



*American Genetic  
Association*



# Seeing the forest and the trees: visualizing next generation sequence data

Dave Clements  
GMOD Help Desk  
National Evolutionary Synthesis Center (NESCent)  
[clements@nescent.org](mailto:clements@nescent.org)

Next Generation Genome Analysis in  
Non-Model Organisms  
An AGA Special Event  
June 13, 2009

# Overview

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



# 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

Backs Garfield &  
modENCODE

*E. coli* landing page

Overview:  
chromosome wide

Region:  
intermediate zoom

Details:  
current area

Tracks:  
current  
configuration



## E. coli strains: 1 kbp from Ancestral:1..1,000

### Instructions

[\[Bookmark this\]](#) [\[Add custom tracks\]](#) [\[Share these tracks\]](#) [\[Send to Galaxy\]](#) [\[Link to Image\]](#)  
[\[High-res Image\]](#) [\[Download PDF\]](#) [\[Help\]](#) [\[Reset\]](#) [\[Make an Error\]](#)

### Search

Landmark or Region:

Data Source

E. coli strains

Scroll/Zoom:

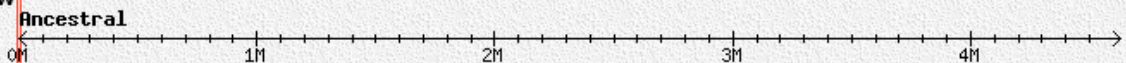


Show 1 kbp



Flip

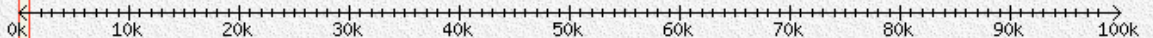
### Overview



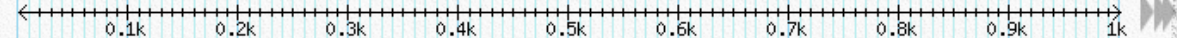
### Variation



### Region



### Details



### Ancestral DNA



### Tracks

Overview  All on  All off

Variation

Region  All on  All off

Variation

Basic features  All on  All off

Ancestral 6-frame translation

Ancestral DNA

Reads  All on  All off

Period1 coverage (density)

Period1A coverage (density)

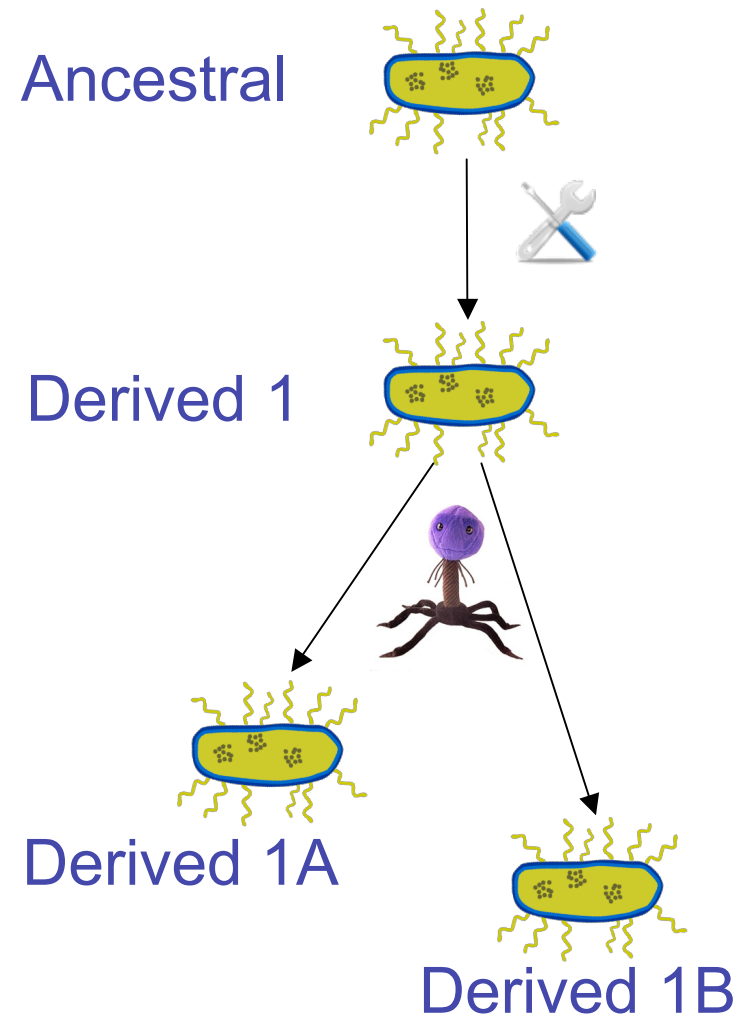
Period1B coverage (density)



