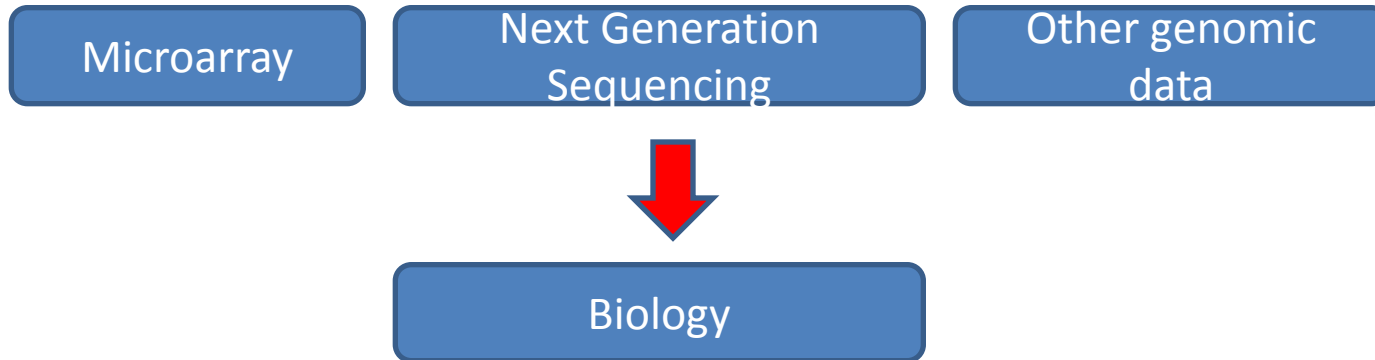




Juggling Genome Coordinates

Bingbing Yuan

April 14, 2011



Sample data and question:

Chromosome	start	end	peak	value
chr1	3521606	3522356	MACS_peak_1	398.3
chr1	3660375	3662829	MACS_peak_2	3100
chr1	4481520	4484198	MACS_peak_3	3100
chr1	4486231	4488053	MACS_peak_4	719.2

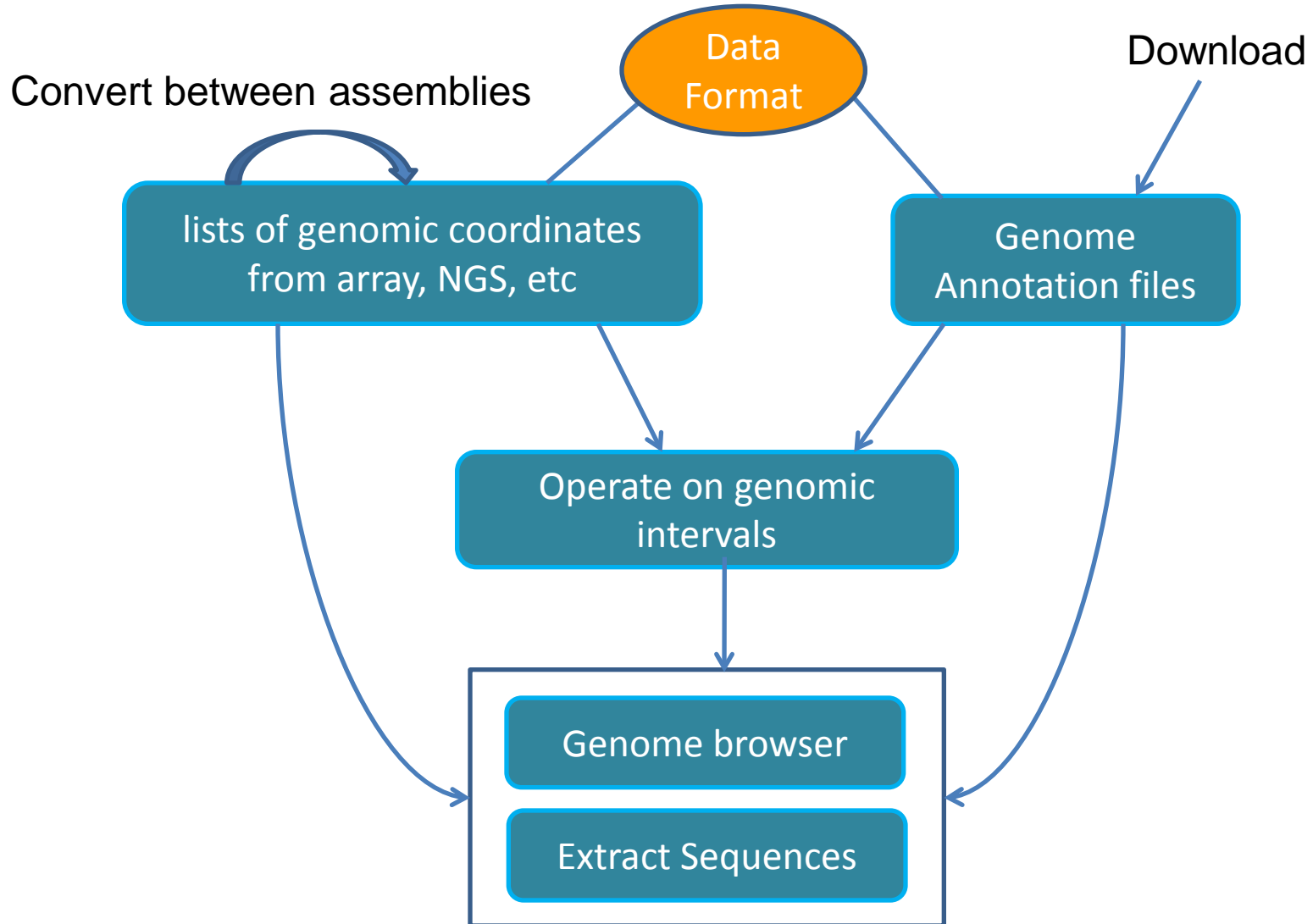
Where are these peaks found relative to genomic features?

Tools



- UCSC table and genome browser
 - <http://genome.ucsc.edu/>
 - Local Mirror: <http://membrane.wi.mit.edu>
- BioMart
 - <http://www.ensembl.org/biomart>
- IGV
 - <http://www.broadinstitute.org/software/igv/>
- Galaxy
 - <http://main.g2.bx.psu.edu/>
 - Previous Hot Topics:
http://iona.wi.mit.edu/bio/education/hot_topics/galaxy/Galaxy.pdf
- BedTools: (Installed on tak)
 - <http://code.google.com/p/bedtools/>
- Samtools: (Installed on tak)
 - <http://samtools.sourceforge.net/samtools.shtml>

