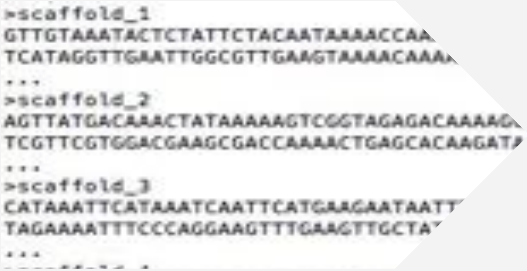
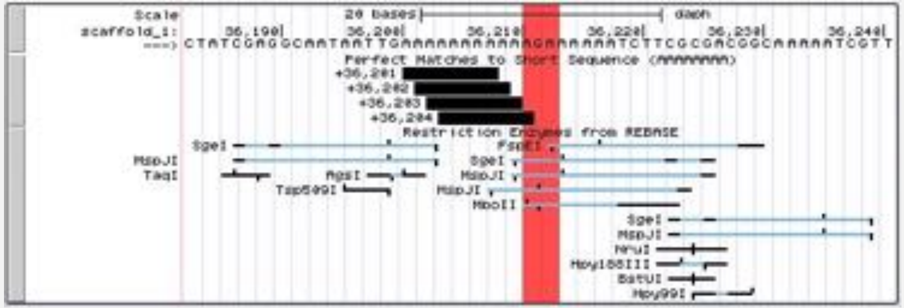
UCSC Browser Assembly Hubs
To Visualize and Analyze New Genomes

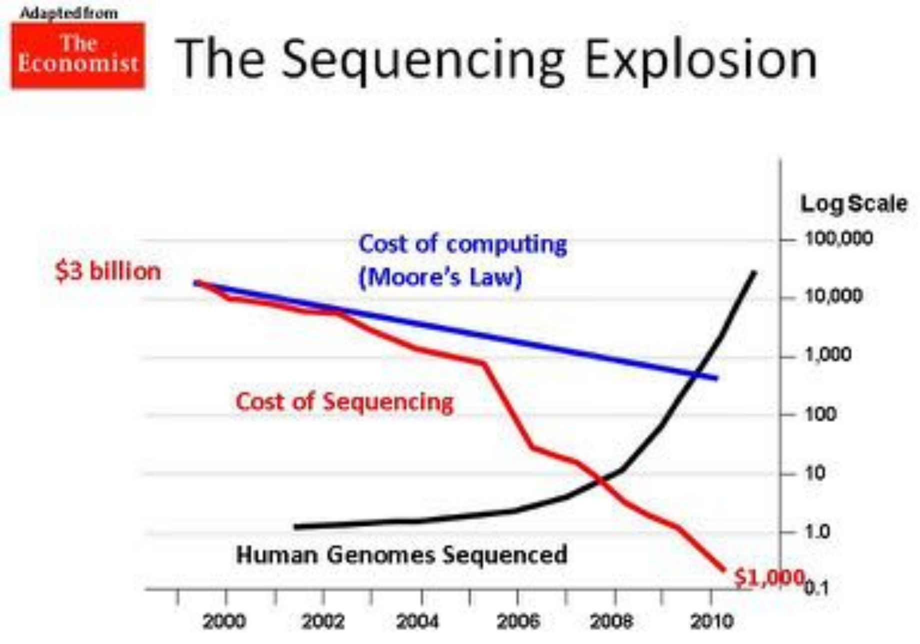


```
>scaffold_1
GTTGTAAATACTCTATTCTACAATAAAACCAA
TCATAGGTTGAATTGGCGTTGAAGTAAAACAAA
...
>scaffold_2
AGTTATGACAAACTATAAAAAGTCGGTAGAGACAAAAG
TCGTTCGTGGACGAAGCGACCAAAACTGAGCACAAGAT
...
>scaffold_3
CATAAATTCATAAATCAATTCATGAAGAATAATT
TAGAAAATTTCCCAGGAAGTTTGAAGTTGCTA
...
```

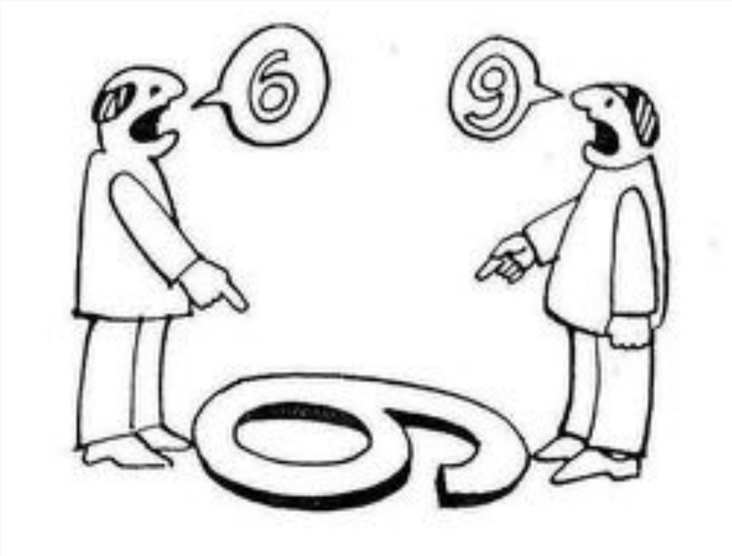
```
hubDirectory
|__hub.txt
|__genomes.txt
|__hg19
        |__trackDb.txt
```



UCSC Browser Assembly Hubs
Increased Availability of New Assemblies



UCSC Browser Assembly Hubs
Chance to View Data From a New Perspective



1. Binary Indexed Files used for Visualizing Data in Hubs

2. Text Files used to define Track Hubs

3. Examples of Existing Assembly Hubs and Analyzing Variants

Assembly Track Hub Summary:

Remotely Hosted Data

Binary indexed files:

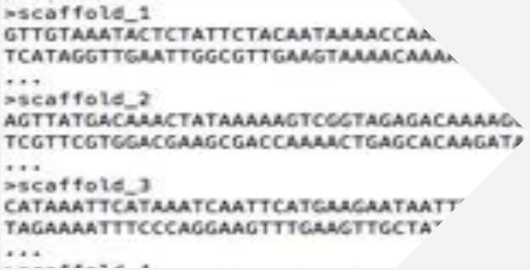
twoBitPath,

bigDataUrl

Text files: **hub.txt,**

genomes.txt,

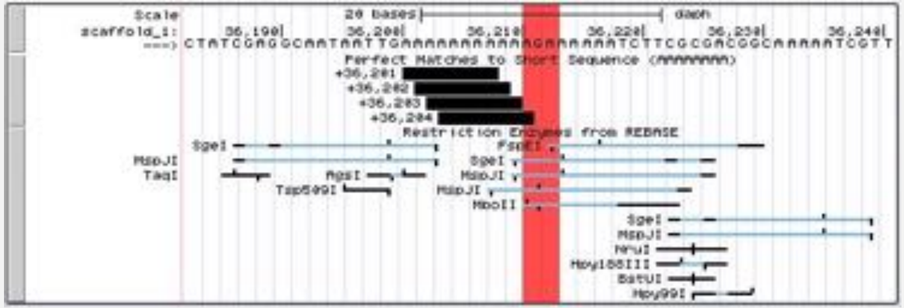
trackDb.txt



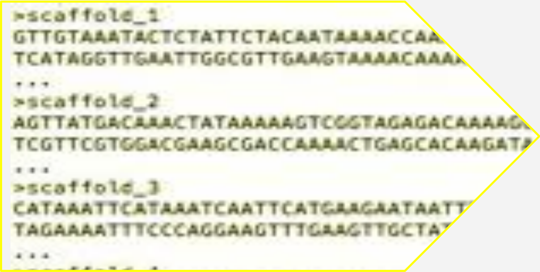
```
>scaffold_1
GTTGTAAATACTCTATTCTACAATAAAACCAA
TCATAGGTTGAATTGGCGTTGAAGTAAAACAAA
...
>scaffold_2
AGTTATGACAAACTATAAAAAGTCGGTAGAGACAAAAG
TCGTTCGTGGACGAAGCGACCAAAACTGAGCACAAGAT
...
>scaffold_3
CATAAATTCATAAATCAATTCATGAAGAATAATT
TAGAAAATTTCCCAGGAAGTTTGAAGTTGCTA
...
```

```
hubDirectory
|__hub.txt
|__genomes.txt
|__hg19
        |__trackDb.txt
```

