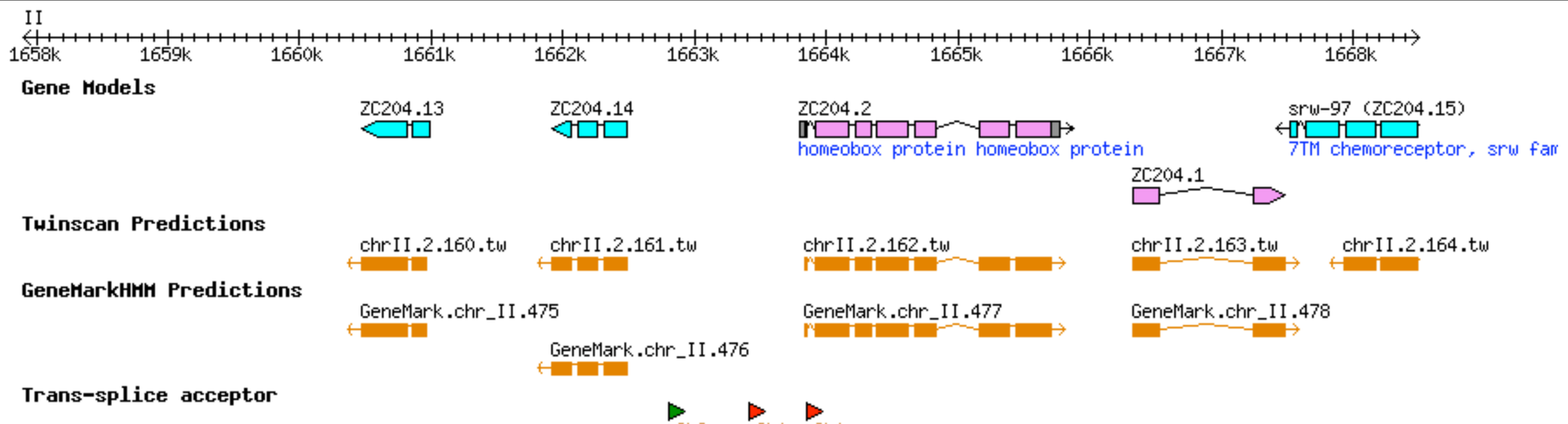
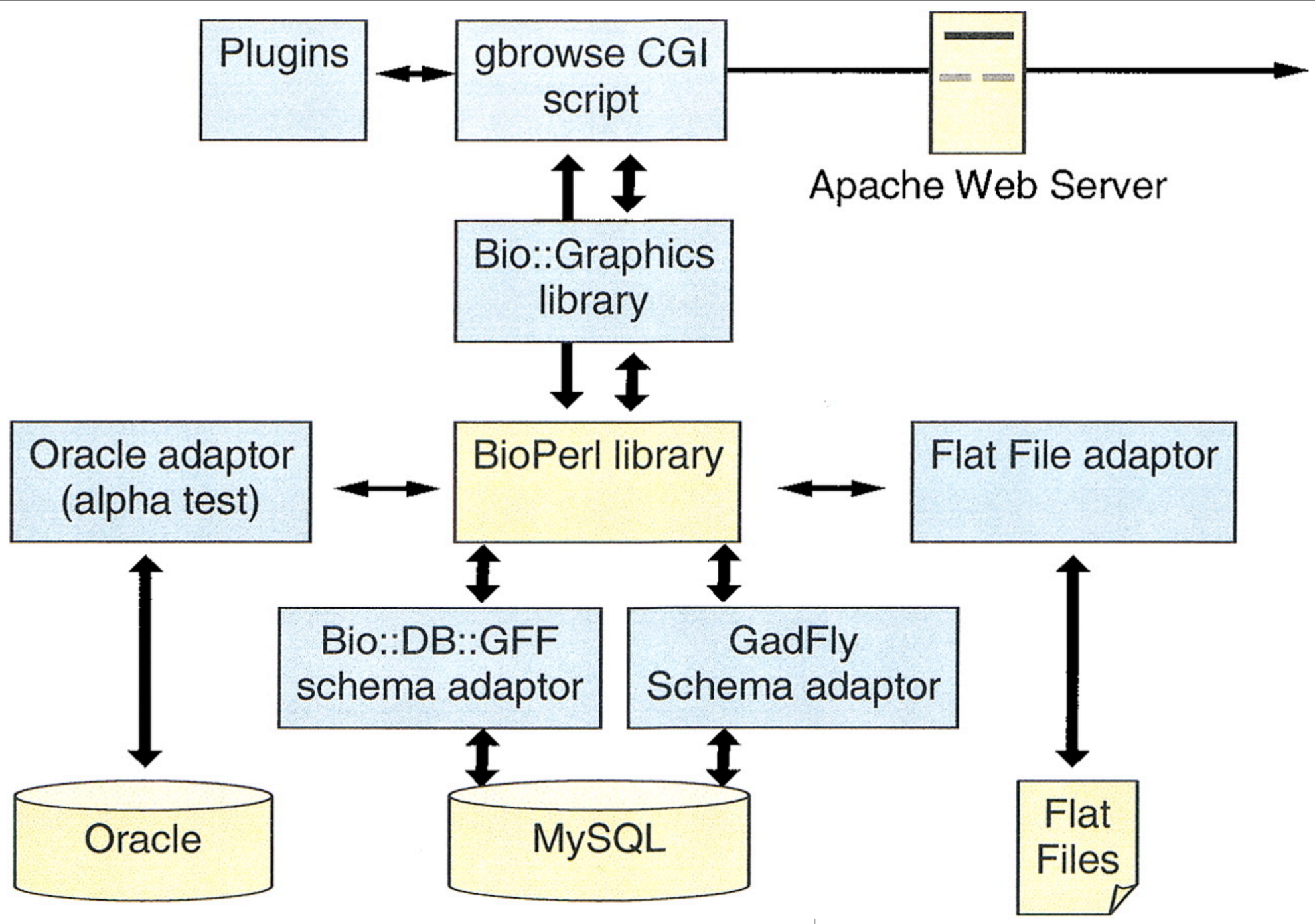
Gbrowse Workshop