Work Flow



Bed

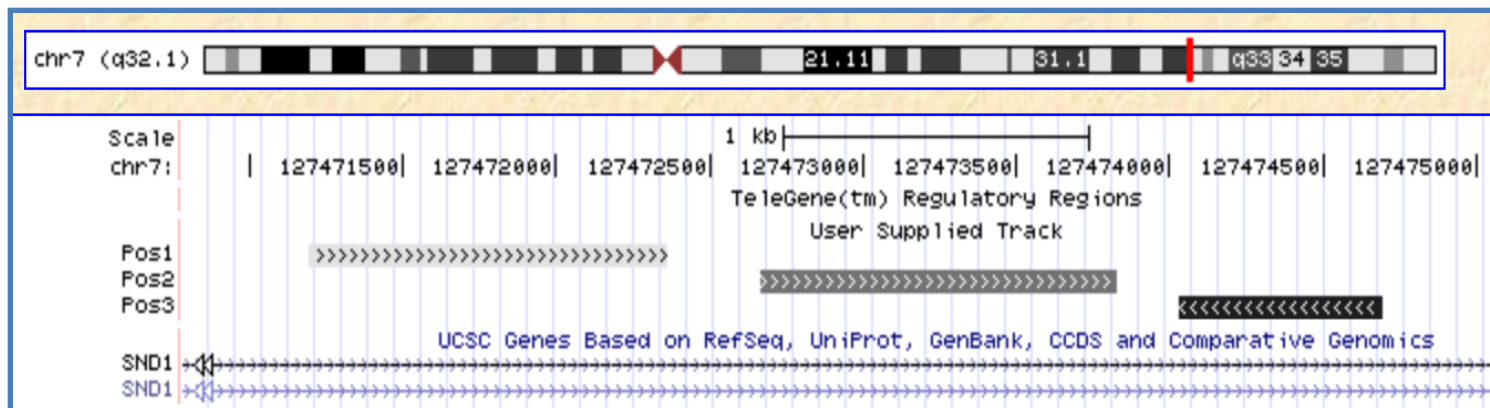


- Created by UCSC team
- The first 100 bases of a chromosome are defined as chromStart=0, chromEnd=100
- First 3 columns are required

Track type=bed useScore=1

Score: 0-1000

```
chr7      127471196 127472363 Pos1 1      +
chr7      127472663 127473830 Pos2 500   +
chr7      127474030 127474697 Pos3 900   -
```



More complex formats



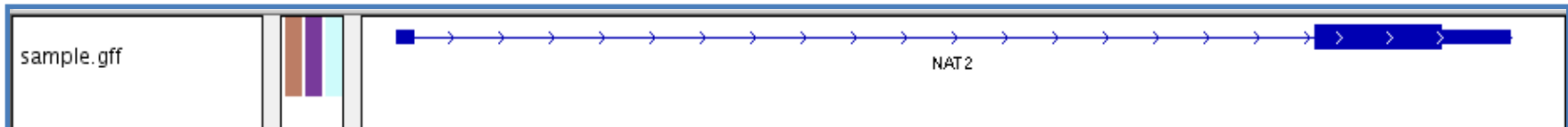
- GFF (General Feature Format)

- Must be tab-separated

		Feature	Start	End	Score	Strand (+/-/.)	Frame (0/1/2/.)	Group
chr8	mm9_refGene	exon	70018847	70018972	0	+	.	NAT2
chr8	mm9_refGene	start_codon	70025139	70025141	0	+	.	NAT2
chr8	mm9_refGene	CDS	70025139	70026008	0	+	0	NAT2
chr8	mm9_refGene	stop_codon	70026009	70026011	0	+	.	NAT2
chr8	mm9_refGene	exon	70025133	70026477	0	+	.	NAT2

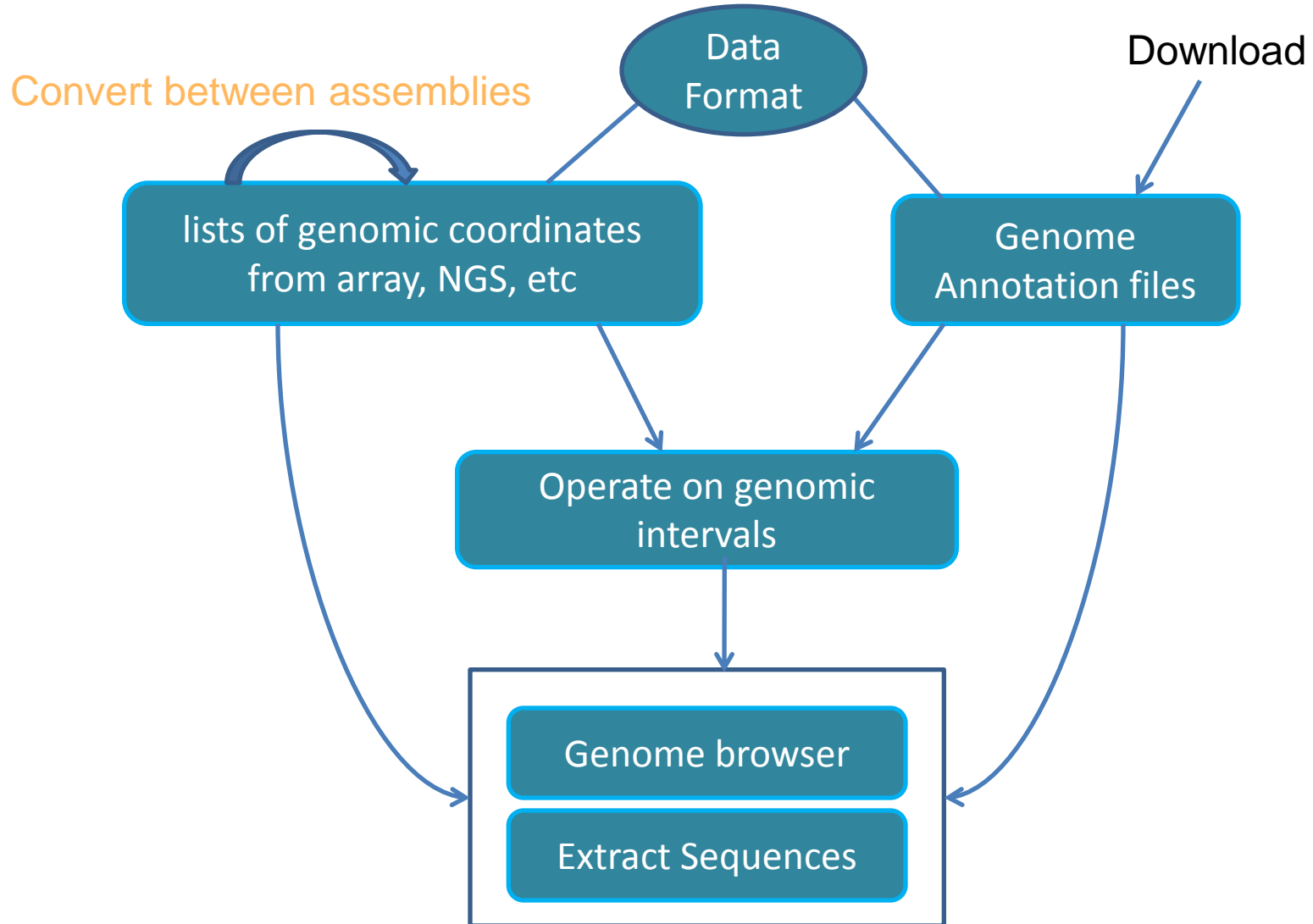
- GTF (Gene Transfer Format)

- `gene_id "SGIP1"; transcript_id "NM_032291";`



Commonly used .gtf files are in `/nfs/genome/genomeBuild/gtf`, eg:
`/nfs/genomes/human_gp_feb_09/gtf/hg19.refgene.gtf`

Work Flow



Chromosome nomenclature



- Assembled chromosomes: chr1, chr2 ...
- chr*_random: unplaced sequence on those reference chromosomes
- chrUn_* : unlocalized sequences where the corresponding reference chromosome has not been determined.
- haplotype chromosomes: chr6_cox_hap2.fa

LiftOver



- UCSC: Utilities ->liftOver

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help

Lift Genome Annotations

This tool converts genome coordinates and genome annotation files between assemblies. The input data can be pasted into the text box, or uploaded from a file. If a pair of assemblies cannot be selected from the pull-down menus, a direct lift between them is unavailable. However, a sequential lift may be possible. Example: lift from Mouse, May 2004, to Mouse, Feb. 2006, and then from Mouse, Feb. 2006 to Mouse, July 2007 to achieve a lift from mm5 to mm9.

