

Annotation of Ehux ESTs

Ehux Jamboree, June 17, 2009

Alan Kuo and Igor Grigoriev

DOE Joint Genome Institute

akuo@lbl.gov

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

Conclusions of my 2008 Jamboree talk

- **7809 scaffolds (168 Mbp)**
- **39126 gene predictions.**
- **33341 genes in Reduced Model set.**
- **Assembly and Annotation released in JGI portal, with exceptionally diverse variety of community data types.**
- **“usefulGeneLists” of expressed genes and gene families may really be useful.**

But there is a dark side ...

A challenging genome

- **Challenging genome project**
 - high GC content -> hard sequencing stops
 - high repeat content
 - diploid and polymorphic
 - source not clonal (?!)
- **Draft genomic assembly has short contigs and misassemblies, including artifactual tandem repeats and separated alleles.**

Are we missing genes because of assembly?

Are there missing genes?

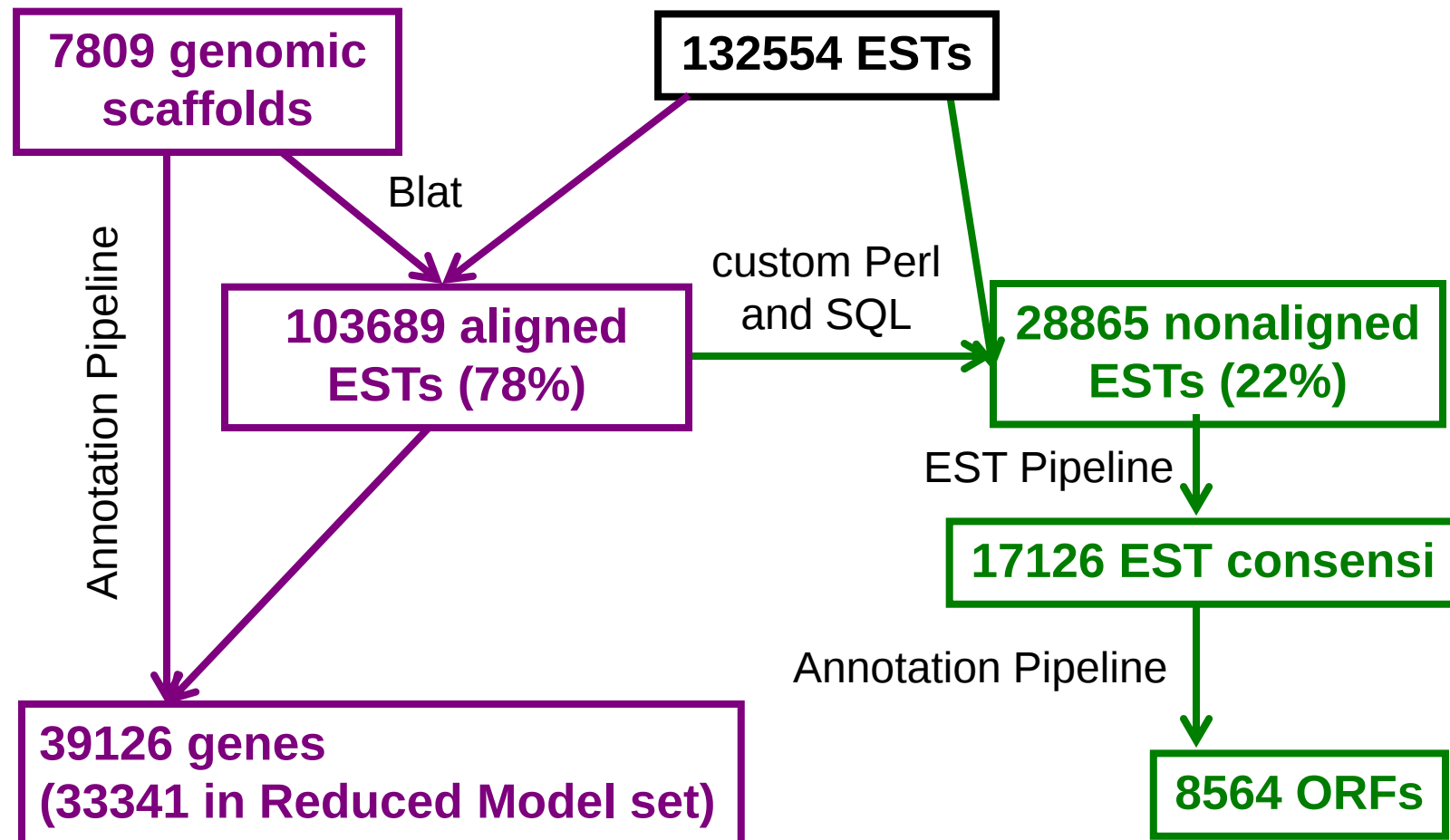
JGI genomes	Phatr1 Phaeodactylum	Naegr1 Naegleria	Chlre4 Chlamydomonas	Emihu1 Emiliana
# ESTs queried	91787	38282	202044	94168
# ESTs aligned	86553 (94%)	38010 (99%)	145544 (72%)	72992 (76%)

