H3ABioNet

Pan African Bioinformatics Network for H3Africa

Using public sources of data

H3ABioNet Data Management Workshop

Data sources

- Genotypes
  - 1000 Genomes
  - HapMap
  - dbGaP
  - EGA
  - GEO
  - WTCCC
  - PGP

Data sources

- Annotations
  - Chip manufacturers
  - ENSEMBL
  - RefSeq
  - dbNSFP

Tools

- Google
- Galaxy
- PLINK (“PLINK 2”)
- Scripting
  - sed, awk, grep, cut, sort, uniq, parallel
- Databases

Data QC issues

- Genome build
- Annotation builds
- Identifiers
- Strand

Question 1

Hi

I would like to know how many SNPs contained in the Immunochip are in the Illumina 2.5 Million SNPs chip.

The Immunochip is an Illumina Infinium genotyping chip, containing 196,524 polymorphisms (718 small insertion deletions, 195,806 SNPs), designed to perform deep replication of major autoimmune and inflammatory diseases.

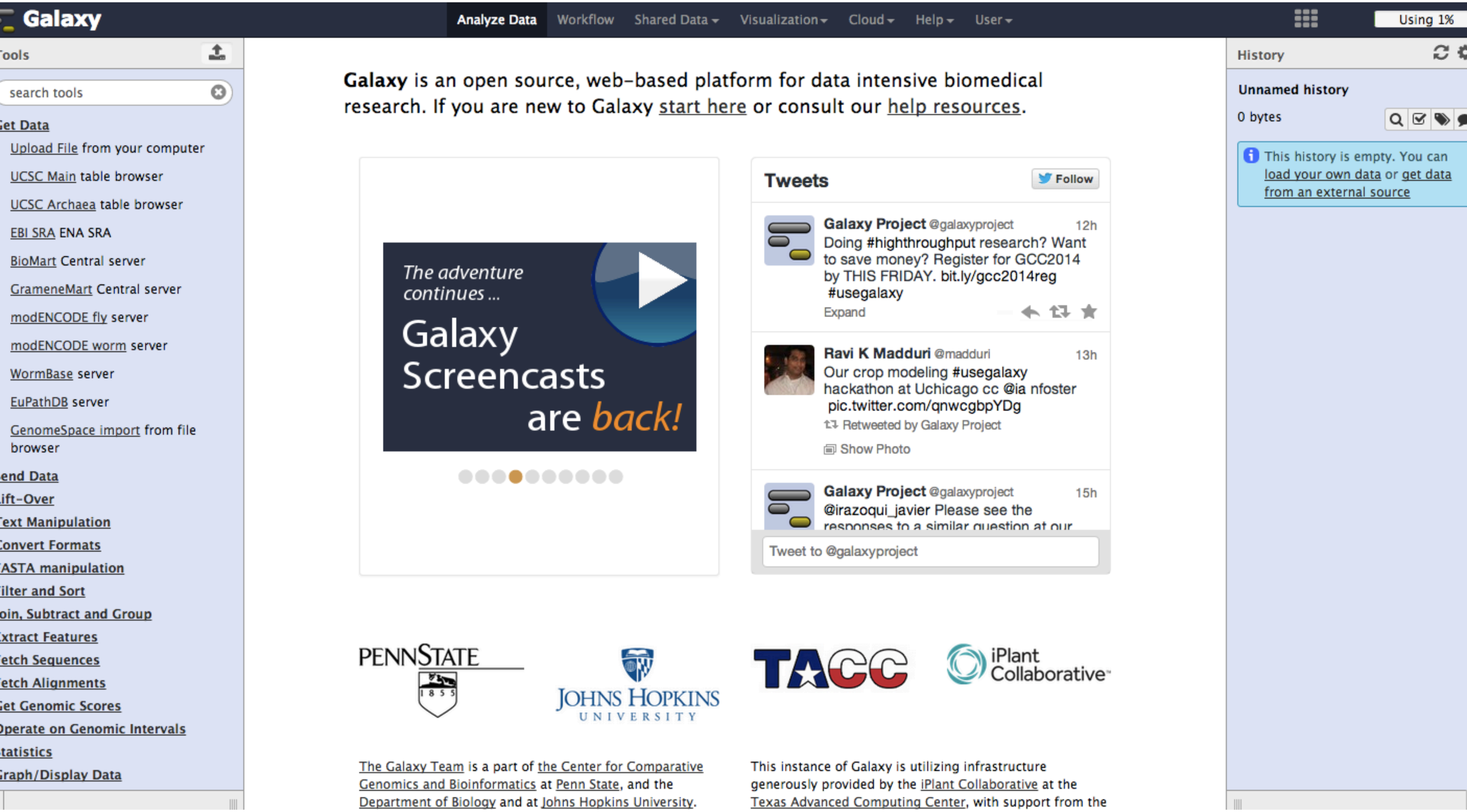
The 2.5M Illumina chip promises to perform well in typing both common and rare SNP content from the 1kGP (MAF>2.5%) for diverse world populations. This array contains tagSNP data from recently released 1000 Genomes Project pilot data.

Usenet post, 1997

Some people, when confronted with a problem, think
“I know, I'll use regular expressions.”

Now they have two problems.

Galaxy



et Data
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
EBI SRA ENA SRA
BioMart Central server
GrameneMart Central server
modENCODE fly server
modENCODE worm server
WormBase server
EuPathDB server
GenomeSpace import from file browser
end Data
ift-Over
ext Manipulation
onvert Formats
ASTA manipulation
ilter and Sort
oin, Subtract and Group
xtract Features
etch Sequences
etch Alignments
et Genomic Scores
perate on Genomic Intervals
tatistics
raph/Display Data

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

The adventure continues ...
Galaxy Screencasts are *back!*

Tweets

Galaxy Project @galaxyproject 12h
Doing #highthroughput research? Want to save money? Register for GCC2014 by THIS FRIDAY. bit.ly/gcc2014reg #usegalaxy
Expand

Ravi K Madduri @madduri 13h
Our crop modeling #usegalaxy hackathon at Uchicago cc @ia nfoster pic.twitter.com/qnwcgbpYDg
Retweeted by Galaxy Project
Show Photo

Galaxy Project @galaxyproject 15h
@irazoqui_javier Please see the responses to a similar question at our

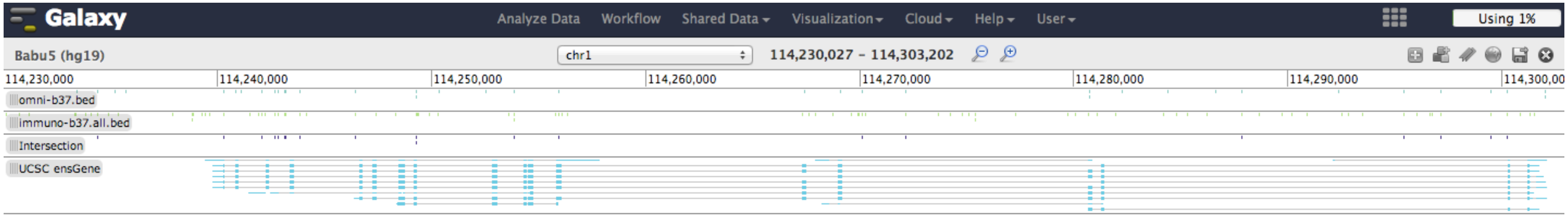
PENNSTATE 1855
JOHNS HOPKINS UNIVERSITY
TACC
iPlant Collaborative

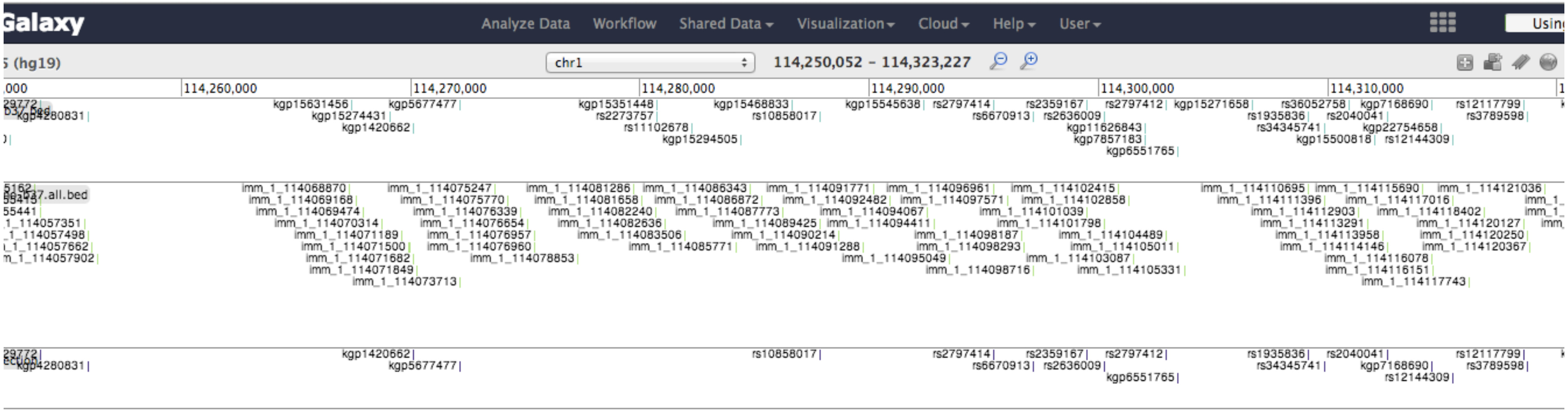
The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Biology and at Johns Hopkins University.

This instance of Galaxy is utilizing infrastructure generously provided by the iPlant Collaborative at the Texas Advanced Computing Center, with support from the

Unnamed history
0 bytes
This history is empty. You can load your own data or get data from an external source

Visualisation in Trackster



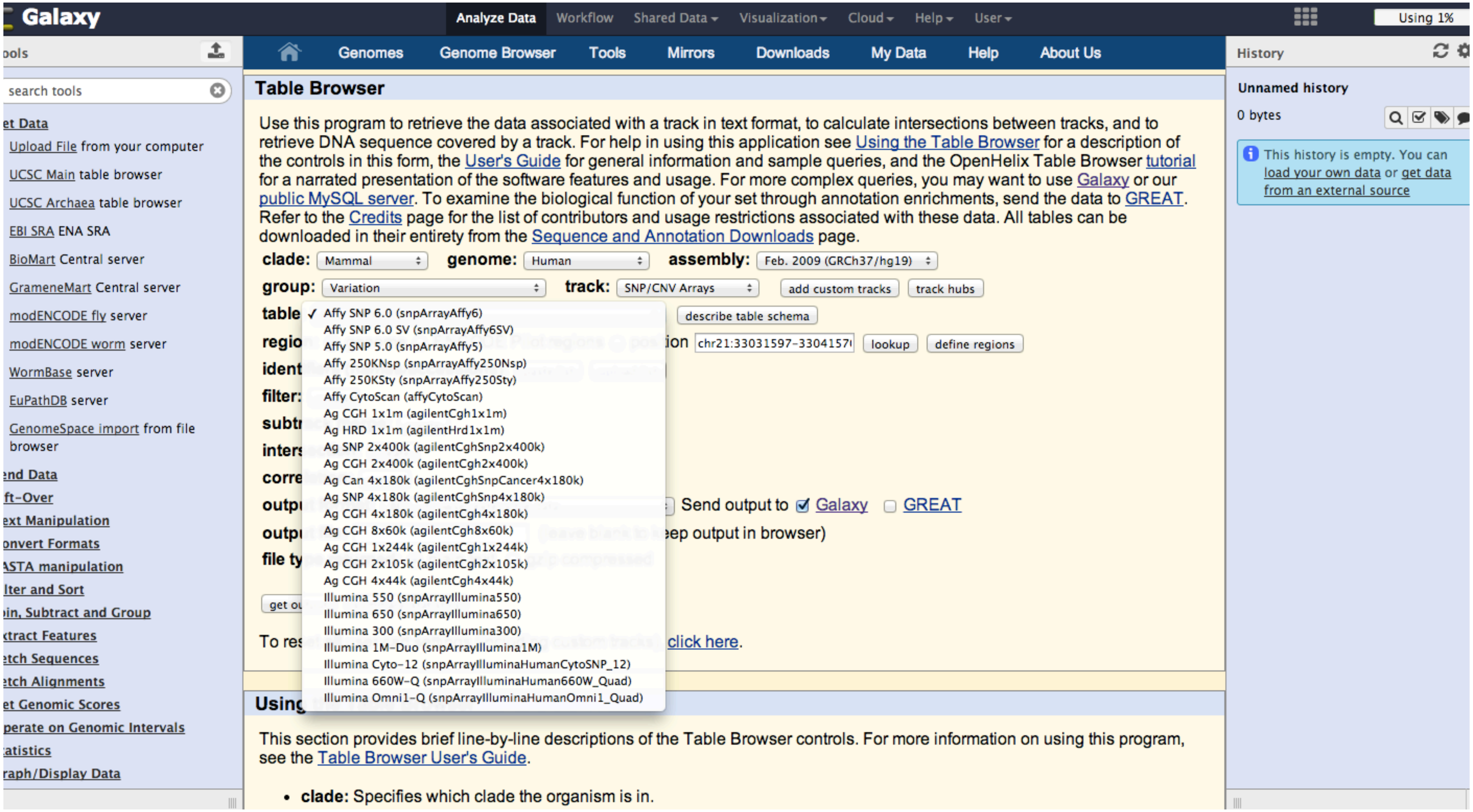


1st attempt

UCSC table browser

Galaxy

UCSC table browser



This history is empty. You can load your own data or get data from an external source

