# Visualising and Integrating Next Generation Sequence Data using GMOD

Evolutionary Genetics - The Impact of Next Generation Sequencing Technologies

2 April 2009

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

# Outline

- GBrowse as an alignment viewer
  - *E. coli*
  - Whole genome resequencing
- Next Generation Sequencing & Bioinformatics
- GBrowse for population genetics
  - Threespine Stickleback
  - Deep sequencing of select regions
  - looking for SNPs
- Other Visualisations
- GMOD Project
- Followed by …


# Visualisation Panel & Discussion

| | |
|---|---|
| **Chuck Cannon** | Chinese Academy of Sciences<br>Assembly free approaches |
| **Philip Johnson** | UC Berkeley<br>Metagenomics and gene flow |
| **Phillip Morin** | NOAA Fisheries & Scripps Institution of Oceanography<br>Natural diversity and Geolocation |
| **Korbinian Schneeberger** | Max Planck Institute for Developmental Biology<br>Arabidopsis: Deep and Wide |

GMOD

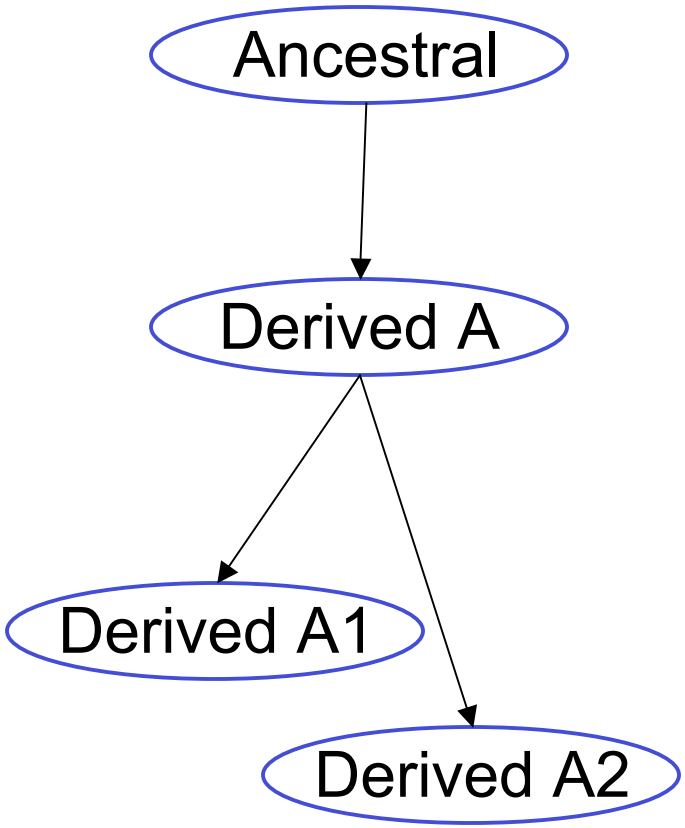
National Evolutionary Synthesis Center
NESCent

# Outline

- GBrowse as an alignment viewer
  - *E. coli*
  - Whole genome resequencing
- Next Generation Sequencing & Bioinformatics
- GBrowse for population genetics
  - Threespine Stickleback
  - Deep sequencing of select regions
  - looking for SNPs
- Other Visualisations
- GMOD Project
- Panel & Discussion


# *E. coli*: Resequencing

- Tale of 4 strains
  - Ancestral:
    - reference
  - Derived A:
    - manipulated in two places (neutral, metabolic)
    - exposed to phage yielding 2 resistant strains
  - Derived A1
    - 1bp change
  - Derived A2
    - 2-3Kbp deletion




Work done at U Oregon by Brendan Bohanon, Liz Perry, and Nick Stiffler

GMOD

National Evolutionary Synthesis Center
NESCent

# Process

- Extract DNA
- Sonicate it, aiming for 500bp fragments
- Unpaired end run on an Illumina GA2
- Filter results for quality
- Align it with MAQ
- Visualize it with GBrowse


