Sponsored by

Database Tools for Evolutionary Genomics

An introduction to GMOD software for managing, annotating, and visualizing genomic data

Annual Meeting of the Society for Molecular Biology and Evolution (SMBE 2009)

June 6, 2009

Agenda

2:30 Dave Clements, National Evolutionary Synthesis Center

Using GMOD for evolutionary genomics and next generation sequence data

3:10 Ben Faga, University of Iowa

The CMap comparative map browser and displaying population distributions with GBrowse and PhyloGeoViz

3:45 Sheldon McKay, Cold Spring Harbor Laboratory

Comparative genomics with GBrowse_syn

4:20 Mark Yandell, University of Utah

Simplifying genome annotation and functional genomics with MAKER – the easy-to-use genome annotation pipeline

Sponsored by

Using GMOD for Evolutionary Genomics and Next Generation Sequence Data

Dave Clements
GMOD Help Desk
National Evolutionary Synthesis Center
clements@nescent.org

Annual Meeting of the Society for Molecular Biology and Evolution (SMBE 2009)

June 6, 2009

Overview

- GMOD Project Overview
  - Software
  - Community
- Visualizing Next Generation Sequence in GBrowse
  - SAMtools and GBrowse as a short read viewer
    - Whole genome resequencing of E. coli strains
  - GBrowse for population genetics
    - SNPs in threespine stickleback
  - Other visualizations
  - Next Generation Sequencing & Bioinformatics

GMOD

SMBE 2009

GMOD is …

- A set of interoperable open-source **software** components for visualizing, annotating and managing biological data.

- An active **community** of developers and users that are addressing common challenges with their biological data.
  - Mailing lists, meetings, support staff, wiki, …