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Variation **track:** SNP/CNV Arrays add custom tracks track hubs

table describe table schema

- ✓ Affy SNP 6.0 (snpArrayAffy6)
- Affy SNP 6.0 SV (snpArrayAffy6SV)
- Affy SNP 5.0 (snpArrayAffy5)
- Affy 250KNsp (snpArrayAffy250Nsp)
- Affy 250KSty (snpArrayAffy250Sty)
- Affy CytoScan (affyCytoScan)
- Ag CGH 1x1m (agilentCgh1x1m)
- Ag HRD 1x1m (agilentHrd1x1m)
- Ag SNP 2x400k (agilentCghSnp2x400k)
- Ag CGH 2x400k (agilentCgh2x400k)
- Ag Can 4x180k (agilentCghSnpCancer4x180k)
- Ag SNP 4x180k (agilentCghSnp4x180k)
- Ag CGH 4x180k (agilentCgh4x180k)
- Ag CGH 8x60k (agilentCgh8x60k)
- Ag CGH 1x244k (agilentCgh1x244k)
- Ag CGH 2x105k (agilentCgh2x105k)
- Ag CGH 4x44k (agilentCgh4x44k)
- Illumina 550 (snpArrayIllumina550)
- Illumina 650 (snpArrayIllumina650)
- Illumina 300 (snpArrayIllumina300)
- Illumina 1M-Duo (snpArrayIllumina1M)
- Illumina Cyto-12 (snpArrayIlluminaHumanCytoSNP_12)
- Illumina 660W-Q (snpArrayIlluminaHuman660W_Quad)
- Illumina Omni1-Q (snpArrayIlluminaHumanOmni1_Quad)

regio ion chr21:33031597-3304157 lookup define regions

ident

filter:

subtr

inter

corre

outp Send output to ☑ Galaxy □ GREAT

outp ep output in browser)

file ty

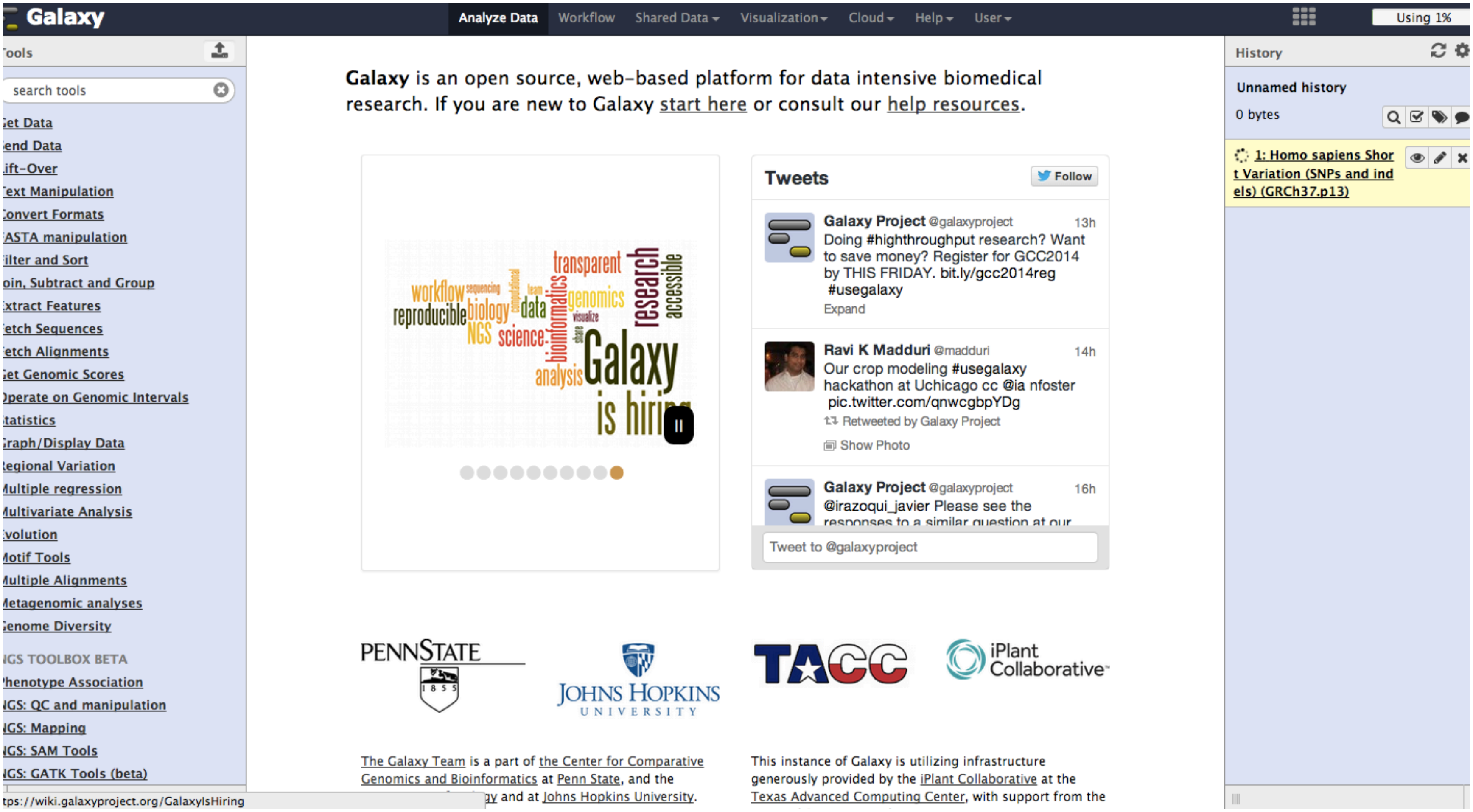
get ou

To res click here.

Using

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the Table Browser User's Guide.

- **clade:** Specifies which clade the organism is in.



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

Galaxy Project @galaxyproject 13h
Doing #highthroughput research? Want to save money? Register for GCC2014 by THIS FRIDAY. bit.ly/gcc2014reg #usegalaxy
Expand

Ravi K Madduri @madduri 14h
Our crop modeling #usegalaxy hackathon at Uchicago cc @ia nfoster pic.twitter.com/qnwcgbpYDg
Retweeted by Galaxy Project
Show Photo

Galaxy Project @galaxyproject 16h
@irazoqui_javier Please see the responses to a similar question at our

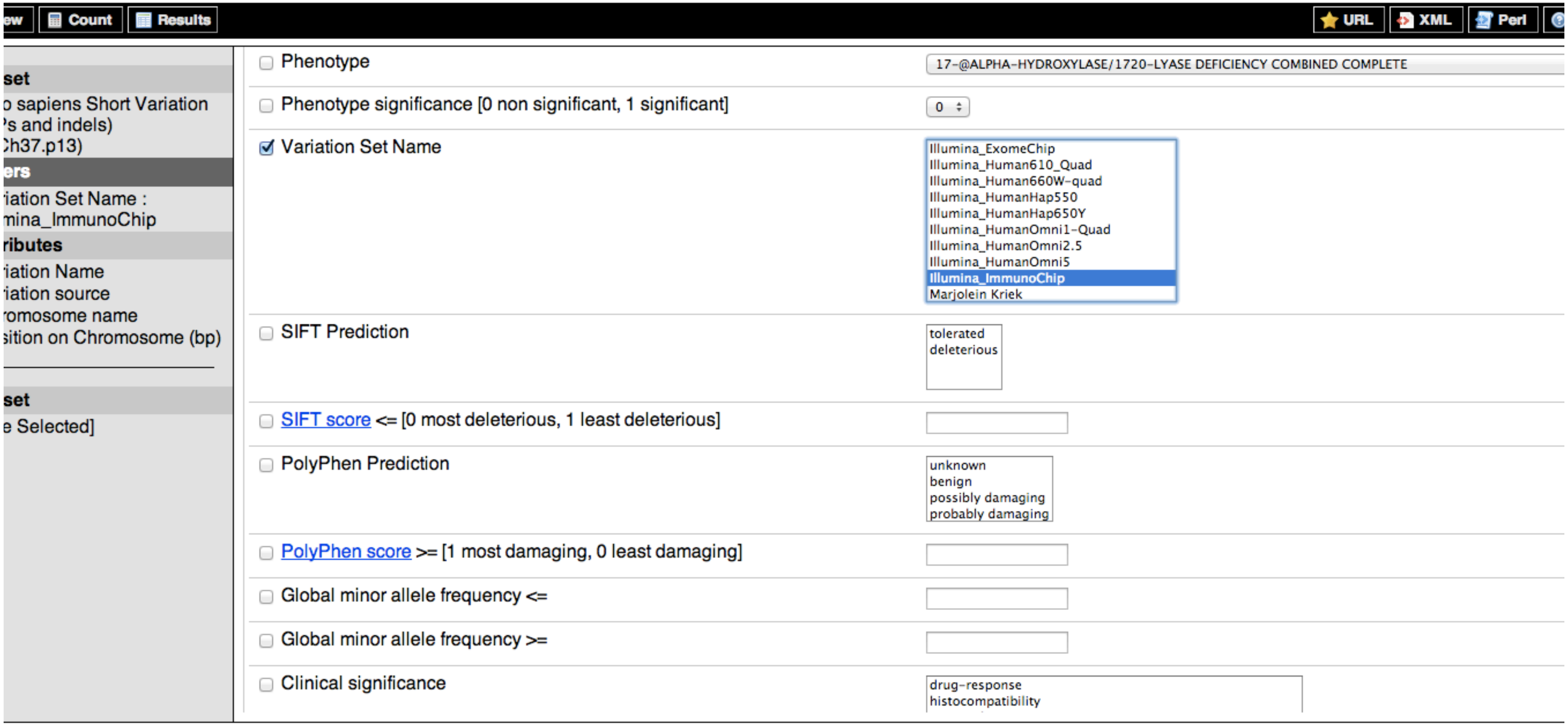
The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the ...gy and at Johns Hopkins University.

This instance of Galaxy is utilizing infrastructure generously provided by the iPlant Collaborative at the Texas Advanced Computing Center, with support from the

2nd attempt

BioMart

Galaxy



set
o sapiens Short Variation
's and indels)
Ch37.p13)

ers
iation Set Name :
mina_ImmunoChip

ributes
iation Name
iation source
romosome name
sition on Chromosome (bp)

set
e Selected]

□ Phenotype — 17-@ALPHA-HYDROXYLASE/1720-LYASE DEFICIENCY COMBINED COMPLETE

□ Phenotype significance [0 non significant, 1 significant] — 0

☑ Variation Set Name

- Illumina_ExomeChip
- Illumina_Human610_Quad
- Illumina_Human660W-quad
- Illumina_HumanHap550
- Illumina_HumanHap650Y
- Illumina_HumanOmni1-Quad
- Illumina_HumanOmni2.5
- Illumina_HumanOmni5
- **Illumina_ImmunoChip**
- Marjolein Kriek

□ SIFT Prediction

- tolerated
- deleterious

□ SIFT score <= [0 most deleterious, 1 least deleterious]

□ PolyPhen Prediction

- unknown
- benign
- possibly damaging
- probably damaging

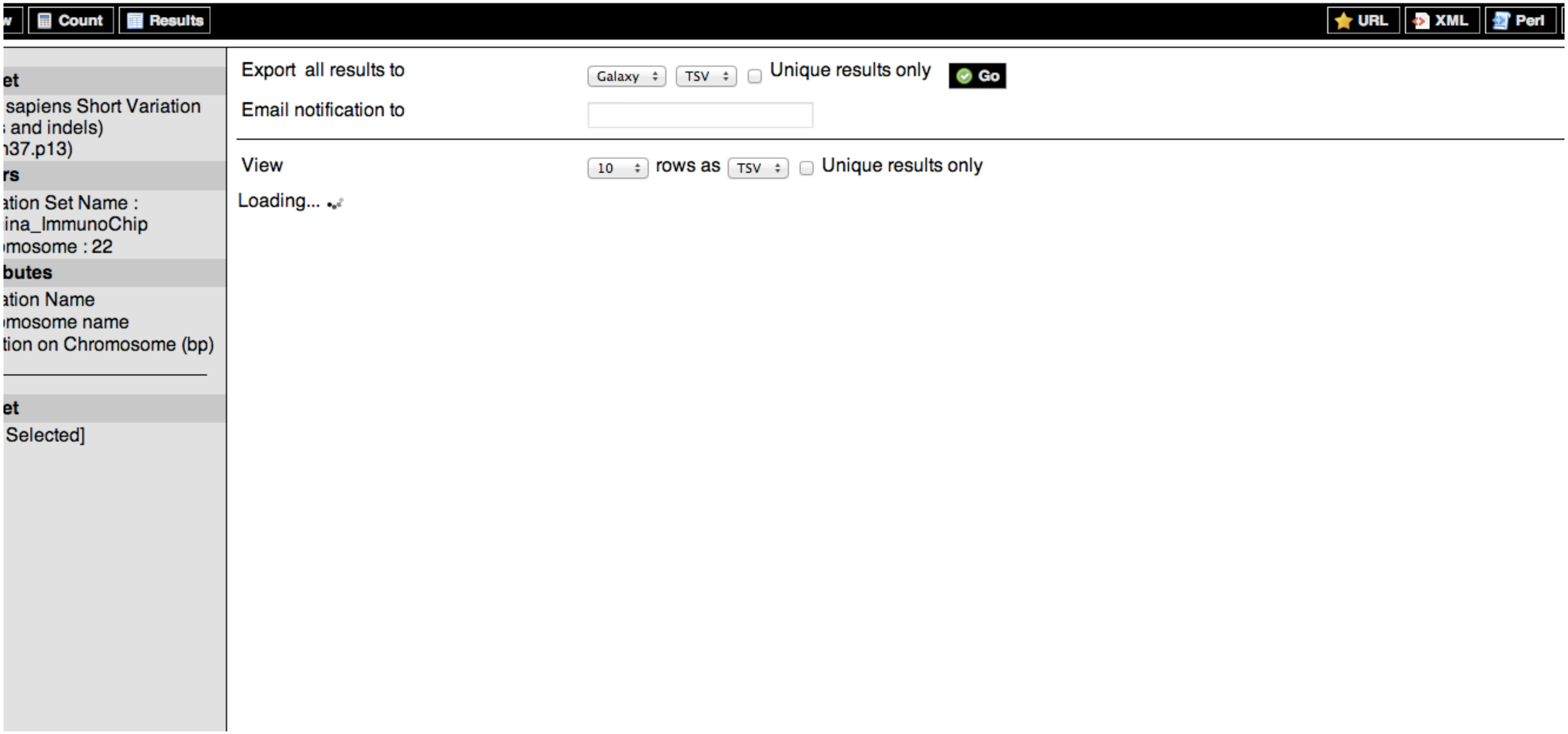
□ PolyPhen score >= [1 most damaging, 0 least damaging]

□ Global minor allele frequency <=

□ Global minor allele frequency >=

□ Clinical significance

- drug-response
- histocompatibility



w | Count | Results | URL | XML | Perl

et

sapiens Short Variation
; and indels)
ı37.p13)

rs

ation Set Name :
ina_ImmunoChip
ımosome : 22

butes

ation Name
ımosome name
tion on Chromosome (bp)

et

Selected]

Export all results to Galaxy TSV □ Unique results only Go

Email notification to

View 10 rows as TSV □ Unique results only

Loading...