# GBrowse

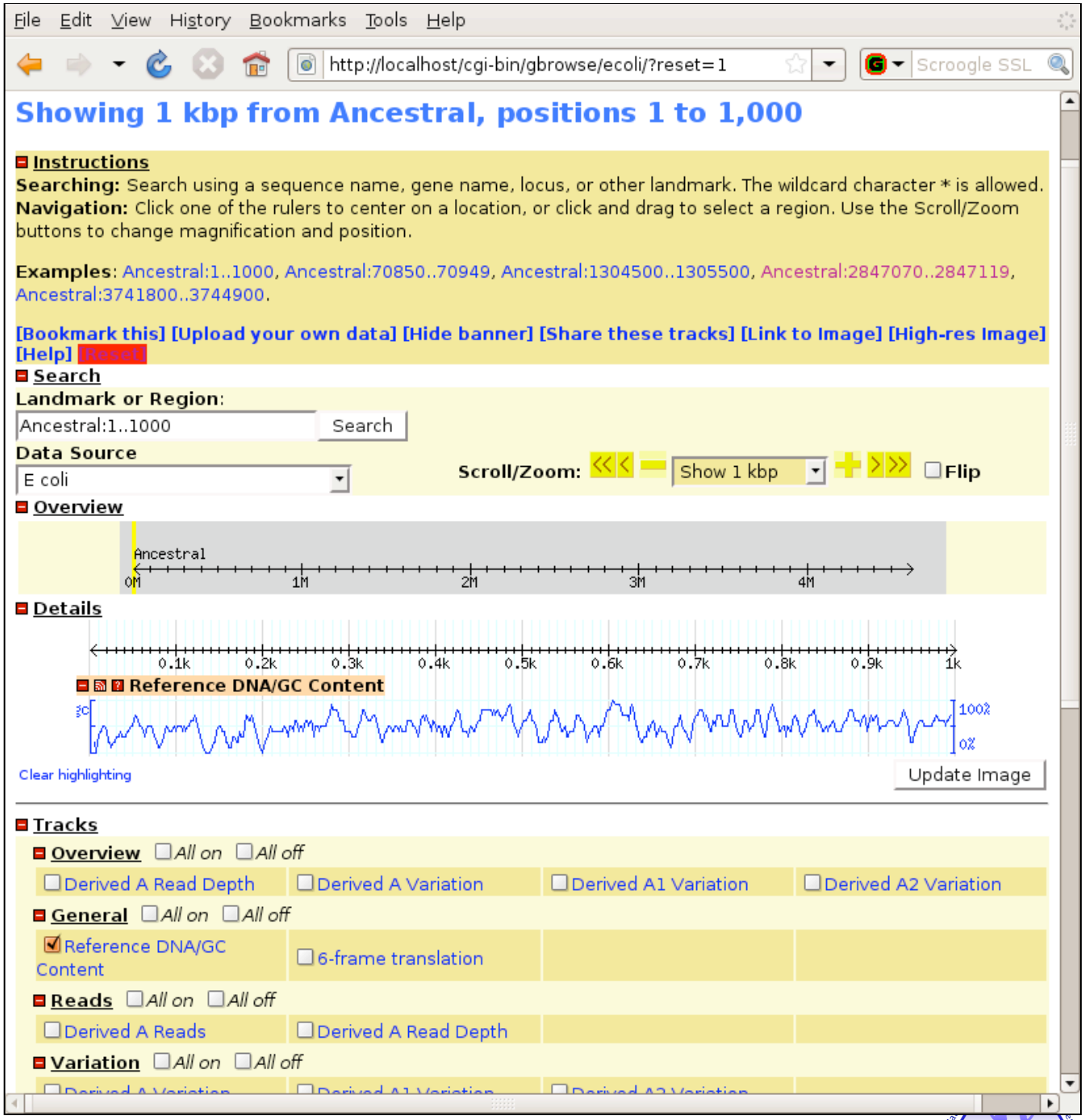
GMOD's main genome browser

*E. coli* landing page

Overview:
chromosome wide

Details:
current region

Tracks:
current configuation




NESCent

# GBrowse as an Alignment Viewer

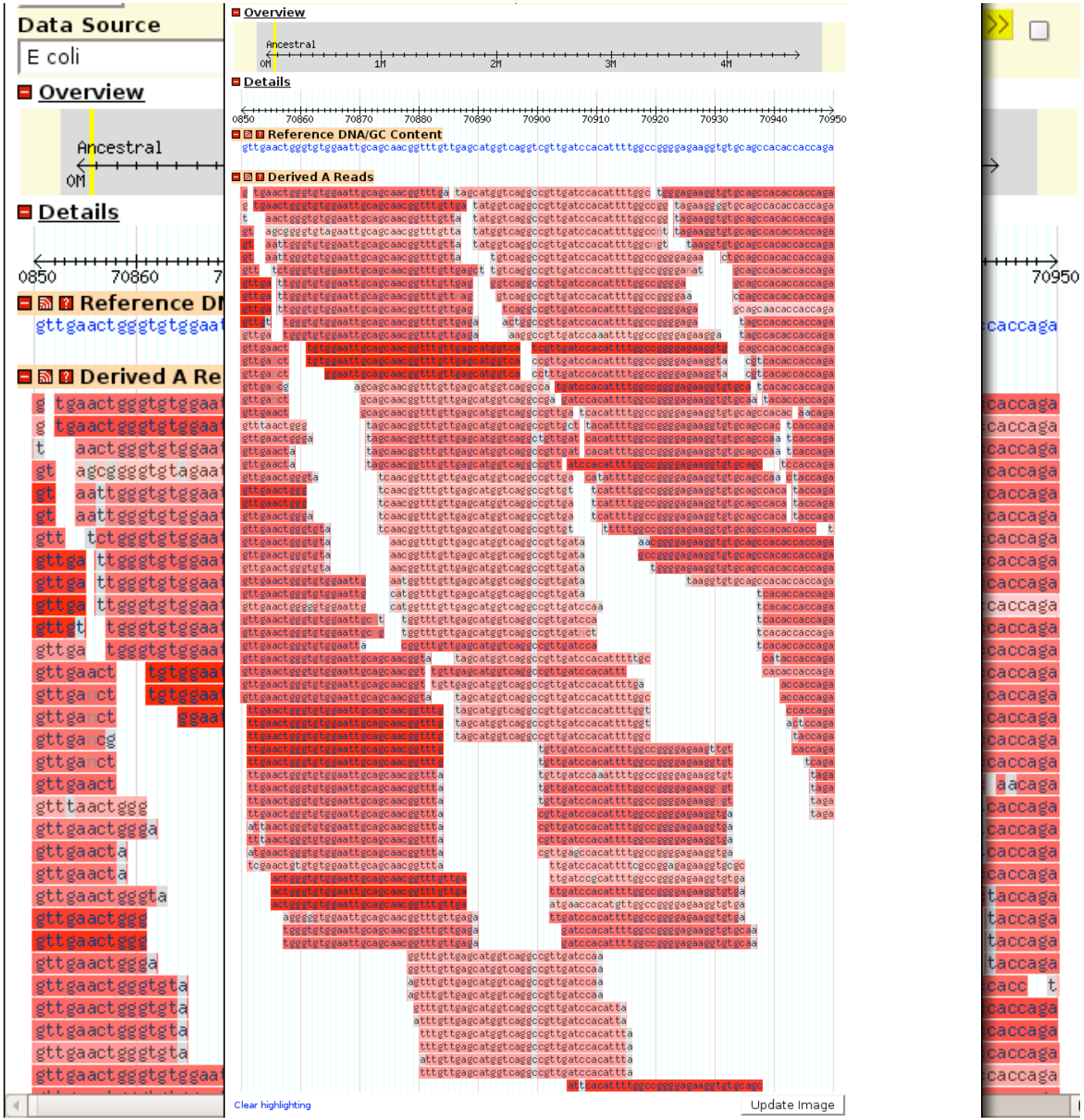
High magnification view: 100bp
Shows a SNP

