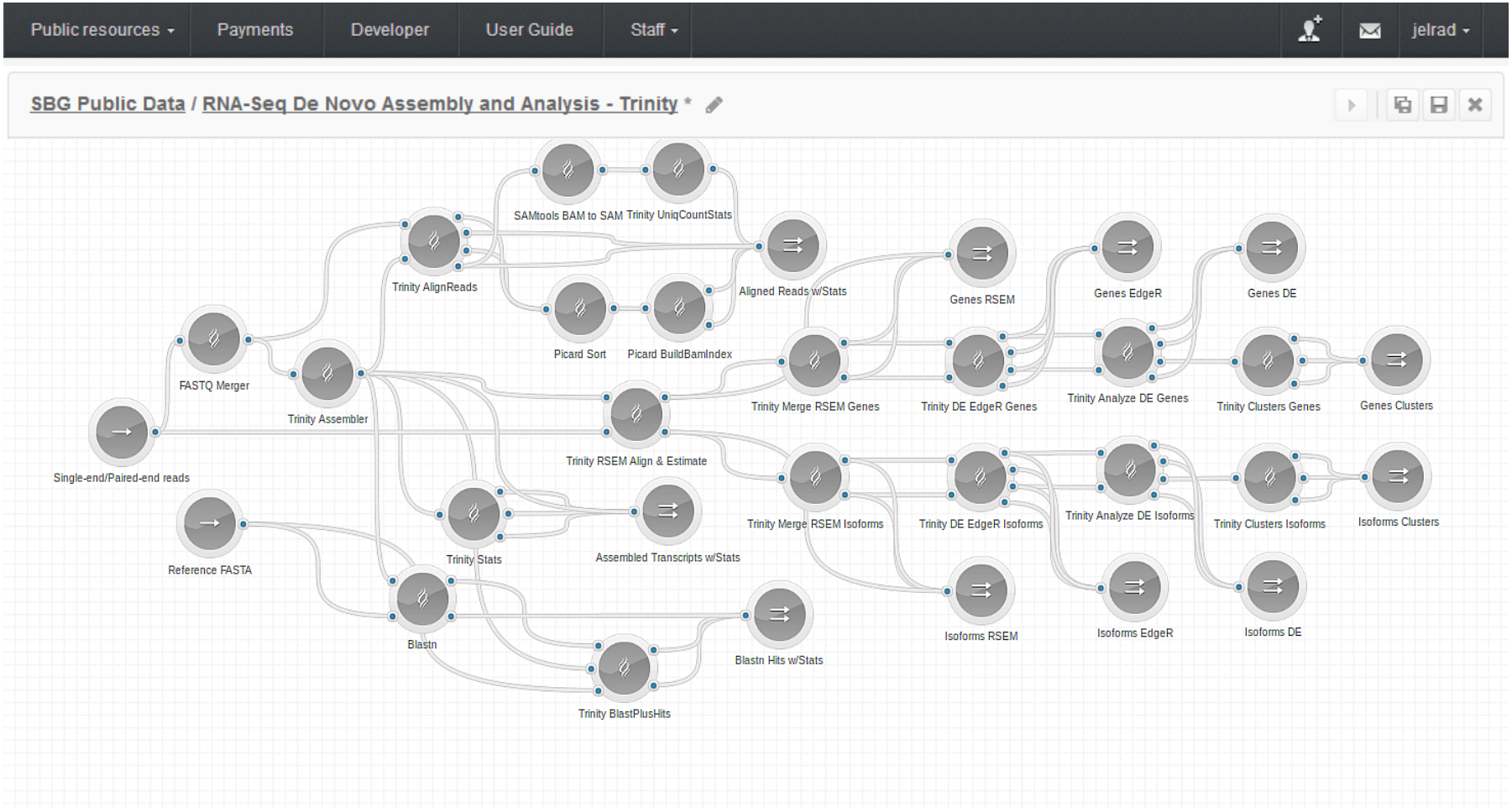
Trinity assembler - run on Amazon ec2 cluster instance, with 256GB RAM, 32 CPUs.