# The Trees: *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



## *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

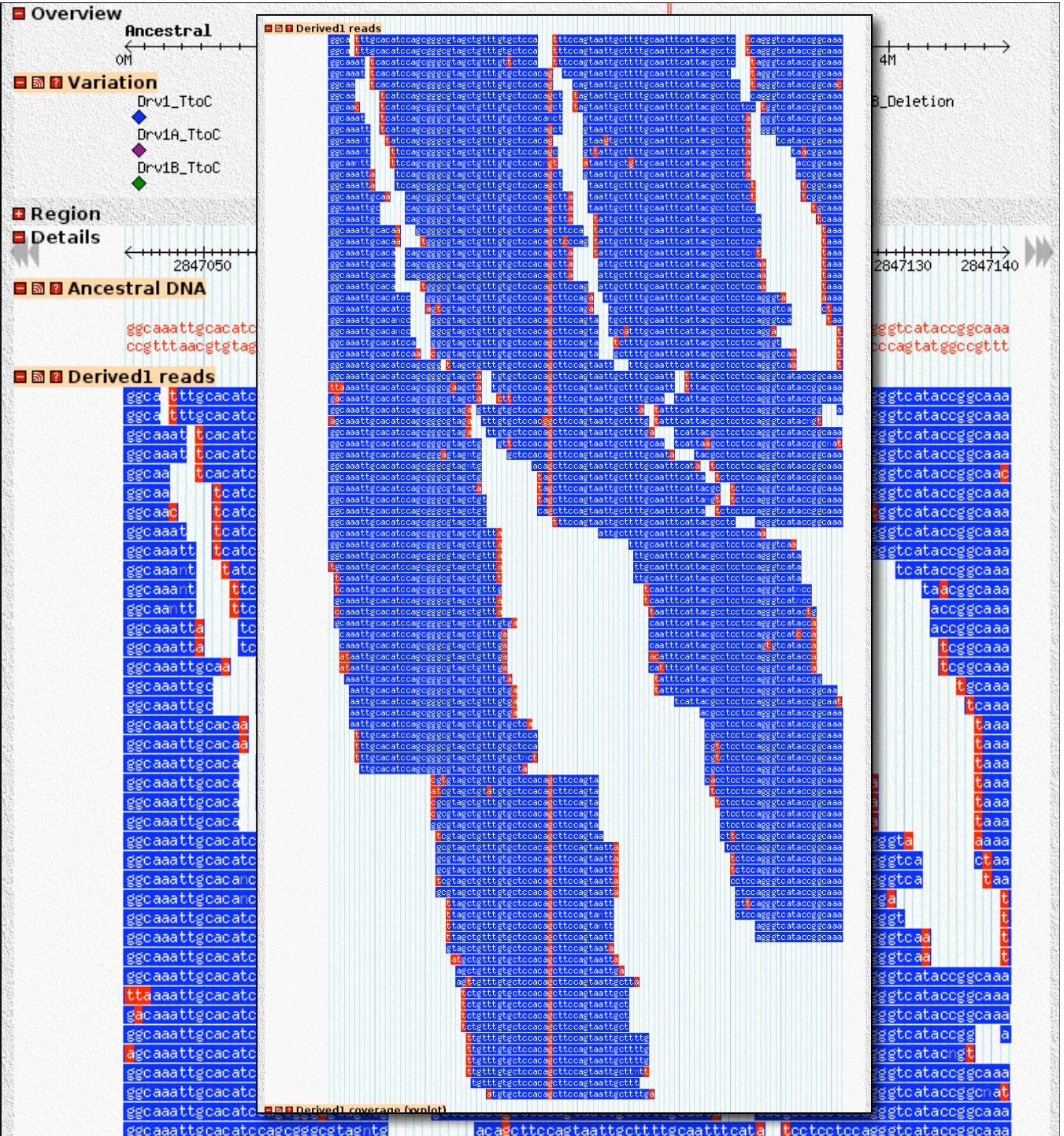


\* Li H., Ruan J. and Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18 (11), 1851-1858.  
<http://maq.sf.net>

# GBrowse as an Alignment Viewer

High magnification view: 100bp

Uses GBrowse 2 (Beta) and the Bio-SamTools GBrowse database adaptor (Alpha).