Colors: mapping quality

Dark red: high

Light red: Low

Currently no good way to show both base and read quality

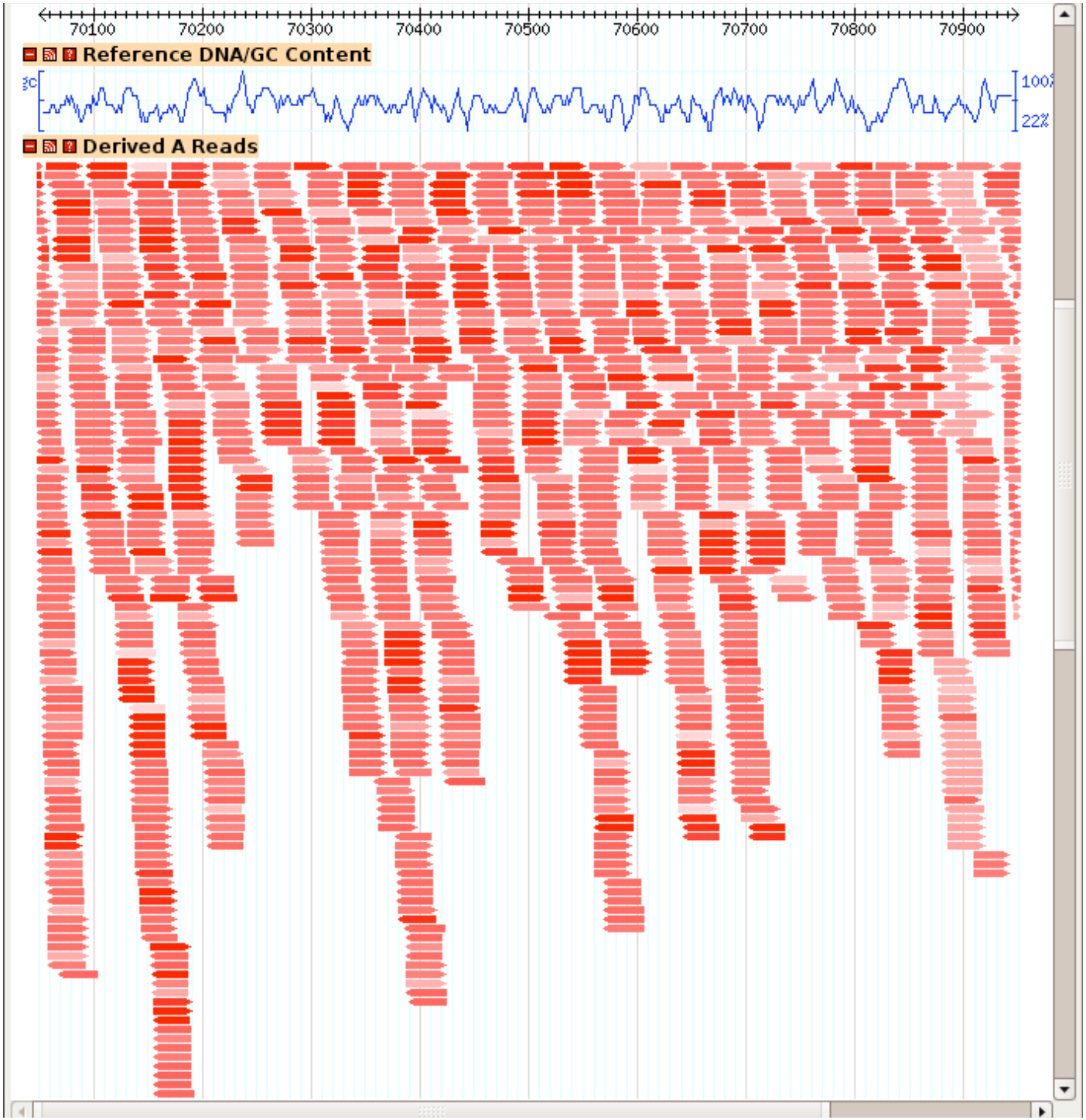


# GBrowse as an Alignment Viewer

Magnification:
900bp

Letters go away at
~110bp

On my laptop:
out to ~2-3kbp
5kbp times out



# GBrowse as a Short Read Viewer?

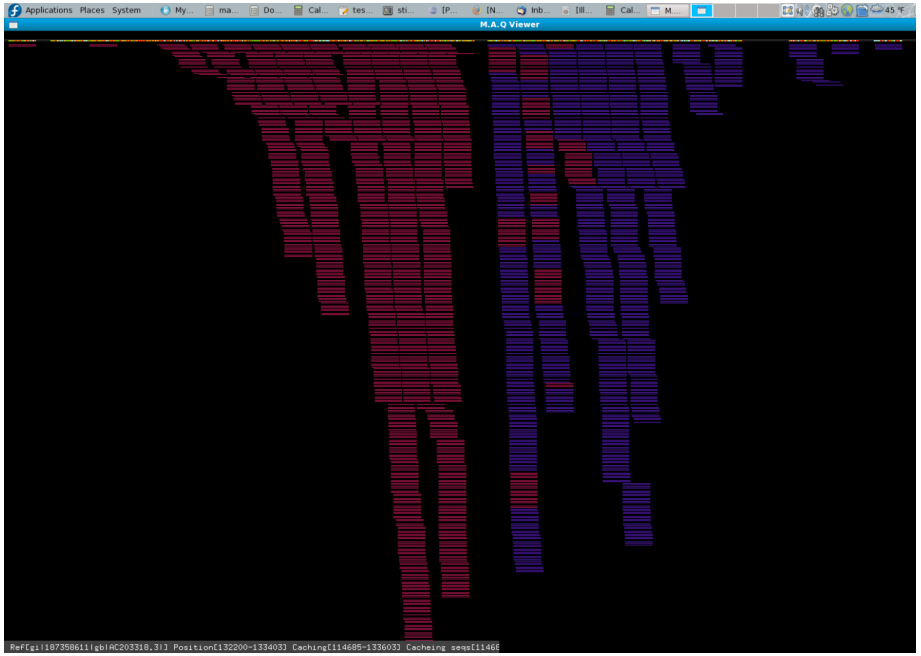
## Does it make sense to show individual reads?