Overview Of Trinity Pipeline

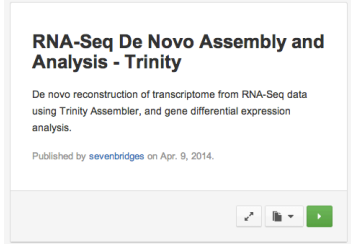
RNA-Seq De Novo Assembly and Analysis - Trinity

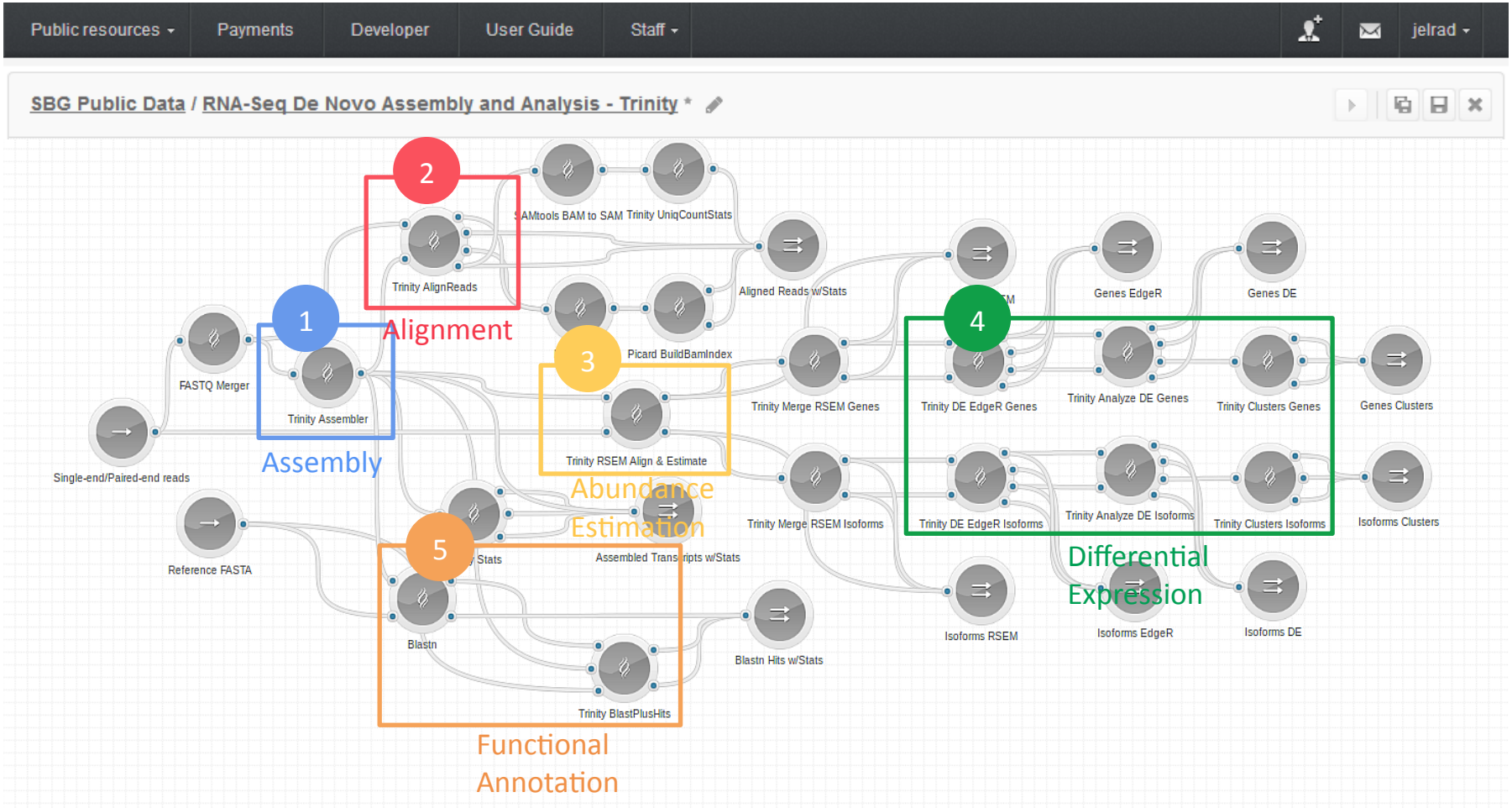
De novo reconstruction of transcriptome from RNA-Seq data using Trinity Assembler, and gene differential expression analysis.

Published by sevenbridges on Apr. 9, 2014.



Overview Of Trinity Pipeline





How to use this pipeline?