At least we have lots of ESTs....

Can we find missing genes?

- 1. Align all ESTs to genomic scaffolds.**
- 2. Recover all 'nonaligned' ESTs.**
- 3. Cluster ESTs and build cluster consensi.**
- 4. Predict ORFs on the EST consensi.**
- 5. Determine if ORFs are in fact new.**

Maybe we can



Consensus properties

	Genomic scaffolds	EST consensi
# consensi or scaffolds	7809	17126
Avg. length (nt)	21479	630
Min. length (nt)	1000	100
Max. length (nt)	3018814	3406
% gap space	7%	0%
% repeat space	>= 26%*	6%

* Piler results only

Do consensi complement scaffolds?

	JGI	CSUSM	Iowa	Genoscope
Strain	CCMP1516	CCMP1516	CCMP371	RCC1216/7
# ESTs	72513	7649	14006	38386
% aligned to scaffolds	82%	31%	73%	78%
% aligned to consensi	24%	61%	32%	20%
% aligned to both	9%	4%	6%	3%

ORF properties

	on scaffolds	on consensi
# ORFs or genes	33341	8564
Avg. protein length (aa)	347	148
Max. protein length (aa)	5478	984
% w/ start+stop	82%	44%
% w/ Pfam domain	35%	35%
% w/ NR support	70%	56%
% w/ EST support	47%	100%
% "bacterial"	5%	1%