At low resolution?

No - better ways to do it and computationally prohibitive

At high resolution?

No - that's what your alignment software is for.




M.A.Q Viewer screenshot

## Why are you using GBrowse?

## To visualize, share and integrate your data

GMOD

National Evolutionary Synthesis Center
NESCent

# GBrowse as an Alignment Viewer!

Summarize alignments

Overview:

- Read coverage
- Variation

Details:

- Read coverage, total and for both strands
- Variation in all 3 derived lines

What's lost?

- Mapping quality, and that doesn't have to be.

# Outline

- GBrowse as an alignment viewer
  - *E. coli*
  - Whole genome resequencing
- Next Generation Sequencing & Bioinformatics
- GBrowse for population genetics
  - Threespine Stickleback
  - Deep sequencing of select regions
  - looking for SNPs
- Other Visualisations
- GMOD Project
- Followed by …


# Bioinformatics Support & Knowledge are Key!

## 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 for his group has been “data management, interrogation, and visualization.”

## Don Gilbert

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.**

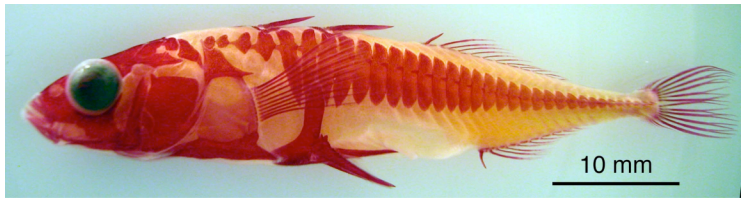
* http://www.genomeweb.com/sequencing/users-weigh-next-gen-platforms-over-half-consider-adding-systems-%E2%80%9808

# Outline

- GBrowse as an alignment viewer
  - *E. coli*
  - Whole genome resequencing
- Next Generation Sequencing & Bioinformatics
- GBrowse for population genetics
  - Threespine Stickleback
  - Deep sequencing of select regions
  - looking for SNPs
- Other Visualisations
- GMOD Project
- Panel & Discussion