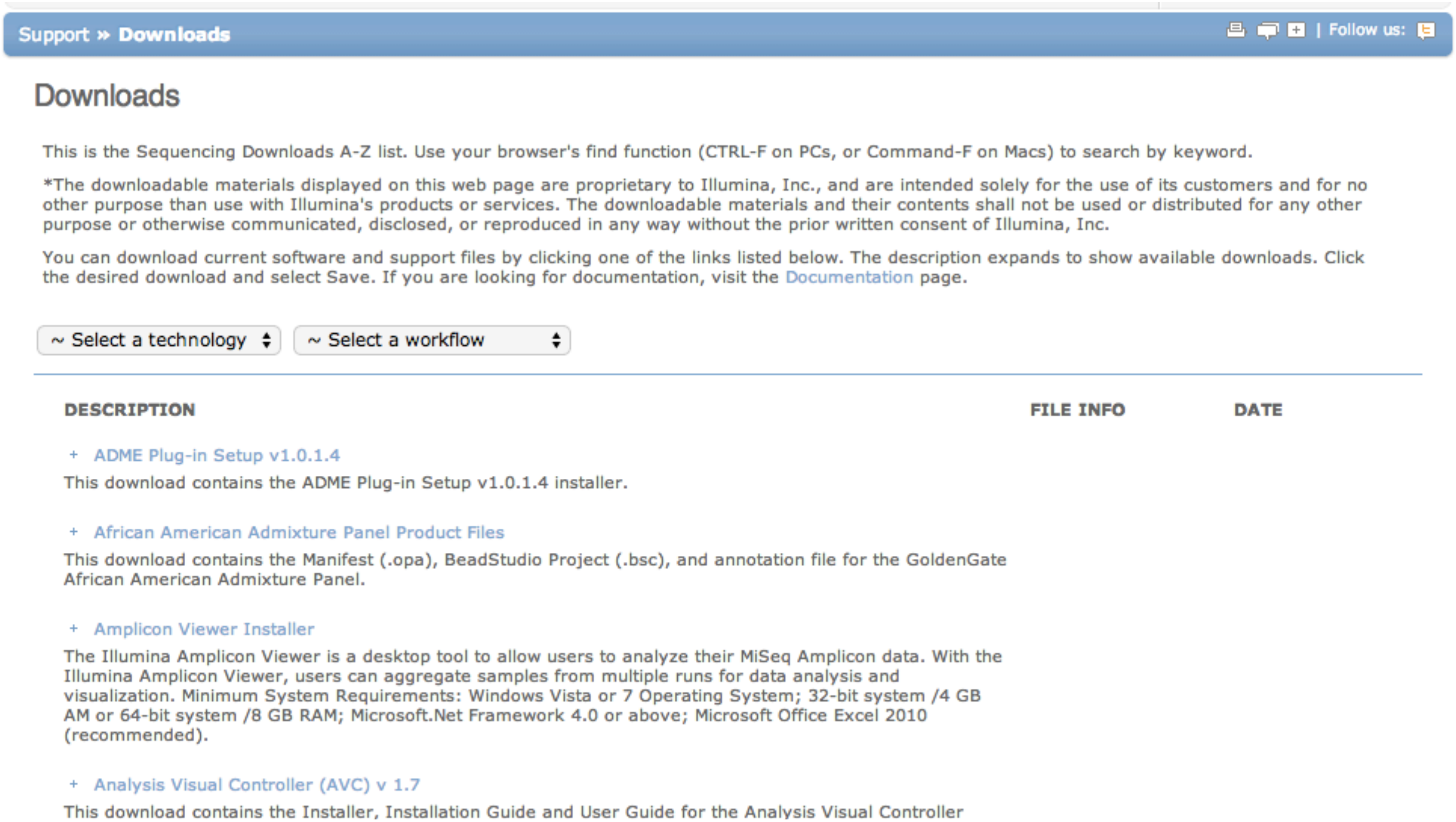
3rd attempt

Illumina website

Custom cleanup

Galaxy

Manufacturer annotations



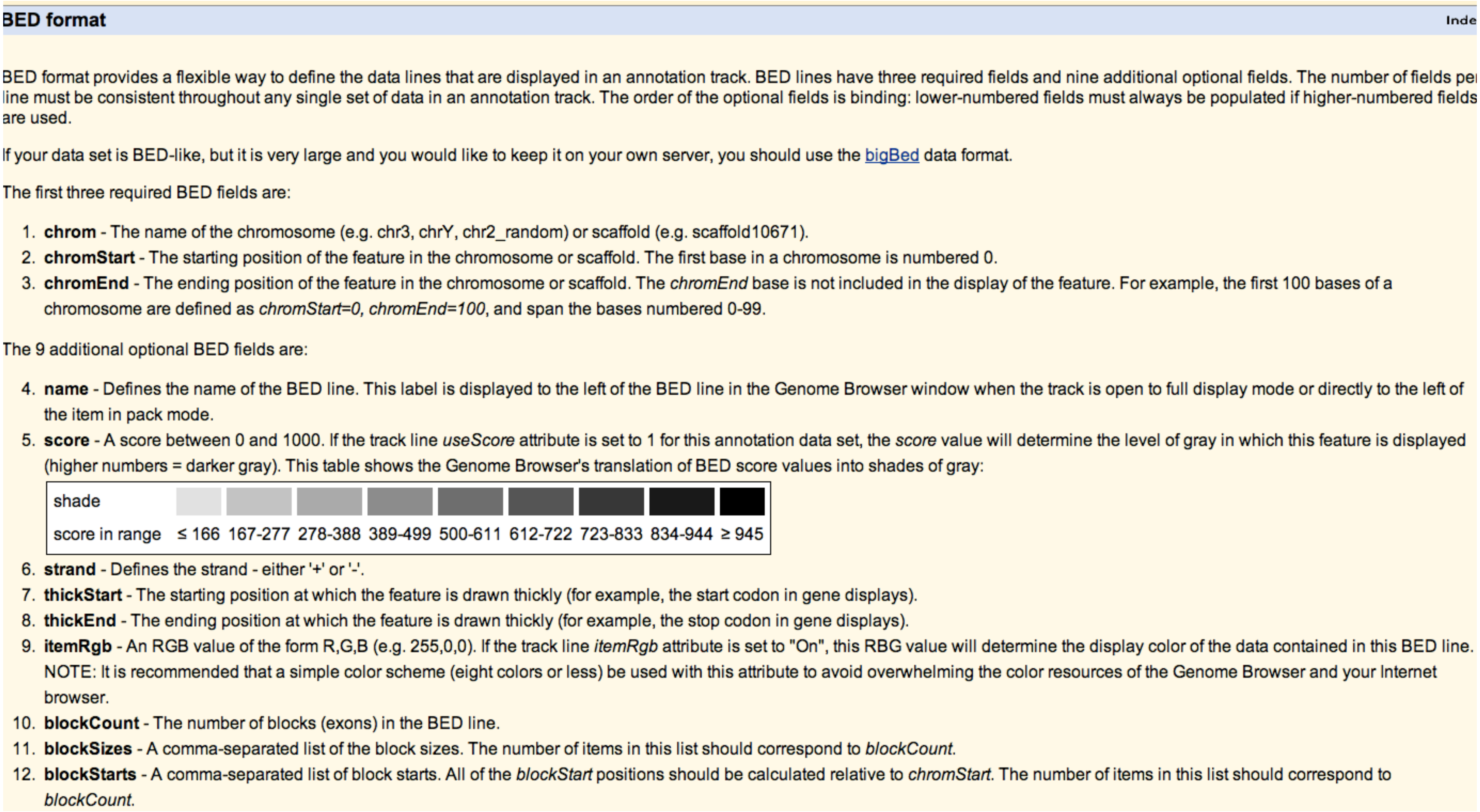
Downloads

This is the Sequencing Downloads A-Z list. Use your browser's find function (CTRL-F on PCs, or Command-F on Macs) to search by keyword.

*The downloadable materials displayed on this web page are proprietary to Illumina, Inc., and are intended solely for the use of its customers and for no other purpose than use with Illumina's products or services. The downloadable materials and their contents shall not be used or distributed for any other purpose or otherwise communicated, disclosed, or reproduced in any way without the prior written consent of Illumina, Inc.

You can download current software and support files by clicking one of the links listed below. The description expands to show available downloads. Click the desired download and select Save. If you are looking for documentation, visit the Documentation page.

DESCRIPTION	FILE INFO	DATE
+ ADME Plug-in Setup v1.0.1.4 This download contains the ADME Plug-in Setup v1.0.1.4 installer.		
+ African American Admixture Panel Product Files This download contains the Manifest (.opa), BeadStudio Project (.bsc), and annotation file for the GoldenGate African American Admixture Panel.		
+ Amplicon Viewer Installer The Illumina Amplicon Viewer is a desktop tool to allow users to analyze their MiSeq Amplicon data. With the Illumina Amplicon Viewer, users can aggregate samples from multiple runs for data analysis and visualization. Minimum System Requirements: Windows Vista or 7 Operating System; 32-bit system /4 GB AM or 64-bit system /8 GB RAM; Microsoft.Net Framework 4.0 or above; Microsoft Office Excel 2010 (recommended).		
+ Analysis Visual Controller (AVC) v 1.7 This download contains the Installer, Installation Guide and User Guide for the Analysis Visual Controller		



BED format

Inde

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

If your data set is BED-like, but it is very large and you would like to keep it on your own server, you should use the bigBed data format.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade	■	■	■	■	■	■	■	■	■
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays).
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

```
head immuno.csv
Illumina, Inc.
[Heading]
Descriptor File Name,Immuno_BeadChip_11419691_B.bpm
Assay Format,Infinium HD Ultra
Date Manufactured,11/01/2010
Loci Count ,196524
[Assay]
IlmnID,Name,IlmnStrand,SNP,AddressA_ID,AlleleA_ProbeSeq,AddressB_ID,AlleleB_ProbeSeq,GenomeBuild,Chr,MapInf
o,Ploidy,Species,Source,SourceVersion,SourceStrand,SourceSeq,TopGenomicSeq,BeadSetID
1-159076491-G-DELETION-1_P_F_1767851002,1-159076491-G-DELETION,PLUS,[D/I],
0049656396,ACAGCAATCCTGTGAGGTACTTATTATCACCCCCATTTTACTCAAGGGGG,,,36,1,159076491,diploid,Homo
sapiens,WTCCCseq,1,PLUS,TTTACTCTTAACAGCAATCCTGTGAGGTACTTATTATCACCCCCATTTTACTCAAGGGGG[-/
G]AAGAAAATTGAGGCTCAGAGAGGTTAATGAATCTGCCAGAGATCACAGAGCTTCTTTTTT,TTTACTCTTAACAGCAATCCTGTGA
GGTACTTATTATCACCCCCATTTTACTCAAGGGGG[-/
G]AAGAAAATTGAGGCTCAGAGAGGTTAATGAATCTGCCAGAGATCACAGAGCTTCTTTTTT,285
1-159093319-A-DELETION-1_M_R_1767851004,1-159093319-A-DELETION,MINUS,[I/D],
0043648403,TCTATTCTGCATATTAGTTGCCTGTAGGATTCATAGTTTGCAATTTTTTT,,,36,1,159093319,diploid,Homo
sapiens,WTCCCseq,1,PLUS,ATGTACAGTAAAGGAAACAATTCACAGAGTGAAAAGGCAACCAATAGAATAGGAAAAAAA[-/
A]TTGCAAACTATGAATCCTACAGGCAACTAATATGCAGAATAGACAAGAAATTCAAACGTC,ATGTACAGTAAAGGAAACAATTCAC
AGAGTGAAAAGGCAACCAATAGAATAGGAAAAAAA[-/
A]TTGCAAACTATGAATCCTACAGGCAACTAATATGCAGAATAGACAAGAAATTCAAACGTC,285
```

Steps

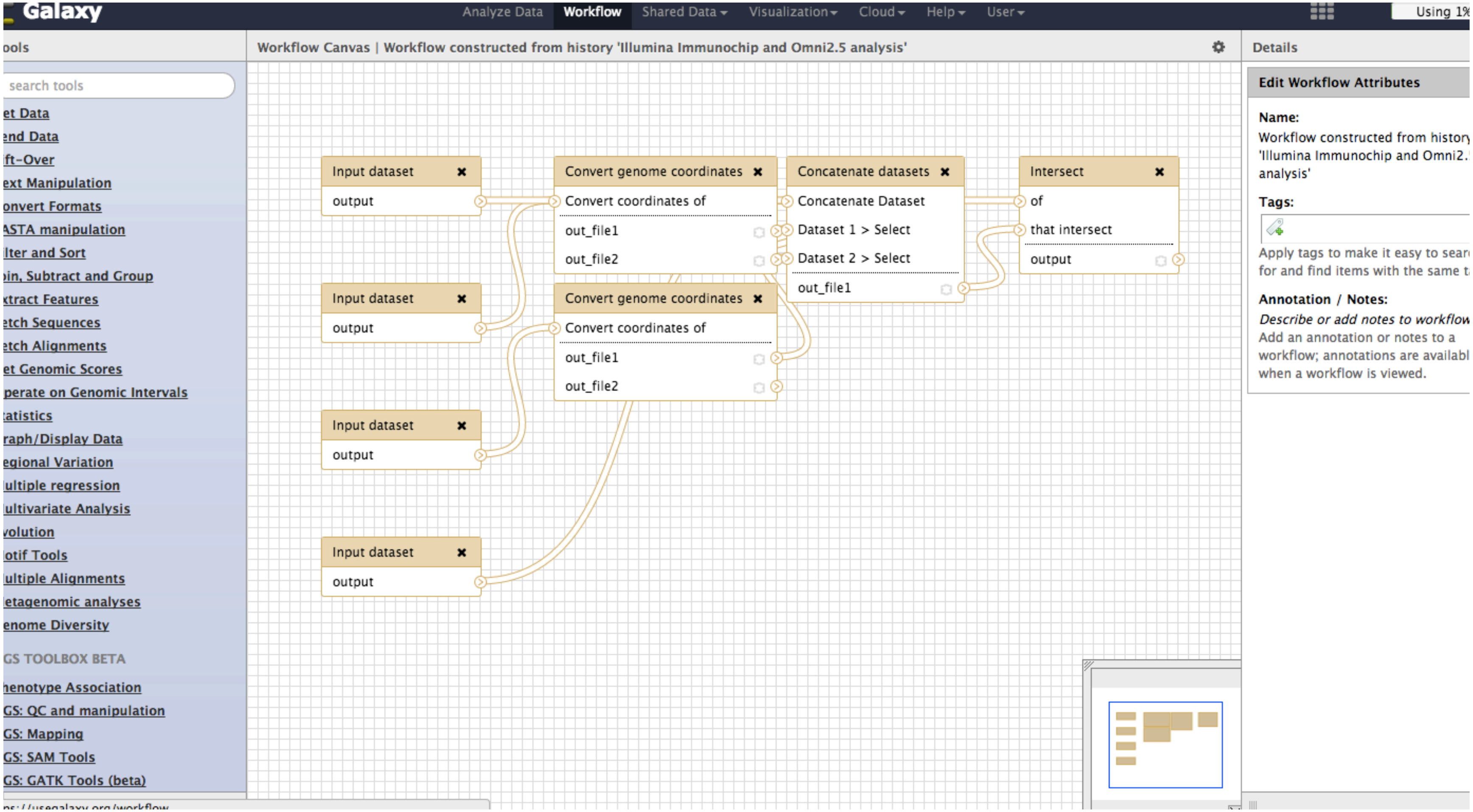
- Extract columns wanted
- Add end position
- Split by build numbers
- Upload to Galaxy
- Concatenate B36 annotations
- Liftover to B37

```
#!/bin/bash

# for some reason, the immuno annotation file has both build 36 and 37.1 annotations
# need to split them

cat input/annot/immuno_beadchip_11419691_b.csv \
        | grep -e ".*\[[ACGT]/[ACGT]" \
        | awk -F',' '{if ($9=="36"  )print "chr"$10"\t"$11"\t"$11+1"\t"$2}' \
        | grep -v -f skip > tmp/immuno-b36.bed
cat input/annot/immuno_beadchip_11419691_b.csv \
        | grep -e ".*\[[ACGT]/[ACGT]" \
        | awk -F',' '{if ($9=="36.2")print "chr"$10"\t"$11"\t"$11+1"\t"$2}' \
        | grep -v -f skip > tmp/immuno-b36.2.bed
cat input/annot/immuno_beadchip_11419691_b.csv \
        | grep -e ".*\[[ACGT]/[ACGT]" \
        | awk -F',' '{if ($9=="37.1")print "chr"$10"\t"$11"\t"$11+1"\t"$2}' \
        | grep -v -f skip > tmp/immuno-b37.bed
cat input/annot/humanomni25m-8v1-1_b.csv \
        | grep -e ".*\[[ACGT]/[ACGT]" \
        | awk -F',' '{if ($9=="37.1")print "chr"$10"\t"$11"\t"$11+1"\t"$2}' \
        | grep -v -f skip > tmp/omni-b37.bed
```

```
head tmp/omni-b37.bed
chrY    21762685 21762686 200610-147
chrY    21779251 21779252 200610-148
chrY    21867854 21867855 200610-149
chrY    21751440 21751441 200610-150
chrY    21888865 21888866 200610-151
chrY    21730357 21730358 200610-152
chrY    21740450 21740451 200610-153
chrY    21753199 21753200 200610-155
chrY    21868776 21868777 200610-156
chrY    17286006 17286007 200610-158
```



