GBrowse and Next Generation Sequencing Data

Scott Cain[1], David Clements[2], Lincoln Stein[1]
and the GMOD Consortium

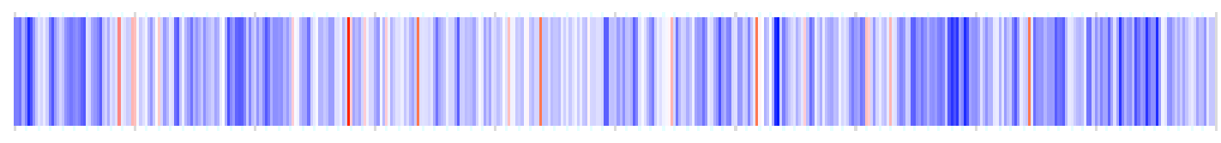
[1]Ontario Institute for Cancer Research, 101 College St, Toronto, Ontario, Canada, M5G 0A3
[2]National Evolutionary Synthesis Center, 2024 W. Main Street, Suite A200, Durham, NC 27705-4667 USA

The widespread adoption of next generation sequencing (NGS) technologies is generating large volumes of data that researchers need to visualize in order to fully exploit it. The new Bio::DB::Sam data adaptor enables the popular Generic Genome Browser (GBrowse)[1] (http://gmod.org/GBrowse) to present short read data from a SAMtools[2] (http://samtools.sourceforge.net/) generated database. SAMtools is an open source toolkit and common file format for storing NGS alignment data. Here we present examples of GBrowse using the Bio::DB::Sam adaptor with *E. coli* resequencing data as a proof of concept for using GBrowse as a NGS browser.

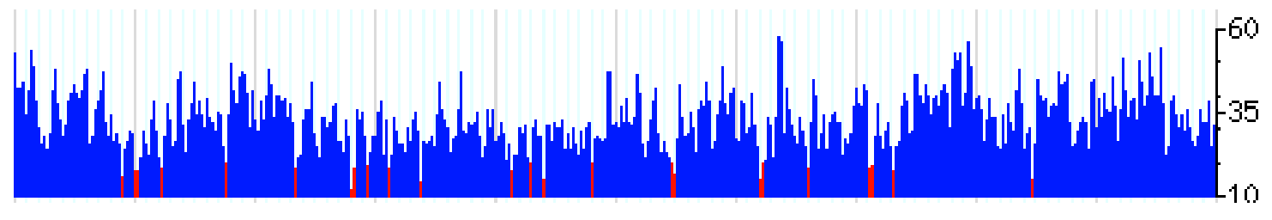
What SAMtools and Bio::DB::Sam Provide

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. BAM is a space-efficient indexed binary representation of SAM that is optimized for rapid retrieval of mapped alignments that overlap a region of interest. Bio::DB::Sam is a GBrowse data adaptor that allows GBrowse to use the data in a BAM data file to produce four data representations:

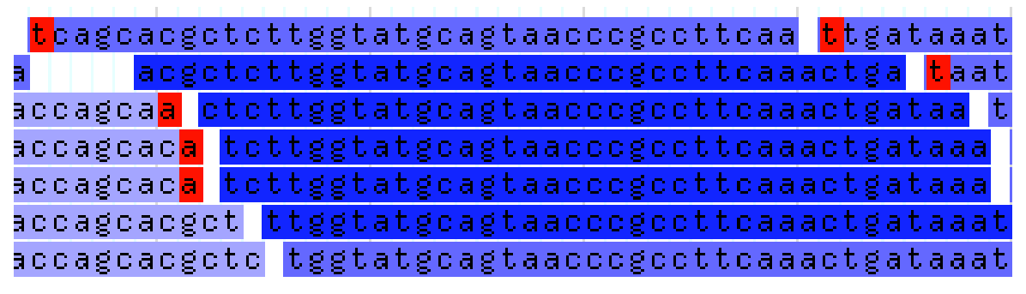
Coverage Density Plots



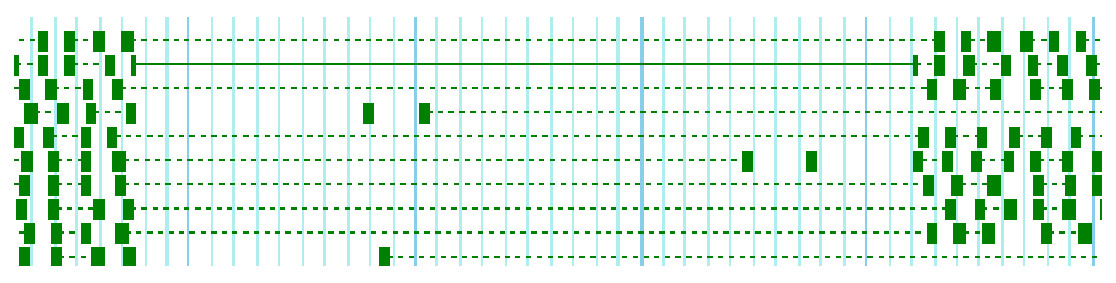
Coverage XY Plots



Individual Read Glyphs



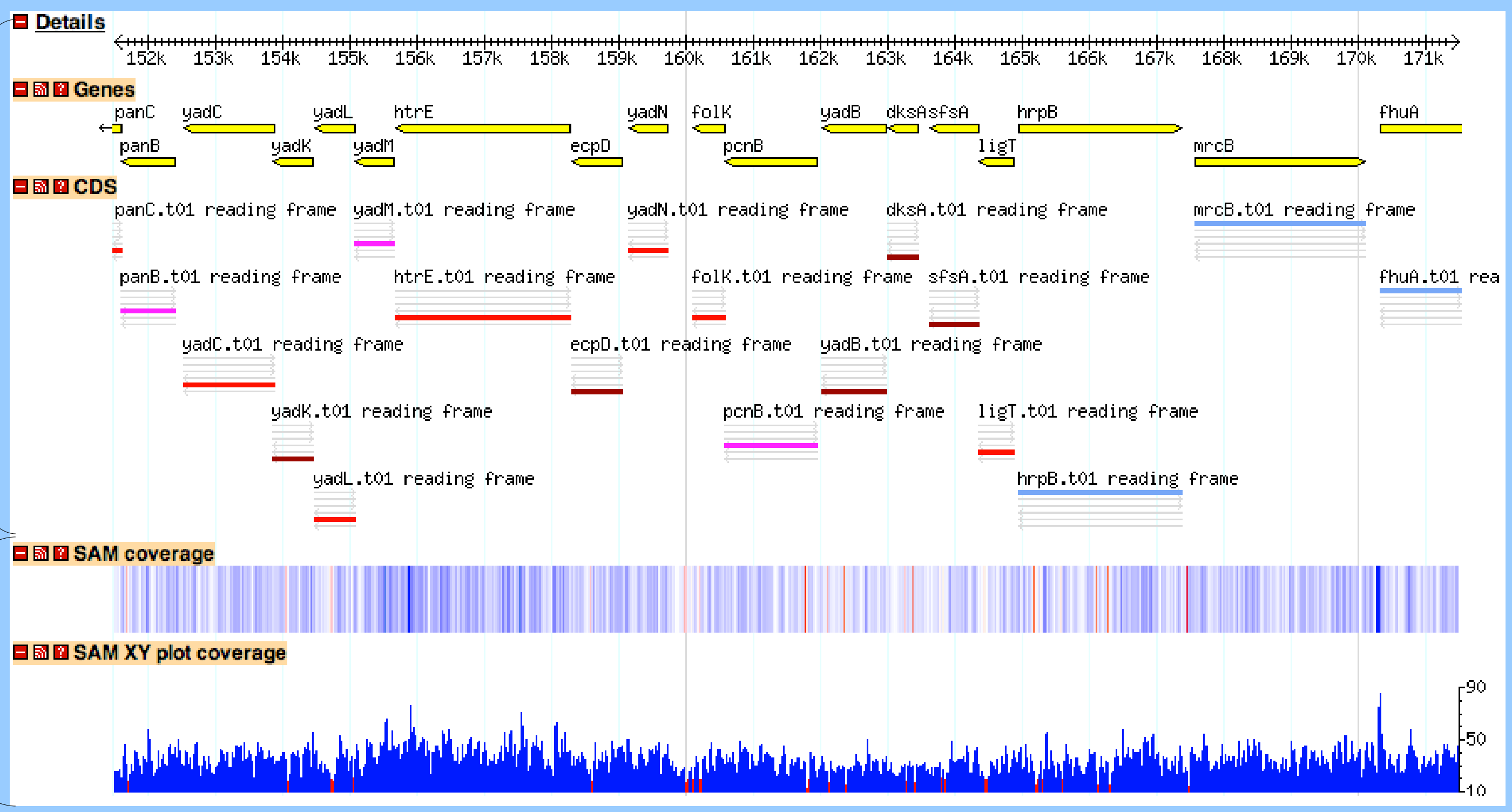
Paired Read Glyphs



GBrowse 1.70

Because the GBrowse 1.70 infrastructure requires one configuration file per data source, each BAM data set must have its own configuration file (*i.e.*, its own GBrowse source). Data can then be shared between GBrowse sources using the "share tracks" facility that is included with GBrowse. This is referred to as *gbgff sharing*, after the name of the cgi script that makes it possible. Sharing can be done either by the user by clicking on the "Share this track" link for a given track, or by the administrator placing a "remote feature" directive in the configuration file. Sharing data like this does have some drawbacks for BAM data, as BAM read glyphs loose their sequence and quality score data, so glyphs cannot be colored by quality or have their sequence mismatches highlighted.