Example Analysis
Objective: Reproduce Haas et al. (2013) Experiment

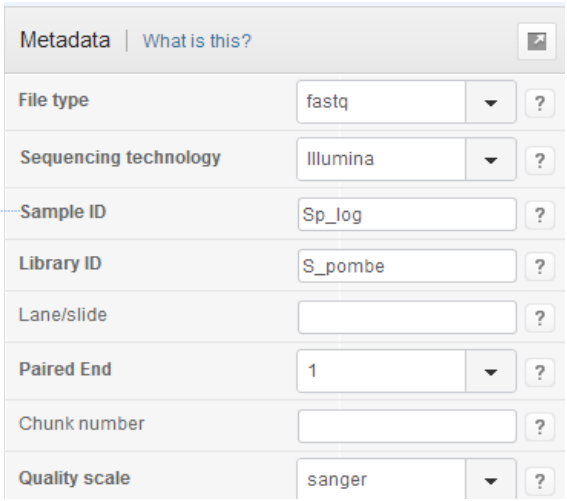
Why is reproducibility validation important?

Example Analysis
Objective: Reproduce Haas et al. (2013) Experiment

Input Data:

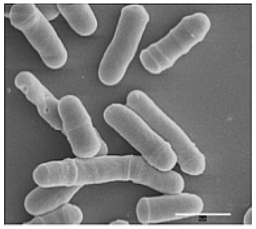
RNA-Seq reads of *S. pombe* from 4 conditions:

- Logarithmic growth
  - Sp.log.1M.left.fq, Sp.log.1M.right.fq
- Plateau phase
  - Sp.plat.1M.left.fq, Sp.plat.1M.right.fq
- Diauxic shift
  - Sp.ds.1M.left.fq, Sp.ds.1M.right.fq
- Heat shock
  - Sp.hs.1M.left.fq, Sp.hs.1M.right.fq



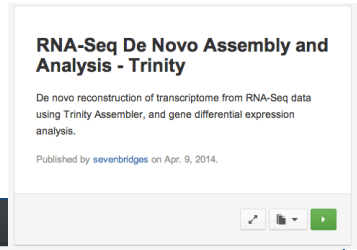
S_pombe_refTrans.fasta

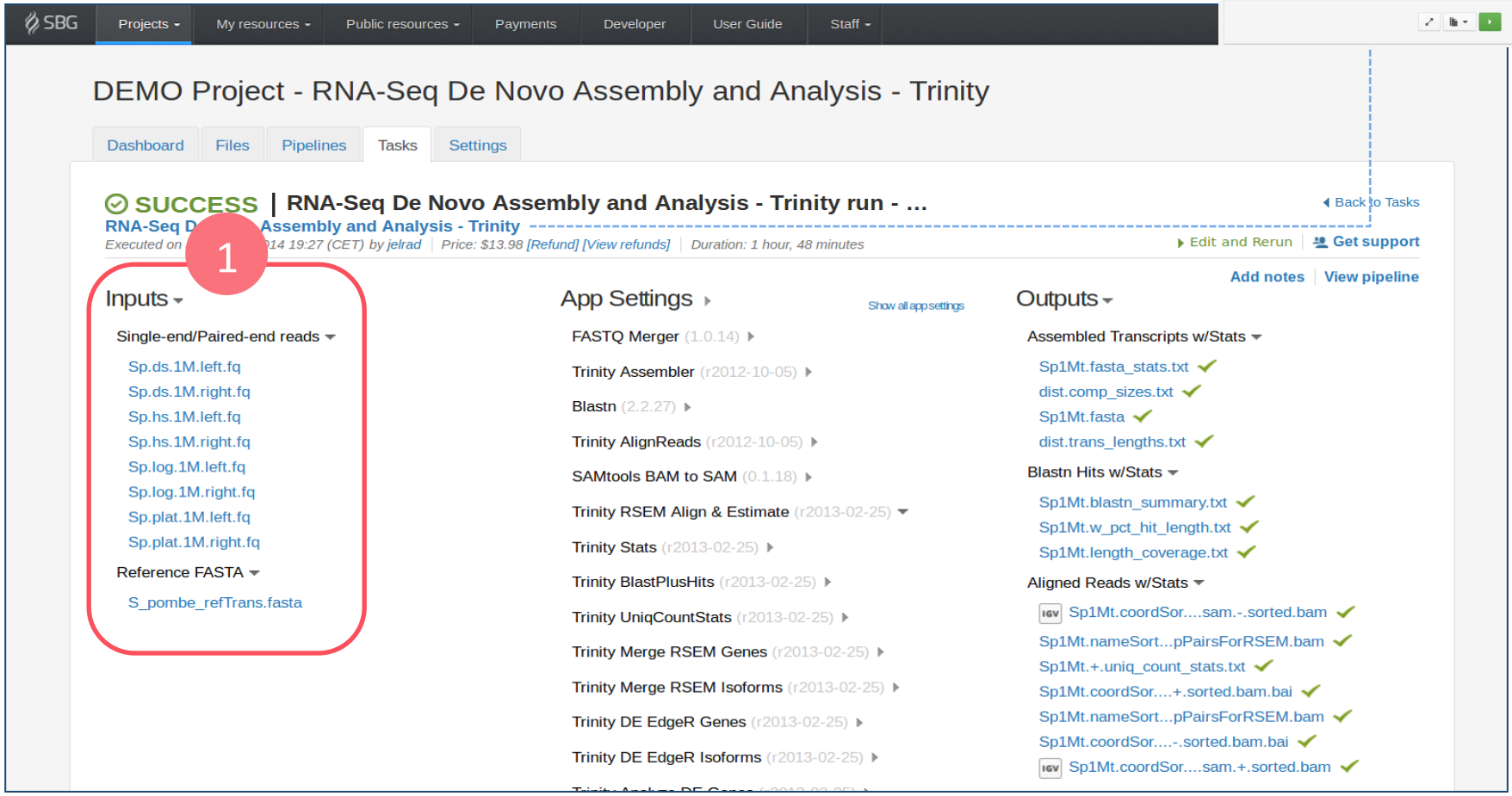
Or transcript sequences from closely related species, or adjust BLASTN param to use NCBI's NT database.