Visualizable at UCSC

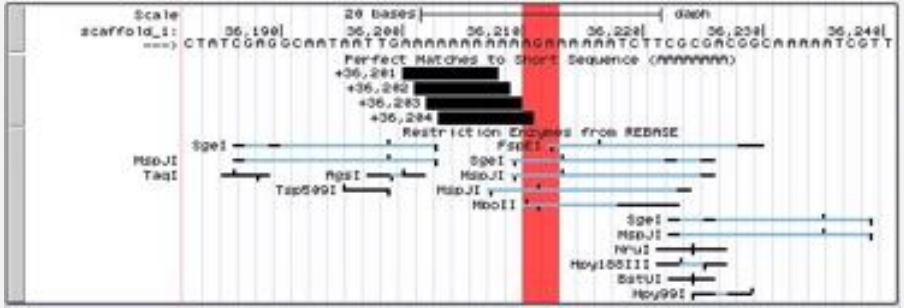


Building a 2bit file: binary indexed FASTA



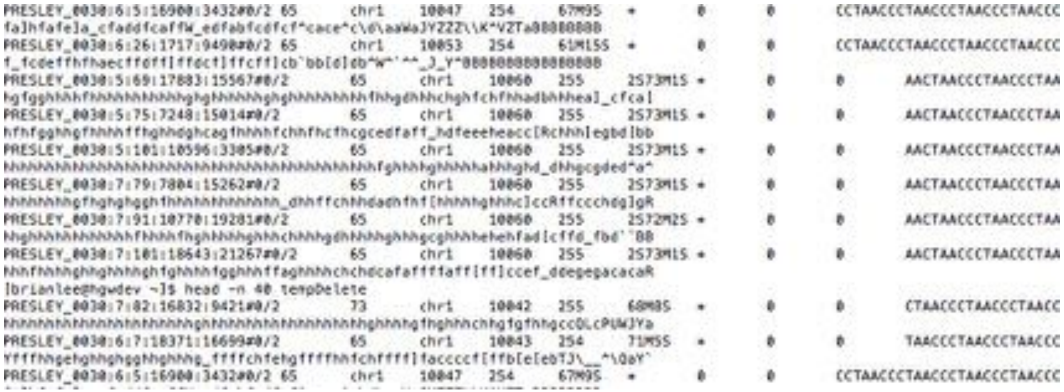
```
>scaffold_1
GTTGTAAATACTCTATTCTACAATAAAACCAA
TCATAGGTTGAATTGGCGTTGAAGTAAAACAAAA
...
>scaffold_2
AGTTATGACAAACTATAAAAAGTCGGTAGAGACAAAAG
TCGTTCGTGGACGAAGCGACCAAAACTGAGCACAAGATA
...
>scaffold_3
CATAAATTCATAAATCAATTCATGAAGAATAATT
TAGAAAATTTCCCAGGAAGTTTGAAGTTGCTA
...
```

```
hubDirectory
|__hub.txt
|__genomes.txt
|__hg19
      |__trackDb.txt
```



A BAM file (.bam) is the binary version of a SAM file.

```
samtools view -S -b sample.sam > sample.bam
```



```
PRESLEY_0030:6:5:16900:3432#0/2 65      chr1    10047   254     67M9S   *       0       0       CCTAACCCTAACCCTAACCCTAACCC
fa]hfafela_cfaddfcaffW_edfabfcdfcf^cace^c\d\aaWaJYZZZ\\K^VZTa0BBBBBBB
PRESLEY_0030:6:26:1717:9490#0/2 65      chr1    10053   254     61M15S  *       0       0       CCTAACCCTAACCCTAACCCTAACCC
f_fcdeffhfhaecffdff[ffdcf]ffcff]cb`bb[d]db^W^'^^_J_Y^BBBBBBBBBBBBBBBB
PRESLEY_0030:5:69:17883:15567#0/2       65      chr1    10060   255     2S73M1S *       0       0       AACTAACCCTAACCCTAA
hgfgghhhhfhhhhhhhhhhghghhhhhhghghhhhhhhhhfhhgdhhhchghfchfhhadbhhheal_cfca]
PRESLEY_0030:5:75:7248:15014#0/2        65      chr1    10060   255     2S73M1S *       0       0       AACTAACCCTAACCCTAA
hfhfgghhgfhhhhffhghhdghcagfhhhhfchhfhcfhcgcedfaff_hdfeeeheacc[RchhhlegbdIbb
PRESLEY_0030:5:101:10596:3305#0/2       65      chr1    10060   255     2S73M1S *       0       0       AACTAACCCTAACCCTAA
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhfghhhhghhhhhahhhghd_dhhgcgded^a^
PRESLEY_0030:7:79:7804:15262#0/2        65      chr1    10060   255     2S73M1S *       0       0       AACTAACCCTAACCCTAA
hhhhhhhhgfhghghgghfhhhhhhhhhhhhh_dhhffchhhdaehfhf[hhhhhghhhc]ccRffccchdg]gR
PRESLEY_0030:7:91:10770:19281#0/2       65      chr1    10060   255     2S72M2S *       0       0       AACTAACCCTAACCCTAA
hhghhhhhhhhhhhfhhhhfhghhhhhghhhchhhhgdhhhhghhhgcghhhhehehfad[cffd_fbd``BB
PRESLEY_0030:7:101:18643:21267#0/2      65      chr1    10060   255     2S73M1S *       0       0       AACTAACCCTAACCCTAA
hhhfhhhhghhghhhhghfghhhhfgghhhffaghhhhchchdcafafffaff[ff]ccef_ddegepacacaR
[brianlee@hgwdev ~]$ head -n 40 tempDelete
PRESLEY_0030:7:82:16832:9421#0/2        73      chr1    10042   255     68M8S   *       0       0       CTAACCCTAACCCTAACC
hhhhhhhhhhhhhhhhhhhhhhhghhhhhhhhhhhhhhhhhhhhhhghhhhgfhghhhchhgfgfhhgccQLcPUWJYa
PRESLEY_0030:6:7:18371:16699#0/2        65      chr1    10043   254     71M5S   *       0       0       TAACCCTAACCCTAACCC
Yfffhhgehghhghgghhghhhg_ffffchfehgffffhhfchffff]faccccf[ffb[e[ebTJ\__^\QaY`
PRESLEY_0030:6:5:16900:3432#0/2 65      chr1    10047   254     67M9S   *       0       0       CCTAACCCTAACCCTAACCCTAACCC
```

The resulting binary file **sample.bam** can have data more easily extracted and can also be viewed in Genome Browsers.

Sequence Alignment Map (SAM) is a text-based format for storing biological sequences ... The binary equivalent of a SAM file is a Binary Alignment Map (BAM) file,

A 2bit file is a binary indexed version of a FASTA file (stores sequence ACGT as 00 01 11 10)

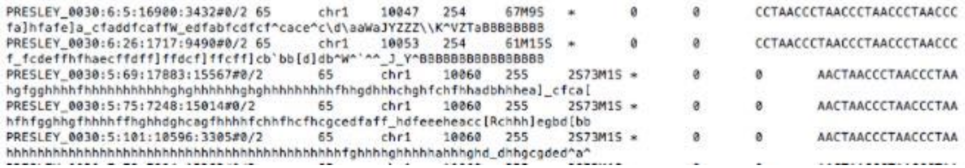
```
faToTwoBit input.fasta output.2bit
```

```
>scaffold_1
GTTGTAAATACTCTATTCTACAATAAAACCAAAAGATAACTCTTTATCAG
TCATAGGTTGAATTGGCGTTGAAGTAAAACAAAAAATACTGCTCAAAAGG
...
>scaffold_2
AGTTATGACAAACTATAAAAAGTCGGTAGAGACAAAAGCGAAAAGGATCT
TCGTTCGTGGACGAAGCGACCAAAACTGAGCACAAGATAAATCCCGAATA
...
>scaffold_3
CATAAATTCATAAATCAATTCATGAAGAATAATTTTAGAAAAATGGTTCA
TAGAAAATTTCCCAGGAAGTTTGAAGTTGCTATAATGATTATTTCTCTTG
...
>scaffold_4
ATACTGCCCATGCCTCATCTACTCCCTTCTTAACCTTTTCATCAACAACT
TGCTTAAGTTGGTCATCACTGACTGAGTGAGCCCAAACTGGAATAGCTGT
...
```

The resulting binary file **output.2bit** can have data more easily extracted and can also be viewed in Genome Browsers.

