Example Science Applications

Far too many for the time we have in this session

Download Slides and Follow Along

mcbios.readthedocs.org

RNA-Seq with Tuxedo Pipeline

Overview

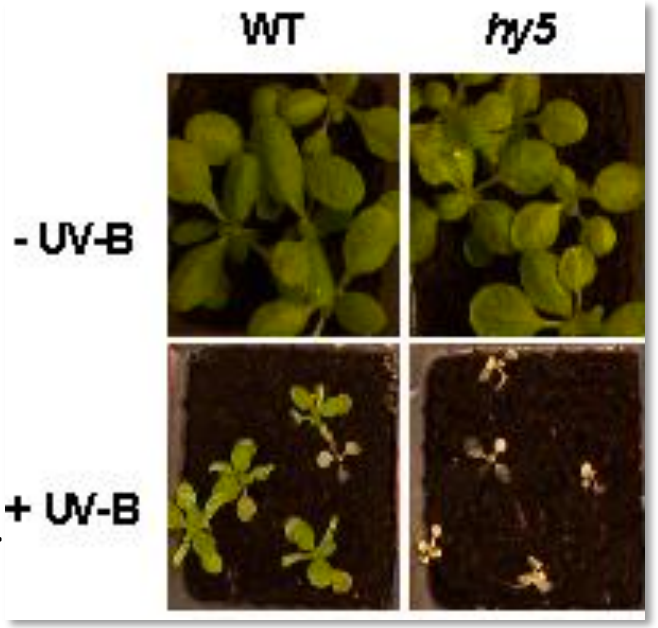
Determine differential ~~expression~~ abundance of transcripts in between a WT and mutant organism

See Full tutorial in CyVerse Learning Center

Experiment Overview

Example experiment

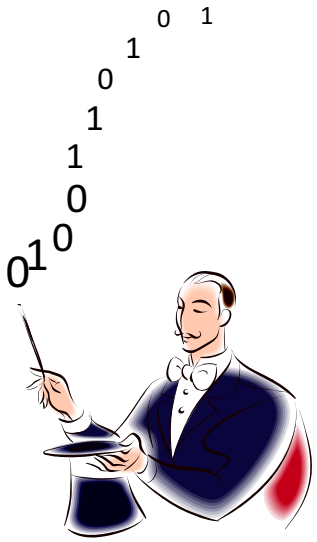
- LONG HYPOCOTYL 5 (HY5) is a basic leucine zipper transcription factor (TF).
- Mutations cause aberrant phenotypes in Arabidopsis morphology, pigmentation and hormonal response.
- We will use RNA-Seq to compare WT and *hy5* to identify HY5-regulated genes.



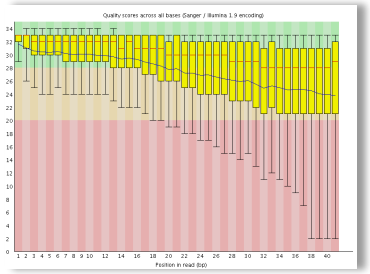
Source: http://www.gla.ac.uk/media/media_73736_en.jpg

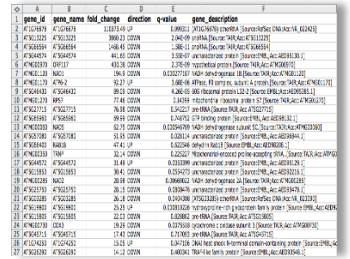
Now what?

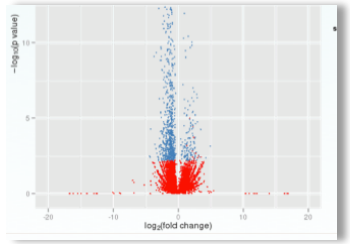
```
@SRR070570.4 HWUSI-EAS455:3:1:1:1096 length=41
CAAGGCCCGGGAACGAATTCACCGCCGTATGGCTGACCGGC
+
BA?39AAA933BA05>A@A=?4,9################
@SRR070570.12 HWUSI-EAS455:3:1:2:1592 length=41
GAGGCGTTGACGGGAAAAGGGATATTAGCTCAGCTGAATCT
+
@=:9>5+.5=?@<6>A?@6+2?:</7>,%1/=0/7/>48##
@SRR070570.13 HWUSI-EAS455:3:1:2:869 length=41
TGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCA
+
A;BAA6=A3=ABBBA84B<&78A@BA=(@B>AB2@>B@/9?
@SRR070570.32 HWUSI-EAS455:3:1:4:1075 length=41
CAGTAGTTGAGCTCCATGCGAAATAGACTAGTTGGTACCAC
+
BB9?A@>AABBBB@BCA?A8BBBAB4B@BC71=?9;B:3B?
@SRR070570.40 HWUSI-EAS455:3:1:5:238 length=41
AAAAGGGTAAAAGCTCGTTTGATTCTTATTTTCAGTACGAA
+
BBB?06-8BB@B17>9)=A91?>>8>*@<A<>>@1:B>(B@
@SRR070570.44 HWUSI-EAS455:3:1:5:1871 length=41
GTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTGTAAG
+
BBBCBCCBBBBBA@BBCCB+ABBCB@B@BB@:BAA@B@BB>
@SRR070570.46 HWUSI-EAS455:3:1:5:1981 length=41
GAACAACAAAACCTATCCTTAACGGGATGGTACTCACTTTC
+
?A>-?B;BCBBB@BC@/>A<BB:?<?B?=75?:9@@@3=>:
```



