Genome annotation: Bioinformatics for high-throughput genomics and beyond

Frank Larimer

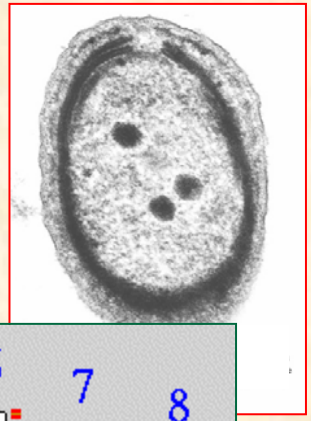
Oak Ridge National Laboratory

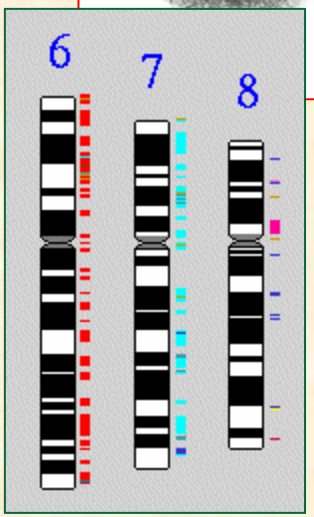
9th DOE Contractor and Grantee Workshop

Oakland, CA, 28 January 2002

Comparative Genomics Resource

- Microbial
  - JGI & public finished genomes (~80)
  - JGI & Collaborative draft genomes (~25)
- Human
  - Finished NT sequences
  - Draft Genbank entries
  - Draft Human Genome Builds
  - JGI chromosomes - 5 - 16 - 19
- Mouse and model organisms
  - Finished mouse clones / NTs
  - Mouse draft as available
  - JGI model organisms

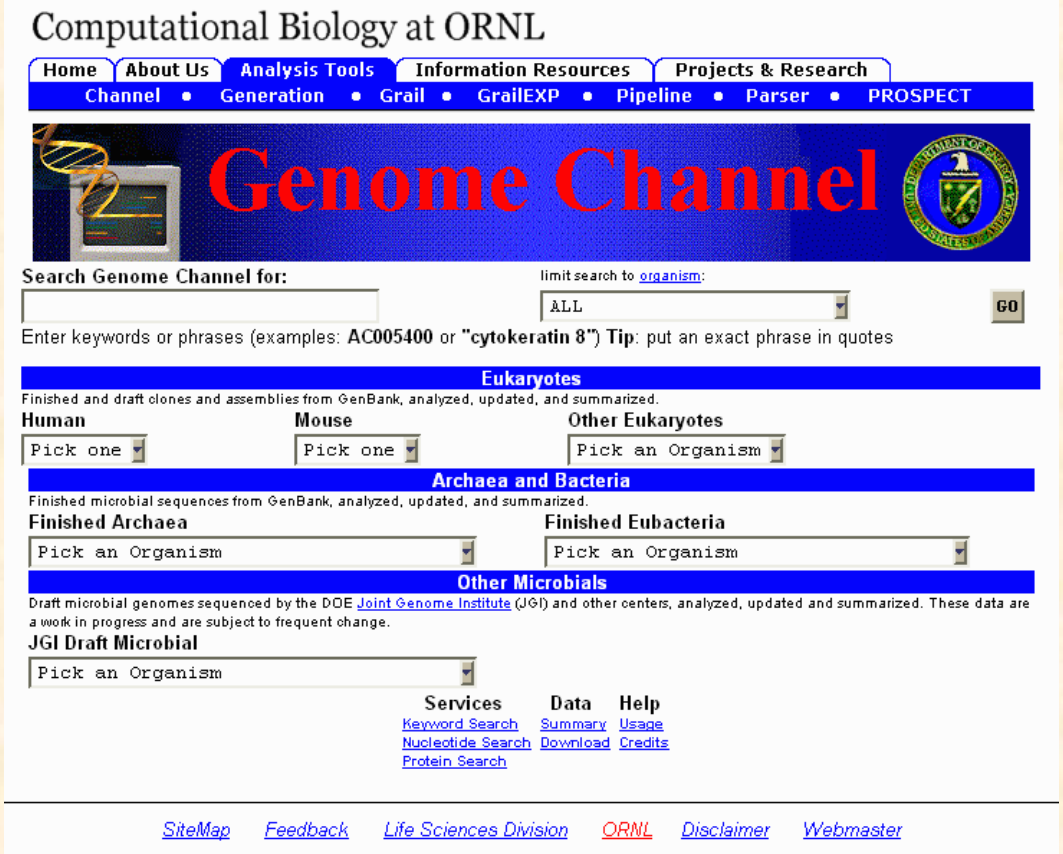




Community Tools and Services

- GRAIL EXP **eukaryotic genome analysis**
- Microbial Analysis **Generation, Critica, Glimmer**
- HPC Blast, Pfam, InterPro & other Tools
- Genome Analysis Pipeline **human and microbial**
- PROSPECT **protein structure prediction server**
- SRS, GDB, Swissprot, HOBACGEN **database servers**
- Genome Channel / Catalog **eukaryotic and microbial genomes / draft chromosomes**

Web access to annotation



Web access to annotation

Computational Biology at ORNL

Home | About Us
Channel • G

Search Genome Chann

Enter keywords or phrase

Finished and draft clones and .

Human

Pick one

Finished microbial sequences

Finished Archaea

Pick an Organism

Draft microbial genomes sequ
a work in progress and are sub

JGI Draft Microbial

Pick an Organism

SiteMap

Genome Channel

Computational Biology at ORNL

Home | About Us | Analysis Tools | Information Resources | Projects & Research

Channel • Generation • Grail • GrailEXP • Pipeline • Parser • PROSPECT

Search This Organism for:

Enter keywords or phrases (examples: **AC005400** or **"cytokeratin 8"**) **Tip**: put an exact phrase in quotes

GO

Human Data Sets

HTML or ASCII.
Many links may take a long time to finish.

Contigs	HA
Clones	HA
GenBank CDSs	HA
GrailEXP Gene Models	HA
GrailEXP PolyAs	HA
Genscan Gene Models	HA
Genscan PolyAs	HA
tRNA Genes	HA
STS	HA
Repeats	HA
CpGs	HA

Genome Channel

Human Overview

GO to another organism | Pick an Organism

Links

- Chromosomes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 Unknown X Y

GenBank Contigs (Accession Number format NT_xxxxxx)

- March 2001 build of contigs.
- Estimated 4% coverage using **FINISHED** and **DRAFT** data
- 29531 genomic clone sequences (4428.18 Mb total)
- Assembled into 583 contigs (120.83 Mb assembled, nonoverlapping)
- 11883 GenBank gene models
- 74297 Genscan gene models
- 92197 GRAIL-EXP gene models of which 48332 genes were aligned with one or more ESTs

Web access to annotation

Computational Biology at ORNL

Home | About Us | Analysis Tools | Information Resources | Projects & Research

Channel • Generation • Grail • GrailEXP • Pipeline • Parser • PROSPECT

Genome Channel / Human / Chromosome 19

Chromosome Overview

GO to another organism: Pick an Organism

GO to another chromosome: Pick a Chr

Summary

59 Contigs (6.79 MB), 794 Clones (75.97 MB), 755 GenBank CDSs, 2284 Genscan Genes (1590 with BLAST hits), 2998 GrailEXP Genes (1711 with BLAST hits and 2376 with EST evidence)

Links to Contigs

Contig	Begin Position	End Position	Length	Band	Color
NT_025171.1	1	43299	43299	p13.3	
NT_025146.2	93300	216761	123462	p13.3	
NT_011287.1	266762	339739	72978	p13.3	
NT_011269.3	389740	889240	499501	p13.3	
NT_025196.2	939241	1079045	139805	p13.3	
NT_011284.2	1129046	1255132	126087	p13.3	
NT_011227.3	1305133	1532235	227103	p13.3	
NT_011286.1	1582236	1655184	72949	p13.3	

Search This Organism for:

Enter keywords or phrases (examples: **AC005400** or **"cytokeratin 8"**) **Tip**: put an exact phrase in quotes

GO

Chromosome 19 Data Sets

HTML or ASCII

Many links may take a long time to finish.

Contigs	HA
Clones	HA
GenBank CDSs	HA
GrailEXP Gene Models	HA
GrailEXP PolyAs	HA
Genscan Gene Models	HA
Genscan PolyAs	HA
tRNA Genes	HA
STS	HA
Repeats	HA
CpGs	HA