Jason Stajich
University of California, Berkeley
&
University of California, Riverside (July 2009)



Genome Browser

wormbase



Gbrowse. Stein et al. Genome Res 2002

Data formats

- Simple tabular formats for genome annotation data
- GFF is the main format for Gbrowse databases
- Tabular format for

```
contig101 GeneMark   gene    4718 5134  .   -   .   ID=gene000000;Name=9397_g
contig101 GeneMark   mRNA    4718 5134  .   -   .   ID=mRNA000000;Parent=gene000000;Name=9397_t
contig101 GeneMark   CDS     4718 5134  .   -   0   ID=exon000000;Parent=mRNA000000
contig101 GeneMark   exon    4718 5134  .   -   0   ID=exon000001;Parent=mRNA000000
contig101 GeneMark   CDS     4718 5134  .   -   0   ID=cds000000;Parent=mRNA000000
```

- FASTA is the sequence format

```
>contig101
ATTCCAATATAAGGAGTTTATTTTAGTCTACCGGCTATATAAAATTTATA
AAAGTAGGTTATTTAAATTGGATTTATTCTTATAAATATAGTCTATAGTA
ATTAAAGGTTTCTTTTACTAAAGTAGTTAAATTGTTAGTACTATAATTAT
ATAATTTTATTAGATAGATAGTAAACCTTAAATATAATTATTATAGTATA
GGGTTATTTAATTTAATCGAAGAAGTATAATCCTCTAACTTCTATACTAG
```

Dense numerical data

- Per-base information
  - Per-base PhastCons conservation or % identity
  - Microarray data for all probes
  - Sliding window calculations
- Next-Gen High Throughput Sequence
  - ChIP-Seq, RNA-Seq, smallRNA-Seq
  - Resequencing data

