UCSC Genome Browser Track Data Hubs

Ann S. Zweig, Brian J. Raney, Galt Barber, Angie S. Hinrichs, Donna Karolchik,
David Haussler and W. James Kent
Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California, 95064, USA

Track Data Hubs

The University of California, Santa Cruz (UCSC) Genome Browser[1,2] (*http://genome.ucsc.edu*) is a public, freely available web-based graphical viewer for the display of genome sequences and their annotations with links to external public databases. Track Data Hubs are web-accessible directories of genomic data that can be viewed on the UCSC Genome Browser alongside the native annotation tracks.

Track Data Hubs are useful for projects that generate large amounts of genome-wide data sets. For smaller data sets the Custom Track mechanism of displaying data in the genome browser is often easier. However when a project has more than a half dozen wiggle plots or other tracks to display, the Data Hub allows the tracks to be organized into composite (grouped) tracks. This makes it possible to show data for a large collection of tissues and experimental conditions in an elegant way, similar to how the ENCODE data is displayed at UCSC.

Because the data files remain on the remote server, we have eliminated the need to transfer large data sets across the Internet. Labs and individual users format their data sets using one of the browser binary data types (bigBed, bigWig[3] or Binary Alignment Map[4] (BAM)), make the files available on an Internet-connected computer (a.k.a. Track Data Hub), then register the data collection with UCSC. The Genome Browser's Data Hub Portal lists registered data collections from around the world of interest to all types of scientists. Genome Browser users select data sets to display on the Portal page, then access the data hub tracks in the same way that they currently access the native annotation tracks: through the track controls. If for some reason the data tracks from a remote site are not available, the Genome Browser provides an informative error message in the space where the track would normally appear.

This distributed data model allows Genome Browser users to view data sets from scientists worldwide using the familiar Genome Browser interface. Support for Track Data Hubs is in development, and a prototype version of a Data Hub for the Epigenomics Roadmap Project is available on *http://genome-preview.ucsc.edu*.

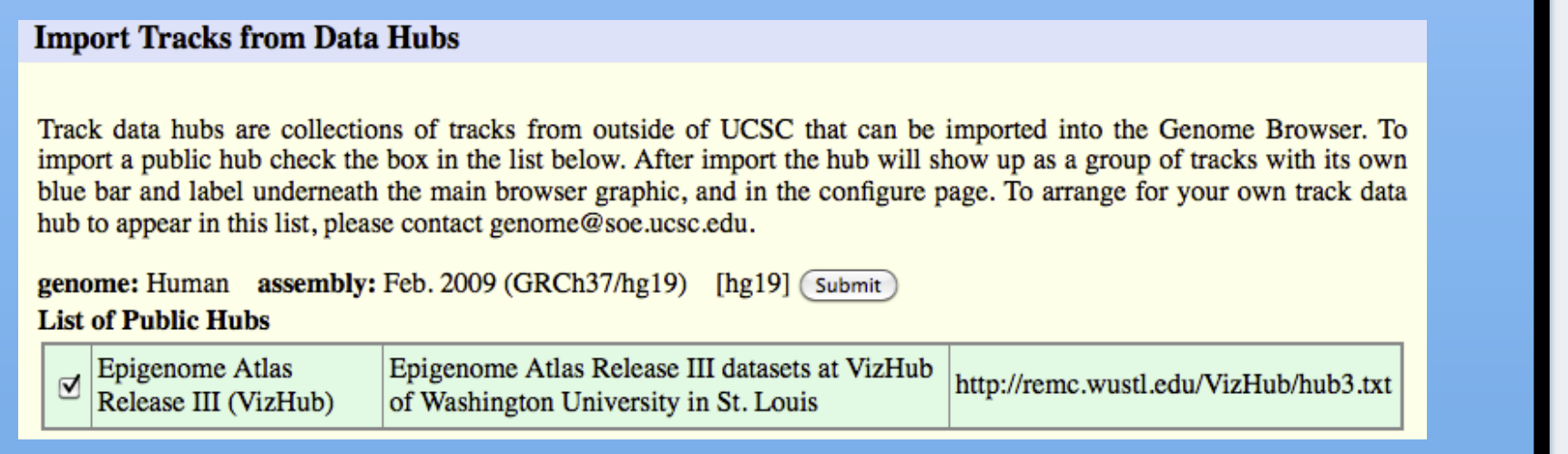
Track Data Hubs vs. Custom Tracks

Custom Tracks work well for those who have only a few small data sets and do not require persistent data. Custom Tracks are uploaded to the UCSC Genome Browser and viewed alongside the native annotation tracks. They are saved for a short amount of time on the Genome Browser website, and can be shared with colleagues using the Session tool. Custom Tracks can be built from several data types such as bigBed, bigWig, BAM, BED, wiggle, bedGraph, GTF, PSL, MAF, BEDdetail, etc.