http://gmod.org

GMOD Software: A whirlwind tour

Apollo

GBrowse & JBrowse

CMap

MAKER & DIYA

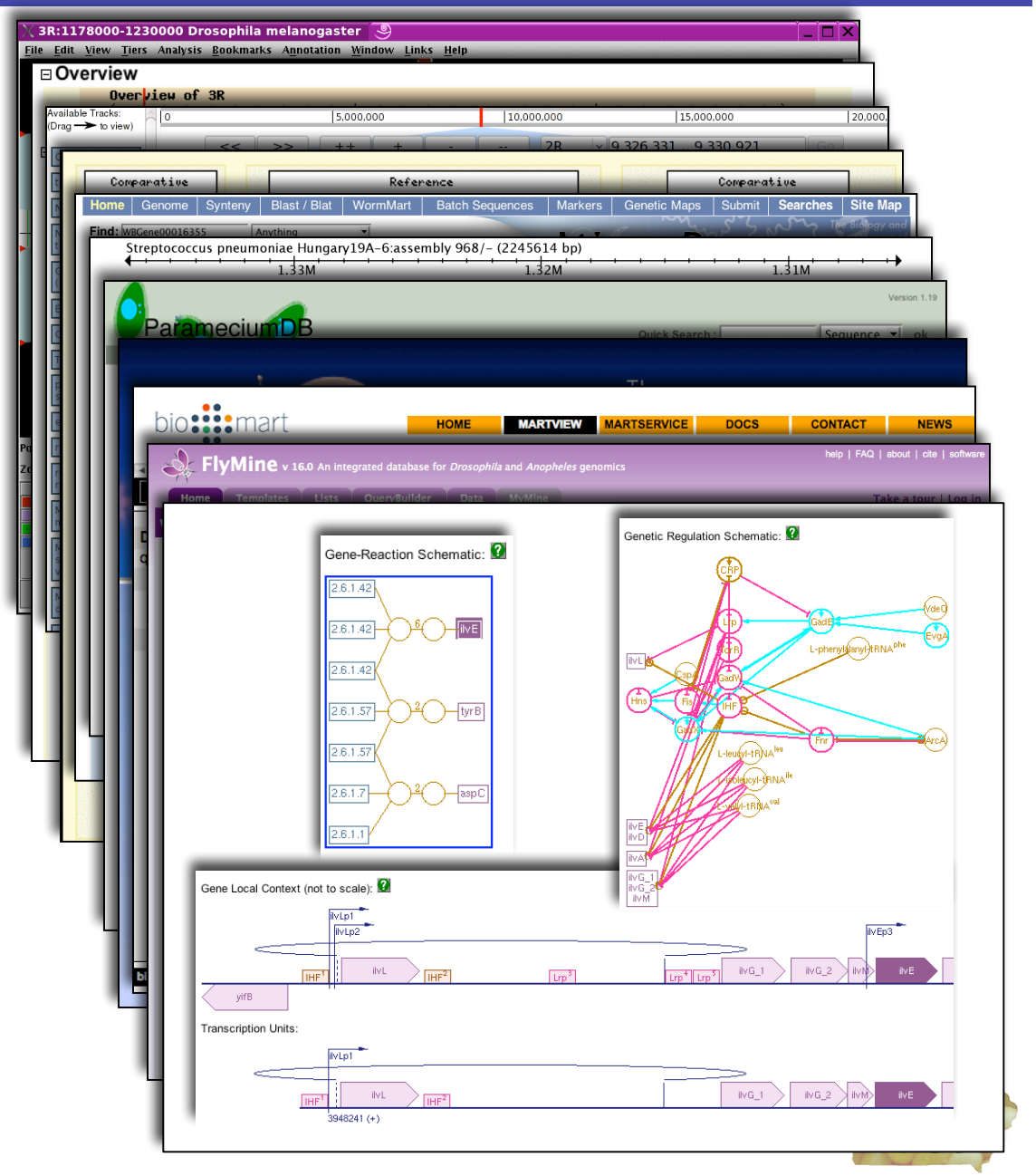
GBrowse_syn, Sybil, …

Chado, GMODWeb, Tripal

BioMart, InterMine

Pathway Tools

…



GMOD: Community

Next GMOD Meeting

University of Oxford, UK

6-7 August 2009

Part of GMOD Europe 2009

GMOD Courses

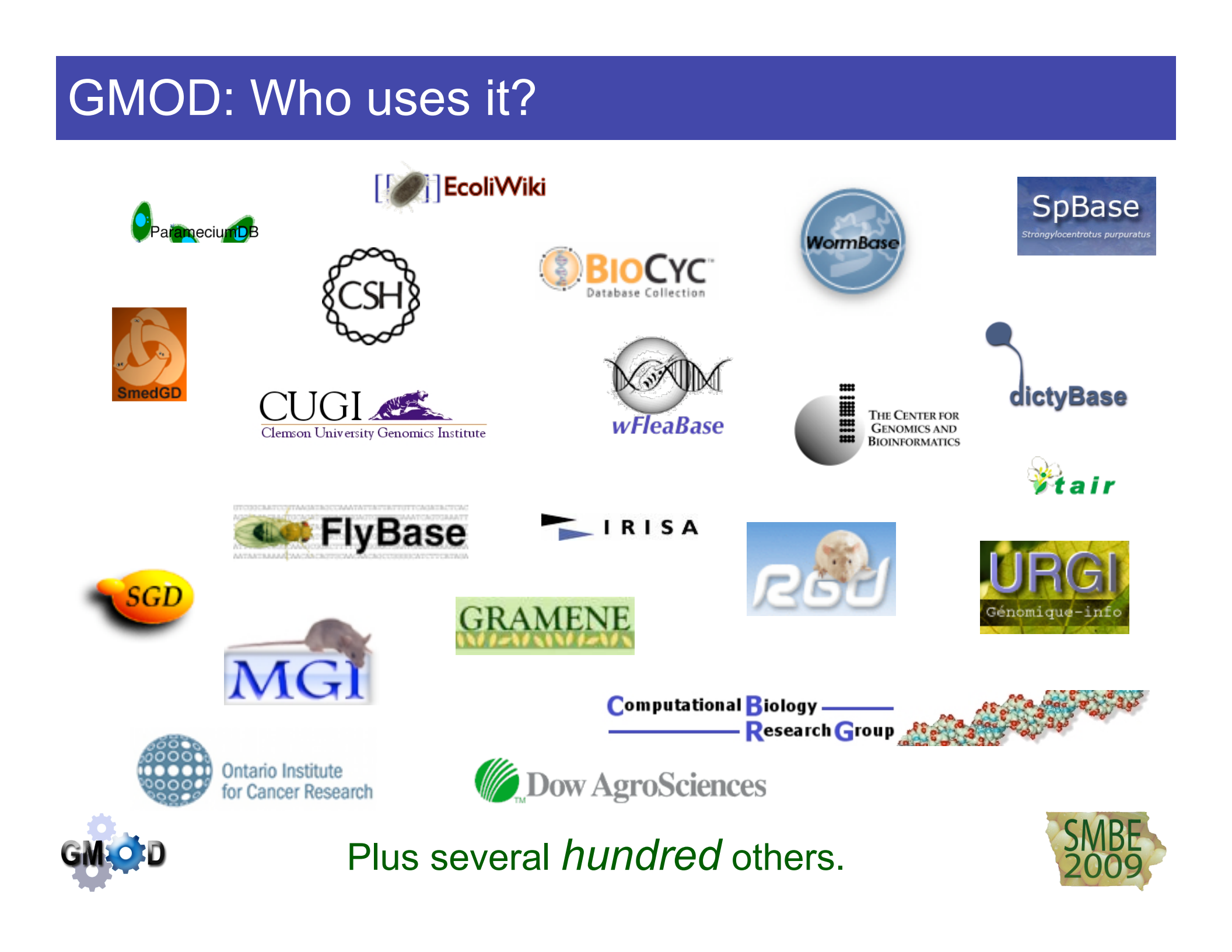
Multi-day, hands-on tutorials covering installation and configuration of GMOD components

Both 2009 courses are full

Offered again in 2010

Considering one in Asia/Pacific

GMOD: Who uses it?

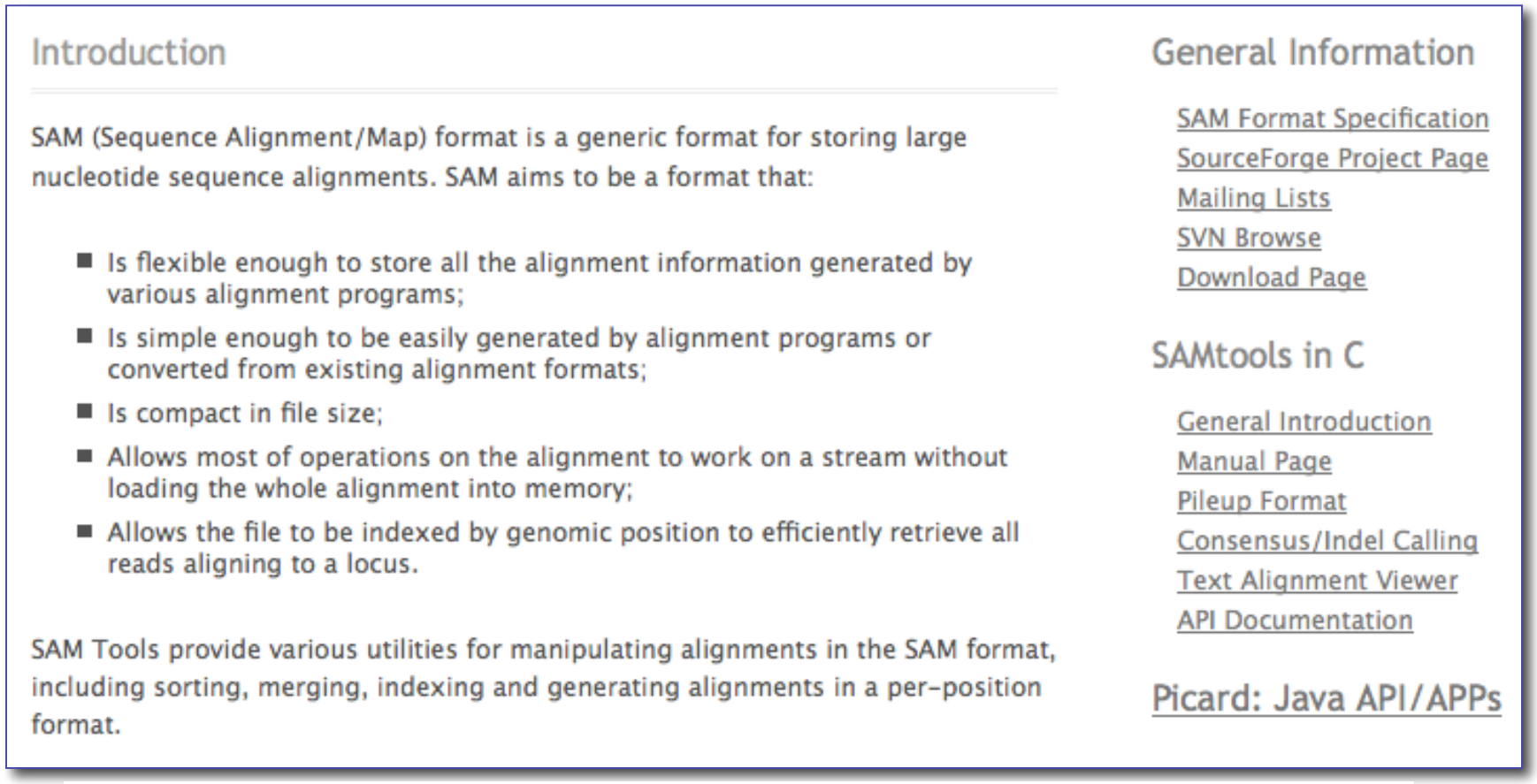


Plus several *hundred* others.

