The Data Integrator: a new way to combine data sources underlying the UCSC Genome Browser

Angie S. Hinrichs[1], Kate R. Rosenbloom[1], Matthew L. Speir[1], Donna Karolchik[1], Ann S. Zweig[1], David Haussler[1,2], Robert M. Kuhn[1], W. James Kent[1]
[1]University of California Santa Cruz Genomics Institute, Santa Cruz, CA; [2]Howard Hughes Medical Institute, UCSC, Santa Cruz, CA

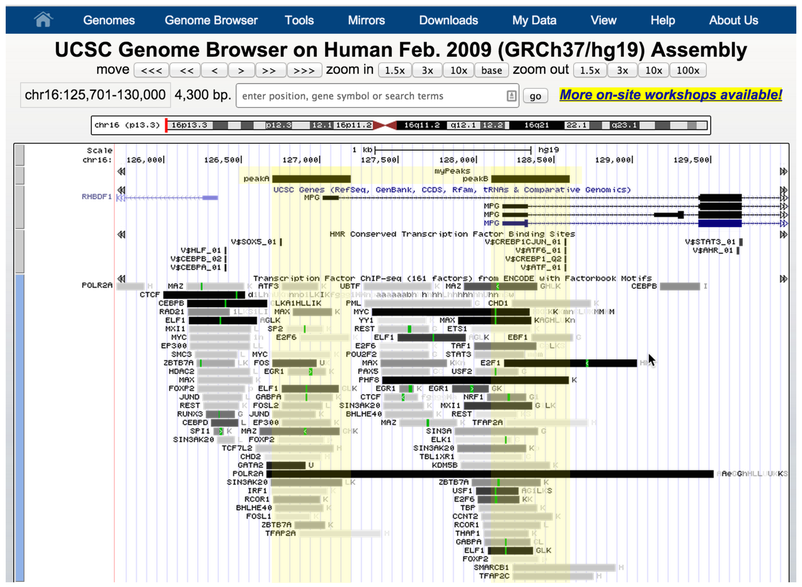
Introduction

Users of the UCSC Genome Browser have asked us variations of the same question many times over the years:

How can I get the { … score / gene name / TFBS } for each of my { peaks / regions / probes } ?

While the Table Browser tool can compute the intersections of two tracks by genomic position, it does so by reducing the second track to a set of genomic regions. It does not include any annotation columns of the second track in the output.

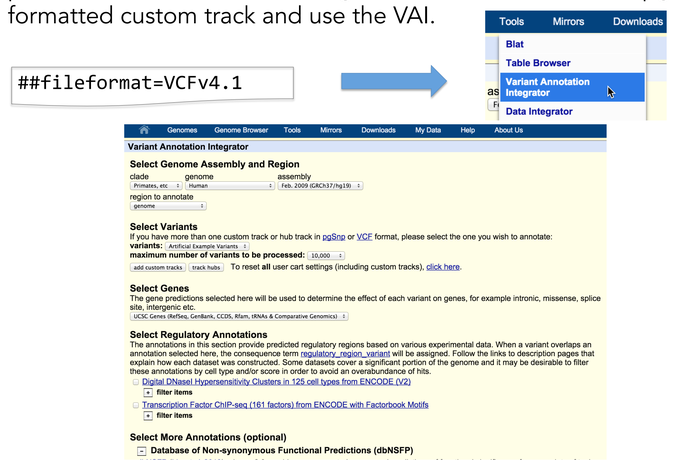
We have developed a new tool, the Data Integrator, that can combine data from up to five tracks based on overlap with items in the first track, and output all columns or a selected subset of columns. The output is tab-separated text that can be viewed in the web browser window or downloaded to a local file, optionally compressed by gzip. Like the Genome Browser and Table Browser, the Data Integrator can combine data from the browser database, user custom tracks and hub tracks.