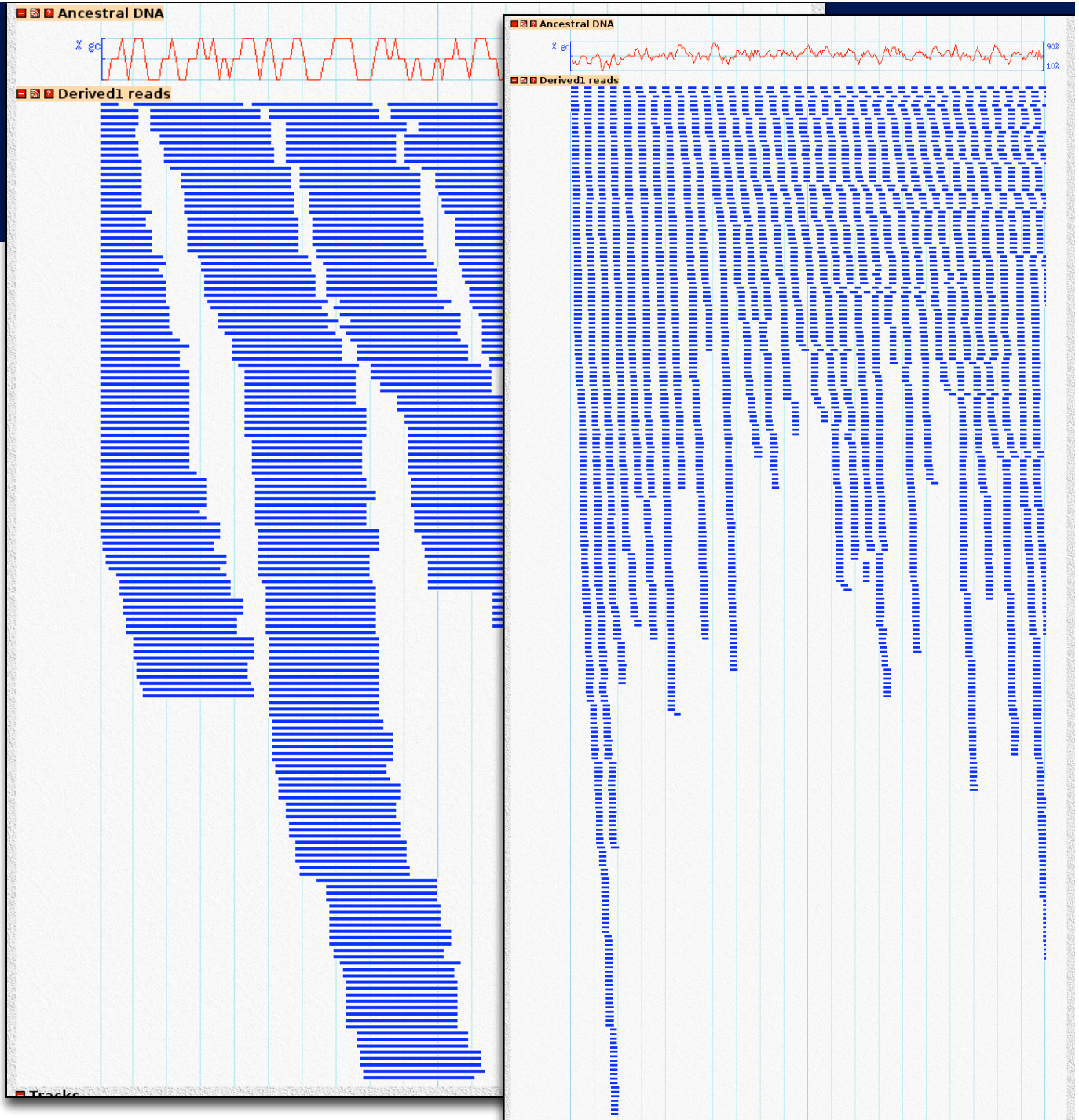


# GBrowse as an Alignment Viewer

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



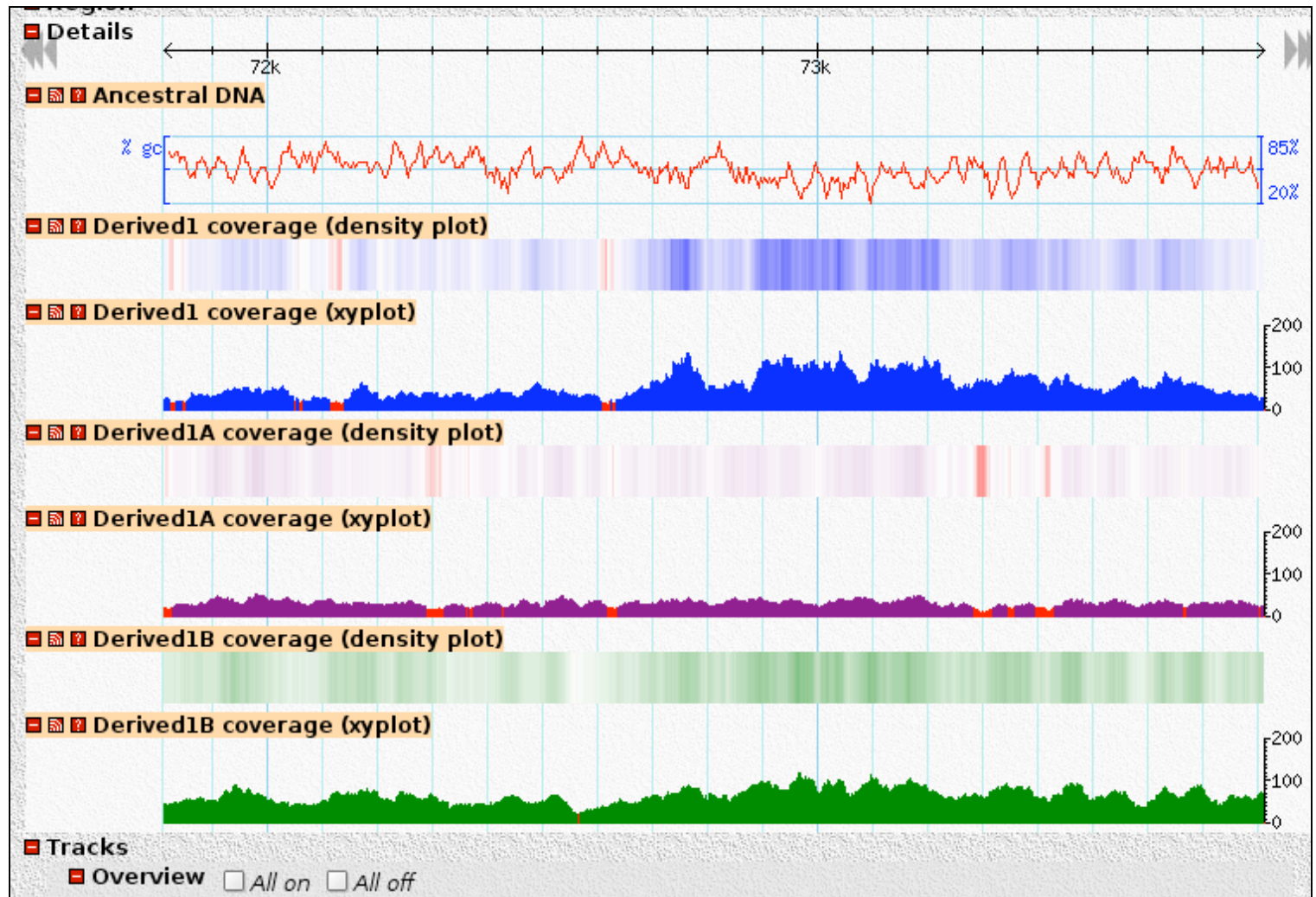


# GBrowse as an Alignment Viewer

Summarize!