19p13.3

19q13.1

Finished and draft clones and

Human

Pick one

Finished microbial sequences

Finished Archaea

Pick an Organism

Draft microbial genomes sequ

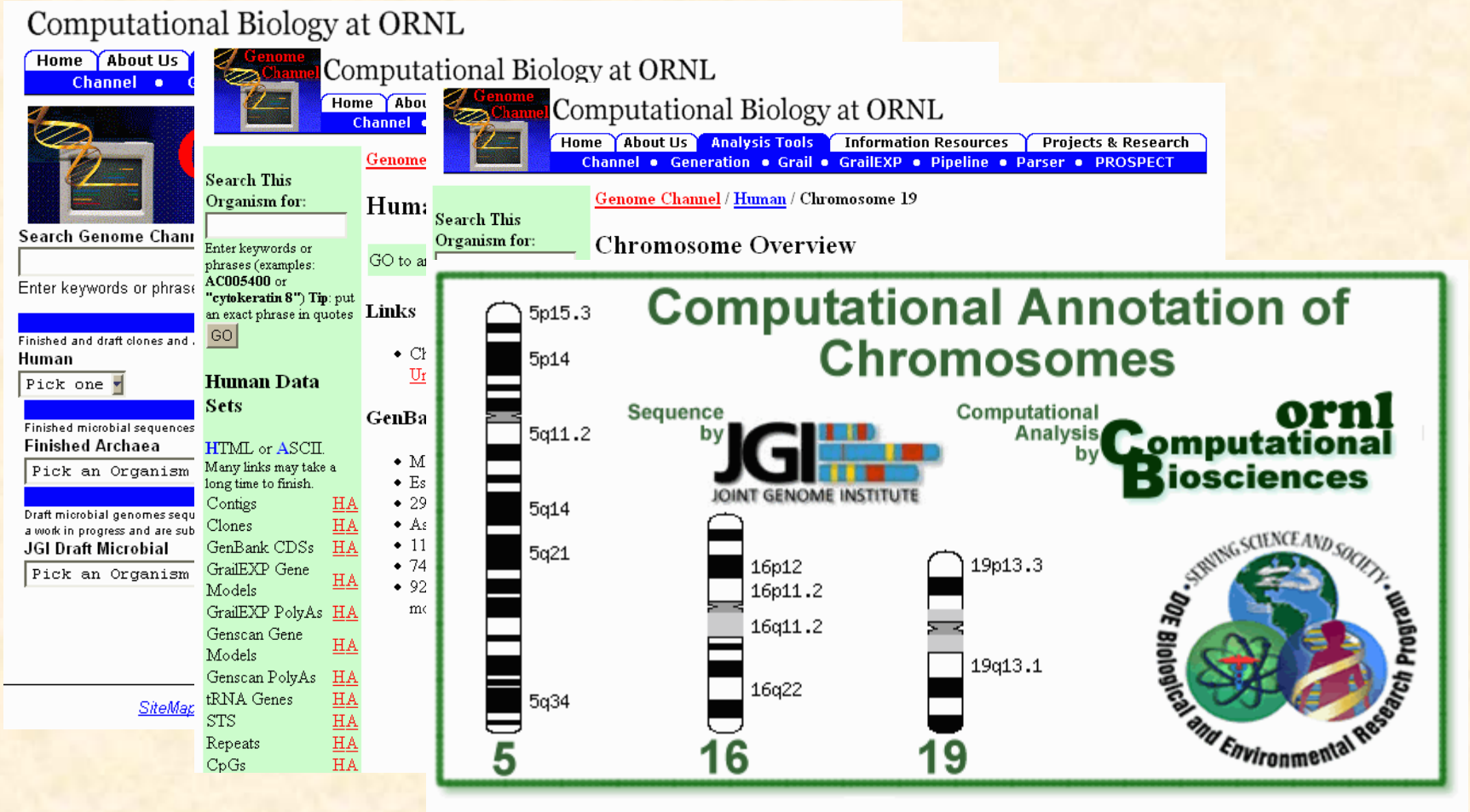
a work in progress and are sub

JGI Draft Microbial

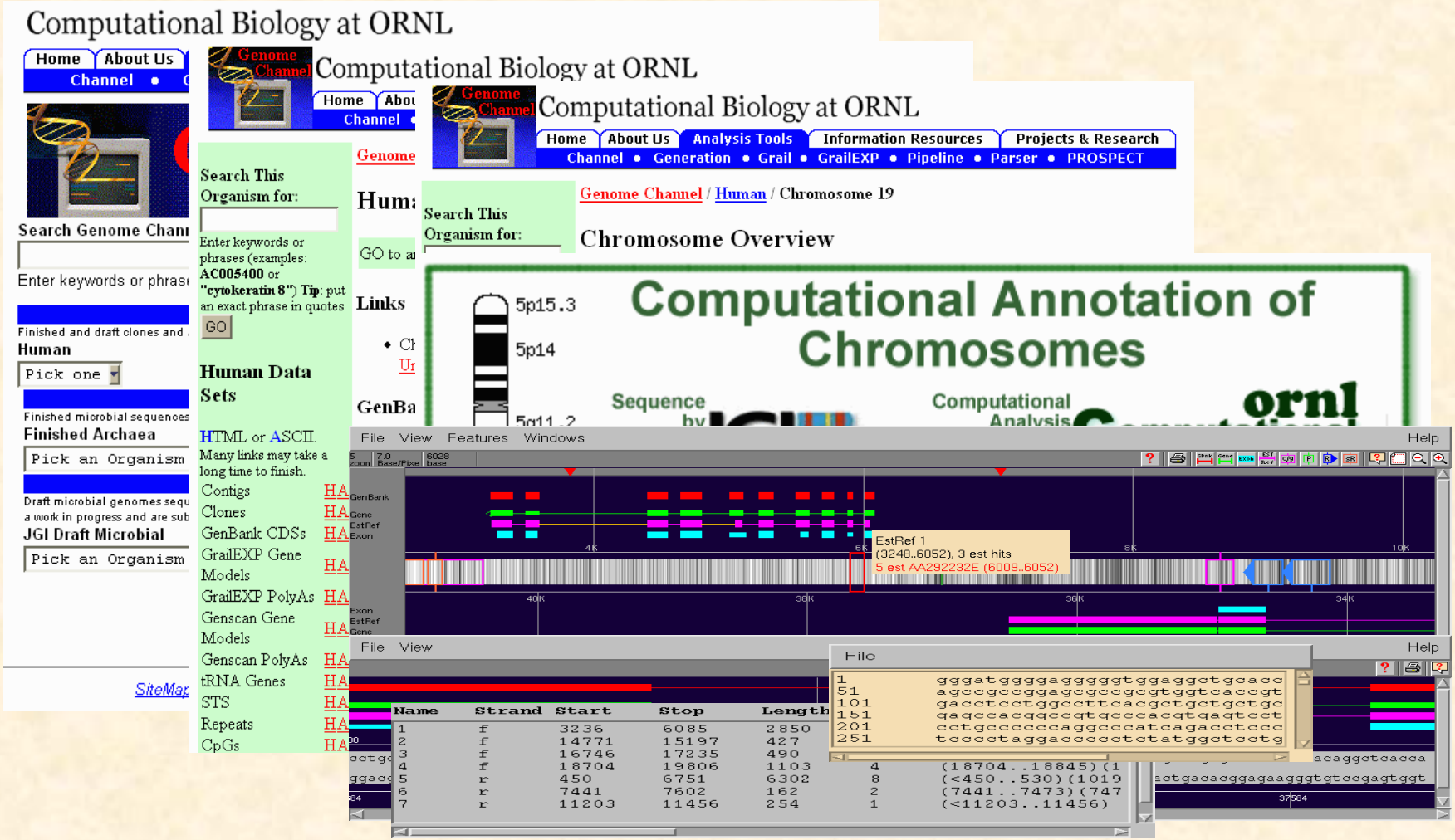
Pick an Organism

SiteMap

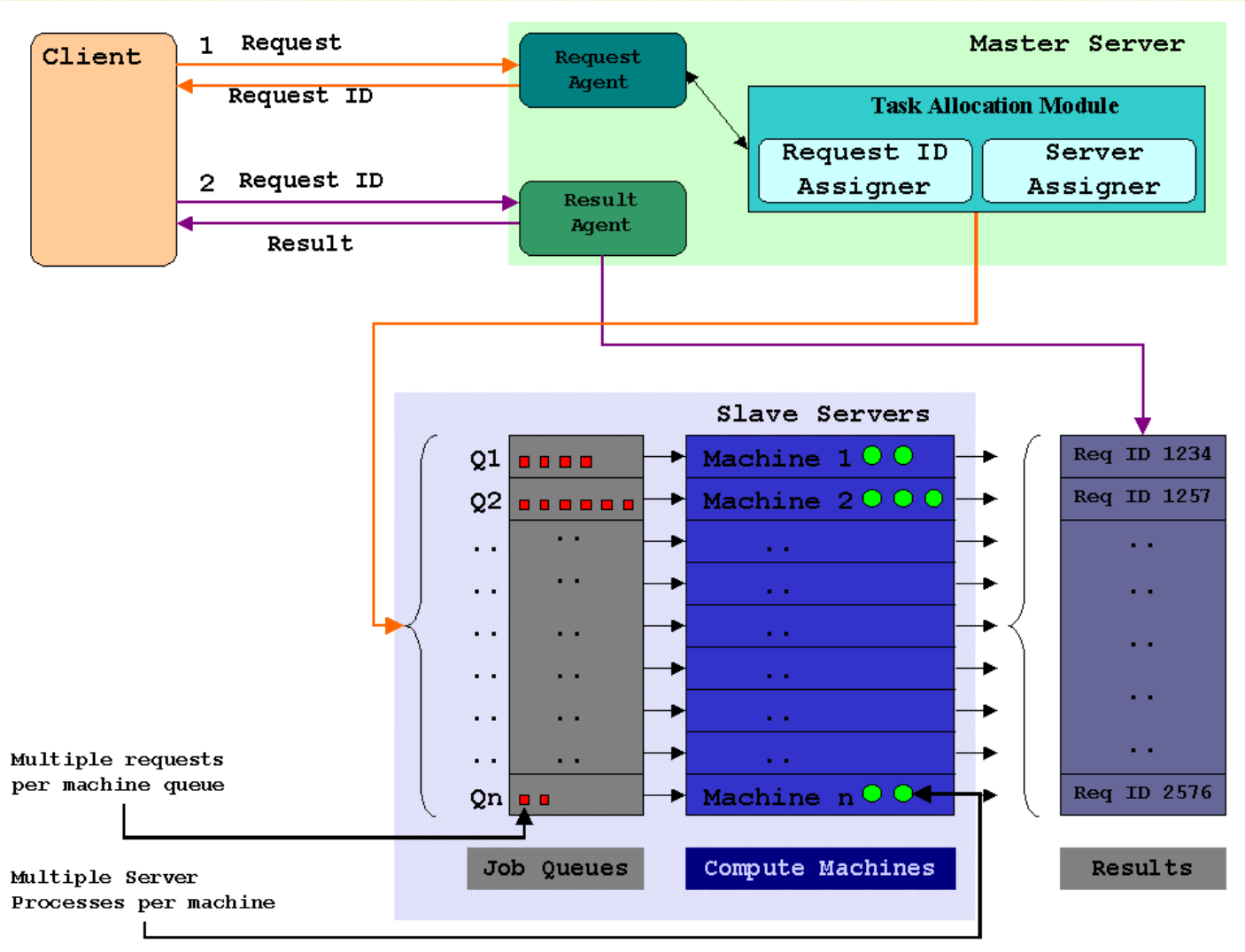
Web access to annotation



Web access to annotation



Web-accessible Analysis Pipeline



Infrastructure Detail

The High-performance Computing Environment

- Hardware:
  - **Eagle:IBM SP3 184 nodes (736 Power 3+ cpus), ~1 Tflops**
  - **Falcon:Compaq 64 nodes (256 Alpha cpus), ~300 Gflops**
  - **Colt :Compaq 16 nodes (64 Alpha cpus), ~100 Gflops**
  - **Htorc : Intel/Linux 64 nodes (128 Pentium III cpus)**

- Storage:
  - **HPSS (High Performance System),**
  - **up to 100TBytes.**

- Applications:
  - **GIST (Genome Integrated Supercomputing Toolkit) , includes MPP NCBI BLAST, MPP HmmerPfam, GrailEXP, PROSPECT, InterPro, genome assembly**

JGI Microbial Sequencing

- Final annotation:
  - *Prochlorococcus marinus* MED4
  - *Rhodopseudomonas palustris*
  - *Synechococcus* WH8102
  - *Prochlorococcus marinus* MIT9313
  - *Nitrosomonas europaea*
- Drafted, finishing:
  - *Nostoc punctiforme* - 198 contigs
  - *Cytophaga hutchinsonii* - 98 contigs
- Microbe Month: 15 species sequenced to deep draft (6-8x coverage)
- Draft output
  - ~93 million bp analyzed
  - ~85,000 gene models
- Microbe drafting ongoing in 2002...

DOE JOINT GENOME INSTITUTE

Operated by The University of California for The US Department of Energy

News and Events

Who We Are
- Mission Statement
- JGI Members/Partners
- Production Genomics Facility
- Organizational Chart
- Contact Us / Employment
- FAQs

JGI Programs
- Human Genome Project
- Microbial Genomics
- HSA19/mouse Comparative Sequencing
- Fugu Genome Project
- Ciona Genome Project
- White Rot Genome Project
- Comparative Genomics
- Instrumentation
- Computational Genomics
- Research and Development

JGI Home
- Production Statistics
- Production Protocols
- JGI Internal Site

Download JGI Sequences

Genome Links

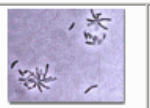
J

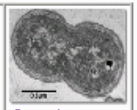
Completed Microbial Genomes

Nitrosomonas europaea

Prochlorococcus marinus strain MED4

Prochlorococcus marinus strain MIT9313

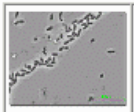

Rhodopseudomonas palustris

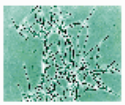

Synechococcus

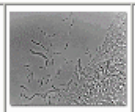
2002 Microbial Genomes

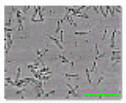
Lactobacillus gasseri

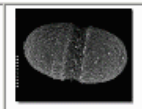
2001 Microbial Genomes

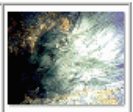

Burkholderia fungorum

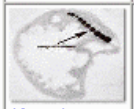

Chloroflexus aurantiacus


Cytophaga hutchinsonii

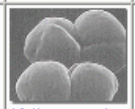

Desulfitobacterium hafniense


Enterococcus faecium


Ferroplasma acidarmanus


Magnetococcus MC-1

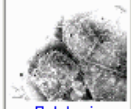
Magnetospirillum magnetotacticum

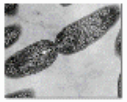

Methanosarcina barkeri

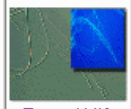
Nostoc punctiforme

Novosphingobium aromaticivorans

Pseudomonas fluorescens


Ralstonia metallidurans


Rhodobacter sphaeroides


Thermobidifa fusca

Xylella fastidiosa strain ann1

Xylella fastidiosa strain dixon

Genome Channel Computational Biology at ORNL

Home | About Us | Analysis Tools | Information Resources | Projects & Research

Channel • Generation • Grail • GrailEXP • Pipeline • Parser • PROSPECT

Produced for the Joint Genome Institute Microbial Sequencing program

DOE Microbial Genomes

DRAFT * DRAFT * DRAFT

Prochlorococcus marinus sp. MED4 analysis files

Version 01sep00

1,657,995 bp in 1 contigs of 20 reads or greater, 31 % GC
1694 candidate protein-encoding gene models

Modeled Genes Organized by:	**FASTA Files**
• Metabolic Pathway • Functional Category • Contig • COGs Functional Group	• Sequence Data (used for gene modeling) • CDS models • CDS translations
Tab-Delimited Files	**RNAs**
• Gene model summary detail tab-delimited • NR blast top hit tab-delimited • COGs tab-delimited • Pfam tab-delimited • EC top hit tab-delimited • GC content, GC and AT skew, per gene • KEGG Genes database top hit	• tRNA content • 16S rRNA • 23S rRNA
Search the Sequence	**Repeats**
• Blast search engine	• Summary of Repeats

Genome Channel Computational Biology at ORNL

Home | About Us | Analysis Tools | Information Resources | Projects & Research

Channel • Generation • Grail • GrailEXP • Pipeline • Parser • PROSPECT

Produced for the Joint Genome Institute Microbial Sequencing program

DOE

DRAFT * DRAFT * DRAFT

Prochlorococcus marinus sp. MED4 files

Version 01sep00

1,657,995 bp in 1 contigs of 20 reads or greater, 31 % G

1694 candidate protein-encoding gene models

Modeled Genes Organized by:	FASTA Files
• Metabolic Pathway • Functional Category • Contig • COGs Functional Group	• Sequence Data (used for gene model • CDS models • CDS translations
Tab-Delimited Files	**RNAs**
• Gene model summary detail tab-delimited • NR blast top hit tab-delimited • COGs tab-delimited • Pfam tab-delimited • EC top hit tab-delimited • GC content, GC and AT skew, per gene • KEGG Genes database top hit	• tRNA content • 16S rRNA • 23S rRNA
Search the Sequence	**Repeats**
• Blast search engine	• Summary of Repeats