Example: get a text file with the overlapping items shown here

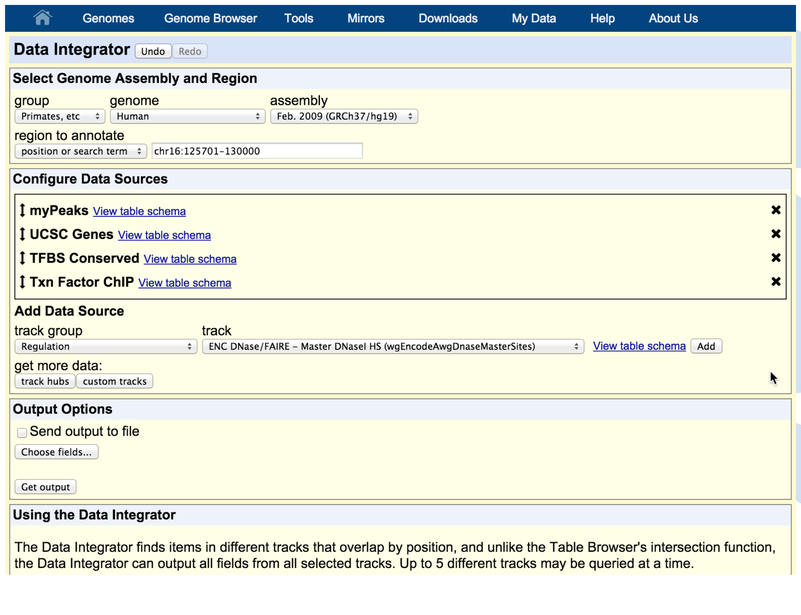
Got variants?

The Data Integrator is a more flexible and open-ended tool complementary to the **Variant Annotation Integrator** (VAI).
If you are starting with a list of variant calls, and would like to get predicted functional effects on genes, then create a VCF or pgSnp-formatted custom track and use the VAI.



Data Integrator

```
track name=myPeaks
chr16   126700   127200  peakA
chr16   128100   128600  peakB
...
```




```
#ct_myPeaks_9149.chrom  ct_myPeaks_9149.chromStart      ct_myPeaks_9149.chromEnd        ct_myPeaks_9149.name    knownGene.name
knownGene.geneSymbol    tfbsConsSites.name      tfbsConsSites.strand    tfbsConsSites.zScore    wgEncodeRegTfbsClusteredV3.name
wgEncodeRegTfbsClusteredV3.score     wgEncodeRegTfbsClusteredV3.expCount     wgEncodeRegTfbsClusteredV3.cellType
chr16   126700  127200  peakA   uc002cfm.4      MPG     V$SOX9_B1       -       2.05    TCF7L2  230     1       HeLa-S3,
chr16   126700  127200  peakA                   V$SOX5_01       -       2.36    MAZ     604     3       GM12878,HeLa-S3,K562,
chr16   126700  127200  peakA                   V$AREB6_04      -       1.89    CHD2    128     1       HeLa-S3,
chr16   126700  127200  peakA                   V$SPZ1_01       -       2.17    GATA2   851     1       HUVEC,
chr16   126700  127200  peakA                   V$OLF1_01       -       2.04    POLR2A  1000    51
A549,A549,ECC-1,GM12878,GM12878,HCT-116,HeLa-S3,HepG2,HepG2,HUVEC,HUVEC,K562,K562,SK-N-SH,GM10847,GM12878,GM12891,GM12892,GM15510,
GM18505,GM18526,GM18951,GM19099,GM19193,HCT-116,HEK293,HeLa-S3,HepG2,HUVEC,IMR90,K562,K562,K562,K562,K562,MCF10A-Er-Src,MCF10A-Er-Src,
NB4,PBDE,Raji,A549,Gliobla,GM12878,HeLa-S3,HepG2,HUVEC,K562,MCF-7,MCF-7,MCF-7,ProgFib,
...
```