SAMtools and  
GBrowse help  
here

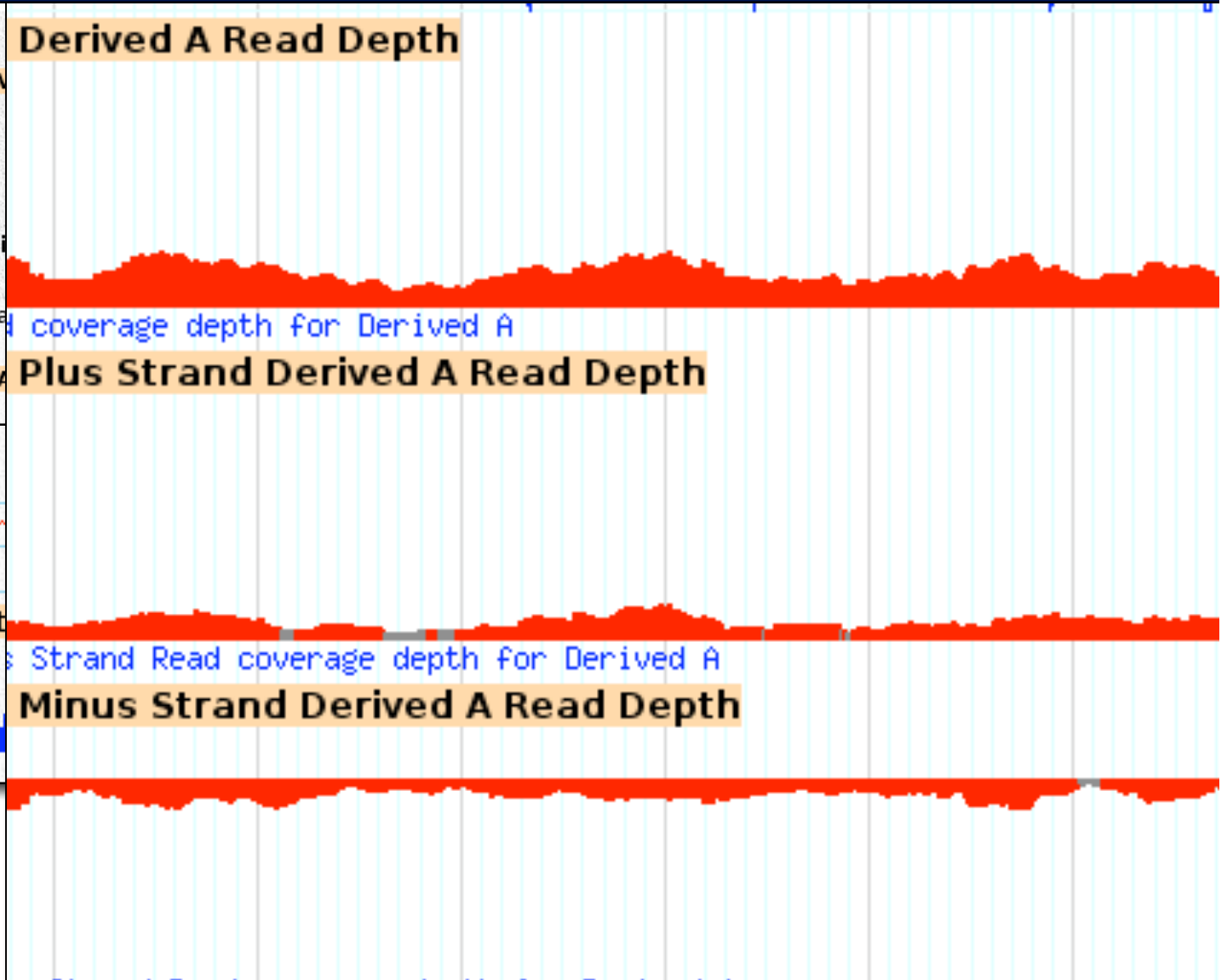
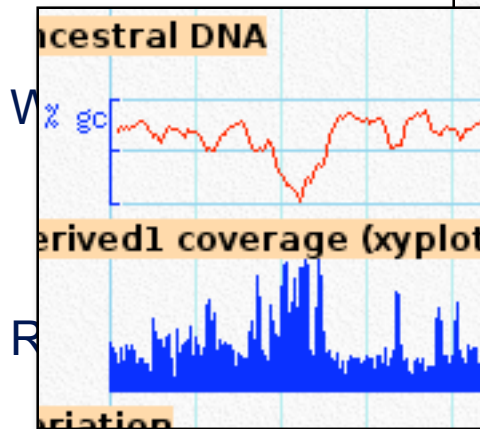
SAMtools can  
summarize  
several values  
from the data.



# GBrowse as an Alignment Viewer

Even basic summarization can be enough to gain insight.

GC content & read coverage for first 200Kbp of *E. coli*



# Overview

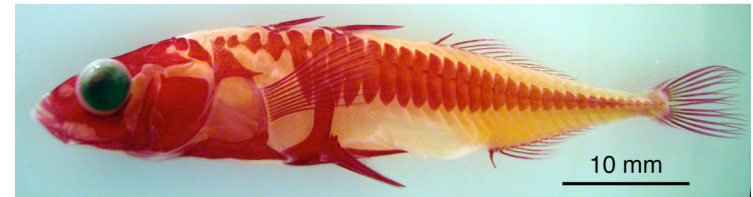
- Visualizing Next Generation Sequence in GBrowse
  - The trees: SAMtools, and GBrowse as a short read viewer
    - Whole genome resequencing of E. coli strains
  - The forest: GBrowse for population genetics
    - SNPs in threespine stickleback
  - Other visualizations
  - Next Generation Sequencing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



# The Forest: 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





# 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



# GBrowse for Population Genetics

Shows

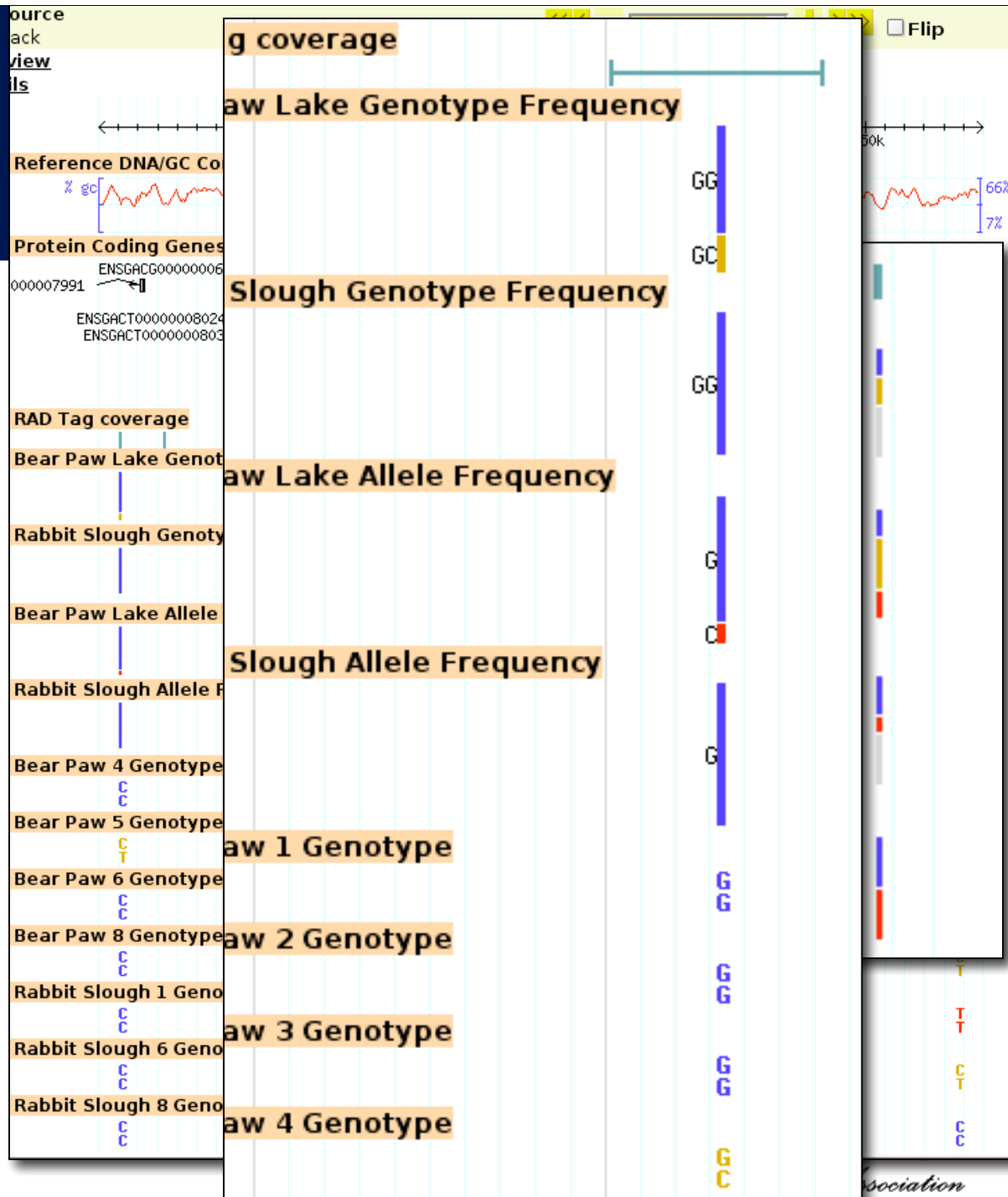
Where we looked  
Allele & genotype  
frequencies

