

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

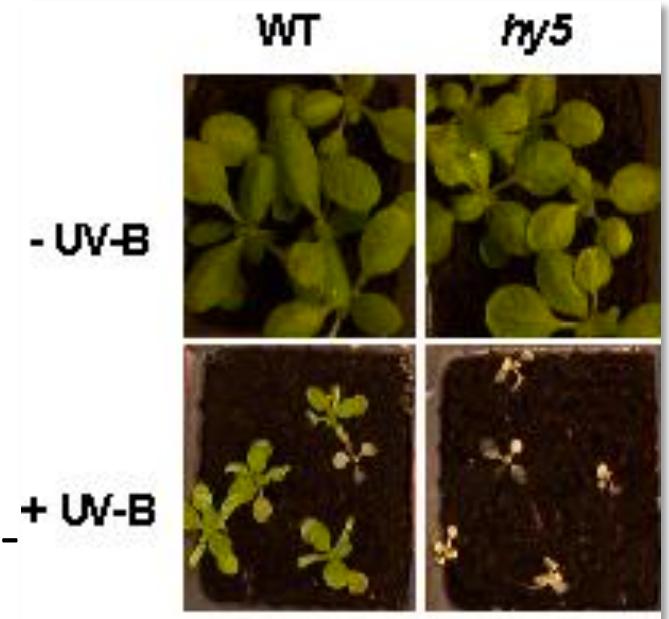




Experiment Overview

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.



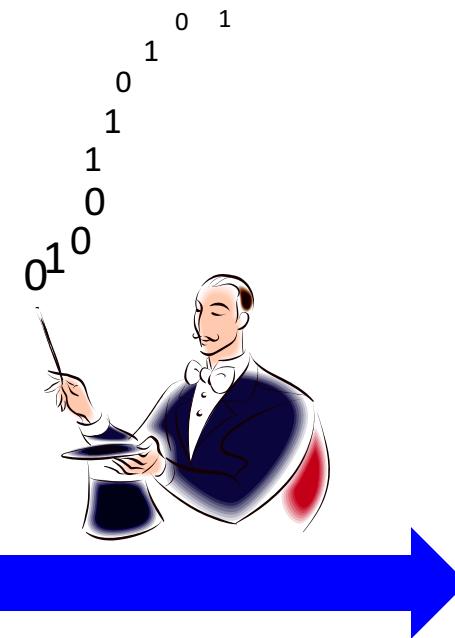
Source: http://www.gla.ac.uk/media/media_73736_en.jpg



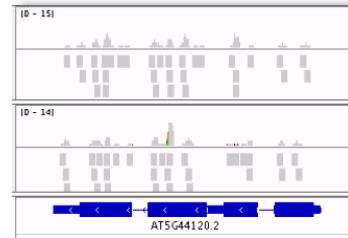
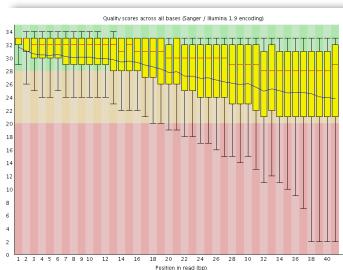


Now what?

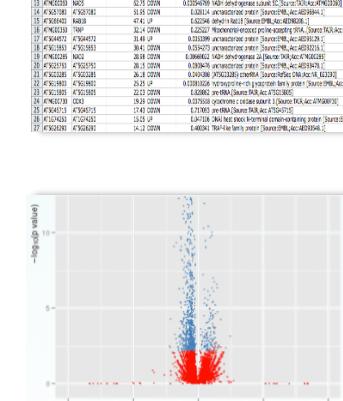
```
@SRR070570.4 HWUSI-EAS455:3:1:1:1096 length=41
CAAGGCCCGGGAACGAATTCACCGCCGTATGGCTGACCGGC
+
BA?39AAA933BA05>A@A=?4,9#####
@SRR070570.12 HWUSI-EAS455:3:1:2:1592 length=41
GAGGC GTTGACGGGAAAAGGGATATTAGCTCAGCTGAATCT
+
@=:9>5+.5=?@<6>A?@6+2?:</7>,%1/=0/7/>48##
@SRR070570.13 HWUSI-EAS455:3:1:2:869 length=41
TGCCAGTAGTCATATGCTTGTCTAAAGATTAAGCCATGCA
+
A;BAA6=A3=ABBBA84B<&78A@BA=( @B>AB2@>B@/9?
@SRR070570.32 HWUSI-EAS455:3:1:4:1075 length=41
CAGTAGTTGAGCTCATGCGAAATAGACTAGTTGGTACCAC
+
BB9?A@>AABBBA@BCA?A8BBBAB4B@BC71=?9;B:3B?
@SRR070570.40 HWUSI-EAS455:3:1:5:238 length=41
AAAAGGGTAAAGCTCGTTGATTCTTATTTCAGTACGAA
+
BBBB?06-8BB@B17>9)=A91?>>8>*<A<>>@1:B>(B@
@SRR070570.44 HWUSI-EAS455:3:1:5:1871 length=41
GTCATATGCTTGTCTAAAGATTAAGCCATGCATGTGTAAG
+
BBBBCBCCBBBBBA@BBCB+ABBCB@B@BB@:BAA@B@BB>
@SRR070570.46 HWUSI-EAS455:3:1:5:1981 length=41
GAACAACAAAACCTATCCTAACGGATGGTACTCACTTTC
+
?A>-?B;BCBBB@BC@/>A<BB:>?<?B?=75?:9@@3=>:
```