However, if users have large data sets that would be prohibitive to upload, need to ensure the persistence of their data, or would like to take advantage of additional track functionality, Data Hubs are a better solution. Annotation tracks in Data Hubs can be grouped into composite or super-tracks and configured to display the data using the most appropriate methods. Although they are restricted to only the remote binary data types (bigBed, bigWig, and BAM), the extensive configuration options overshadow this restriction.

Accessing Data from a Track Data Hub

Users of the UCSC browser can access tracks from remote Track Data Hubs. The Data Hub Portal page lists all of the publicly available Data Hubs from which to choose:



Import Tracks from Data Hubs

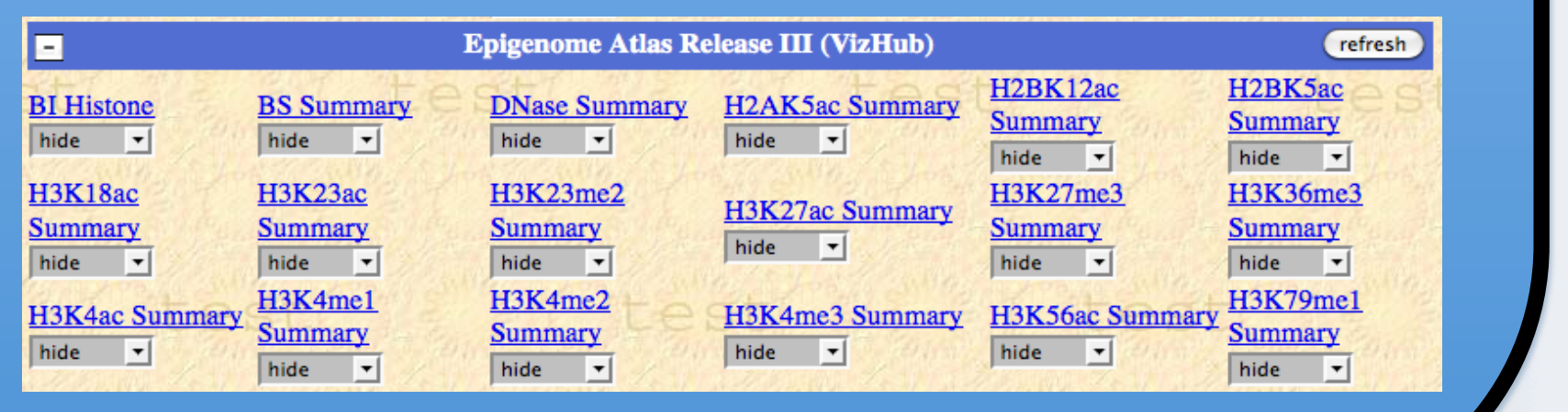
Track data hubs are collections of tracks from outside of UCSC that can be imported into the Genome Browser. To import a public hub check the box in the list below. After import the hub will show up as a group of tracks with its own blue bar and label underneath the main browser graphic, and in the configure page. To arrange for your own track data hub to appear in this list, please contact genome@soe.ucsc.edu.