Overview

- GMOD Project Overview
  - Software
  - Community
- Visualizing Next Generation Sequence in GBrowse
  - SAMtools, and GBrowse as a short read viewer
    - Whole genome resequencing of E. coli strains
  - GBrowse for population genetics
    - SNPs in threespine stickleback
  - Other visualizations
  - Next Generation Sequencing & Bioinformatics

SAMtools



Introduction

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

General Information

SAM Format Specification
SourceForge Project Page
Mailing Lists
SVN Browse
Download Page

SAMtools in C

General Introduction
Manual Page
Pileup Format
Consensus/Indel Calling
Text Alignment Viewer
API Documentation

Picard: Java API/APPs

Heng Li, *et al*., http://samtools.sf.net

Platform neutral set of programs and file formats specifically for short reads.

GBrowse

GMOD's main genome browser

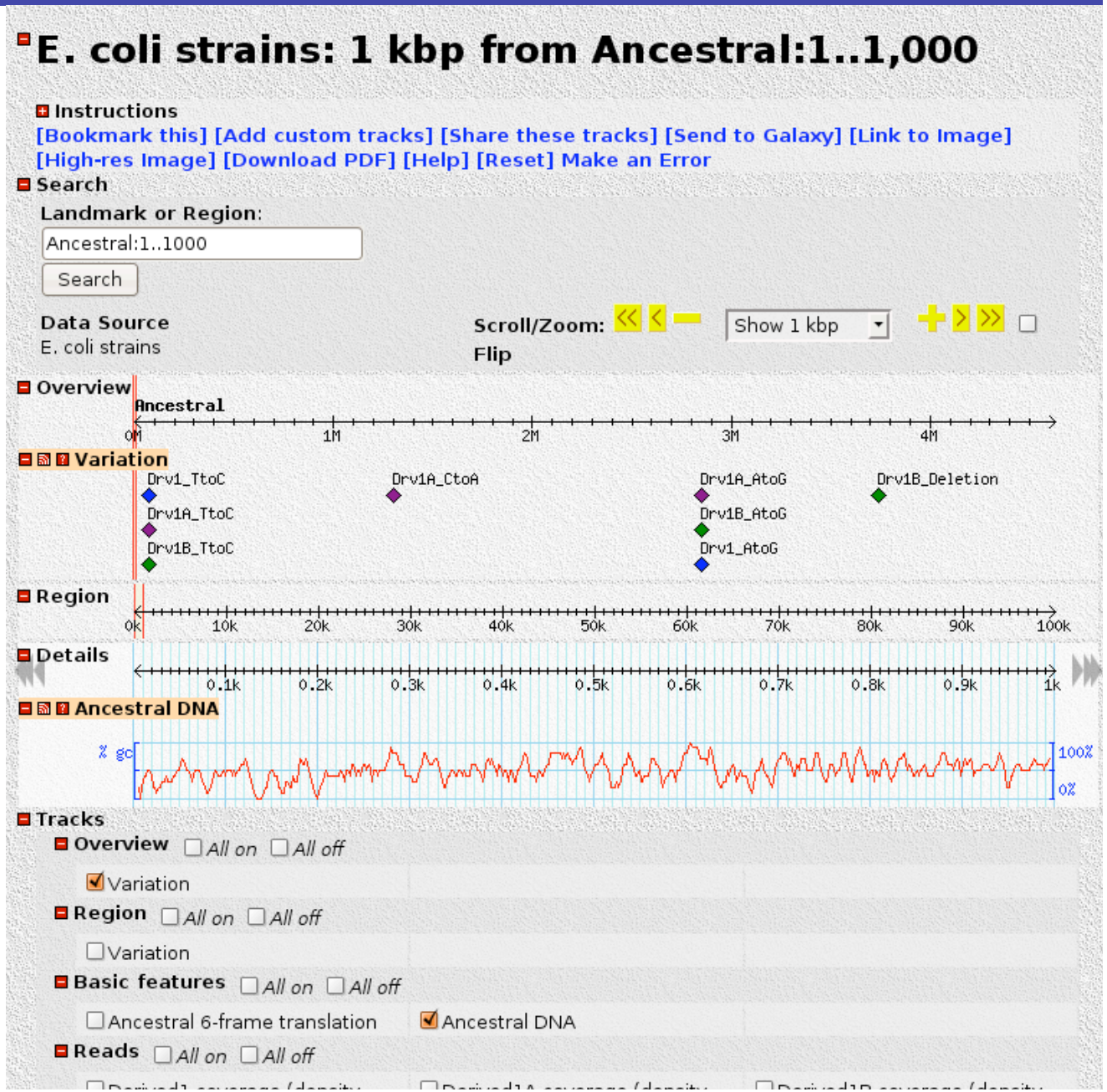
E. coli landing page

Overview: chromosome wide

Region: intermediate zoom

Details: current area

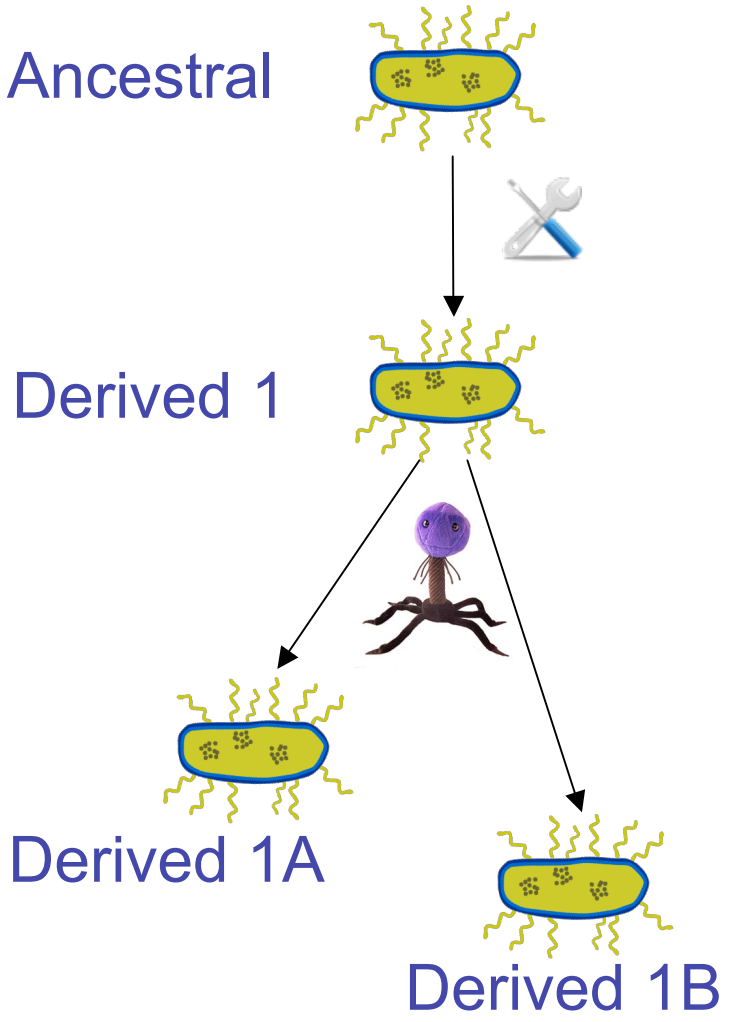
Tracks: current configuration



E. coli: Whole Genome Resequencing

Tale of 4 strains

- Ancestral
  - Reference
- Derived 1
  - Manipulated in two places (neutral, metabolic)
  - Exposed to phage yielding 2 resistant strains
- Derived 1A
  - 1bp change
- Derived 1B
  - 2-3kbp deletion