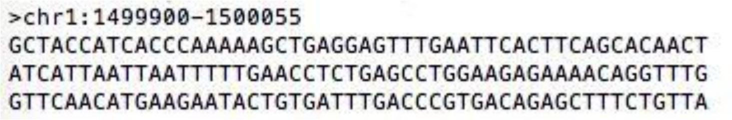
Extracting a specific window location of data from a BAM and 2bit file

```
samtools view http://location_of/file.bam  "chr1:1499900-1500055" >
output.sam
```



```
PRESLEY_0030:6:5:16900:3432#0/2 65      chr1    10047   254     67M9S   *       0       0       CCTAACCCTAACCCTAACCCTAACCC
fa]hfafe]a_cfaddfcaffW_edfabfcdfcf^cace^c\d\aaWaJYZZZ\\K^VZTaBBBBBBBB
PRESLEY_0030:6:26:1717:9490#0/2 65      chr1    10053   254     61M15S  *       0       0       CCTAACCCTAACCCTAACCCTAACCC
f_fcdeffhfhaecffdff]ffdcf]ffcff]cb`bb[d]db^W^'^^_J_Y^BBBBBBBBBBBBBBBB
PRESLEY_0030:5:69:17883:15567#0/2       65      chr1    10060   255     2S73M1S *       0       0       AACTAACCCTAACCCTAA
hgfgghhhhfhhhhhhhhhhhghghhhhhhghghhhhhhhhhfhhgdhhhchghfchfhhadbhhhea]_cfca[
PRESLEY_0030:5:75:7248:15014#0/2        65      chr1    10060   255     2S73M1S *       0       0       AACTAACCCTAACCCTAA
hfhfgghhgfhhhhffhghhdghcagfhhhhfchhfhcfhcgcedfaff_hdfeeeheacc[Rchhh]egbd[bb
PRESLEY_0030:5:101:10596:3305#0/2       65      chr1    10060   255     2S73M1S *       0       0       AACTAACCCTAACCCTAA
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhfghhhhghhhhhahhhghd_dhhgcgded^a^
```

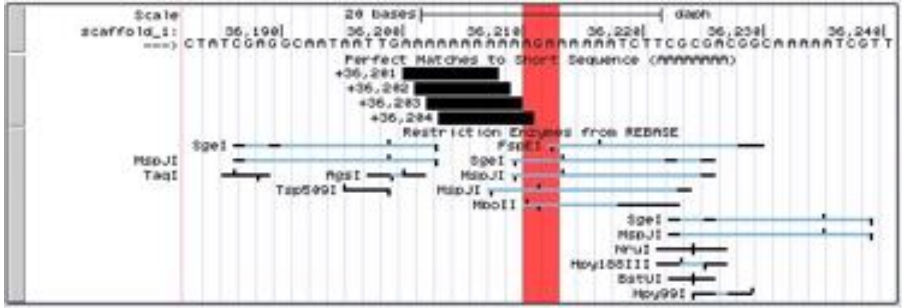
```
twoBitToFa -seq=chr1 -start=1499900 -end=1500055
twoBitPath http://yourGenome/your.2bit output.fa
```



```
>chr1:1499900-1500055
GCTACCATCACCCAAAAAGCTGAGGAGTTTGAATTCACTTCAGCACAACT
ATCATTAATTAATTTTTTGAACCTCTGAGCCTGGAAGAGAAAACAGGTTTG
GTTCAACATGAAGAATACTGTGATTTGACCCGTGACAGAGCTTTCTGTTA
```

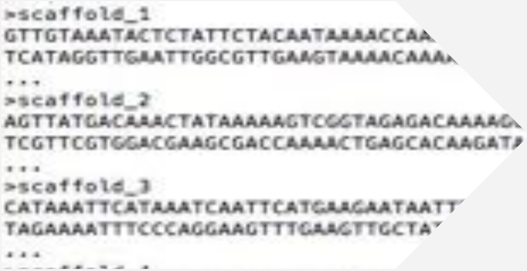
Assembly Track Hubs

bigDataUrl *http://location_of/file.bam*



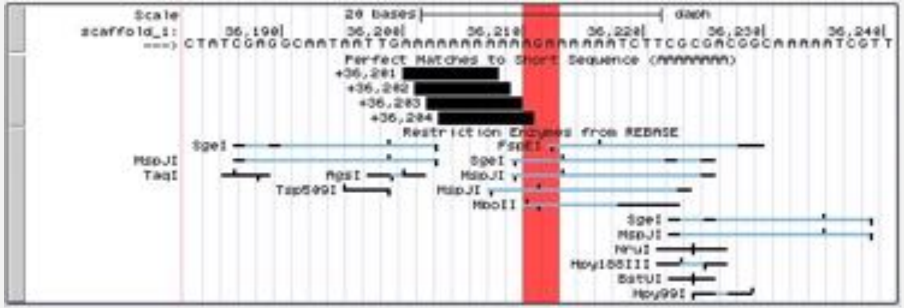
twoBitPath *http://yourGenome/your.2bit*

Building the Hub Text (.txt) Files



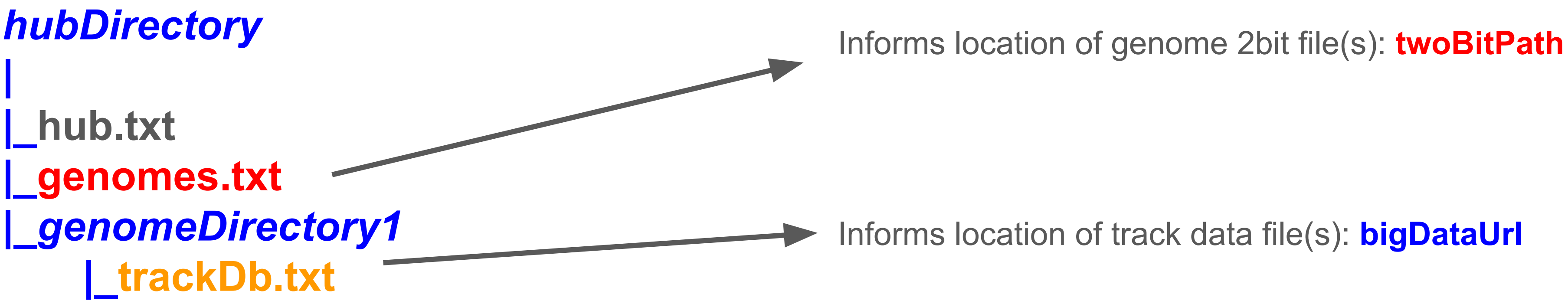
```
>scaffold_1
GTTGTAAATACTCTATTCTACAATAAAACCAA
TCATAGGTTGAATTGGCGTTGAAGTAAAACAAAA
...
>scaffold_2
AGTTATGACAAACTATAAAAAGTCGGTAGAGACAAAAG
TCGTTCGTGGACGAAGCGACCAAAACTGAGCACAAGAT
...
>scaffold_3
CATAAATTCATAAATCAATTCATGAAGAATAATT
TAGAAAATTTCCCAGGAAGTTTGAAGTTGCTA
...
```

```
hubDirectory
|__hub.txt
|__genomes.txt
|__hg19
        |__trackDb.txt
```



Track Hubs

Track Hubs are text files (**hub.txt, genomes.txt, trackDb.txt**) that describe and point to the location of binary indexed data files (**twoBitPath, bigDataUrl**)



hubDirectory
|
|_hub.txt
|_genomes.txt
|_*genomeDirectory1*
    |_trackDb.txt

Informs location of genome 2bit file(s): **twoBitPath**

Informs location of track data file(s): **bigDataUrl**

Track Hubs

Track Hubs are text files (**hub.txt, genomes.txt, trackDb.txt**) that describe and point to the location of binary indexed data files (**twoBitPath, bigDataUrl**)

hub.txt

```
hub MyHubsNameWithoutSpaces
shortLabel My Hub's Name
longLabel  Longer label about my hub.
email myEmail@address
genomesFile http://location_of/genomes.txt
```

genomes.txt

```
genome yourGenome
trackDb http://location_of/trackDb.txt
twoBitPath http://location_of/output.2bit
...
```

trackDb.txt

```
track cytoBandIdeo
type bigBed
longLabel Chromosome ideogram with cytogenetic bands
shortLabel cytoBandIdeo
bigDataUrl http://location_of/yourCytoBandIdeo.bigBed

track bam1
type bam
shortLabel BAM example
longLabel A BAM file for my genome data
visibility dense
bigDataUrl http://location_of/file.bam
...
```

Each new track gets a new stanza

bigDataUrl lines point to location of binary indexed data

More trackDb parameters can be defined too (color, ect.)