# 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

Work done at U Oregon by Bill Cresko, Paul Hohenlohe, and Nick Stiffler

GMOD

National Evolutionary Synthesis Center
NESCent

# Process

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

GMOD

National Evolutionary Synthesis Center
NESCent

# GBrowse for Population Genetics

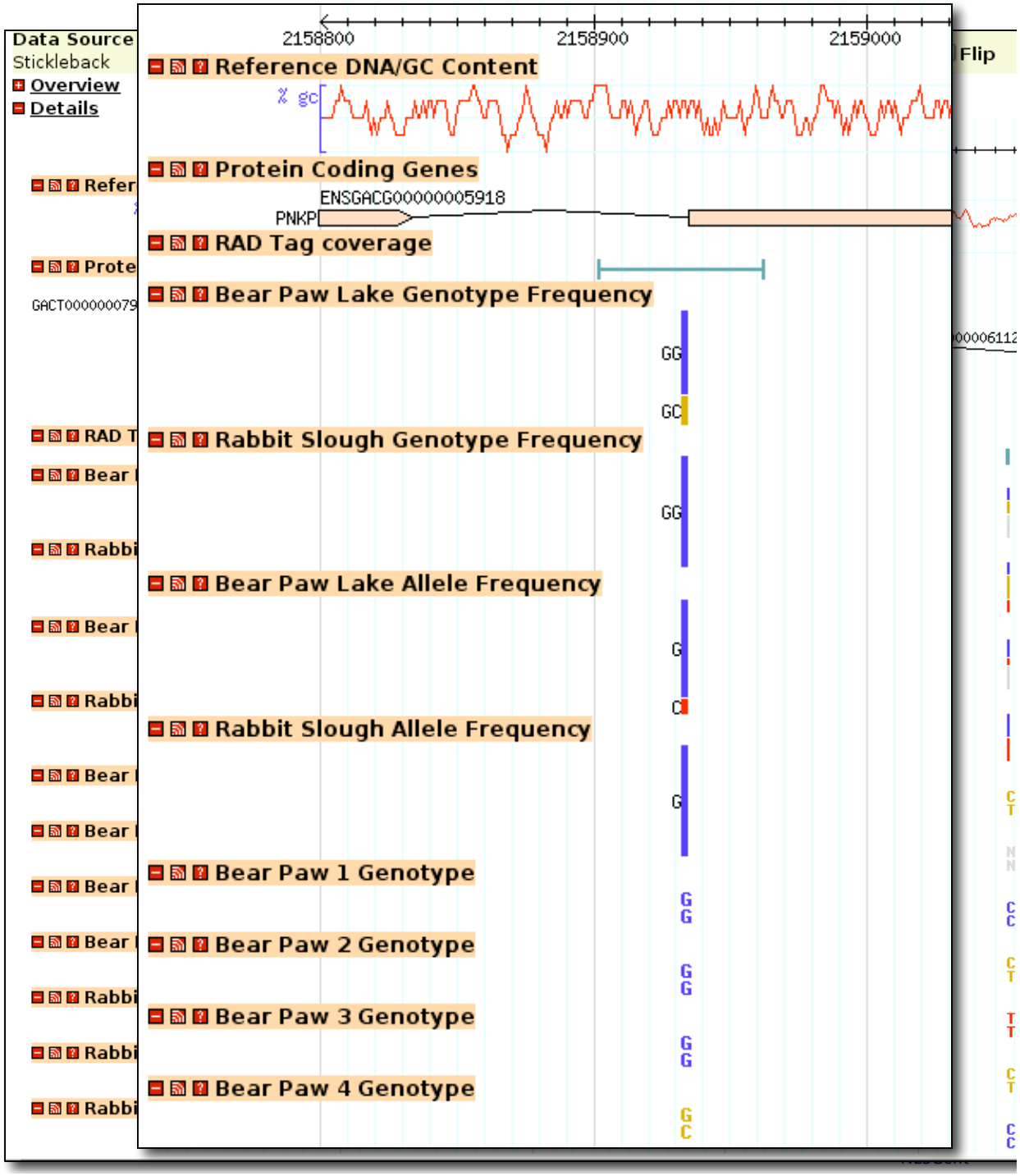
Allele & genotype frequencies

- By population
- Individual genotypes

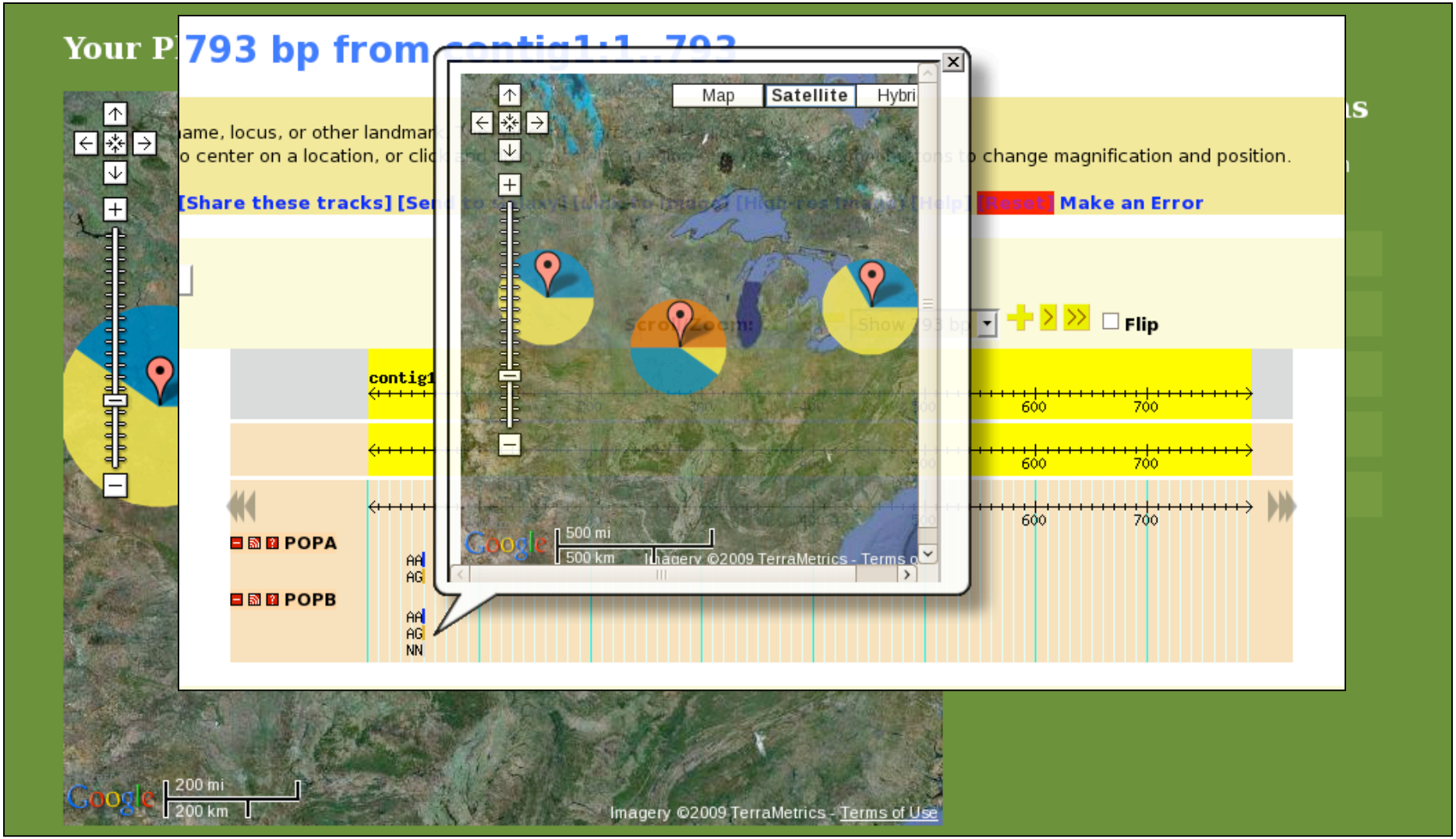
Where we looked

Could also show:

