Emiliania huxleyi v1.0*

(aka Emihu1)

Alan Kuo

Genome Annotation Group, JGI, Walnut Creek

akuo@lbl.gov

*This work was supported by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Outline

- How we annotate
- Our annotation of Ehux
- Using the Ehux annotation

The JGI way

Sequence	Assemble	Annotate
genomic DNA reads cDNA reads	genomic scaffolds ESTs/EST clusters	Genes and their functions

---

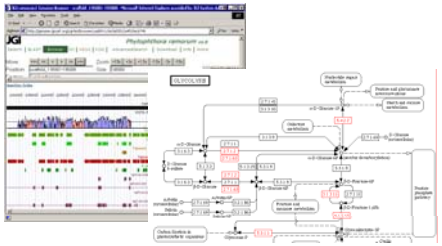
?

Discuss with community

Annotation strategy

Annotation pipeline

Automated annotation



Portal release

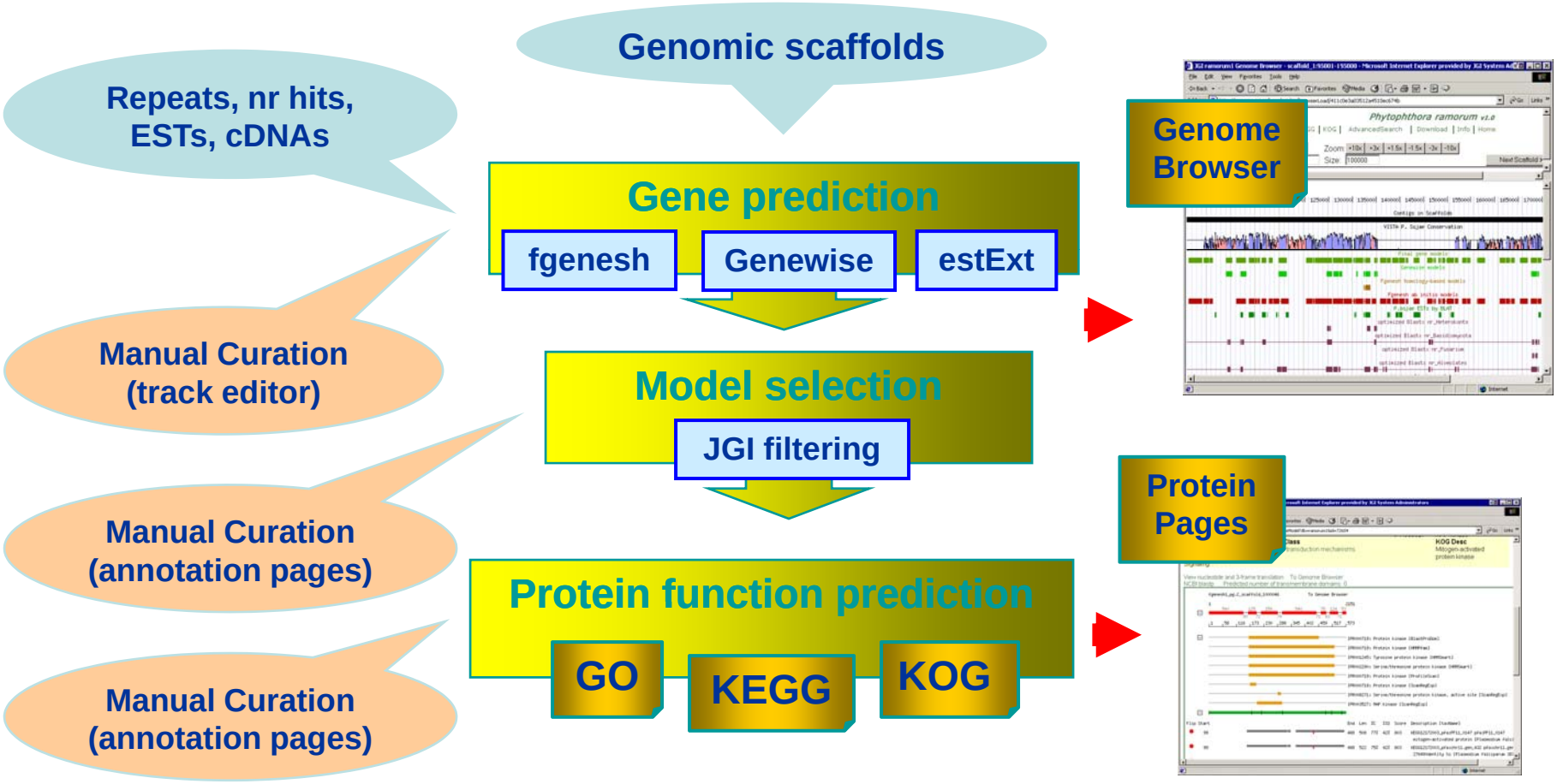
Manual curation

This jamboree

Scientific discoveries!

Adapted from I. Grigoriev

JGI annotation pipeline + portal



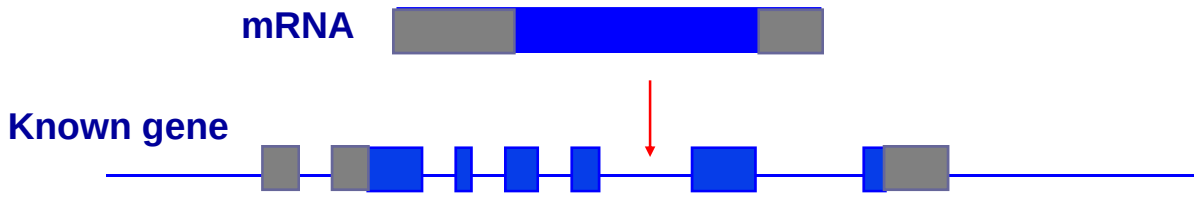
Pipeline

Portal

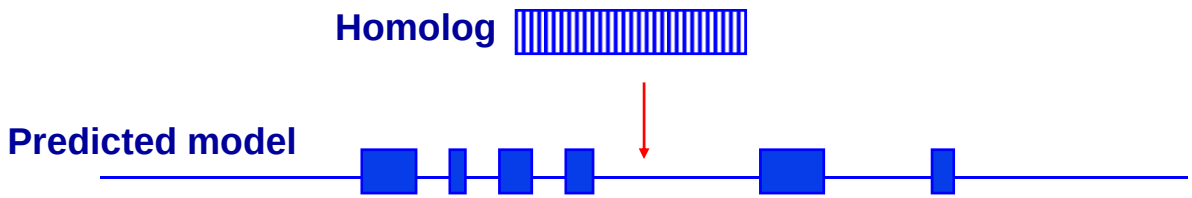
Adapted from I. Grigoriev

Structural annotation I

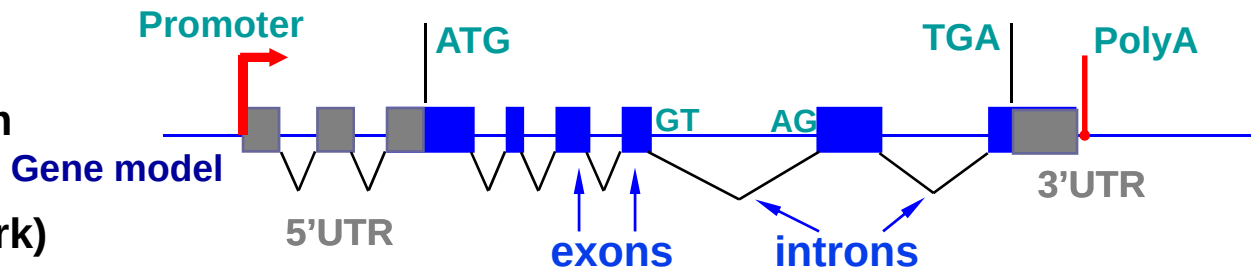
cDNA-based **methods map transcripts onto the genome (fgenesh_kg, est_orfs)**