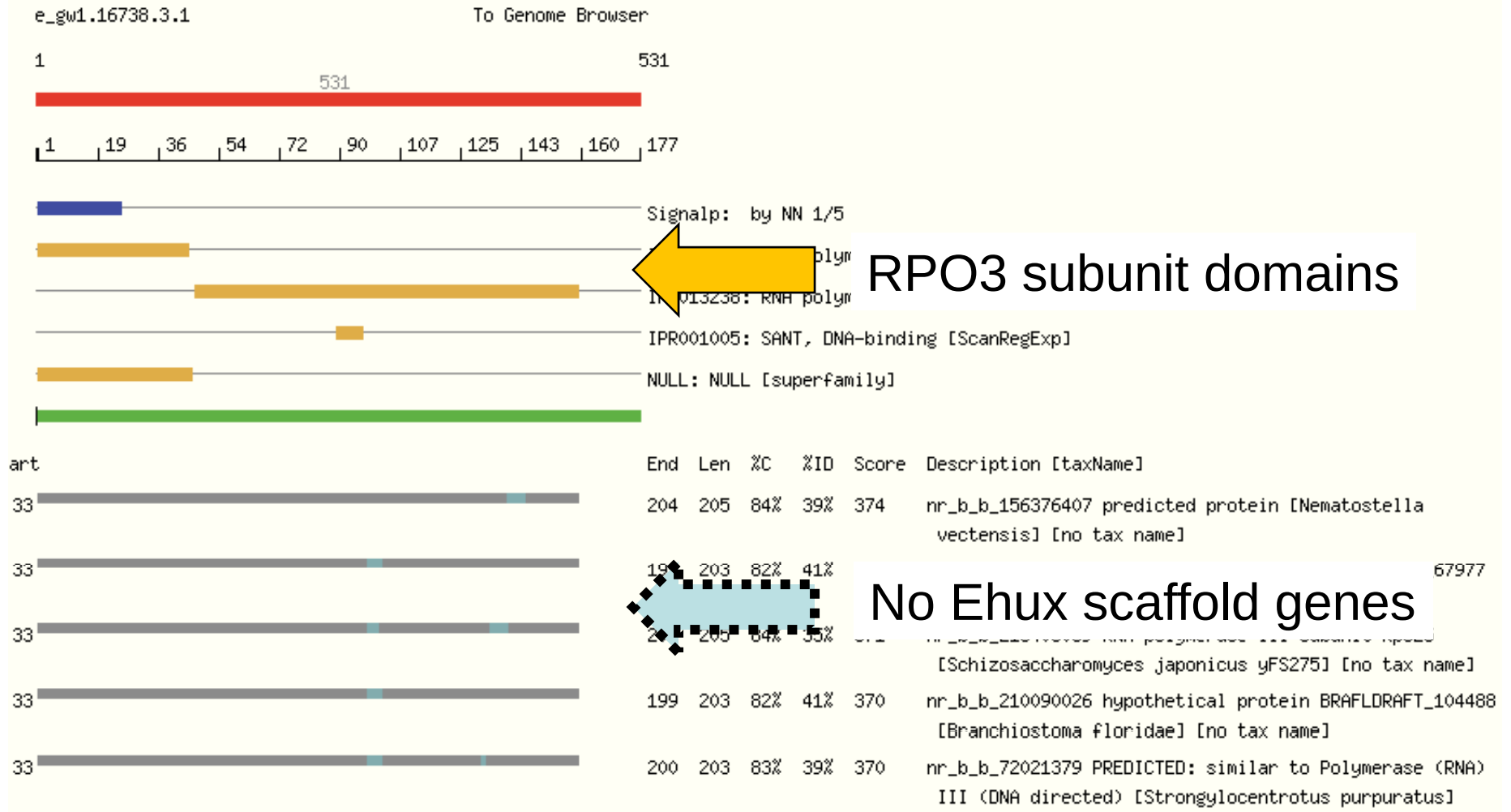
Do we find 'missing' domains?