Viewing the Assembly Hub (2bit file) at UCSC

Browser requests *AGCTs* for only the genomic window currently viewing

UCSC Genome Browser on Your organism Feb. 2017 Assembly Assembly (yourGenome)

PAG:30-99 70 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Base Position: full

Restr Enzymes: hide

Short Match: pack

Adding *Browser Extensible Data* (BED) Annotations

text-based BED custom tracks: chromosome start end

track name=ex1 description="Blue example BED3 track" color=0,0,255

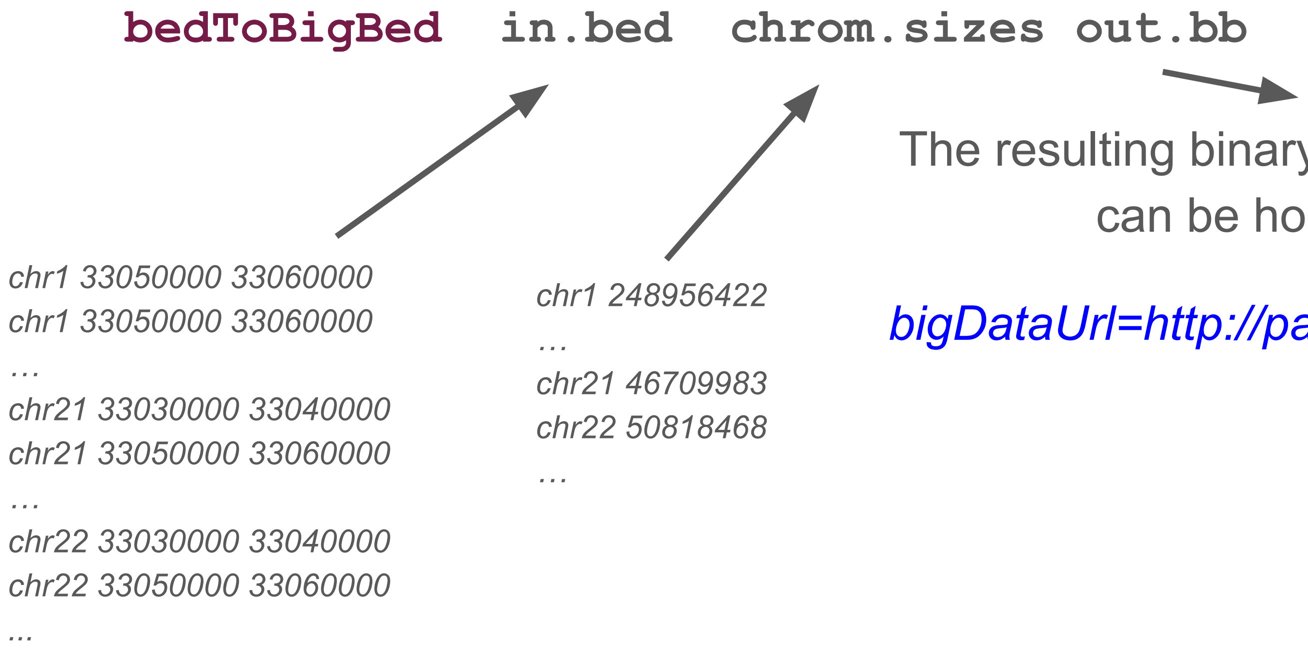
chr21 33030000 33040000

chr21 33050000 33060000

…

Adding tracks inside Track Hubs

create binary indexed versions of text files (`bedToBigBed`)



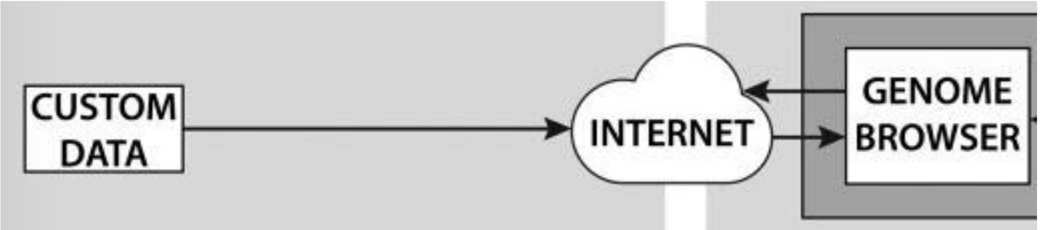
```
bedToBigBed  in.bed  chrom.sizes out.bb
```

```
chr1 33050000 33060000
chr1 33050000 33060000
…
chr21 33030000 33040000
chr21 33050000 33060000
…
chr22 33030000 33040000
chr22 33050000 33060000
...
```

```
chr1 248956422
…
chr21 46709983
chr22 50818468
…
```

The resulting binary file **out.bb** (bigBed) can be hosted online:

*bigDataUrl=http://path.lab.edu/to/**out.bb***



```
bigDataUrl http://path.lab.edu/to/out.bb
```

The **trackDb.txt** stanzas descriptions and where to find binary indexed tracks

```
track track1bam
bigDataUrl http://location_of/file.bam

track exampleBigBed
type bigBed
shortLabel ex1
longLabel   Blue example BED3 track
color 0,0,255
bigDataUrl http://path.lab.edu/to/out.bb
...
```

View Data Tracks at UCSC

A wide variety of track binary indexed formats exist:

- *BAM, CRAM, VCF*
- *bigGenePred, bigBarChart*
- *bigPsl, bigChain, bigMaf,*
- *bigNarrowPeak, halSnake*
- *bigBed, bigWig*

Additional Track Types In Hubs Example

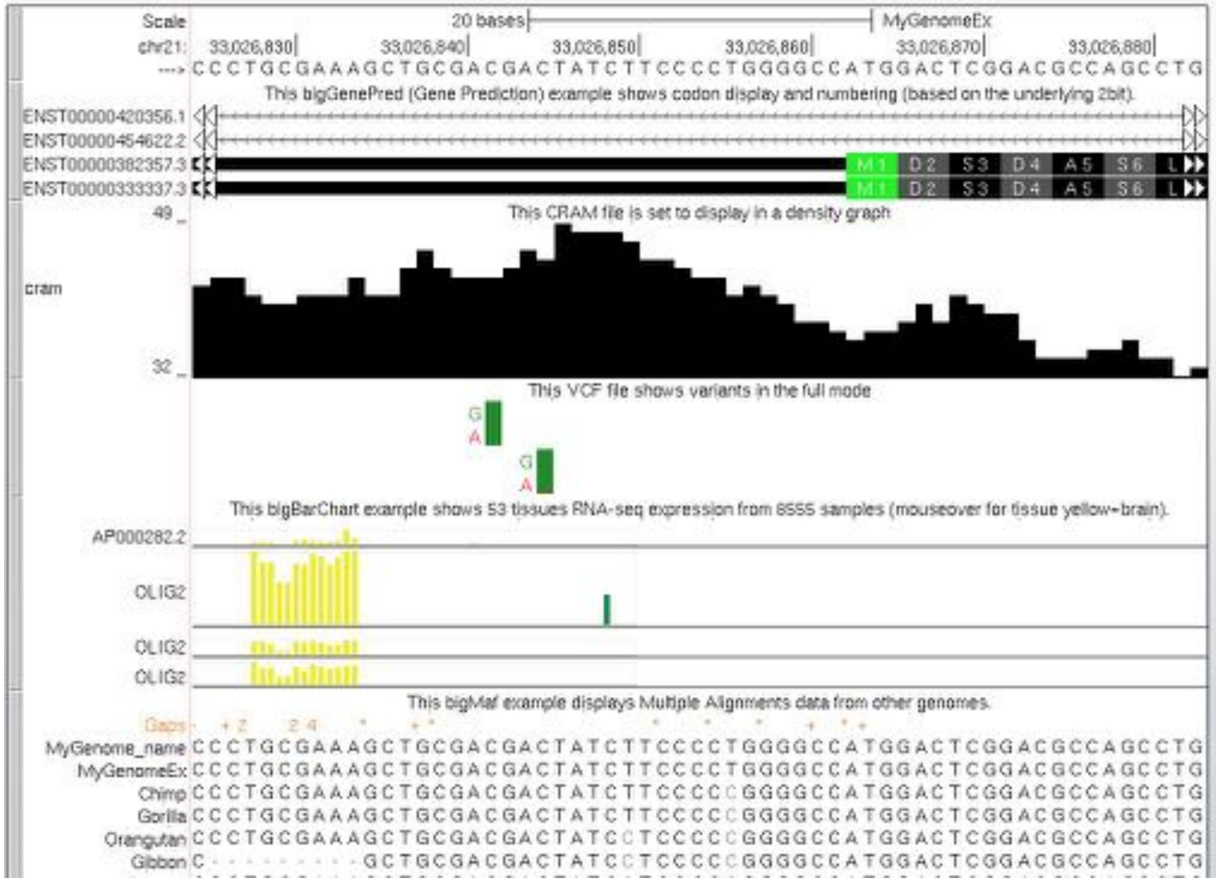
bigGenePred

CRAM

VCF

bigBarChart

bigMaf



This bigGenePred (Gene Prediction) example shows codon display and numbering (based on the underlying 2bit).