Homology-based **methods build exons around alignments with known genes (fgenesh_pm, Genewise)**



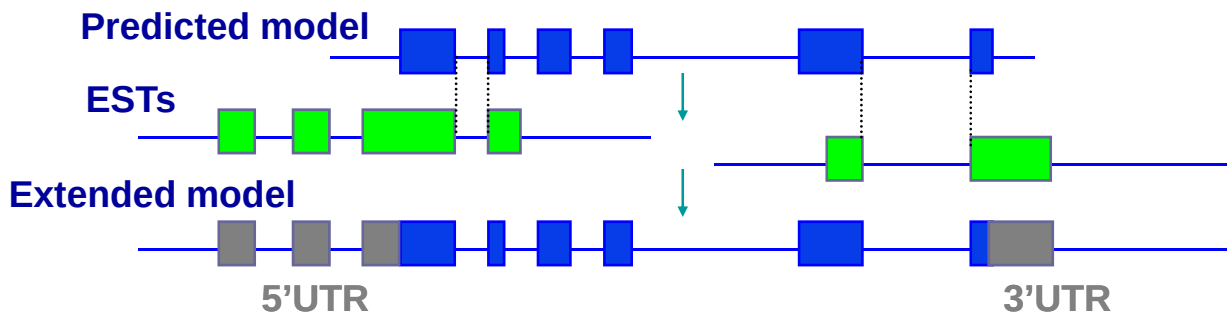
ab initio **methods use knowledge of gene structure gleaned from known genes (fgenesh_pg, Genemark)**



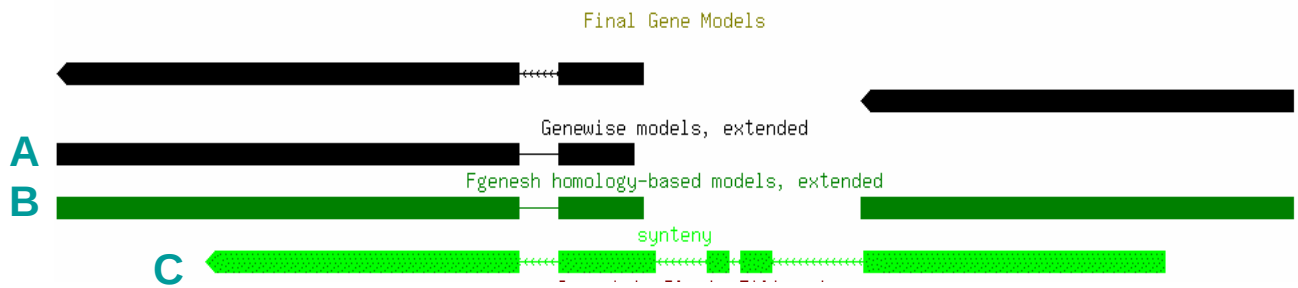
Adapted from I. Grigoriev

Structural annotation II

Use *ESTs/cDNAs* to extend and correct gene models (estExt)



***A representative set of gene models* is selected on basis of homology support, model completeness and EST support (Filtered Models)**



Adapted from I. Grigoriev

Functional annotation

Find similar proteins:

- double-affine Smith-Waterman alignments against known proteins: Swissprot, KEGG, KOG

Predict domains:

- InterProScan against manually curated domains: Pfam, ProSite, ProDom, etc

Group gene models into categories:

- Gene Ontology terms, KOG clusters of orthologous groups, EC numbers, KEGG pathways

Adapted from I. Grigoriev

Break for questions

We know how to annotate, now let's annotate

What went in…

- 7809 scaffolds (168 Mbp)
  - high GC (66%) -> hard stops -> short reads
  - highly repetitive and polymorphic diploid
- 94168 ESTs,
  - > 4 distinct conditions or timepoints
  - 41086 cluster consensi and unigenes
- 629 manually curated cDNAs
- nr: nothing related, so use:
  - all ‘protists’
  - some model organisms

Genomic assembly

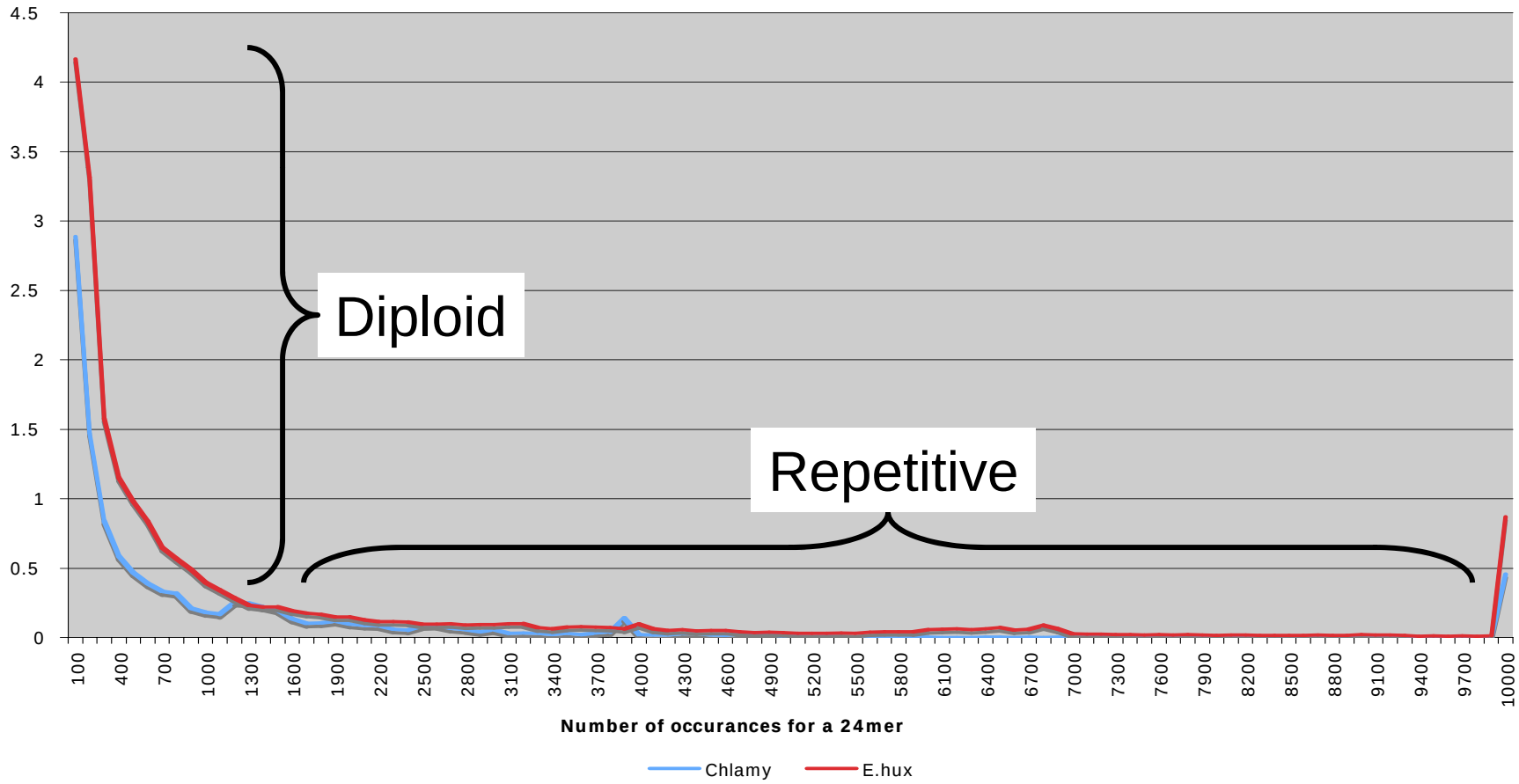
```
Main genome scaffold total: 7809
Main genome contig total:   16942
Main genome scaffold sequence total: 167.7 MB
Main genome contig sequence total:   156.0 MB
   (->  7.0% gap)
Main genome scaffold N/L50: 109/404.8 KB
Main genome contig N/L50:   1315/29.7 KB
Number of scaffolds > 50 KB: 487
% main genome in scaffolds > 50 KB: 83.9%

coverage- 10.00 x on average.

Included Reads          2,991,780  76.51%
```