Hmmpfam both ORFs and Reduced Model set
-> 88 ORF-specific domains

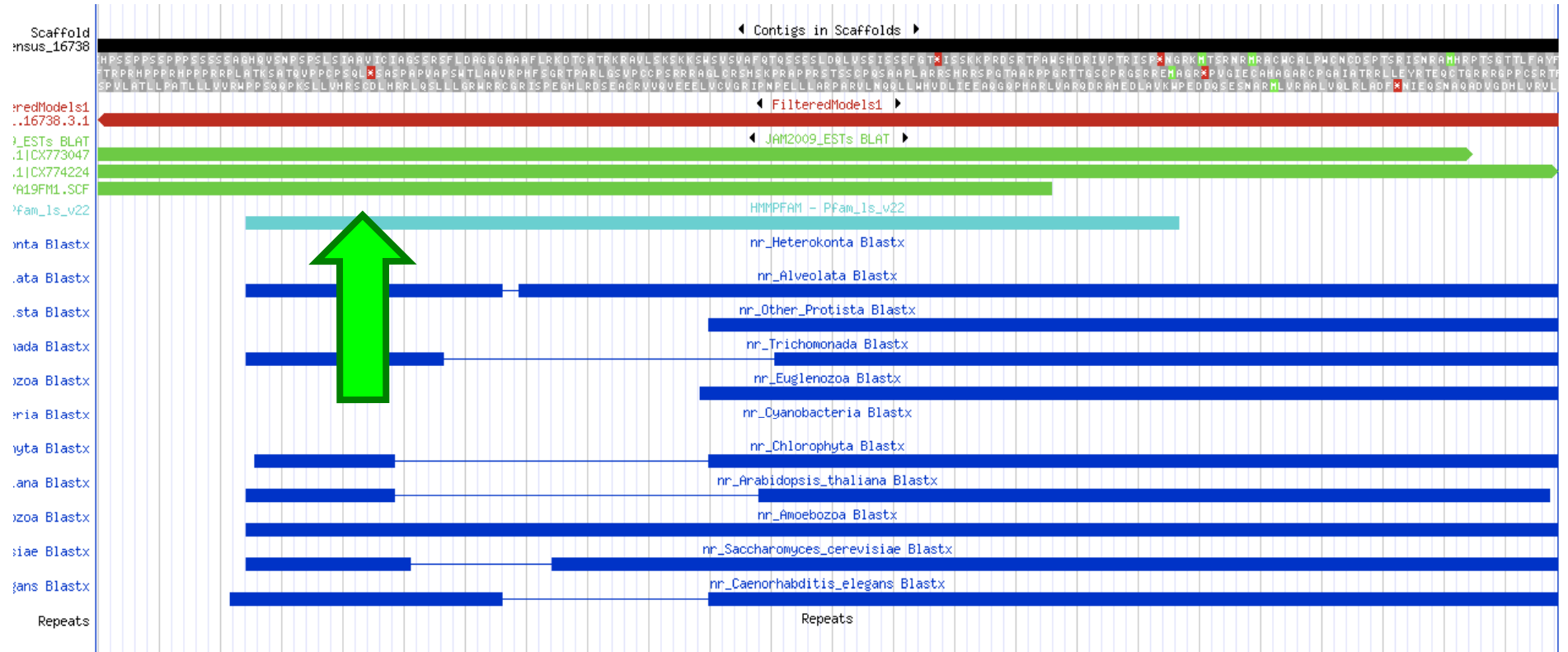
Pfam_Doma	Emihu	EmihuEXT	domain
PF08827.8	6	6	Ribosomal L15
PF08292.3	0	6	RNA polymerase III subunit Rpc25
PF09273.2	6	6	Rubisco LSMT substrate binding
PF00158.17	0	3	Sigma-54 interaction domain
PF02653.7	0	3	Branched-chain amino acid transport system / permease component
PF02772.7	0	3	S-adenosylmethionine synthetase, central domain
PF03466.11	0	3	LysR substrate binding domain
PF07715.6	0	3	TonB-dependent Receptor Plug Domain

6 ORFs have Pfam RNA_pol_Rbc25
A universal RPO3 subunit domain

RPO3 subunit protein



RPO3 subunit ORF



ESTs are from CCMP371, no other strain

Do we find 'missing' sequences?

Cluster ORFs with Reduced Model gene set

-> 2074 ORF-only clusters

-> 3135 ORFs (37%)

[Run 351](#)