Netscape

File Edit View Go Communicator Help

Prochlorococcus marinus MED4 genes g

```
Photosystem I
   Contig26 126- slr1834 - d1018170 - P700 apoprotein subuni
   Contig26 127- slr1835 - d1018171 - P700 apoprotein subuni
   Contig26 61- ssl0563 - d1010687 - photosystem I subunit V
   Contig26 82- slr0737 - d1017421 - photosystem I subunit I
   Contig26 539- ssr2831 - d1019116 - photosystem I subunit
   Contig26 434- sll0819 - d1018841 - photosystem I subunit
   Contig26 435- sml0008 - d1018840 - photosystem I subunit
   Contig24 160- ssr0390 - d1017452 - photosystem I subunit
   Contig26 129- slr1655 - d1019506 - photosystem I subunit
Photosystem II
   Contig26 1081- sll0247 - d1018681 - chlorophyll a/b bindi
   Contig26 833- sll1867 - d1018963 - photosystem II D1 prot
   Contig26 551- slr0906 - d1011109 - photosystem II CP47 pr
   Contig25 192- sll0851 - d1018532 - photosystem II CP43 pr
   Contig25 191- sll0849 - d1018533 - photosystem II D2 prot
   Contig26 870- ssr3451 - d1017825 - cytochrome b559 a subu
   Contig26 871- smr0006 - d1017826 - cytochrome b559 b subu
   Contig26 567- slr1280 - d1019017 - NADH dehydrogenase sub
   Contig26 590- ssl2598 - d1018362 - photosystem II PsbH pr
   Contig26 847- sml0001 - d1010736 - photosystem II PsbI pr
   Contig26 873- smr0008 - d1017828 - photosystem II PsbJ pr
   Contig26 858- sml0005 - d1017808 - photosystem II PsbK pr
   Contig26 872- smr0007 - d1017827 - photosystem II PsbL pr
   Contig26 878- sml0003 - d1017441 - photosystem II PsbM pr
   Contig26 846- smr0009 - d1018363 - photosystem II PsbN pr
   Contig26 604- sll0427 - d1019207 - photosystem II mangane
   Contig26 552- smr0001 - d1010784 - photosystem II PsbT pr
   Contig24 175- sll1398 - d1017351 - photosystem II 13 kD p
   Contig26 413- slr1645 - d1019195 - photosystem II 11 kD p
Phycobilisome
   Contig26 875- sll1577 - d1017965 - class III phycoerythri
Soluble electron carriers
   Contig23 110- sll1245 - d1018905 - cytochrome CytM (cytM)
   Contig25 101- sll0248 - d1018680 - flavodoxin (isiB)
   Contig26378- sll0199 - d1010878 - plastocyanin (petE)
```

Document: Done

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

Genome Channel

Computational Biology at ORNL

Home | About Us | Analysis Tools | Information Resources | Projects & Research

Channel • Generation • Grail • GrailEXP • Pipeline • Parser • PROSPECT

Produced for the Joint Genome Institute Microbial Sequencing program

DOE

DRAFT * DRAFT * DRAFT

Prochlorococcus marinus sp. MED4

files

Version 01sep00

1,657,995 bp in 1 contigs of 20 reads or greater, 31 % G(

1694 candidate protein-encoding gene models

Modeled Genes Organized by:

- Metabolic Pathway
- Functional Category

FASTA Files

- Sequence Data (used for gene model
- CDS models

Tab-

Sear

Netscape

File Edit View Go Communicator Help

Prochlorococcus marinus MED4 genes g

```
Photosystem I
   Contig26 126- slr1834 - d1018170 - P700 apoprotein subuni
   Contig26 127- slr1835 - d1018171 - P700 apoprotein subuni
   Contig26 61- ssl0563 - d1010687 - photosystem I subunit V
   Contig26 82- slr0737 - d1017421 - photosystem I subunit I
   Contig26 539- ssr2831 - d1019116 - photosystem I subunit
   Contig26 434- sll0819 - d1018841 - photosystem I subunit
   Contig26 435- sml0008 - d1018840 - photosystem I subunit
   Contig24 160- ssr0390 - d1017452 - photosystem I subunit
   Contig26 129- slr1655 - d1019506 - photosystem I subunit
Photosystem II
   Contig26 1081- sll0247 - d1018681 - chlorophyll a/b bindi
```

Draft pmar Data - Netscape

File Edit View Go Communicator Help

Gene	Strand	Generation	Glimmer	Critica	Stop	Best Hit	Enzyme Maps
193	r	202304(ATG)..200775	202304(ATG)..200775	202304(ATG)..200775	200775 (TGA)	>gi|2499981|sp|Q55860|CBIP_SYNY3 COBYRIC ACID SYNTHASE >gi|1001778|dbj|BAA10617| (D64004) cobyric acid synthase [Synechocystis sp.] e-118	
109	f	202371(GTG)..202997	202371(GTG)..202997	202371(GTG)..202997	202997 (TAA)	>gi|1653512|dbj|BAA18425| (D90914) MAF [Synechocystis sp.] 2e-31	
192	r	204422(GTG)..203040	204449(TTG)..203040	204422(GTG)..203040	203040 (TAA)	>gi|227610|prf|1707315B photosystem II CP43 protein [Synechococcus sp.] 0.0	
191	r	205482(ATG)..204406	205482(ATG)..204406	205482(ATG)..204406	204406 (TAA)	>gi|131299|sp|P11005|PSBD_SYNP7 PHOTOSYSTEM II D2 PROTEIN (PHOTOSYSTEM Q(A) PROTEIN) >gi|79676|pir||JT0321 photosystem II protein D2 - Synechococcus sp. (PCC 7942) >gi|154586 (M20815) D2 thylakoid protein [Synechococcus sp.] >gi|154588 (M20814) psbDI thylakoid protein [Synechococcus sp.] 0.0	
110	f	205667(ATG)..206224	205667(ATG)..206224	205667(ATG)..206224	206224 (TAA)	>gi|1723362|sp|P51220|YCF4_PORPU HYPOTHETICAL 21.2 KD PROTEIN YCF4 (ORF186) >gi|2147565|pir||S73141 hypothetical protein 4 - Porphyra purpurea chloroplast >gi|1276686 (U38804) hypothetical chloroplast ORF 4. [Porphyra purpurea] 2e-34	
190	r	207401(ATG)..206877	207401(ATG)..206877	207401(ATG)..206877	206877 (TAA)	>gi|1708470|sp|P51230|ILVH_PORPU ACETOLACTATE SYNTHASE SMALL SUBUNIT (AHAS) (ACETOHYDROXY-ACID SYNTHASE SMALL SUBUNIT) (ALS) >gi|2147069|pir||S73151 acetohydroxyacid synthase small chain - Porphyra purpurea chloroplast >gi|1276696 (U38804) acetohydroxyacid synthase small subunit [Porphyra purpurea] 4e-56	4.1.3.18 - map00290 map00650 map00660 map00770
111	f	206398(GTG)..206880	206347(TTG)..206880	206398(GTG)..206880	206880 (TAA)	>gi|3184556 (AF052290) peptidyl-prolyl cis-trans isomerase B [Synechococcus PCC7002] 5e-18	

OAK
U. S. I

Genome Channel

Computational Biology at ORNL

Home | About Us | Analysis Tools | Information Resources | Projects & Research

Channel • Generation • Grail • GrailEXP • Pipeline • Parser • PROSPECT

Produced for the Joint Genome Institute Microbial Sequencing program

Prochl

Modeled Genes Organiz

- Metabolic Pathway
- Functional Category

Tab-

Sear

OAK
U. S. I

Prochlorococcus marinus files - Netscape

File Edit View Go Communicator Help

Version 10dec99 - Contig26 Gene 875

Gene Finders

Strand = r
Stop Location = 279721
Stop Codon = TAG

Gene Modeler	Start Location	Start Codon
Generation	279194	ATG
Glimmer	279194	ATG
Critica	279194	ATG

Blast against NR

Best Hit	Evalue	Percent Identity
gi|1669716|emb|CAA93118| (Z68890) class III phycoerythrin beta-subunit [Prochlorococcus marinus] >gi|5734498|emb|CAB52705.1| (AJ001230) phycoerythrin type III beta subunit [Prochlorococcus marinus]	9e-29	39

Refresh BLAST results

PFam Model Comparison

Model	Description	Evalue	Score
Phycobilisome	Phycobilisome proteins	9.4e-29	103.1

Refresh PFAM results

Best NR Hit with an EC Number

No EC Pathway hits

***Synechocystis PCC6803* Comparison**

Synecho ID	Description	Evalue	Percent Identity	Category
/	phycocyanin b subunit	4e-18	31.617647	

MRNA

```
ATGACAGTTTCAAAGAGTAATCAAATTTTATCAAATGATAGAGATTTAGAAAATATAAGTAATAAAAATATTGAAGACAT
AAAAGAATTTATTAATACTGCAAACTCAAGATTAGATGCAATAGATTCAATAACAAATAATTCTCACGCAATTGCCGCTG
ATGCTGTGACTGCAATGATTTGTGAAAATCAAGATTCAGTTAATACAAAAATATCTTTAGATACCACCAATAAGATGTCT
GTTTGTCTAAGAGATGGAGAAATAATTTTAAGGATTGTTTCTTACCTTTTGATTTCTGATGACGAATCAGTTTTATCTAA
AAACTGTTTAAAGGATCTTAAAAATACTTATTTGGCCCTTGGTGTACCTCTGAAAAATGCTATCCGAGTTTTTGAATTGA
TGAGAGATGCAACGATTTCTGATTTAAAGTCTACTGTAAATTCTATGAAAGGAGAAAAAGAATTTCTTTCTGATTTAATT
TCTAATACAGAGTTTCAATTTGAGAGAATAATTAATCTTTTAAGATAG
```

PROTEIN

```
MTVSKSNQILSNDRDLENISNKNIEDIKEFINTANSRLDAIDSITNNSHAIAADAVTAMICENQDSVNTKISLDTTNKMS
VCLRDGEIILRIVSYLLISDDESVLSKNCLKDLKNTYLALGVPLKNAIRVFELMRDATISDLKSTVNSMKGEKEFLSDLI
SNTEFQFERIINLLR*
```

Document: Done

Draft pmar Data - Netscape

File Edit View Go Communi

Gene	Strand	C
193	r	20230
109	f	20237
192	r	20442
191	r	20548
110	f	20566
190	r	20740
111	f	20639

OE

Help

s marinus **MED4 genes g**

```
834 – d1018170 – P700 apoprotein subuni
835 – d1018171 – P700 apoprotein subuni
63 – d1010687 – photosystem I subunit V
37 – d1017421 – photosystem I subunit I
831 – d1019116 – photosystem I subunit
819 – d1018841 – photosystem I subunit
008 – d1018840 – photosystem I subunit
390 – d1017452 – photosystem I subunit
655 – d1019506 – photosystem I subunit

0247 – d1018681 – chlorophyll a/b bindi
```

Hit	Enzyme Maps
YRIC ACID SYNTHASE ic acid synthase [Synechocystis sp.]	
[Synechocystis sp.] 2e-31	
43 protein [Synechococcus sp.] 0.0	
OSYSTEM II D2 PROTEIN 76|pir|JT0321 photosystem II protein D2 (M20815) D2 thylakoid protein sbDI thylakoid protein [Synechococcus	
OTHETICAL 21.2 KD PROTEIN pothetical protein 4 - Porphyra purpurea al chloroplast ORF 4. [Porphyra	
TOLACTATE SYNTHASE SMALL CID SYNTHASE SMALL SUBUNIT) acid synthase small chain - Porphyra cetohydroxyacid synthase small subunit	4.1.3.18 - map00290 map00650 map00660 map00770
rans isomerase B [Synechococcus	

D1 prot
CP47 pr
CP43 pr
D2 prot
a subu
b subu
ase sub
PsbH pr
PsbI pr
PsbJ pr
PsbK pr
PsbL pr
PsbM pr
PsbN pr
mangane
PsbT pr
13 kD p
11 kD p
erythri
(cytM)
B)
tE)

Genome Channel

Computational Biology at ORNL

Home | About Us | Analysis Tools | Information Resources | Projects & Research

Channel • Generation • Grail • GrailEXP • Pipeline • Parser • PROSPECT

Produced for the Joint Genome Institute Microbial Sequencing program

Prochl

Prochlorococcus marinus files - Netscape

File Edit View Go Communicator Help

Version 10dec99 - Contig26 Gene 875

Gene Finders

Strand = r
Stop Location = 279721
Stop Codon = TAG

Gene Modeler	Start Location	Start Codon
Generation	279194	ATG
Glimmer	279194	ATG
Critica	279194	ATG