Adapted from J. Schmutz

Repetitive and diploid



Adapted from J. Schmutz

…what came out

	Auran	Chlre	Emihu
Scaffolds (Mbp)	57	120	168
Repeats (Mbp)	5 (9%)	9 (8%)	17 (10%)
Genes (Mbp)	25 (43%)	63 (52%)	67 (40%)
# genes	11501	15256	39126
# genes / Mbp	203	127	233

How were the genes predicted?

method	Auran	Chlre	Emihu
cDNA	230 (2%)	4910 (32%)	7792 (20%)
homology	5622 (49%)	3974 (26%)	4973 (13%)
ab initio	5649 (49%)	6372 (42%)	26361 (67%)

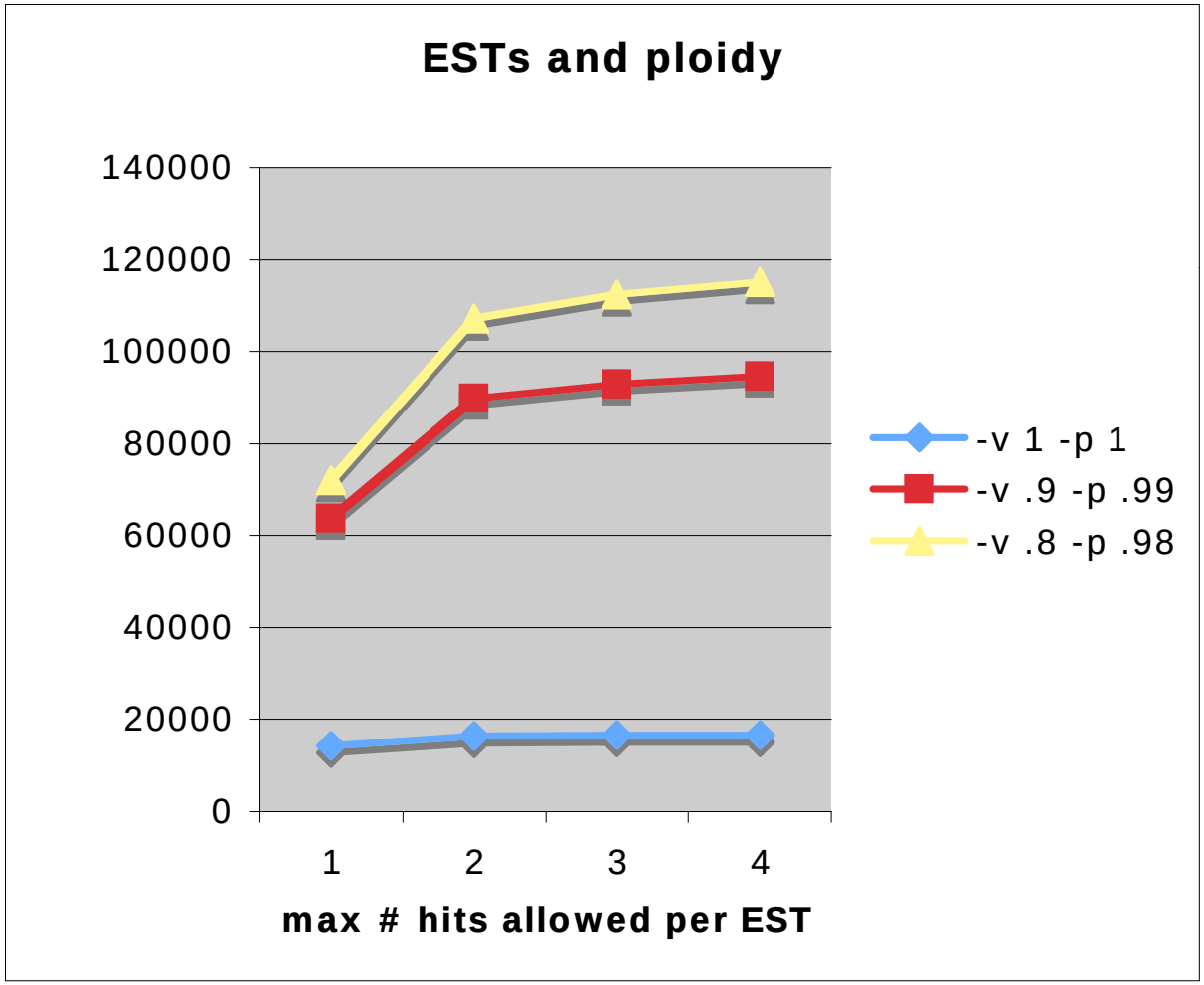
Are the predictions any good?

metric	Auran	Chlre	Emihu
complete	7243 (63%)	8919 (58%)	32174 (82%)
ESTs	2435 (21%)	7894 (51%)	18259 (47%)
nr	9232 (80%)	8845 (56%)	27402 (70%)
Pfam	6427 (56%)	7143 (47%)	13620 (35%)