Original Genome: Original Assembly: New Genome: New Assembly:

Minimum ratio of bases that must remap:
Minimum chain size in target:
Minimum hit size in query:
Allow multiple output regions:
Min ratio of alignment blocks/exons that must map:
If thickStart/thickEnd is not mapped, use the closest mapped base:

For descriptions of the supported data formats, see the bottom of this page.
Data Format:

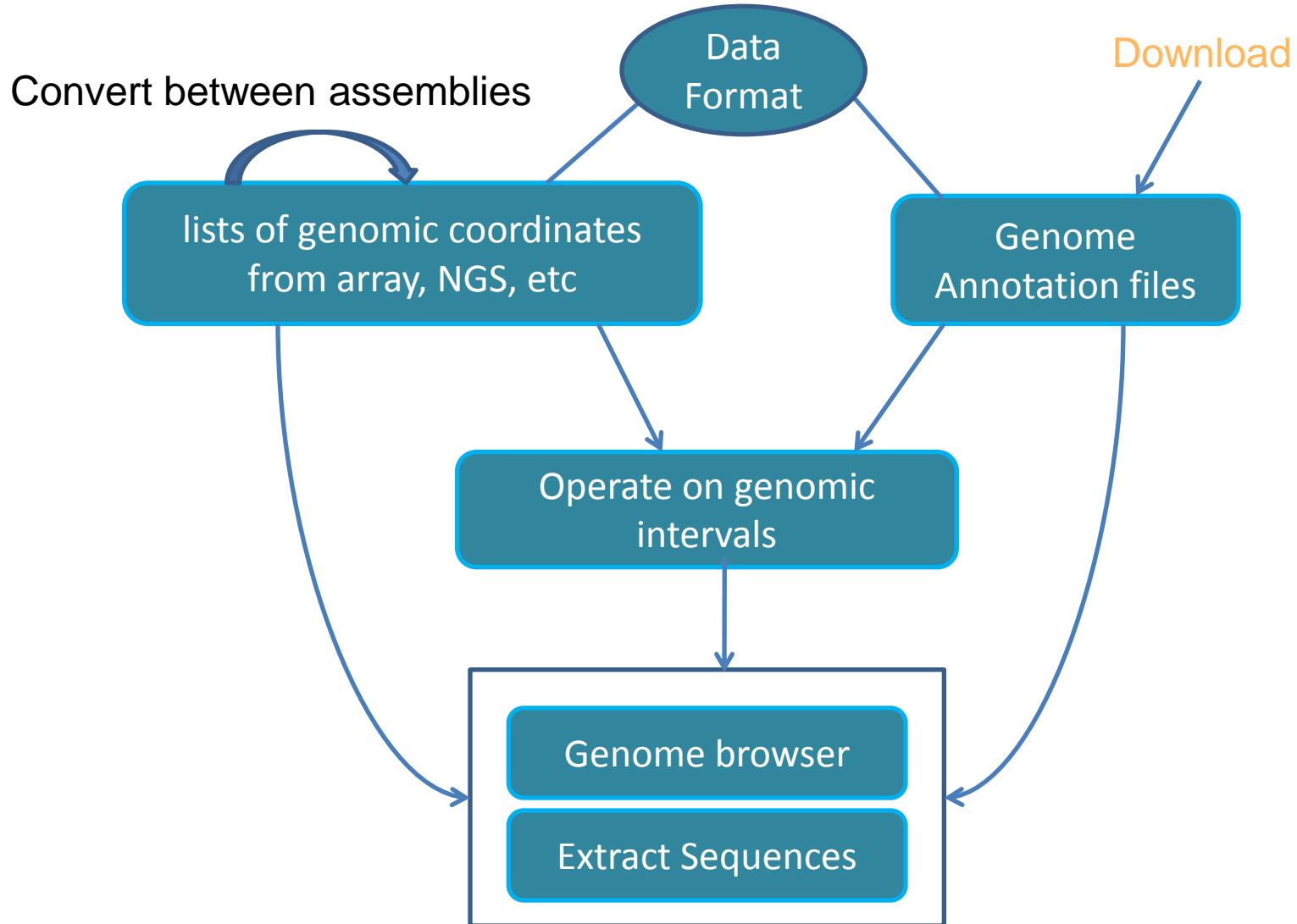
Paste in data:

```
chr1 4481008 4486494 A
chr1 4763278 4775807 B
chr1 4797973 4836816 C
chr1 4847774 4887987 D
```

Or upload data from a file:

Tak: \$ liftOver foo.bed mm8ToMm9.over.chain foo.mm9.bed foo.NOTmm9.bed

Work Flow



UCSC Table Browser

(Local mirror: <http://membrane.wi.mit.edu>)



[Home](#) [Genomes](#) [Genome Browser](#) [Blat](#) [Tables](#) [Gene Sorter](#) [PCR](#) [Session](#) [FAQ](#) [Help](#)

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.

clade: **genome:** **assembly:**

group: **track:**

table:

region: genome position

identifiers (names/accessions):

filter:

intersection:

correlation:

output format: Send output to [Galaxy](#) [GREAT](#)

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

UCSC Table Browser

(Local mirror: <http://membrane.wi.mit.edu>)



Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Output refGene as BED

Include [custom track](#) header:

name=

description=

visibility= ▾

url=

Create one BED record per:

- Whole Gene
- Upstream by bases
- Exons plus bases at each end
- Introns plus bases at each end
- 5' UTR Exons
- Coding Exons
- 3' UTR Exons
- Downstream by bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

BioMart



New Count Results URL XML Perl Help

Export all results to TSV Unique results only

Email notification to

View rows as Unique results only


Chromosome Name	Transcript Start (bp)	Transcript End (bp)	Ensembl Transcript ID
17	74730199	74733413	ENST00000508921
17	75464643	75468852	ENST00000508979
17	75543023	75559325	ENST00000510620
17	75554224	75559074	ENST00000510484
17	75718954	75724641	ENST00000504504
17	77889984	77900524	ENST00000507040
17	78313698	79329659	ENST00000505044
17	78775440	78779420	ENST00000501711
17	79604197	79606203	ENST00000499078
17	79885705	79888628	ENST00000500627

Dataset: Homo sapiens genes (GRCh37.p2)

Filters: [None selected]

Attributes: Chromosome Name, Transcript Start (bp), Transcript End (bp), Ensembl Transcript ID

Dataset: [None Selected]



Our previous hot topics on BioMart:

http://iona.wi.mit.edu/bio/education/hot_topics/galaxy/Galaxy.pdf

Customizing Genome Features



Galaxy Analyze Data Workflow Shared Data Help User

Tools Options ▾

Get Genomic Scores

Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- Join intervals of two datasets side-by-side
- Get flanks** returns flanking region/s for every gene
- Fetch closest non-overlapping feature for every interval
- Profile Annotations for a set of genomic intervals

Get flanks

Select data: 30: peak2.bed

Region: Around Start

Location of the flanking region/s: Downstream

Offset: 0

Length of the flanking region(s): 200

Execute

This tool finds the upstream and/or downstream flanking region(s) of all the selected regions in the input file.