Sharing from a SAM data source into a genome annotation data source



From the main "genome" GBrowse data source

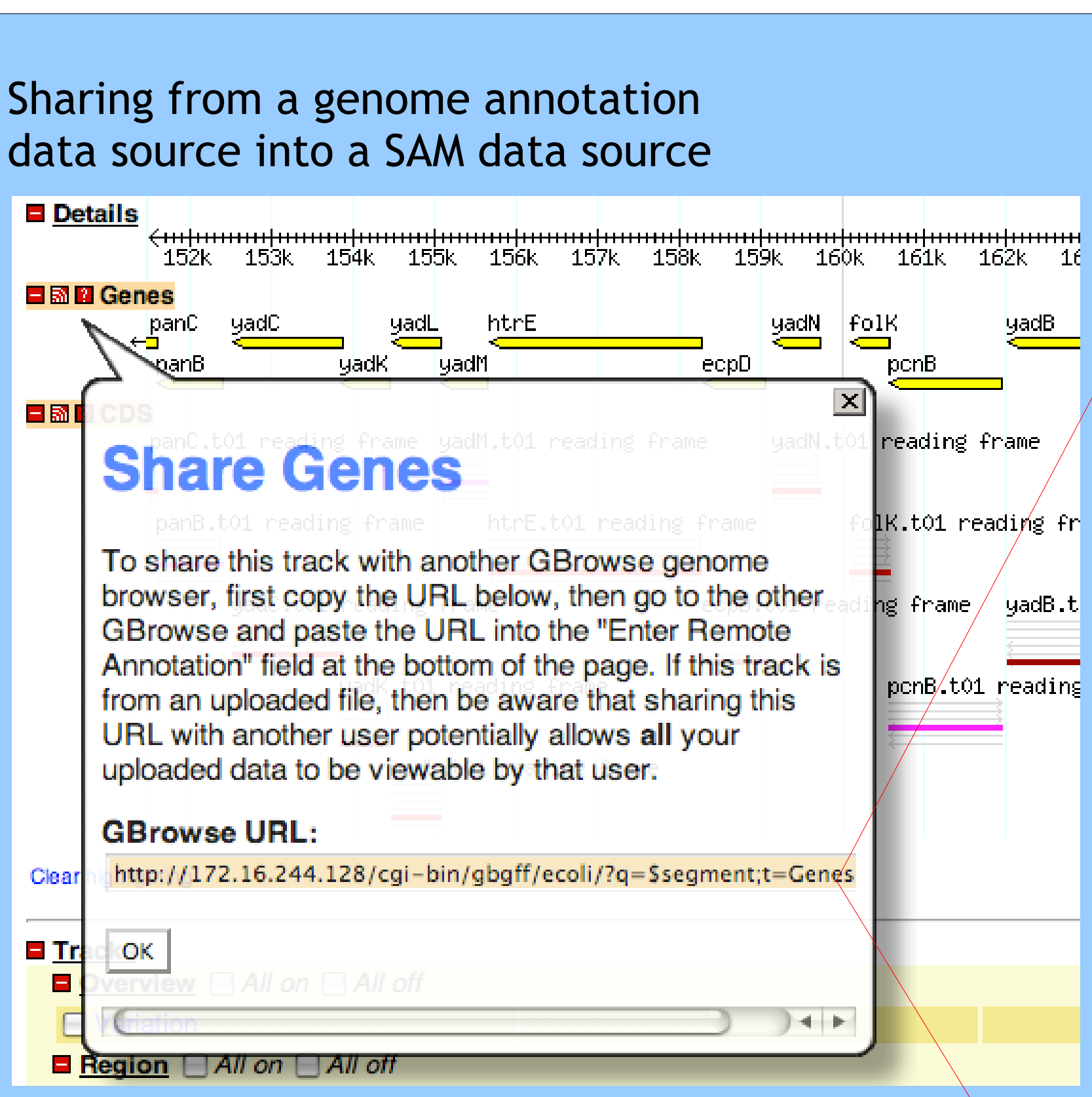
Shared from a Bio::DB::Sam data source

This method of sharing works fine for displaying coverage plots and allows multiple SAM data sources to be included. However, when zoomed in, individual read information like sequence and quality are lost.

Simple "remote feature" track definition:

```
[sam_xy_coverage]
remote feature = http://server.com/cgi-bin/gbgff/ecolisam/?q=$segment;t=Derived1CoverageXyplot;c=1
key            = SAM XY plot coverage
```