This CRAM file is set to display in a density graph

This VCF file shows variants in the full mode

This bigBarChart example shows 53 tissues RNA-seq expression from 8555 samples (mouseover for tissue yellow=brain).

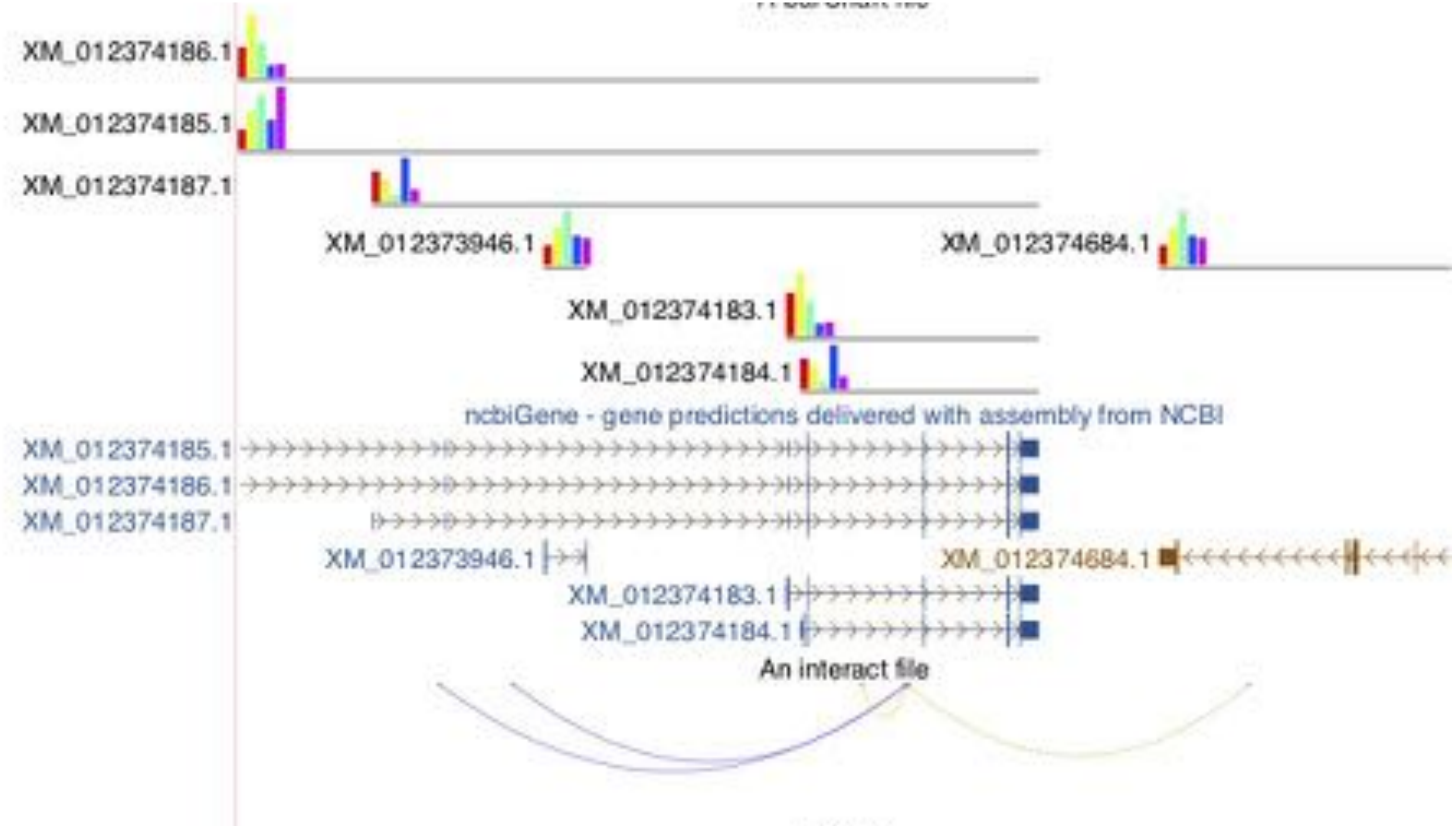
This bigMaf example displays Multiple Alignments data from other genomes.

Additional Track Types In Hubs Example

bigBarChart

bigGenePred

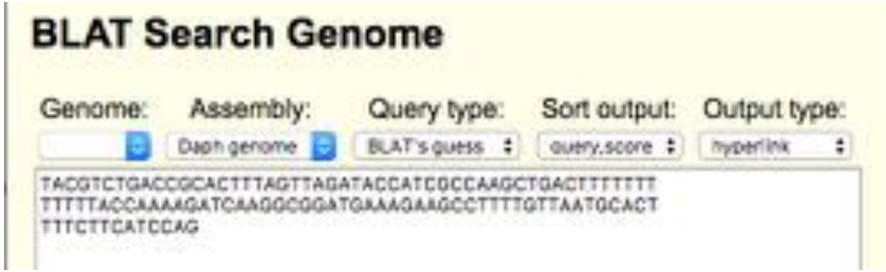
bigInteract



http://genome.ucsc.edu/s/brianlee/PAG_examples

Additional Tools Available on Assembly Hubs

- BLAT search sequence or amino-acid matched



BLAT Search Genome

Genome:

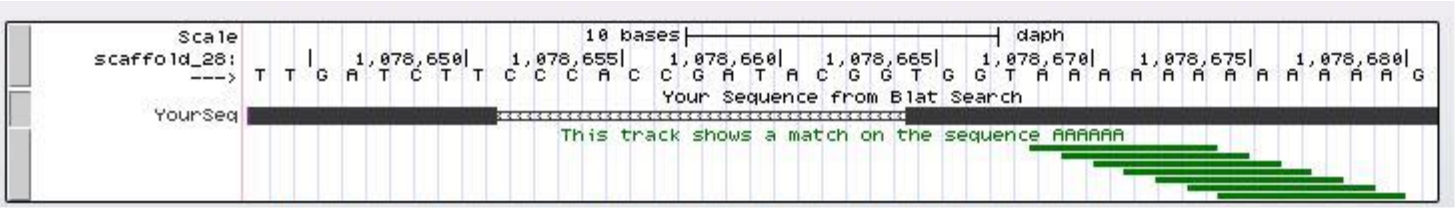
Assembly: Daph genome

Query type: BLAT's guess

Sort output: query,score

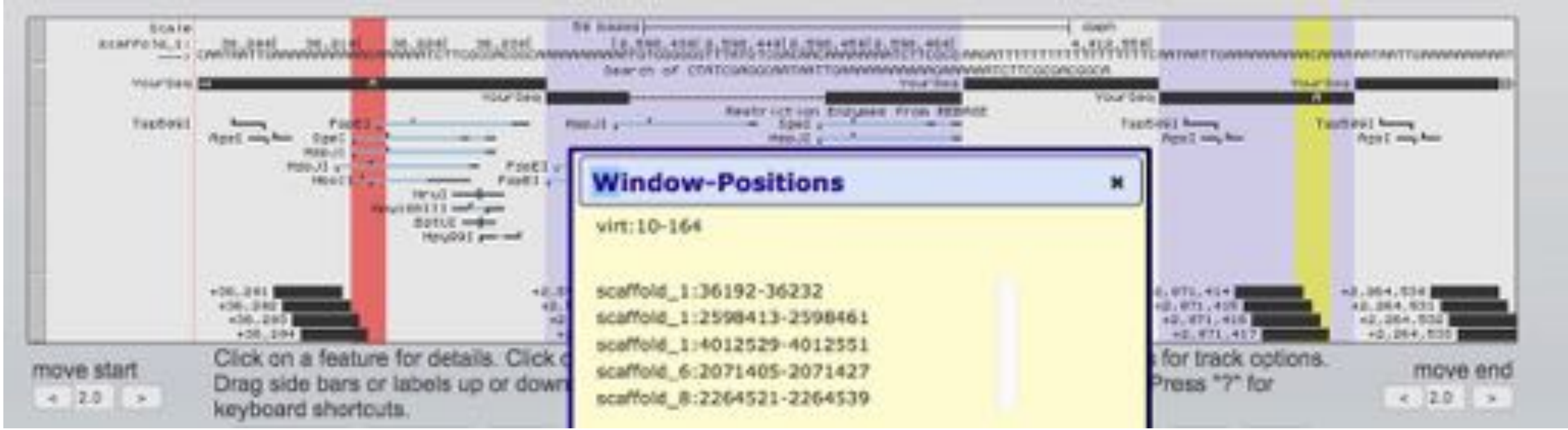
Output type: hyperlink

```
TACGTCTGACCGCACTTTAGTTAGATACCATCGCCAAGCTGACTTTTTTT
TTTTTACCAAAAGATCAAGGCGGATGAAAGAAGCCTTTTGTTAATGCACT
TTTCTTCATCCAG
```



Additional Tools Available on Assembly Hubs

- BLAT search sequence or amino-acid match
- Multi-region: simultaneously view non-adjacent regions across chromosomes



virt:10-164

scaffold_1:36192-36232
scaffold_1:2598413-2598461
scaffold_1:4012529-4012551
scaffold_6:2071405-2071427
scaffold_8:2264521-2264539

Click on a feature for details. Click c
Drag side bars or labels up or down
keyboard shortcuts.

s for track options.
Press "?" for

Summary of Assembly Track Hubs

Remotely Hosted Data

Binary indexed files:

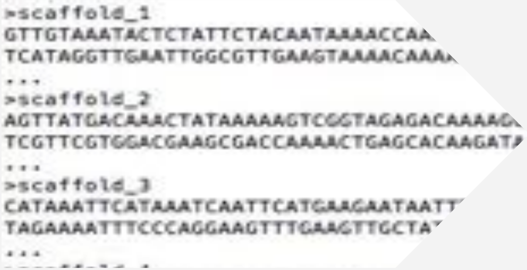
twoBitPath,

bigDataUrl

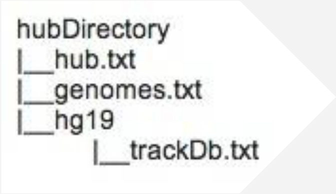
Text files: **hub.txt,**

genomes.txt,

trackDb.txt

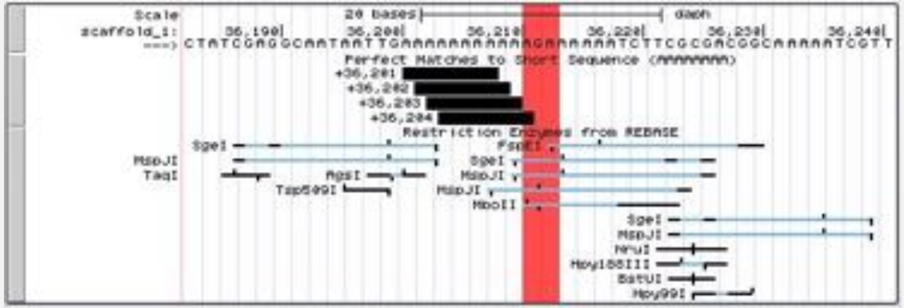


```
>scaffold_1
GTTGTAAATACTCTATTCTACAATAAAACCAA
TCATAGGTTGAATTGGCGTTGAAGTAAAACAAAA
...
>scaffold_2
AGTTATGACAAACTATAAAAAGTCGGTAGAGACAAAAG
TCGTTCGTGGACGAAGCGACCAAAACTGAGCACAAGATA
...
>scaffold_3
CATAAATTCATAAATCAATTCATGAAGAATAATT
TAGAAAATTTCCCAGGAAGTTTGAAGTTGCTA
...
```



```
hubDirectory
|__hub.txt
|__genomes.txt
|__hg19
        |__trackDb.txt
```

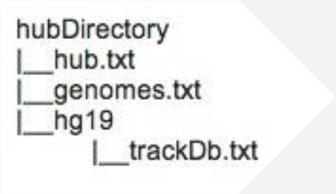
Visualizable at UCSC



Examples of Hubs and Resources

Genbank Assembly Hubs

http://genome-test.gi.ucsc.edu/gbdb/hubs/genbank/



```
hubDirectory
|__hub.txt
|__genomes.txt
|__hg19
        |__trackDb.txt
```

Ensembl Track Hub Registry

https://trackhubregistry.org/

UCSC Public Hubs (Links to Hub Documentation)

http://genome.ucsc.edu/cgi-bin/hgHubConnect

G-OnRamp

http://gonramp.wustl.edu/

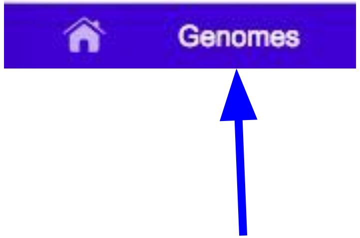
CyVerse Discovery Environment byte-range supported data hosting

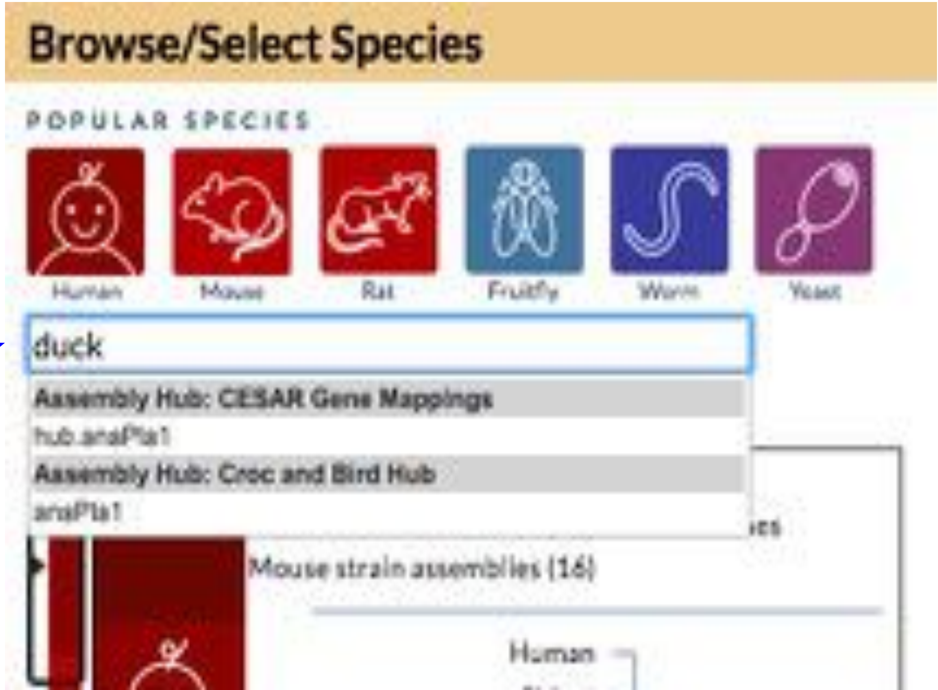
https://de.cyverse.org/de/

Public Track Hubs

UCSC Public Hubs *moderated collection of track hubs*

http://genome.ucsc.edu/cgi-bin/hgGateway





duck

Search term

Public Track Hubs

UCSC Public Hubs *moderated collection of track hubs*

http://genome.ucsc.edu/cgi-bin/hgHubConnect

Search term

Enter search terms to find in public track hub description pages:

duck

Filter hubs by assembly:

Displayed list **restricted by search terms:** duck

When exploring the detailed search results for a hub, you may right-click on a

Clicking Connect redirects to the gateway page of the selected hub's default a

Display	Hub Name	Description
Disconnect	Croc and Bird Hub	Croc, Bird, and Archosaur Assembly Hub

Search details ...
- 1 Matching Assembly
  - Apr. 2013 (BGI_duck_1.0/anaPla1) (anaPla1)
  - *Assembly Description:* ... This is the mallard duck Anas platyrynchos genom