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



# Geolocation

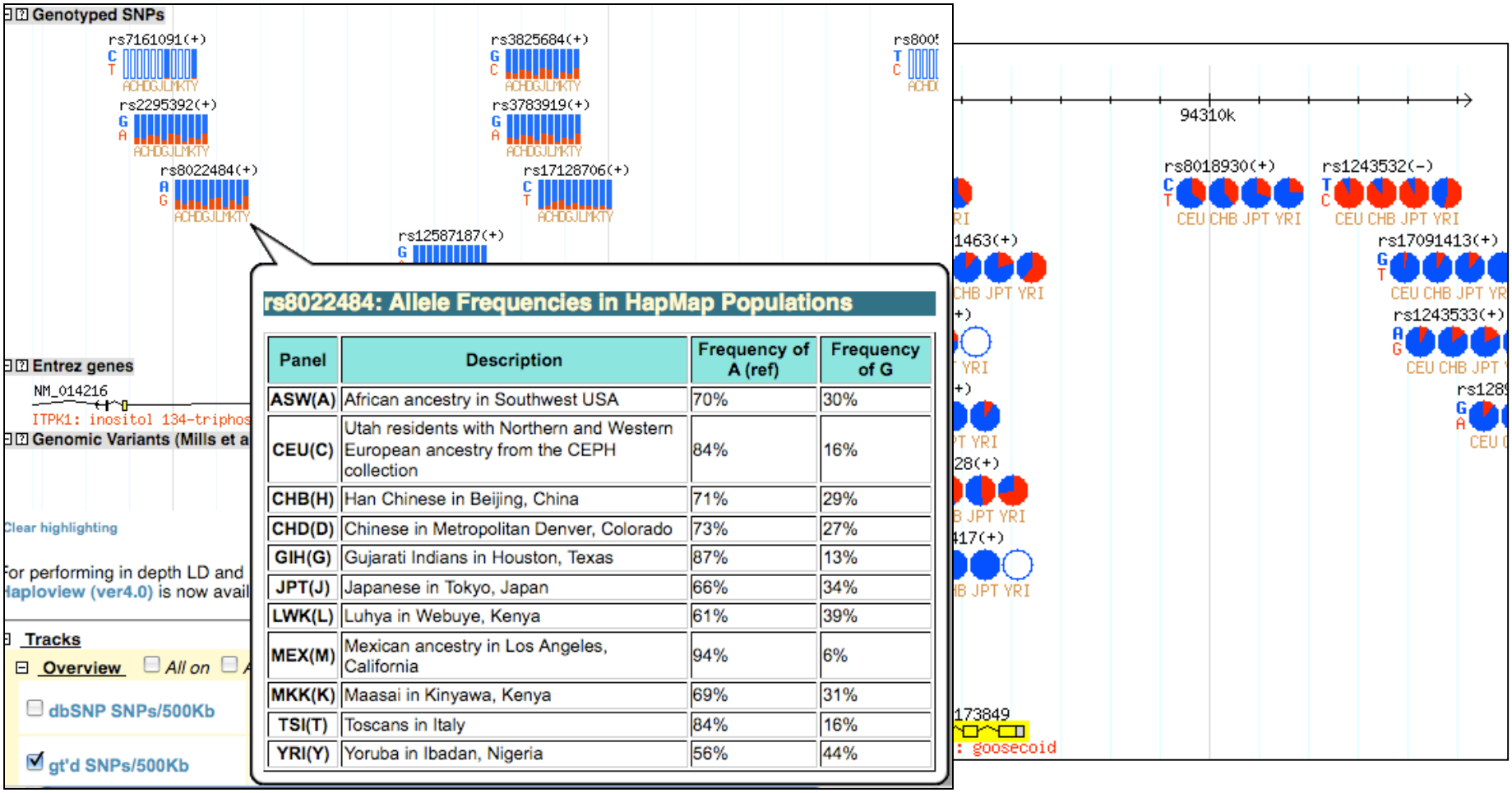



Ongoing work by Ben Faga, using PhyloGeoViz

GMOD

National Evolutionary Synthesis Center
NESCent

# 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% |

HapMap.org

GMOD

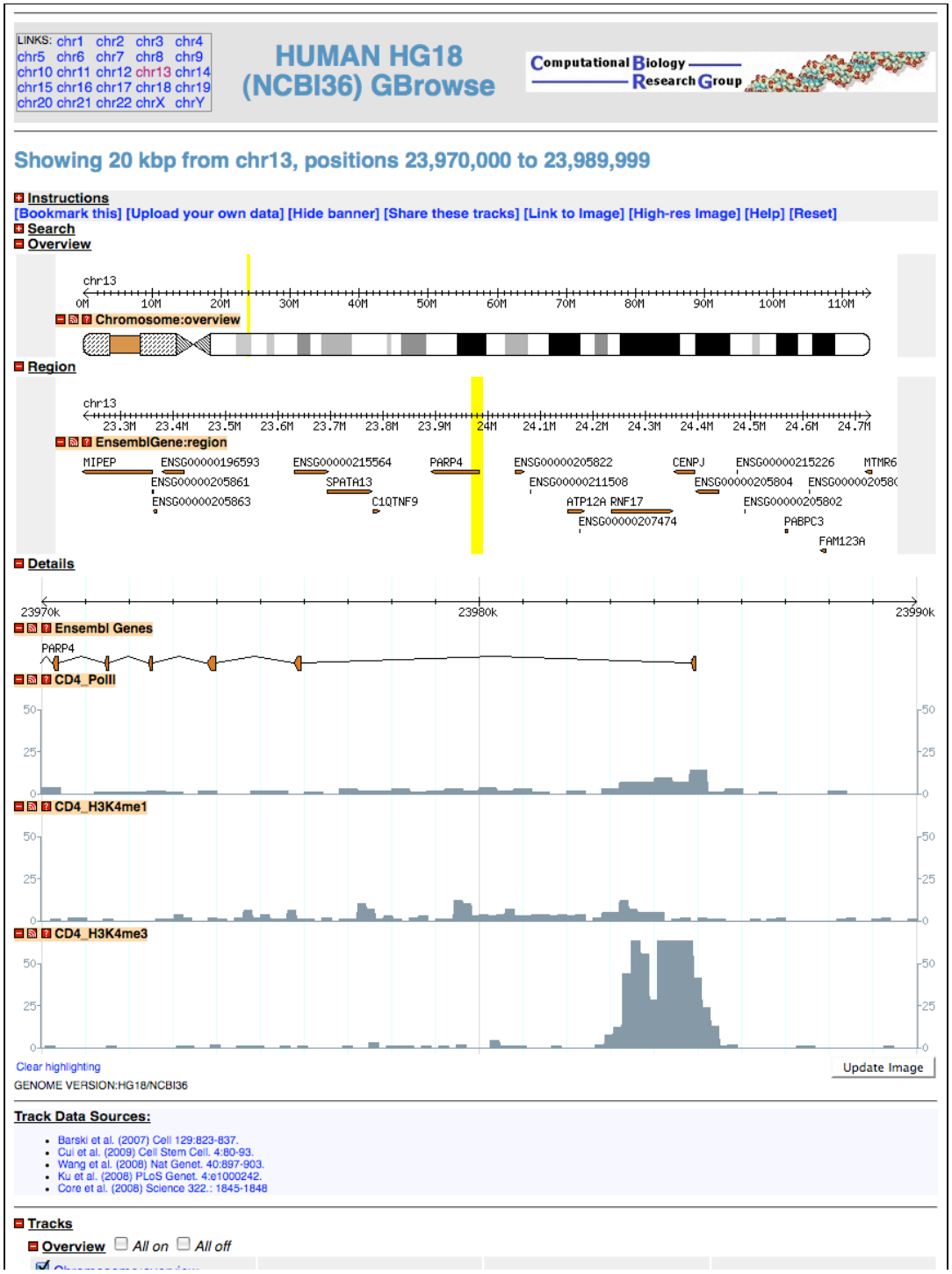
National Evolutionary Synthesis Center
NESCent