Sharing from a genome annotation data source into a SAM data source



Share Genes

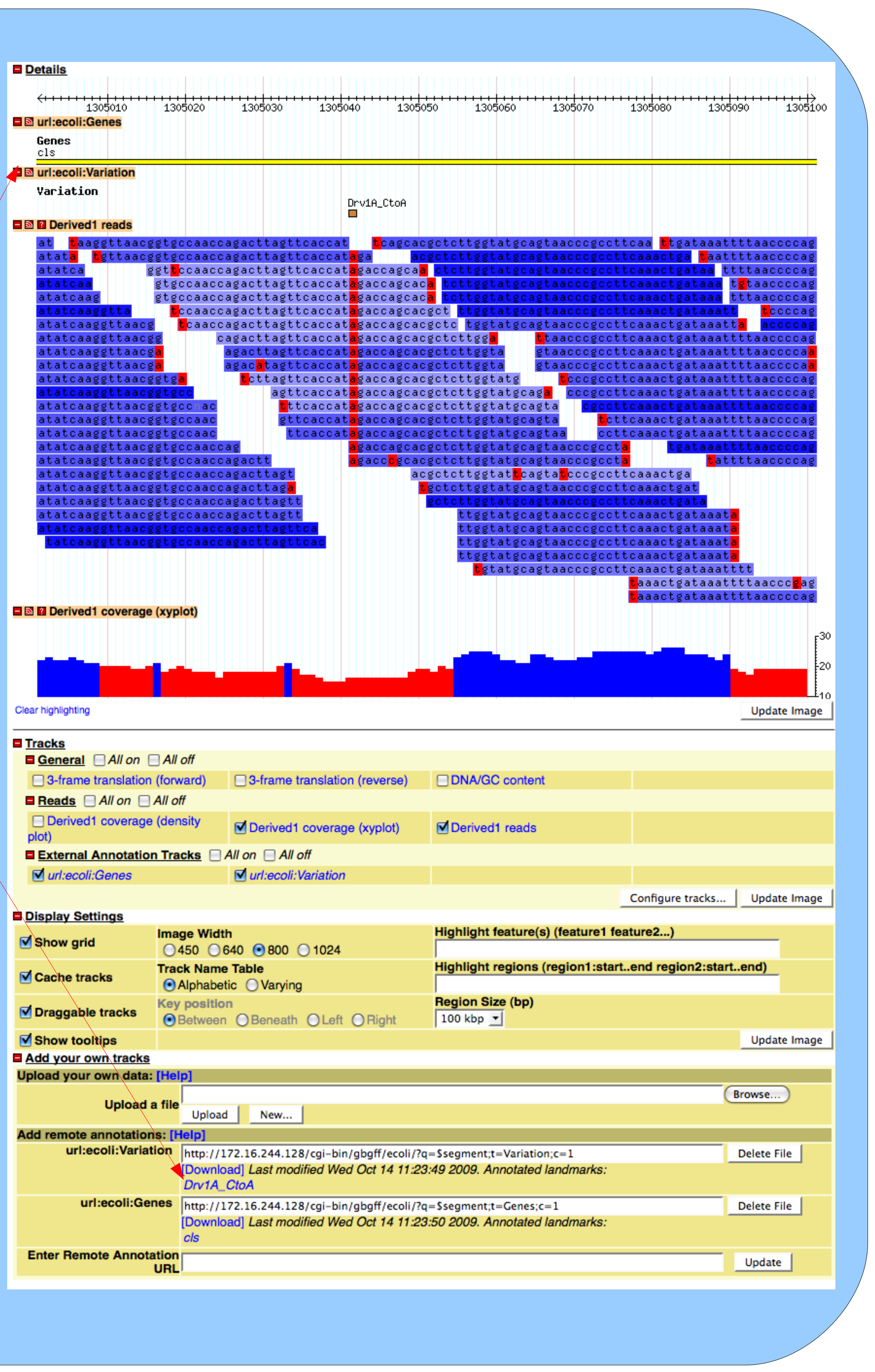
To share this track with another GBrowse genome browser, first copy the URL below, then go to the other GBrowse and paste the URL into the "Enter Remote Annotation" field at the bottom of the page. If this track is from an uploaded file, then be aware that sharing this URL with another user potentially allows **all** your uploaded data to be viewable by that user.

GBrowse URL:

http://172.16.244.128/cgi-bin/gbgff/ecoli/?q=$segment;t=Genes

OK

The user shares the "Genes" and "Variation" tracks from the genome data source to the Bio::DB::Sam data source, allowing the read track to display quality and mismatch information. This allows the individual read information to be displayed, but sharing in this direction limits the display to one SAM data source at a time. Here a zoomed in view shows a SNP identified with the sequencing run.



Tracks
- General □ All on □ All off
  - □ 3-frame translation (forward) □ 3-frame translation (reverse) □ DNA/GC content
- Reads □ All on □ All off
  - □ Derived1 coverage (density plot) ☑ Derived1 coverage (xyplot) ☑ Derived1 reads
- External Annotation Tracks □ All on □ All off
  - ☑ url:ecoli:Genes ☑ url:ecoli:Variation

Configure tracks... Update Image

Display Settings
- ☑ Show grid — Image Width ○ 450 ○ 640 ● 800 ○ 1024 — Highlight feature(s) (feature1 feature2...)
- ☑ Cache tracks — Track Name Table ● Alphabetic ○ Varying — Highlight regions (region1:start..end region2:start..end)
- ☑ Draggable tracks — Key position ● Between ○ Beneath ○ Left ○ Right — Region Size (bp) 100 kbp
- ☑ Show tooltips — Update Image

Add your own tracks

Upload your own data: [Help]

Upload a file — Browse... Upload New...