Bioinformatician



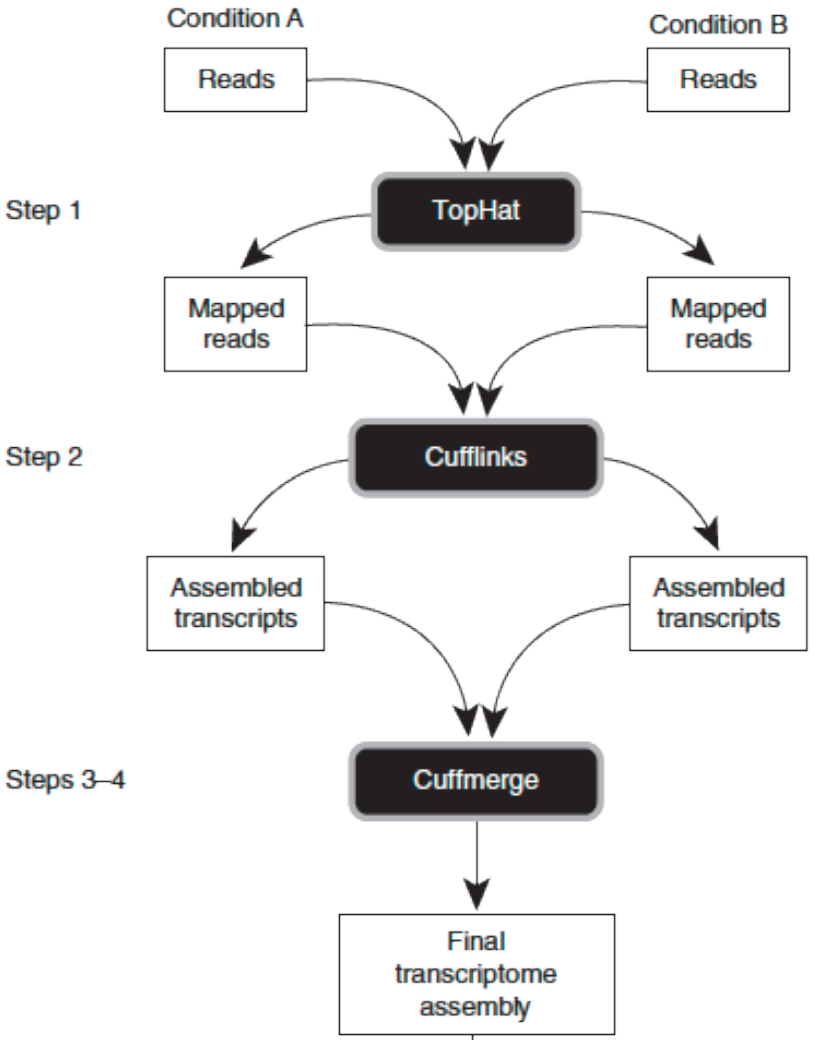


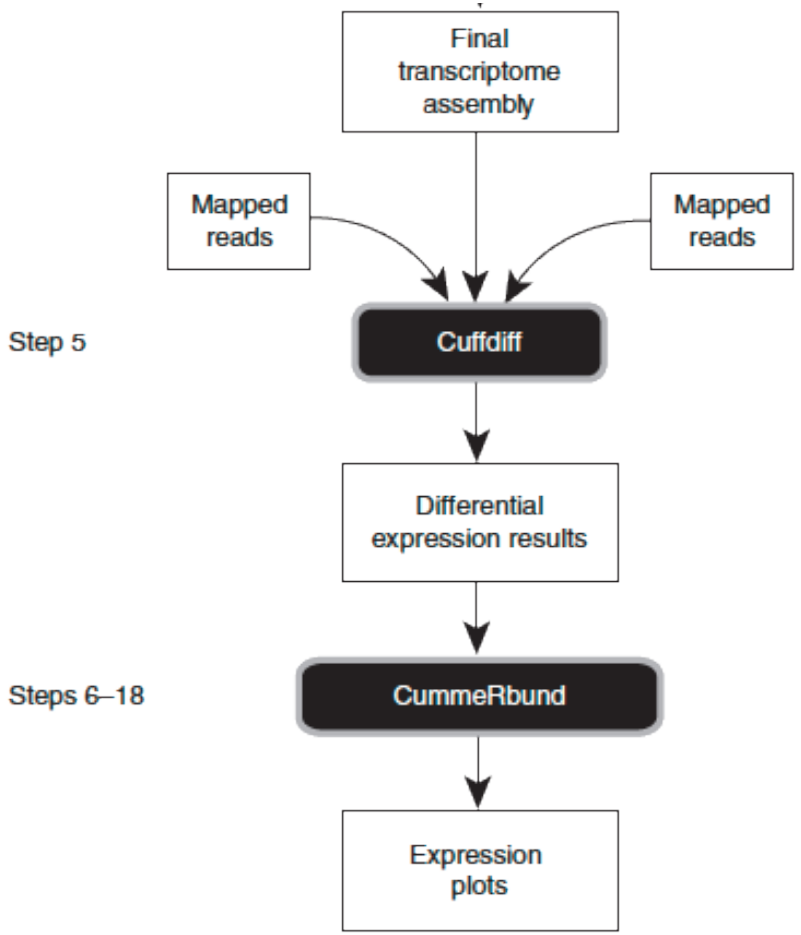


Tuxedo Workflow

Differential expression

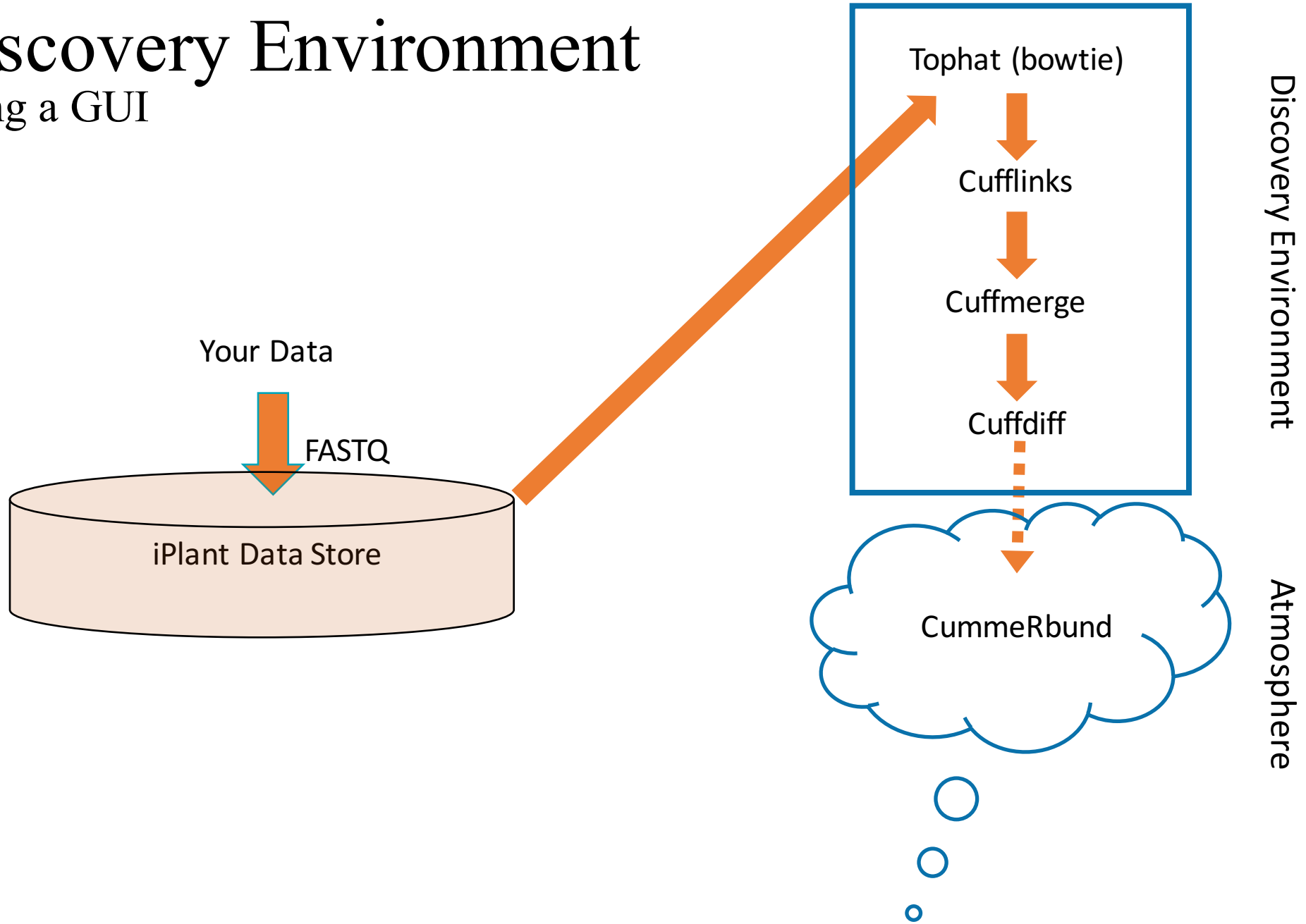
*TopHat and Cufflinks require a sequenced genome