Brendan Bohannon, Liz Perry, Nick Stiffler, University of Oregon

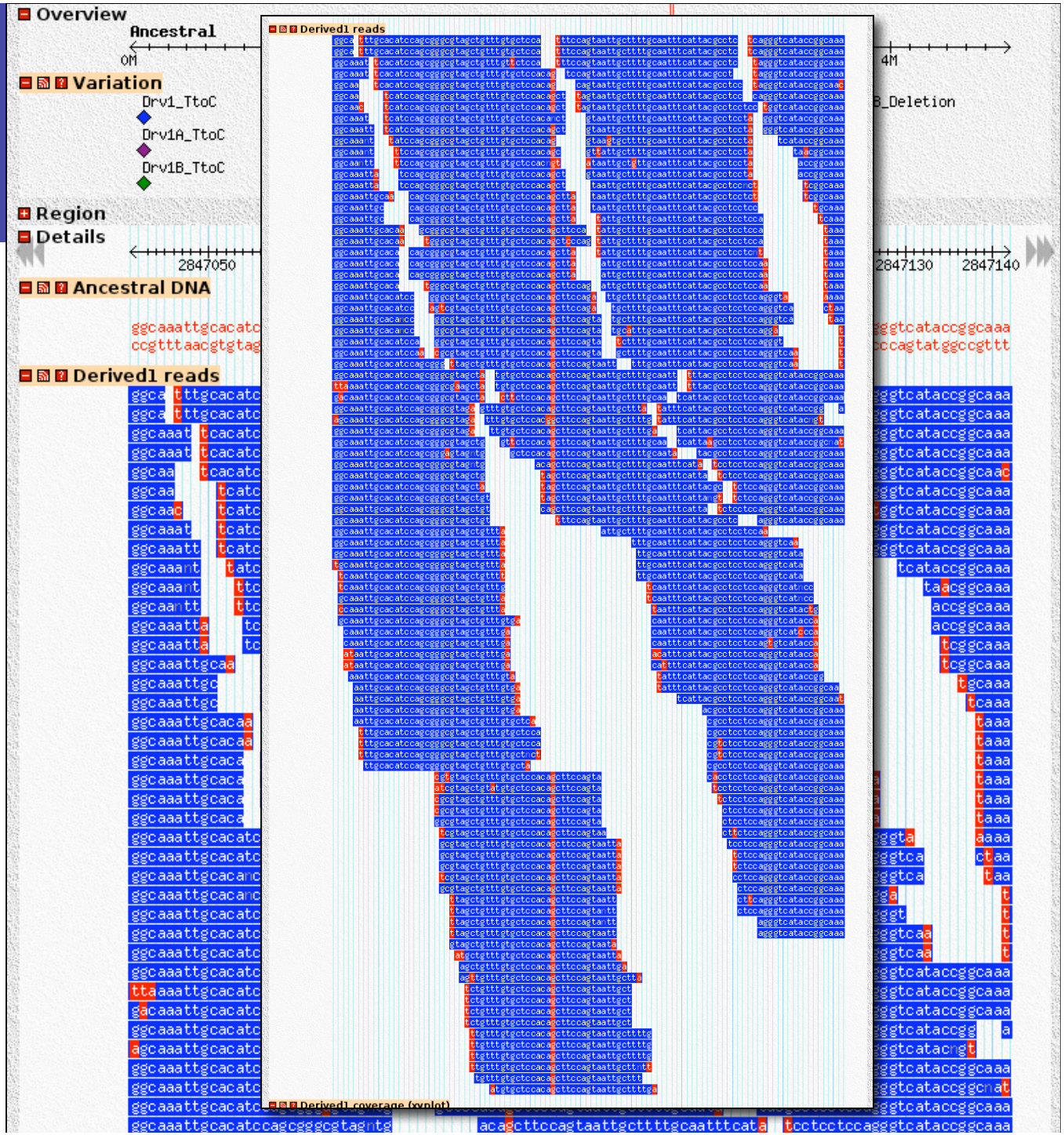
E. coli: Process

- Extract DNA from 3 derived strains
- Sonicate, aiming for 500bp fragments
- Unpaired end run on an Illumina GA2
- Filter results for quality
- Align with MAQ*
- Convert to BAM (SAMtools binary format)
- Visualize with GBrowse

* Li H., Ruan J. and Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res., 18 (11), 1851-1858. http://maq.sf.net

GBrowse as an Alignment Viewer

- High magnification view: 100bp
- Uses GBrowse 2 (Beta) and the Bio-SamTools GBrowse database adaptor (Alpha).