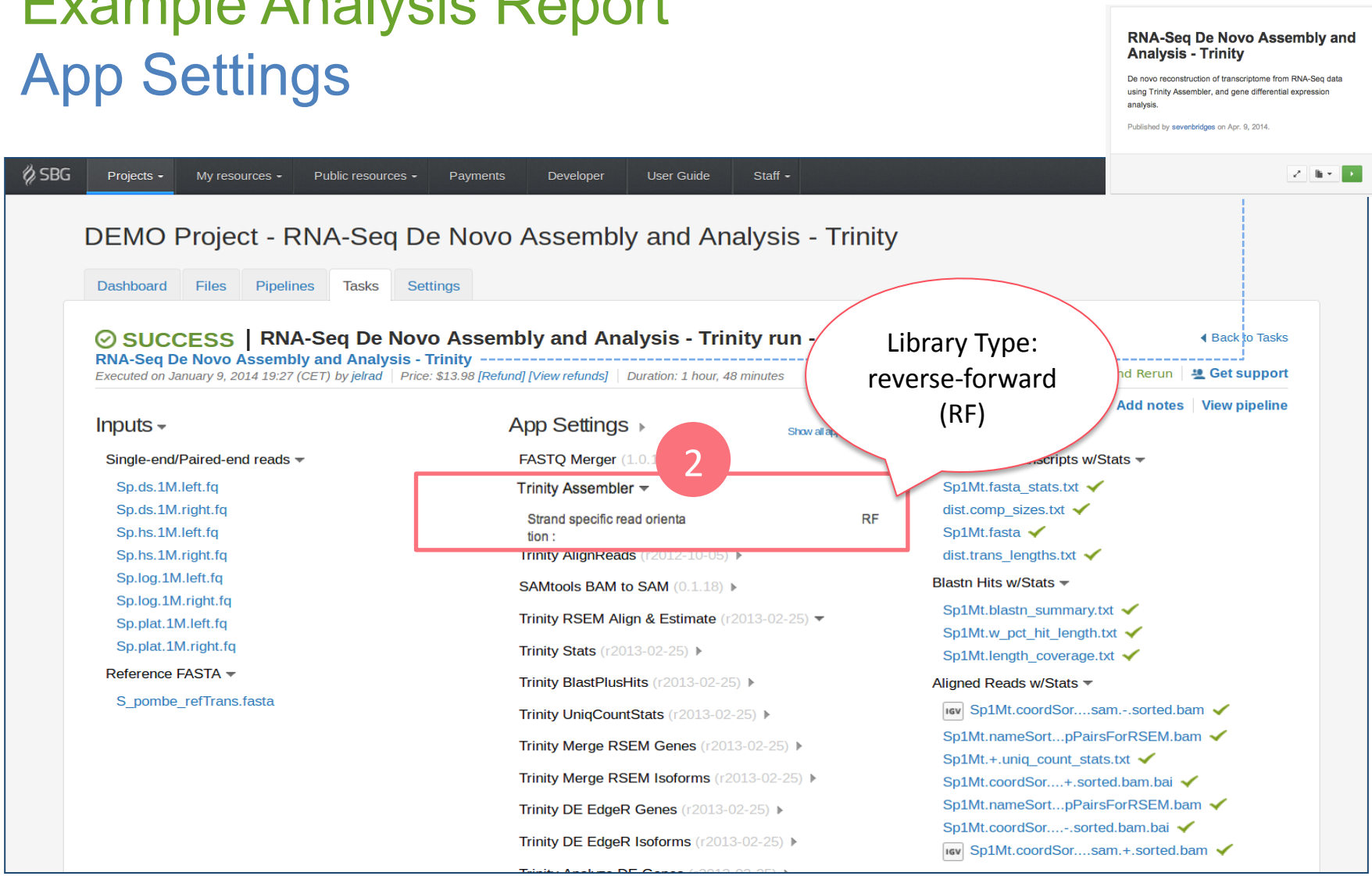
Fission yeast
(*Schizosaccharomyces pombe*)