High density data formats

- UCSC format BED and Wiggle formats

- BED format for alignments -- UCSC's own GFF flavor

  - http://genome.ucsc.edu/FAQ/FAQformat

```
track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

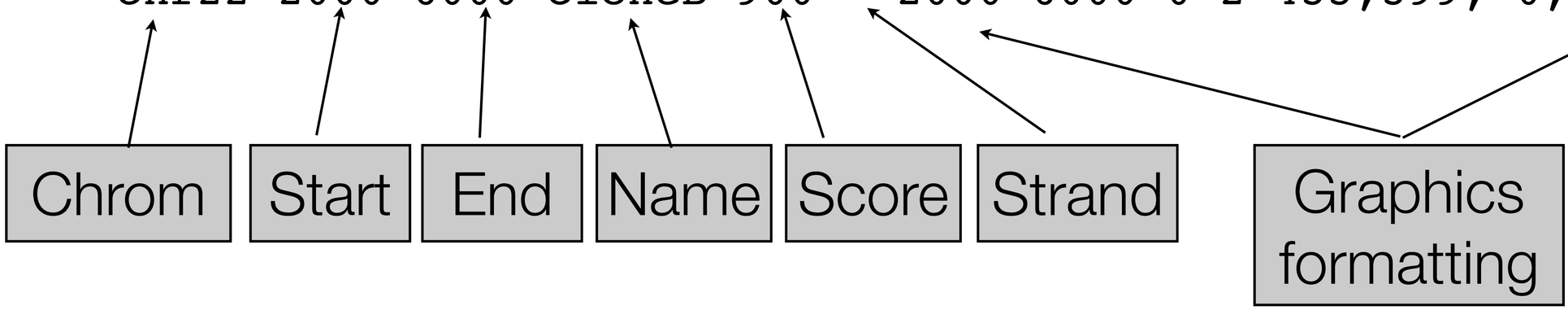
High density data formats

- UCSC format BED and Wiggle formats

- BED format for alignments -- UCSC's own GFF flavor

  - http://genome.ucsc.edu/FAQ/FAQformat

```
track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```



High density data formats

- UCSC format BED and Wiggle formats

- BED format for alignments -- UCSC's own GFF flavor

  - http://genome.ucsc.edu/FAQ/FAQformat

```
track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

High density formats

- Wiggle Track Format (WIG) is per-base format that can be efficiently indexed

- ```
  track type=wiggle_0 name="variableStep" description="variableStep format"
  variableStep chrom=chr19 span=150
  59304701 10.0
  59304901 12.5
  59305401 15.0
  59305601 17.5
  ```