Note: Every line should contain at least 3 columns: Chromosome number, Start and Stop co-ordinates. If any of these columns is missing or if start and stop co-ordinates are not numerical, the tool may encounter exceptions and such lines are skipped as invalid. The number of invalid skipped lines is documented in the resulting history item as a "Data issue".

History Options ▾

32: Get flanks on data 30

2 regions
format: interval, database: hg19
Info: Location: Downstream, Region: start, Flank-length: 200, Offset: 0
view in GeneTrack
display at Ensembl Current

1.Chrom	2.Start	3.End	4.Name	5	6.Strand
chr22	600	800	NM_028946	0	-
chr22	1000	1200	NM_174568	0	+

30: peak2.bed

2 regions
format: bed, database: hg19
Info: uploaded bed file
view in GeneTrack
display at Ensembl Current

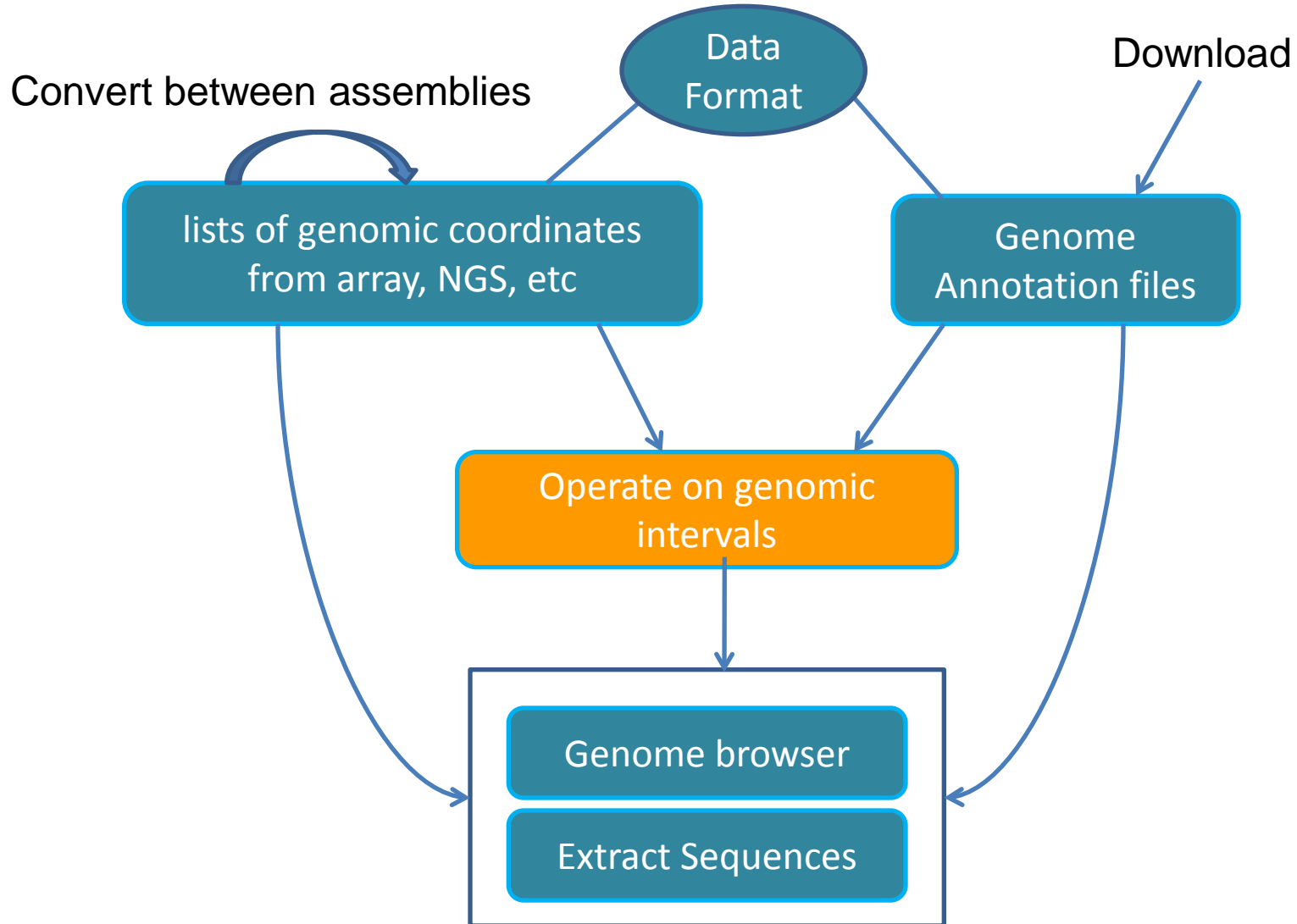
1.Chrom	2.Start	3.End	4.Name	5	6.Strand
chr22	500	800	NM_028946	0	-
chr22	1000	7000	NM_174568	0	+

•Excel

•Bedtools:

Tak: \$ slopBed -i foo.bed -g hg19.genome -l 0 -r 200 -s > out.bed

Work Flow



Galaxy

Tools Options ▾

- [Get Data](#)
- [Send Data](#)
- [ENCODE Tools](#)
- [Lift-Over](#)
- [Text Manipulation](#)
- [Convert Formats](#)
- [FASTA manipulation](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Fetch Alignments](#)
- [Get Genomic Scores](#)
- [Operate on Genomic Intervals](#)
- [Statistics](#)
- [Graph/Display Data](#)
- [Regional Variation](#)
- [Multiple regression](#)
- [Multivariate Analysis](#)
- [Evolution](#)
- [Motif Tools](#)
- [Multiple Alignments](#)
- [Metagenomic analyses](#)
- [Human Genome Variation](#)
- [EMBOSS](#)

Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the intervals of a dataset
- Join the intervals of two datasets side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest non-overlapping feature for every interval
- Profile Annotations for a set of genomic intervals





Annotate genomic coordinates



Join

Join:
3: exp.peak.bed
First dataset

with:
9: UCSC Main on Mous..ne (genome)
Second dataset

with min overlap:
1
(bp)

Return:
Only records that are joined (INNER JOIN)

Execute

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!
See Galaxy Interval Operation [Screencasts](#) (right click to open this link in another window).