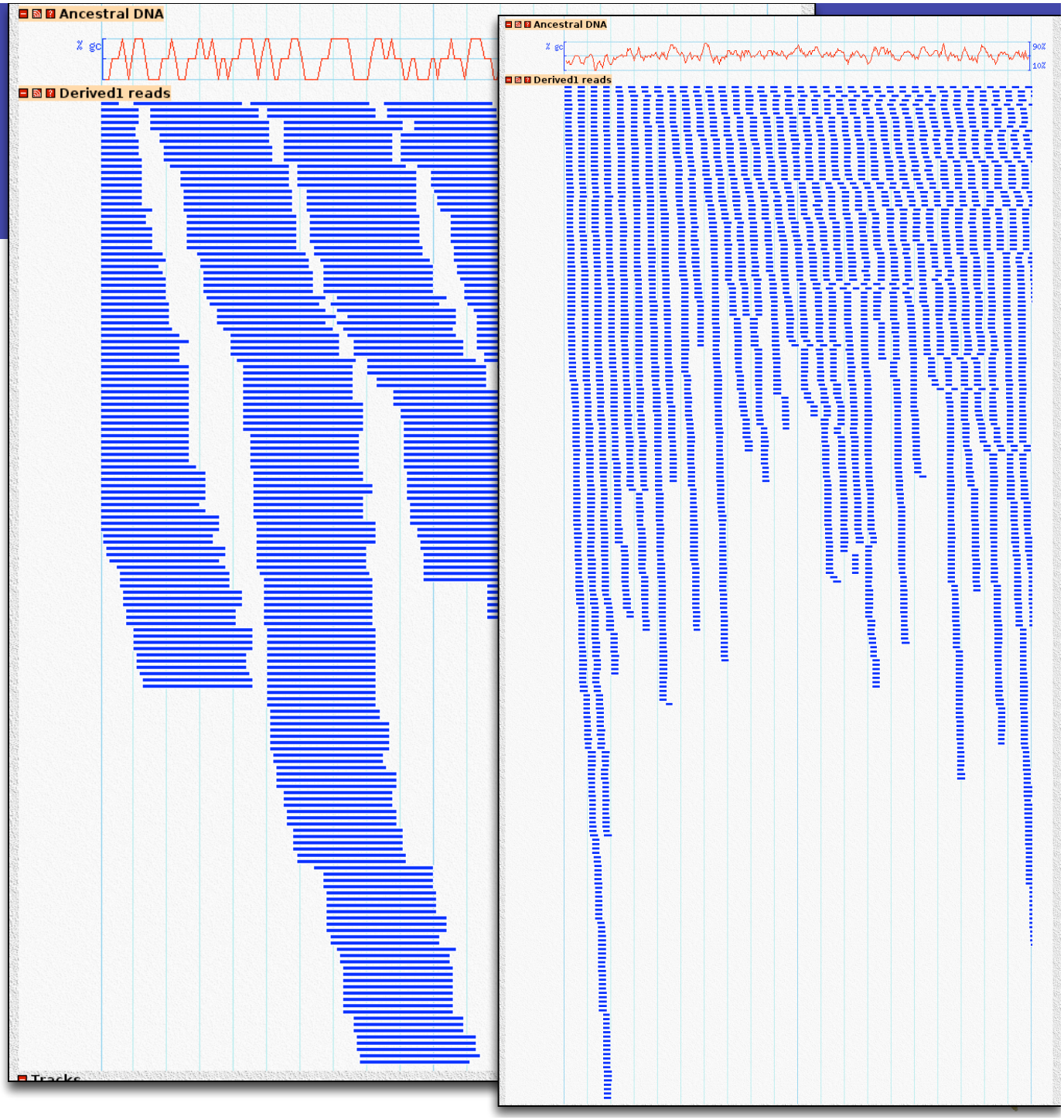


GBrowse as an Alignment Viewer

As you zoom out to 200bp you lose letters.

As you zoom out to 2000bp the view becomes much less useful.

SAMtools, GBrowse 2, & Bio-SamTools adaptor make this volume of data computationally tractable

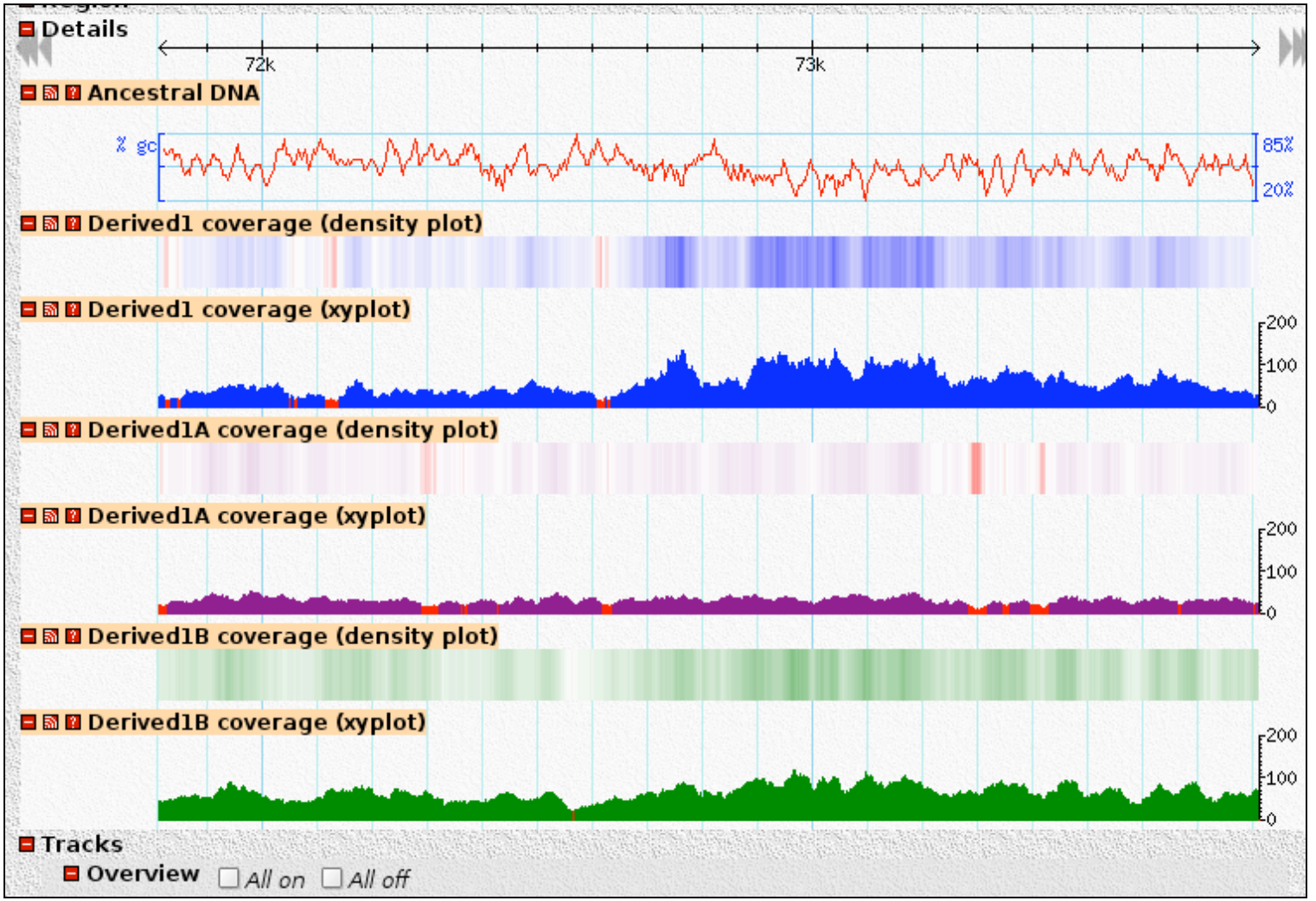


GBrowse as an Alignment Viewer

Summarize!

SAMtools and GBrowse help here

SAMtools can summarize several values from the data.



GMOD

SMBE 2009

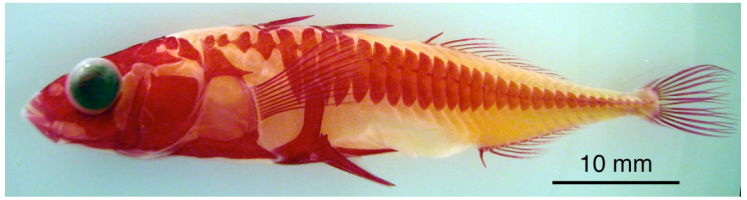
Overview

- GMOD Project Overview
  - Software
  - Community
- Visualizing Next Generation Sequence in GBrowse
  - SAMtools, and GBrowse as a short read viewer
    - Whole genome resequencing of E. coli strains
  - GBrowse for population genetics
    - SNPs in threespine stickleback
  - Other visualizations
- Next Generation Sequencing & Bioinformatics

GBrowse for Population Genetics

Threespine Stickleback

- Tale of 2 populations, 8 (or 12) fish from each
  - Rabbit Slough, marine
    - ancestral, reference
    - High body plating
  - Bearpaw Lake, freshwater
    - Diverged in last 10-15,000 years
    - Low body plating