- ```
  track type=wiggle_0 name="fixedStep" description="fixedStep format"
  variableStep chr19 start=59307401 step=1 span=500
  91
  87
  21
  20
  20
  ```

Neurospora_crassa_Linkage_Group_IV

1980k 1981k 1982k 1983k 1984k 1985k 1986k 1987k 1988k 1989k 1990k 1991k 1992k 1993k 1994k 1995k 1996k 1997k 1998k 1999k 2000k 2001k 200

DNA

% gc

NCU Ids

NCU07710 NCU07709 NCU10542 NCU07708 NCU07706 NCU07707

Named Genes (Radford laboratory)

trm-32

GeneMark+PASA updated

NCU07710 NCU07707 NCU07706 NCU07709

PhastCons (transcript-Intron)

phastcons

MUMMER+PECAN N.tetrasperma

MUMMER_PECAN_Nt

MUMMER+PECAN N.discreta

MUMMER_PECAN_Nd

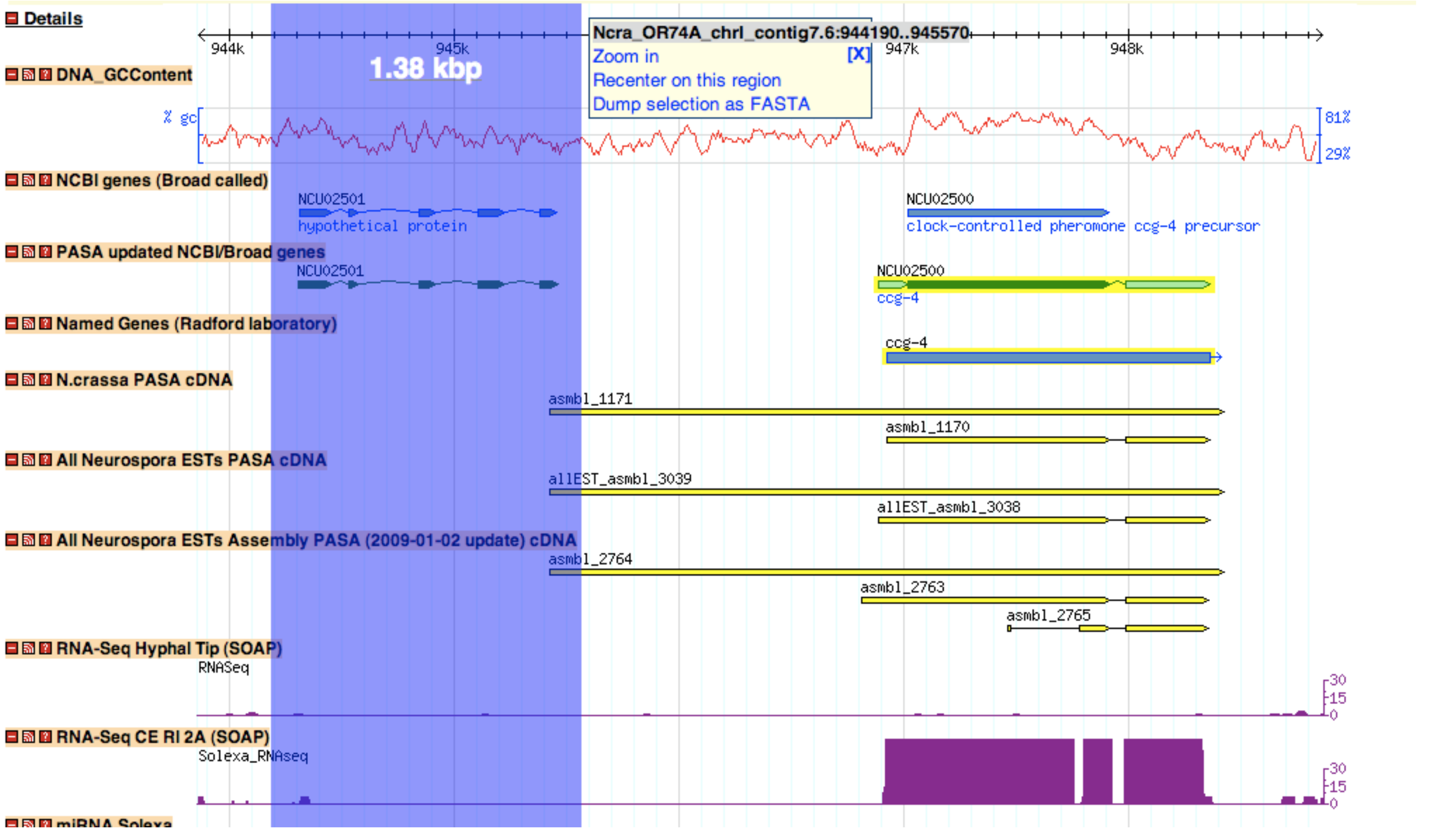
Data Loading

- Data is loaded into the mysql database
    - Scripts are bp_seqfeature_load.pl
- Genome Browser view is controlled by a configuration file

Genome Browser interaction

- Clicking on the browser to recenter
- Using the “rubber banding” to zoom in
- Using the “rubber banding” to select sequence
- Popup views of genome data

Rubber Banding



Ncra_OR74A_chrI_contig7.6:944190..945570

>Ncra_OR74A_chrI_contig7.6:944190..945570
acctgatatcatgtcggctttgcatagtacgattggtacgtgaggcatatggccggagag
ctcggtcgttgggtcccgttcggatcggcgcctgacacgttcaaaccatcaagacaatgc
cgggtgtgctaccgagtgttcgtcaaggggttctgacgctgtggtgccgggcatctttg
ctgagattttgagccgatcaggcgaggccagcggttgcgggcgaacccgaatggtcatcg
cgggcgcgttgaaaaacccgagtgaagtgaaaggccggaatccggtcgttggtccattat
acacatcagtcctggtaagttcgttagcttggtacctacagacaggggttttaagcaggg
gagatggggaatgttacagcccgctgtaagtaccctgttattaccctataattataccgc
ttatagagggtgcgctgggtaattctagttgatttagttagaattcgagggtttgtatgt