Blast against NR

Best Hit
gi|1669716|emb|CAA93118| (Z68890) class III phycoerythrin beta-subunit [Prochlorococcus m
>gi|5734498|emb|CAB52705.1| (AJ001230) phycoerythrin type III beta subunit [Prochlorococ

Refresh BLAST results

PFam Model Comparison

Model	Description	Evalue	Score
Phycobilisome	Phycobilisome proteins	9.4e-29	103.1

Refresh PFAM results

Best NR Hit with an EC Number

No EC Pathway hits

***Synechocystis PCC6803* Comparison**

Synecho ID	Description	Evalue	Percent Identity	Category
/	phycocyanin b subunit	4e-18	31.617647	

MRNA

```
ATGACAGTTTCAAAGAGTAATCAAATTTTATCAAATGATAGAGATTTAGAAAATATAAGTAATAAAAATATTGAAGACAT
AAAAGAATTTATTAATACTGCAAACTCAAGATTAGATGCAATAGATTCAATAACAAATAATTCTCACGCAATTGCCGCTG
ATGCTGTGACTGCAATGATTTGTGAAAATCAAGATTCAGTTAATACAAAAATATCTTTAGATACCACCAATAAGATGTCT
GTTTGTCTAAGAGATGGAGAAATAATTTTAAGGATTGTTTCTTACCTTTTGATTTCTGATGACGAATCAGTTTTATCTAA
AAACTGTTTAAAGGATCTTAAAAATACTTATTTGGCCCTTGGTGTACCTCTGAAAAATGCTATCCGAGTTTTTGAATTGA
TGAGAGATGCAACGATTTCTGATTTAAAGTCTACTGTAAATTCTATGAAAGGAGAAAAAGAATTTCTTTCTGATTTAATT
TCTAATACAGAGTTTCAATTTGAGAGAATAATTAATCTTTTAAGATAG
```

PROTEIN

```
MTVSKSNQILSNDRDLENISNKNIEDIKEFINTANSRLDAIDSITNNSHAIAADAVTAMICENQDSVNTKISLDTTNKMS
VCLRDGEIILRIVSYLLISDDESVLSKNCLKDLKNTYLALGVPLKNAIRVFELMRDATISDLKSTVNSMKGEKEFLSDLI
SNTEFQFERIINLLR*
```

Document: Done

Rhodopseudomonas palustris 1 - ORNL

METABOLIC PATHWAYS

Metabolism of Complex Carbohydrates
Metabolism of Cofactors and Vitamins
Nucleotide Metabolism
Metabolism of Complex Lipids
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Other Substances

01100 5/1/00

Modeled Genes Organiz

- Metabolic Pathway
- Functional Category

Tab-

Sear

Draft pmar Data - Netscap

File Edit View Go Communi

Gene	Strand	C
193	r	20230
109	f	20237
192	r	20442
191	r	20548
110	f	20566
190	r	20740
111	f	20639

Enzyme Maps
4.1.3.18 - map00290 map00650 map00660 map00770

[Synechocystis sp.] 2e-31

43 protein [Synechococcus sp.] 0.0

OSYSTEM II D2 PROTEIN
76|pir|JT0321 photosystem II protein D2
(M20815) D2 thylakoid protein
sbDI thylakoid protein [Synechococcus

OTHETICAL 21.2 KD PROTEIN
pothetical protein 4 - Porphyra purpurea
al chloroplast ORF 4. [Porphyra

TOLACTATE SYNTHASE SMALL
CID SYNTHASE SMALL SUBUNIT)
acid synthase small chain - Porphyra
cetohydroxyacid synthase small subunit

rans isomerase B [Synechococcus

MED4 genes g

- P700 apoprotein subuni
- P700 apoprotein subuni
photosystem I subunit V
photosystem I subunit I
- photosystem I subunit
- photosystem I subunit
- photosystem I subunit
- photosystem I subunit
- photosystem I subunit

– chlorophyll a/b bindi

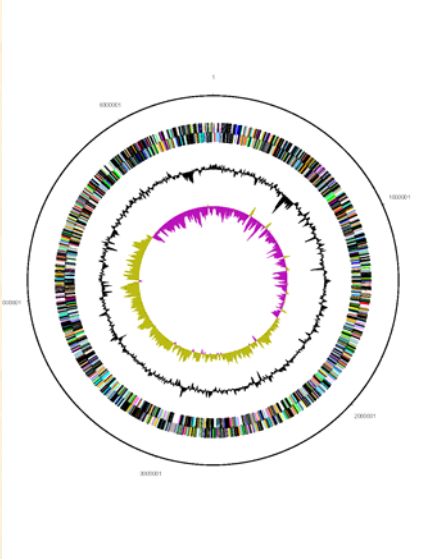
D1 prot
CP47 pr
CP43 pr
D2 prot
a subu
b subu
ase sub
PsbH pr
PsbI pr
PsbJ pr
PsbK pr
PsbL pr
PsbM pr
PsbN pr
mangane
PsbT pr
13 kD p
11 kD p

erythri

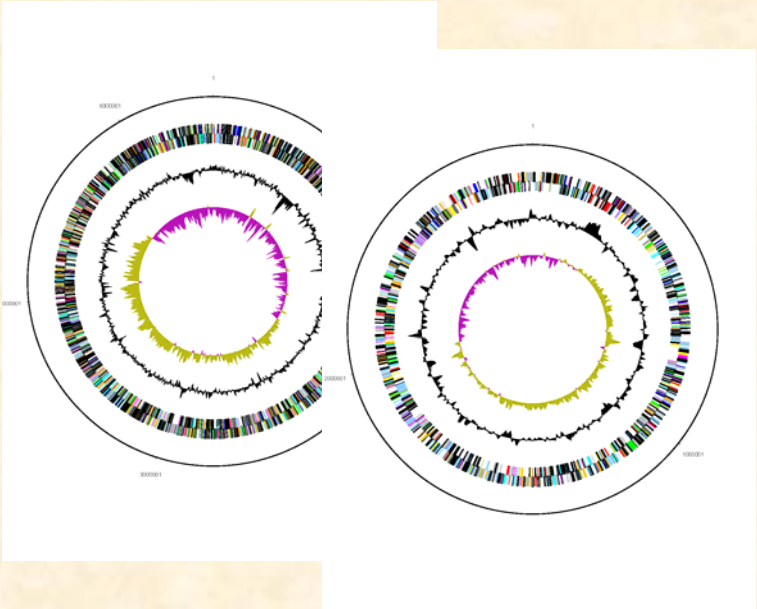
(cytM)
B)
tE)

OAK
U. S. I

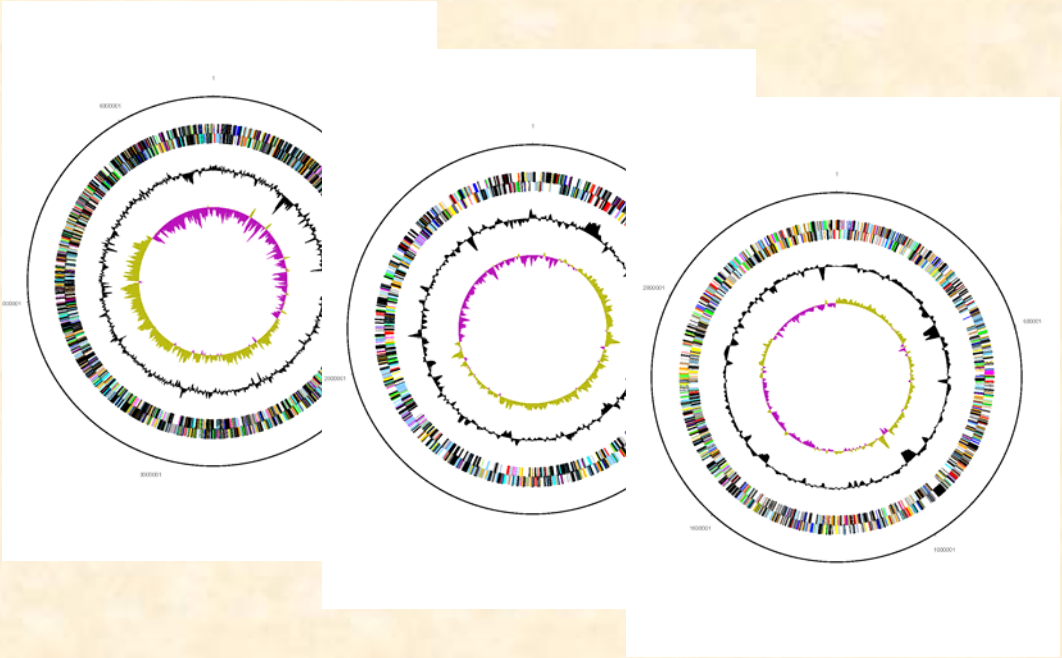
Perspective has shifted from genes to genomes



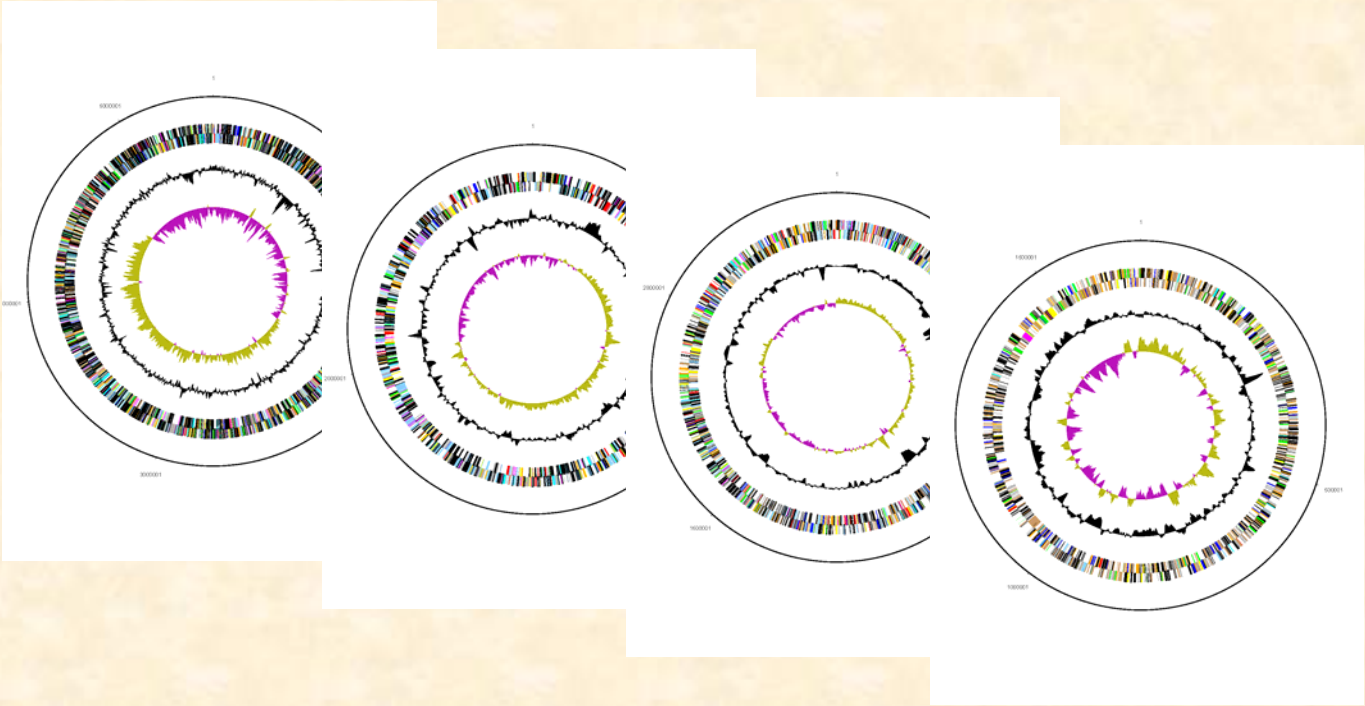
Perspective has shifted from genes to genomes