Workflow Canvas | Workflow constructed from history 'Illumina Immunochip and Omni2.5 analysis'

Edit Workflow Attributes

Name:
Workflow constructed from history 'Illumina Immunochip and Omni2.5 analysis'

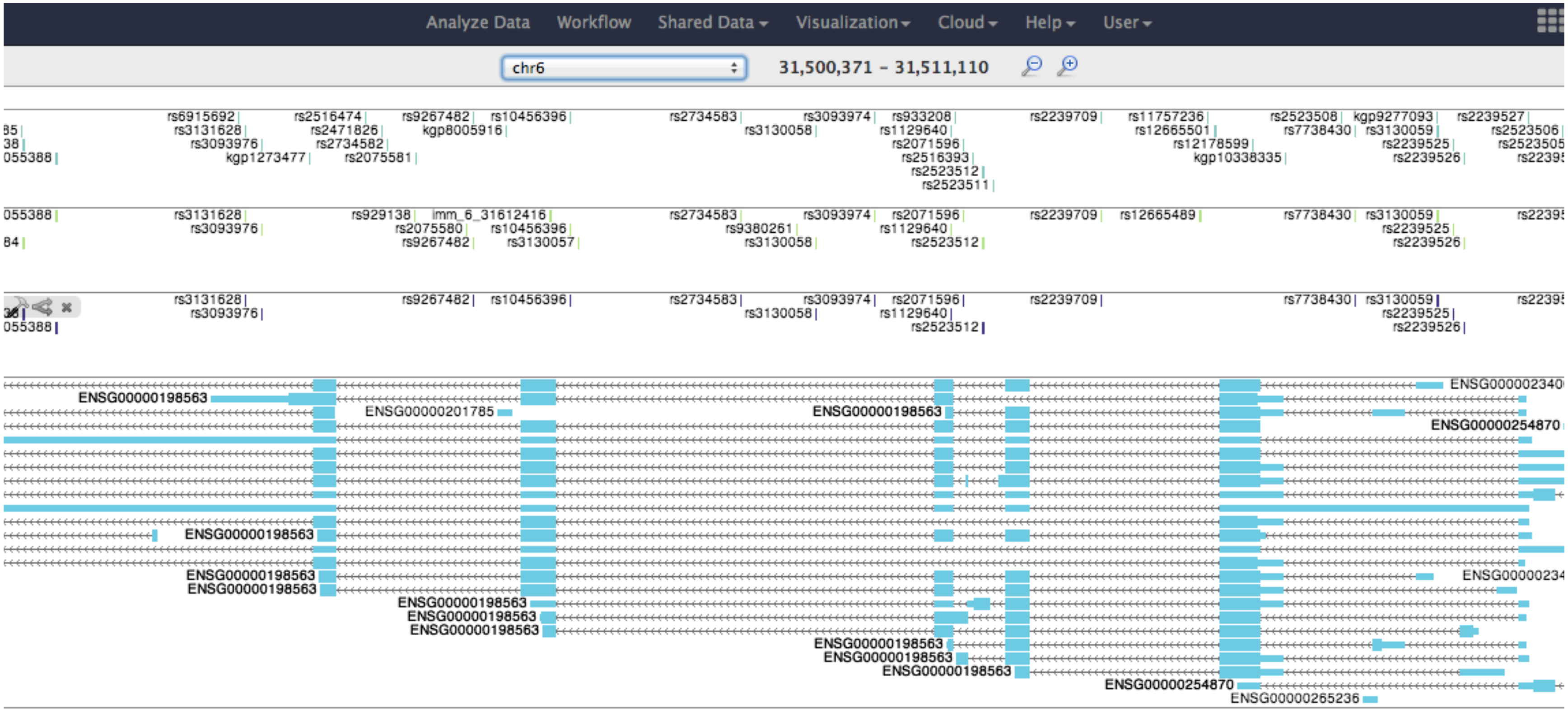
Tags:

Apply tags to make it easy to sear for and find items with the same t

Annotation / Notes:

Describe or add notes to workflow

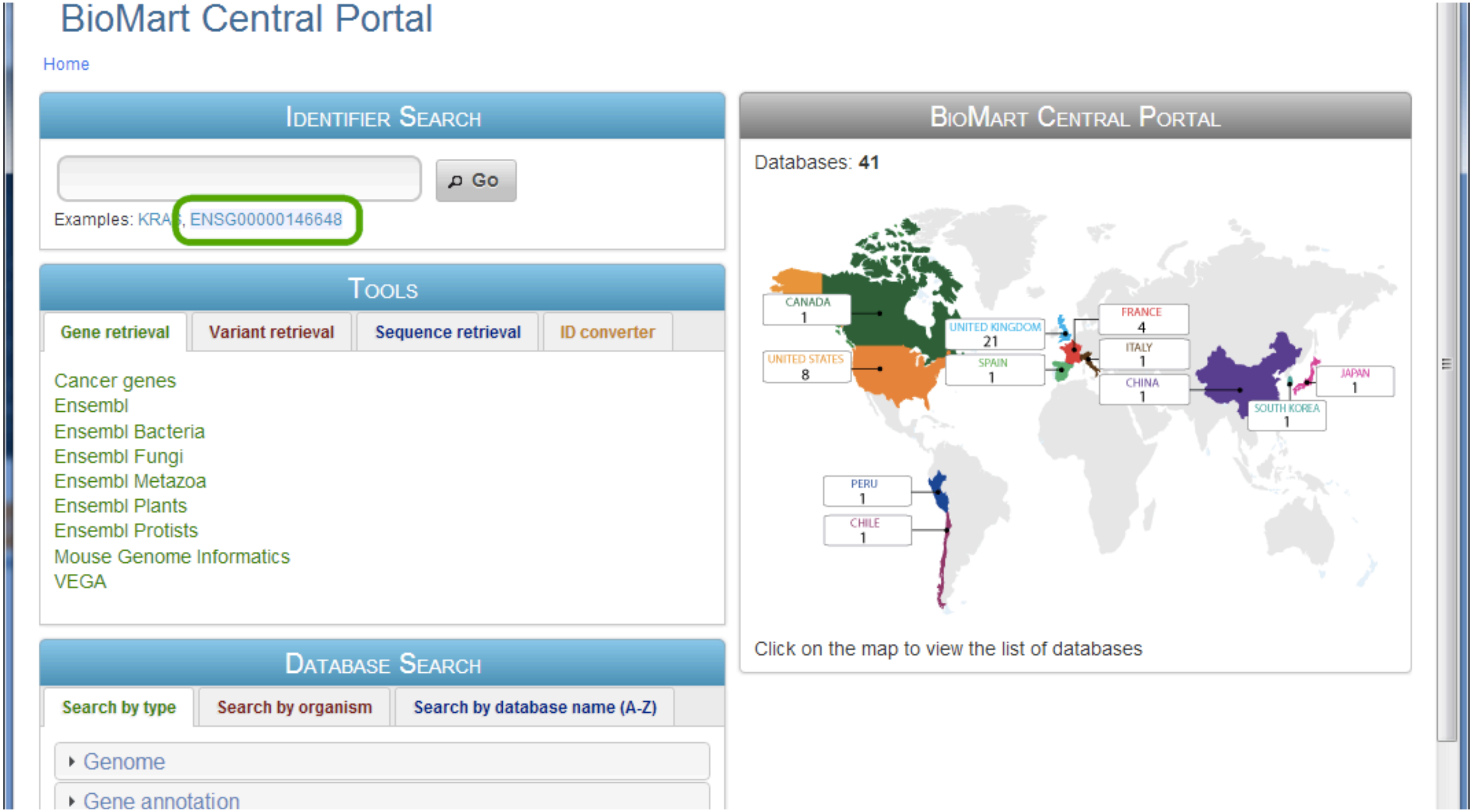
Add an annotation or notes to a workflow; annotations are availabl when a workflow is viewed.


Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
chr6
31,500,371 - 31,511,110
rs6915692
rs2516474
rs9267482
rs10456396
rs2734583
rs3093974
rs933208
rs2239709
rs11757236
rs2523508
kgp9277093
rs2239527
rs3131628
rs2471826
kgp8005916
rs3130058
rs1129640
rs12665501
rs7738430
rs3130059
rs2523506
rs3093976
rs2734582
rs2071596
rs12178599
rs2239525
rs2523505
kgp1273477
rs2075581
rs2516393
kgp10338335
rs2239526
rs2523512
rs2523511
rs929138
imm_6_31612416
rs12665489
rs2075580
rs9380261
rs3130057
ENSG00000198563
ENSG00000201785
ENSG00000254870
ENSG00000265236

Question 2

I would like to view the allele frequencies of my study populations and compare them to the frequencies found in the HapMap populations, for the following genes

BioMart 0.8



BioMart Central Portal

Home

Identifier Search

⌕ Go

Examples: KRAS, ENSG00000146648

Tools

Gene retrieval | Variant retrieval | Sequence retrieval | ID converter

Cancer genes
Ensembl
Ensembl Bacteria
Ensembl Fungi
Ensembl Metazoa
Ensembl Plants
Ensembl Protists
Mouse Genome Informatics
VEGA

Database Search

Search by type | Search by organism | Search by database name (A-Z)

