The Genomic HyperBrowser

Statistical genome analysis made transparent and accessible

Eivind Hovig and Geir Kjetil Sandve





Challenges and opportunities in the post genomic era

- I will next year generate more data than all up to now
- But asking the crucial questions is still difficult!
- Need analysis methodology to match the power of data generation systems

Or take the enormous increase in public data

GNF Ratio CpG Islands I TFBS Conserved GRD View ROD Human QTL SNPs (129) DNase GN659 Raw DNase Heg2 Raw DNase Heg2 Raw DNase Heg2 Pval DNase GP Val DNase Heg2 Pval	Ch	nr13 (q12.3) p13 p12 p11.2
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics With Hill Hill Hill Hill Hill Hill Hill Hil	chail 2 i	
GNF Ratio CpG Islands (Islands < 300 Bases are Light Green) TFBS Conserved GRD View ROD Human QTL SNFs (129) T-Hay Most Cons I DNase GM659 Raw DNase Heg2 Raw DNase Heg3 Pval	UCSC Genes	UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
GNF Ratio CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands < 300 Bases are Light Green CpG Islands (Islands < 300 Bases are Light Green CpG Islands (Islands < 300 Bases are Light Green CpG Islands (Islands < 300 Bases are Light Green CpG Islands (Islands < 100 Islands Islands - 100 I	Simple Repeats	
CpG Islands Image: CpG Islan	GNF Ratio	
GAD View Human Quantitative Trait Locus from RGD OSkeponitimitatis QiL 7 SNPS (129) SNPS (129) PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment DNase CM069 Raw DNase CHLa Raw DNase Hea Raw DNase Hea Raw DNase Hea Raw DNase Hea Raw DNase CHea Pval IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII		
Osteoarthritis QTL 7 SNPs (129) Simple Nucleotide Polymorphisms (dbSNP build 129) 17-Way Most Cons PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment 17-Way Most Cons PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment 17-Way Most Cons Image: State S	GAD View	
SNPs (129) PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment DNase GM069 Raw DNase GM069 Raw DNase HeLa Raw DNase HepG2 Raw DNase HepG2 Raw DNase K562 Raw DNase K562 Pval UIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	RGD Human QTL	Osteoarthritis QTL 7
17-Way Most Cons I	SNPs (129)	
FAIRE Signal FAIRE Signal FAIRE PeakFinder FAIRE ChIPOTIe BinCons Chai Cons TBA Align NHGRI DIPS	DNase GM069 Raw DNase CD4 Raw DNase HeLa Raw DNase HepG2 Raw DNase H9Raw DNase IMR90 Raw DNase GM069 Pva1 DNase CD4 Pva1 DNase HeLa Pva1 DNase HepG2 Pva1 DNase HepG2 Pva1 DNase H9P0 Pva1 DNase IMR90 Pva1 DNase K562 Pva1 FAIRE Signa1 FAIRE Signa1 FAIRE ChIPOT 1e BinCons Chai Cons TBA Align	TEA Alignments and Conservation of 36 Vertebrates in the ENCODE Regions