Example Analysis Report
Input Data

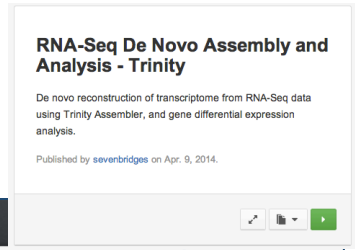




Example Analysis Report
App Settings



Example Analysis Report
Output Results



SBG | Projects | My resources | Public resources | Payments | Developer | User Guide | Staff

DEMO Project - RNA-Seq De Novo Assembly and Analysis - Trinity

Dashboard | Files | Pipelines | Tasks | Settings

SUCCESS | RNA-Seq De Novo Assembly and Analysis - Trinity run - ...

Back to Tasks

RNA-Seq De Novo Assembly and Analysis - Trinity

Executed on January 9, 2014 19:27 (CET) by jelrad | Price: $13.98 [Refund] [View refunds] | Duration: 1 hour, 48 minutes

...and Rerun | Get support | Add notes | View pipeline

3

Click to view results on website or download

Inputs

Single-end/Paired-end reads
- Sp.ds.1M.left.fq
- Sp.ds.1M.right.fq
- Sp.hs.1M.left.fq
- Sp.hs.1M.right.fq
- Sp.log.1M.left.fq
- Sp.log.1M.right.fq
- Sp.plat.1M.left.fq
- Sp.plat.1M.right.fq

Reference FASTA
- S_pombe_refTrans.fasta

Trinity Stats (r2013-02-25)

Trinity BlastPlusHits (r2013-02-25)

Trinity UniqCountStats (r2013-02-25)

Trinity Merge RSEM Genes (r2013-02-25)

Trinity Merge RSEM Isoforms (r2013-02-25)

Trinity DE EdgeR Genes (r2013-02-25)

Trinity DE EdgeR Isoforms (r2013-02-25)