Bioinformatician



#	read_id	gene_name	fold_change	direction	q-value	base description
1	F000001		100.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
2	F000002		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
3	F000003		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
4	F000004		100.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
5	F000005		400.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
6	F000006		400.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
7	F000007		100.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
8	F000008		100.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
9	F000009		200.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
10	F000010		70.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
11	F000011		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
12	F000012		900.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
13	F000013		900.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
14	F000014		900.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
15	F000015		900.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
16	F000016		900.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
17	F000017		200.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
18	F000018		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
19	F000019		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
20	F000020		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
21	F000021		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
22	F000022		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
23	F000023		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
24	F000024		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
25	F000025		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
26	F000026		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
27	F000027		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
28	F000028		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
29	F000029		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
30	F000030		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
31	F000031		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
32	F000032		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
33	F000033		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
34	F000034		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
35	F000035		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
36	F000036		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
37	F000037		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
38	F000038		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
39	F000039		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
40	F000040		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
41	F000041		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
42	F000042		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
43	F000043		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
44	F000044		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
45	F000045		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
46	F000046		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
47	F000047		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
48	F000048		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
49	F000049		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
50	F000050		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
51	F000051		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
52	F000052		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
53	F000053		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
54	F000054		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
55	F000055		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
56	F000056		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
57	F000057		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
58	F000058		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
59	F000059		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
60	F000060		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
61	F000061		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
62	F000062		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
63	F000063		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
64	F000064		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
65	F000065		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
66	F000066		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
67	F000067		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
68	F000068		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
69	F000069		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
70	F000070		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
71	F000071		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
72	F000072		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
73	F000073		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
74	F000074		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
75	F000075		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
76	F000076		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
77	F000077		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
78	F000078		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
79	F000079		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
80	F000080		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
81	F000081		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
82	F000082		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
83	F000083		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
84	F000084		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
85	F000085		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
86	F000086		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
87	F000087		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
88	F000088		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
89	F000089		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
90	F000090		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
91	F000091		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
92	F000092		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
93	F000093		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
94	F000094		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
95	F000095		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
96	F000096		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
97	F000097		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
98	F000098		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
99	F000099		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)
100	F000100		300.0	Up	0.0001	Primer sequence, Sanger 1.9 (Sanger)