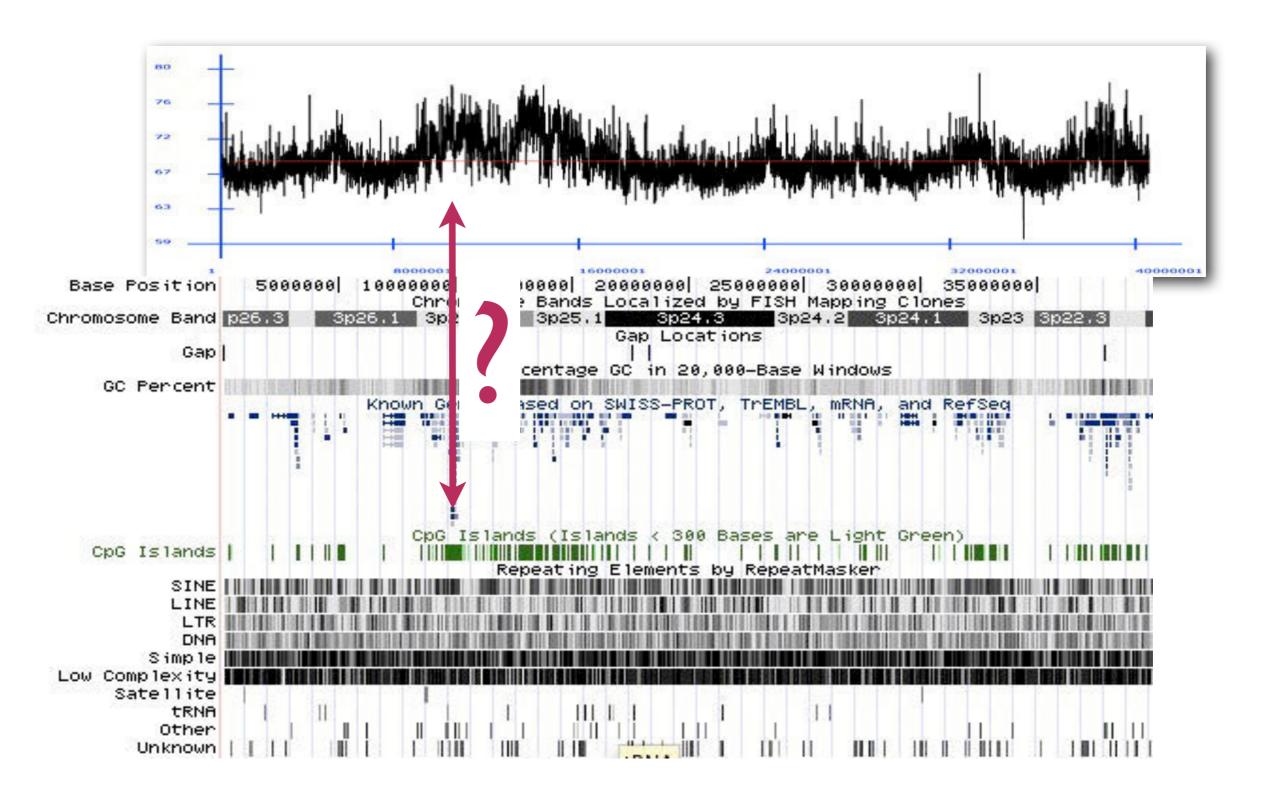
ENCODE, FANTOM, GEO, Roadmap Epigenomics ...

High melting temperature Low melting temperature Chr. 17: 93.25 Chr. 19: 47.75 77 71 87 65 81 74 5 68 53 61 47 78786525 78790525 78794525 78798525 78802525 78806525 47833453 47837453 47841453 47845453 47849453 47853453 Base Position 78798888 78795668 75566666 78865666 Chronosome Bands Localized by FISH Mapping Clones 47840000 47845888 47858688 Base Position Chromosome Band 17025.3 Chromosome Bands Localized by FISH Mapping Clones Gap Locat ions Chromosome Band 19913.2 Gap Gap Locations Fercentage GC in 20,000-Base Windows Gap GC Percent Percentage GC in 20,000-Base Windows known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq GC Percent Known Genes Based on SWISS-PROT, TrENBL, mRNA, and RefSed CpC Islands (Islands < 300 Bases are Light Green) CpG Islands (Islands 300 Bas are Light Green) CpG Islands CpG Islands Repeating Elements by RepeatMasker Repeating E peatMasker ments by SINE B 10. SINE LINE LINE LTR 10.8 LTR DNA 1 DNR Simple 101 Simple 100 Low Complexity Low Complexity Satellite Sate11ite TRNA tRNA Other Other Unknown Unknoun

CpG Islands

AT simple repeats

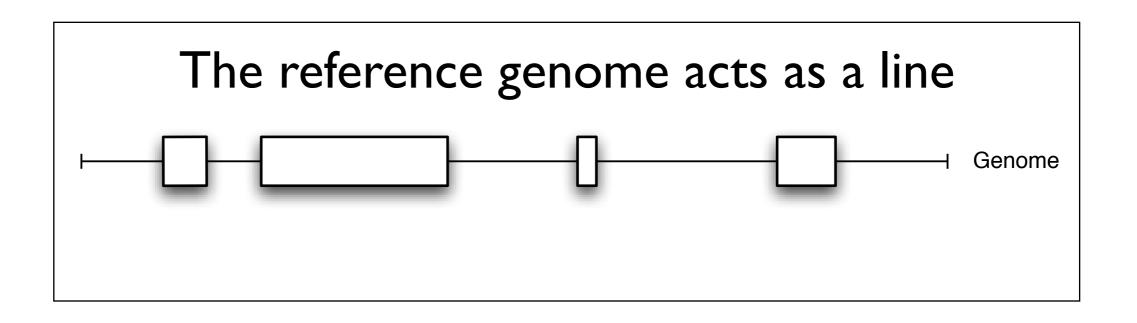
This can't be it ?!