Syntax

- **Where overlap** specifies the minimum overlap between intervals that

History Options

10: Join on data 9 and data 3

9: UCSC Main on Mouse: refGene (genome)

6: Base Coverage on data 5

5: Subtract on data 3 and data 4

4: Remove beginning on data 2

3: exp.peak.bed
30,540 regions
format: interval, database: mm9
Info: uploaded interval file
display at UCSC [main](#) view in [GeneTrack](#)
display at Ensembl [Current](#)

1. Chrom	2. Start	3. End	4	5
chr1	3521606	3522356	MACS_peak_1	398.28
chr1	3660375	3662829	MACS_peak_2	3100.00
chr1	4481520	4484198	MACS_peak_3	3100.00
chr1	4486231	4488053	MACS_peak_4	719.23
chr1	4512877	4513242	MACS_peak_5	61.62
chr1	4561215	4562439	MACS_peak_6	861.39



chr1	3521606	3522356	MACS_peak_1	398.28	chr1	3204562	3661579	NM_001011874	0	-
chr1	3660375	3662829	MACS_peak_2	3100.00	chr1	3204562	3661579	NM_001011874	0	-
chr1	4481520	4484198	MACS_peak_3	3100.00	chr1	4481008	4486494	NM_011441	0	-
chr1	4486231	4488053	MACS_peak_4	719.23	chr1	4481008	4486494	NM_011441	0	-

Tak: \$ intersectBed -a A.bed -b B.bed -wa -wb

Find genes closed to peaks



1. Chrom	2. Start	3. End	4. Name	5	6. Strand	7	8	9	10
chr1	134212701	134230065	NM_028778	0	+	134212806	134228958	0	7
chr1	134212701	134230065	NM_001195025	0	+	134212806	134228958	0	8
chr1	33510655	33726603	NM_008922	0	-	33510930	33725856	0	14
chr1	58714963	58752833	NM_175370	0	-	58715267	58749257	0	15
chr1	8352741	9289811	NM_027671	0	-	8353555	8794024	0	21

Fetch closest non-overlapping feature

For every interval in:

Fetch closest feature(s) from:

Located:

3: exp.peak.bed
 30,540 regions
 format: interval, database: mm9
 Info: uploaded interval file

display at UCSC [main view](#) in [GeneTrack](#)
 display at Ensembl [Current](#)

1. Chrom	2. Start	3. End	4	5
chr1	3521606	3522356	MACS_peak_1	398.28
chr1	3660375	3662829	MACS_peak_2	3100.00
chr1	4481520	4484198	MACS_peak_3	3100.00
chr1	4486231	4488053	MACS_peak_4	719.23
chr1	4512877	4513242	MACS_peak_5	61.62
chr1	4561215	4562439	MACS_peak_6	861.39

chr1	3521606	3522356	MACS_peak_1	398.28	chr1	4280926	4399322	NM_001195662	0	-	4283061	4399268	0
chr1	3660375	3662829	MACS_peak_2	3100.00	chr1	4280926	4399322	NM_001195662	0	-	4283061	4399268	0
chr1	4481520	4484198	MACS_peak_3	3100.00	chr1	4280926	4399322	NM_001195662	0	-	4283061	4399268	0
chr1	4486231	4488053	MACS_peak_4	719.23	chr1	4280926	4399322	NM_001195662	0	-	4283061	4399268	0
chr1	4512877	4513242	MACS_peak_5	61.62	chr1	4481008	4486494	NM_011441	0	-	4481796	4483487	0

How many peaks overlap CpG islands?



1.Chrom	2.Start	3.End	4	5
chr1	3660375	3662829	MACS_peak_2	3100.00
chr1	4481520	4484198	MACS_peak_3	3100.00
chr1	4773811	4776506	MACS_peak_9	2950.10
chr1	4797191	4799453	MACS_peak_10	2506.20
chr1	4846355	4849267	MACS_peak_12	3100.00
chr1	5007623	5011557	MACS_peak_14	3100.00

?

1.Chrom	2.Start	3.End	4	5	6	7	8
chr1	3660375	3662829	MACS_peak_2	3100.00	cpgIslandExt	957	2
chr1	4481520	4484198	MACS_peak_3	3100.00	cpgIslandExt	1972	1
chr1	4773811	4776506	MACS_peak_9	2950.10	cpgIslandExt	438	1
chr1	4797191	4799453	MACS_peak_10	2506.20	cpgIslandExt	544	1
chr1	4846355	4849267	MACS_peak_12	3100.00	cpgIslandExt	907	1
chr1	5007623	5011557	MACS_peak_14	3100.00	cpgIslandExt	1154	1

Profile Annotations



Profile Annotations ←

Choose Intervals:
18: sample_peak.bed

Keep Region/Table Pairs with 0 Coverage:
Discard

Output per Region/Summary:
Per Region

Choose Tables to Use:

- [+] Comparative Genomics
- [+] Genes and Gene Prediction Tracks
- [+] Mapping and Sequencing Tracks
- [+] Phenotype and Allele
- [+] Expression and Regulation
- [+] mRNA and EST Tracks
- [+] Variation and Repeats
- [+] Uncategorized Tables

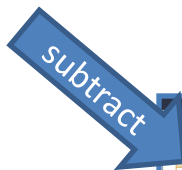
Selecting no tables will result in using all tables.

Execute

[+] Expression and Regulation

- ORegAnno
- NHGRI BiP
- Affy Exon Probes
- GNF Atlas 2
- GNF U74B
- GNF U74C
- GNF U74A
- Affy MOE430
- REST
- [+] agilentCgh
- Affy Exon Tissues
- Affy GNF1M
- Affy U74
- CpG Islands
- Allen Brain

Remove overlapping regions between two datasets



Subtract

Subtract:
3: exp.peak.bed
Second dataset

from:
4: Remove beginning on data 2
First dataset

Return:
Non-overlapping pieces of intervals
of the first dataset (see figure below)

where minimal overlap is:
1
(bp)

Execute

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!
See Galaxy Interval Operation [Screencasts](#) (right click to open this link in another window).

Syntax

- **Where overlap is at least** sets the minimum length (in base pairs) of overlap between elements of the two datasets.
- **Intervals with no overlap** returns entire intervals from the first dataset that do not overlap the second dataset. The returned intervals are completely unchanged, and this option only filters out intervals that overlap with the second dataset.
- **Non-overlapping pieces of intervals** returns intervals from the first dataset that have the intervals from the second dataset removed. Any overlapping base pairs are removed from the range of the interval. All fields besides start and end are guaranteed to remain unchanged.

Example

History

Options

- 10: Join on data 9 and data 3
- 9: UCSC Main on Mouse: refGene (genome)
- 6: Base Coverage on data 5
- 5: Subtract on data 3 and data 4
- 4: Remove beginning on data 2
22,437 regions
format: bed, database: mm9
display at UCSC [main view](#) in [GeneTrack](#)
display at Ensembl [Current](#)
- 3: exp.peak.bed
30,540 regions
format: interval, database: mm9
Info: uploaded interval file
display at UCSC [main view](#) in [GeneTrack](#)
display at Ensembl [Current](#)
- 2: exp2_peaks.bed

1. Chrom	2. Start	3. End	4. Name	5
chr1	3052582	3053252	MACS_peak_1	1073.97
chr1	3330773	3331061	MACS_peak_2	74.78
chr1	3333447	3334015	MACS_peak_3	149.54
chr1	3472706	3473645	MACS_peak_4	548.91
chr1	3638938	3639583	MACS_peak_5	274.11
chr1	3671336	3672045	MACS_peak_6	442.06

1. Chrom	2. Start	3. End	4	5
chr1	3521606	3522356	MACS_peak_1	398.28
chr1	3660375	3662829	MACS_peak_2	3100.00
chr1	4481520	4484198	MACS_peak_3	3100.00
chr1	4486231	4488053	MACS_peak_4	719.23
chr1	4512877	4513242	MACS_peak_5	61.62
chr1	4561215	4562439	MACS_peak_6	861.39

Tak: `$ subtractBed -a exp1.bed -b exp2.bed`

Calculating the depth and breadth of sequence coverage across defined "windows" in a genome



Coverage ←

What portion of:

4: chr16_1000_100.bed

First dataset

is covered by:

6: mapped_chr16.bed

Second dataset

Execute

History

6: mapped_chr16.bed
 ~20,000 regions
 format: bed, database: mm9
 Info: uploaded bed file
 display at UCSC [main](#)
 view in [GeneTrack](#)
 display at Ensembl [Current](#)

1.Chrom	2.Start	3.End
chr16	3001103	3001129
chr16	3001106	3001132
chr16	3001540	3001566
chr16	3001779	3001805
chr16	3002334	3002360
chr16	3003174	3003200

4: chr16_1000_100.bed
 ~110,000 regions
 format: bed, database: mm9
 Info: uploaded bed file
 display at UCSC [main](#)
 view in [GeneTrack](#)
 display at Ensembl [Current](#)

1.Chrom	2.Start	3.End
chr16	0	1000
chr16	900	1900
chr16	1800	2800
chr16	2700	3700
chr16	3600	4600

chr16	2997000	2998000	0	0.0
chr16	2997900	2998900	0	0.0
chr16	2998800	2999800	0	0.0
chr16	2999700	3000700	0	0.0
chr16	3000600	3001600	55	0.055
chr16	3001500	3002500	78	0.078
chr16	3002400	3003400	26	0.026
chr16	3003300	3004300	180	0.18
chr16	3004200	3005200	445	0.445
chr16	3005100	3006100	310	0.31

Tak: \$ coverageBed -a mapped_chr16.bed -b chr16_1000_100.bed

Feature Distribution



Redundant List Analysis

[Click here for directions](#)

INPUT OPTION #1: Paste your list

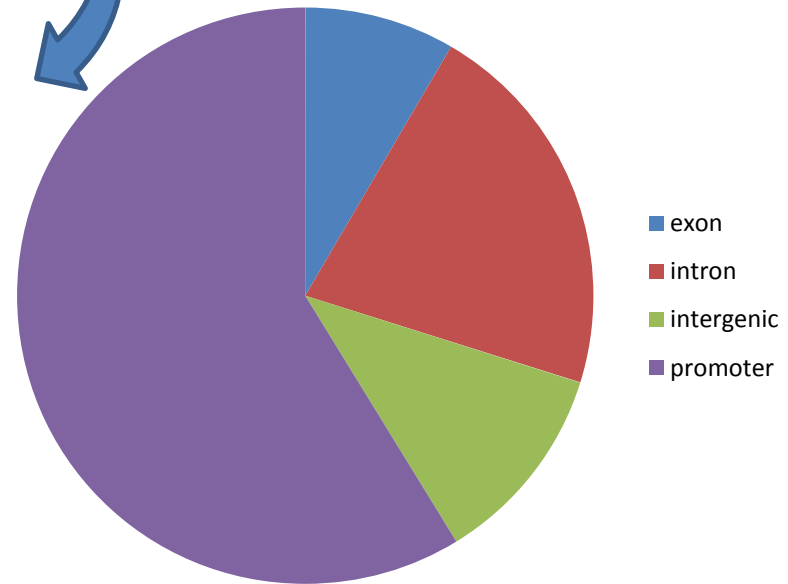
[Sample input](#) [Sample output](#)

```
exon
exon
intron
promoter
intron
intergenic
promoter
promoter
promoter
promoter
exon
exon
```

INPUT OPTION #2: Upload a file
(Note: Data should be in column format)

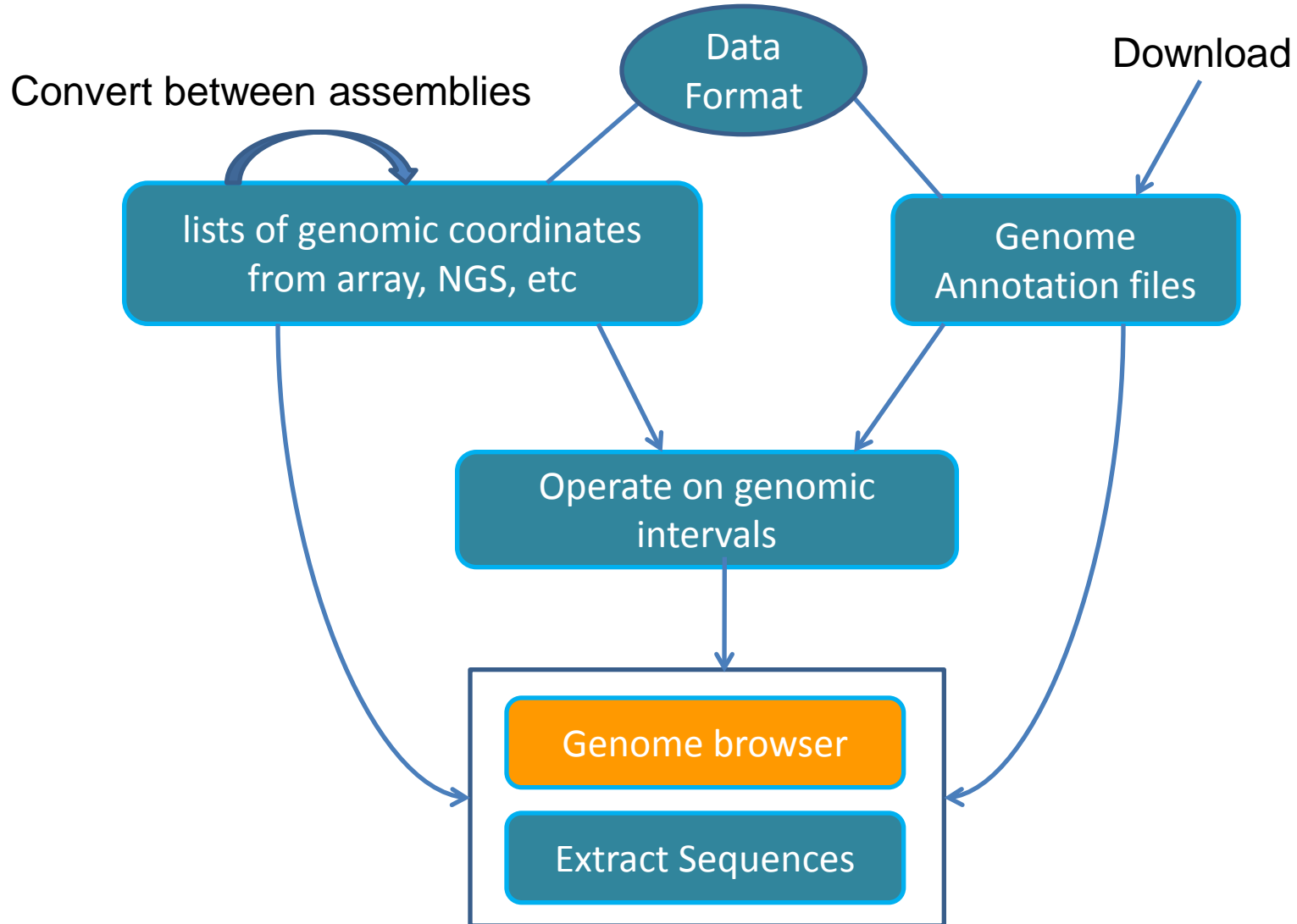
Click "Submit Data" below to analyze

exon	26
intron	66
intergenic	35
promoter	181



<http://iona.wi.mit.edu/cedrone/redundant/>

Work Flow



Create Genome Browser Tracks



- <http://genome.ucsc.edu/goldenPath/help/customTrack.html>
- Define the Genome Browser display characteristics
 - browser position chr22:1000-10000
- Define the annotation track display characteristics: type, name, description, color, etc.
 - track type=bedGraph name=myExp
- Format the data set: GFF, bedGraph, GTF, BED, WIG, etc