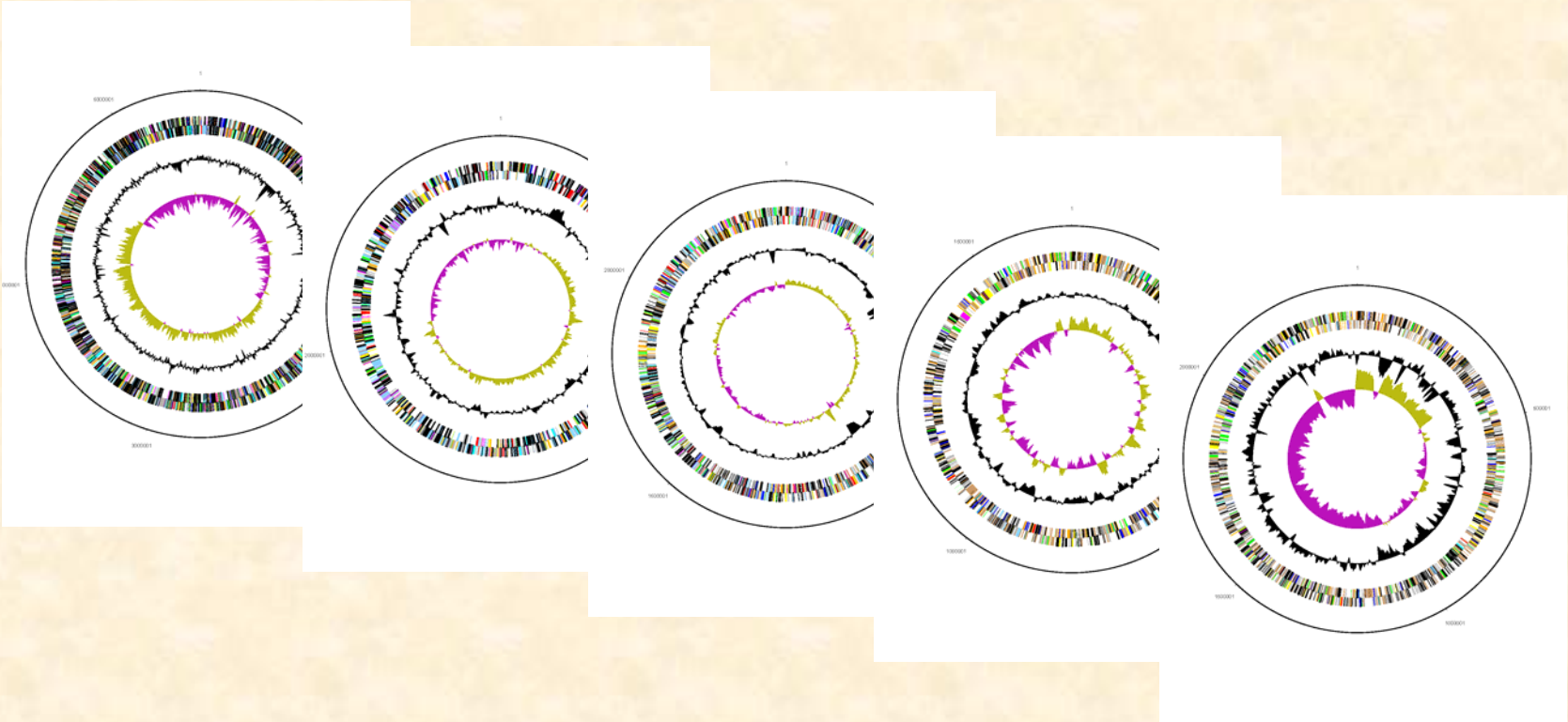
Perspective has shifted from genes to genomes



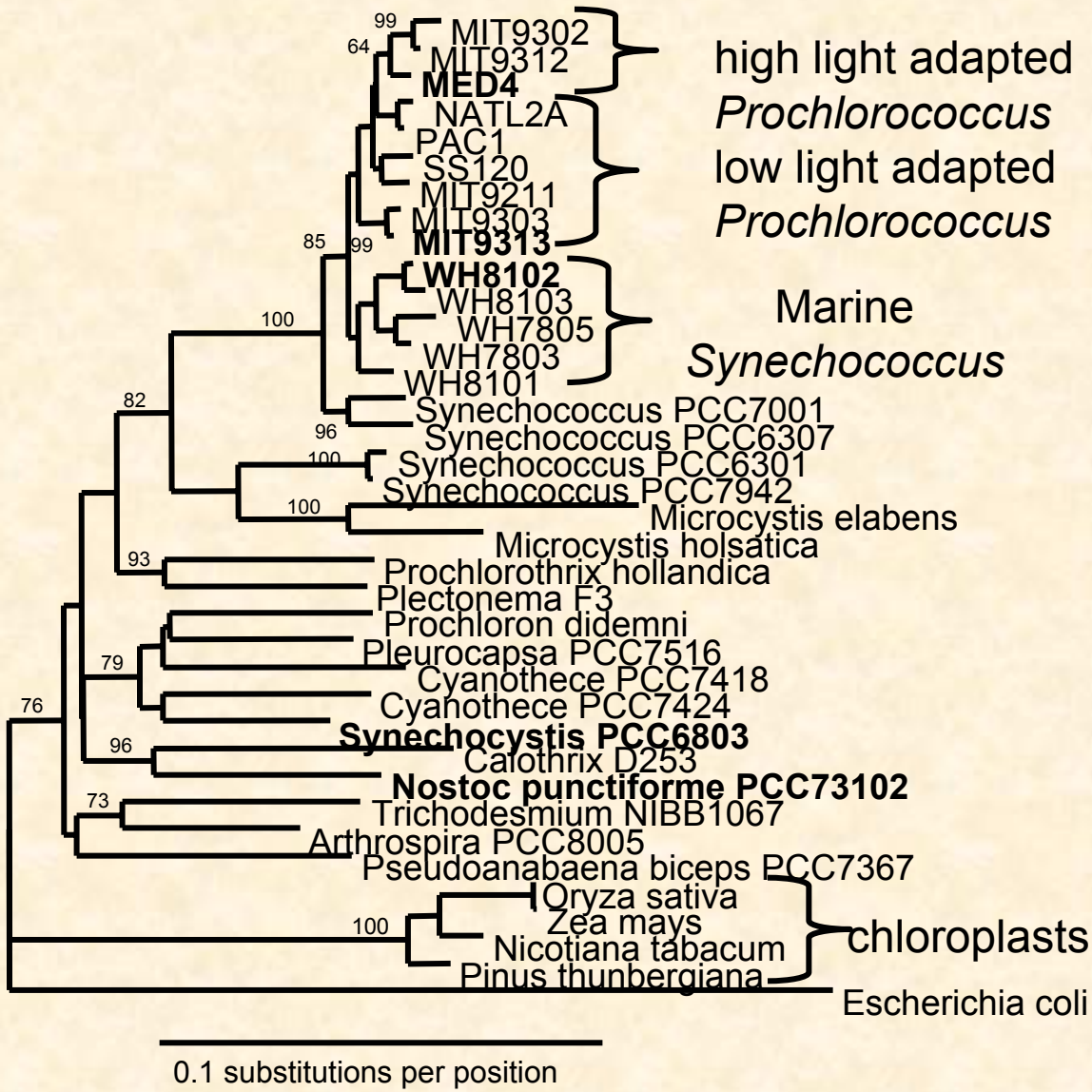
Perspective has shifted from genes to genomes



Perspective has shifted from genes to genomes



Comparative genomics

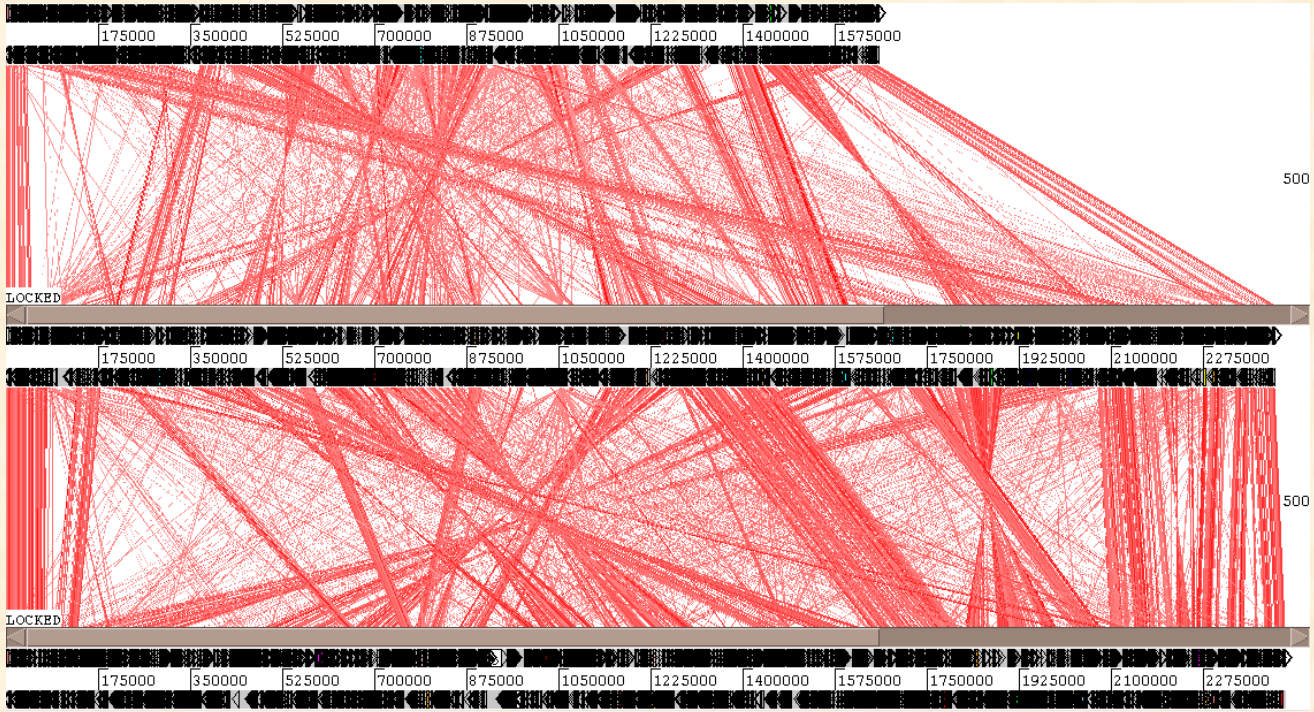


Comparative Genomics: Global alignment

P. marinus MED4

P. marinus MIT9313

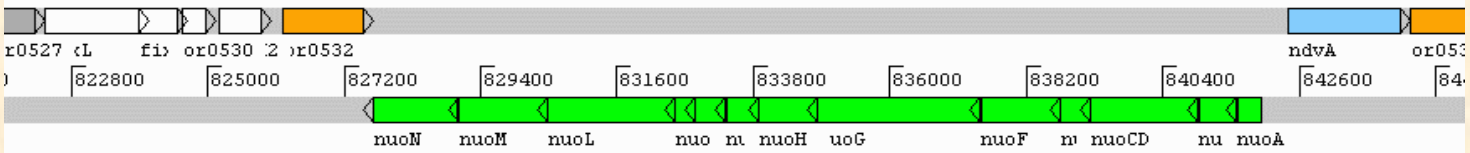
Synechococcus WH8102



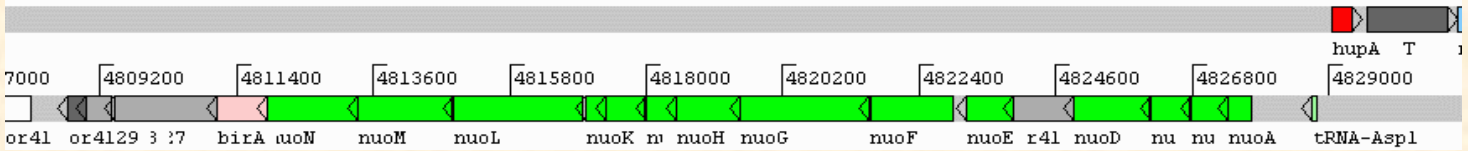
NADH-ubiquinone oxidoreductase operon organization