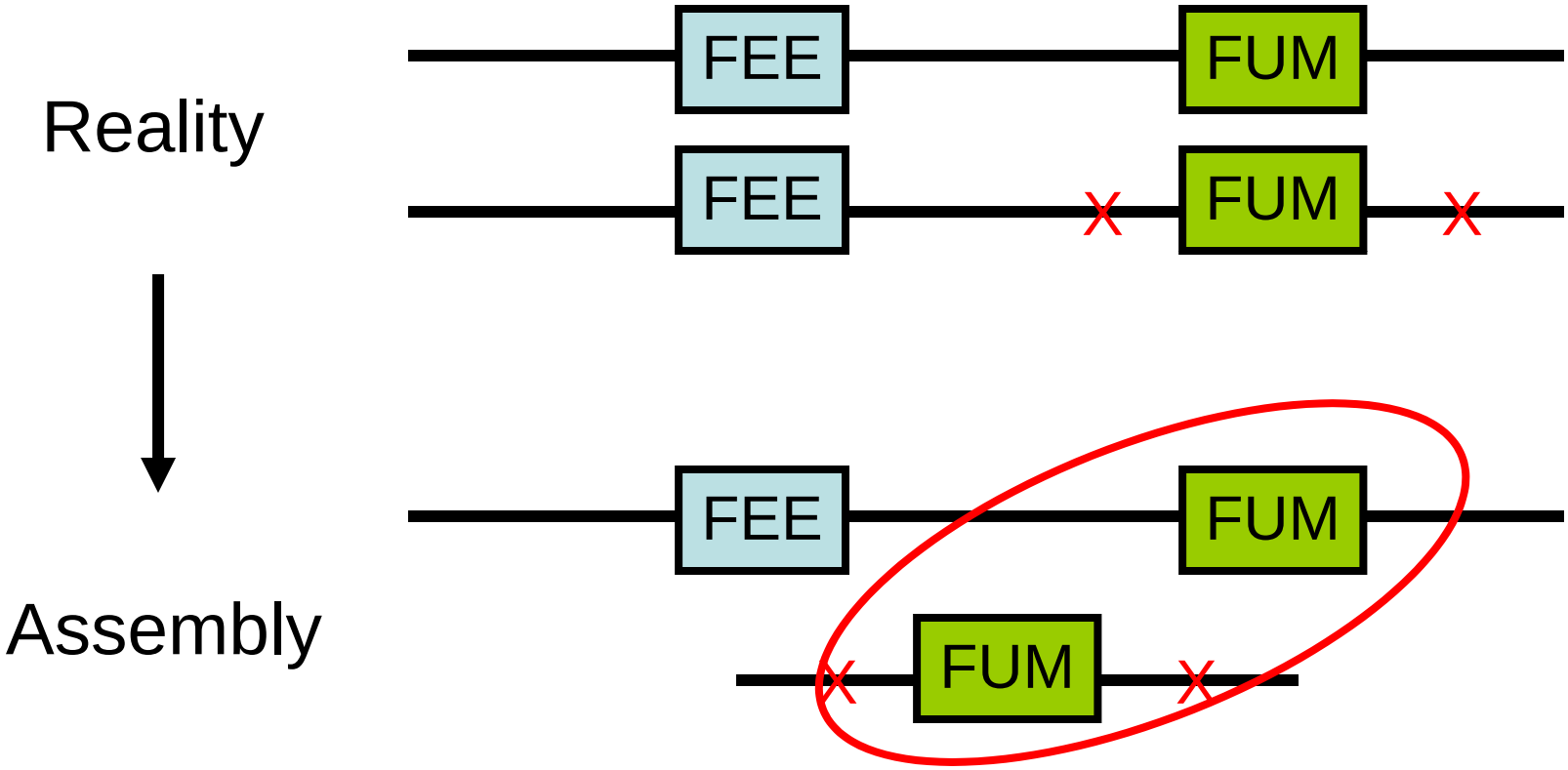
What do the genes look like?

Average values	Auran	Chlre	Emihu
Gene length (nt)	2138	4096	1719
Transcript length (nt)	1601	1481	1112
Protein length (aa)	523	420	341
Exon length (nt)	694	188	307
Intron length (nt)	412	382	238
Exons / gene	2.31	7.90	3.60