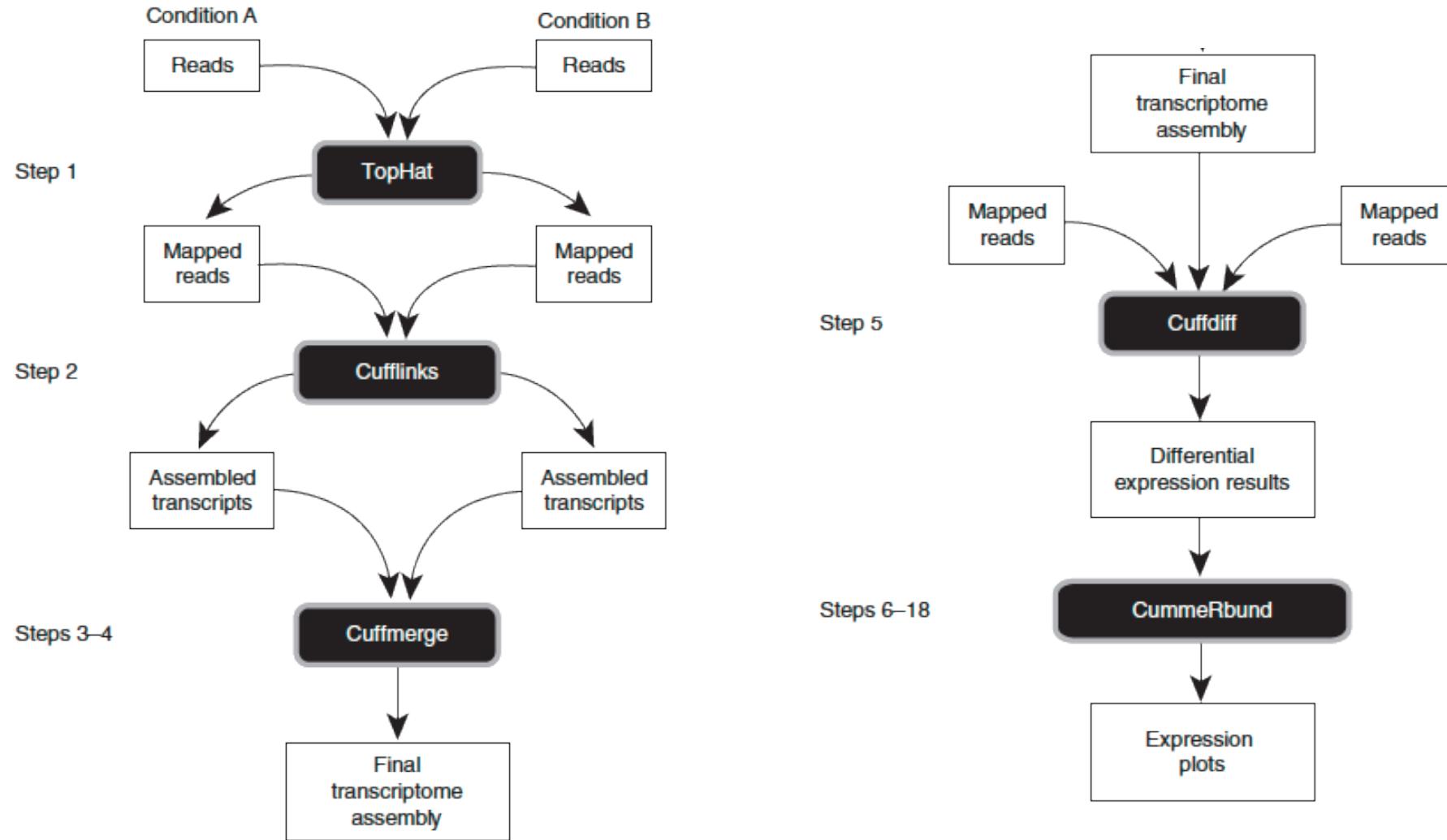




Tuxedo Workflow

Differential expression

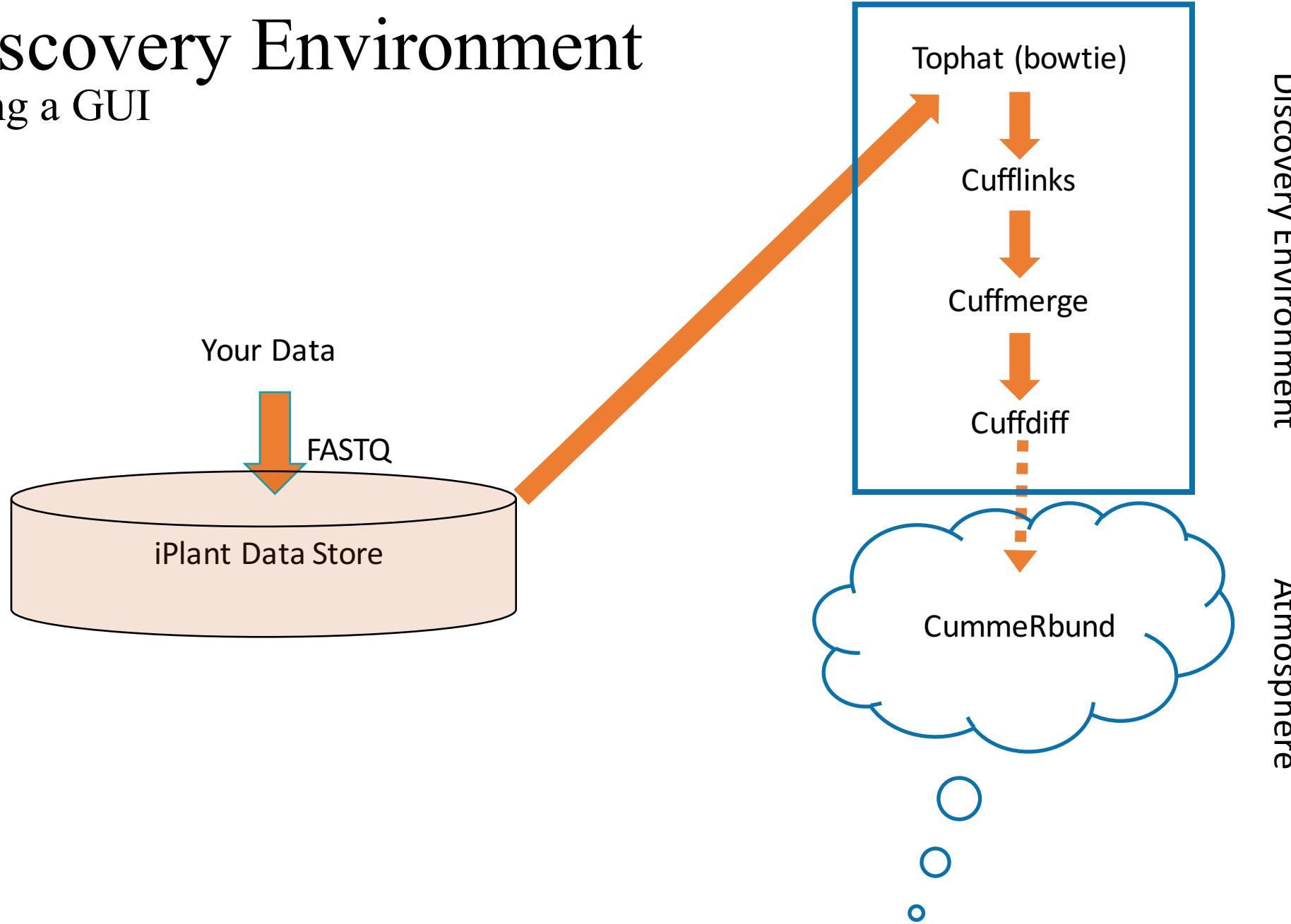
*TopHat and Cufflinks require a sequenced genome





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

The screenshot shows the Discovery Environment web application. At the top, there's a header bar with the title "Discovery Environment" and icons for notifications, user profile, and help. Below the header, there are three main navigation buttons: "Data" (with a cloud icon), "Apps" (with a grid icon), and "Analyses" (with a monitor icon). A large, semi-transparent watermark featuring a brain and a DNA double helix is centered on the page. In the bottom right corner, there's a yellow "Feedback" button.