Discovery Environment

Using a GUI



Genome annotation with MAKER-P

Overview

Annotate your Genome of Choice (in your lifetime)

See Full tutorial in

CyVerse Learning Center

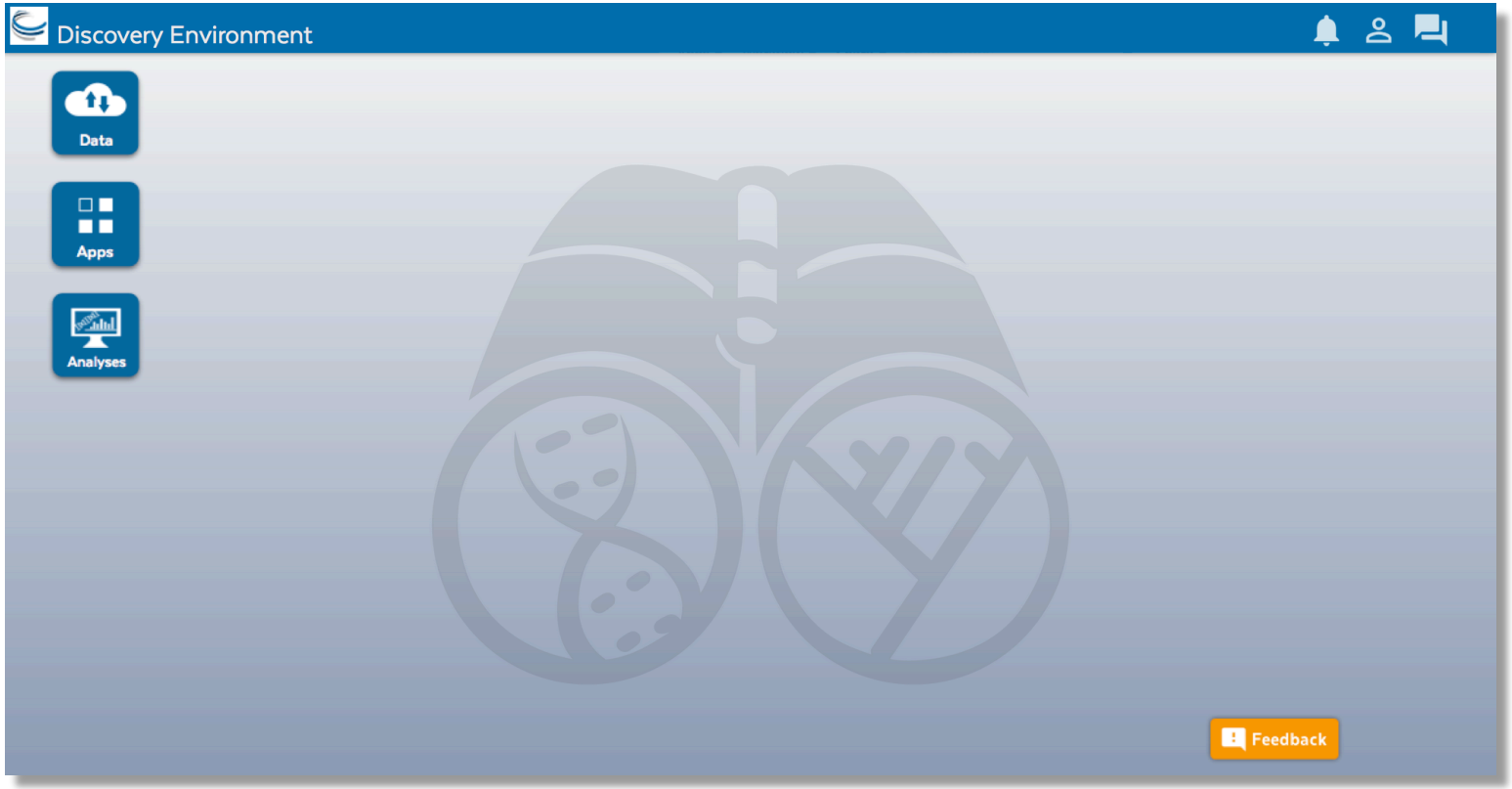
Genome annotation with MAKER-P

MAKER-P_2.28
c5104d19-b4a2-4304-beb2-4921ac61c1ca

MAKER-P_2.31_3_JBrowse
7888b8e1-c006-4794-82d9-4c940ddbf4c6
- AUGUSTUS
- BLAST
- blat
- RepeatMasker
- exonerate
- SNAP
- maker-p
- jbrowse

MAKER-P_2.31.3-07 resize
92ac582a-e2cc-4f76-8ac5-cc31fba99a76

MAKER with workqueue
9557128c-ef11-4218-a02c-eb1c590e90ed
- MAKER



What Are Annotations?