Assembly is not haploid



Separation of alleles

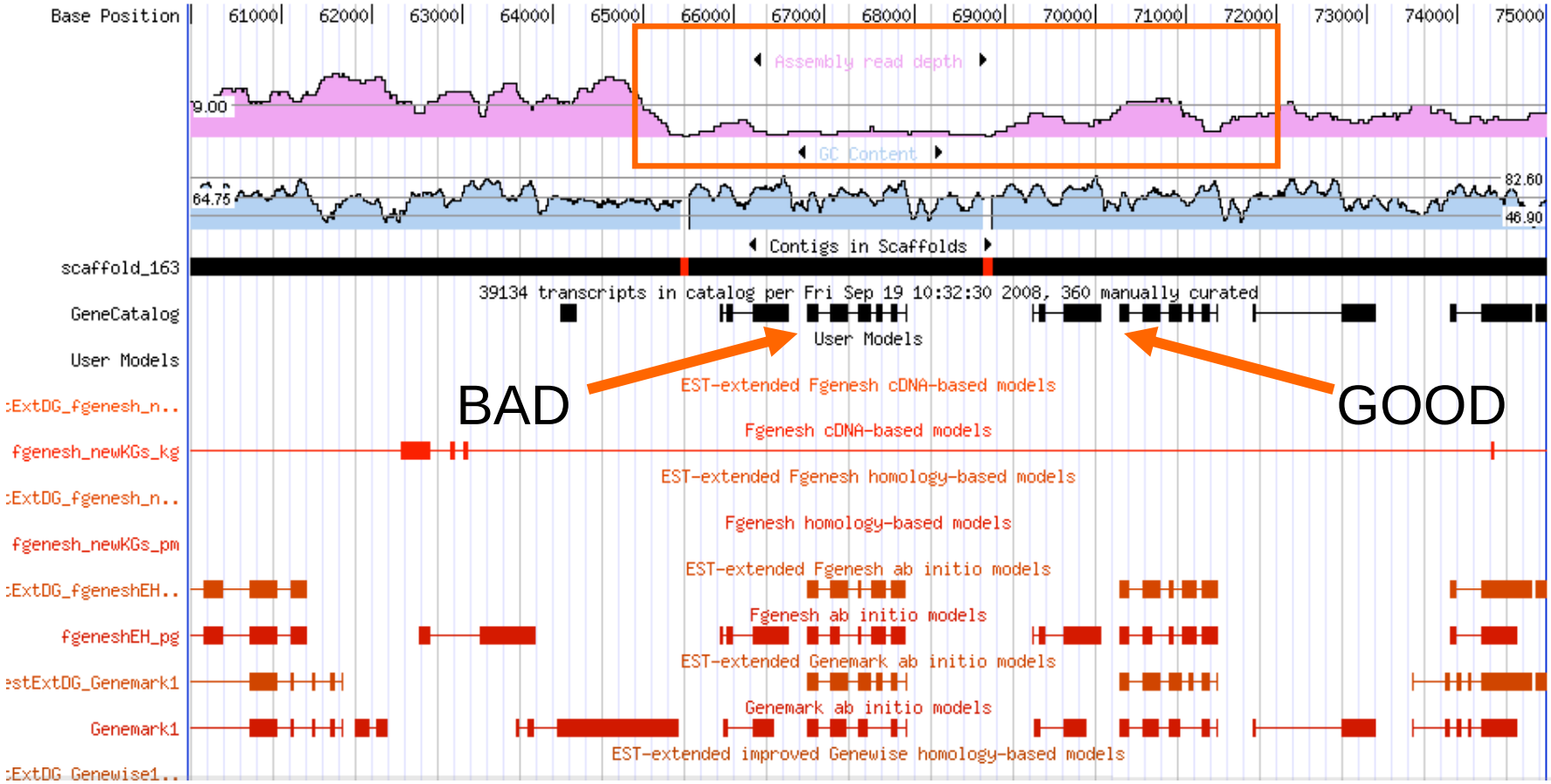


2931 pairs (or 5862 candidate alleles)

Reduced gene set

	Not reduced	Reduced
# genes	39126	33341 (⇓15%)
# clusters	20875	20542 (⇓2%)
# clusters w/ 1 gene (singlets)	11484	14917 (⇑30%)
# clusters w/ 2 genes (doublets)	7427	4258 (⇓43%)
# clusters w/ > 2 genes (multis)	1965	1367 (⇓30%)
# genes in multis	12788	9908 (⇓23%)

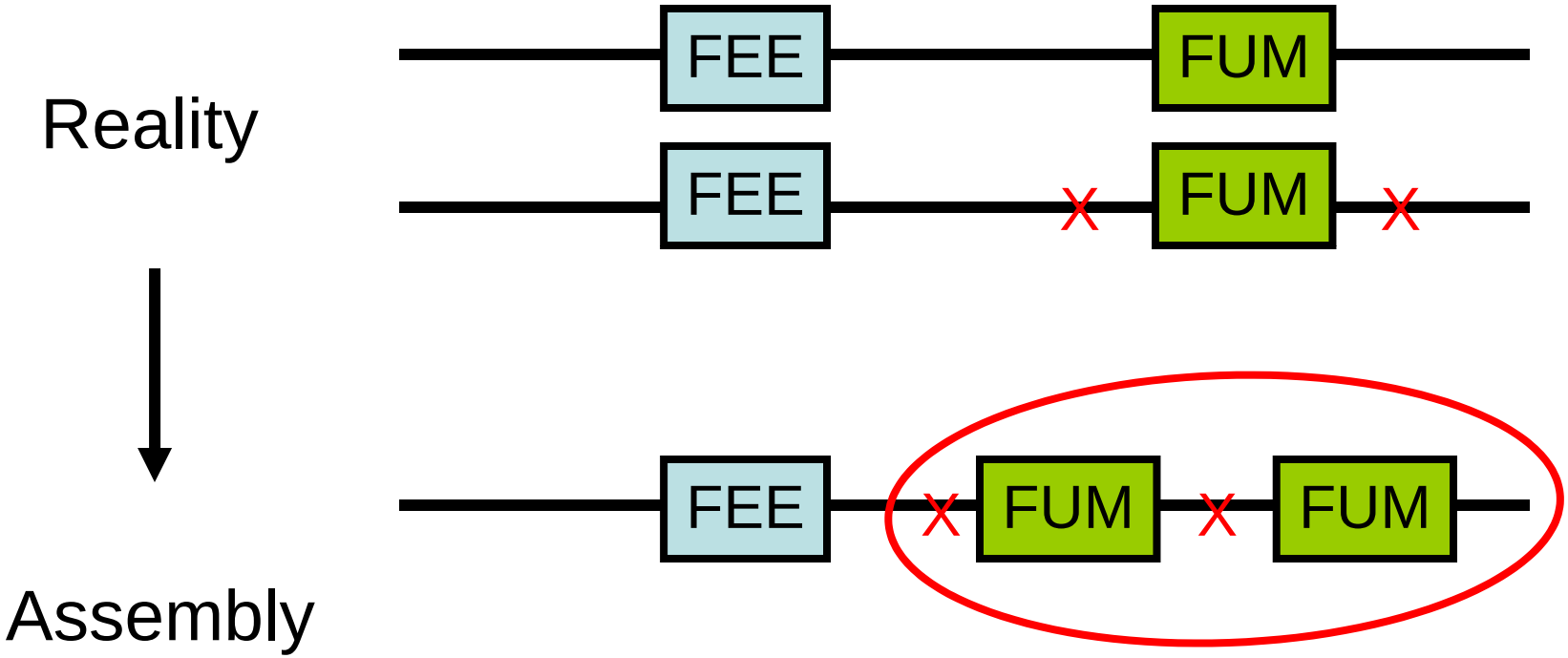
Artifactual tandem repeat



"The bad copies have severe low coverage and are bounded by gaps"

Adapted from J. Schmutz

Tandem duplication



40 pairs (or 80 candidate duplicates)

Break for questions

We have genes, now let's do something with them.