# Outline

- GBrowse as an alignment viewer
  - *E. coli*
  - Whole genome resequencing
- Next Generation Sequencing & Bioinformatics
- GBrowse for population genetics
  - Threespine Stickleback
  - Deep sequencing of select regions
  - looking for SNPs
- Other Visualisations
- GMOD Project
- Panel & Discussion


# Other Visualisations

- Methylation in human
- Mostly ChIP-Seq results
- Visualisation by Computational Biology Research Group at Oxford

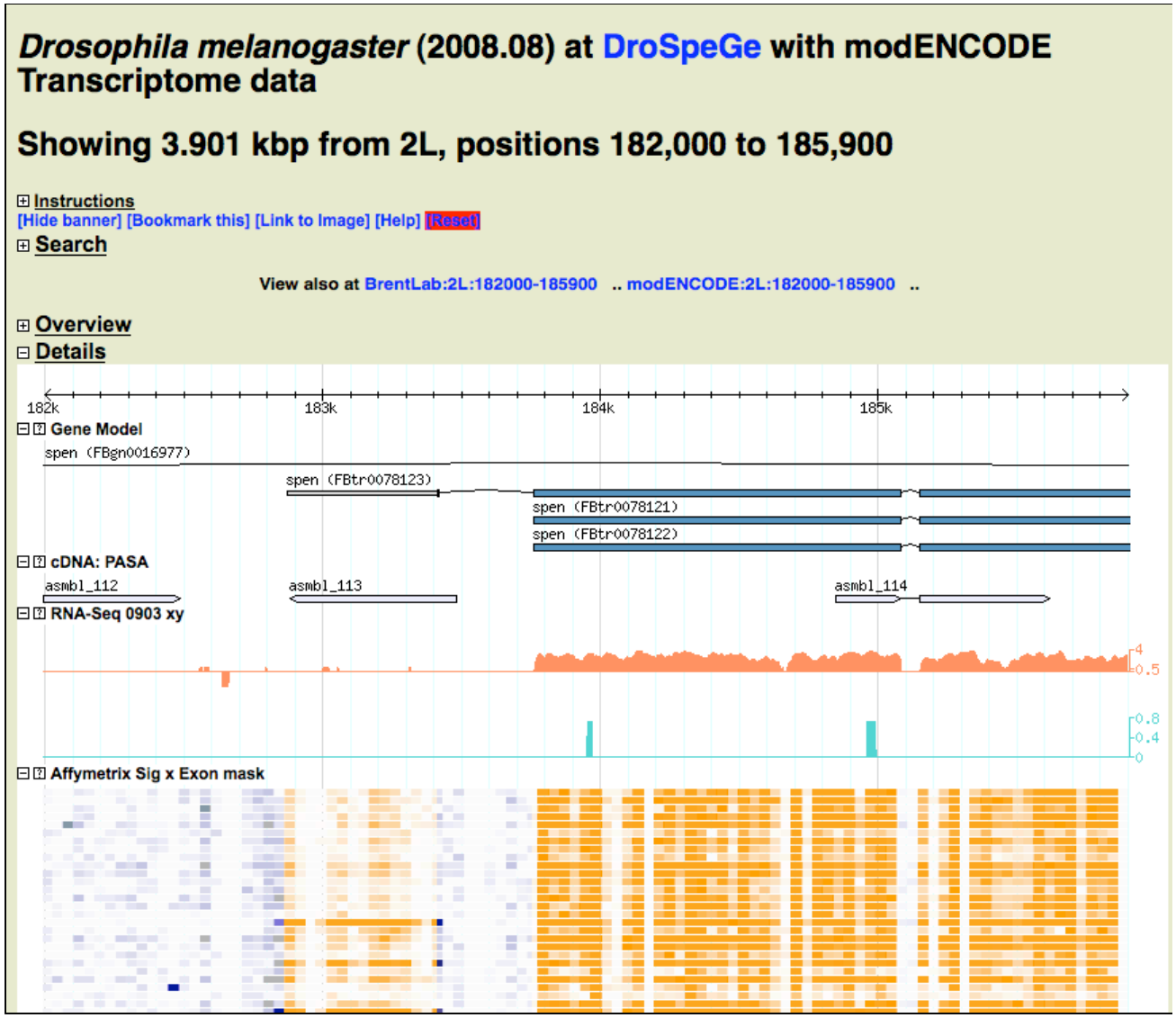


# Other Visualisations

Transriptome analysis for modENCODE

Custom modifications to some glyph code

Visualisation by Don Gilbert

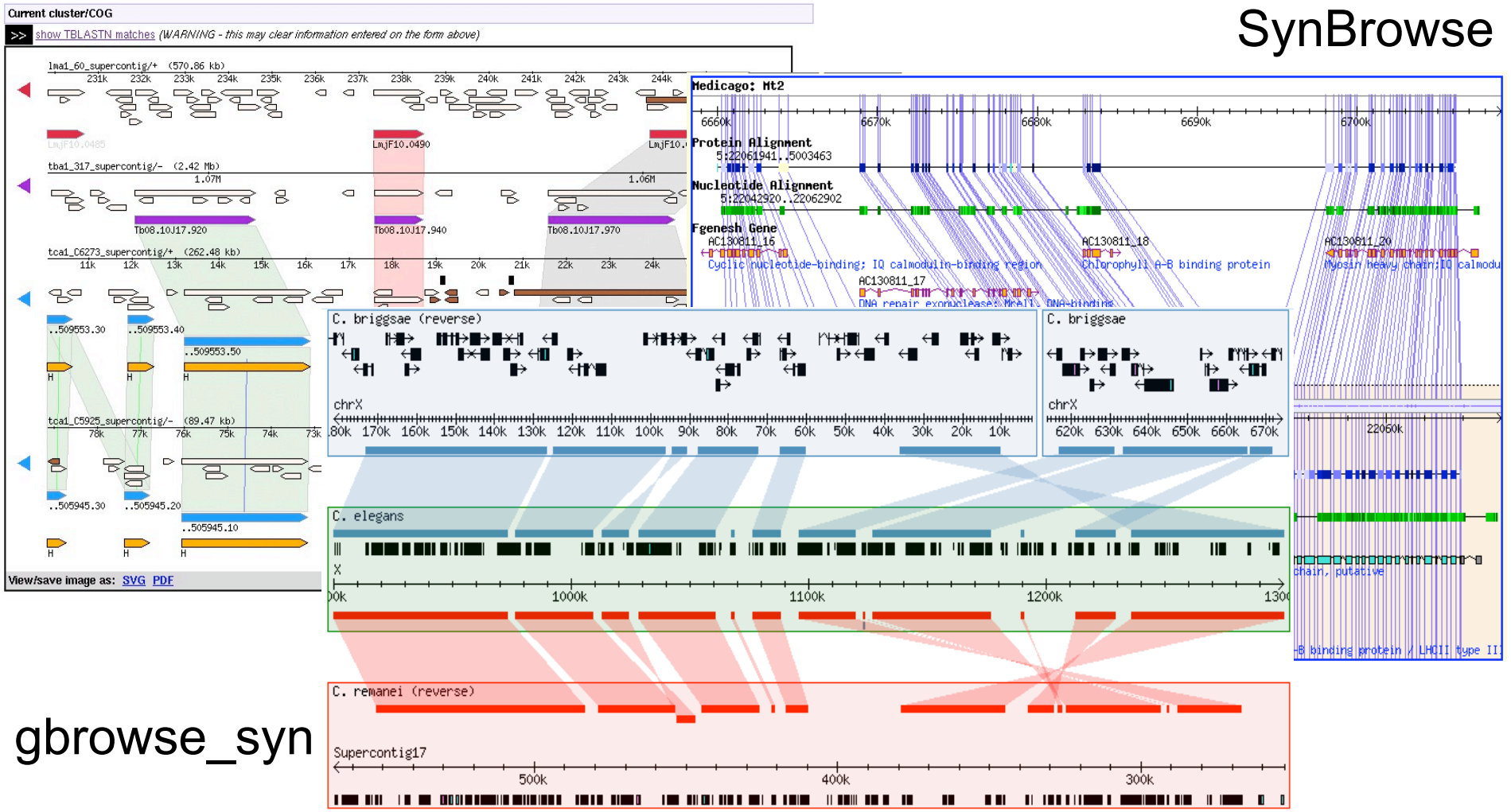



http://insects.eugenes.org/species/data/dmel5/modencode/bigmap/

# Comparative Genomics

Sybil

SynBrowse




gbrowse_syn

# Visualisation Conclusions

- You will need bioinformatics support
- Valuable insights come from analysis and summarization
- Anything that can be associated with sequence or genomic region can be visualized in a genome browser.


# Outline

- GBrowse as an alignment viewer
  - *E. coli*
  - Whole genome resequencing
- Next Generation Sequencing & Bioinformatics
- GBrowse for population genetics
  - Threespine Stickleback
  - Deep sequencing of select regions
  - looking for SNPs
- Other Visualisations
- GMOD Project
- Panel & Discussion


# GMOD is …

- A set of interoperable open source tools for many common biological database needs
  - GBrowse, Chado, Apollo, BioMart, CMap, Pathway Tools, Sybil, InterMine, MAKER …
- Active community of users and developers
  - Mailing lists, semi-annual meetings