Quick Review

Annotations are descriptions of features of the genome

- Structural: exons, introns, UTRs, splice forms etc.
- Coding & non-coding genes
- Expression, repeats, transposons

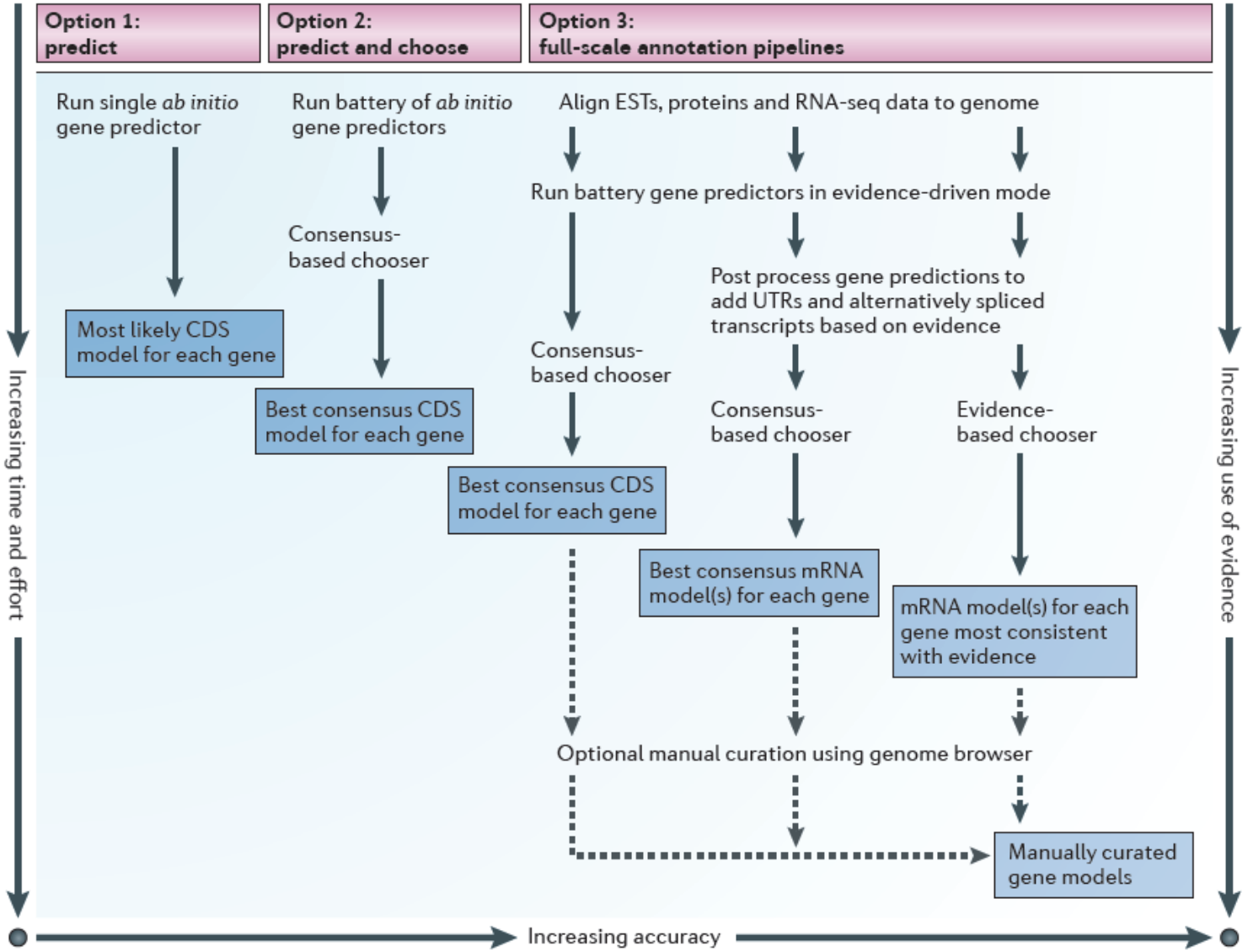
Annotations should include evidence trail

- Assists in quality control of genome annotations

Examples of evidence supporting a structural annotation:

- *Ab initio* gene predictions
- ESTs
- Protein homology

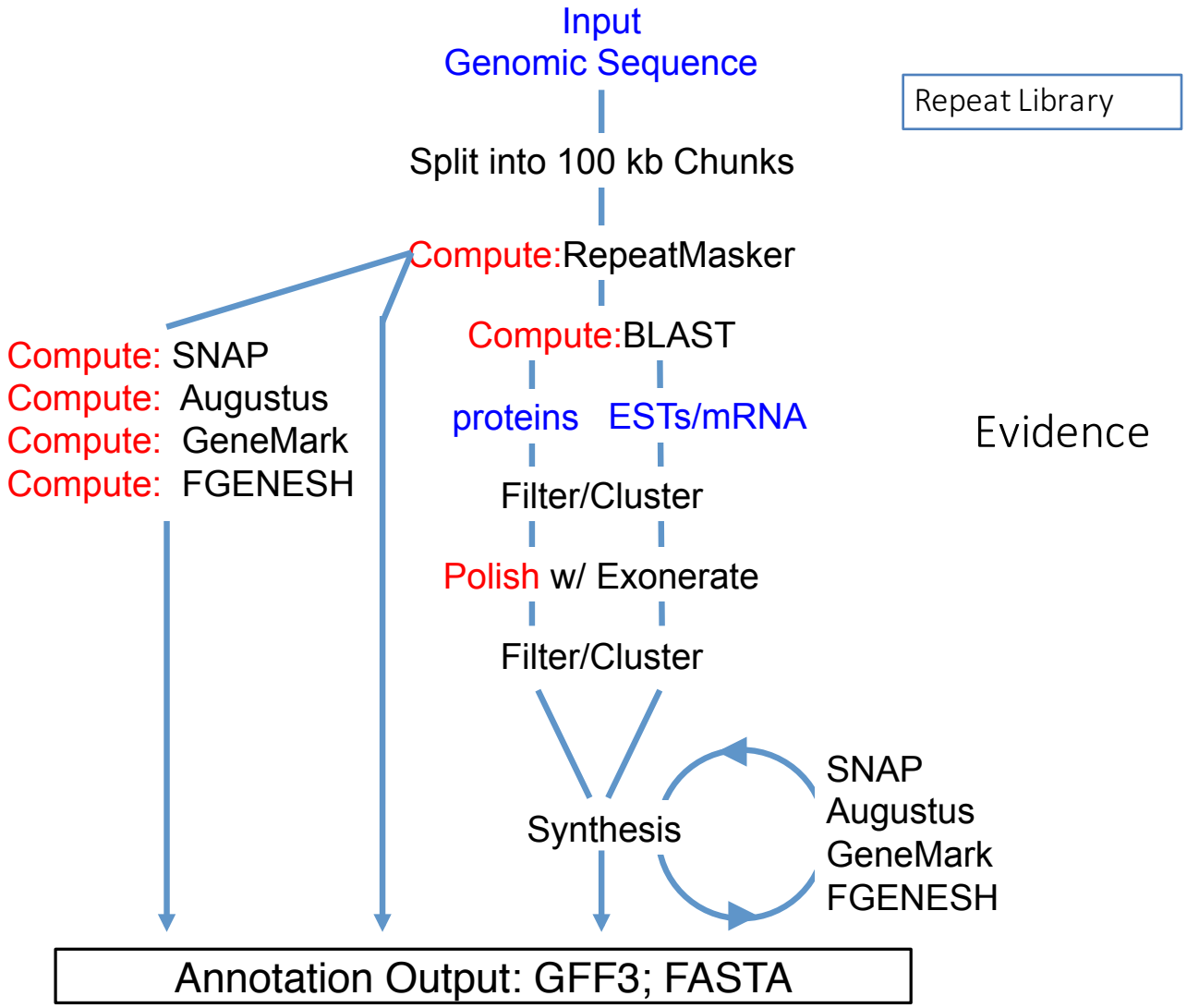
Options for Protein-coding Gene Annotation