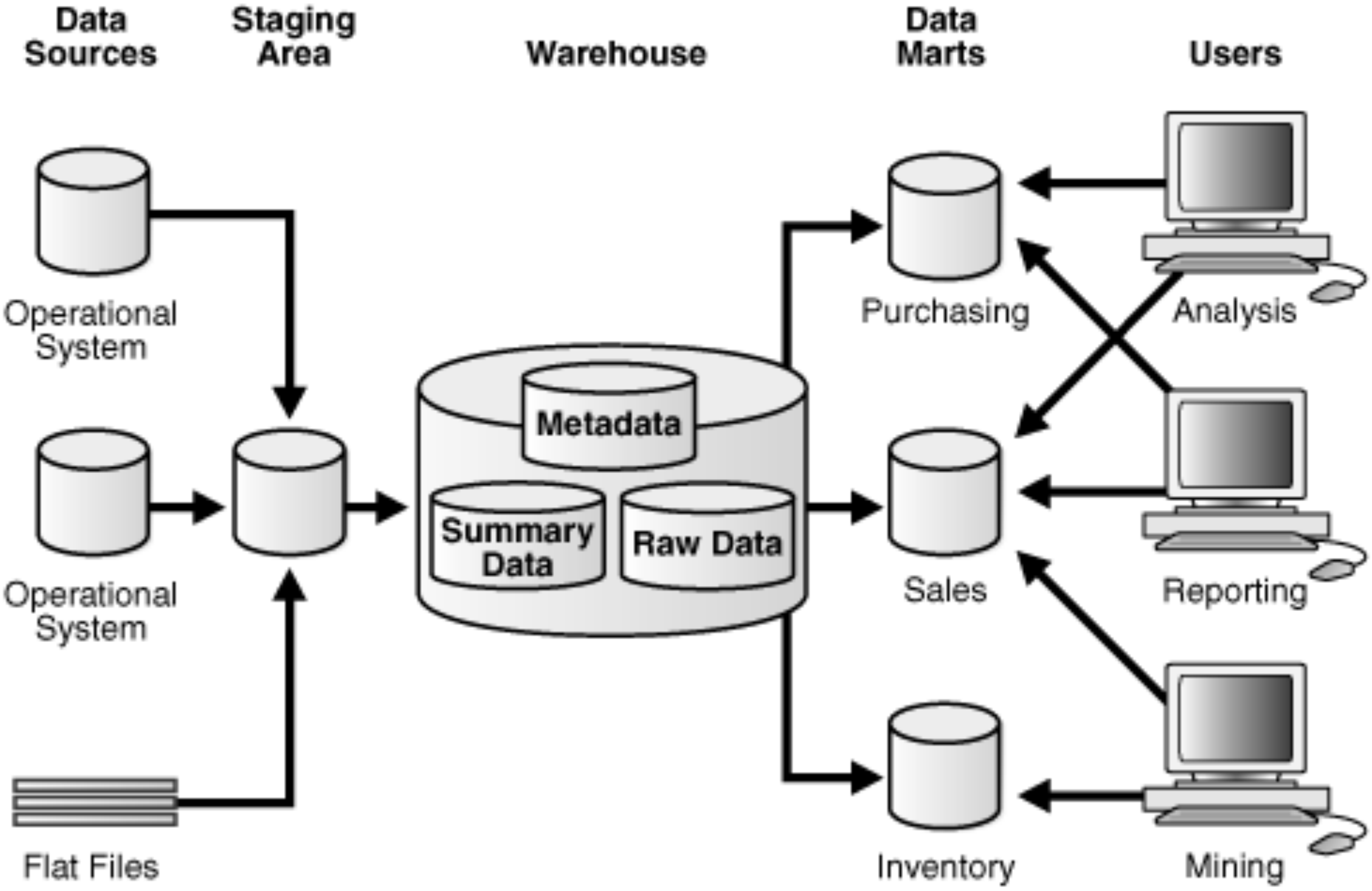
‣ Genome
‣ Gene annotation

BioMart Central Portal

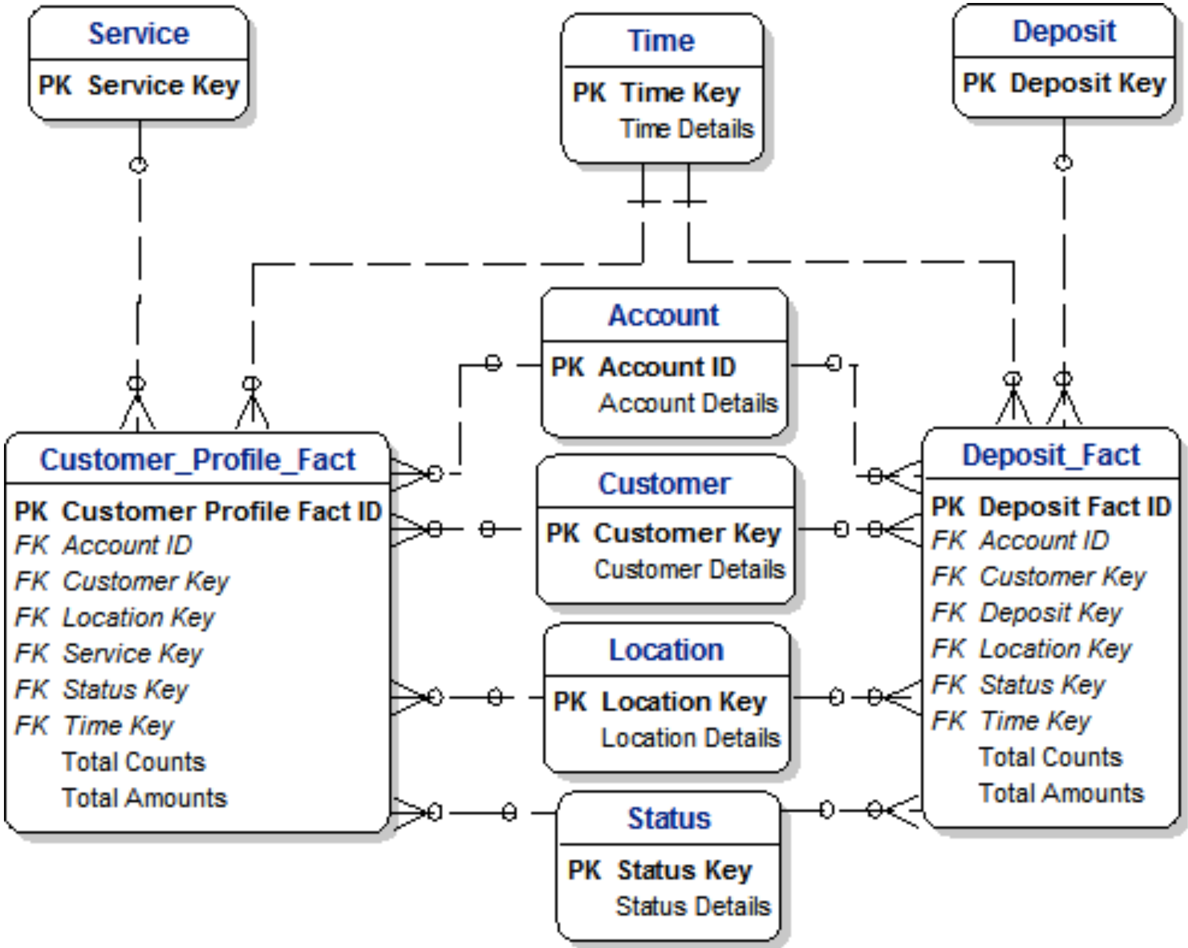
Databases: **41**

Click on the map to view the list of databases

Data marts


Data Sources
Staging Area
Warehouse
Data Marts
Users
Operational System
Operational System
Flat Files
Metadata
Summary Data
Raw Data
Purchasing
Sales
Inventory
Analysis
Reporting
Mining

Databases



Service
PK Service Key
Time
PK Time Key
Time Details
Deposit
PK Deposit Key
Account
PK Account ID
Account Details
Customer_Profile_Fact
PK Customer Profile Fact ID
FK Account ID
FK Customer Key
FK Location Key
FK Service Key
FK Status Key
FK Time Key
Total Counts
Total Amounts
Customer
PK Customer Key
Customer Details
Deposit_Fact
PK Deposit Fact ID
FK Account ID
FK Customer Key
FK Deposit Key
FK Location Key
FK Status Key
FK Time Key
Total Counts
Total Amounts
Location
PK Location Key
Location Details
Status
PK Status Key
Status Details

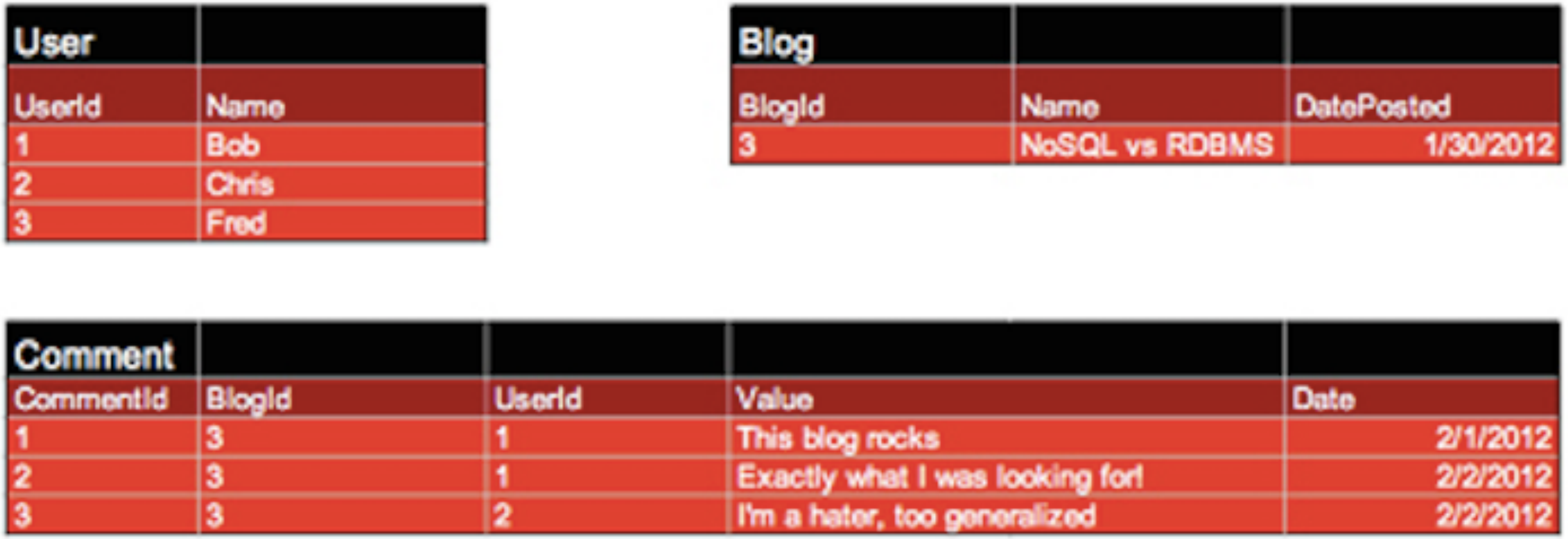
BioMart

- Integration
- Not “live”
- R interface (biomaRt)

SQL query

```
SELECT main.rsid_106, main.chromosome_106, main.position_106, afmap_mart.marker__allele_frequency_KHS__dm.a_freq_101,
afmap_mart.marker__allele_frequency_HER__dm.a_freq_101, afmap_mart.marker__allele_frequency_STS__dm.a_freq_101,
afmap_mart.marker__allele_frequency_CON__dm.a_freq_101, afmap_mart.marker__allele_frequency_ZUL__dm.a_freq_101,
afmap_mart.marker__allele_frequency_ASW__dm.a_freq_101, afmap_mart.marker__allele_frequency_CEU__dm.a_freq_101,
afmap_mart.marker__allele_frequency_CHB__dm.a_freq_101, afmap_mart.marker__allele_frequency_CHD__dm.a_freq_101,
afmap_mart.marker__allele_frequency_GIH__dm.a_freq_101, afmap_mart.marker__allele_frequency_JPT__dm.a_freq_101,
afmap_mart.marker__allele_frequency_LWK__dm.a_freq_101, afmap_mart.marker__allele_frequency_MEX__dm.a_freq_101,
afmap_mart.marker__allele_frequency_MKK__dm.a_freq_101, afmap_mart.marker__allele_frequency_TSI__dm.a_freq_101,
afmap_mart.marker__allele_frequency_YRI__dm.a_freq_101 FROM afmap_mart.marker__allele_frequency_TSI__dm,
afmap_mart.marker__allele_frequency_CHD__dm, afmap_mart.marker__allele_frequency_ZUL__dm,
afmap_mart.marker__allele_frequency_KHS__dm, afmap_mart.marker__allele_frequency_LWK__dm,
afmap_mart.marker__allele_frequency_CHB__dm, afmap_mart.marker__allele_frequency_ASW__dm,
afmap_mart.marker__allele_frequency_JPT__dm, afmap_mart.marker__allele_frequency_STS__dm,
afmap_mart.marker__allele_frequency_HER__dm, afmap_mart.marker__allele_frequency_CEU__dm,
afmap_mart.marker__allele_frequency_MEX__dm, afmap_mart.marker__allele_frequency_MKK__dm,
afmap_mart.marker__allele_frequency_CON__dm, afmap_mart.marker__allele_frequency_YRI__dm,
afmap_mart.marker__allele_frequency_GIH__dm, afmap_mart.marker__association__main main WHERE (main.genesymbol_103 like '%BRCA2%')
AND main.id_106_key=afmap_mart.marker__allele_frequency_TSI__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_KHS__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_ASW__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_CEU__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_YRI__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_ZUL__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_CHD__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_JPT__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_GIH__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_STS__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_CON__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_MKK__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_MEX__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_LWK__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_HER__dm.id_106_key AND
main.id_106_key=afmap_mart.marker__allele_frequency_CHB__dm.id_106_key
```

Traditional Relational DB Schema

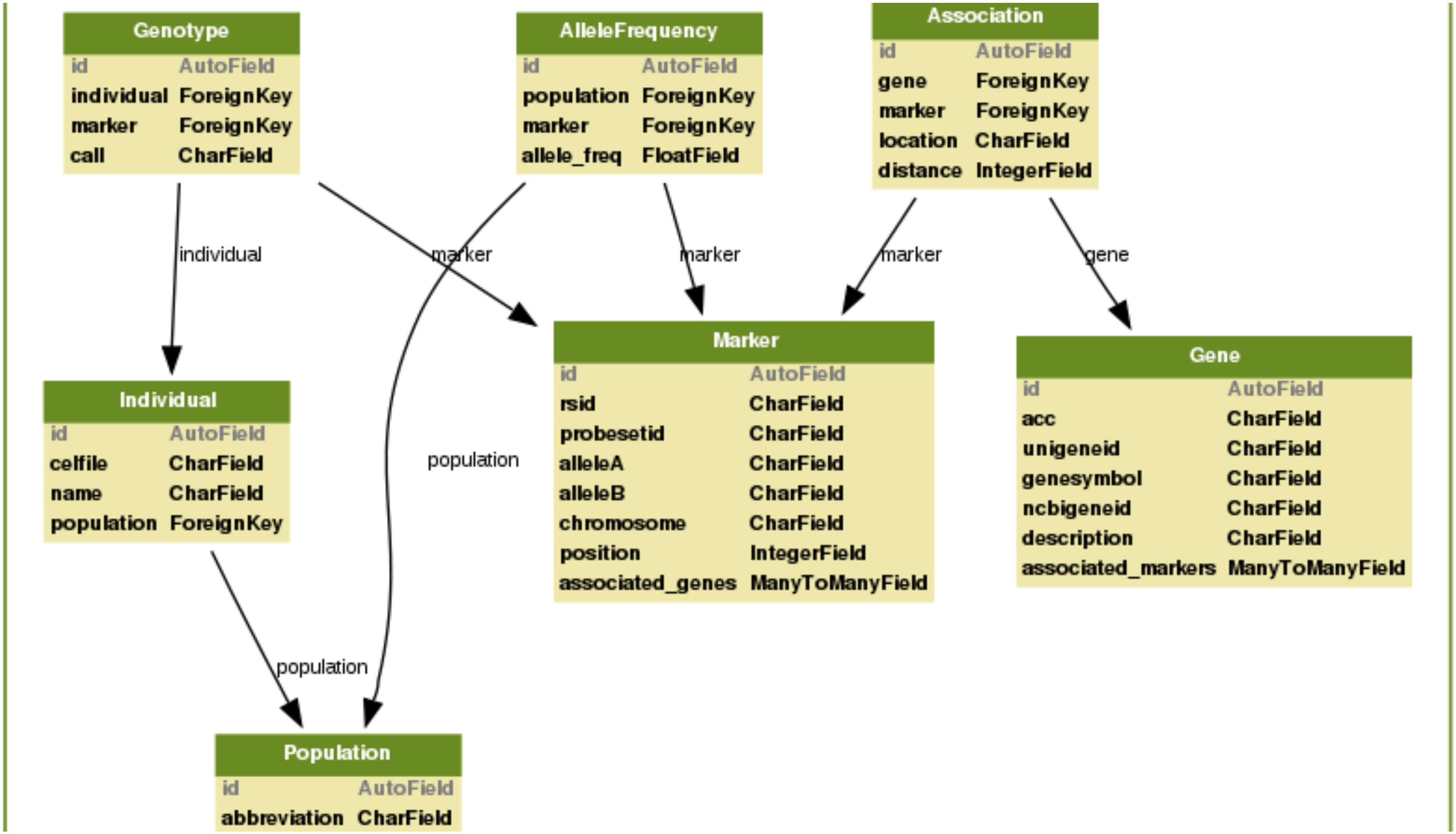


User	
UserId	Name
1	Bob
2	Chris
3	Fred

Blog		
BlogId	Name	DatePosted
3	NoSQL vs RDBMS	1/30/2012

Comment				
CommentId	BlogId	UserId	Value	Date
1	3	1	This blog rocks	2/1/2012
2	3	1	Exactly what I was looking for!	2/2/2012
3	3	2	I'm a hater, too generalized	2/2/2012

Relational



Genotype

id	AutoField
individual	ForeignKey
marker	ForeignKey
call	CharField

AlleleFrequency

id	AutoField
population	ForeignKey
marker	ForeignKey
allele_freq	FloatField

Association

id	AutoField
gene	ForeignKey
marker	ForeignKey
location	CharField
distance	IntegerField

Individual

id	AutoField
celfile	CharField
name	CharField
population	ForeignKey

Marker

id	AutoField
rsid	CharField
probesetid	CharField
alleleA	CharField
alleleB	CharField
chromosome	CharField
position	IntegerField
associated_genes	ManyToManyField