cluster 27

61 proteins found

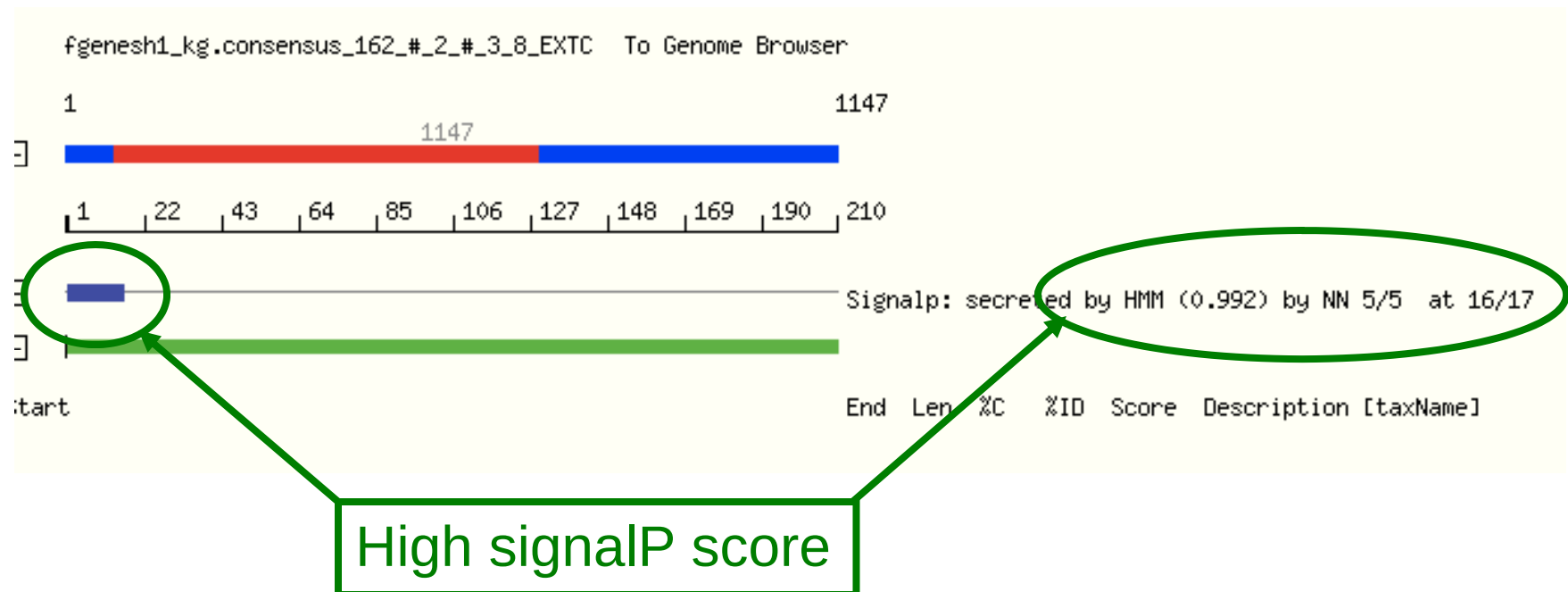
download selected proteins

Do Clustalw

Protein Id	Location	Gene length	Protein length	Number of domains	Organism	DOMAINS	SYNTENY	GENE MODEL
15463 ☐	consensus_162:1-1147	1147	210	0	E. huxleyi (Emhu1EXTC)		1147	
20779 ☐	consensus_12260:1-561	561	147	0	E. huxleyi (Emhu1EXTC)		561	
15494 ☐	consensus_179:1-556	556	146	0	E. huxleyi (Emhu1EXTC)		556	
18911 ☐	consensus_10368:1-528	528	141	0	E. huxleyi (Emhu1EXTC)		528	
19575 ☐	consensus_11021:1-525	525	139	0	E. huxleyi (Emhu1EXTC)		525	