gggtggttgaatgcctcctatcaaccgattctcactttgaattgaaccaatcagagctcg
tagatagacctatcccgccttgcagccccccctgttacccccttgtgagaagatgtgggg
cgtgtatggcgctaacaaacttttaaggactgatgtgtaaatgccgttgaaaagctttgg
ggtcccgggccgcttagttgggatgacgcactgtctgcagcccgcttccatgttgcagat
tttagtgtacgtactgtgtaccttactagttcagggtccgaccactgagccagccagccc
tatcaggccgggcagtataatcagcgttgctagatacggagcgctcctcgcagcttctga
agactgataactgacctggtgctggtttgccgtagaaattgttagtacgtgagtgctctg
gatccacctgtacagtgtaaaggaaatccacggcactttcacaatggaaactttgcatgg
cgcctcggagctgcccgaagtcctctcgagacctgatgggaatactgtaccggtaaaccg
aataccatatgcttggtgagttgccagacgtgcagaacccgaccaaccgcattttataat
gatctgcgaacactccagtcctatgaatcacattgatcgatctgggctgctggcccaaat
ttcacttccgcaacaccaccatcgacaggactgcaacccatcacaactcgcagtatcaat
accgaagactttaccagcggagcatctgctatacttcgctgcaattgcgaacgaaacatg
ttcaactgaaacataacccatcacgaatttcccgggtaaattaaccacagaaagagcacc
agttgttcgtggaaaaccccccaagttgaactgaacagtaactcacgtctcgagagtgagc

Details

947.1k 947.2k 947.3k 947.4k 947.5k 947.6k 947.7k 947.8k 947.9k 948k 948.1k 948.2k 948.3k 948.4k 948.5k

DNA_GCContent

% gc

88
6%

NCBI genes (Broad called)

NCU02500

clock-controlled pheromone ccg-4 precursor

PASA updated NCBI/Broad genes

NCU02500

ccg-4

Named Genes (Radford laboratory)

ccg-4

N.crassa PASA cDNA

asmbl_1171

asmbl_1170

All Neurospora ESTs PASA cDNA

allEST_asmbl_3039

allEST_asmbl_3038

All Neurospora ESTs Assembly PASA (2009-01-02 update) cDNA

asmbl_2764

asmbl_2763

asmbl_2765

RNA-Seq Hyphal Tip (SOAP)

NCU02500

clock-controlled pheromone ccg-4 precursor

MIPS Functional Categories