Unified operon

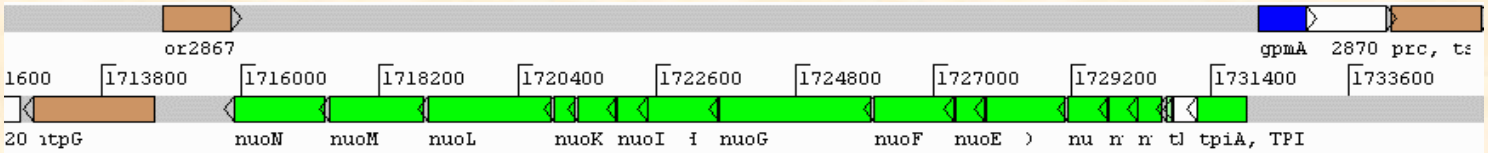
R. palustris



R. palustris

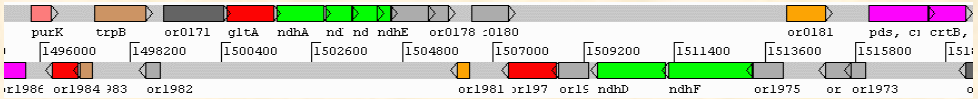


N. europaea

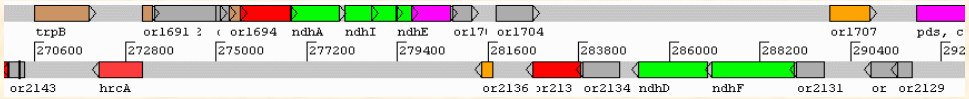


NADH-ubiquinone oxidoreductase operon organization

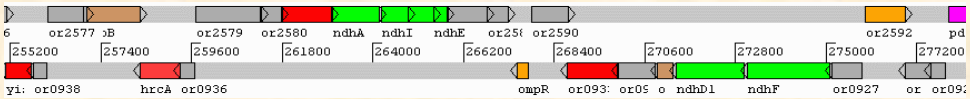
Dispersed operons



P. marinus MED4

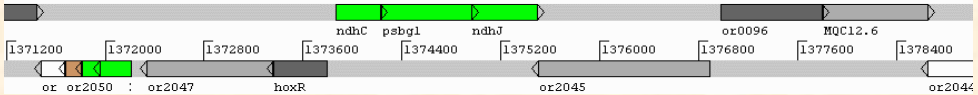


P. marinus MIT9313

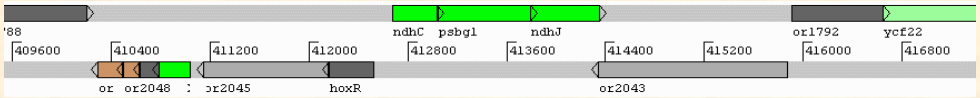


Synechococcus WH8102

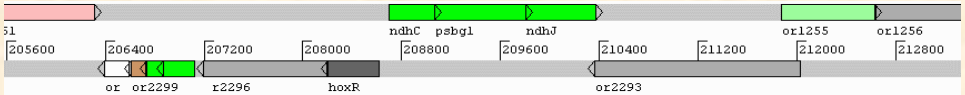
P. marinus MED4



P. marinus MIT9313

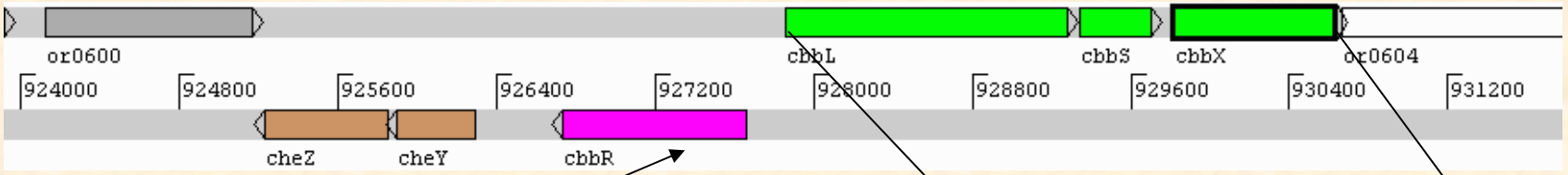


Synechococcus WH8102

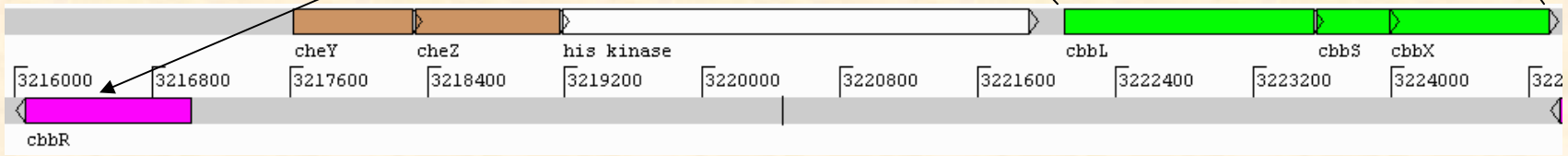


Differing organization of RuBisCO operons

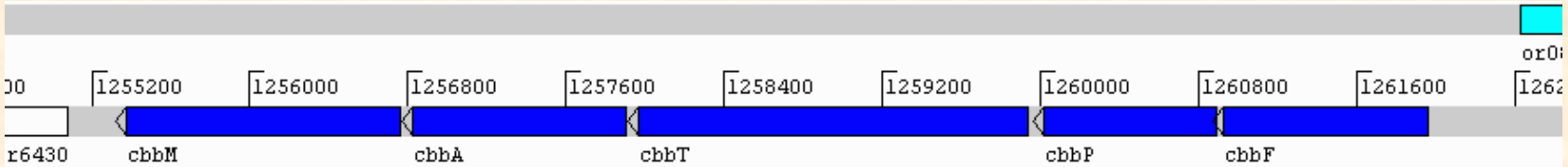
N. europaea



R. palustris

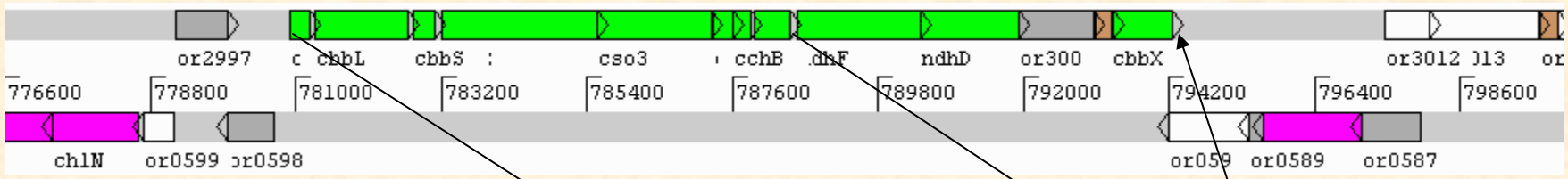


R. palustris

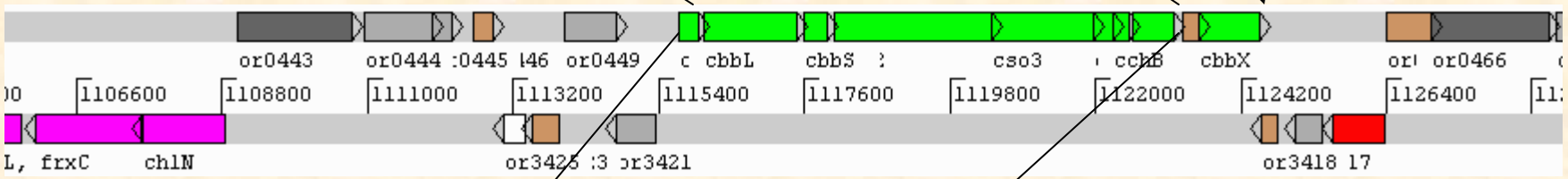


Differing organization of RuBisCO operons

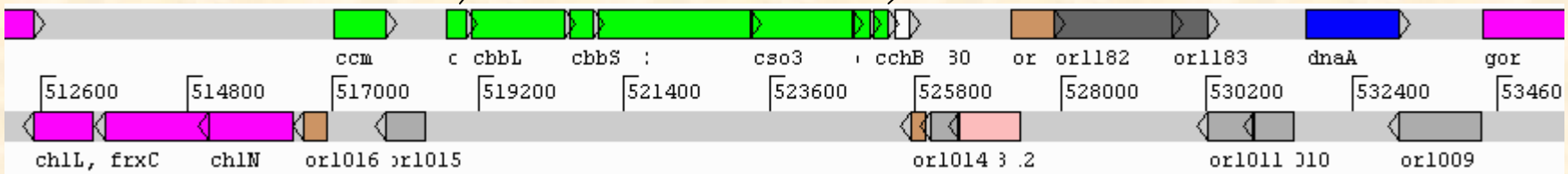
Synechococcus WH8102



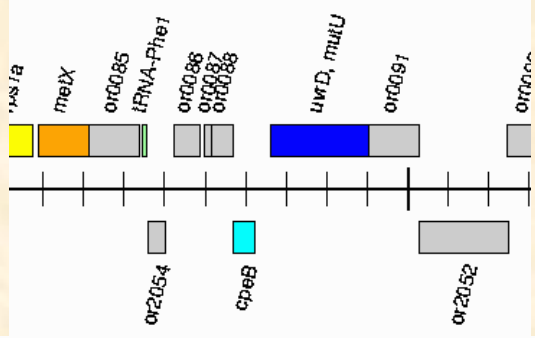
P. marinus MIT9313



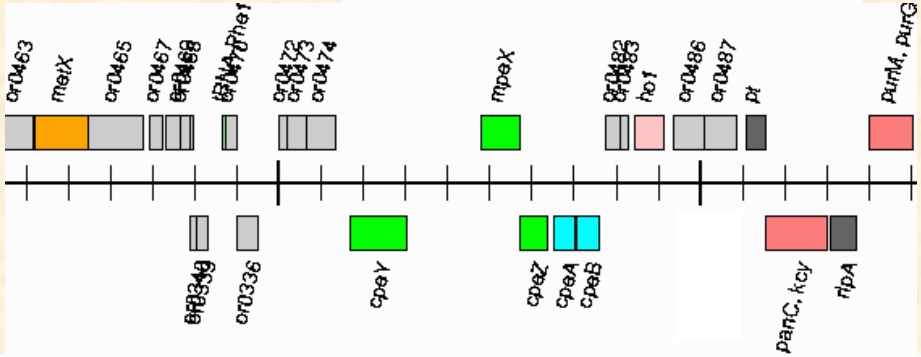
P. marinus MED4



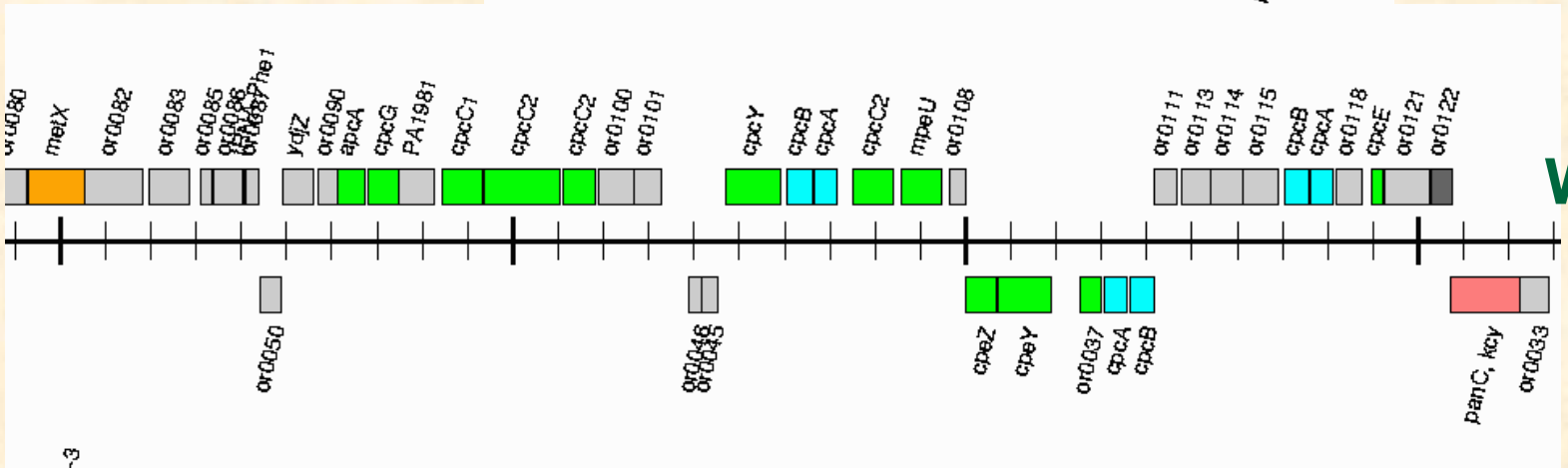
Phycoerythrin and phycocyanin gene clusters



MED4



MIT9313



WH8102

Futures

- High-throughput microbial genomics
- White rot fungus
- Finished human chr5, 16, and 19
- Poplar
- Comparative genomics
  - genome evolution
  - operons, regulons, and control networks
  - proteomics
  - community perspective - “genomes of ecosystems”

Acknowledgements

ORNL Genome Analysis

Erich Baker
Gwo-Liang Chen
Michael Galloway
Loren Hauser
Douglas Hyatt
Miriam Land
Phil Locascio
Victor Olman
Denise Schmoyer
Manesh Shah
Jay Snoddy
Inna Volker
Ed Uberbacher

Collaborators

Dan Arp
Ron Atlas
Penny Chisholm
Dan Cullen
Carrie Harwood
Bernard Henrissat
Mike Himmel
Allen Hooper
Martin Klotz
Mark McBride
Ron Mackenzie
Jack Meeks
Brian Palenik
Gabrielle Rocap
Bob Tabita