By population

Individual genotypes

Could also show:

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

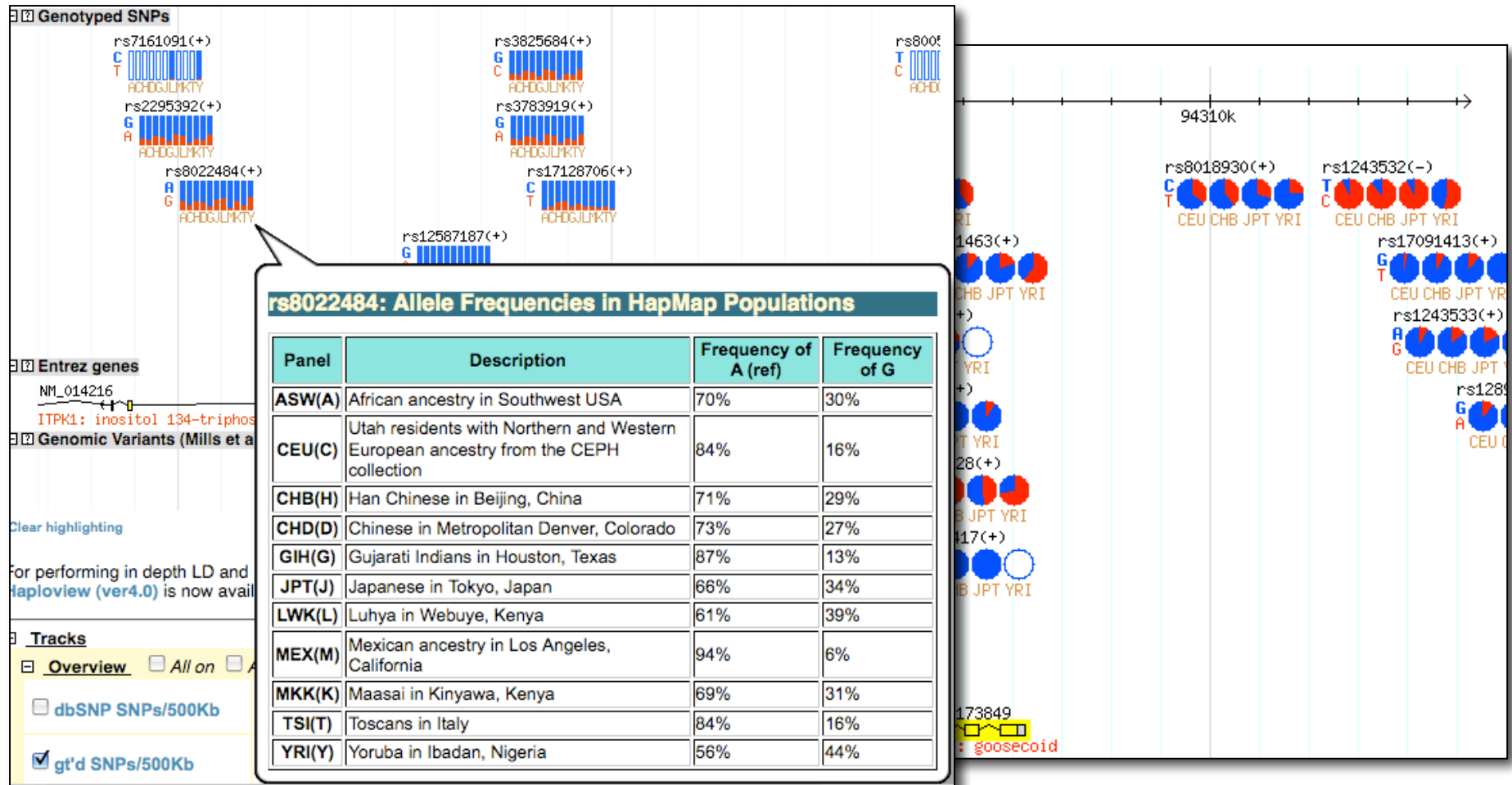


# Overview

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



# HapMap Allele Frequencies



<http://hapmap.org>



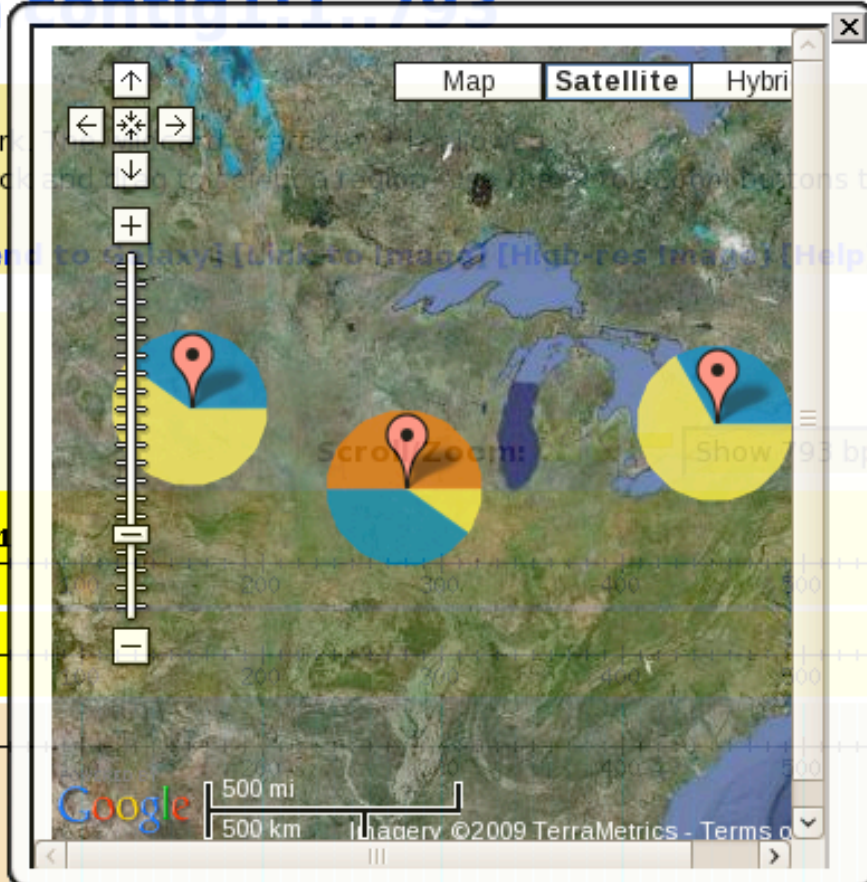


# Geolocation data

793 bp from contig1:1\_793

Name, locus, or other landmark  
to center on a location, or click

[Share these tracks] [Send to Galaxy] [Click to Image] [High-res Image] [Help] [Reset] Make an Error



to change magnification and position.

+ > >>  Flip



Ben Faga

American Genetic Association

# Methylation in Human

Mostly ChIP-Seq results

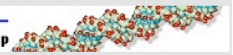
Visualization by Computational Biology Research Group at Oxford



LINKS: chr1 chr2 chr3 chr4  
chr5 chr6 chr7 chr8 chr9  
chr10 chr11 chr12 chr13 chr14  