- Pattern repeats all over northern hemisphere
- Deep sequencing around restriction sites
- Aiming to identify SNPs at a minimum density, genome wide

Bill Cresko, Paul Hohenlohe, and Nick Stiffler, all @ U Oregon

GBrowse for Population Genetics

Process for threespine stickleback

- Extract DNA from each fish
- Break it up with restriction enzymes.
- Apply RAD tags with bar code
- Do an unpaired-end run on an Illumina GA2
- Filter results for quality
- Align it with MAQ
- Make SNP calls
- Visualize it with GBrowse

GMOD

SMBE 2009

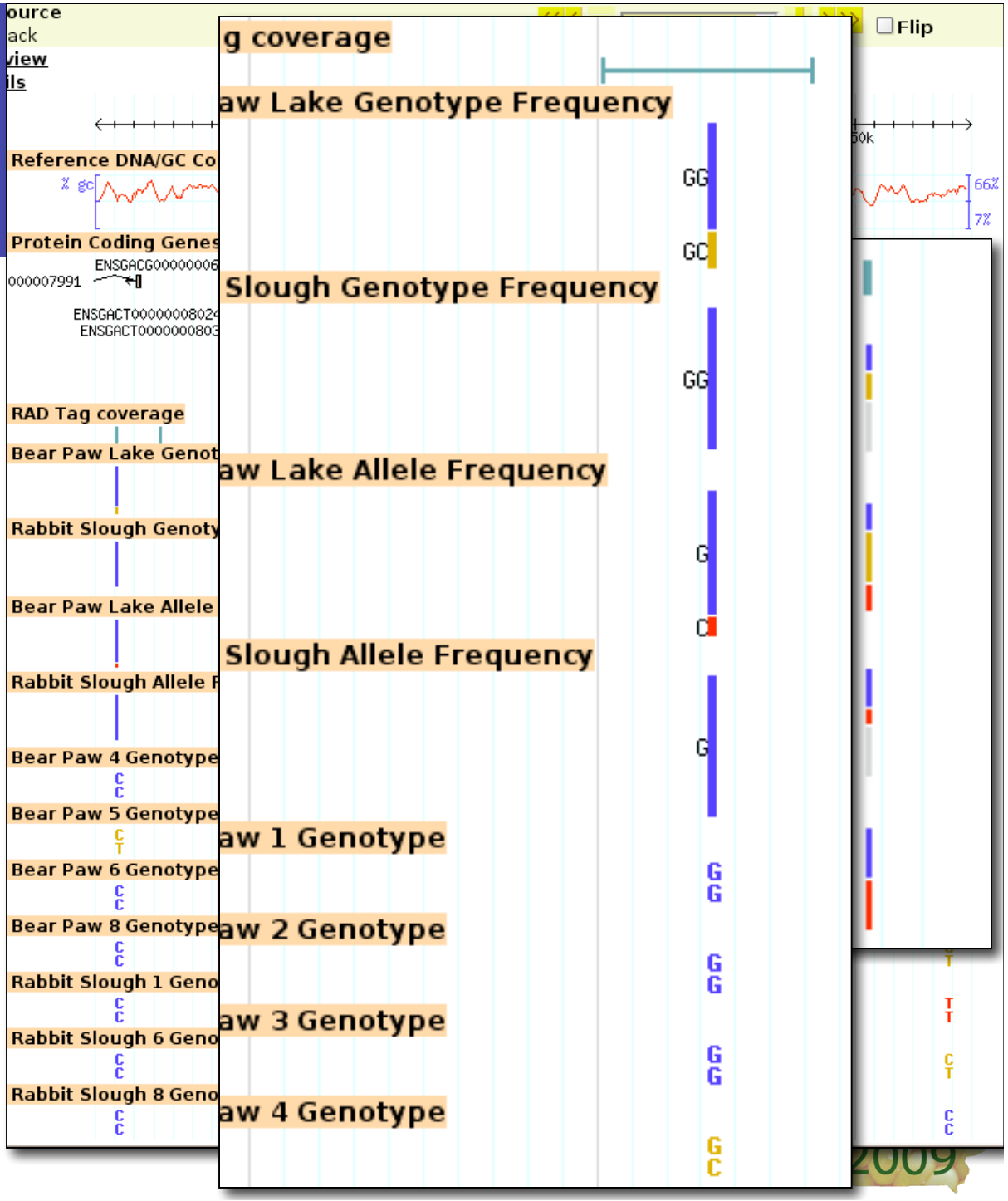
GBrowse for Population Genetics

Shows

- Where we looked
- Allele & genotype frequencies
  - By population
  - Individual genotypes

Could also show:

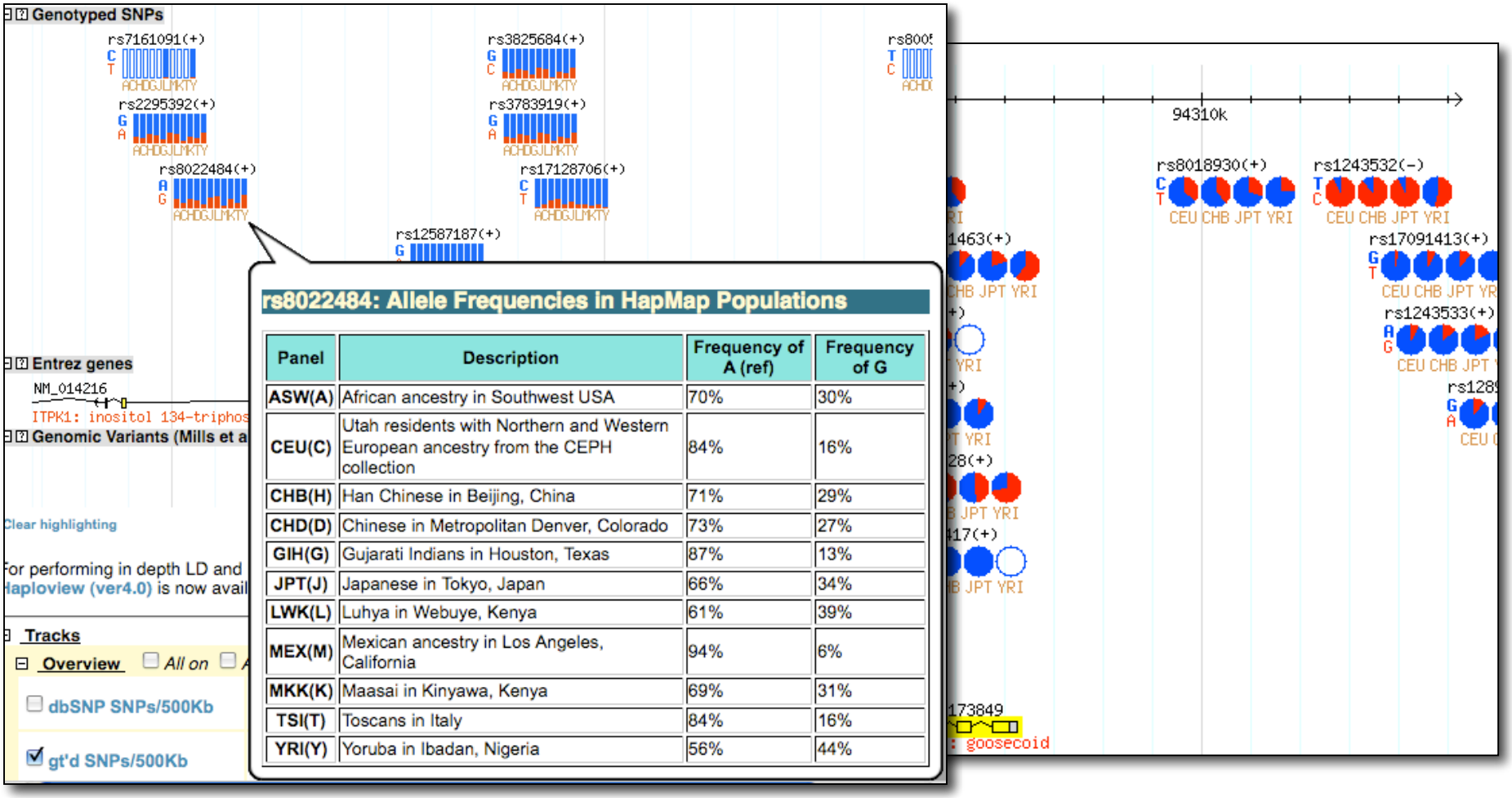
- Frequency by phenotype or any other characteristic
- Sliding window stats