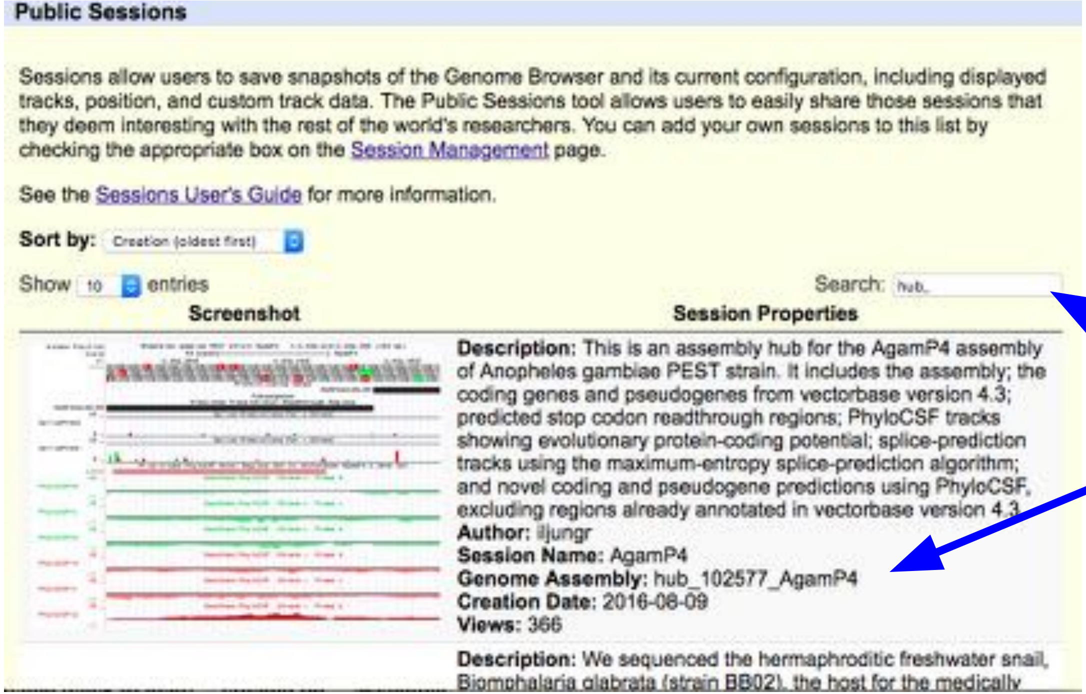
Display	Hub Name	Description
Disconnect	CESAR Gene Mappings	Human Exons mapped by CESAR

Search details ...
- 1 Matching Assembly
  - Apr 2013 (BGI_duck_1.0/anaPla1) (hub

Assembly Hubs via Sessions

Public Sessions *community contributed views of the Browser*

http://genome.ucsc.edu/cgi-bin/hgPublicSessions



Public Sessions

Sessions allow users to save snapshots of the Genome Browser and its current configuration, including displayed tracks, position, and custom track data. The Public Sessions tool allows users to easily share those sessions that they deem interesting with the rest of the world's researchers. You can add your own sessions to this list by checking the appropriate box on the Session Management page.

See the Sessions User's Guide for more information.

Sort by: Creation (oldest first)

Show 10 entries

Search: hub_

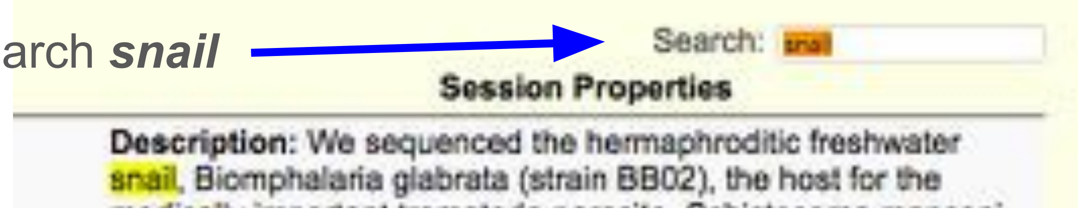
Screenshot	Session Properties
	Description: This is an assembly hub for the AgamP4 assembly of Anopheles gambiae PEST strain. It includes the assembly; the coding genes and pseudogenes from vectorbase version 4.3; predicted stop codon readthrough regions; PhyloCSF tracks showing evolutionary protein-coding potential; splice-prediction tracks using the maximum-entropy splice-prediction algorithm; and novel coding and pseudogene predictions using PhyloCSF, excluding regions already annotated in vectorbase version 4.3. **Author:** iljungr **Session Name:** AgamP4 **Genome Assembly:** hub_102577_AgamP4 **Creation Date:** 2016-08-09 **Views:** 366
	Description: We sequenced the hermaphroditic freshwater snail, Biomphalaria glabrata (strain BB02), the host for the medically

*hub_*

Assembly Hubs via Sessions

Public Sessions *community contributed views of the Browser*

1) Search ***snail***



Search: snail

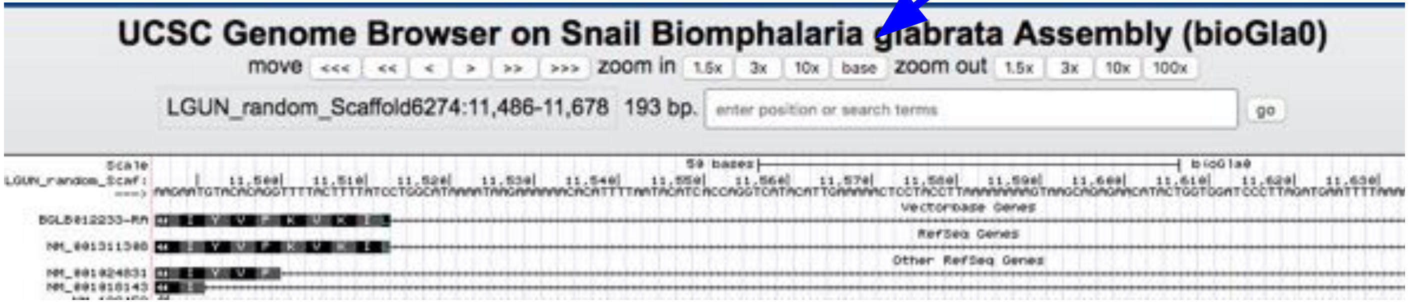
Session Properties

Description: We sequenced the hermaphroditic freshwater snail, Biomphalaria glabrata (strain BB02), the host for the

2) Click image under screenshot to load the session

3) Click the “***base***” zoom-in button

base

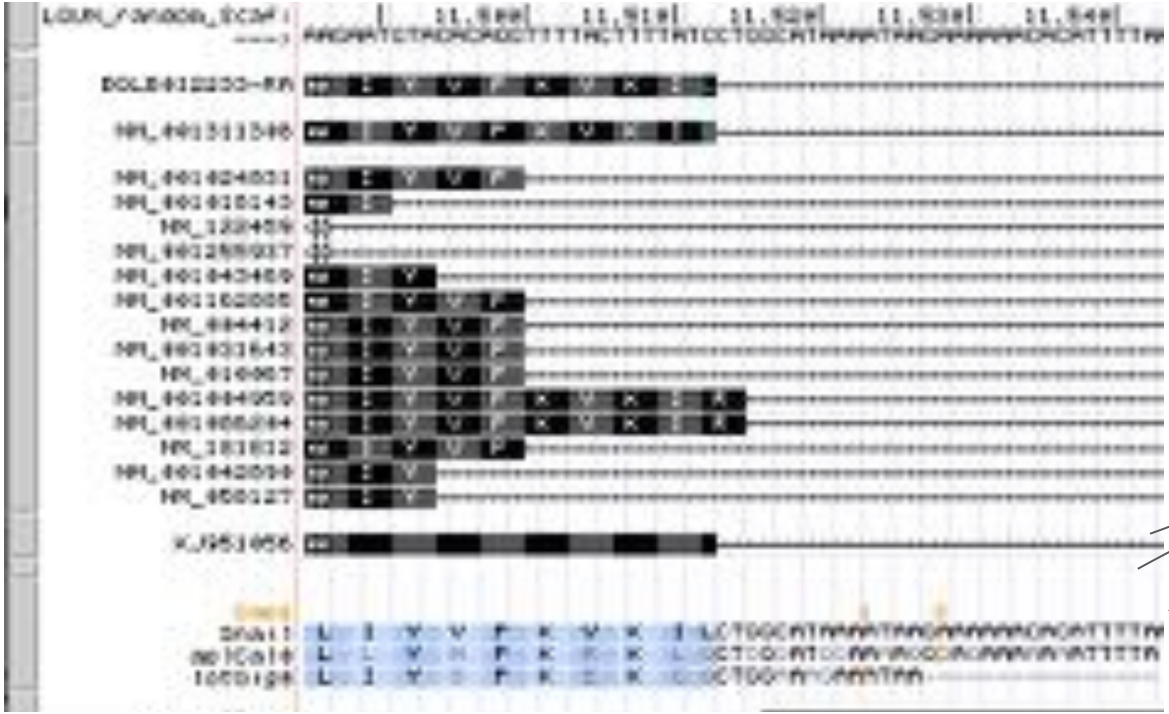


UCSC Genome Browser on Snail Biomphalaria glabrata Assembly (bioGla0)

LGUN_random_Scaffold6274:11,486-11,678 193 bp.