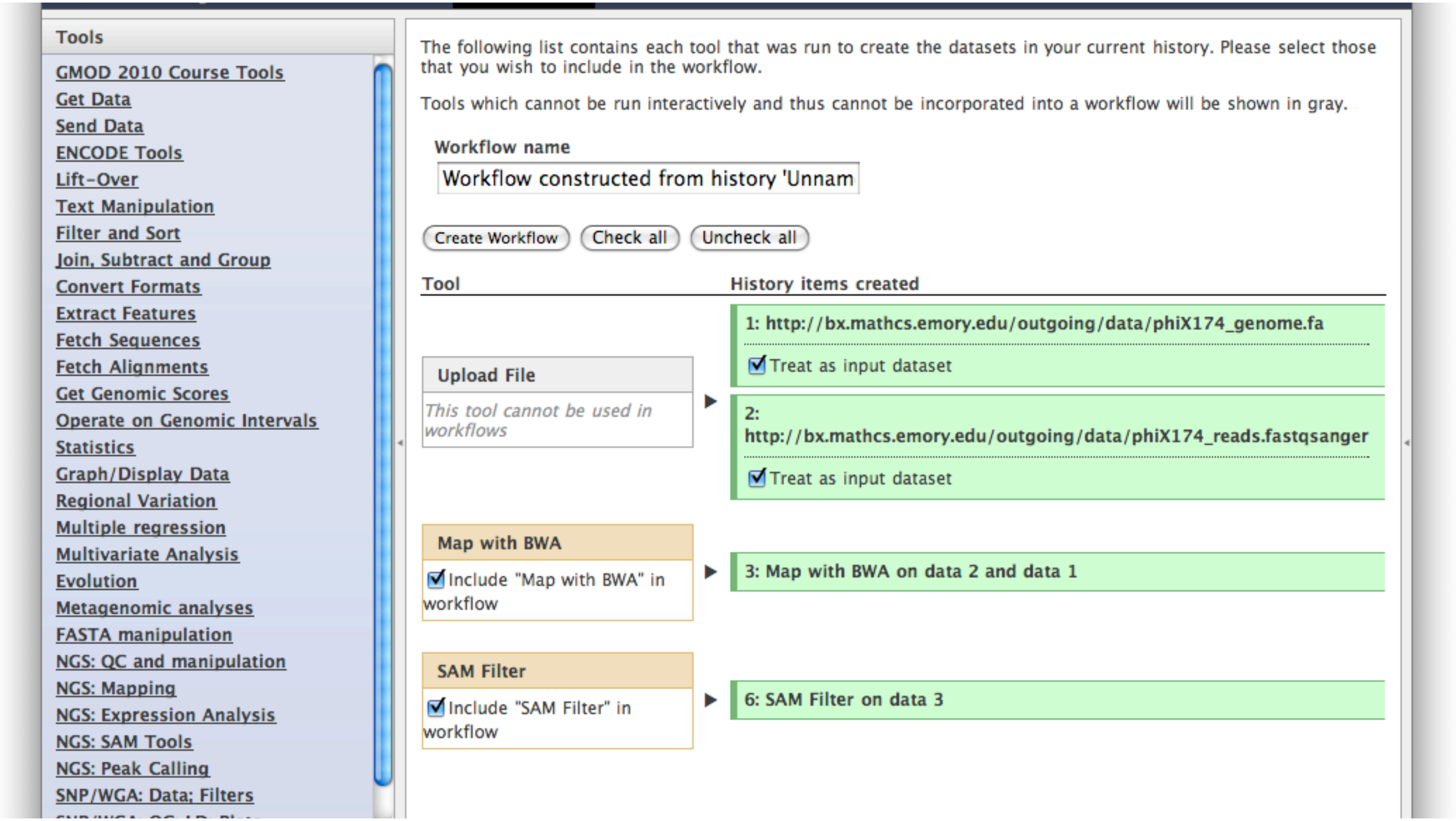
Gene

id	AutoField
acc	CharField
unigeneid	CharField
genesymbol	CharField
ncbigeneid	CharField
description	CharField
associated_markers	ManyToManyField

Population

id	AutoField
abbreviation	CharField

Galaxy

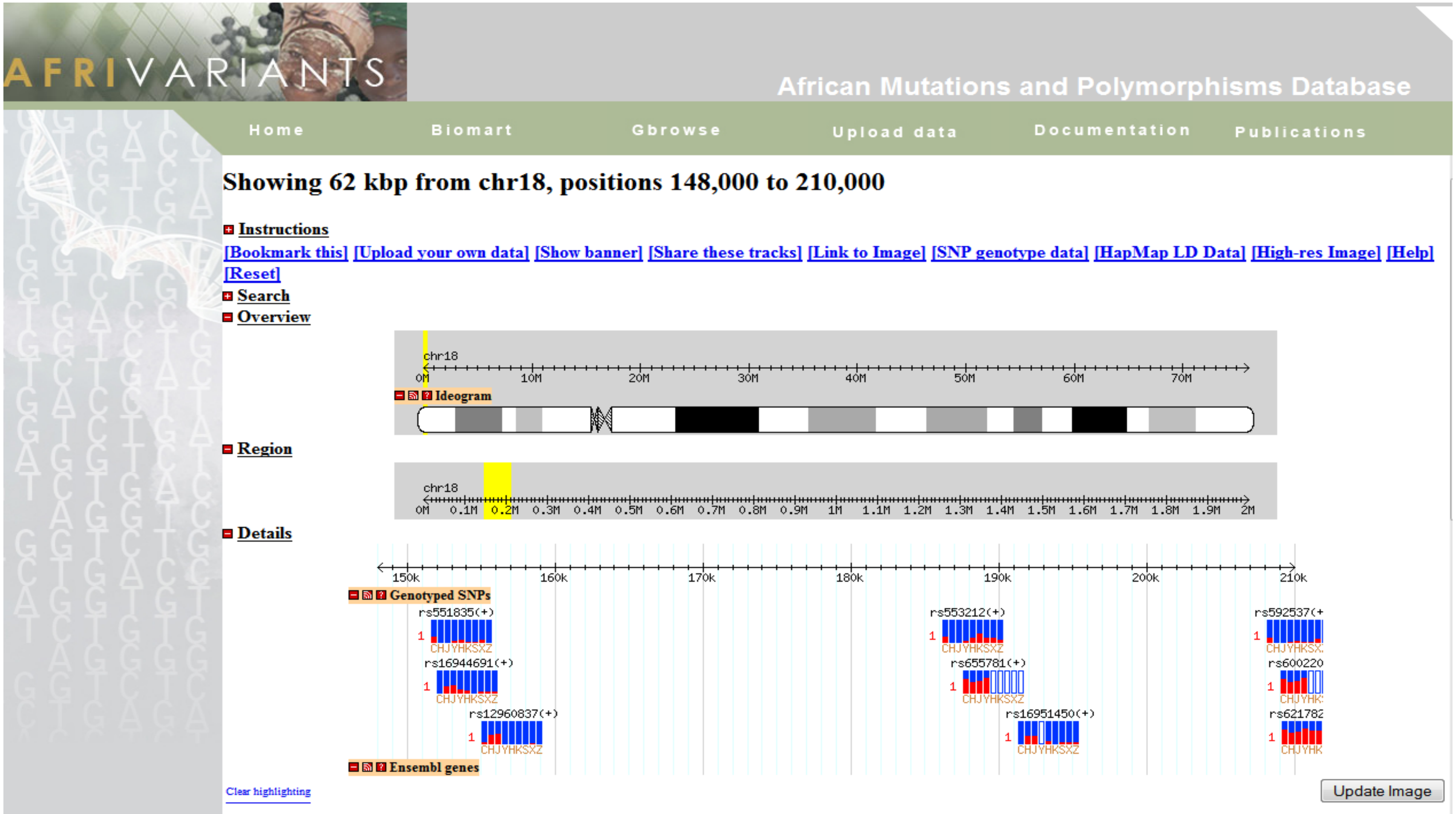


The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name: Workflow constructed from history 'Unnam

Tool	History items created
Upload File *This tool cannot be used in workflows*	1: http://bx.mathcs.emory.edu/outgoing/data/phiX174_genome.fa ☑ Treat as input dataset 2: http://bx.mathcs.emory.edu/outgoing/data/phiX174_reads.fastqsanger ☑ Treat as input dataset
Map with BWA ☑ Include "Map with BWA" in workflow	3: Map with BWA on data 2 and data 1
SAM Filter ☑ Include "SAM Filter" in workflow	6: SAM Filter on data 3



AFRIVARIANTS

African Mutations and Polymorphisms Database

Home Biomart Gbrowse Upload data Documentation Publications

Showing 62 kbp from chr18, positions 148,000 to 210,000

⊞ **Instructions**

[Bookmark this] [Upload your own data] [Show banner] [Share these tracks] [Link to Image] [SNP genotype data] [HapMap LD Data] [High-res Image] [Help] [Reset]

⊞ **Search**

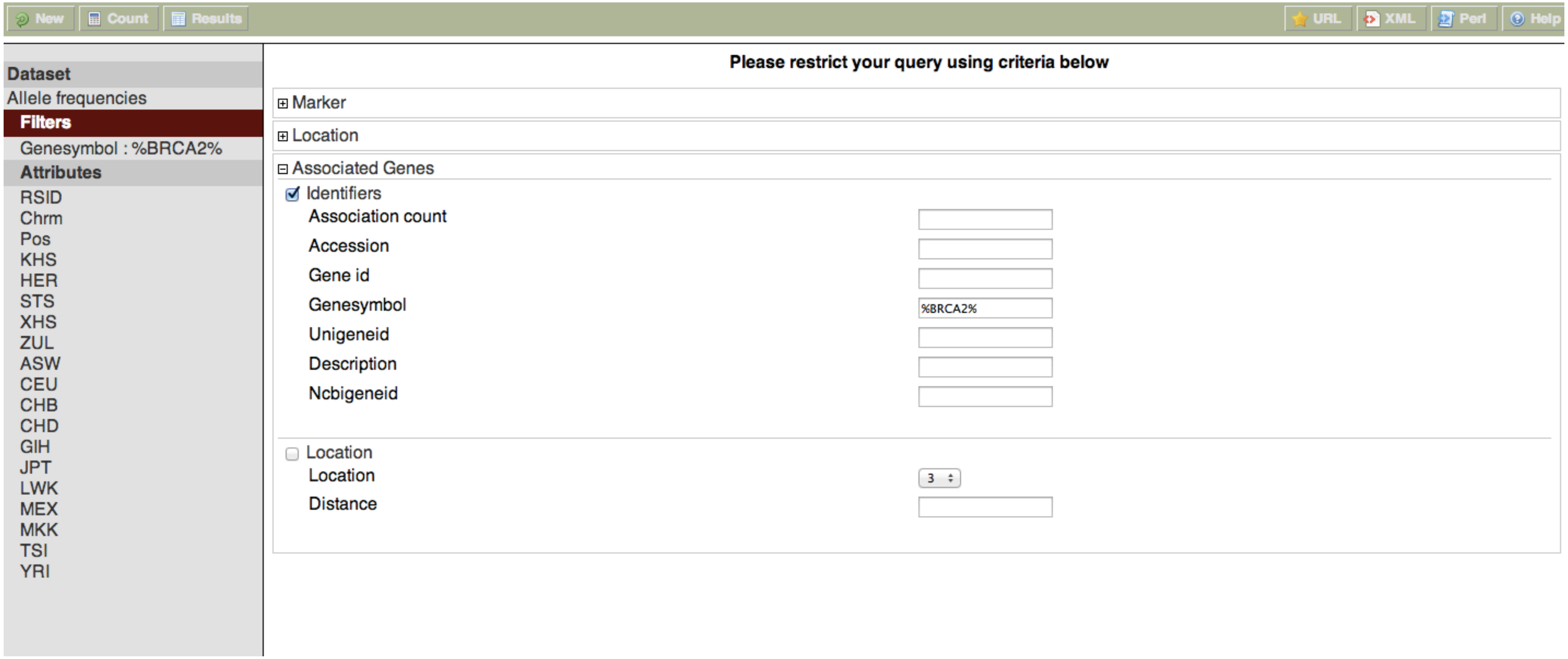
⊟ **Overview**

⊟ **Region**

⊟ **Details**

Clear highlighting

Update Image



Dataset

Allele frequencies

Filters

Genesymbol : %BRCA2%

Attributes

RSID
Chrm
Pos
KHS
HER
STS
XHS
ZUL
ASW
CEU
CHB
CHD
GIH
JPT
LWK
MEX
MKK
TSI
YRI

Please restrict your query using criteria below

⊞ Marker

⊞ Location

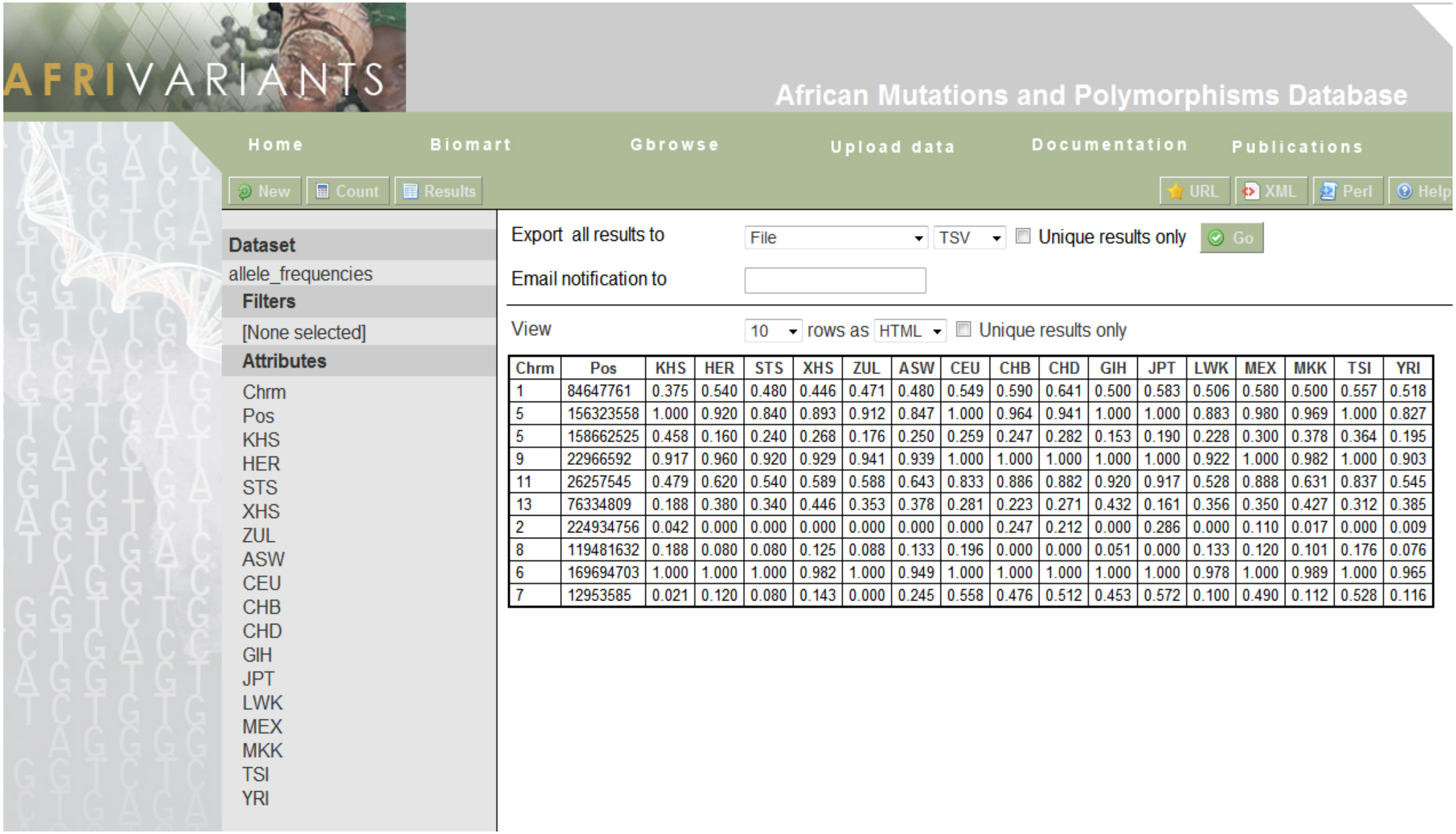
⊟ Associated Genes

☑ Identifiers

Association count	
Accession	
Gene id	
Genesymbol	%BRCA2%
Unigeneid	
Description	
Ncbigeneid	