- 34 INTERACTION WITH THE ENVIRONMENT

Gbrowse details

Name: NCU02500
Type: gene
Description:
Source: NCBI_PASA_allnames
Position: Ncra_OR74A_chrI_contig7.6:946888..948363 (+ strand)
Length: 1476
Alias: ccg-4
load_id: pasa_gene007463

Parts:

- **Type:** mRNA
  Description:
  Source: NCBI_PASA_allnames
  Position: Ncra_OR74A_chrI_contig7.6:946888..948363 (+ strand)
  Length: 1476
  Alias: ccg-4.T0
  load_id: pasa_mrna007189
  parent_id: pasa_gene007463

  Parts:

  - **Type:** five_prime_utr
    Description:
    Source: NCBI_PASA_allnames
    Position: Ncra_OR74A_chrI_contig7.6:946888..947016 (+ strand)
    Length: 129
    load_id: pasa_utr5p_of_mrna007189
    parent_id: pasa_mrna007189
  - **Type:** exon
    Description:
    Source: NCBI_PASA_allnames
    Position: Ncra_OR74A_chrI_contig7.6:946888..947916 (+ strand)
    Length: 1029
    load_id: pasa_mrna007189.exon1
    parent_id: pasa_mrna007189
  - **Type:** CDS
    Description:
    Source: NCBI_PASA_allnames
    Position: Ncra_OR74A_chrI_contig7.6:947017..947916 (+ strand)
    Length: 900
    load_id: pasa_mrna007189.cds1
    parent_id: pasa_mrna007189
  - **Type:** exon
    Description:
    Source: NCBI_PASA_allnames
    Position: Ncra_OR74A_chrI_contig7.6:947986..948363 (+ strand)
    Length: 378
    load_id: pasa_mrna007189.exon2
    parent_id: pasa_mrna007189
  - **Type:** three_prime_utr
    Description:
    Source: NCBI_PASA_allnames
    Position: Ncra_OR74A_chrI_contig7.6:947986..948363 (+ strand)
    Length: 378
    load_id: pasa_utr3p_of_mrna007189
    parent_id: pasa_mrna007189