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

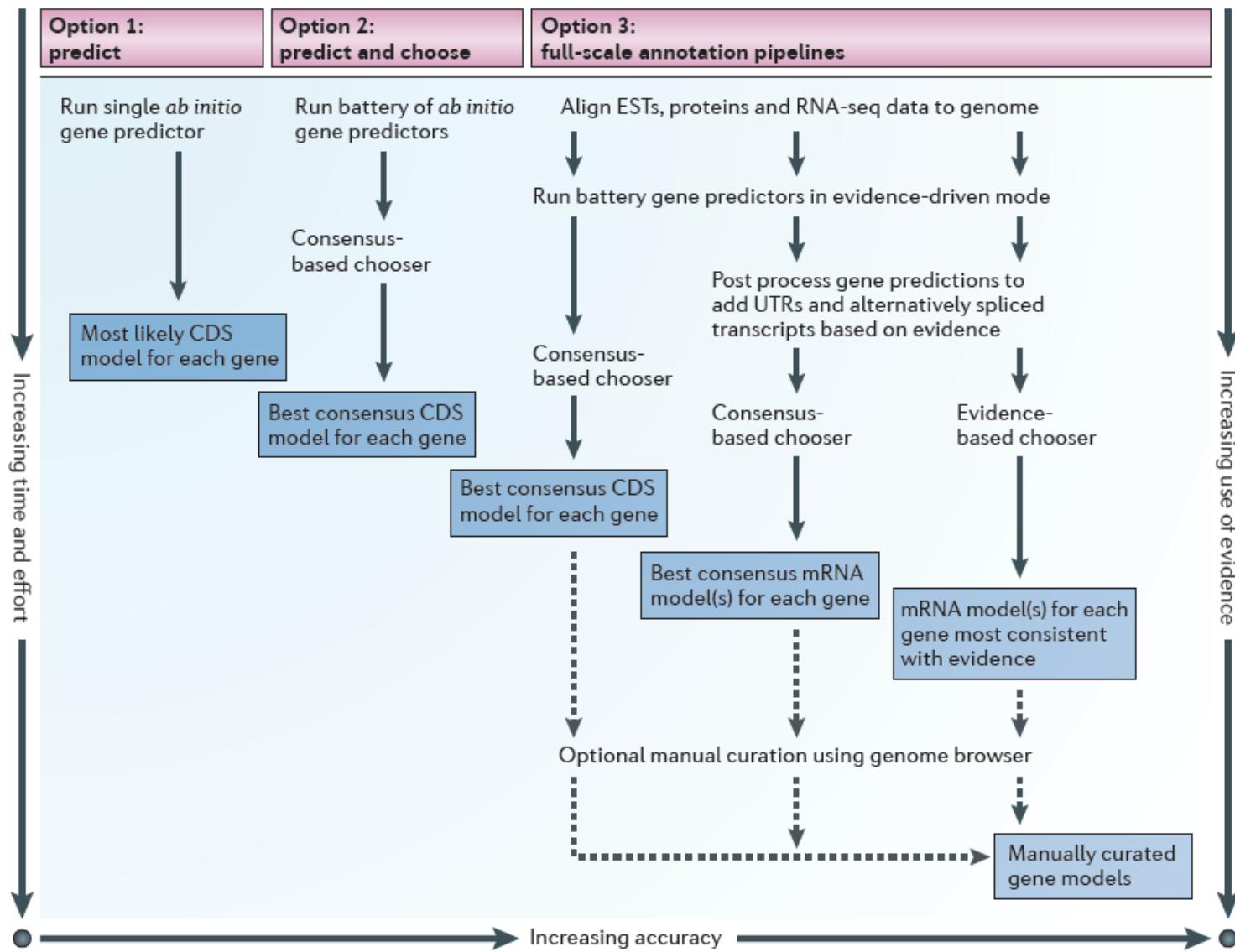
Examples of evidence supporting a structural annotation:

- *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

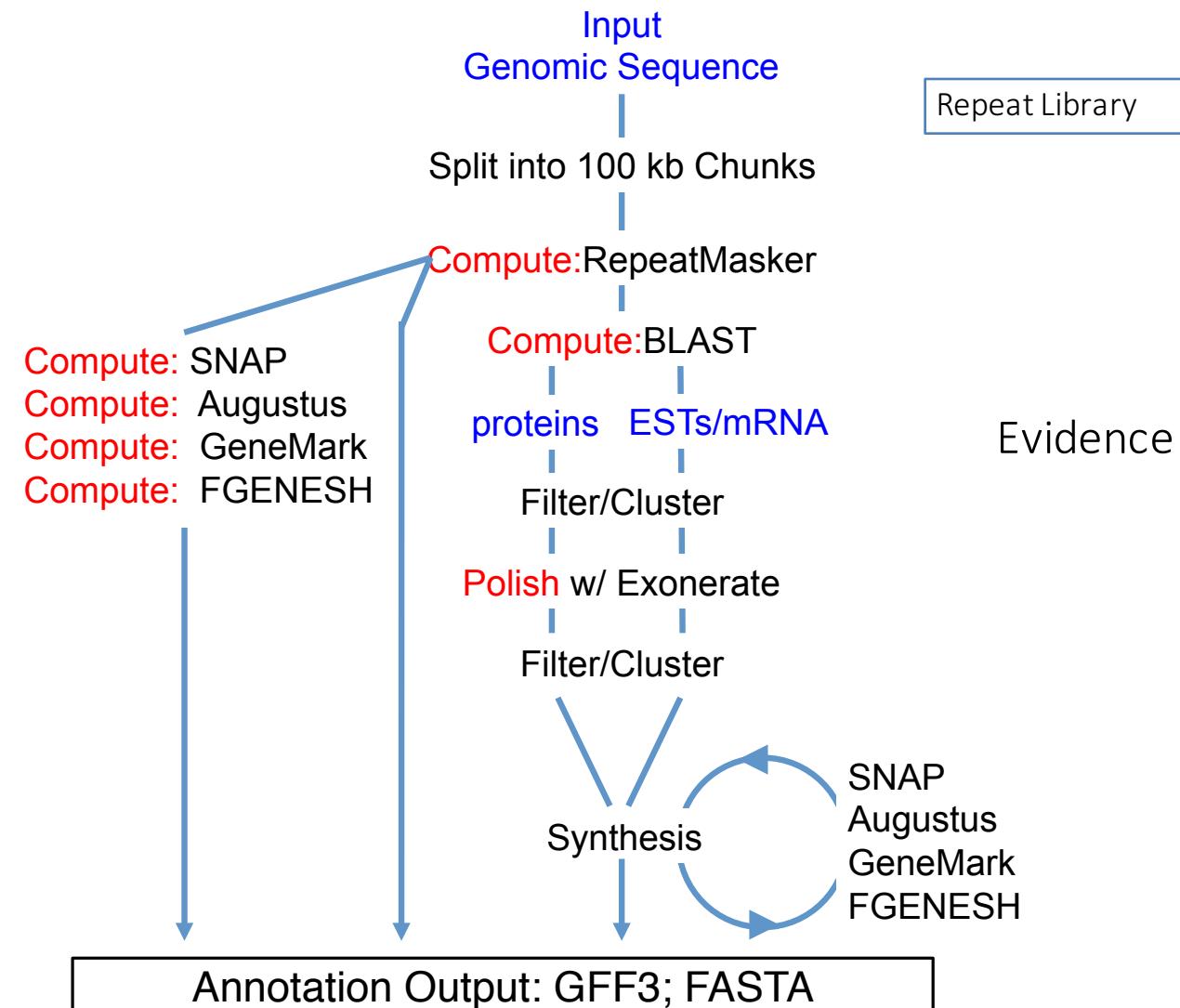




MAKER-P Automated Pipeline

MPI-enabled to allow parallel operation on large compute clusters

Ab initio
prediction





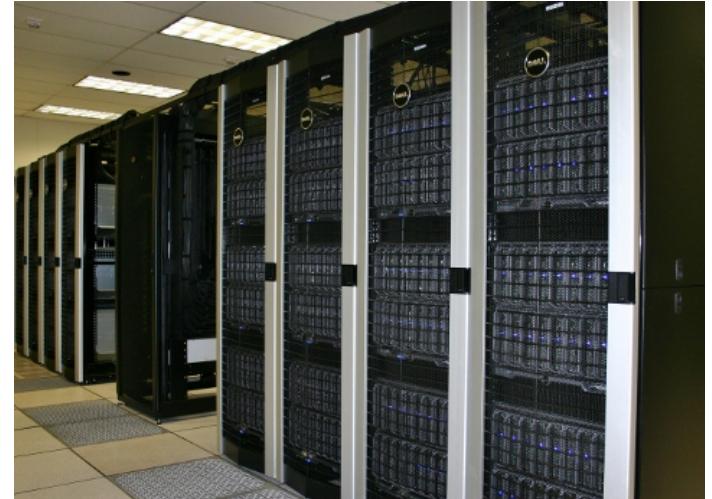
MAKER-P at iPlant

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)						
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.

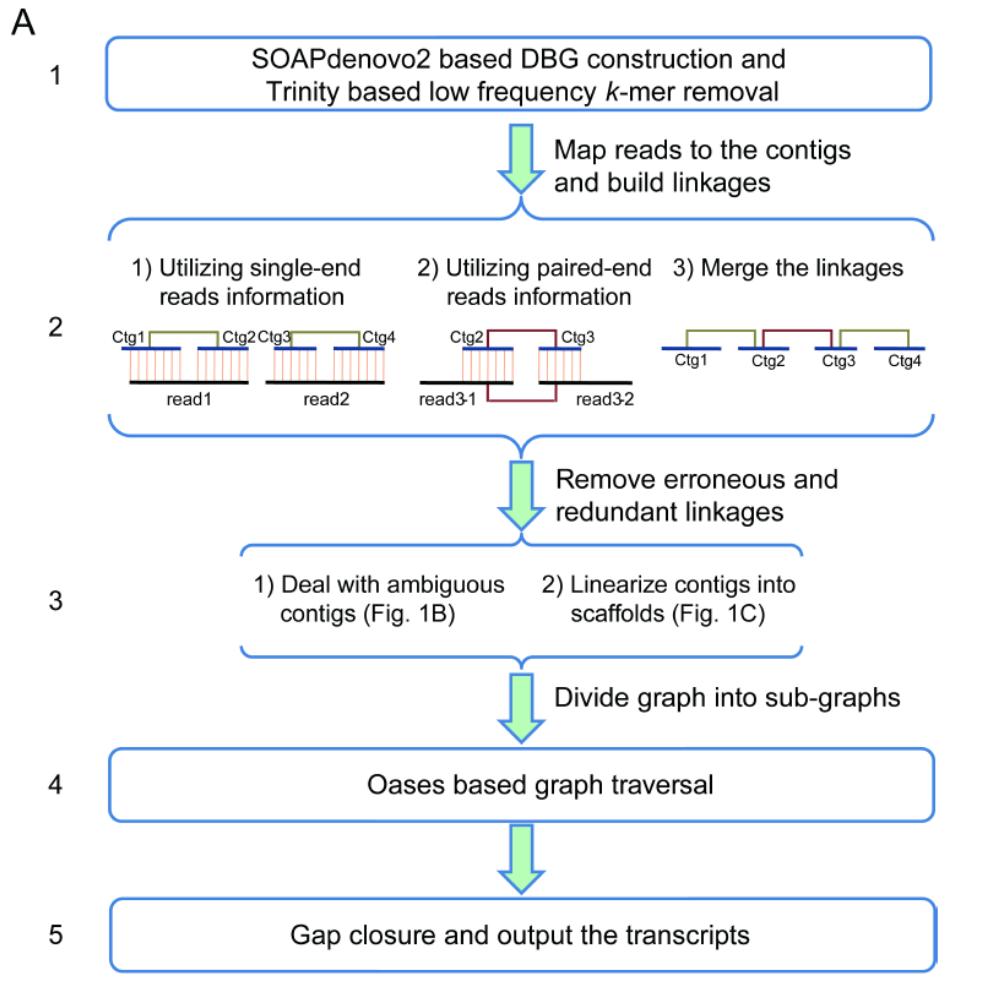
- *Runs more quickly (easier to refine parameters)
- Less memory demands
- Good quality (Software changes rapidly, so “clear winners” will always change, you can too when the time comes)