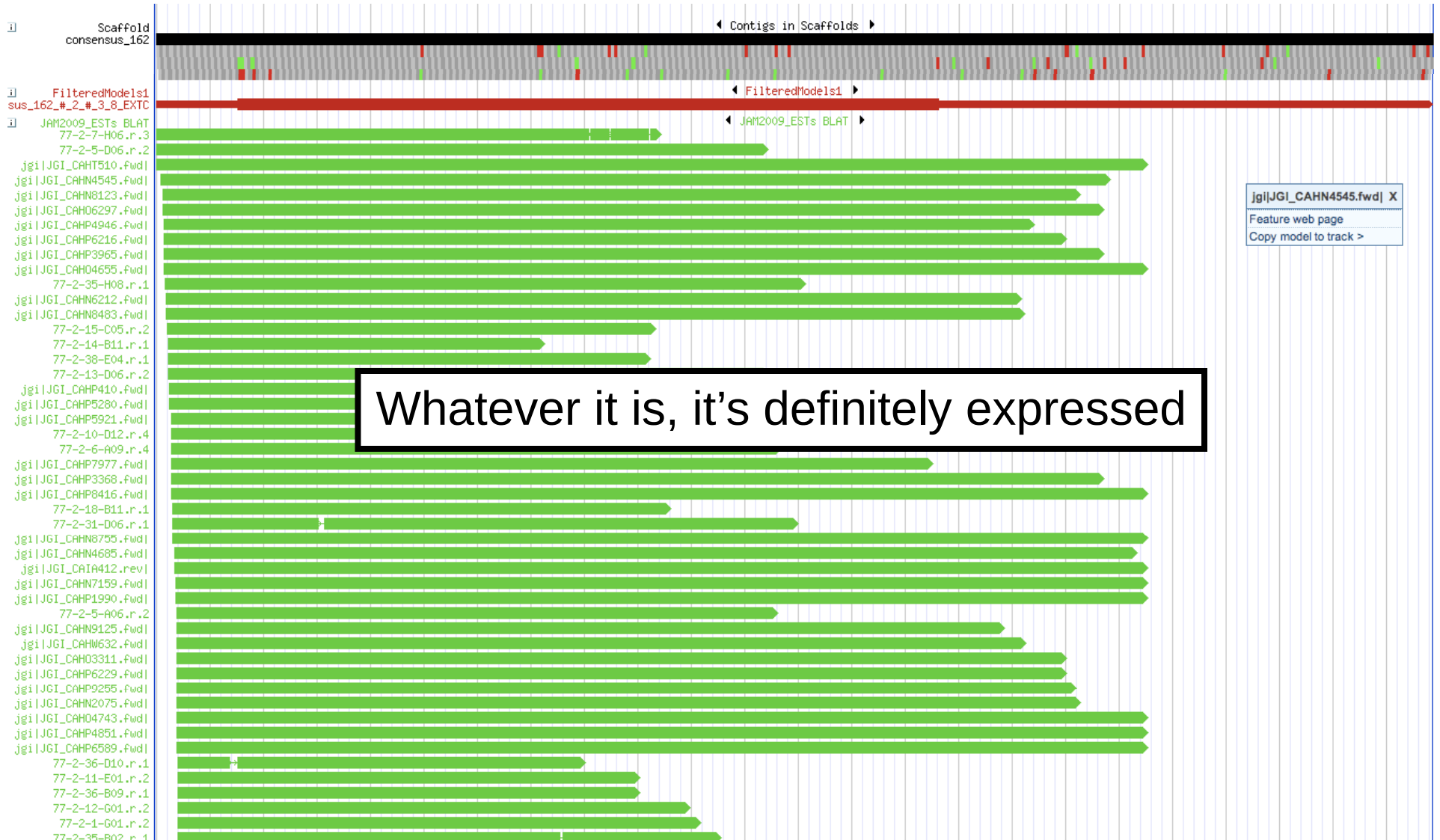
Largest protein in an ORF-only cluster

A consensi-specific protein ...