Add remote annotations: [Help]

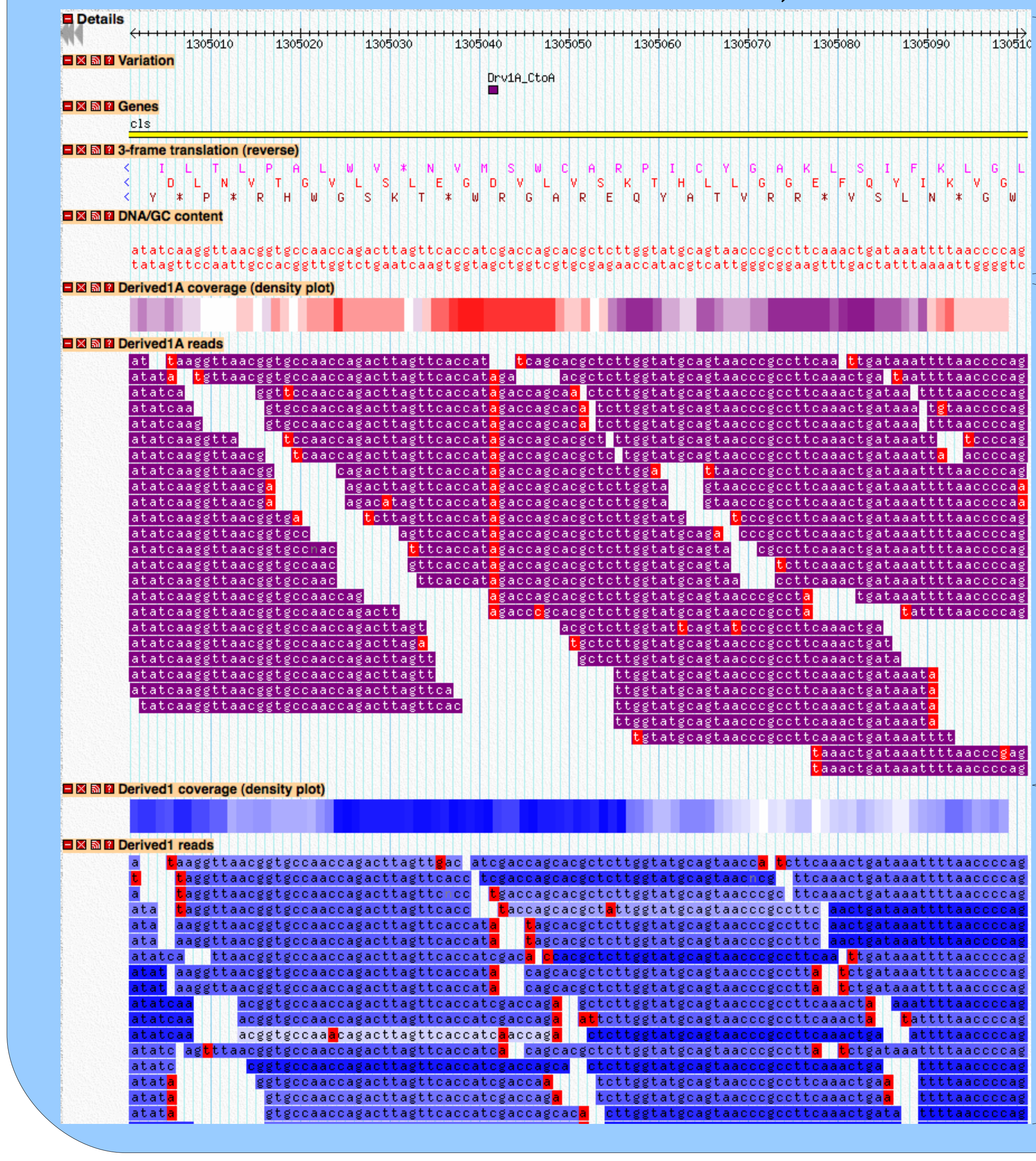
url:ecoli:Variation — http://172.16.244.128/cgi-bin/gbgff/ecoli/?q=$segment;t=Variation;c=1 — Delete File
[Download] Last modified Wed Oct 14 11:23:49 2009. Annotated landmarks: Drv1A_CtoA