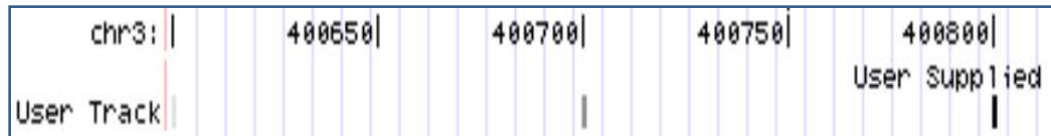
```
browser position chr19:59302001-59304701
track type=bedGraph name="Exp 1"
chr19 59302000 59302300 -1.0
chr19 59302300 59302600 -0.75
```

Wig (Wiggle)



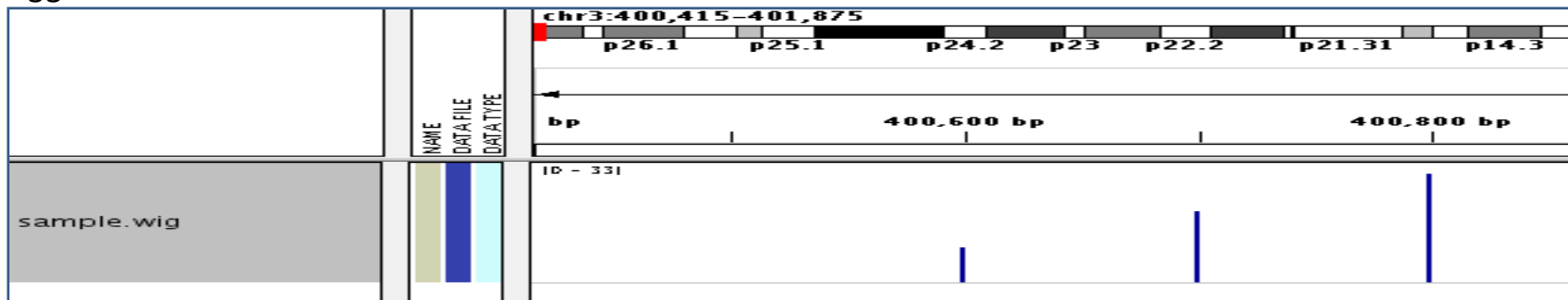
- Created by UCSC team
- Optimized for storing “levels”.

```
track type=wiggle_0  
variableStep chrom=chr2  
300701 12.5  
300702 12.5  
300703 12.5  
300704 12.5  
300705 12.5
```



```
track type=wiggle_0  
variableStep chrom=chr2 span=5  
300701 12.5
```

```
track type=wiggle_0  
fixedStep chrom=chr3 start=400601 step=100 span=1  
11  
22  
33
```



BedGraph



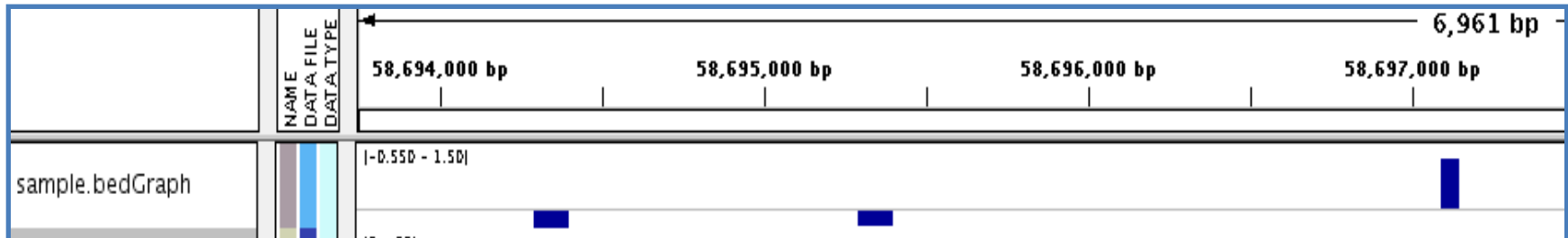
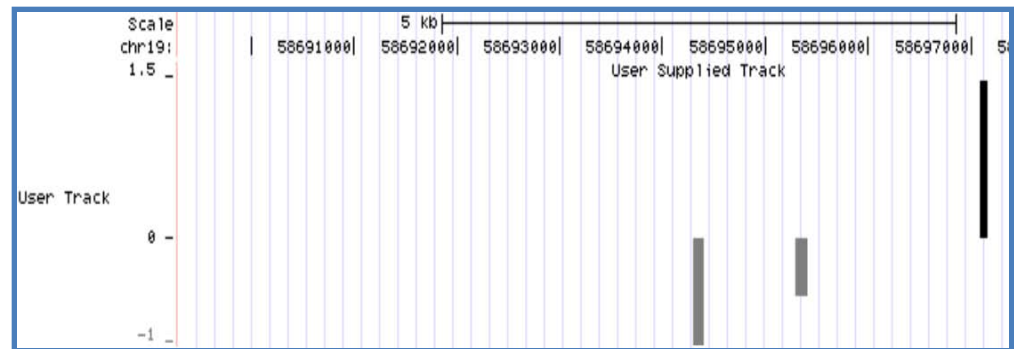
- The first chromosome position is 0. The last position in a chromosome of length N would be N – 1
- Display continuous data that is sparse or contains elements of varying size
- variableStep wig: chromStarts >100 bases apart