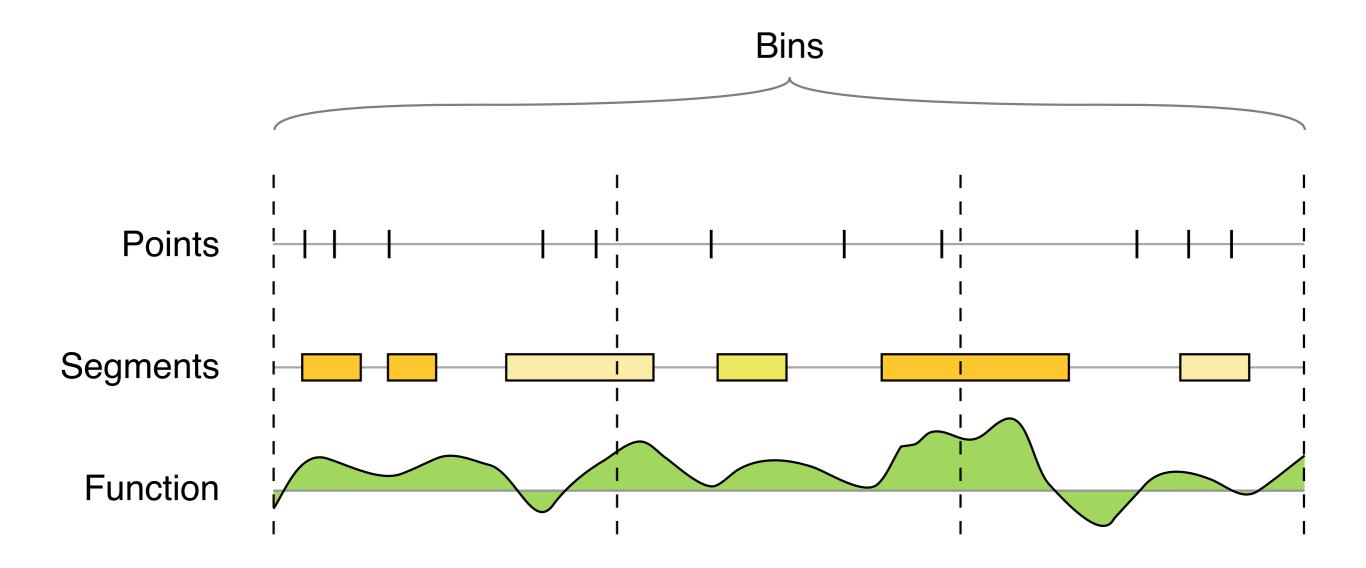
The whiteboard view of genome data

Reference genome acts like coordinate system for genomic data

chr21 10079666 10120808 NM_001187 chr21 13332357 13412442 NR_026916 chr21 13700575 13700652 NR_036164 chr21 13904368 13935777 NM_174981 chr21 14137324 14142556 NR_026755



Delineating basic types of genomic tracks



Cataloguing generic analyses of relations between the formal data types

Ρ	Р	Different frequencies?
P	P	Located nearby?
P	S	Located inside?
P	S	Located nonuniformly inside?
P	S	Located nearby?
S	S	Similar segments?
S	S	Overlap?
S	S	Located nearby?
F	F	Correlated?
P	F	Higher values at locations?
S	F	Higher values inside?
Ρ	VS	Located in segments with high values?
S	VP	Higher values inside segments?
VP	VP	Nearby values similar?
P	VS (c/c)	Located in case segments
VS (c/c)	S	Preferential overlap?
VP (cat)	VS (cat)	Category pairs differentially co-located?
LGP	P	Colocalized in 3D?

Biological example

• B-cells important for the pathology of multiple sclerosis?

From history (bed, wig,) ‡ 1: imported: MS regions [hg18]	\$	
	What is a	genomic track
Second Track		
Chromatin [221] ‡		
L Chromatin state segmentation [144] ‡		
_ wgEncodeBroadHmmGm12878HMM [16] ‡		
I_ 1 Active Promoter ‡		
9		
Analysis		
Category: Hypothesis testing ‡ Overlap? ‡ ?		
Are 'imported. MS regions (1)' overlapping '1 Active Promoter		
(wgEncodeBroadHmmGm12878HMM)', more than expected by chance?		
Ĩ		8
Segments		
Segments		

You asked:

Are 'imported: MS regions' overlapping '1 Active Promoter (wgEncodeBroadHmmGm12878HMM)', more than expected by chance?