Yandell & Ence. *Nature Reviews Genetics* **13**, 329-342 (May 2012) | doi:10.1038/nrg3174

MAKER-P Automated Pipeline

MPI-enabled to allow parallel operation on large compute clusters



Ab initio prediction

Evidence

Collaboration with Yandell Lab

MAKER-P at iPlant

TACC Lonestar Supercomputer

22,656 CPU cores on 1,888 nodes

Genome	Assembly	Size (Mb)	CPU	Run Time
Arabidopsis thaliana	TAIR10	120	600	2:44
Arabidopsis thaliana	TAIR10	120	1500	1:27
Zea mays	RefGen_v2	2067	2172	2:53

Campbell et al. Plant Physiology. December 4, 2013, DOI:10.1104/pp.113.230144

PAG 2014:

W559 - Annotation of the Lobolly Pine Megagenome—Jill Wegrzyn

- 20.15 Gb assembly—split into 40 jobs—216 CPU/job (8640 CPU total)—17 hours

P157 - Disease Resistance Gene Analysis on Chromosome 11 Across Ten Oryza Species

- 10 rice species (each w/12 chromosome pseudomolecules)
- 96 CPU per chromosome (1152 CPU total) ~2hr per genome

XSEDE
Extreme Science and Engineering Discovery Environment

Agave API

Transcriptome assembly with SOAPdenovo-Trans

Overview

Generate and validate your transcriptome

See Full tutorial in

CyVerse Learning Center

Why SOAPdenovo-Trans?

Some comparisons

Table 1. Computational requirements.

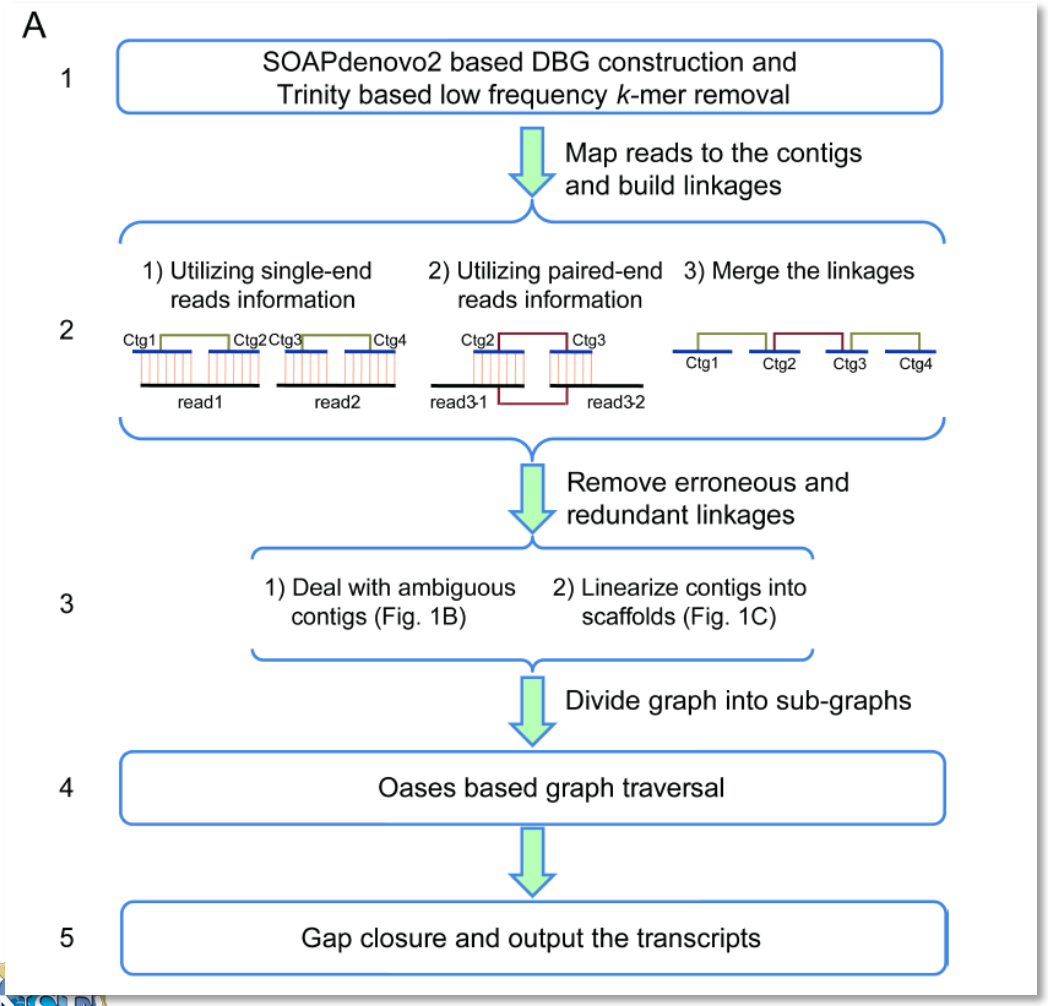
Method	Rice				Mouse			
	Small dataset		Large dataset		Small dataset		Large dataset	
	Peak memory (GB)	Time (hr)	Peak memory (GB)	Time (hr)	Peak memory (GB)	Time (hr)	Peak memory (GB)	Time (hr)
SOAPdenovo-Trans	10.7	0.2	29.3	0.8	10.5	0.3	16.7	1.0
Trinity	10.1	17.7	37.6	35.6	10.5	16.6	26.3	47.5
Oases	9.0	0.6	53.2	3.0	8.8	0.7	35.1	2.7

All assemblies were processed with 10 threads, on a computer with two Quad-core Intel 2.8GHz CPUs and 70GB of memory, running CentOS 5.