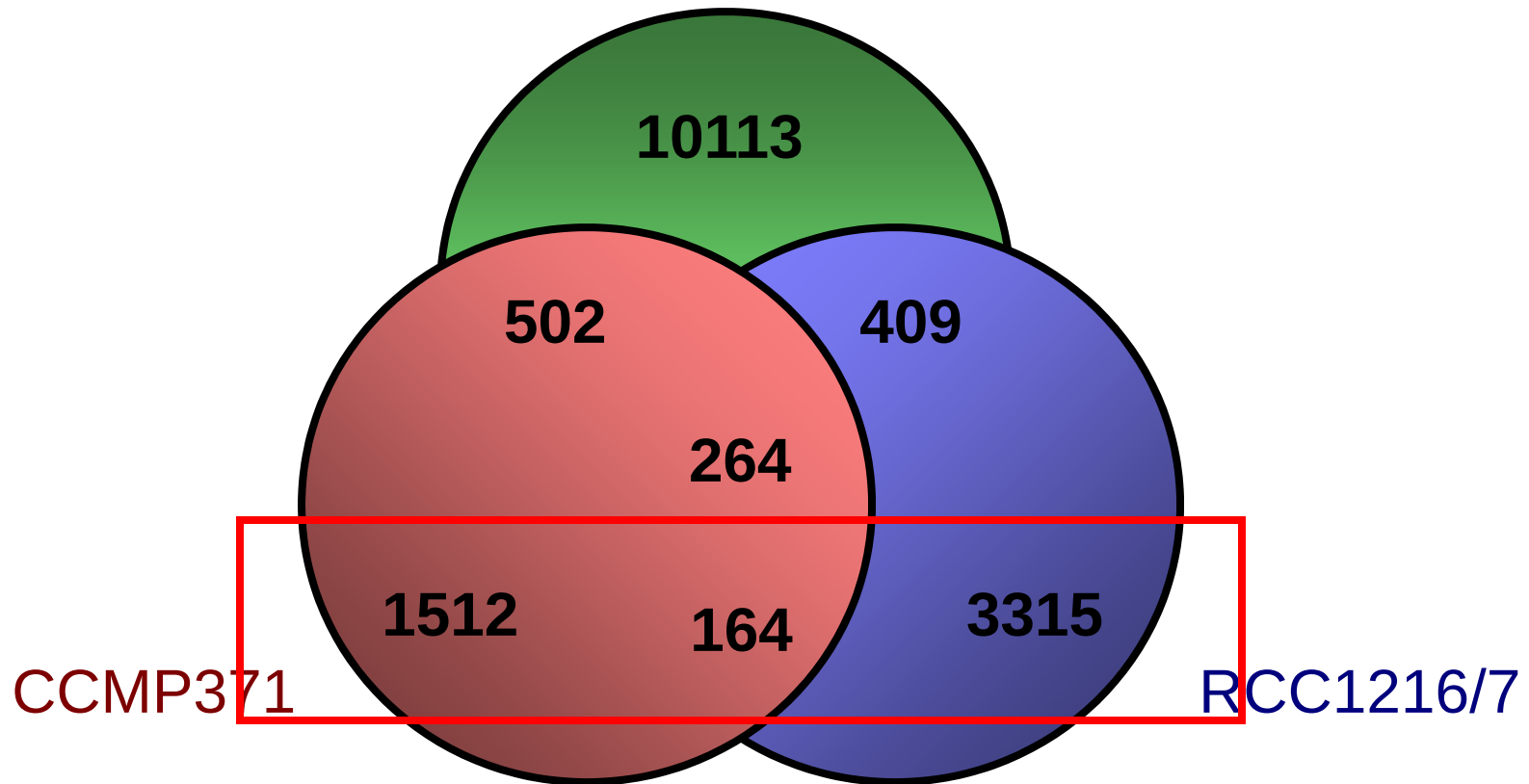
No similarity with any genomic scaffold gene

... and its expression



Strain-specific consensi???

CCMP1516



Possible non-CCMP1516 genes?

Summary

- 1. 22% ESTs do not align with scaffolds.**
- 2. EST Pipeline assembles 17126 consensi from the nonaligned ESTs.**
- 3. Annotation Pipeline predicts 8564 ORFs on the consensi.**
- 4. Domain analysis of ORFs reveals 'missing' genes.**
- 5. Cluster analysis reveals 'missing' genes.**
- 6. Expression analysis reveals potential strain-specific genes.**

Acknowledgments

CSUSM

Xiaoyu Zhang
Betsy Read

Genoscope

Peter von Dassow
Hiroyuki Ogata
Colomban de Vargas

JGI EST Pipeline

Jasmyn Pangilinan
Jeff Froula
Erika Lindquist

JGI Annotation Pipeline

Bobby Otilar
Frank Korzeniewski
Asaf Salamov
Andrea Aerts