SOAPdenovo-Trans

Overview



- *De Bruijn* 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 short RNA-Seq reads Yinlong Xie1,2,3,†, Gengxiong Wu1,†, et.al.1BGI-Shenzhen, Shenzhen, China.



SOAPdenovo-Trans

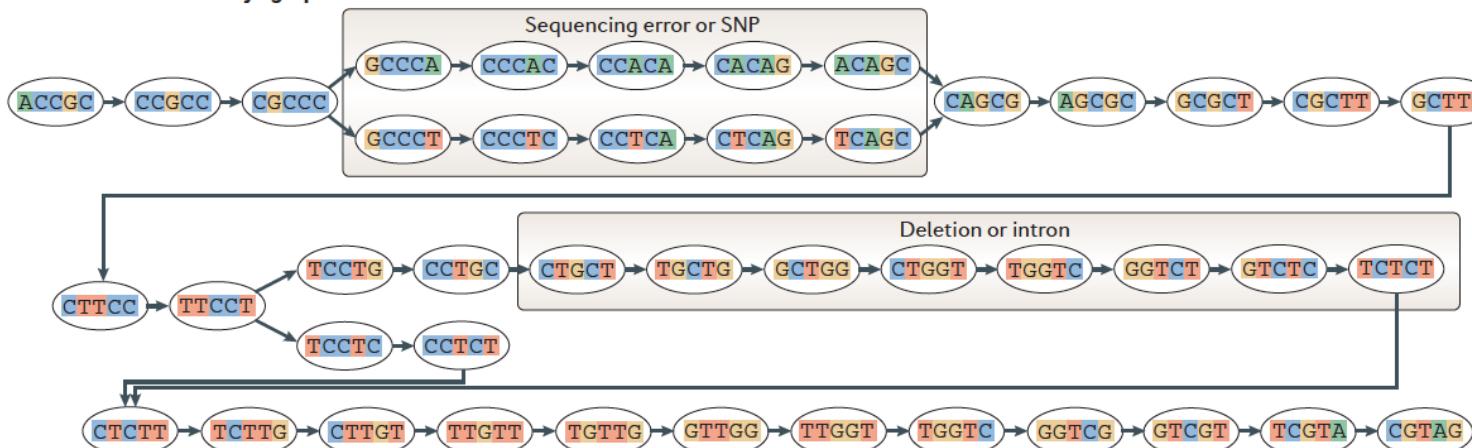
Kmers and De Bruijn graphs

a Generate all substrings of length k from the reads

ACAGC	TCC TG	GTCTC				
CACAG	TT CCT	GGT CT				
CCACA	CTT CC	TGG TC	TGTTG			
CCCAC	GCT TC	CT GGT	TTG TT			
GCCCC	CG CTT	GCT GG	CTT GT			
CGCCC	GCG CT	TG CTG	TCT TG			
CCGCC	AG CGC	CT GCT	CT CTT			
ACCGC	CAG CG	CCT GC	TCT CT			
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTG	CGCCCTCAGCGCTTCCTCTTGTTGGTCGTAG	Reads				

} k-mers (k=5)

b Generate the De Bruijn graph



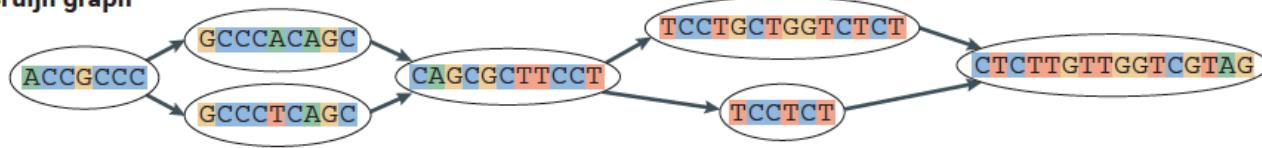
- Reads split into k-mers
- De Bruijn graph constructed from kmers