Simplistic answer:

Yes - the data suggests this (p-value: 0.004975)

Precise answer:

The p-value is 0.004975 for the test

H0: The segments of track 1 are located independently of the segments of track 2 with respect to overlap

VS

H1: The segments of track 1 tend to overlap the segments of track 2

Low p-values are evidence against H0.

The test was also performed for <u>each bin separately</u>, resulting in 12 significant bins out of 26, at 10% FDR* (17 bins excluded from FDR-analysis due to lacking p-values).

Please note that both the effect size and the p-value should be considered in order to assess the practical significance of a result.

* False Discovery Rate: The expected proportion of false positive results among the significant bins is no more than 10%.

@REGION= chrArms

@BINNING=*

@TN1=Chromatin:Chromatin%20state
%20segmentation:wgEncodeBroadHmmGm12878HMM:1%20Active
%20Promoter

@TN2=Phenotype and disease associations:Assorted experiments:Multiple Sclerosis, Sawcer et al. (2011)

@ANALYSIS=RandomizationManagerStat(tf1=TrivialFormatConv erter,tail=more,assumptions=PermutedSegsAndSampledInters egsTrack_,rawStatistic=TpRawSegsOverlapStat,maxSamples=1 00,numResamplings=100,tf2=TrivialFormatConverter)

hg18|@REGION|@BINNING|@TN1|@TN2|@ANALYSIS

@REGION=___chrArms__

@BINNING=*

@TN1=Chromatin:Chromatin%20state
%20segmentation:wgEncodeBroadHmmGm12878HMM:*1%20Active
%20Promoter

@TN2=Phenotype and disease associations:Assorted experiments:Multiple Sclerosis, Sawcer et al. (2011)

@ANALYSIS=RandomizationManagerStat(tf1=TrivialFormatConv erter,tail=more,assumptions=PermutedSegsAndSampledInters egsTrack_,rawStatistic=TpRawSegsOverlapStat,maxSamples=1 00,numResamplings=100,tf2=TrivialFormatConverter)

hg18|@REGION|@BINNING|@TN1|@TN2|@ANALYSIS

B-cells important for MS?

- MS-associated regions (GWAS)
- Active regions in B-cells (chromatin state AP)
- Do MS overlap unexpectedly with B-cell AP?
 - But: They may also overlap other-cell AP..
 - Must use case-control analysis

Combine two BED files into single case-control track
Genome build: Human Mar. 2006 (hg18/NCBI36) ‡
Select track to be used as case: 8 - Extract track from HyperBrowser repc ‡
Select track to be used as control: 9 - Extract track from HyperBrowser repc ‡
Shared regions should be: removed \$
Execute
Corresponding batch command line:

From history (bed, wig,) * 1: imported: MS regions [hg18] * What is a genomic track Second Track From history (bed, wig,) * 10: Combine two BED files into single cas *	he Genomic HyperBrowser (v1.6)	
From history (bed, wig,)	enome build: Human Mar. 2006 ((hg18/NCBI36) ‡
What is a genomic track Second Track From history (bed, wig,) + 10: Combine two BED files into single cas + Analysis Category: Hypothesis testing + Preferential overlap? + ? Are 'Combine two BED files into single case-control track (10)' marked as case overlapping unexpectedly more with 'imported. MS regions (1)' than 'Combine two BED files into single case-control track (10)' marked as control? Track type Treat 'Combine two BED files into single case-control track (10)' as: Original format (") + ? Options Alternative hypothesis: more +	First Track	
From history (bed, wig,) 10: Combine two BED files into single cas Analysis Category: Hypothesis testing Preferential overlap? ? Are 'Combine two BED files into single case-control track (10)' marked as case overlapping unexpectedly more with 'imported. MS regions (1)' than 'Combine two BED files into single case-control track (10)' marked as control? Track type Treat 'Combine two BED files into single case-control track (10)' as: Original format (") ; Options Alternative hypothesis: more ; 10: Combine two BED files into single case-control track (10)' as:	From history (bed, wig,)	1: imported: MS regions [hg18] What is a genomic track
Analysis Category: Hypothesis testing ‡ Preferential overlap? ‡? Are 'Combine two BED files into single case-control track (10)' marked as case overlapping unexpectedly more with 'imported. MS regions (1)' than 'Combine two BED files into single case-control track (10)' marked as control? Track type Treat 'Combine two BED files into single case-control track (10)' as: Original format (") ‡ Treat 'imported. MS regions (1)' as: Original format (") ‡ ? Options Alternative hypothesis: more ‡	Second Track	
Category: Hypothesis testing Preferential overlap? Are 'Combine two BED files into single case-control track (10)' marked as case overlapping unexpectedly more with 'imported. MS regions (1)' than 'Combine two BED files into single case-control track (10)' marked as control? Track type Treat 'Combine two BED files into single case-control track (10)' as: Original format (") reat 'imported. MS regions (1)' as: Original format (") Options Alternative hypothesis: more	From history (bed, wig,)	10: Combine two BED files into single cas \$
Are 'Combine two BED files into single case-control track (10)' marked as case overlapping unexpectedly more with 'imported. MS regions (1)' than 'Combine two BED files into single case-control track (10)' marked as control? Track type Treat 'Combine two BED files into single case-control track (10)' as: Original format (") ‡ Treat 'imported. MS regions (1)' as: Original format (") ‡ ? Options Alternative hypothesis: more ‡	Analysis	
unexpectedly more with 'imported. MS regions (1)' than 'Combine two BED files into single case-control track (10)' marked as control? Track type Treat 'Combine two BED files into single case-control track (10)' as: Original format (") ‡ Treat 'imported. MS regions (1)' as: Original format (") ‡ ? Options Alternative hypothesis: more ‡	Category. Hypothesis testing	• Preferential overlap:
Treat 'Combine two BED files into single case-control track (10)' as: Original format (") ‡ Treat 'imported. MS regions (1)' as: Original format (") ‡ ? Options Alternative hypothesis: more ‡	unexpectedly more with 'impo	rted. MS regions (1)' than 'Combine two BED files into single
Original format (") Treat 'imported. MS regions (1)' as: Original format (") ? Options Alternative hypothesis: more \$	Track type	
Treat 'imported. MS regions (1)' as: Original format (") ‡ ? Options Alternative hypothesis: more ‡	Treat 'Combine two BED files int	o single case-control track (10)' as:
? Options Alternative hypothesis: more \$	Original format (")	\$
Alternative hypothesis: more \$	Treat 'imported. MS regions (1)'	as: Original format (") +
Alternative hypothesis: more \$?	
	Options	
		*
		of both tracks: permu 1

You asked:

Are 'Combine two BED files into single case-control track' marked as case overlapping unexpectedly more with 'imported: MS regions' than 'Combine two BED files into single case-control track' marked as control?

Simplistic answer:

Yes - the data suggests this (p-value: 0.004975)

Precise answer:

The p-value is 0.004975.

Low p-values are evidence against H0.

The test was also performed for <u>each bin separately</u>, resulting in 4 significant bins out of 26, at 10% FDR* (17 bins excluded from FDR-analysis due to lacking p-values).

Please note that both the effect size and the p-value should be considered in order to assess the practical significance of a result.

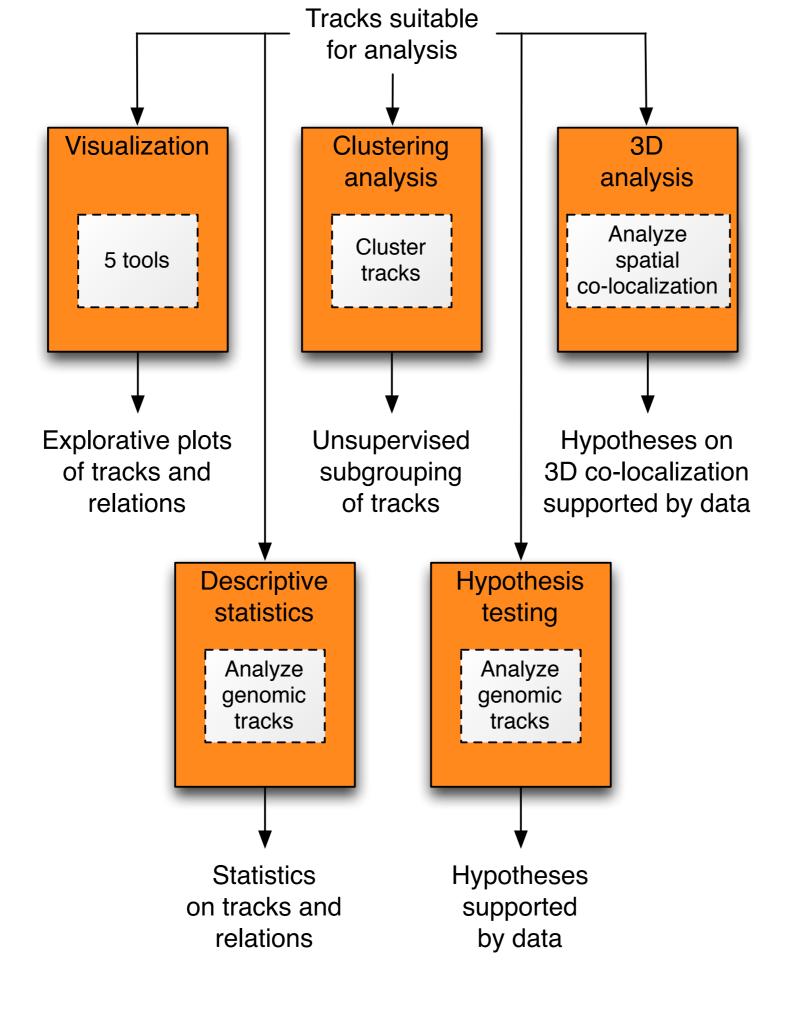
* False Discovery Rate: The expected proportion of false positive results among the significant bins is no more than 10%.