☐ Location

Location	3
Distance	



Dataset

allele_frequencies

Filters

[None selected]

Attributes

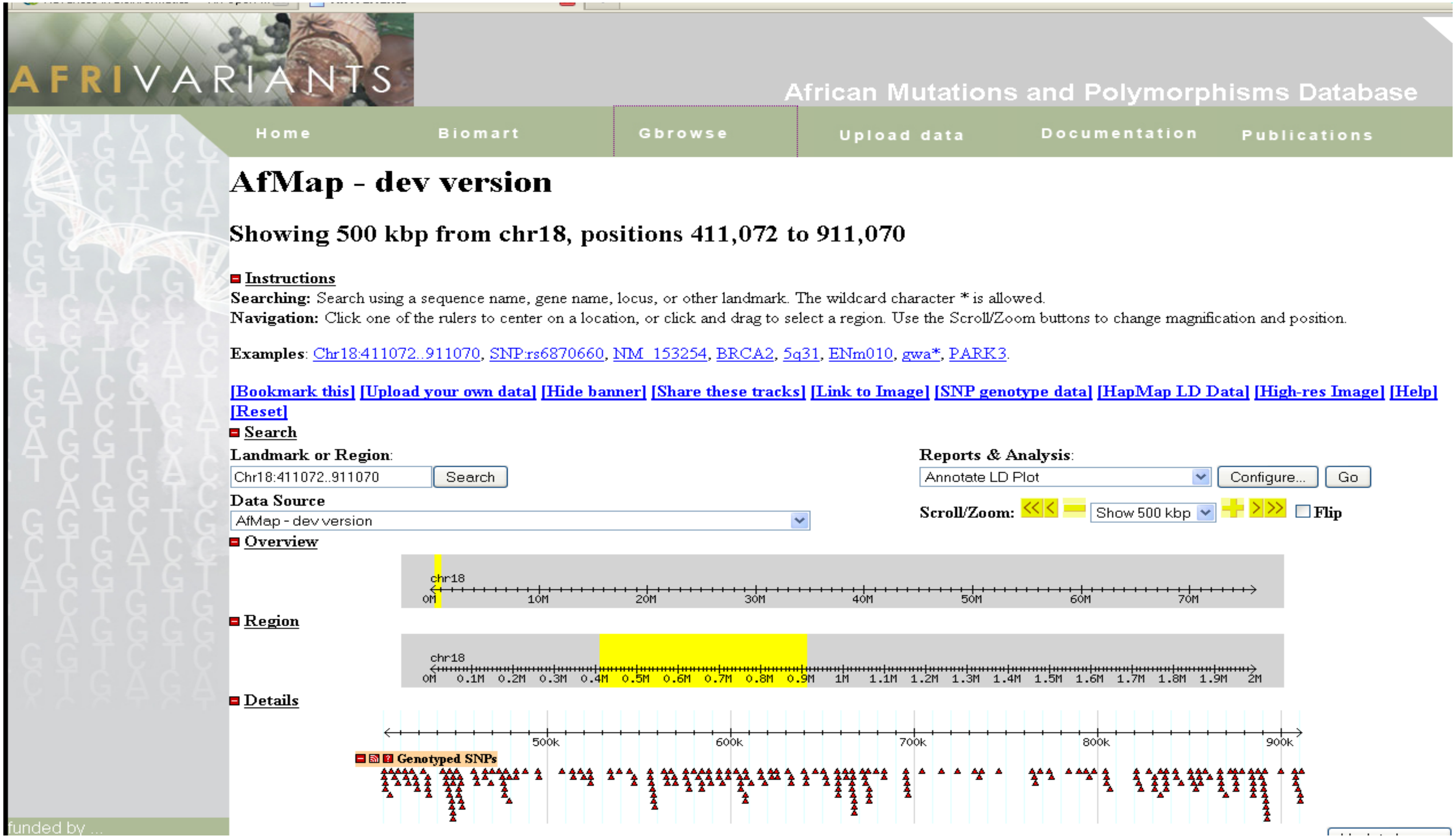
Chrm
Pos
KHS
HER
STS
XHS
ZUL
ASW
CEU
CHB
CHD
GIH
JPT
LWK
MEX
MKK
TSI
YRI

Export all results to File TSV □ Unique results only Go

Email notification to

View 10 rows as HTML □ Unique results only

Chrm	Pos	KHS	HER	STS	XHS	ZUL	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
1	84647761	0.375	0.540	0.480	0.446	0.471	0.480	0.549	0.590	0.641	0.500	0.583	0.506	0.580	0.500	0.557	0.518
5	156323558	1.000	0.920	0.840	0.893	0.912	0.847	1.000	0.964	0.941	1.000	1.000	0.883	0.980	0.969	1.000	0.827
5	158662525	0.458	0.160	0.240	0.268	0.176	0.250	0.259	0.247	0.282	0.153	0.190	0.228	0.300	0.378	0.364	0.195
9	22966592	0.917	0.960	0.920	0.929	0.941	0.939	1.000	1.000	1.000	1.000	1.000	0.922	1.000	0.982	1.000	0.903
11	26257545	0.479	0.620	0.540	0.589	0.588	0.643	0.833	0.886	0.882	0.920	0.917	0.528	0.888	0.631	0.837	0.545
13	76334809	0.188	0.380	0.340	0.446	0.353	0.378	0.281	0.223	0.271	0.432	0.161	0.356	0.350	0.427	0.312	0.385
2	224934756	0.042	0.000	0.000	0.000	0.000	0.000	0.000	0.247	0.212	0.000	0.286	0.000	0.110	0.017	0.000	0.009
8	119481632	0.188	0.080	0.080	0.125	0.088	0.133	0.196	0.000	0.000	0.051	0.000	0.133	0.120	0.101	0.176	0.076
6	169694703	1.000	1.000	1.000	0.982	1.000	0.949	1.000	1.000	1.000	1.000	1.000	0.978	1.000	0.989	1.000	0.965
7	12953585	0.021	0.120	0.080	0.143	0.000	0.245	0.558	0.476	0.512	0.453	0.572	0.100	0.490	0.112	0.528	0.116



AfMap - dev version

Showing 500 kbp from chr18, positions 411,072 to 911,070

Instructions

Searching: Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed.

Navigation: Click one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position.

Examples: Chr18:411072..911070, SNP:rs6870660, NM_153254, BRCA2, 5q31, ENm010, gwa*, PARK3.

[Bookmark this] [Upload your own data] [Hide banner] [Share these tracks] [Link to Image] [SNP genotype data] [HapMap LD Data] [High-res Image] [Help] [Reset]

Search

Landmark or Region: Chr18:411072..911070

Data Source: AfMap - dev version

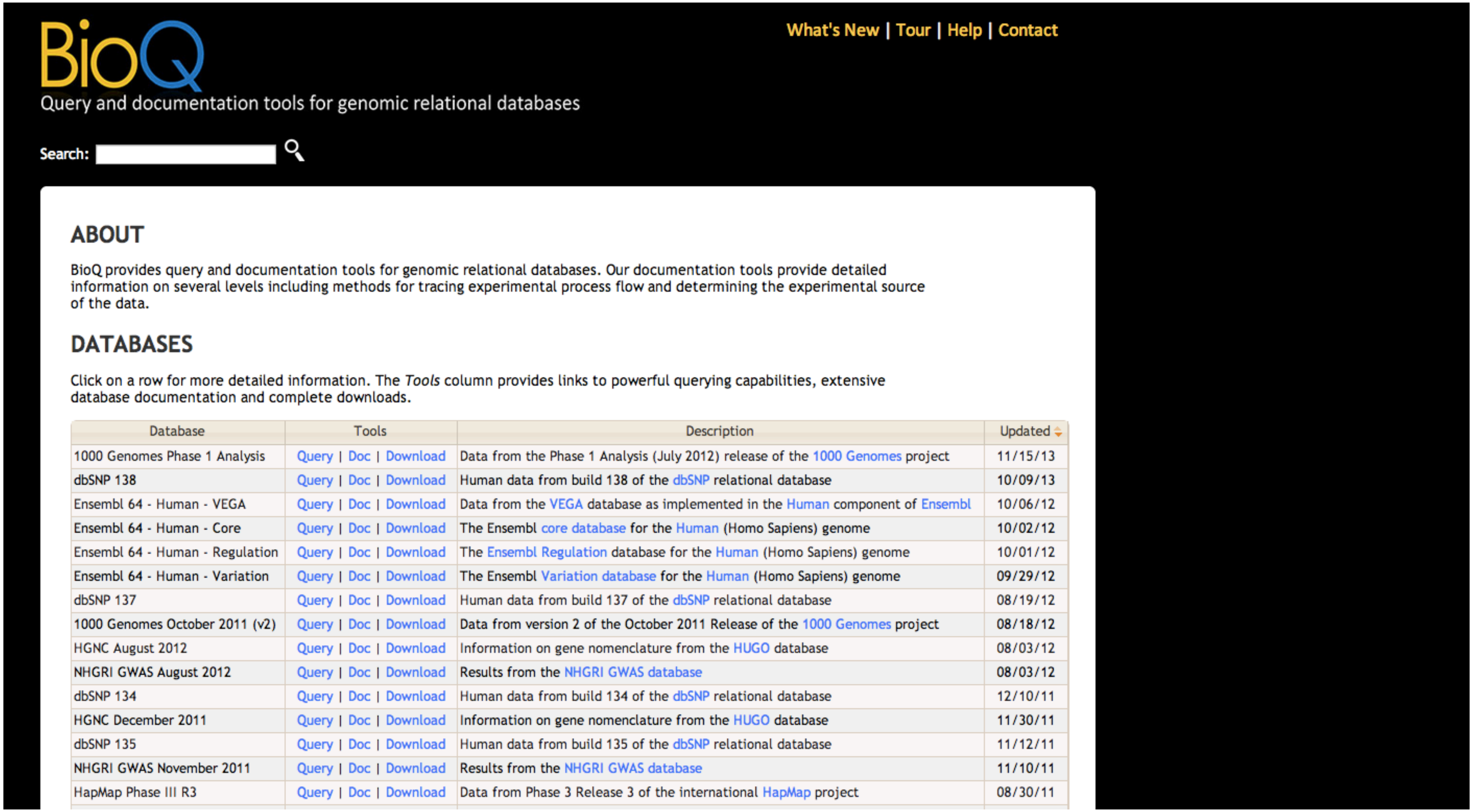
Reports & Analysis: Annotate LD Plot

Scroll/Zoom: Show 500 kbp ☐ **Flip**

What’s next?

- more samples
- more variants
- more data

BioQ (http://bioq.saclab.net/)



Query and documentation tools for genomic relational databases

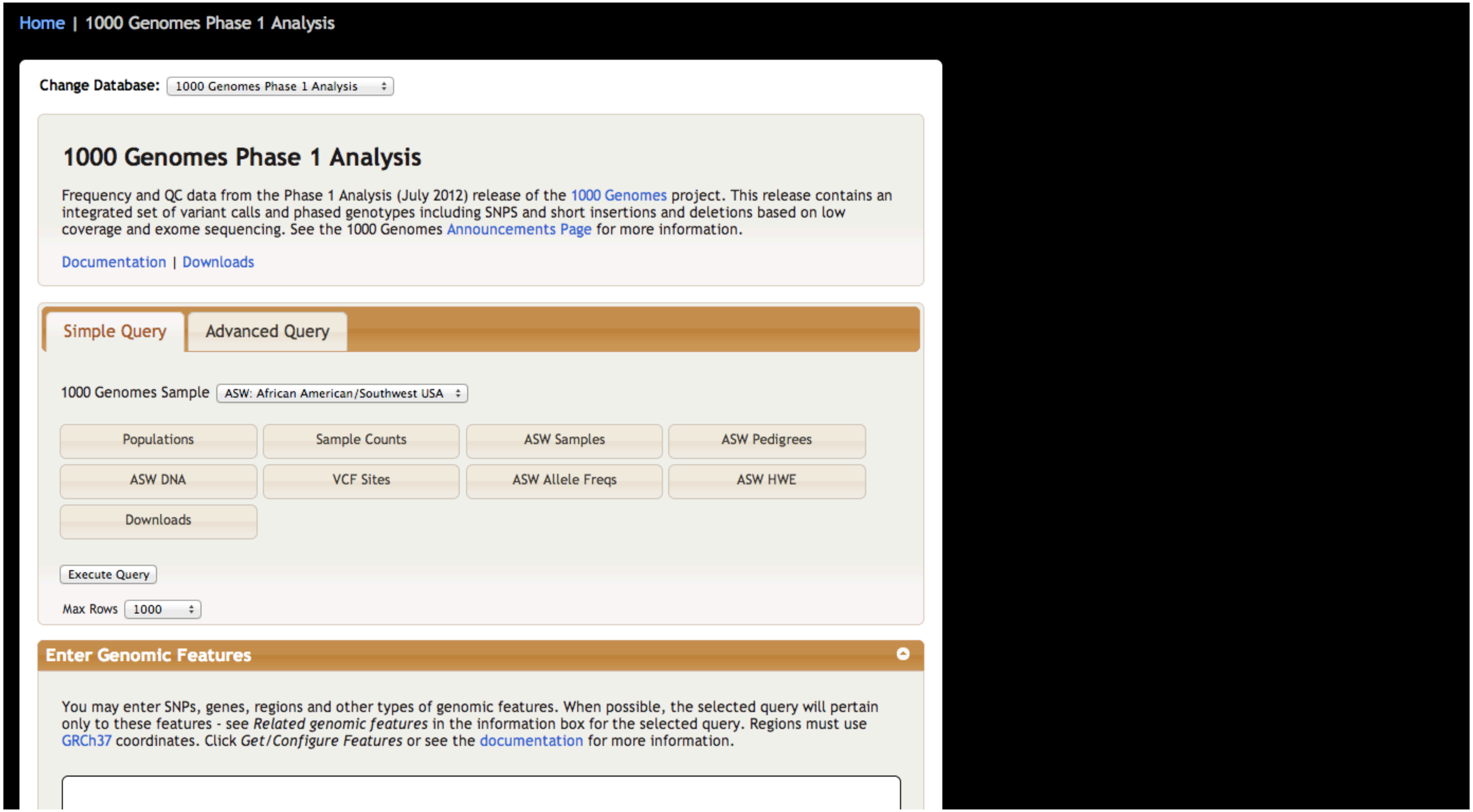
ABOUT

BioQ provides query and documentation tools for genomic relational databases. Our documentation tools provide detailed information on several levels including methods for tracing experimental process flow and determining the experimental source of the data.