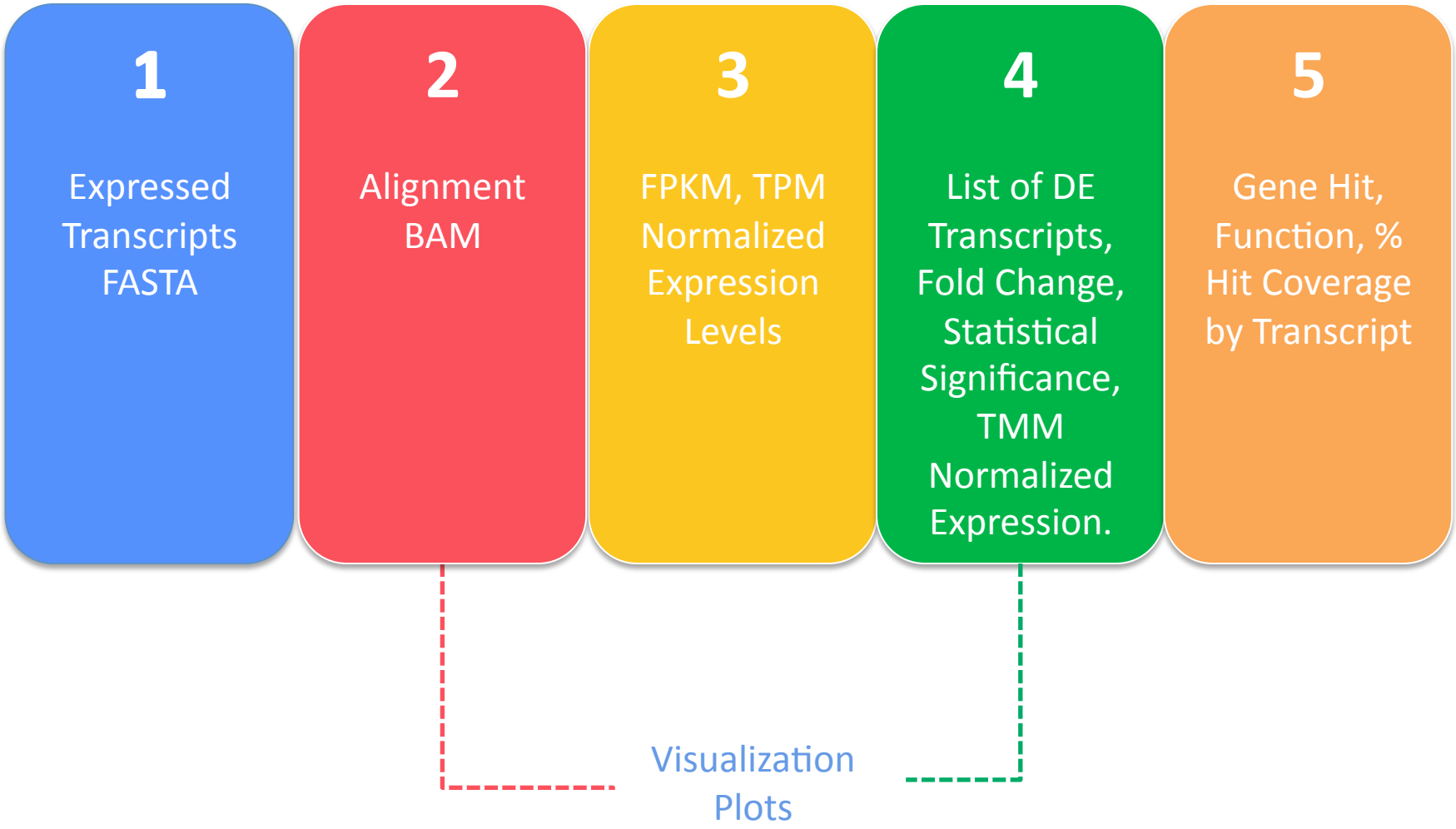
Outputs

Assembled Transcripts w/Stats
- Sp1Mt.fasta_stats.txt
- dist.comp_sizes.txt
- Sp1Mt.fasta
- dist.trans_lengths.txt

Blastn Hits w/Stats
- Sp1Mt.blastn_summary.txt
- Sp1Mt.w_pct_hit_length.txt
- Sp1Mt.length_coverage.txt

Aligned Reads w/Stats
- Sp1Mt.coordSor....sam.-.sorted.bam
- Sp1Mt.nameSort...pPairsForRSEM.bam
- Sp1Mt.+.uniq_count_stats.txt
- Sp1Mt.coordSor....+.sorted.bam.bai
- Sp1Mt.nameSort...pPairsForRSEM.bam
- Sp1Mt.coordSor....-.sorted.bam.bai
- Sp1Mt.coordSor....sam.+.sorted.bam