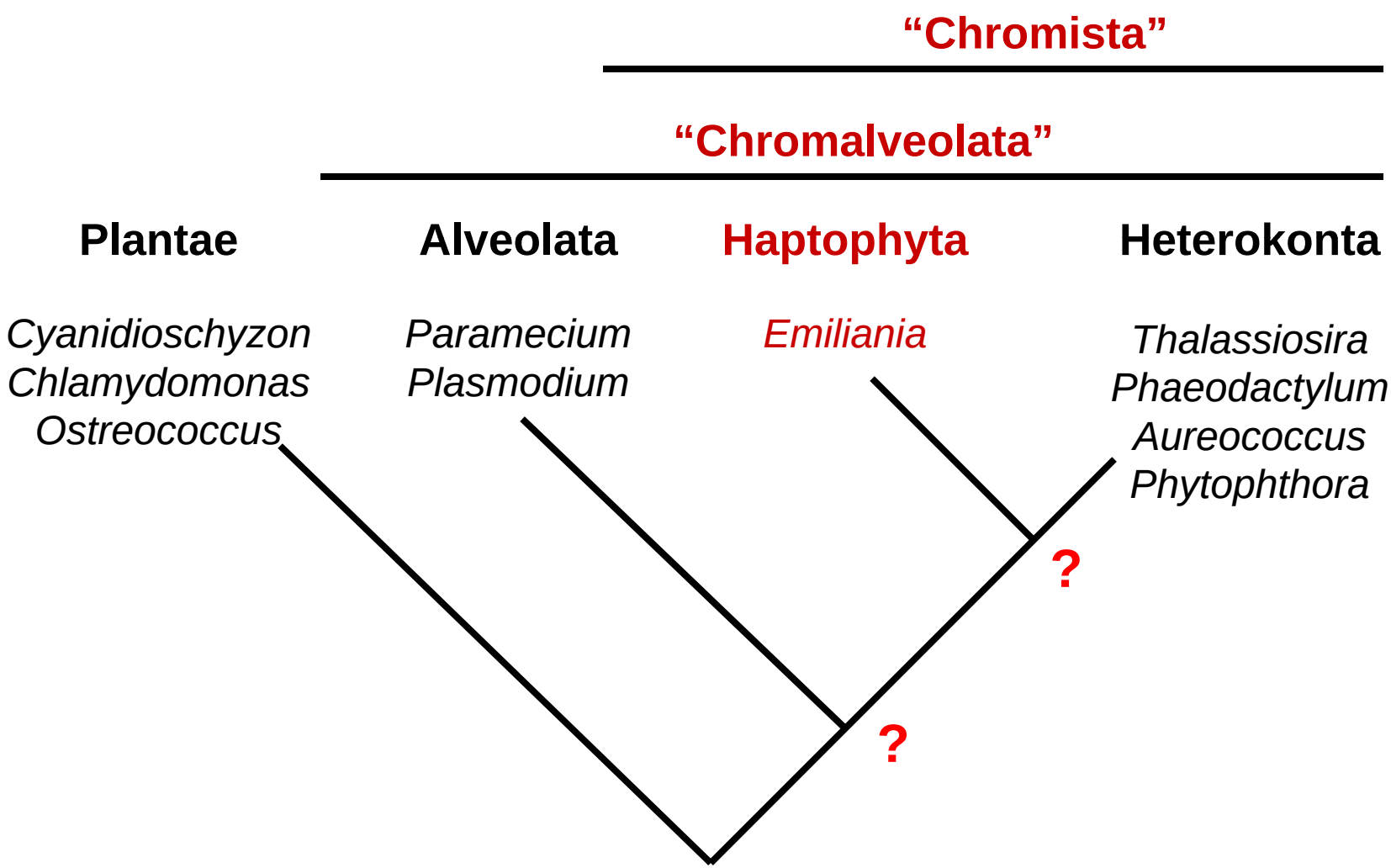
Community data

- manual curation of Gene Catalog
- 35899 genes predicted by EuGene
- 51660 EST-derived Nimblegen probes
- 719059 sRNAs
- library of 86 manually curated TEs
- library of 889 PILER *de novo* repeats
- 188807 PHOBOS tandem repeats
- 139632 SNPs from 6 strains
- still more on the way …

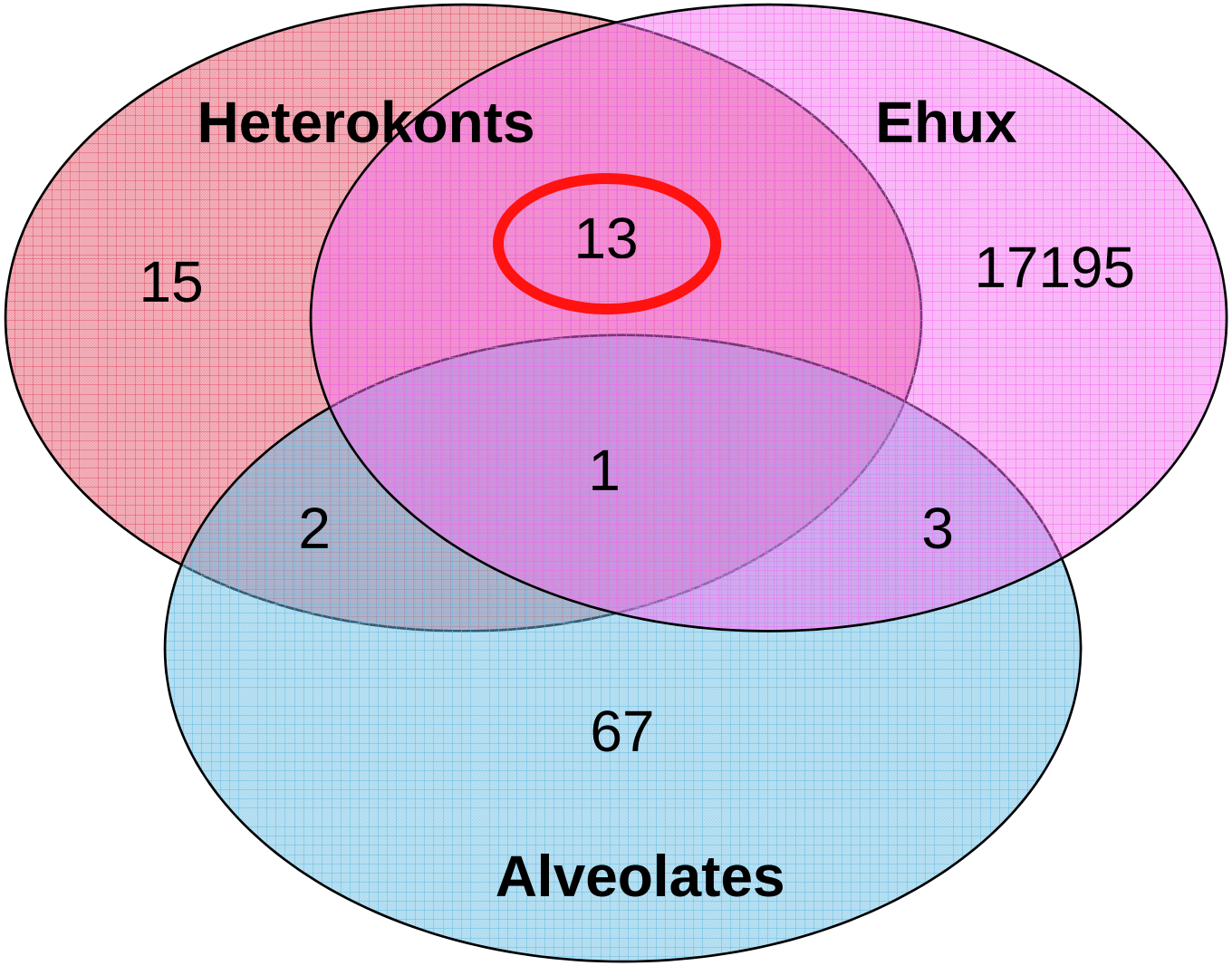
Useful gene lists

- Gene families
- Protein domains
- Conditionally expressed genes
- Clusters with 11 other proteomes