Overview

- GMOD Project Overview
  - Software
  - Community
- Visualizing Next Generation Sequence in GBrowse
  - SAMtools, and GBrowse as a short read viewer
    - Whole genome resequencing of E. coli strains
  - GBrowse for population genetics
    - SNPs in threespine stickleback
  - Other visualizations
- Next Generation Sequencing & Bioinformatics

HapMap Allele Frequencies



rs8022484: Allele Frequencies in HapMap Populations

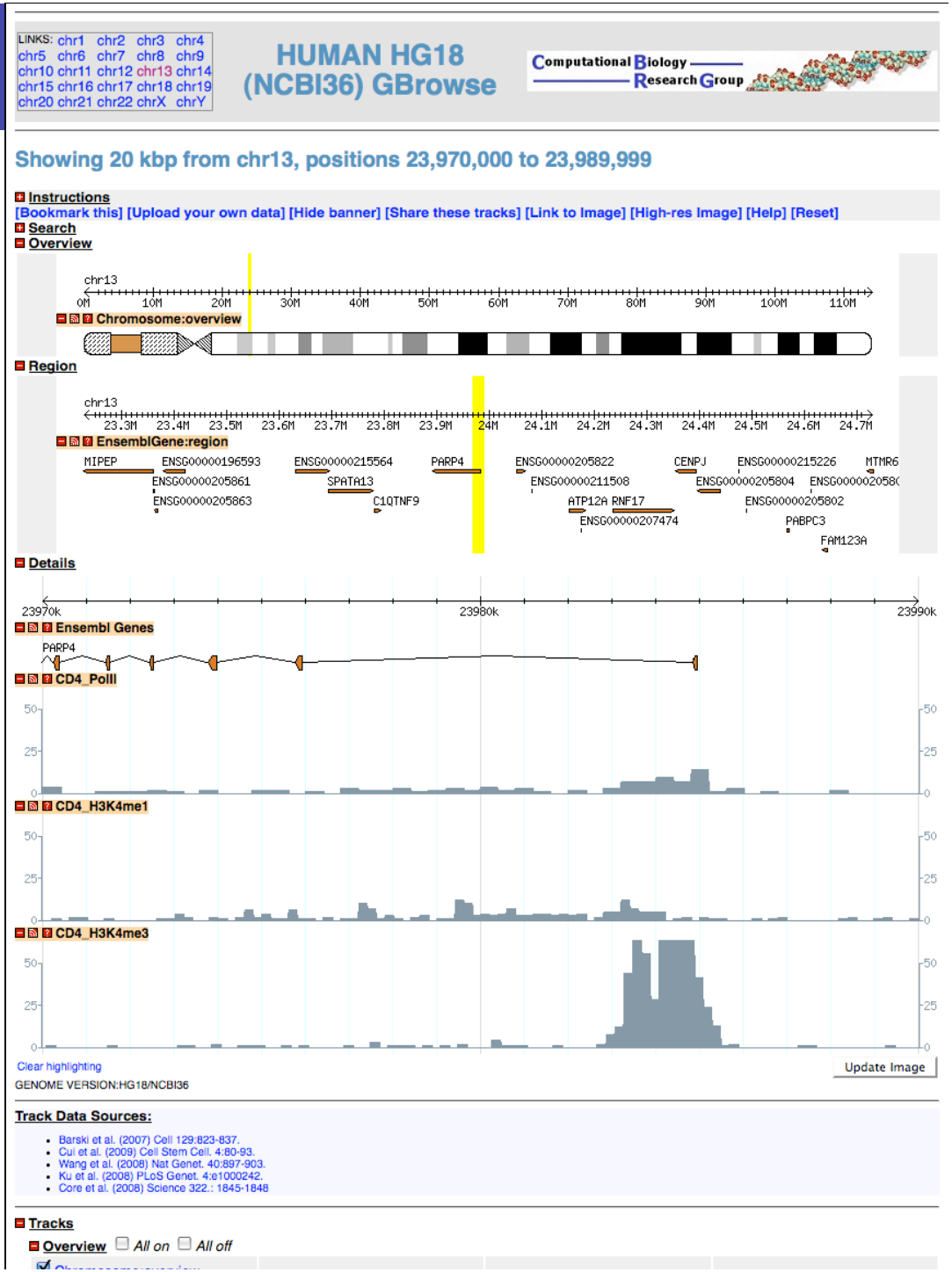
Panel	Description	Frequency of A (ref)	Frequency of G
ASW(A)	African ancestry in Southwest USA	70%	30%
CEU(C)	Utah residents with Northern and Western European ancestry from the CEPH collection	84%	16%
CHB(H)	Han Chinese in Beijing, China	71%	29%
CHD(D)	Chinese in Metropolitan Denver, Colorado	73%	27%
GIH(G)	Gujarati Indians in Houston, Texas	87%	13%
JPT(J)	Japanese in Tokyo, Japan	66%	34%
LWK(L)	Luhya in Webuye, Kenya	61%	39%
MEX(M)	Mexican ancestry in Los Angeles, California	94%	6%
MKK(K)	Maasai in Kinyawa, Kenya	69%	31%
TSI(T)	Toscans in Italy	84%	16%
YRI(Y)	Yoruba in Ibadan, Nigeria	56%	44%