Example Analysis
Important Output Overview

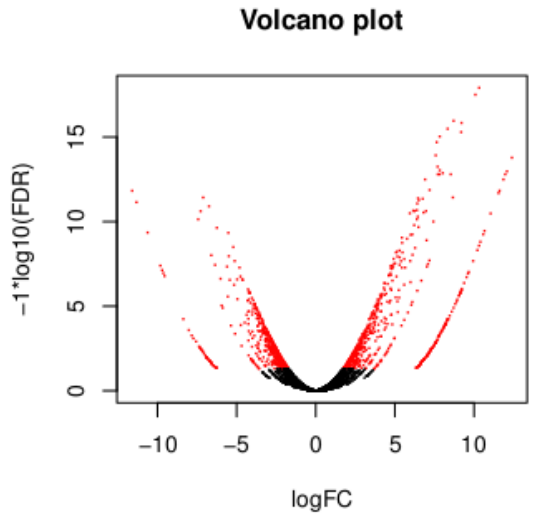


Visualization
Differential Expression Profiles

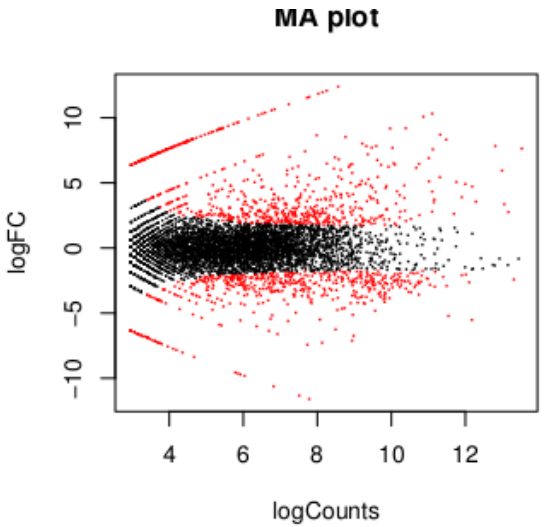
Useful for:

Quickly find which transcripts were differentially expressed.

Detect outliers.



statistical significance by fold change



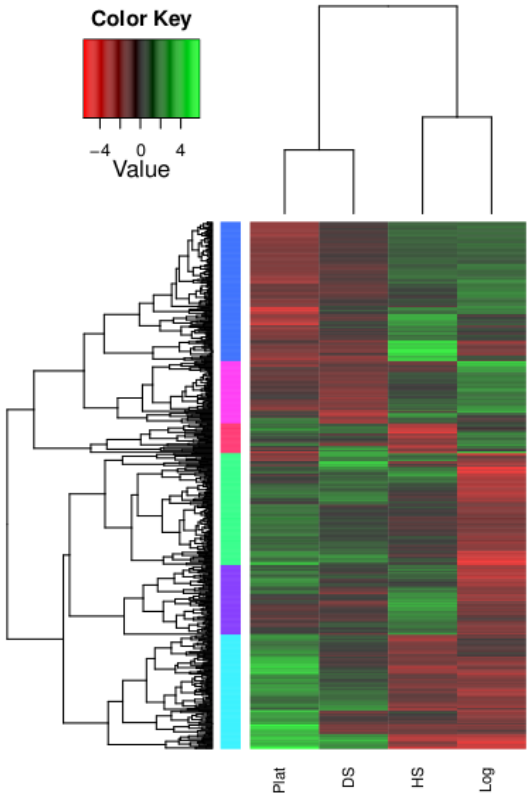
signal intensity over average intensity

Visualization
Clustering Plots

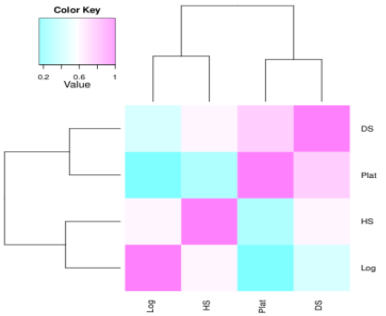
Useful for:

Understand correlation between samples or transcripts.

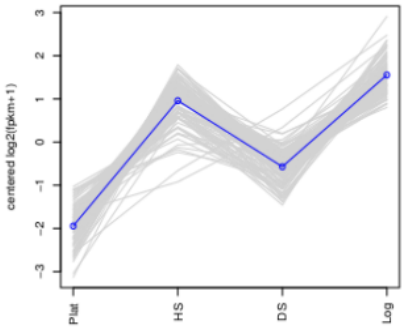
Hierarchical clustering heatmap



An alternative way to see how samples cluster.



An example sub-cluster of transcripts.

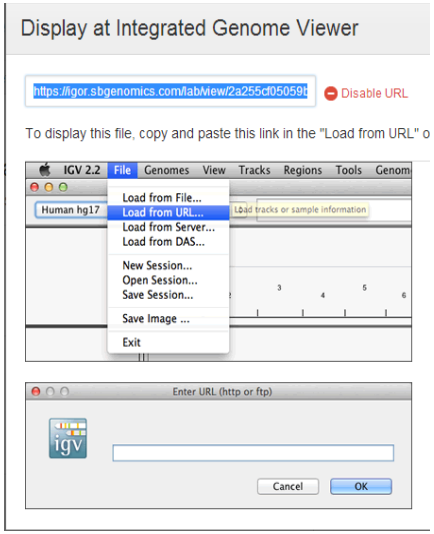


Visualization
Stream Alignment and View In IGV

Useful for:

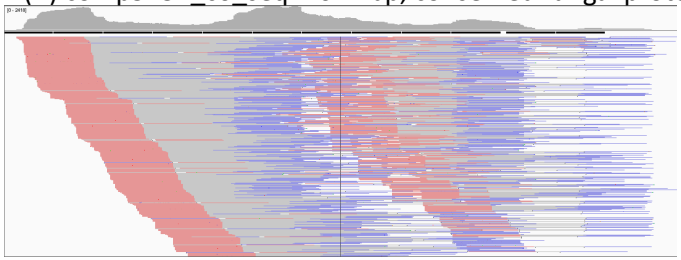
Ensure quality of assembled transcript:
e.g. sufficient read coverage along full sequences of the transcripts

Confirm library type:
e.g. first read in pair (purple) mapped to the reverse of the sense strand, which matches the input data library type: RF

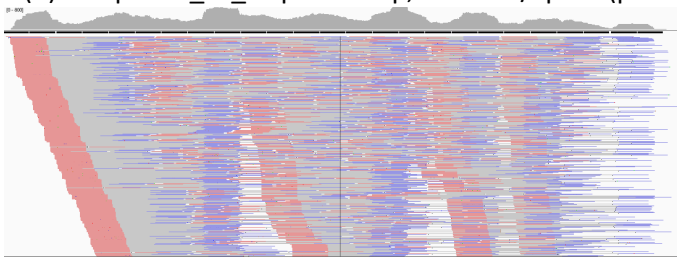


Example top DE transcripts between logarithmic versus plateau conditions.

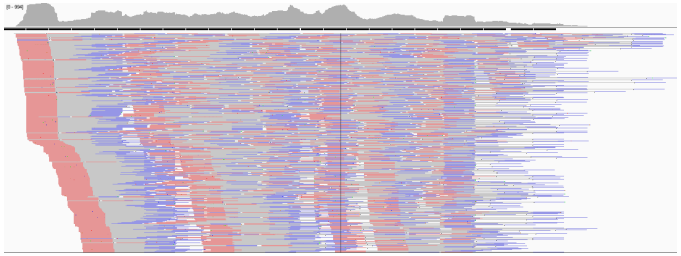
(A) comp5132_c0_seq1: 677 bp, conserved fungal protein



(B) comp5235_c0_seq1: 1279bp, esterase/lipase (predicted)



(C) comp2035_c0_seq1: 1474bp, 3-hydroxyacyl-CoA dehydrogenase (predicted)



Key Findings
Good Reproducibility Between Two Experiments

SBG Experiment

Total trinity transcripts	9349
Total trinity components	8698
Contig N50	1584
# of diff. expressed transcripts	668
Gene hits	4766
'Full length' transcripts	3401

Top 5 DE transcripts between logarithmic versus plateau conditions.

Transcript	logFC	logCPM	p-value	FDR
comp5132_c0_seq1	10.3	11.1	2.13e-22	1.22e-18
comp5235_c0_seq1	10.1	10.9	1.09e-21	3.13e-18
comp5101_c0_seq1	8.7	11.3	5.73e-20	1.10e-16
comp2035_c0_seq1	9.2	10.4	1.01e-19	1.46e-16
comp1972_c0_seq1	8.3	11.5	2.81e-19	3.23e-16

Haas et al. (2013)

Total trinity transcripts	9299
Total trinity components	8649
Contig N50	1585
# of diff. expressed transcripts	659
Gene hits	4765
'Full length' transcripts	3401

Top 5 DE transcripts between logarithmic versus plateau conditions.

Transcript	logFC	logCPM	*P* value	FDR
comp5128_c0_seq1	10.3	11.1	2.13e-22	1.22e-18
comp5231_c0_seq1	10.0	10.9	1.10e-21	3.13e-18
comp5097_c0_seq1	8.7	11.3	5.72e-20	1.10e-16
comp1686_c0_seq1	9.2	10.4	1.01e-19	1.46e-16
comp1012_c0_seq1	8.3	11.5	2.8e-19	3.23e-16

Note: Transcript identifiers are randomly assigned in each run, thus do not match.

Conclusion

- Trinity pipeline is a powerful tool for improving genome annotation, especially for non-model organisms.
- Good reproducibility between two computational experiments using the *S. pombe* data and Trinity pipeline.

Thank You For Attending Our Webinar!

A recording of the webinar, slides, and the whitepaper will be available on our website tomorrow Apr. 18th, 2014 at:

www.sbgenomics.com

Meet us in person!

- Bio-IT World 2014 Apr. 29 - May 1 Showcase (Booth #448) and Workshop

Get in touch!

team@sbgenomics.com

@SBGenomics

blog.sbgenomics.com

facebook.com/sbgenomics