JGI

Dan Rokhsar
Stephanie Stilwagen
Art Kobayashi
Paul Predki

LLNL

Patrick Chain
Jane Lamerdin

USDOE

Marv Frazier
Dan Drell
Anna Palmisano

Genome annotation: Bioinformatics for high-throughput genomics and beyond

Frank Larimer

Oak Ridge National Laboratory

9th DOE Contractor and Grantee Workshop

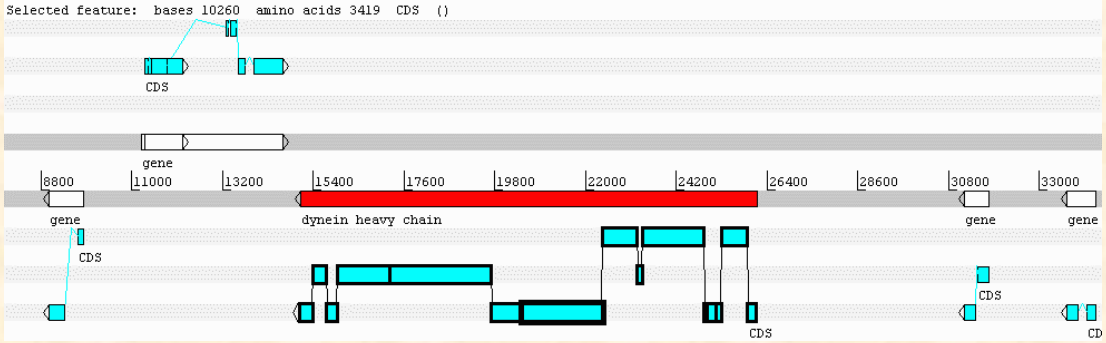
Oakland, CA, 28 January 2002

GrailEXP development for annotation of the white rot genome

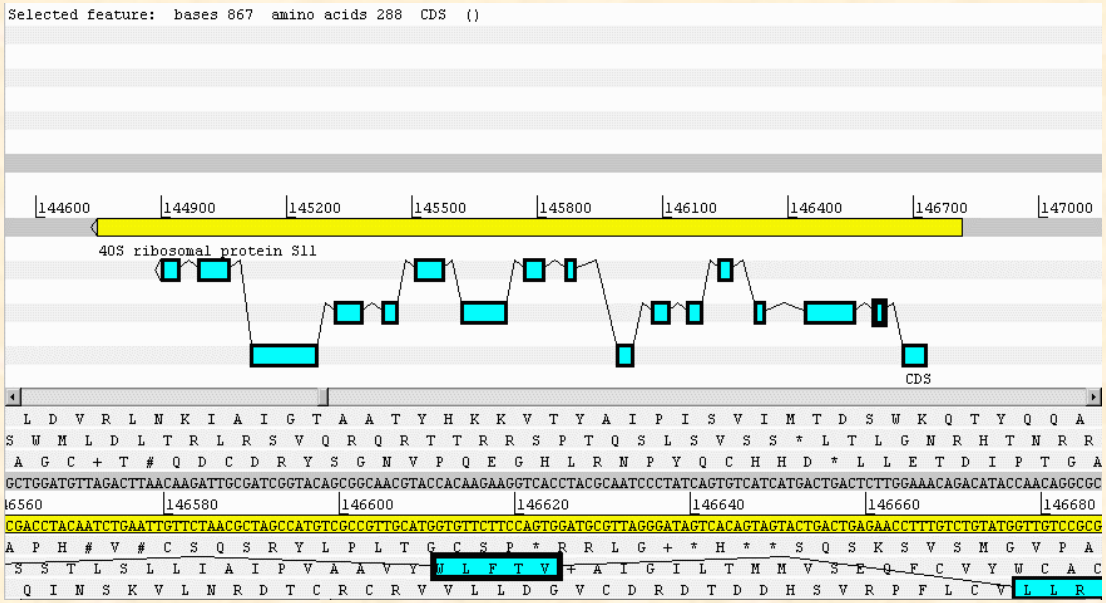
- A set of ~1300 white rot ESTs and mRNAs was aligned with the genomic sequence using GrailEXP
- Using the information from these ESTs, splice site scoring systems were constructed.
- GrailEXP was run on the genomic sequence using the custom white rot splice site scoring system and revised rules for intron sizes (rewarding introns close to the preferred size of ~50 bases).
- Iterative runs, with retraining, incorporating (prototype) protein alignment and heterologous ESTs

Gene modeling: white rot

Dynein: large exons



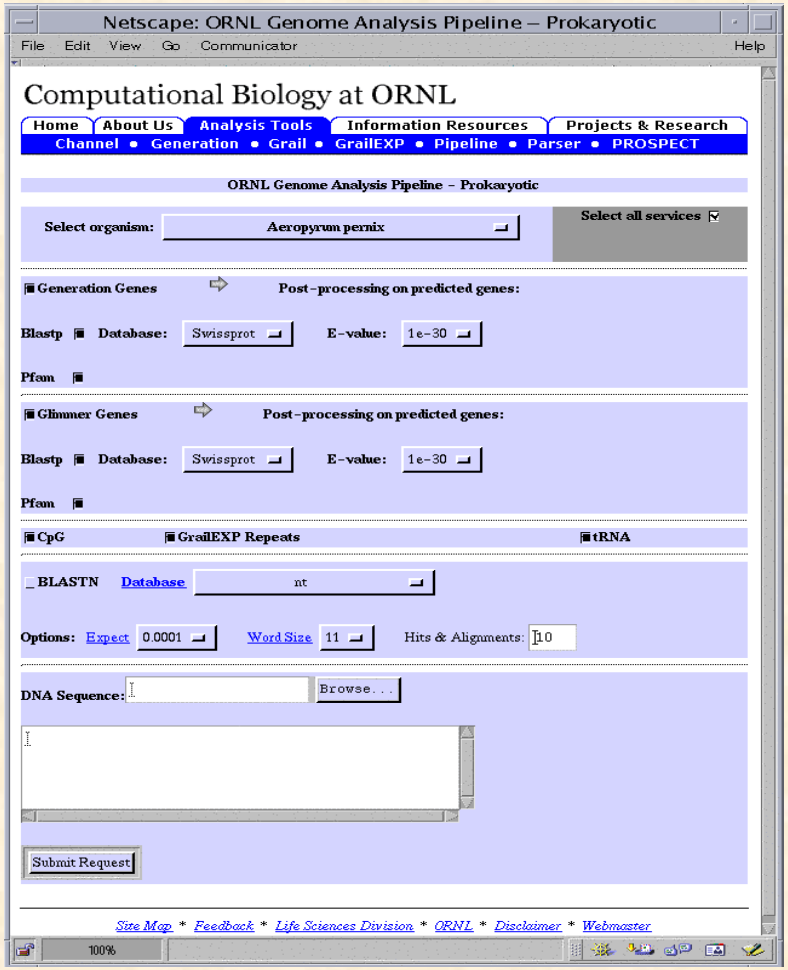
S11: small exons



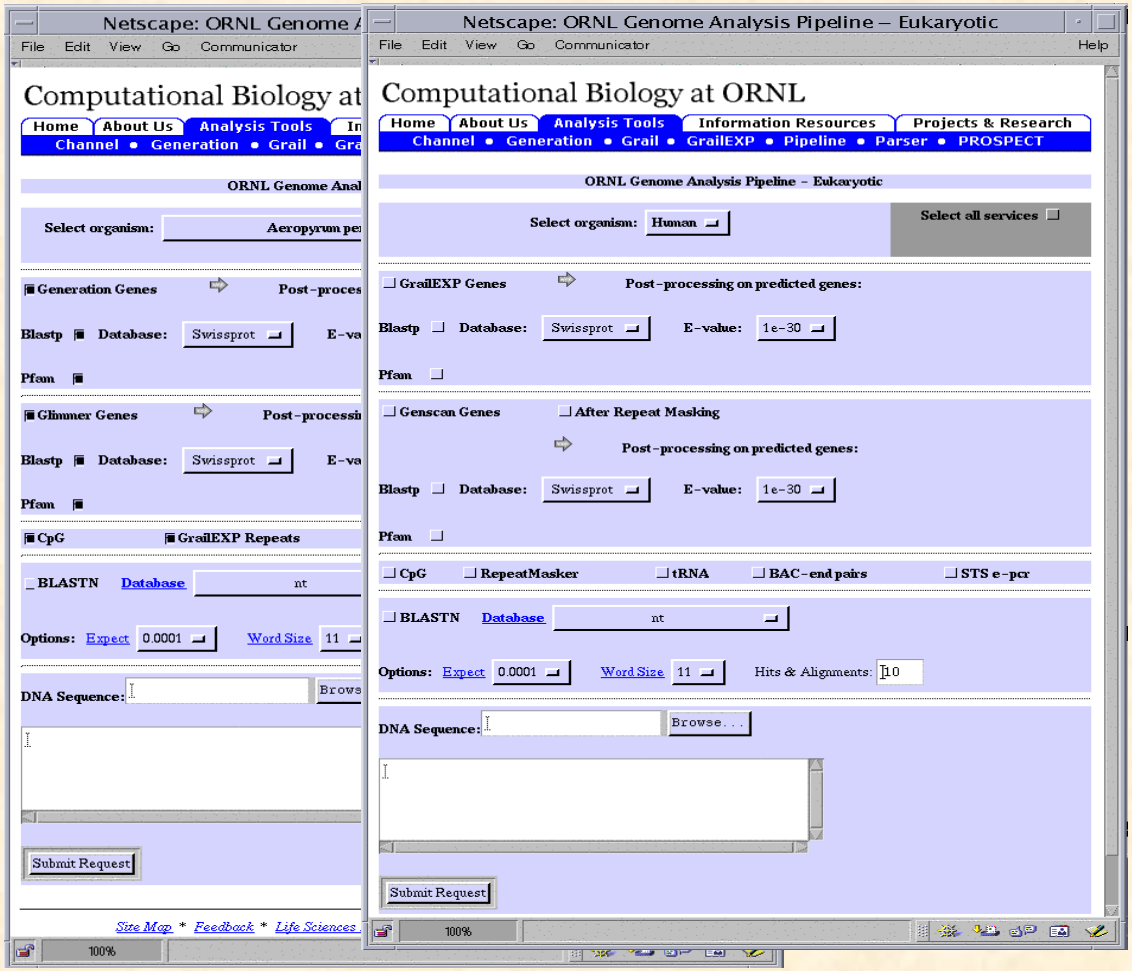
GrailEXP developemnt

- Training of GrailEXP on new genomes
- Streamlining and automation of the training process
- Recognition of more forms of alternative
- splicing
- Addition of protein homology to the system
- Addition of a TBLASTX-based approach for utilizing other genomic sequences, I.e. *Fugu* alignments to locate human genes
- Improvement of user interfaces (JAVA/HTML)
- Support for the DAS (Distributed Annotation System) protocol