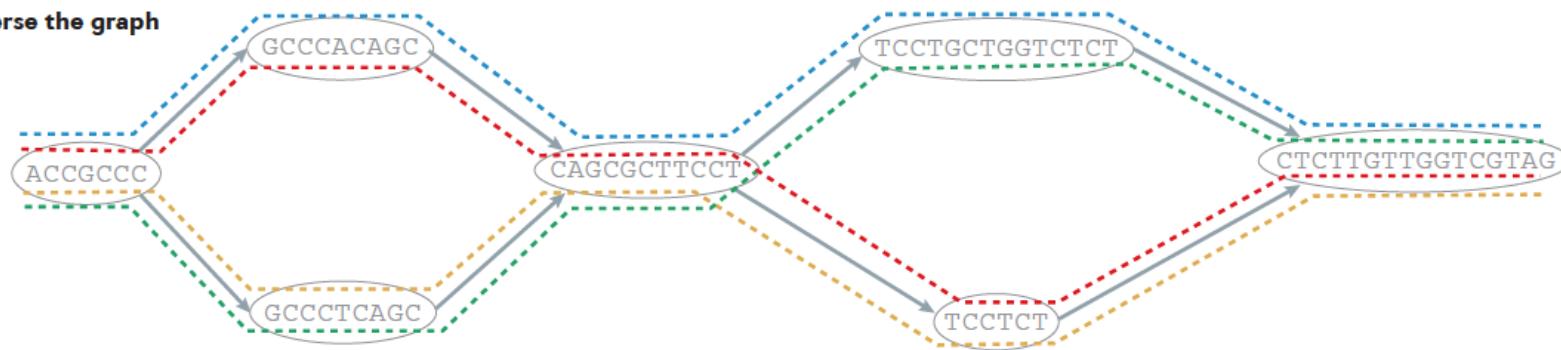
SOAPdenovo-Trans

Kmers and De Bruijn graphs

c Collapse the De Bruijn graph



d Traverse the graph



e Assembled isoforms

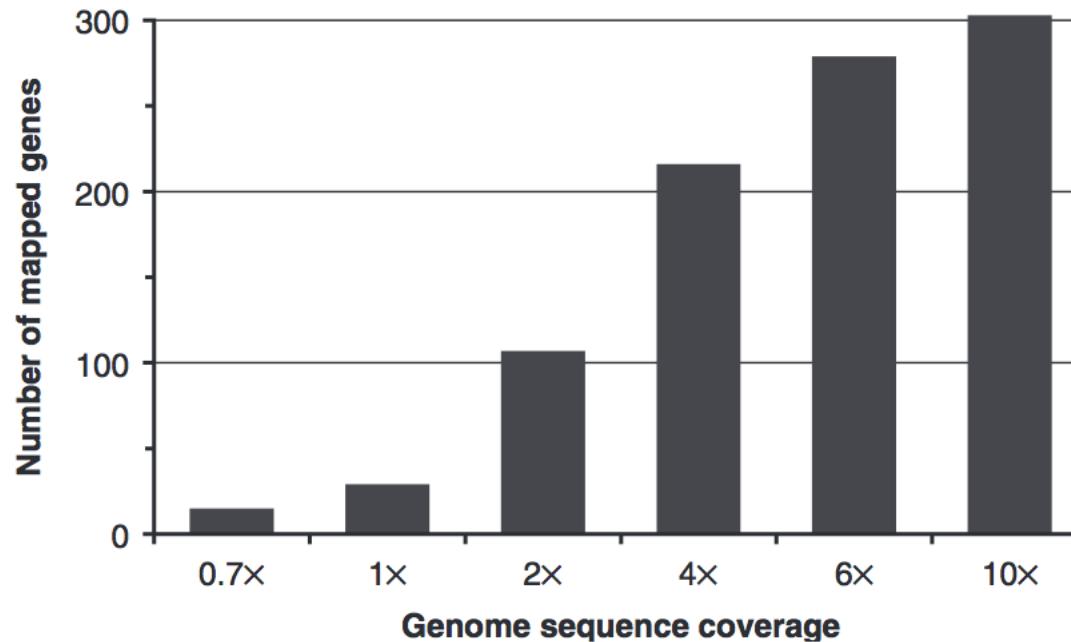
— ACCGGCCCACAGCGCTTCCTGCTGGTCTCTTTGGTGGTCGTAG
- - - ACCGGCCCACAGCGCTTCCT - - - CTTGGTGGTGGTCGTAG
--- ACCGGCCCTCAGCGCTTCCT --- CTTGGTGGTGGTCGTAG
---- ACCGGCCCTCAGCGCTTCCTGCTGGTCTCTTTGGTGGTCGTAG