track type=bedGraph

chr19 58694300 58694400 -1.0

chr19 58695300 58695400 -0.55

chr19 58697100 58697150 1.50

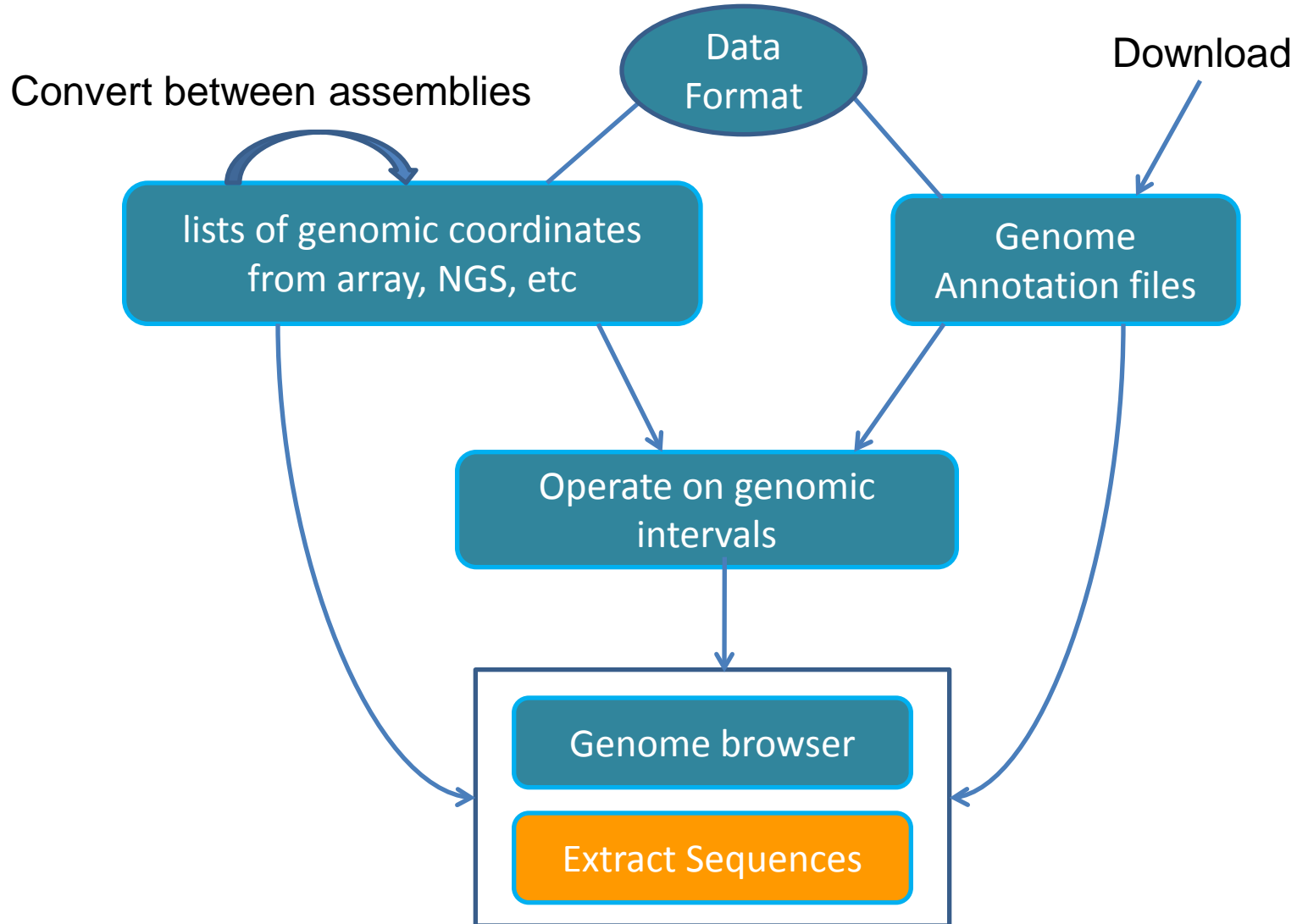


Comparison of formats



- **Bed Graph:**
 - Best used for genome-wide data sets on the order of several million to perhaps 10 million positions
 - Best used when data is not spaced at regular intervals, and the size of the specified regions is not a constant
- **Wig:**
 - Best used for genome-wide data sets on the order of several 10's of million data points
 - Specified regions must be a constant size (specified by the span argument)
- **Large data:**
 - Compressed format: gzip
 - Binary format: bigBed, bigWig
 - IGV, WI UCSC genome mirror

Work Flow



Extract sequences from UCSC



Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

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.

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

group: Genes and Gene Prediction Tracks track: RefSeq Genes [manage custom tracks](#)

table: refGene

region: genome position chr21:33

identifiers (names/accessions): [paste list](#)

filter: [create](#)

intersection: [create](#)

correlation: [create](#)

output format: sequence

output file:

file type returned: plain text gzip

[get output](#)

[summary/statistics](#)

To reset **all** user cart settings (including cu

Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

RefSeq Genes Genomic Sequence

Sequence Retrieval Region Options:

- Promoter/Upstream by 1000 bases
- 5' UTR Exons
- CDS Exons
- 3' UTR Exons
- Introns
- Downstream by 1000 bases
- One FASTA record per gene.
- One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra downstream (3')
 - Split UTR and CDS parts of an exon into separate FASTA records

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Sequence Formatting Options:

- Exons in upper case, everything else in lower case.
- CDS in upper case, UTR in lower case.
- All upper case.
- All lower case.
- Mask repeats: to lower case to N

[get sequence](#)

[cancel](#)

BioMart



e!Ensembl east | BLAST/BLAT | BioMart | Tools | Downloads | More ▾

New | **Count** | **Results** | **★ URL** | **XML** | **Perl** | **Help**

Dataset
Homo sapiens genes (GRCh37.p2)

Filters
Chromosome: 19

Attributes
Ensembl Gene ID
Ensembl Transcript ID
Upstream flank [5000]
Flank (Transcript)

Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

- Features
- Structures
- Transcript Event
- Homologs
- Variation
- Sequences

SEQUENCES:

Sequences (max 1)

- Unspliced (Transcript)
- Unspliced (Gene)
- Flank (Transcript)
- Flank (Gene)
- Flank-coding region (Transcript)
- Flank-coding region (Gene)
- 5' UTR
- 3' UTR
- Exon sequences
- cDNA sequences
- Coding sequence
- Protein

Upstream flank
 Upstream flank 5000

Downstream flank
 Downstream flank

Header Information



Galaxy Analyze Data Workflow Shared Data Visualization Help User

Tools Options ▾

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
 - Extract Genomic DNA using coordinates from assembled/unassembled genomes
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics

Extract Genomic DNA

Fetch sequences for intervals in:
11: peak.txt

Interpret features when possible:
Yes ▾

Only meaningful for GFF, GTF datasets.

Source for Genomic Data:
Locally cached ▾

Output data type:
Interval ▾

Execute

⚠ This tool requires tabular formatted data. If your data is not TAB delimited, use *Text Manipulation* -> *Convert*.

History Options ▾

12: Extract Genomic DNA on data 11

2 regions
format: interval, database: mm9

display at UCSC [main](#)
view in [GeneTrack](#)
display at Ensembl [Current](#)

1. Chrom	2. Start	3. End	4. Name	5. Strand	6
chrM	1	10	forward.1	+	TTAATGTAG
chrM	1	10	reverse.1	-	CTACATTAA

11: peak.txt

2 regions
format: interval, database: mm9
Info: uploaded interval file

display at UCSC [main](#)
view in [GeneTrack](#)
display at Ensembl [Current](#)

1. Chrom	2. Start	3. End	4. Name	5. Strand
chrM	1	10	forward.1	+
chrM	1	10	reverse.1	-

Tak: `$ fastaFromBed -fi /nfs/genomes/human_gp_feb_09/fasta/chrM.fa -bed peak.bed -s -name -fo peak.fa`

Extract genomic sequences



Genomic sequence extractor at WIBR

Extract genomic sequence around any genomic landmark.

Paste in genomic location data in the following format:

ID <tab> chromosome <tab> start <tab> stop <tab> strand

```
myLandmark1 chr1 1008542 1008642 +
myLandmark2 chrX 1009301 1009401 +
```

Species: Human February 2009 (hg19; NCBI 37) ▾

Nt upstream of "start" coordinate:

Nt downstream of "stop" coordinate:

Sequence options:

- Repeats in lower case
- Mask repeats as Ns
- All uppercase

RefSeq/SMED promoter extractor at WIBR

Extract genomic sequence near transcription start and/or stop.

Paste in list of RefSeq IDs (one per line, ex: NM_002864):

```
NM_002864
NM_000015
```

Species: human mouse planarian

Upstream regulatory region:

NT upstream of transcription start:

NT downstream of transcription start:

Downstream regulatory region:

NT upstream of transcription stop:

NT downstream of transcription stop:

Sequence options:

- Repeats in lower case
- Mask repeats as Ns
- All uppercase

Description of the algorithm

http://iona.wi.mit.edu/bell/extract_custom.php

http://iona.wi.mit.edu/bell/refseq_extractor.php

Work Flow