DATABASES

Click on a row for more detailed information. The *Tools* column provides links to powerful querying capabilities, extensive database documentation and complete downloads.

Database	Tools	Description	Updated
1000 Genomes Phase 1 Analysis	Query | Doc | Download	Data from the Phase 1 Analysis (July 2012) release of the 1000 Genomes project	11/15/13
dbSNP 138	Query | Doc | Download	Human data from build 138 of the dbSNP relational database	10/09/13
Ensembl 64 - Human - VEGA	Query | Doc | Download	Data from the VEGA database as implemented in the Human component of Ensembl	10/06/12
Ensembl 64 - Human - Core	Query | Doc | Download	The Ensembl core database for the Human (Homo Sapiens) genome	10/02/12
Ensembl 64 - Human - Regulation	Query | Doc | Download	The Ensembl Regulation database for the Human (Homo Sapiens) genome	10/01/12
Ensembl 64 - Human - Variation	Query | Doc | Download	The Ensembl Variation database for the Human (Homo Sapiens) genome	09/29/12
dbSNP 137	Query | Doc | Download	Human data from build 137 of the dbSNP relational database	08/19/12
1000 Genomes October 2011 (v2)	Query | Doc | Download	Data from version 2 of the October 2011 Release of the 1000 Genomes project	08/18/12
HGNC August 2012	Query | Doc | Download	Information on gene nomenclature from the HUGO database	08/03/12
NHGRI GWAS August 2012	Query | Doc | Download	Results from the NHGRI GWAS database	08/03/12
dbSNP 134	Query | Doc | Download	Human data from build 134 of the dbSNP relational database	12/10/11
HGNC December 2011	Query | Doc | Download	Information on gene nomenclature from the HUGO database	11/30/11
dbSNP 135	Query | Doc | Download	Human data from build 135 of the dbSNP relational database	11/12/11
NHGRI GWAS November 2011	Query | Doc | Download	Results from the NHGRI GWAS database	11/10/11
HapMap Phase III R3	Query | Doc | Download	Data from Phase 3 Release 3 of the international HapMap project	08/30/11



Home | 1000 Genomes Phase 1 Analysis

Change Database: 1000 Genomes Phase 1 Analysis

1000 Genomes Phase 1 Analysis

Frequency and QC data from the Phase 1 Analysis (July 2012) release of the 1000 Genomes project. This release contains an integrated set of variant calls and phased genotypes including SNPS and short insertions and deletions based on low coverage and exome sequencing. See the 1000 Genomes Announcements Page for more information.

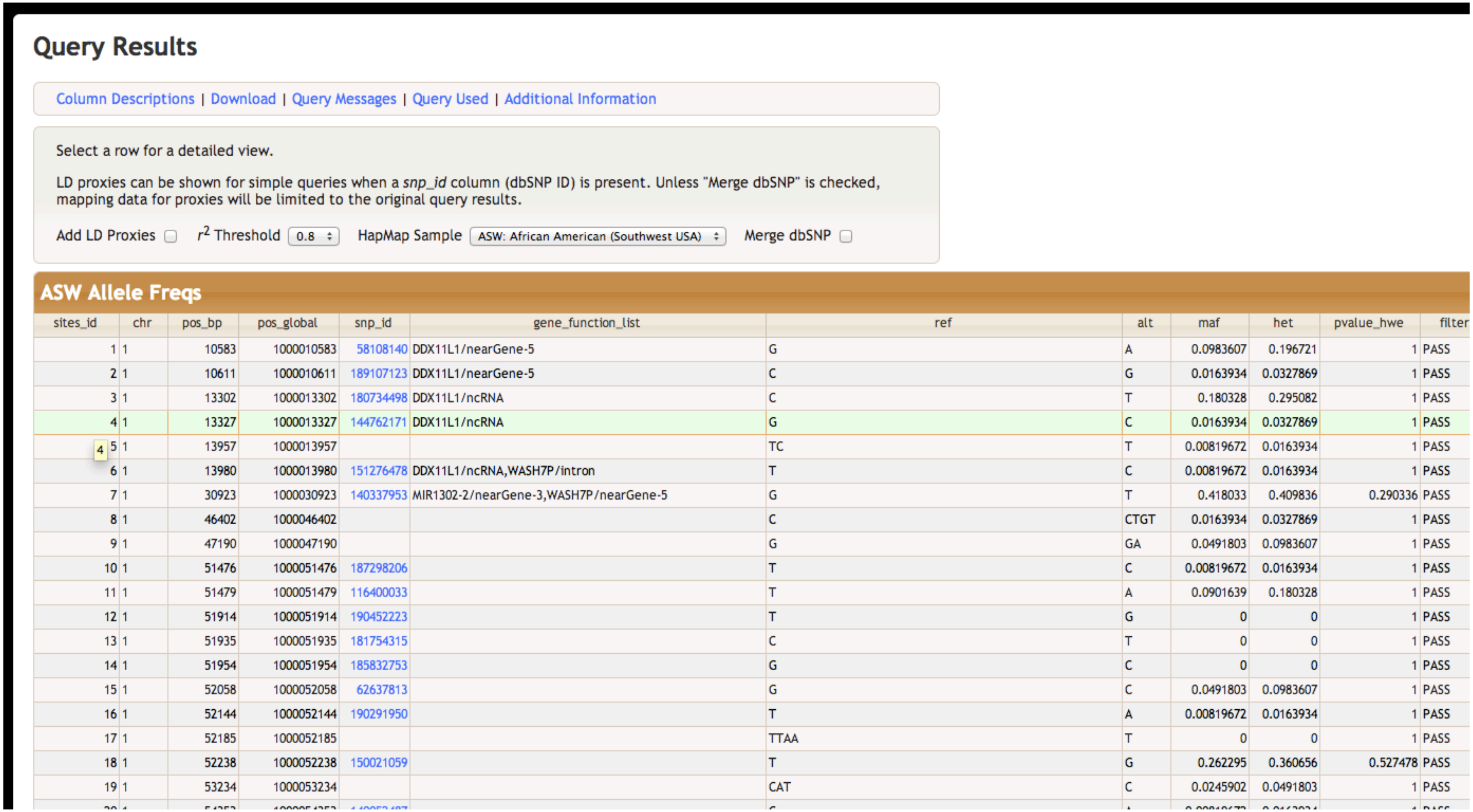
Documentation | Downloads

1000 Genomes Sample: ASW: African American/Southwest USA

Max Rows: 1000

Enter Genomic Features

You may enter SNPs, genes, regions and other types of genomic features. When possible, the selected query will pertain only to these features - see *Related genomic features* in the information box for the selected query. Regions must use GRCh37 coordinates. Click *Get/Configure Features* or see the documentation for more information.



Query Results

Column Descriptions | Download | Query Messages | Query Used | Additional Information

Select a row for a detailed view.

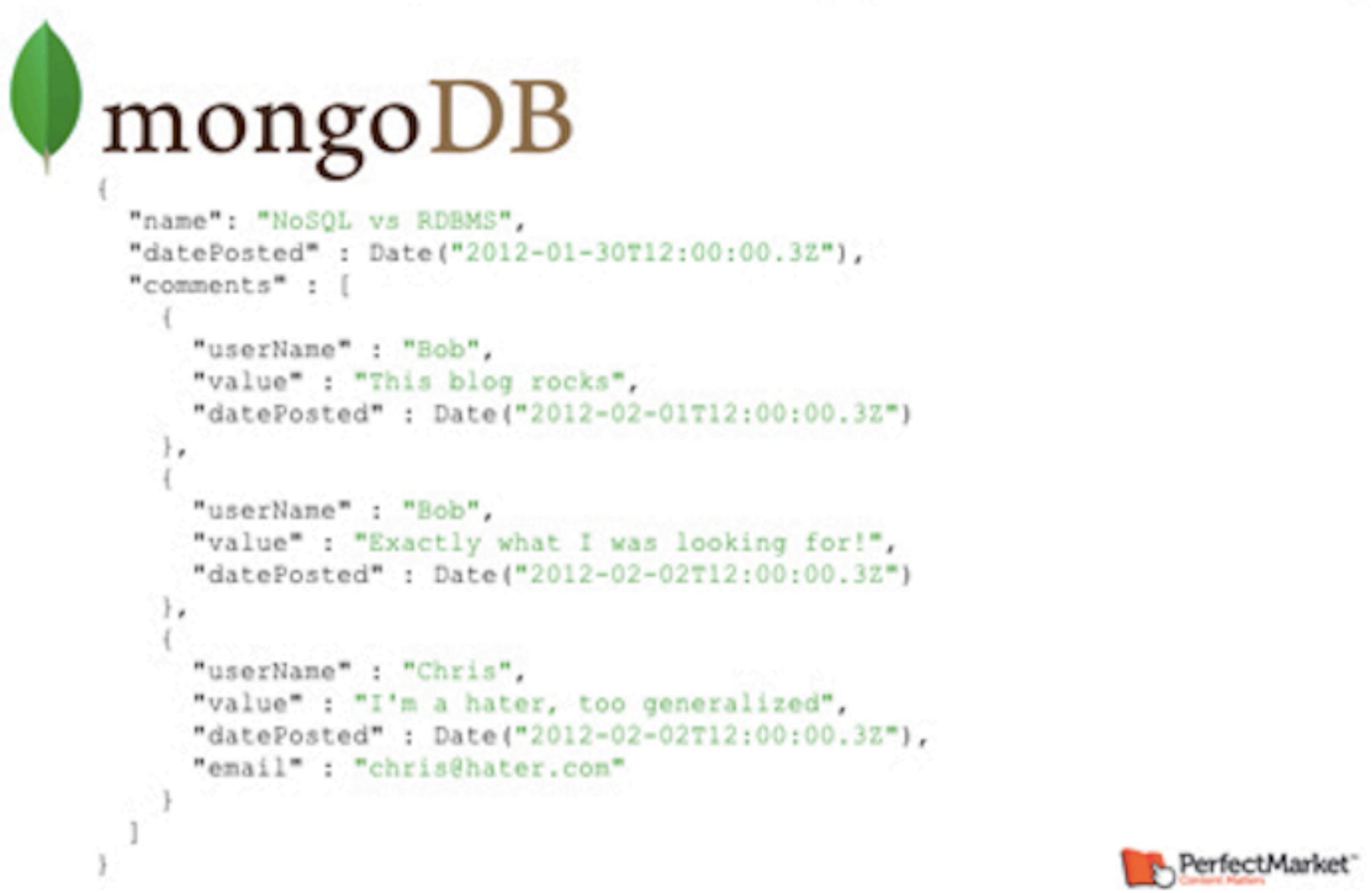
LD proxies can be shown for simple queries when a *snp_id* column (dbSNP ID) is present. Unless "Merge dbSNP" is checked, mapping data for proxies will be limited to the original query results.

Add LD Proxies □ r^2 Threshold 0.8 HapMap Sample ASW: African American (Southwest USA) Merge dbSNP □

ASW Allele Freqs

sites_id	chr	pos_bp	pos_global	snp_id	gene_function_list	ref	alt	maf	het	pvalue_hwe	filter
1	1	10583	1000010583	58108140	DDX11L1/nearGene-5	G	A	0.0983607	0.196721	1	PASS
2	1	10611	1000010611	189107123	DDX11L1/nearGene-5	C	G	0.0163934	0.0327869	1	PASS
3	1	13302	1000013302	180734498	DDX11L1/ncRNA	C	T	0.180328	0.295082	1	PASS
4	1	13327	1000013327	144762171	DDX11L1/ncRNA	G	C	0.0163934	0.0327869	1	PASS
5	1	13957	1000013957			TC	T	0.00819672	0.0163934	1	PASS
6	1	13980	1000013980	151276478	DDX11L1/ncRNA,WASH7P/intron	T	C	0.00819672	0.0163934	1	PASS
7	1	30923	1000030923	140337953	MIR1302-2/nearGene-3,WASH7P/nearGene-5	G	T	0.418033	0.409836	0.290336	PASS
8	1	46402	1000046402			C	CTGT	0.0163934	0.0327869	1	PASS
9	1	47190	1000047190			G	GA	0.0491803	0.0983607	1	PASS
10	1	51476	1000051476	187298206		T	C	0.00819672	0.0163934	1	PASS
11	1	51479	1000051479	116400033		T	A	0.0901639	0.180328	1	PASS
12	1	51914	1000051914	190452223		T	G	0	0	1	PASS
13	1	51935	1000051935	181754315		C	T	0	0	1	PASS
14	1	51954	1000051954	185832753		G	C	0	0	1	PASS
15	1	52058	1000052058	62637813		G	C	0.0491803	0.0983607	1	PASS
16	1	52144	1000052144	190291950		T	A	0.00819672	0.0163934	1	PASS
17	1	52185	1000052185			TTAA	T	0	0	1	PASS
18	1	52238	1000052238	150021059		T	G	0.262295	0.360656	0.527478	PASS
19	1	53234	1000053234			CAT	C	0.0245902	0.0491803	1	PASS

NoSQL based storage


mongoDB
PerfectMarket

```
{
  "name": "NoSQL vs RDBMS",
  "datePosted" : Date("2012-01-30T12:00:00.3Z"),
  "comments" : [
    {
      "userName" : "Bob",
      "value" : "This blog rocks",
      "datePosted" : Date("2012-02-01T12:00:00.3Z")
    },
    {
      "userName" : "Bob",
      "value" : "Exactly what I was looking for!",
      "datePosted" : Date("2012-02-02T12:00:00.3Z")
    },
    {
      "userName" : "Chris",
      "value" : "I'm a hater, too generalized",
      "datePosted" : Date("2012-02-02T12:00:00.3Z"),
      "email" : "chris@hater.com"
    }
  ]
}
```

Advice

- Know the datasets
- Understand what is going on behind the scenes
  - – database queries
  - – where the annotations are from
  - – how measures were calculated
- Make it reproducable