chr15 chr16 chr17 chr18 chr19  
chr20 chr21 chr22 chrX chrY

HUMAN HG18  
(NCBI36) GBrowse

Computational Biology  
Research Group



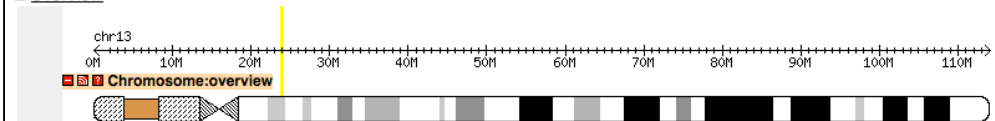
Showing 20 kbp from chr13, positions 23,970,000 to 23,989,999

#### Instructions

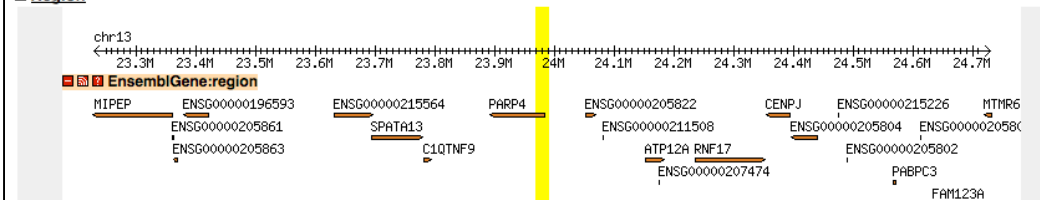
[\[Bookmark this\]](#) [\[Upload your own data\]](#) [\[Hide banner\]](#) [\[Share these tracks\]](#) [\[Link to image\]](#) [\[High-res image\]](#) [\[Help\]](#) [\[Reset\]](#)

#### Search

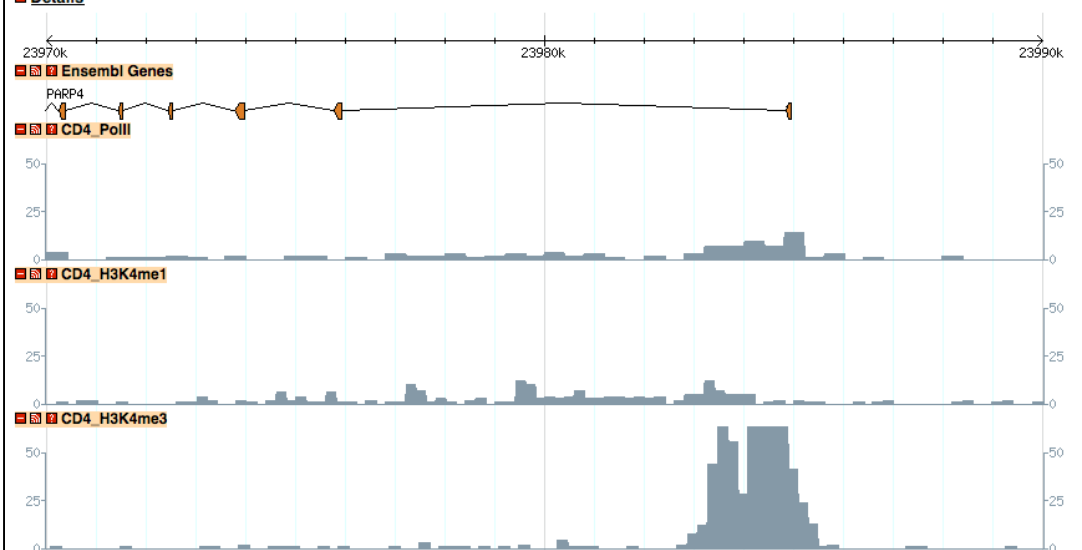
#### Overview



#### Region



#### Details



Clear highlighting

GENOME VERSION:HG18/NCBI36

Update Image

#### Track Data Sources:

- Barski et al. (2007) Cell 129:823-837.
- Cui et al. (2009) Cell Stem Cell. 4:80-93.
- Wang et al. (2008) Nat Genet. 40:897-903.
- Ku et al. (2008) PLoS Genet. 4:e1000242.
- Core et al. (2008) Science 322: 1845-1848