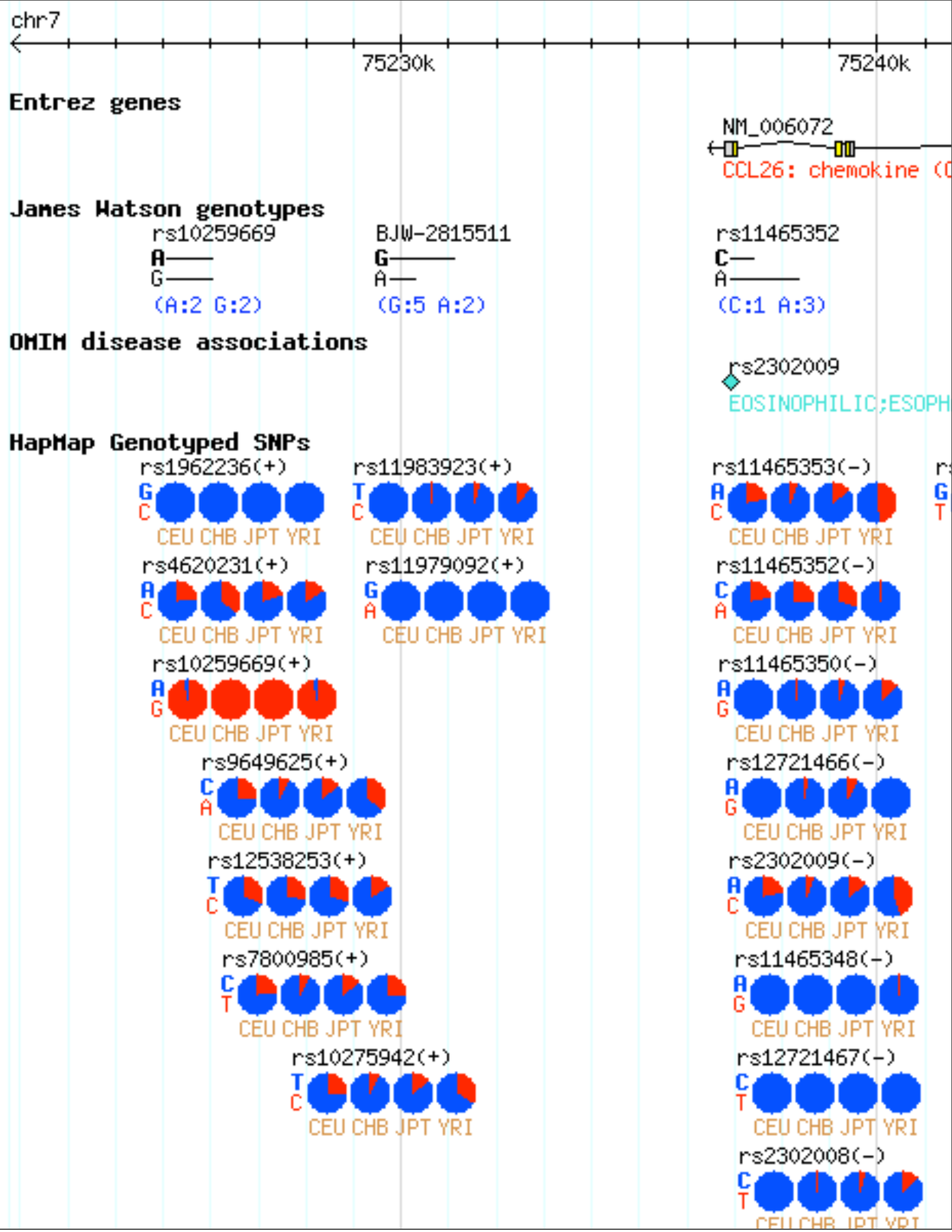
>NCU02500.T0 class=Sequence position=Ncra_OR74A_chrI_contig7.6:946888..948363 (+ strand)
AACACACTTC TTTTTTTCTC CATCACCTTT GACATTGCCA ATCAACCCTC AGAGGTCTTC ATTCTCTCAA TCAACAGGGT
CCTTTCGTTG ACACTTTTTA CATTCTTCAT CCAAGCCGTT TTGTTCAAGA TGAAGTTCAC TCTCCCTCTT GTCATCTTCG
CCGCCGTGGC CTCCGCCACC CCGGTCGCCC AGCCAAATGC CGAGGCCGAA GCCCAGTGGT GCCGGATCCA TGGCCAGTCC
TGCTGGAAGG TCAAGCGTGT TGCCGATGCC TTCGCCAACG CCATCCAGGG CATGGGTGGT CTCCCTCCCC GCGATGAGTC

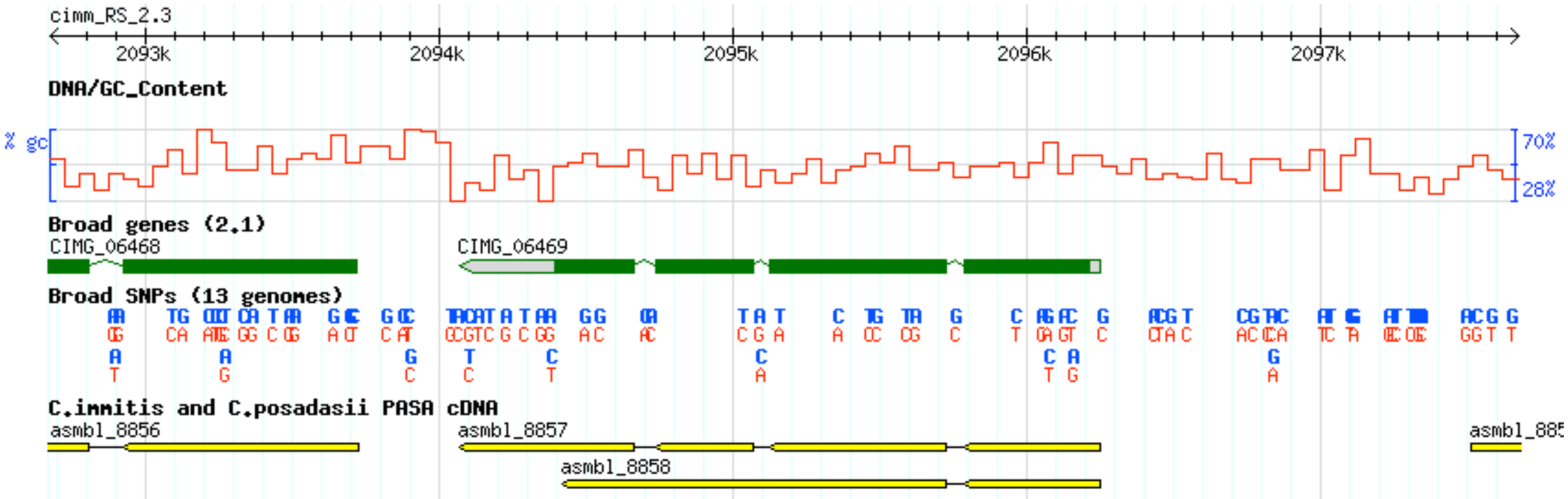
Semantic zooming

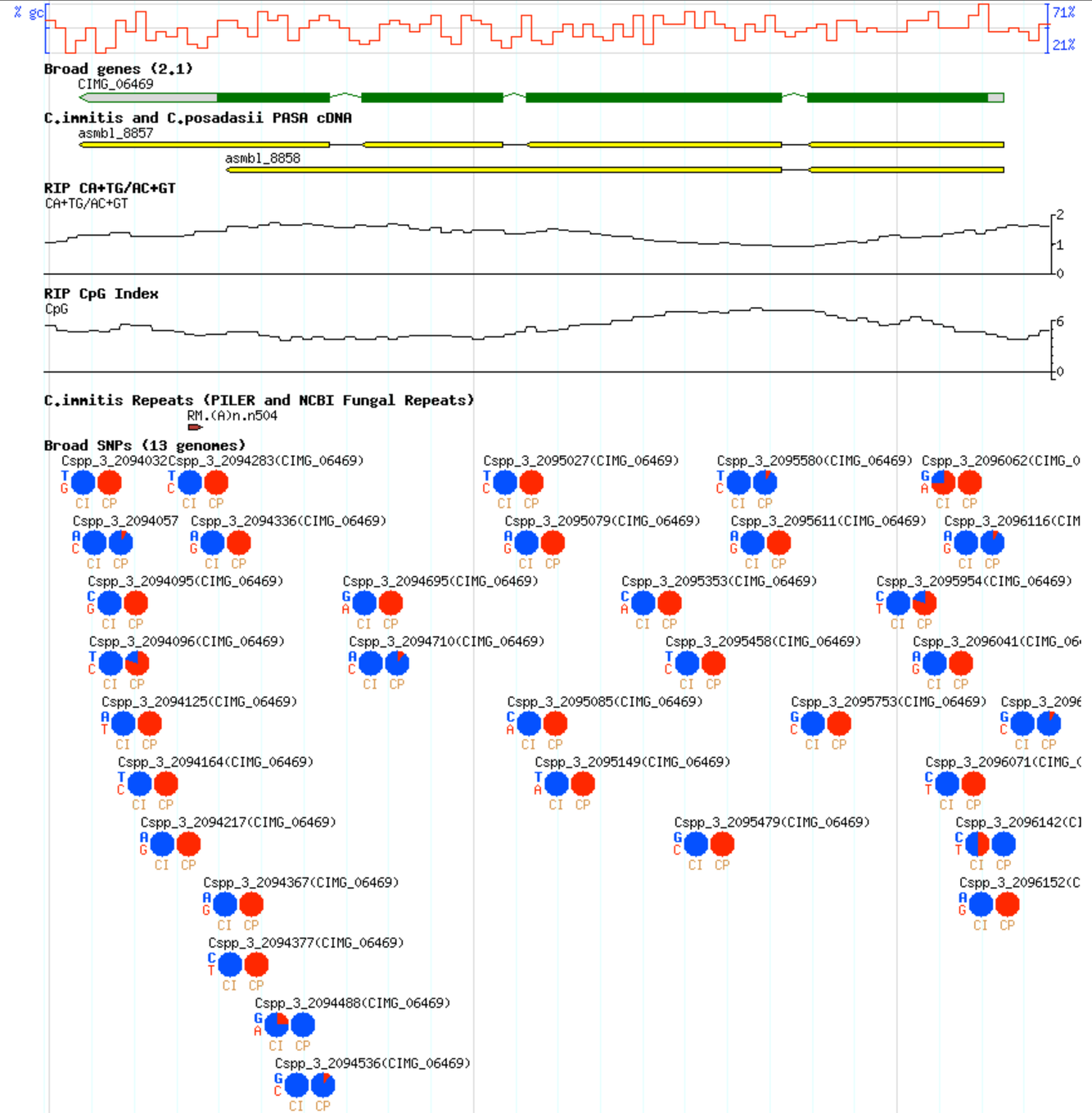
- Displaying glyphs depending on the zoom level
- Useful for very detailed views being simplified at larger zoom level
- Use different glyphs, E.g.
  - Draw gene with transcript information and splicing at zoomed-in but draw as an arrow alone when zoomed out
  - At close-magnification draw detailed per-site data but at zoomed out draw a bar-chart summary of expression or sub-sampled gene

Polymorphism data

- Displaying position of SNPs
- Displaying Allele frequencies as pie-charts
- Comparing genotypes between the reference and other populations




cimm_RS_2.3
2093k
2094k
2095k
2096k
2097k
DNA/GC_Content
% gc
70%
28%
Broad genes (2.1)
CIMG_06468
CIMG_06469
Broad SNPs (13 genomes)
C.immitis and C.posadasii PASA cDNA
asmbl_8856
asmbl_8857
asmbl_8858



Polymorphism data

Polymorphism data

Other tips

- Can construct URL links directly into a region
  http://YOURBROWSER/cgi-bin/gbrowse/DATABASE?name=CHROM:START..STOP

  http://fungalgenomes.org/gb/gbrowse/gb_neurospora_crassa_OR74A_7?name=Ncra_OR74A_chrV_contig7.41:1000..5000