How to build a query

(optional)
Create a custom track or hub track

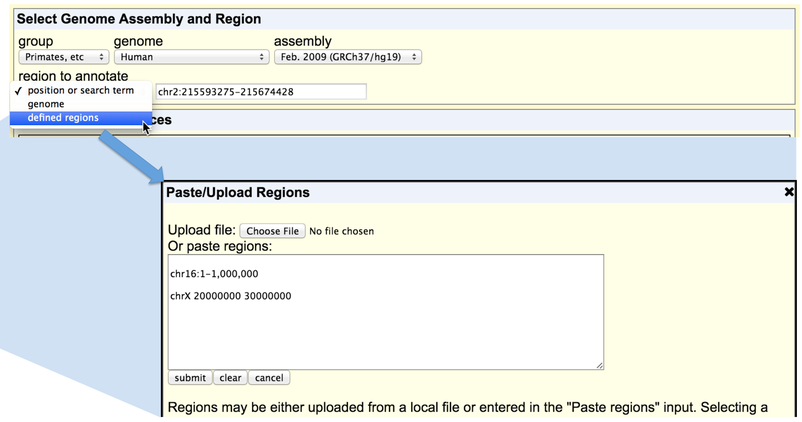
Choose genome assembly and regions of interest

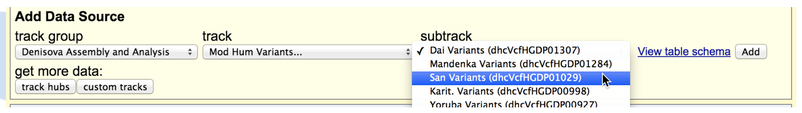
Configure data sources

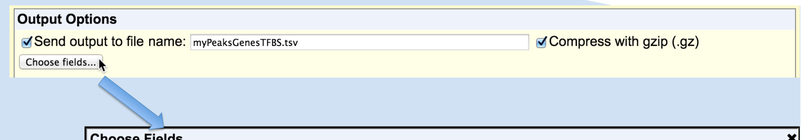
Configure output

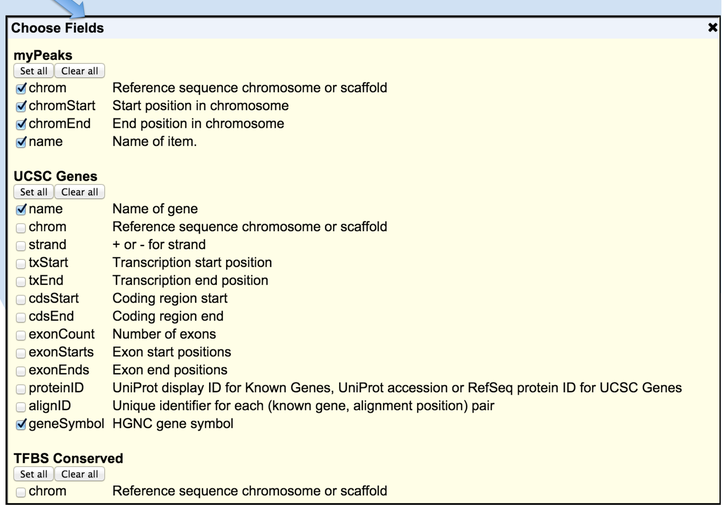
Get data!

Options









Future work

Some of the features we intend to add:

- Filter inputs by value
- Configure overlap rules
- Reorder output columns
- Add columns from related database tables

Reference:
The UCSC Genome Browser database: 2015 update. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit AF, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. *Nucleic Acids Res.* 2015 Jan;43(Database issue):D670-81.

Acknowledgements:
This work was funded by National Human Genome Research Institute (5 U41 HG002371-15 to UCSC Center for Genomic Science). 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.

More information

- Search for answers in our mail list archives: http://genome.ucsc.edu/contacts.html
- Email a new question to our actively monitored list: genome@soe.ucsc.edu
- UCSC training (workshops, videos, tutorials): http://genome.ucsc.edu/training/
- Blog: http://genome.ucsc.edu/blog/

http://genomewiki.ucsc.edu/index.php/BoG2015DataIntegratorPoster