genome: Human **assembly:** Feb. 2009 (GRCh37/hg19) [hg19] Submit

List of Public Hubs

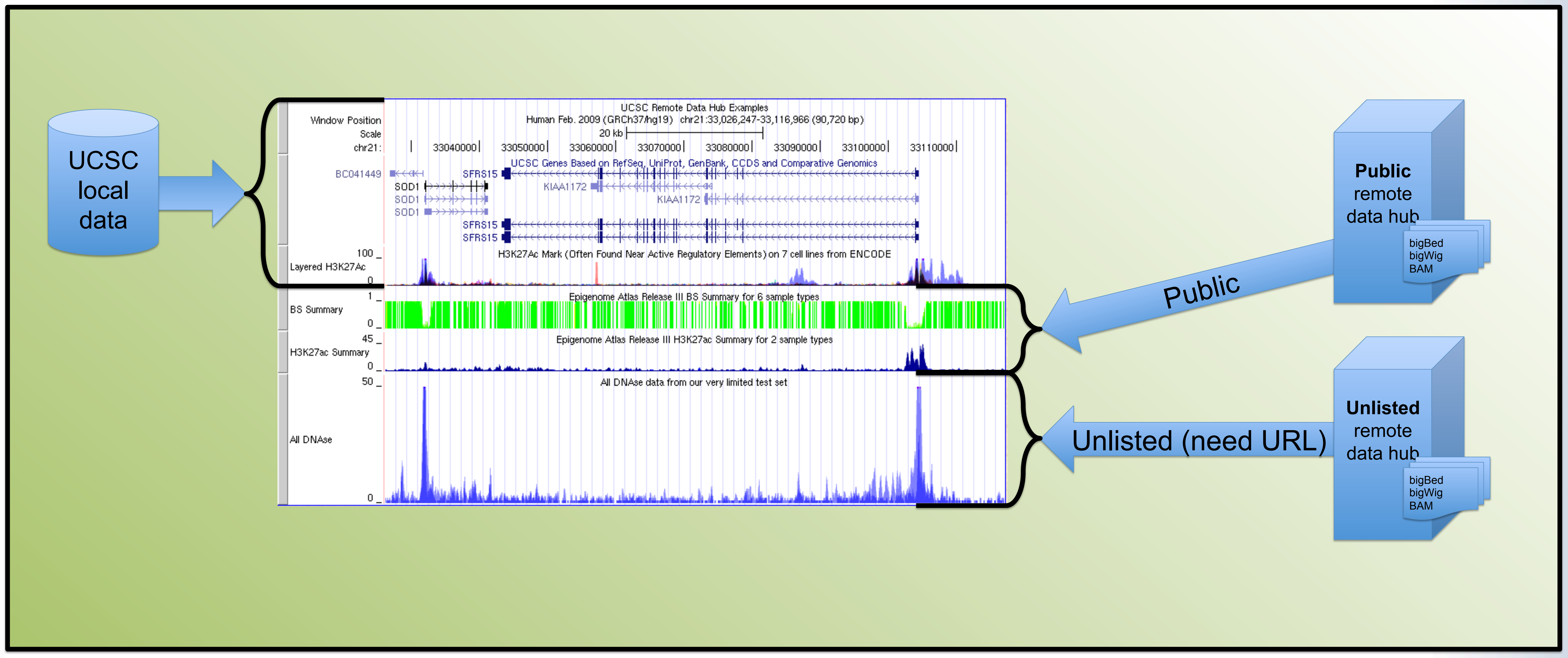
☑	Epigenome Atlas Release III (VizHub)	Epigenome Atlas Release III datasets at VizHub of Washington University in St. Louis	http://remc.wustl.edu/VizHub/hub3.txt

After a Hub is imported, it appears in the track list as a new group that contains all of the tracks in the Data Hub. The individual tracks from the Hub are available for display alongside the native Genome Browser tracks.



Epigenome Atlas Release III (VizHub) refresh

BI Histone hide	BS Summary hide	DNase Summary hide	H2AK5ac Summary hide	H2BK12ac Summary hide	H2BK5ac Summary hide
H3K18ac Summary hide	H3K23ac Summary hide	H3K23me2 Summary hide	H3K27ac Summary hide	H3K27me3 Summary hide	H3K36me3 Summary hide
H3K4ac Summary hide	H3K4me1 Summary hide	H3K4me2 Summary hide	H3K4me3 Summary hide	H3K56ac Summary hide	H3K79me1 Summary hide


UCSC local data
Public remote data hub
bigBed
bigWig
BAM
Public
Unlisted remote data hub
bigBed
bigWig
BAM
Unlisted (need URL)
UCSC Remote Data Hub Examples
Human Feb. 2009 (GRCh37/hg19) chr21:33,026,247-33,116,966 (90,720 bp)
Window Position
Scale
chr21:
20 kb
33040000
33050000
33060000
33070000
33080000
33090000
33100000
33110000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
BC041449
SFRS15
SOD1
KIAA1172
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
1
BS Summary
0
Epigenome Atlas Release III BS Summary for 6 sample types
45
H3K27ac Summary
0
Epigenome Atlas Release III H3K27ac Summary for 2 sample types
50
All DNAse
0
All DNAse data from our very limited test set