Analyzing a variant on an assembly hub

Variant Annotation Integrator (VAI) *process SNPs on provided prediction tracks*



Collection of many Gene PredictionTracks (**file type** **bigGenePred**). Files allows display of AA codons based on reference assembly sequence.

A genomic alignment track **(file type bigPSL).** mRNAs from GenBank aligned to reference.

A Multiple Alignment Track (**file type** **bigMaf**). Allows display of other genomes aligned to this reference.

Analyzing a variant on an assembly hub

Variant Annotation Integrator (VAI) *process SNPs on provided prediction tracks*

```
track vecBase
shortLabel Vectorbase Genes
bigDataUrl vecBaseGenes.bb
longLabel  Vectorbase Genes
group genes
visibility pack
type bigGenePred
baseColorUseCds given
baseColorDefault genomicCodons
priority 1
```

trackDb.txt for *Vectobase* Gene Prediction Track
type bigGenePred

```
track multiz3way
shortLabel Mollusk Align
bigDataUrl bioGlaMaf.bb
summary multizSummary.bb
frames multiz3wayFrames.bb
longLabel  Mollusk Align
group compGeno
visibility pack
type bigMaf
speciesOrder aplCal0 lotGig0
```

trackDb.txt for *Mollusk Align* Multiple Alignment Track that uses multiple bigBeds to display information
type bigMaf

Analyzing a variant on an assembly hub

Variant Annotation Integrator (VAI) *process SNPs on provided prediction tracks*

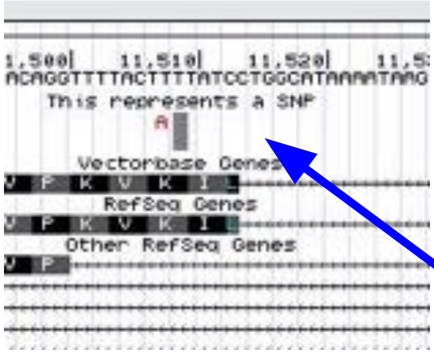
Add a Custom Track

```
track type=pgSnp  name=myData visibility=pack description="This represents a SNP"
LGUN_random_Scaffold6274 11508 11509 A 1 0 0
```

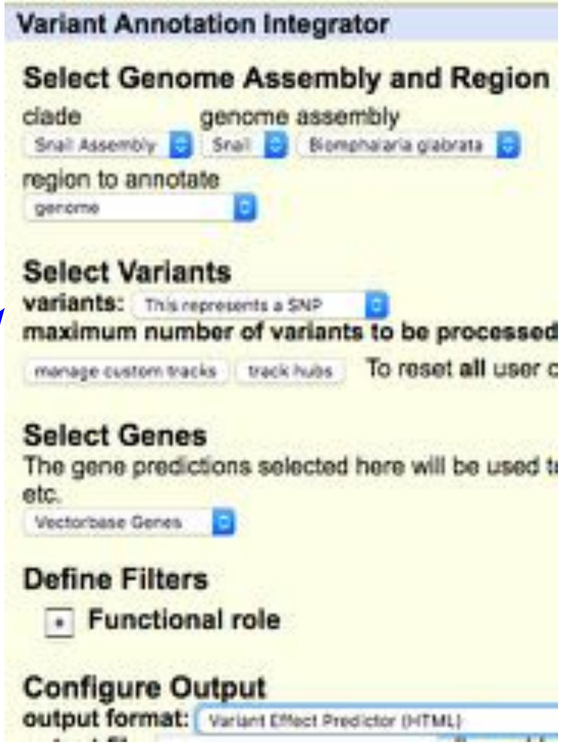
```
track type=pgSnp  name=myData visibility=pack description="This represents a SNP"
LGUN_random_Scaffold6274 11508 11509 A 1 0 0
```

Analyzing a variant on an assembly hub

Variant Annotation Integrator (VAI) *process SNPs on provided prediction tracks*



Custom Track



Variant Annotation Integrator

Select Genome Assembly and Region

clade: Snail Assembly
genome assembly: Snail
Biomphalaria glabrata
region to annotate: genome

Select Variants

variants: This represents a SNP
maximum number of variants to be processed

manage custom tracks | track hubs | To reset all user c

Select Genes

The gene predictions selected here will be used t
etc.

Vectorbase Genes

Define Filters

☐ **Functional role**

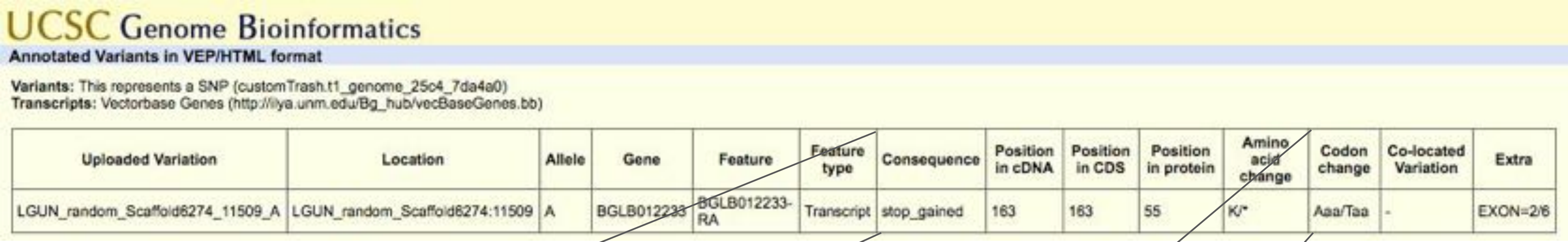
Configure Output

output format: Variant Effect Predictor (HTML)

http://genome.ucsc.edu/s/brianlee/PAG_snail

Analyzing a variant on an assembly hub

Variant Annotation Integrator (VAI) *process SNPs on provided prediction tracks*

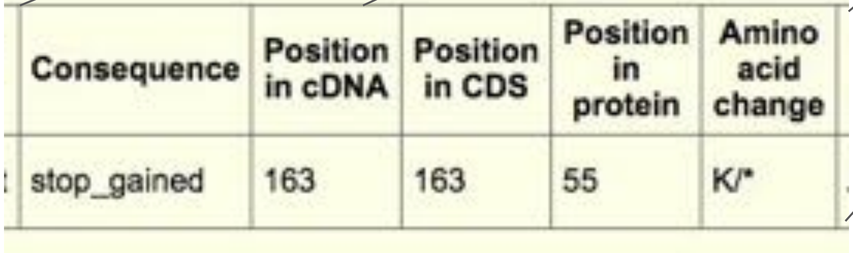


UCSC Genome Bioinformatics

Annotated Variants in VEP/HTML format

Variants: This represents a SNP (customTrash.t1_genome_25c4_7da4a0)
Transcripts: Vectorbase Genes (http://ilya.unm.edu/Bg_hub/vecBaseGenes.bb)

Uploaded Variation	Location	Allele	Gene	Feature	Feature type	Consequence	Position in cDNA	Position in CDS	Position in protein	Amino acid change	Codon change	Co-located Variation	Extra
LGUN_random_Scaffold6274_11509_A	LGUN_random_Scaffold6274:11509	A	BGLB012233	BGLB012233-RA	Transcript	stop_gained	163	163	55	K/*	Aaa/Taa	-	EXON=2/6



Consequence	Position in cDNA	Position in CDS	Position in protein	Amino acid change
stop_gained	163	163	55	K/*

Thank You!