#### Tracks

Overview  All on  All off

CD4\_PoII

# Transcriptome

Transcriptome  
analysis for  
modENCODE

Custom  
modifications to  
some glyph  
code

Visualization by  
Don Gilbert

*Drosophila melanogaster* (2008.08) at [DroSpeGe](#) with modENCODE  
Transcriptome data

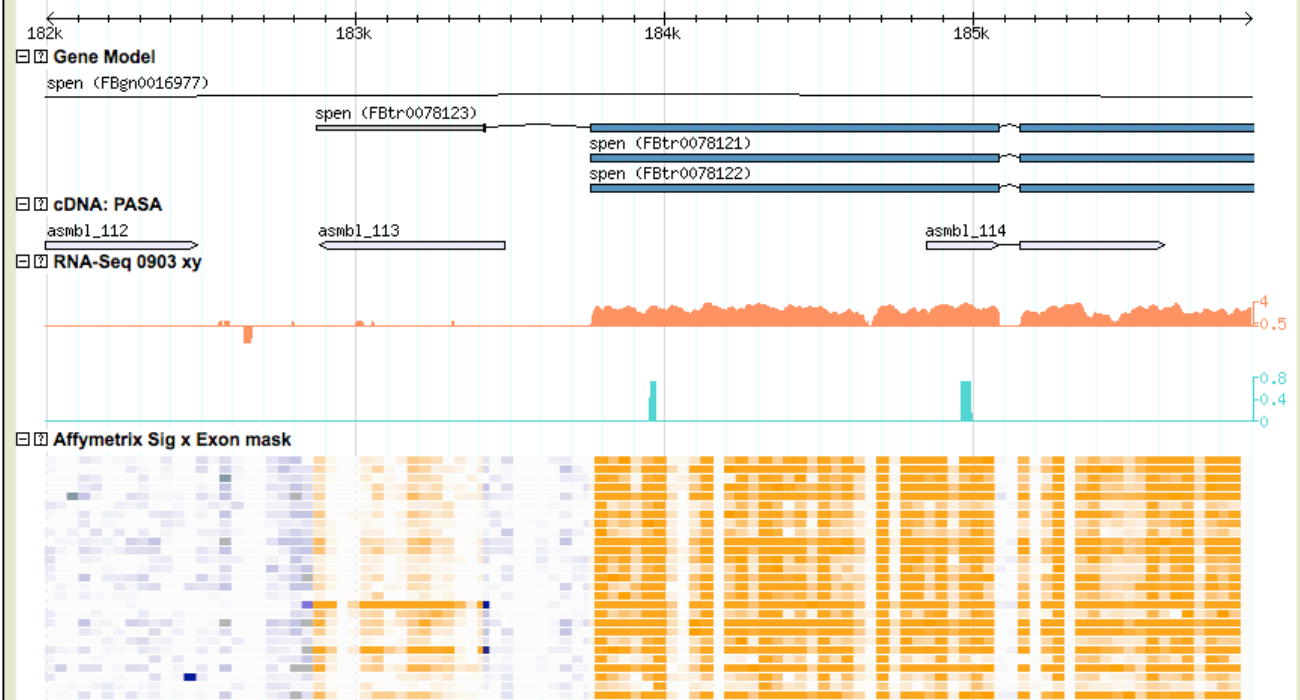
Showing 3.901 kbp from 2L, positions 182,000 to 185,900

[Instructions](#)  
[\[Hide banner\]](#) [\[Bookmark this\]](#) [\[Link to Image\]](#) [\[Help\]](#) [\[Reset\]](#)  
 [Search](#)

View also at [BrentLab:2L:182000-185900](#) .. [modENCODE:2L:182000-185900](#) ..

[Overview](#)

[Details](#)



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



# Overview

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





# Needed Resources

To visualize NGS data you will need

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



# Bioinformatics Support for NGS

Yesterday, *everybody* mentioned their bioinformatics support people.



"LOTS of data."

*Valerie Reinke, yesterday*



# 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.”**



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

*American Genetic  
Association*

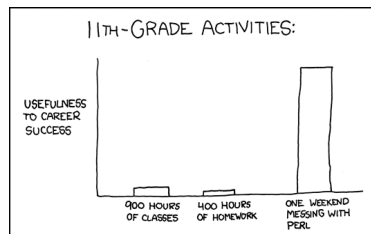
*You should learn some bioinformatics too.*

"Will the next generation of scientists be more trans-disciplinary? Absolutely."

*Eric Green, yesterday*

"One weekend messing with Perl."

*Brent Graveley, yesterday, sort of*



<http://xkcd.com/519/>



# Overview

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



# GMOD for Non-Models

GMOD is two equally important things:

- 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>

*American Genetic  
Association*



# GMOD Software: A whirlwind tour

Apollo

GBrowse & JBrowse

CMap

MAKER & DIYA

GBrowse\_syn, Sybil, ...

Chado, GMODWeb, Tripal

BioMart, InterMine

Pathway Tools

...

3R:1178000-1230000 *Drosophila melanogaster*

Overview  
Overview of 3R

Available Tracks: (Drag → to view)

Comparative Reference Comparative

Home Genome Synteny Blast / Blat WormMart Batch Sequences Markers Genetic Maps Submit Searches Site Map

Find: WBGene00016355 [Anything]

Streptococcus pneumoniae Hungary19A-6:assembly 968/- (2245614 bp)

PalameciumDB

bio::mart

HOME MARTVIEW MARTSERVICE DOCS CONTACT NEWS

FlyMine v 16.0 An integrated database for *Drosophila* and *Anopheles* genomics

Home Templates Lists QueryBuilder Data MyMine

Gene-Reaction Schematic: ?

Genetic Regulation Schematic: ?

Gene Local Context (not to scale): ?