- *Runs more quickly (easier to refine parameters)
- Less memory demands
- Good quality (Software changes rapidly, so “clear winners” will always change, you can too when the time comes)

http://arxiv.org/ftp/arxiv/papers/1305/1305.6760.pdf

SOAPdenovo-Trans
Overview

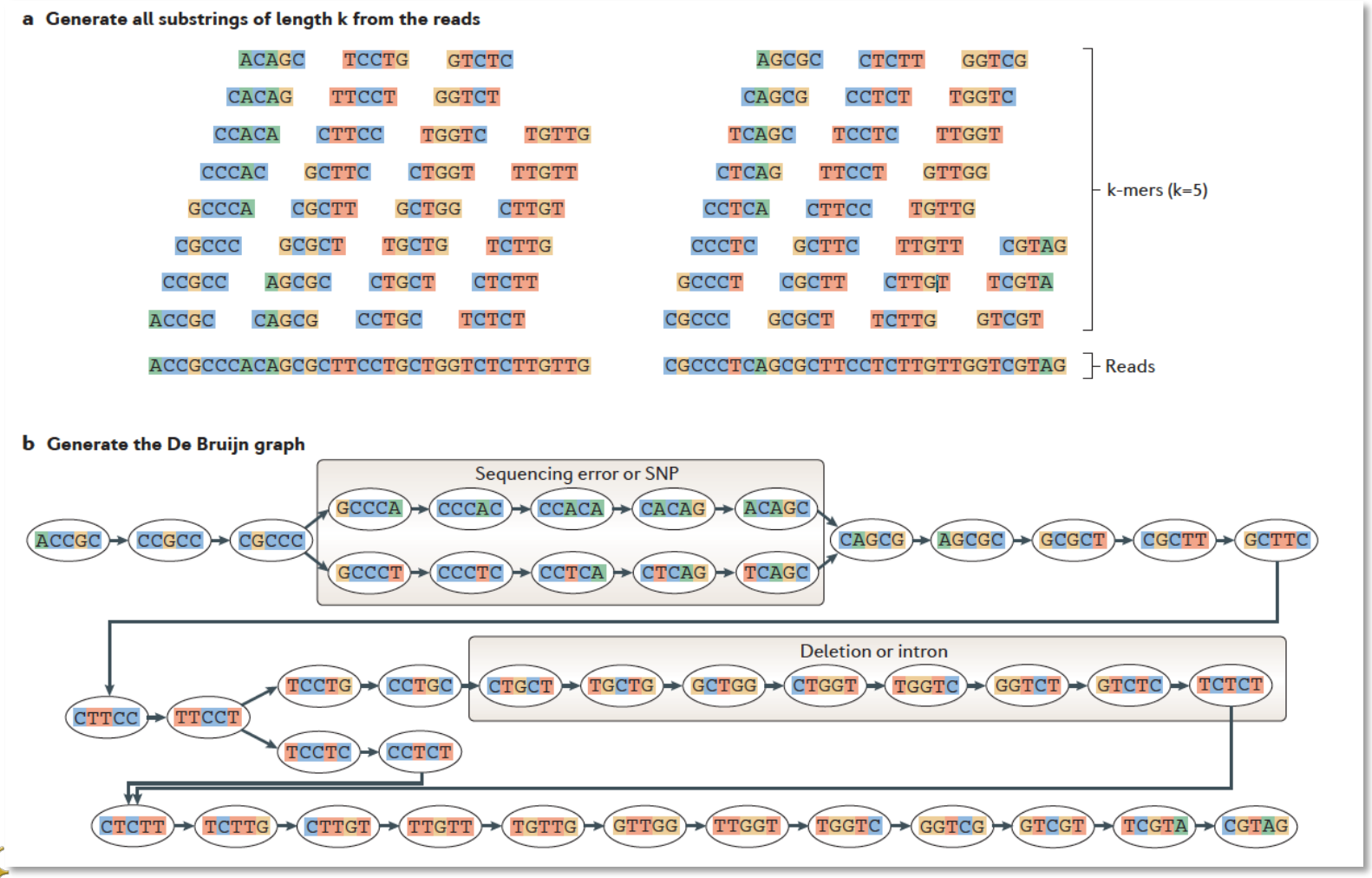


- *De Brujin* graphs are constructed
- Error correction
- Contigs are constructed and single/paired reads are mapped to contigs to make scaffold graphs
- Transcripts are created from scaffold graphs

Figure from: SOAPdenovo-Trans: De novo transcriptome assembly with shortf RNA-Seq reads Yinlong Xie1,2,3,†, Gengxiong Wu1,†, et.al.1BGI-Shenzhen, Shenzhen, China.