http://gmod.org


GMOD AMERICAS 2009 Summer School

16-19 July, 2009

National Evolutionary Synthesis Center
(NESCent)
Durham, North Carolina, USA


3-7 August, 2009

University of Oxford, UK
3-6 August: GMOD Summer School
6-7 August: GMOD Community Meeting

http://gmod.org/wiki/GMOD_Summer_School
http://gmod.org/wiki/GMOD_Europe_2009

Summer school applications due 6 April!

GMOD

National Evolutionary Synthesis Center
NESCent

# Acknowledgements


**Organizers**
Wolfgang Stephan
Diethard Tautz

**Oxford**
Steve Taylor

**OICR**
Lincoln Stein
Scott Cain

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


**Panel**
Chuck Cannon
Philip Johnson
Phillip Morin
Korbinian Schneeberger

**Indiana**
Don Gilbert

**Iowa**
Ben Faga

**NESCent**
Todd Vision
Hilmar Lapp



# Outline

- GBrowse as an alignment viewer
  - *E. coli*
  - Whole genome resequencing
- Next Generation Sequencing & Bioinformatics
- GBrowse for population genetics
  - Threespine Stickleback
  - Deep sequencing of select regions
  - looking for SNPs
- Other Visualisations
- GMOD Project
- Panel & Discussion


# Visualisation Panel & Discussion

| | |
|---|---|
| **Chuck Cannon** | Chinese Academy of Sciences<br>Assembly free approaches |
| **Philip Johnson** | UC Berkeley<br>Metagenomics and gene flow |
| **Phillip Morin** | NOAA Fisheries & Scripps Institution of Oceanography<br>Natural diversity and Geolocation |
| **Korbinian Schneeberger** | Max Planck Institute for Developmental Biology<br>Arabidopsis: Deep and Wide |

GMOD

National Evolutionary Synthesis Center
NESCent

# 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