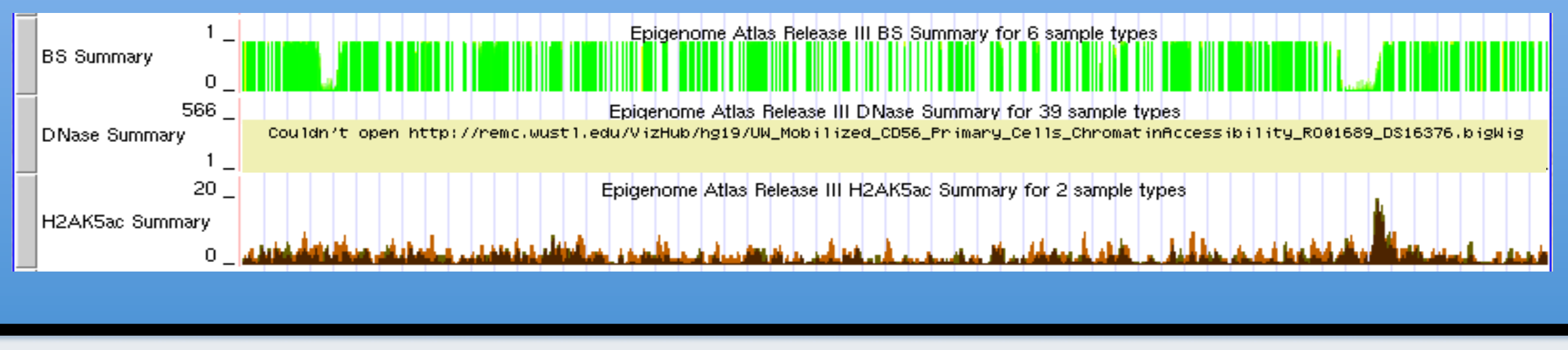
Unlisted Track Data Hubs

For users who would like to see their data alongside the native UCSC Genome Browser annotation tracks without making it available to all users, the solution is an Unlisted Data Hub. An Unlisted Hub is available to those who know the URL, but does not appear on the UCSC Data Hub Portal page.

Please note that Unlisted Data Hubs are not secure in the same way that a bank website (for example) is secure. Rather the URL helps to obfuscate the location of the data; it is a simple barrier to casual users.

Offline Track Data Hubs

Occasionally, for unavoidable reasons, remote Data Hubs may be missing, off-line, or otherwise unavailable. If a Data Hub goes off-line, it is automatically removed from the Portal page, and users browsing data from the remote Hub will encounter an error message with a yellow background instead of the expected data.


BS Summary
1
0
Epigenome Atlas Release III BS Summary for 6 sample types
DNase Summary
566
1
Epigenome Atlas Release III DNase Summary for 39 sample types
H2AK5ac Summary
20
0
Epigenome Atlas Release III H2AK5ac Summary for 2 sample types

```
Couldn't open http://remc.wustl.edu/VizHub/hg19/UW_Mobilized_CD56_Primary_Cells_ChromatinAccessibility_RO01689_DS16376.bigWig
```

Creating a Public Track Data Hub

To share data with hundreds of thousands of Genome Browser users, the data files must be in one of the following formats: bigBed, bigWig, BAM (and soon, VCF[5] indexed by tabix[6]). All data files must be placed on a a web-accessible server (http or ftp). A few supporting files should also be in place, including (for example):

```
myHub/ - directory for the hub as a whole
  hub.txt - text file containing a short description of the hub
  genomes.txt - text file containing a list of genome assemblies
  hg19/ - directory for a recent human assembly
    trackDb.txt - text file containing track display details
                  including names, colors, data types, etc.
    dnase.html - HTML file describing a DNase track to users
    dnaseSignal.bw - wiggle plot of DNAse Signal
    dnaseReads.bam - BAM file of DNAse Reads
```

Data Hubs can be added to the Data Hub Portal page by contacting the Genome Browser mailing list at *genome@soe.ucsc.edu*.

trackDb.txt (example):

```
track dnaseSignal
bigDataUrl dnaseSignal.bw
shortLabel DNAse Signal
longLabel Depth of alignments of DNAse reads
type bigWig

track dnaseReads
bigDataUrl dnaseReads.bam
shortLabel DNAse Reads
longLabel DNAse reads mapped with MAQ
type bam
```

References & Credits

1. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res*. 2002 Jun;12(6):996-1006.
2. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*. 2011 Jan;39(Database issue):D876-82. Epub 2010 Oct 18.
3. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed data sets. *Bioinformatics*. 2010 Sep 1;26(17):2204-7. Epub 2010 Jul 17.
4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map (SAM) Format and SAMtools. *Bioinformatics.* 2009 Aug 15;25(16):2078-9.
5. Variant Call Format (VCF) working specification: *http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40*
6. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics.* 2011 Mar 1;27(5):718-9.

We would like to acknowledge the work of the UCSC Genome Bioinformatics technical staff (*http://genome.ucsc.edu/staff.html*), our many collaborators, and our users for their feedback and support. The project has been funded by a grant from the National Human Genome Research Institute (grant number 5P41 HG002372-04).

The Biology of Genomes, Cold Spring Harbor Laboratory; Ann Zweig; 10 -14 May 2011.
Digital copy of this poster available at: *http://users.soe.ucsc.edu/~ann/BoG2011.pdf*