Chromalveolate hypothesis



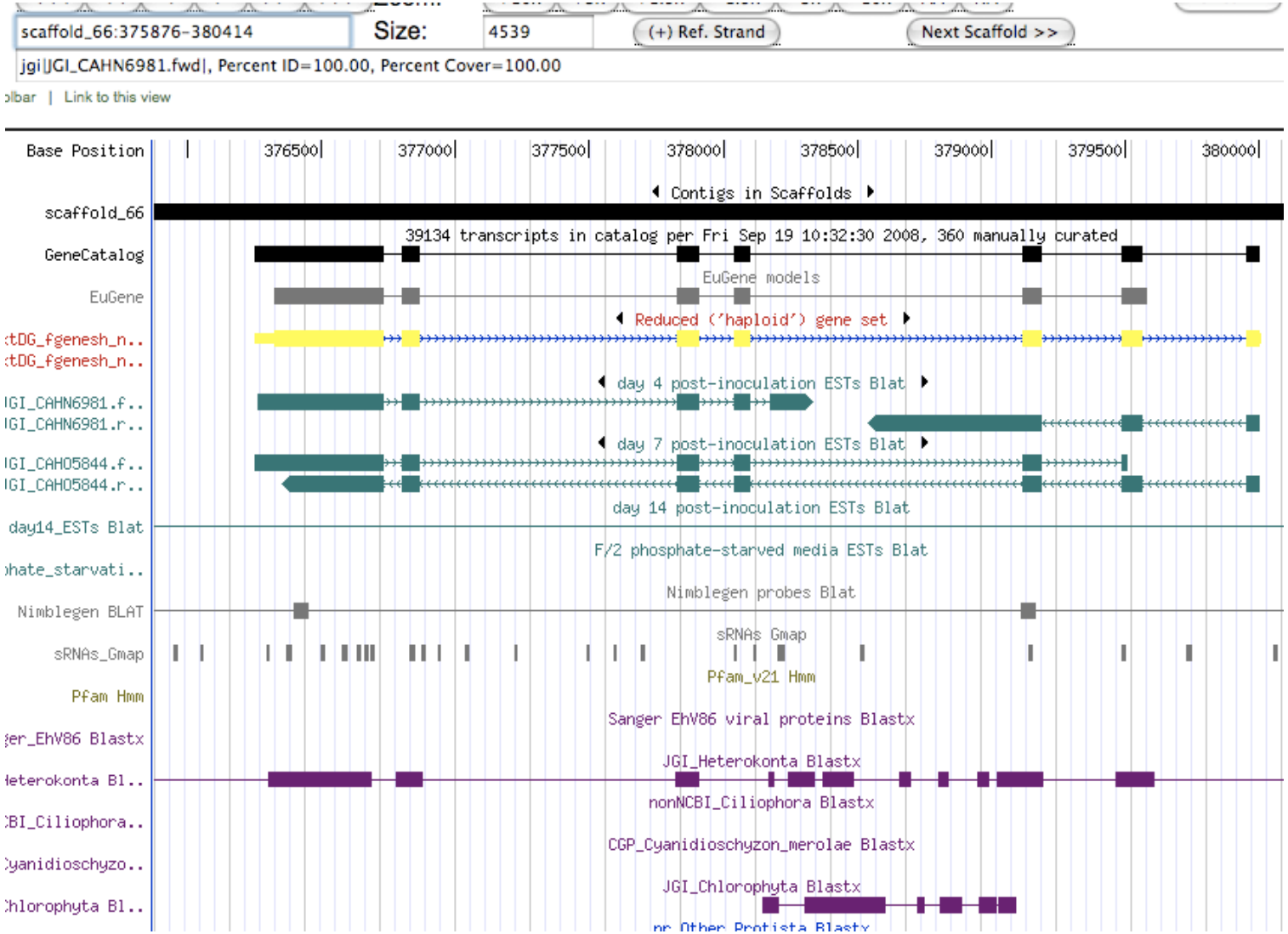
Taxon-specific families



"chromist"-specific families

FamilyId	# genes	# Ehux	
430	23	4	aa-transporter?
1219	**12**	**5**	**cobyrinic acid a,c-diamide synthase**
1395	11	2	exo-1,3-beta-glucanase
1567	10	5	??

Family 1219, gene view

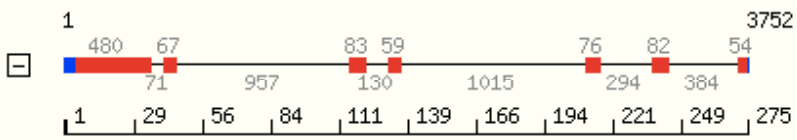


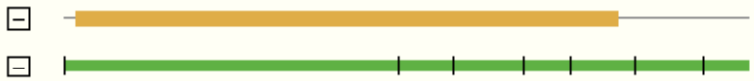
Family 1219, protein view

ASPECT	GO Id	GO Desc	Interpro Id	Interpro Desc
Molecular Function	0042242	cobyrinic acid a,c-diamide synthase activity	IPR002586	Cobyrinic acid a,c-diamide
Biological Process	0009236	vitamin B12 biosynthesis	IPR002586	Cobyrinic acid a,c-diamide

View/modify manual annotation
View nucleotide and 3-frame translation To Genome Browser
NCBI blastp Predicted number of transmembrane domains: 0