- http://fungalgenomes.org/gb/gbrowse/gb_neurospora_crassa_OR74A_7?name=ccg-4

- http://modencode.org/cgi-bin/gbrowse/fly?name=boss

- http://modencode.org/cgi-bin/gbrowse/fly?name=dscam

Gbrowse Sites

- Human
  - Hapmap: http://hapmap.org/cgi-perl/gbrowse
  - Jim Watson: http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/
- C.elegans
  - Wormbase: http://wormbase.org/db/seq/gbrowse/c_elegans/
  - ModEncode: http://modencode.oicr.on.ca/cgi-bin/gbrowse/worm/
- Yeast
  - SGD: http://www.yeastgenome.org/cgi-bin/gbrowse/scgenome/
- Fly
  - Flybase: http://flybase.org/cgi-bin/gbrowse/dmel/
  - Modencode: http://www.modencode.org/cgi-bin/gbrowse/fly/
- Fungi - Neurospora
  - Fungalgenomes http://fungalgenomes.org/gb/gbrowse/neurospora_crassa_OR74A_7/

- Use a Gbrowse genome browser address the following questions

1. Extract the sequence for a promoter of a gene

    1. Zoom into the region around the gene

    2.Look at the upstream region

    3.Use the rubber-band to select the region of sequence and dump as FASTA

2.Extract the centromeric sequence for CEN1 from *Saccharomyces* (http://www.yeastgenome.org/cgi-bin/gbrowse/scgenome/)

1. Try finding what Jim Watson's alleles are for the PARK3 (parkin)

2. How many isoforms are there for DSCAM in the modencode browser for Drosophila?

3. At Wormbase - what protein coding genes is let-7 (miRNA) near or within?

4. At SGD (http://www.yeastgenome.org/cgi-bin/gbrowse/scgenome/) - the snoRNA snR39 is encoded in an intron of what gene? What is the function of the enclosing gene?