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



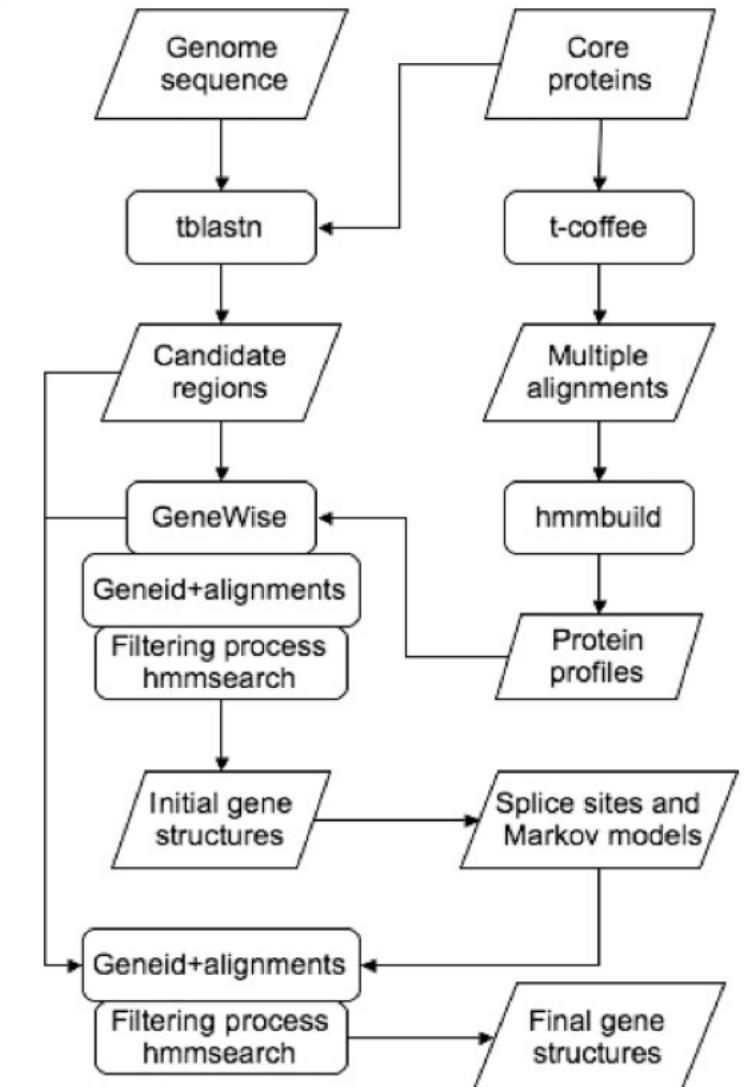
CEGMA

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

The screenshot shows a PLOS ONE article page. At the top, there are navigation links for "Subject Areas", "For Authors", "About Us", and a search bar. Below that, it displays "OPEN ACCESS" and "PEER-REVIEWED" status, along with "RESEARCH ARTICLE". To the right, metrics are shown: "2,909 VIEWS" and "9 CITATIONS". The main title of the article is "Transcriptome Analysis of the Asian Honey Bee *Apis cerana cerana*". Below the title, the authors listed are Zi Long Wang, Ting Ting Liu, Zachary Y. Huang, Xiao Bo Wu, Wei Yu Yan, and Zhi Jiang Zeng. A list of publication details follows: "Published: October 24, 2012" and "DOI: 10.1371/journal.pone.0047954".

Article	About the Authors	Metrics	Comments	Related Content
▼				

► Abstract

Introduction

Results

Discussion

Conclusions

Abstract

Background

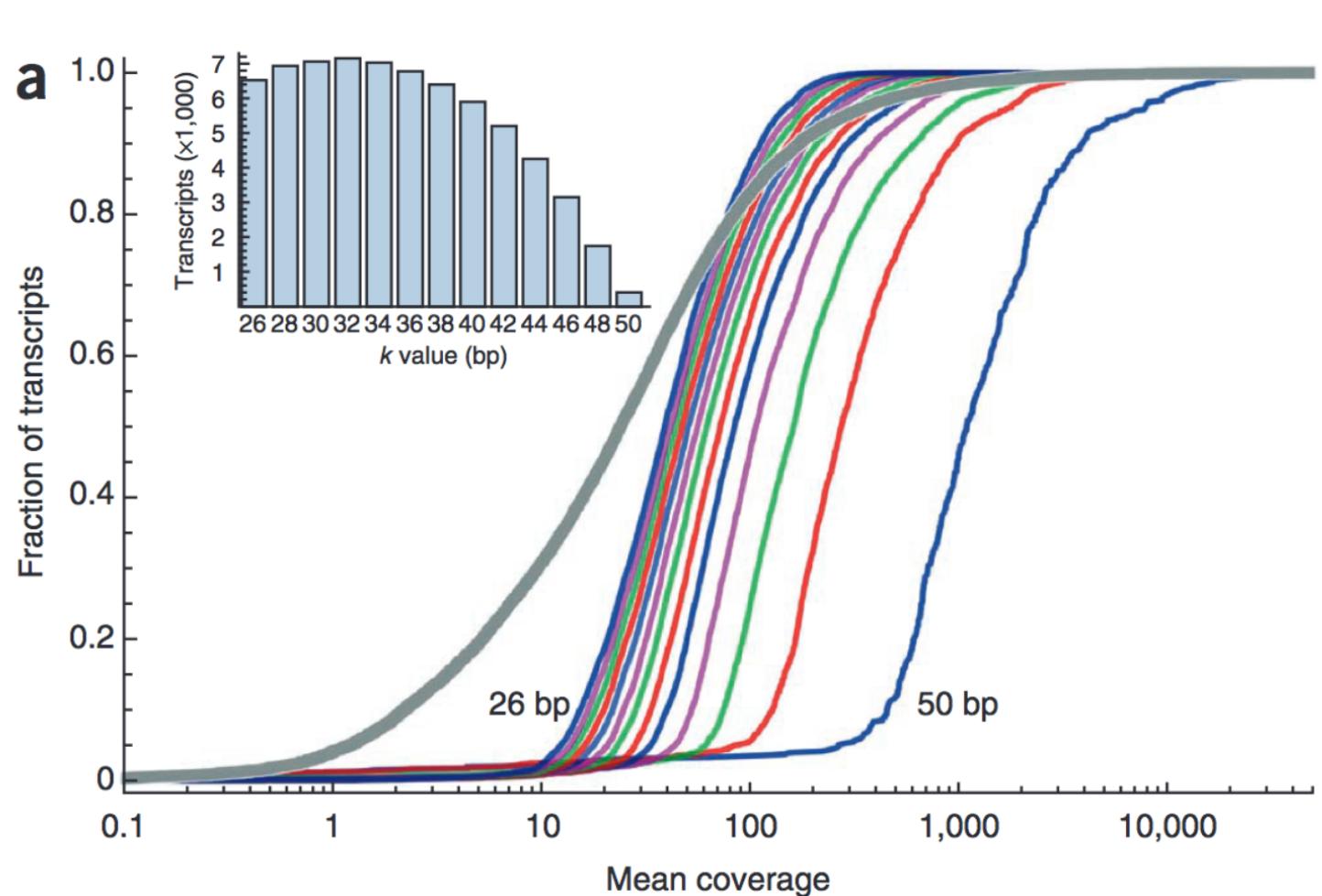
The Eastern hive honey bee, *Apis cerana cerana* is a native and widely bred honey bee species in China. Molecular biology research about this honey bee species is scarce, and





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