SOAPdenovo-Trans

Kmers and De Brujin graphs

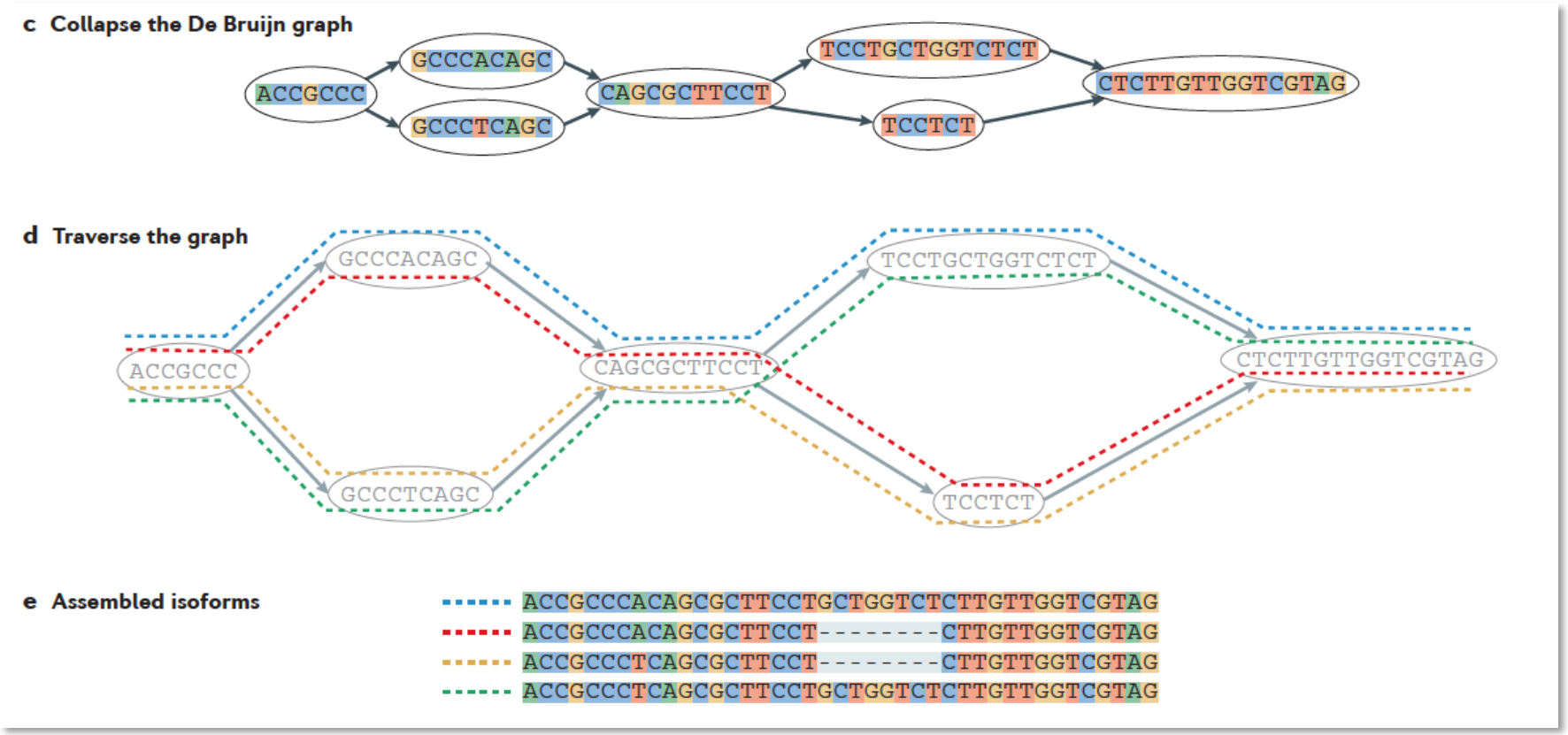


- Reads split into k-mers
- De Brujin graph constructed from kmers

Next-generation transcriptome assembly, *Jeffrey A. Martin and Zhong Wang – Nat.Reviw.Gen* doi:10.1038/nrg3068
Published online 7 September 2011

SOAPdenovo-Trans

Kmers and De Brujin graphs

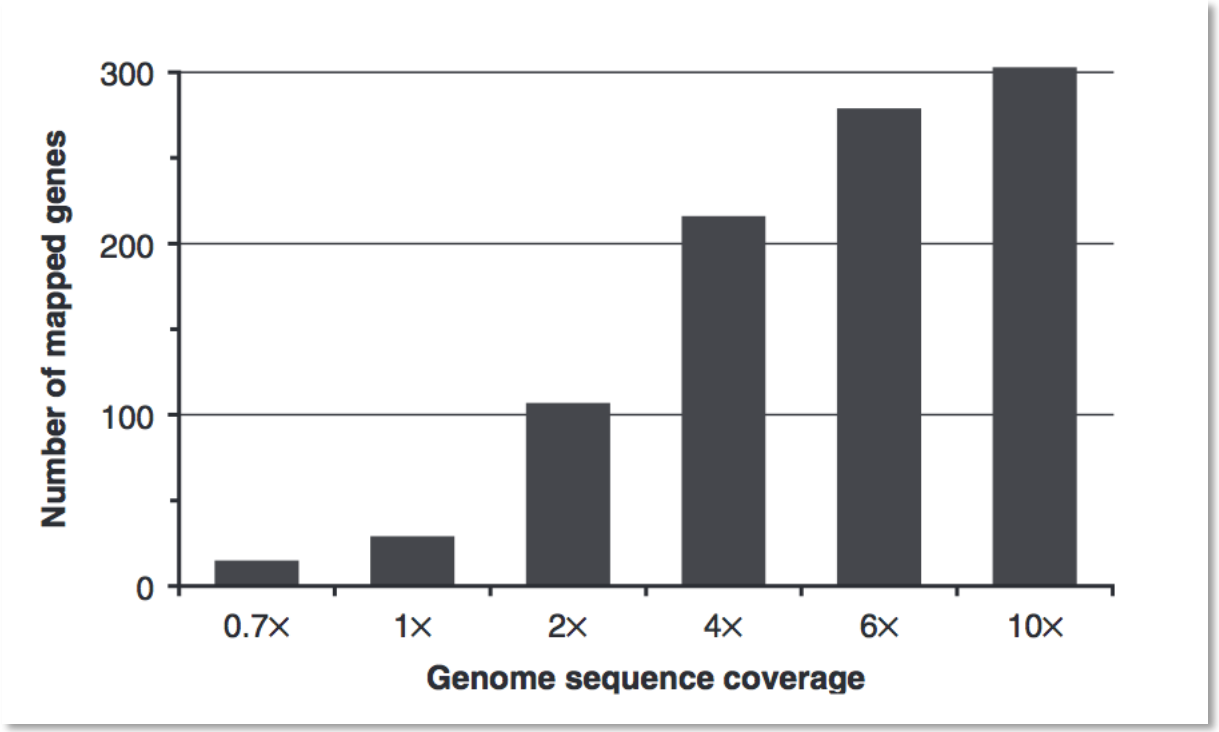


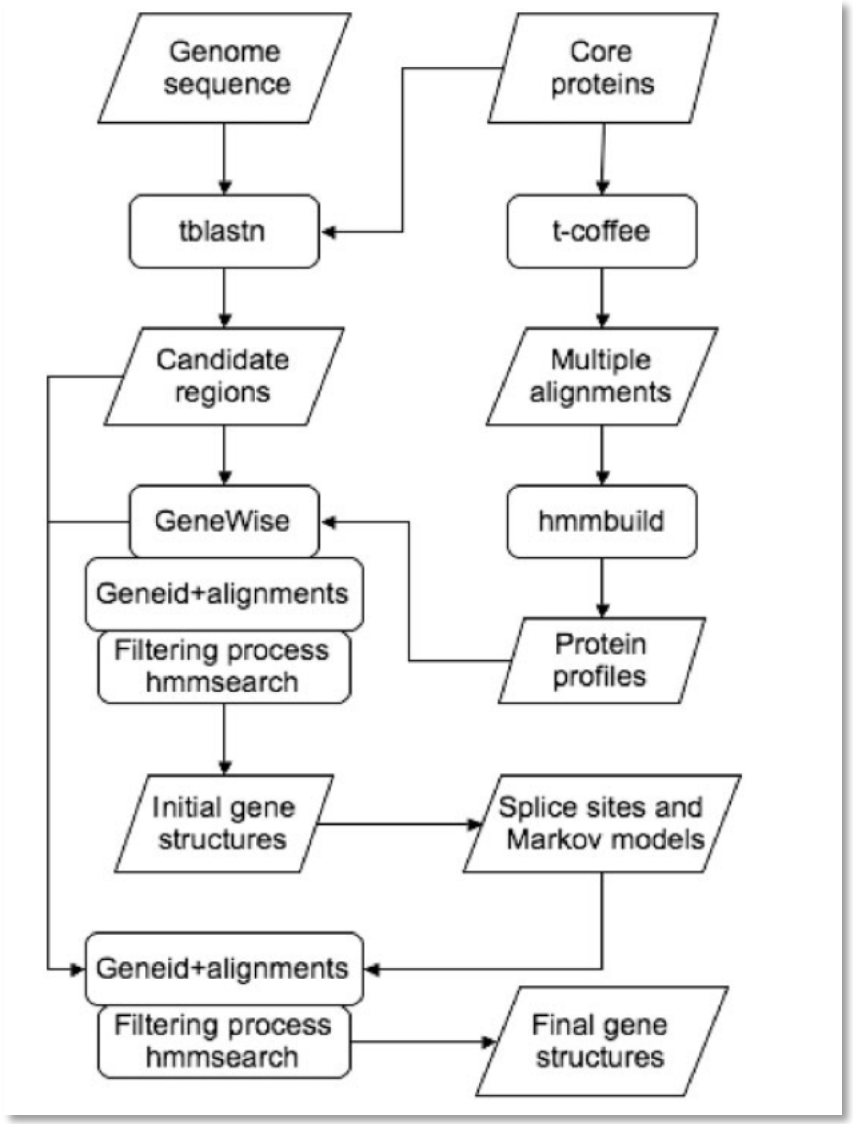
- Redundancies are collapsed
- Paths through the graph that explained the observed sequence generate the alignments