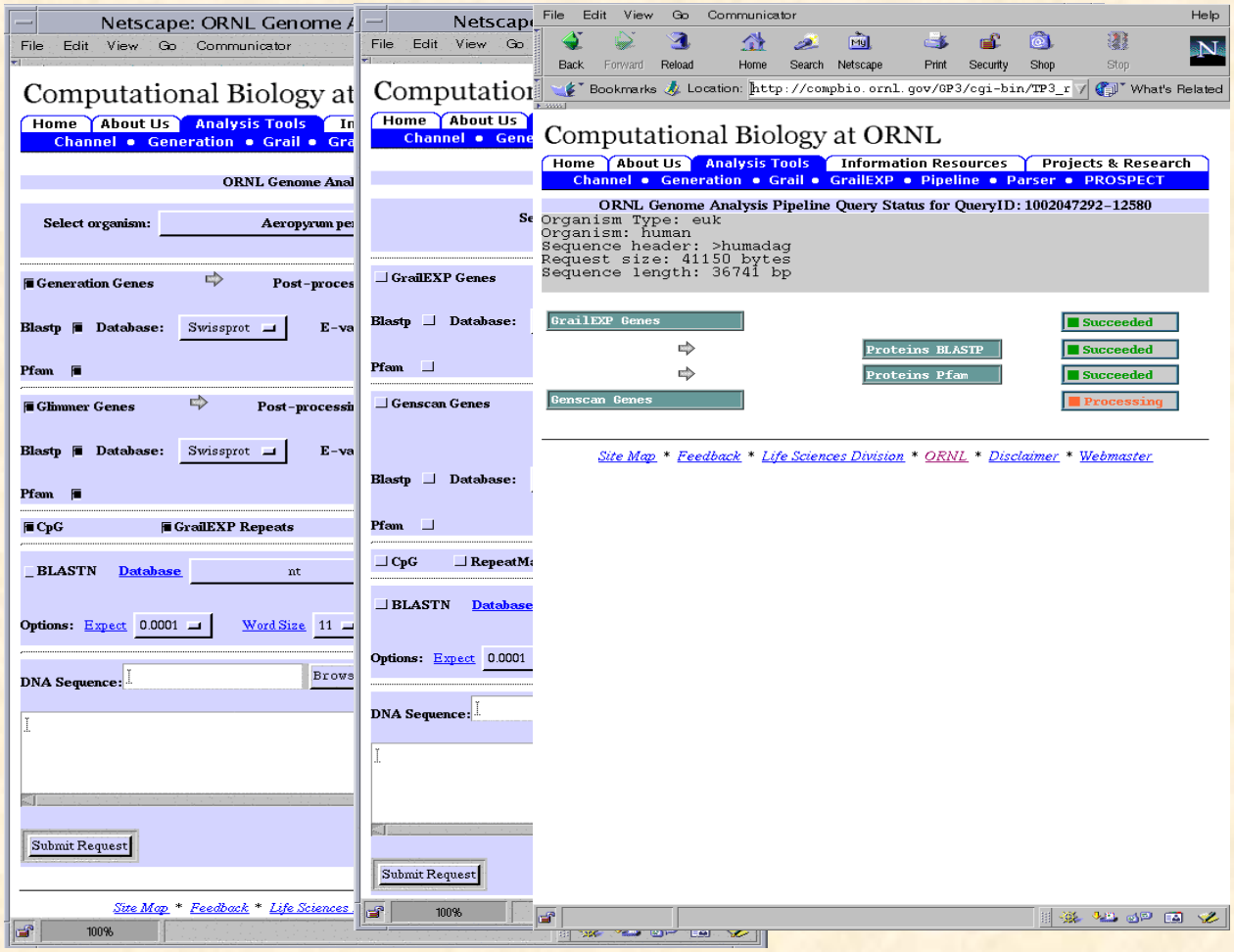
Analysis tools in action



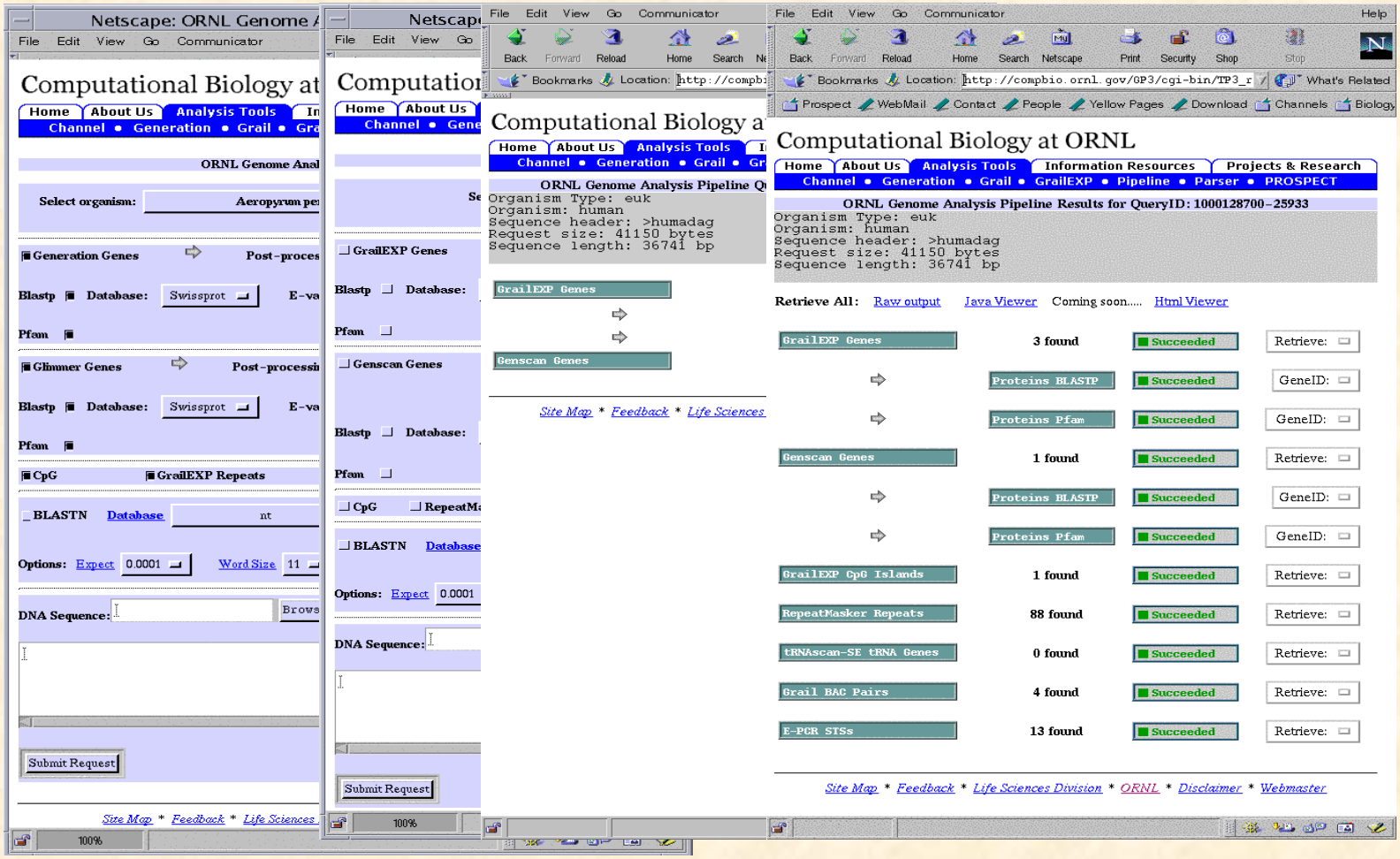
Analysis tools in action



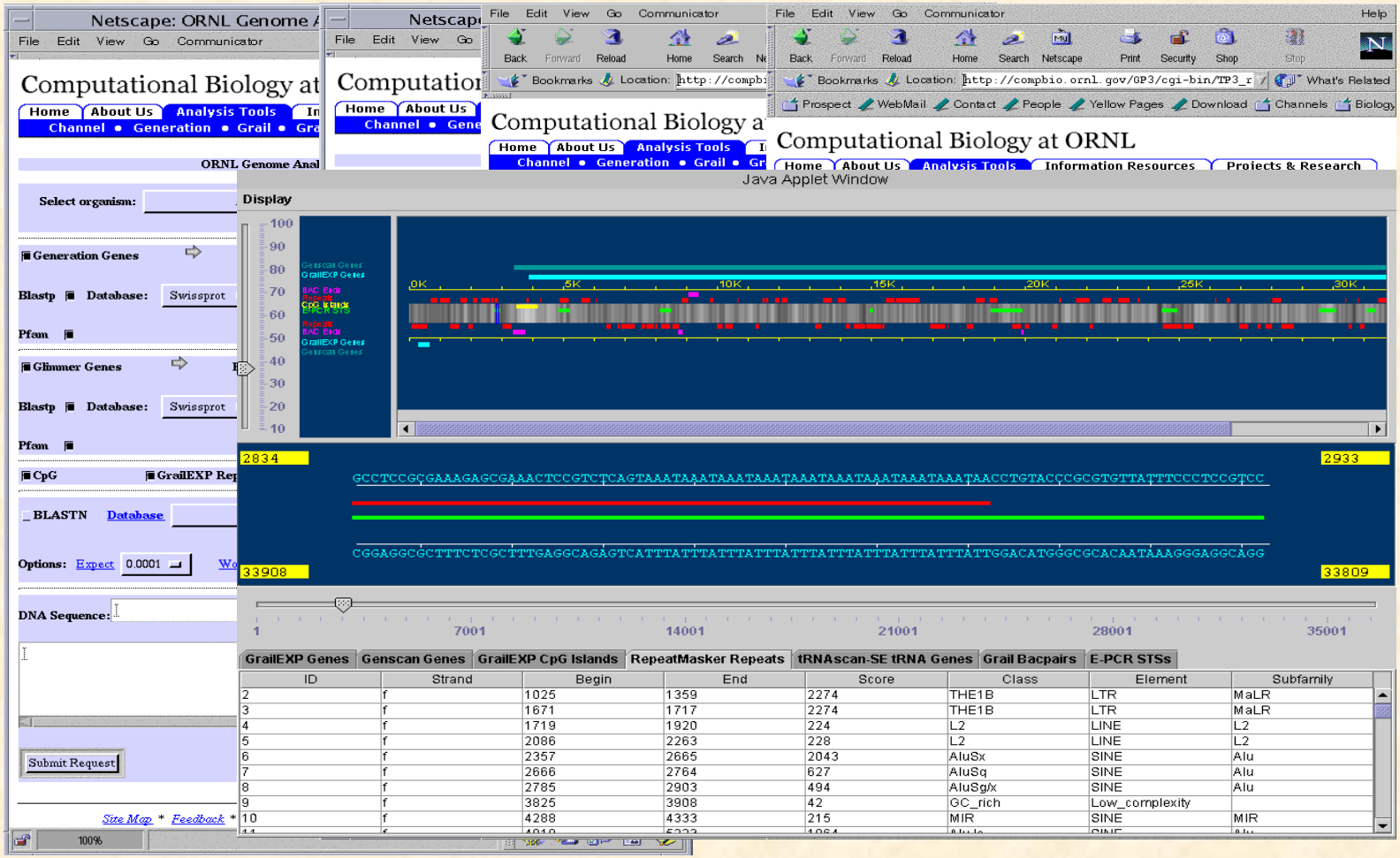
Analysis tools in action



Analysis tools in action

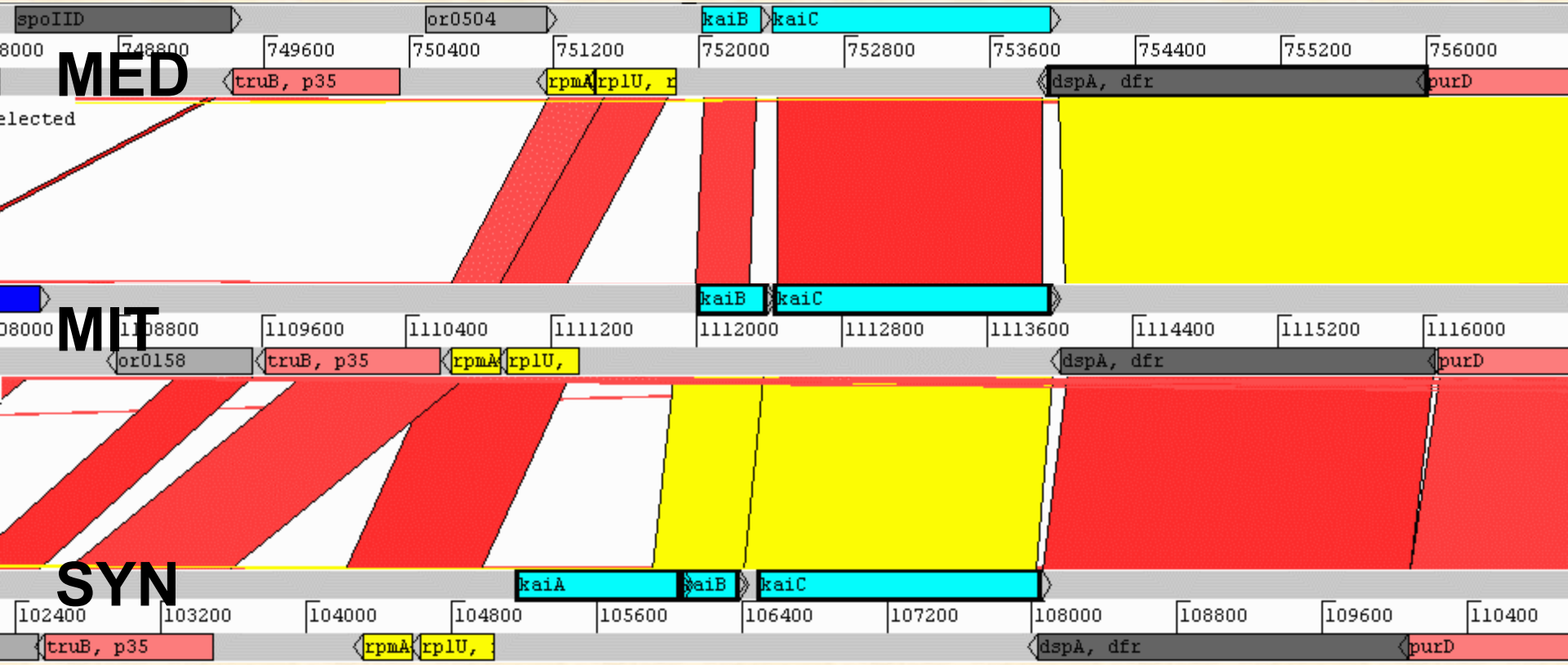


Analysis tools in action



Organization of Phycoerythrin (*cpe*) and Phycocyanin (*cpc*) Clusters

- ***P. marinus*** **MED4**
  - ***cpe*****B only**
- ***P. marinus*** **MIT9313**
  - ***cpe*****A *cpe*****B, seven others in conserved cluster**
- ***Synechococcus***
  - **long *cpe*A *cpe*B cluster, multiple AB sets, inverted repeat, plus 13 others**

- ***Nostoc***
  - **3 sets of *cpc*B *cpc*A clusters**
- ***Synechocystis***
  - **1 *cpc*A *cpc*B, 1 *apc*A *apc*B, solo *apc*B**

*kai*ABC clock operon