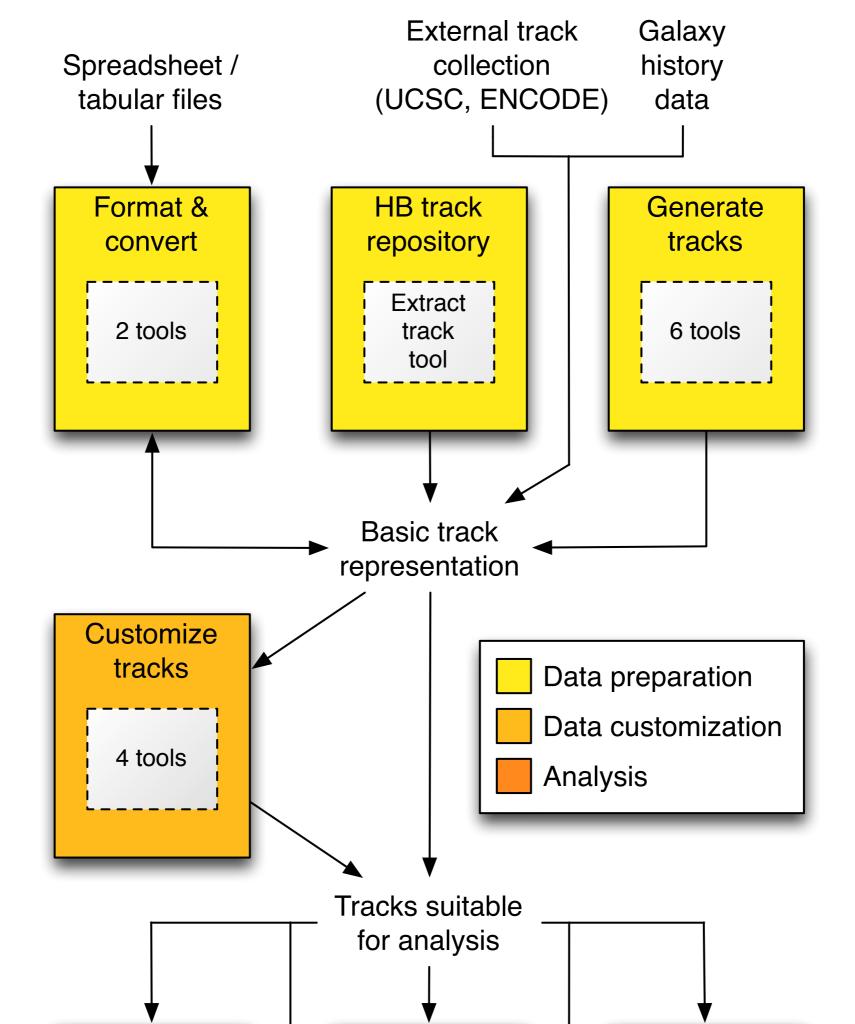
P-values were computed under the **null model** defined by the following preservation and randomization rules:

Preserve segments of both tracks; permute case and control assignment of T1-segments

The Genomic HyperBrowser

- Robust statistical treatment
- Dynamic Galaxy web interface
 - Determines meaningful analyses and options from data
- 76 statistical analyses
- 42 analysis-centric tools















If you have a genomic track, we can analyze it!

If you have a generic question for which we have no answer, we will develop it!

- Google "HyperBrowser" and try out the web system
- PubMed "HyperBrowser" and skim through our 2013 NAR article