http://hapmap.org

Methylation in Human

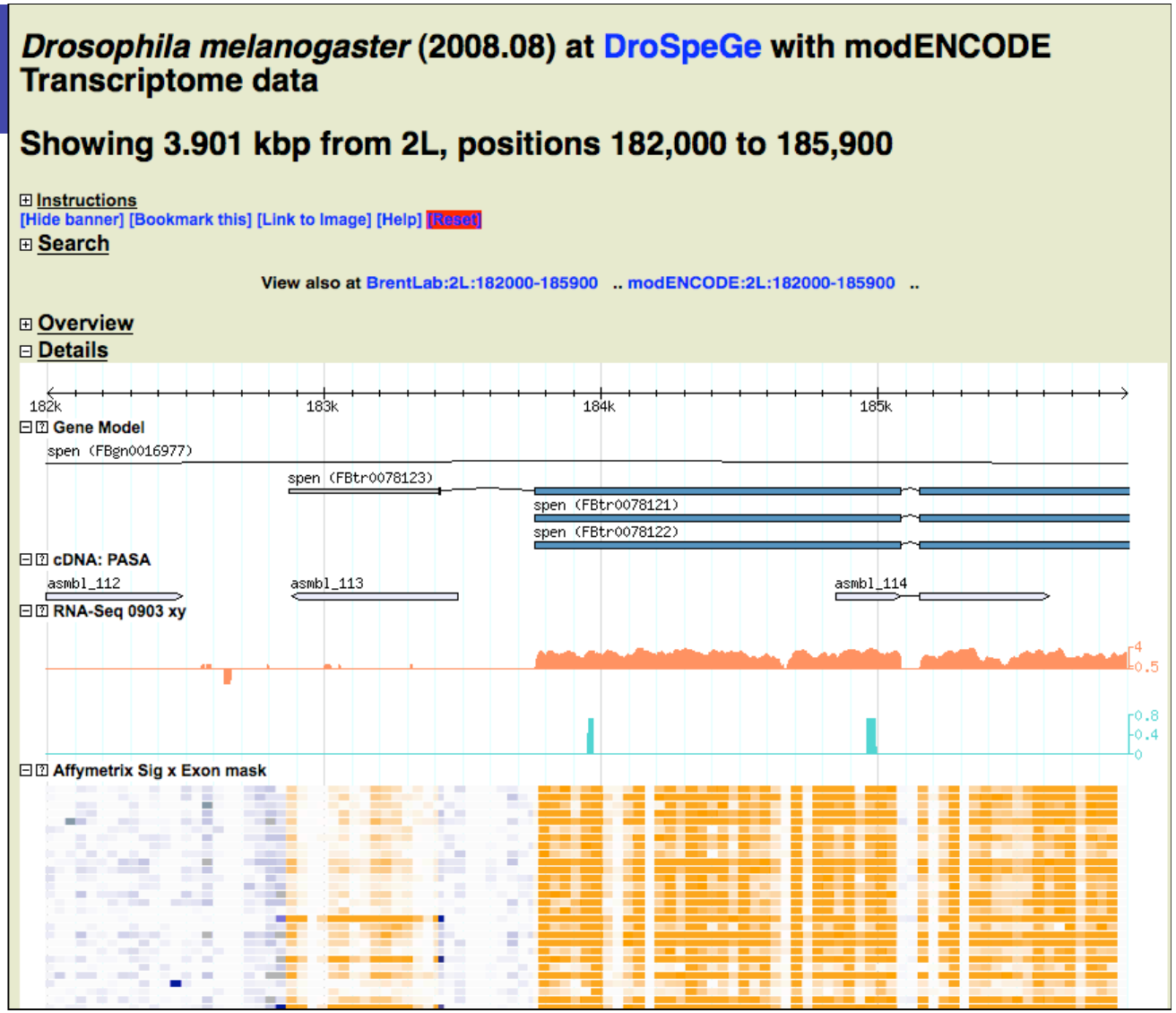
Mostly ChIP-Seq results

Visualization by Computational Biology Research Group at Oxford



Transcriptome

- Transriptome analysis for modENCODE
- Custom modifications to some glyph code
- Visualization by Don Gilbert



http://insects.eugenes.org/species/data/dmel5/modencode/bigmap/

Overview

- GMOD Project Overview
  - Software
  - Community
- Visualizing Next Generation Sequence in GBrowse
  - SAMtools, and GBrowse as a short read viewer
    - Whole genome resequencing of E. coli strains
  - GBrowse for population genetics
    - SNPs in threespine stickleback
  - Other visualizations
- Next Generation Sequencing & Bioinformatics

Needed Resources

To visualize NGS data you will need

- Data
- A computer, or two
  - How big?
- Bioinformatics support
  - How much?

Bioinformatics Support for NGS

- *Examples*
  - *E. coli* data
    - Given aligned reads, and SNP calls
    - Did software installs and configuration, but no programming
  - Threespine Stickleback
    - Given aligned reads, and SNP calls
    - Did installs and configuration, and some scripting (Python and Perl)
- To just *visualize* the data
  - You need someone who can install and configure software, and write scripts.

But it's worse than that …

GMOD

SMBE 2009

You Need to Hire Lots of 

GenomeWeb Survey

Almost all survey respondents pointed out the considerable computational and bioinformatics needs that the new platforms require. **"Anyone thinking of getting these instruments needs a strong IT/informatics group,"** wrote one Illumina user.

"**Our greatest challenge is the lack of bioinformatics support,**" another said.

"**Invest in file servers, computer platforms, and computational biologists,**" a 454 user said.

An ABI SOLiD user said **the greatest challenge has been "data management, interrogation, and visualization."**

But, it's worse than that …

http://www.genomeweb.com/sequencing/users-weigh-next-gen-platforms-over-half-consider-adding-systems-%E2%80%9808

Bioinformatics Support for NGS

You have to learn how to do the informatics too.

My suggestion: **folks should learn to use R, along with Perl, to summarize and quantify these data sets. That also means learning some basic data manipulations** like partitioning …

Many of these data sets have the size of the genome sequences, but the greater complexity of microarray data, as experimenters throw in many treatments and manipulations. **So the lab scientists are the ones who best know contents and likely analyses, more than an informatician just used to processing standard sequence data.**

Don Gilbert, Indiana University

GMOD

SMBE 2009

Acknowledgements

Indiana
Don Gilbert

Sanger
Heng Li

OICR
Lincoln Stein
Scott Cain

Oregon
Nick Stiffler
Liz Perry
Brendan Bohannon
Paul Hohenlohe
Bill Cresko
Patrick Phillips

Speakers
Ben Faga
Sheldon McKay
Mark Yandell

NESCent
Todd Vision
Hilmar Lapp

SMBE
John Logsdon
& Organizing
Committee

Oxford
Steve Taylor

GMOD

SMBE 2009

Thank You!

Dave Clements
GMOD Help Desk

National Evolutionary
Synthesis Center

clements@nescent.org
help@gmod.org

http://gmod.org/GMOD_Help_Desk
http://nescent.org

National Evolutionary Synthesis Center
NESCent

SMBE 2009