estExtDG_fgenesh_newKGs_kg.C_660042 To Genome Browser

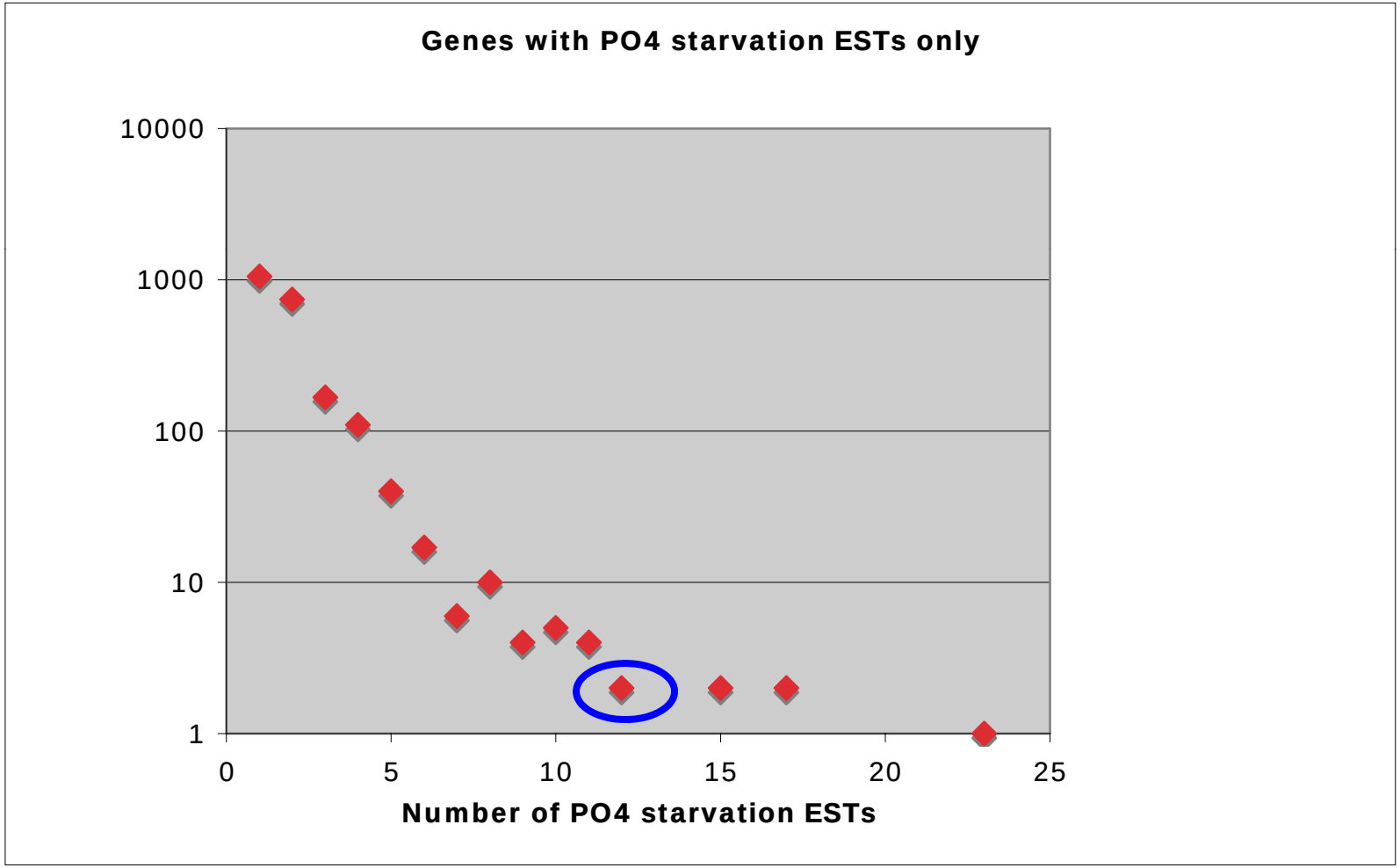




Flip	Start	End	Len	%C	%ID	Score	Description [taxName]
	3	187	364	51%	40%	319	nr_b_b_149196214 cobyrinic acid a,c-diamide synthase family protein [Lentisphaera araneosa [no tax name]
	5	198	364	53%	31%	313	nr_b_b_145620349 cobyrinic acid a,c-diamide synthase family protein [Geobacter bemidjiensis [no tax name]

Cobyrinic acid a,c-diamide synthase

Expression analysis



PO_4 starvation-specific genes

# genes	# ESTs	
1	23	HSP20
2	17	Pfam: proteinase inhibitor I15, antistasin
2	15	cysteinyl-tRNA synthetase
1	**12**	**???**
1	12	Pfam: AAA ATPase

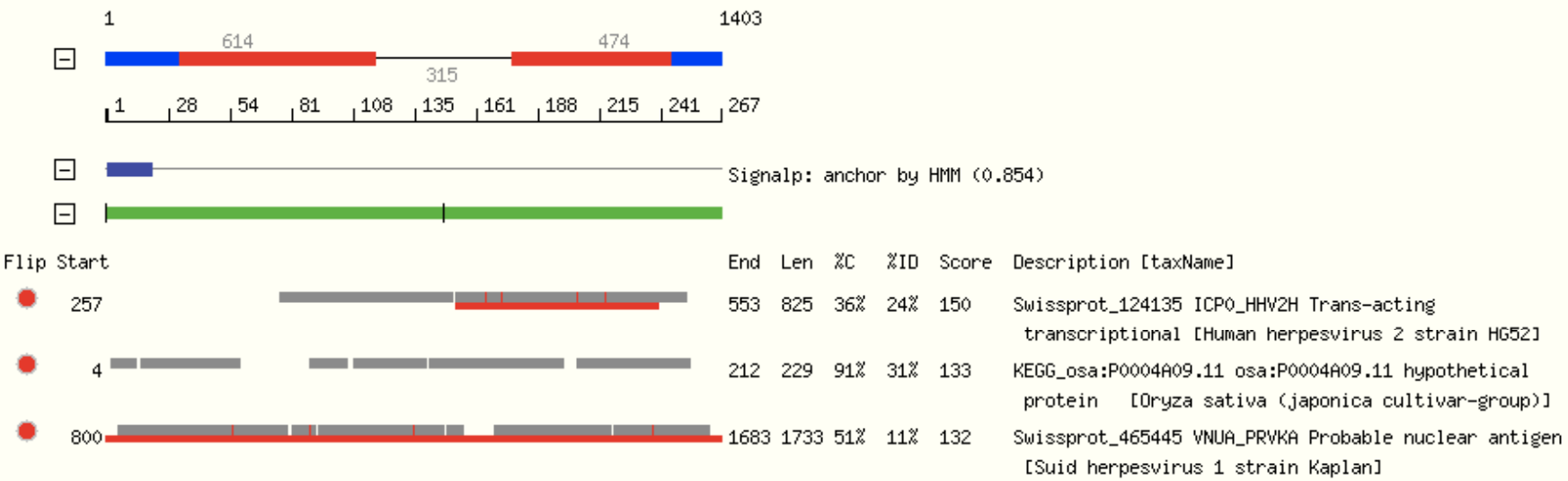
A real starvation-specific gene

A new, Ehux-specific gene

KOG GROUP	KOG Id	KOG Class	KOG Desc
Poorly Characterized	KOG0825	General function prediction only	PHD Zn-finger protein

View/modify manual annotation
View nucleotide and 3-frame translation To Genome Browser
NCBI blastp Predicted number of transmembrane domains: 0

estExtDG_fgenesh_newKGs_pm.C_4440001 To Genome Browser



Results and summary

- 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" may really be useful.
  - 2166 genes with PO_4 starvation ESTs only.
  - clustering with 11 other genomes yields 17195 Ehux-only and 13 'chromist'-only families.

Credits

CSU San Marcos

Betsy Read

Xiaoyu Zhang

Other JGI folks

Jeremy Schmutz

Erika Lindquist

Pete Brokstein

Alex Poliakov

Lukasz Szajkowski

JGI annotation folks

Asaf Salamov

Frank Korzeniewski

Bobby Otillar

Xueling Zhao

Igor Grigoriev

Georgia Tech

Mark Borodovsky

Repetitive genome

	Mbp
Genomic assembly size	168
JGI “Draft Repeats”	7 (4%)
Florian curated TEs	13 (8%)
Christoph PILER repeats	44 (26%)
Christoph PHOBOS repeats	12 (7%)
Genes (Mbp)	67 (40%)

How to find separated alleles

Identify potential allelic pairs.

- Self-Blast FilteredModels set.
- Select excellent hits (identity > 95%, e < 10^{-25}).
- Identical length and exon number.
- If neighbors exist, demand that at least 1 neighbor be also paired with the allele's neighbor.

5862 candidate alleles (or 2931 pairs)

Remove 1 gene of each allelic pair.

- Keep gene on larger scaffold.

33341 genes in Reduced Model (RM) set

How to find tandem duplicates

Identify potential tandem duplicates.

- **Self-Blast Reduced Models set.**
- **Select excellent hits (identity > 95%, e < 10^{-25}).**
- **Identical length and exon number.**
- **Look for adjacent duplicates, up to 1 removed.**

40 candidate artifactual tandem duplicates (80 genes)

How to compare

1. **Self-align all 12 proteomes – [Blastp]**
   - **177463 proteins**
   - **coverage > 90%, e-value < 10^{-5}**
2. **Convert into symmetric similarity matrix**
3. **Cluster into families – [MCL]**
4. **Annotate families with Pfam – [Hmmer]**
5. **Classify families by taxa**
6. **Look for interesting stuff**