Transcription Units:



# GMOD: Community

Next GMOD Meeting  
University of Oxford, UK

6-7 August 2009

Part of GMOD Europe 2009

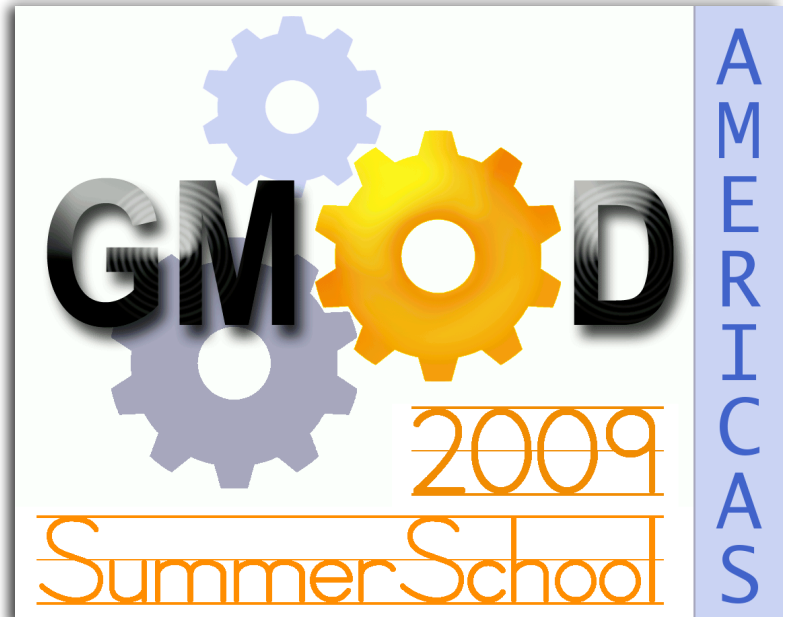
## GMOD Courses

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



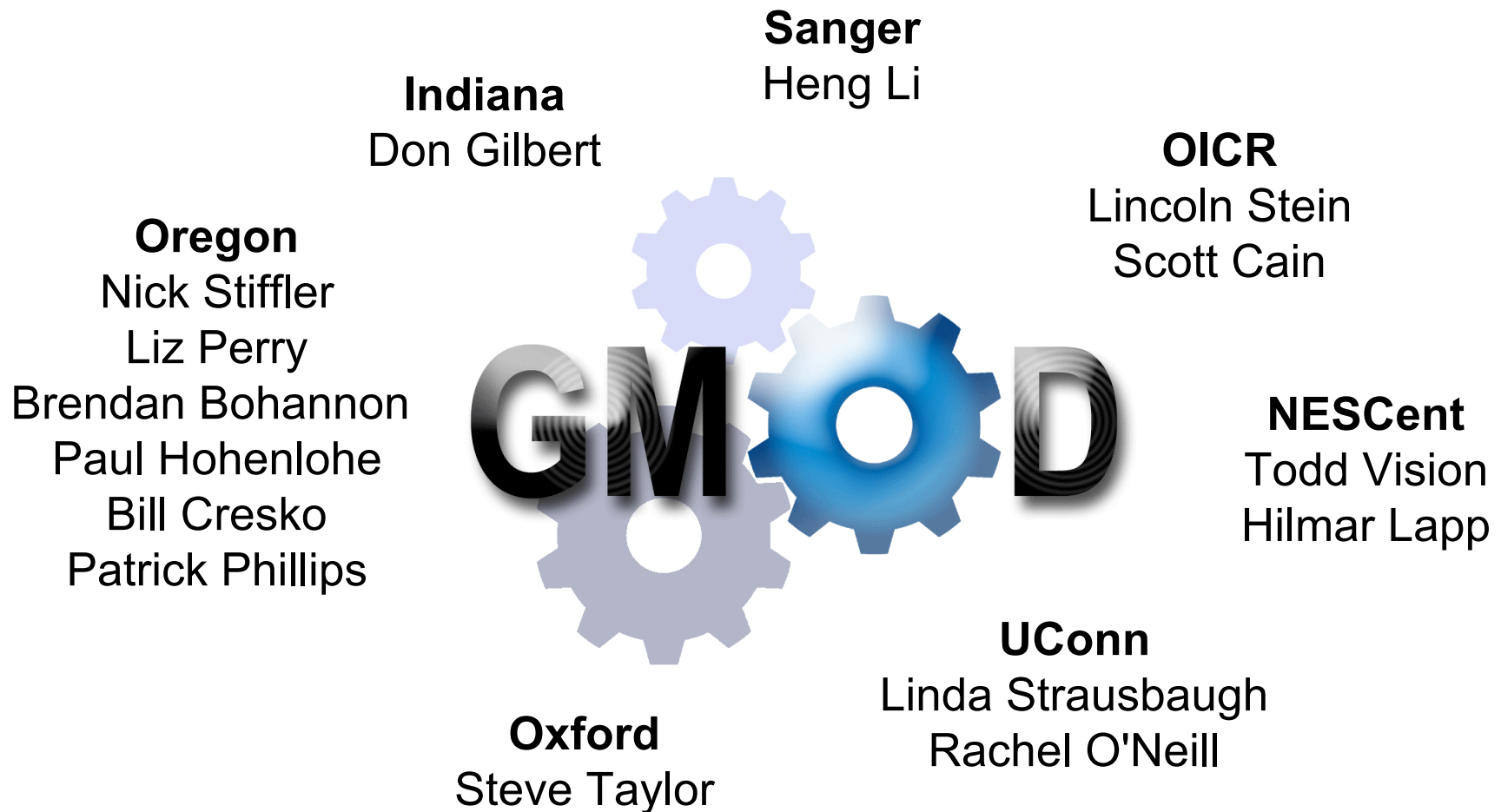
# GMOD: Who uses it?



Plus several *hundreds* others.

American Genetic Association

# Acknowledgements





# Thank You!



Dave Clements  
GMOD Help Desk

National Evolutionary  
Synthesis Center

[clements@nescent.org](mailto:clements@nescent.org)  
[help@gmod.org](mailto:help@gmod.org)

[http://gmod.org/GMOD\\_Help\\_Desk](http://gmod.org/GMOD_Help_Desk)  
<http://nescent.org>



*American Genetic  
Association*

# Bioinformatics Support for NGS

- *Examples*
  - *E. coli* data
    - Given aligned reads, SNP calls
    - Did software installs and configuration, but no programming
  - Threespine Stickleback
    - Given aligned reads, 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.