Next-generation transcriptome assembly, *Jeffrey A. Martin and Zhong Wang – Nat.Reviw.Gen* doi:10.1038/nrg3068
Published online 7 September 2011

CEGMA

How good is assembly coverage?





CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes
Bioinformatics (2007) 23 (9): 1061-1067. doi: 10.1093/bioinformatics/btm071 First published online: March 1, 2007

Sample Data

Asian honeybee transcriptome

plos.org

PLOS ONE | Subject Areas | For Authors | About Us | Searc

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

2,909 VIEWS 9 CITATIONS

Transcriptome Analysis of the Asian Honey Bee *Apis cerana cerana*

Zi Long Wang, Ting Ting Liu, Zachary Y. Huang, Xiao Bo Wu, Wei Yu Yan, Zhi Jiang Zeng

- Published: October 24, 2012
- DOI: 10.1371/journal.pone.0047954

Article | About the Authors | Metrics | Comments | Related Content

- Abstract
- Introduction
- Results
- Discussion
- Conclusions

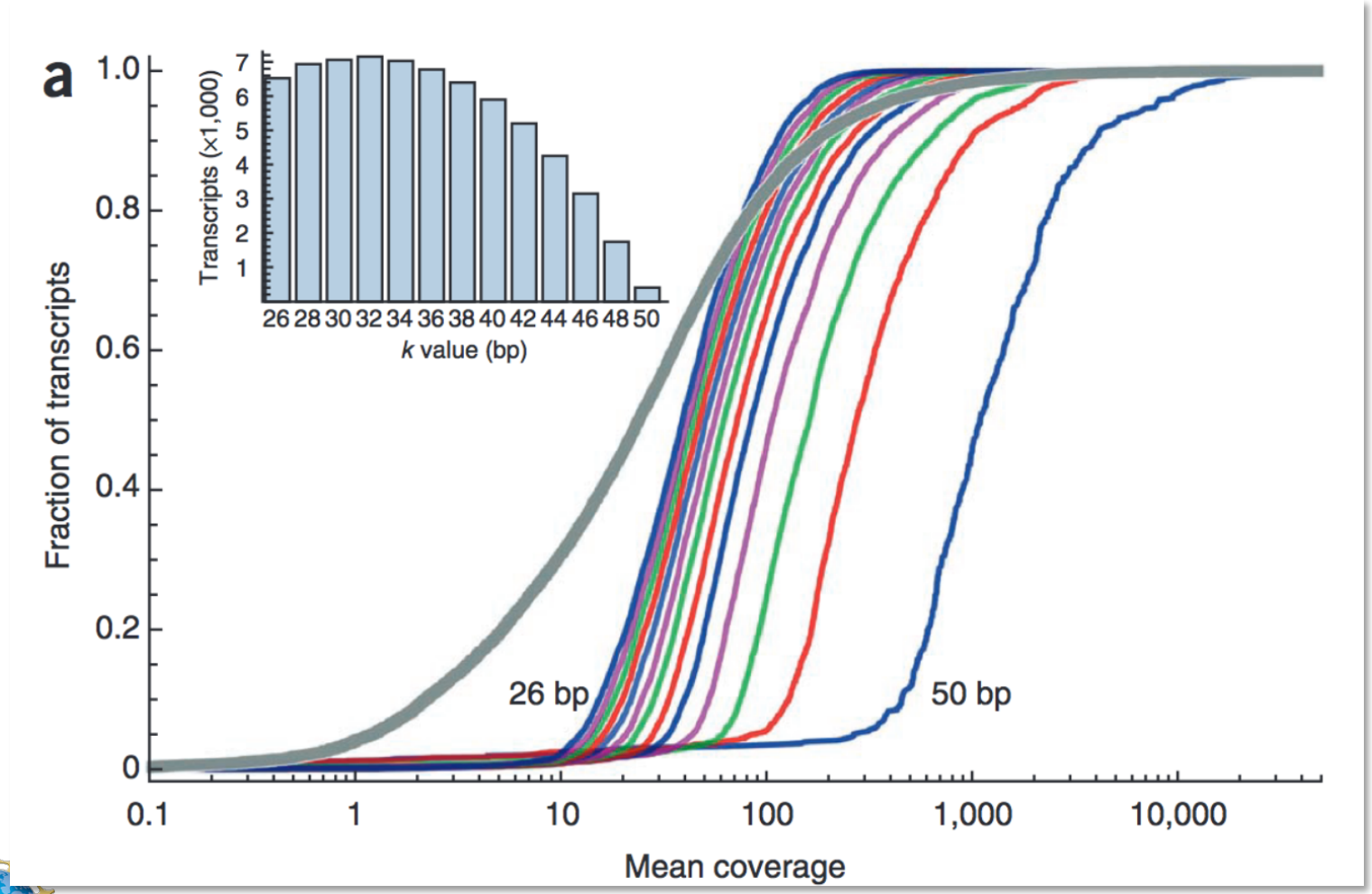
Abstract

Background

The Eastern hive honey bee, *Apis cerana cerana* is a native and widely bred honey bee

SOAPdenovo-Trans

Choosing kmers



- Transcripts with lower read depths were represented better with lower K values
- Transcripts with higher read depth represented better with higher K

***De novo* assembly and analysis of RNA-seq data.** Nature Methods 7, 909–912 (2010) doi:10.1038/nmeth.1517