url:ecoli:Genes — http://172.16.244.128/cgi-bin/gbgff/ecoli/?q=$segment;t=Genes;c=1 — Delete File
[Download] Last modified Wed Oct 14 11:23:50 2009. Annotated landmarks: cls

Enter Remote Annotation URL — Update

GBrowse 2.0

There are several improvements in GBrowse 2 over GBrowse 1.70, including support for distributed databases and image rendering and AJAX image upating (so that the whole page does not need to reloaded to view a new region or data track). GBrowse 2 also allows tracks to come from different data sources in the same page, so one track could come from a flat file, another from a Chado database and a third from a Bio::DB::SeqFeature::Store database. As a result, it is much easier to display next generation sequencing data in GBrowse 2 along with other annotations.



From a Bio::DB::SeqFeature::Store database

Configure GBrowse to use multiple data sources:

```
[genome:database]
db_adaptor    = Bio::DB::SeqFeature::Store
db_args       = -adaptor DBI::mysql
                -dsn dbi:mysql:ecoli
                -user apache
```

Then configure individual tracks to use different data sources for rendering:

```
[Genes]
feature       = gene
database      = genome
glyph         = gene
bgcolor       = yellow
forwardcolor  = yellow
reversecolor  = turquoise
height        = 6
description   = 0
key           = Genes
```

Tracks in GBrowse2 are configured in much the same way as GBrowse 1.70. The main addition is the "database" tag, which specifies where GBrowse2 should look for the data to create the track

From a BAM file

```
[ecolisam:database]
db_adaptor    = Bio::DB::Sam
db_args       = -bam   /var/www/gbrowse2/databases/ecolisam/bam1.bam
                -fasta /var/www/gbrowse2/databases/ecolisam/seq1.fa
search options= default
```

```
[Derived1ACoverageXyplot]
feature       = coverage:2000
glyph         = wiggle_xyplot
database      = ecolisam
height        = 50
fgcolor       = black
bicolor_pivot= 20
pos_color     = purple
neg_color     = red
key           = Derived1A coverage (xyplot)
category      = Reads
label         = 0
```

From another BAM file

```
[ecoliAsam:database]
db_adaptor    = Bio::DB::Sam
db_args       = -bam   /var/www/gbrowse2/databases/ecoliAsam/bam1.bam
                -fasta /var/www/gbrowse2/databases/ecoliAsam/seq1.fa
search options= default
```

```
[Derived1Reads]
feature         = match
glyph           = segments
draw_target     = 1
show_mismatch   = 1
mismatch_color = red
database        = ecoliAsam
bgcolor         = sub {
                 my $feature = shift;
                 my $blueness = sprintf("%X", 255 - $feature->qual * 2.4);
                 my $colour = chr(35) . $blueness . $blueness . "FF";
                 return $colour;
                 }
fgcolor         = black
height          = 5
label           = 0
bump            = fast
key             = Derived1 reads
category        = Reads
```

[1]Stein, L. D. et al. The generic genome browser: a building block for a model organism system database. Genome Res 12: 1599-610. [PMID: 12368253]
[2]Li H.* et al. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

The authors gratefully acknowledge Elizabeth Perry, Nicholas Stiffler, and Brendan Bohannon of the University of Oregon for providing the *E. coli* sequencing data

GMOD is supported by a specific cooperative agreement from the USDA Agricultural Research Service, and by NIH grants co-funded from the National Human Genome Research Institute and the National Institute of General Medical Sciences. GBrowse development is also funded by the Ontario Ministry of Research and Innovation.