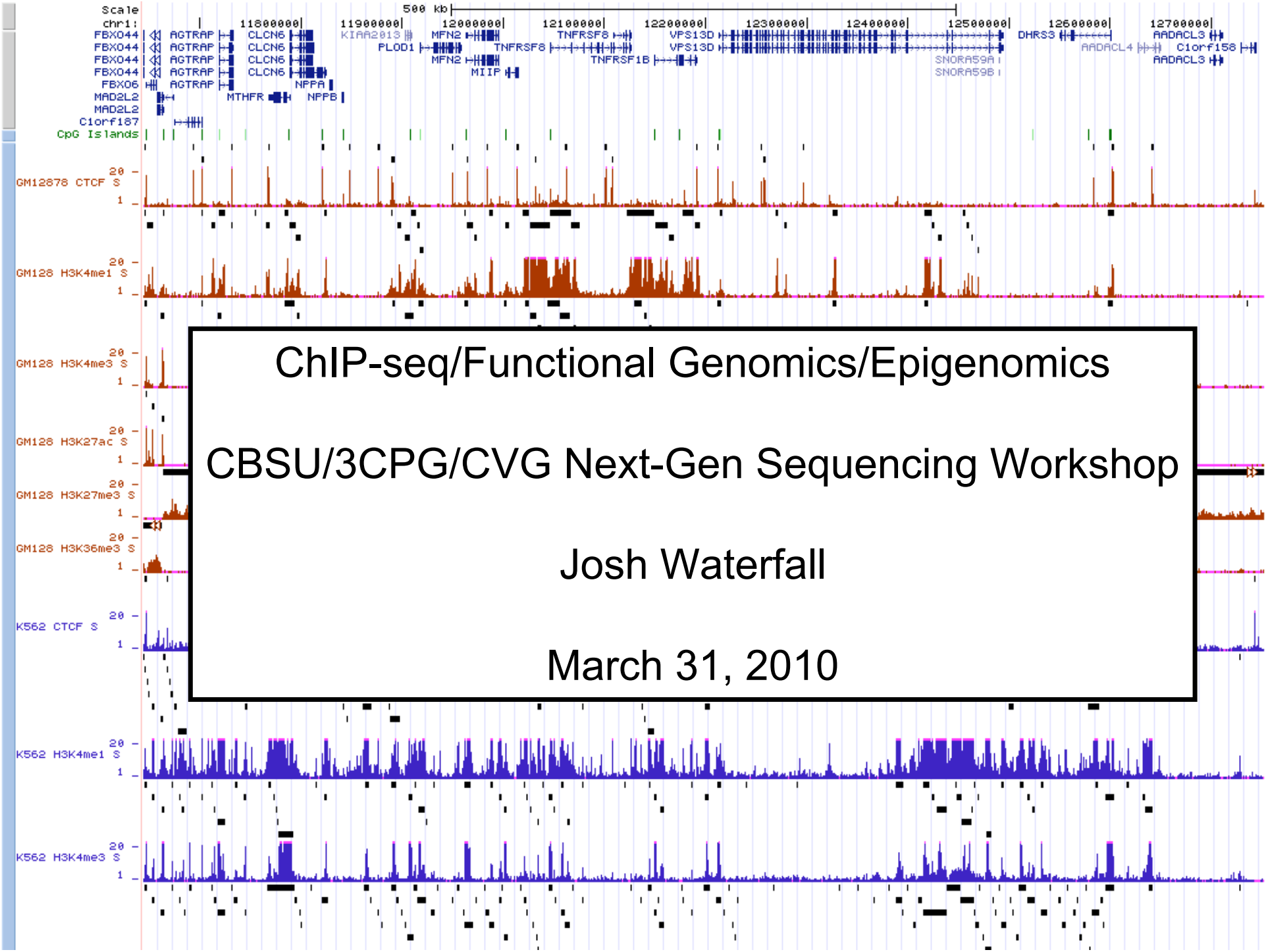


ChIP-seq/Functional Genomics/Epigenomics

CBSU/3CPG/CVG Next-Gen Sequencing Workshop

Josh Waterfall

March 31, 2010

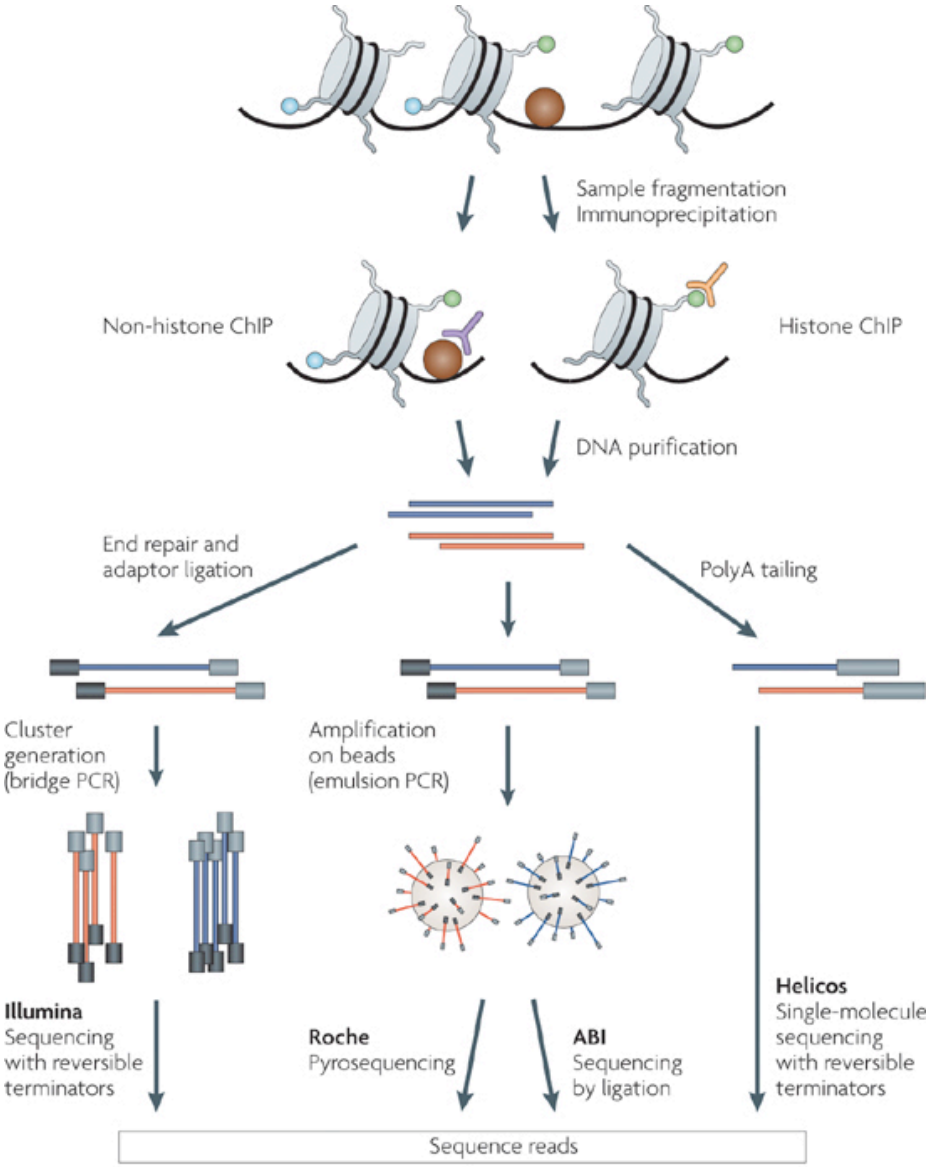
Outline

Introduction to ChIP-seq
- Control data sets
- Peak/enriched region identification

Related functional genomics assays

Useful web resources

Chromatin ImmunoPrecipitation (ChIP-seq)



Park, P.J., Nat. Genetics (2009) 10:669-680

Advantages over array-based methods

No cross-hybridization background
Lower end of sensitivity largely dependent on just sequencing depth
More linear, quantitative assay
Unmappable portion of genome is distinct from, and much smaller than, repeat masked portion
No limit based on probe locations
Needs less starting material
Higher resolution

Important Issues:

Appropriate Controls
Identifying Enriched Regions

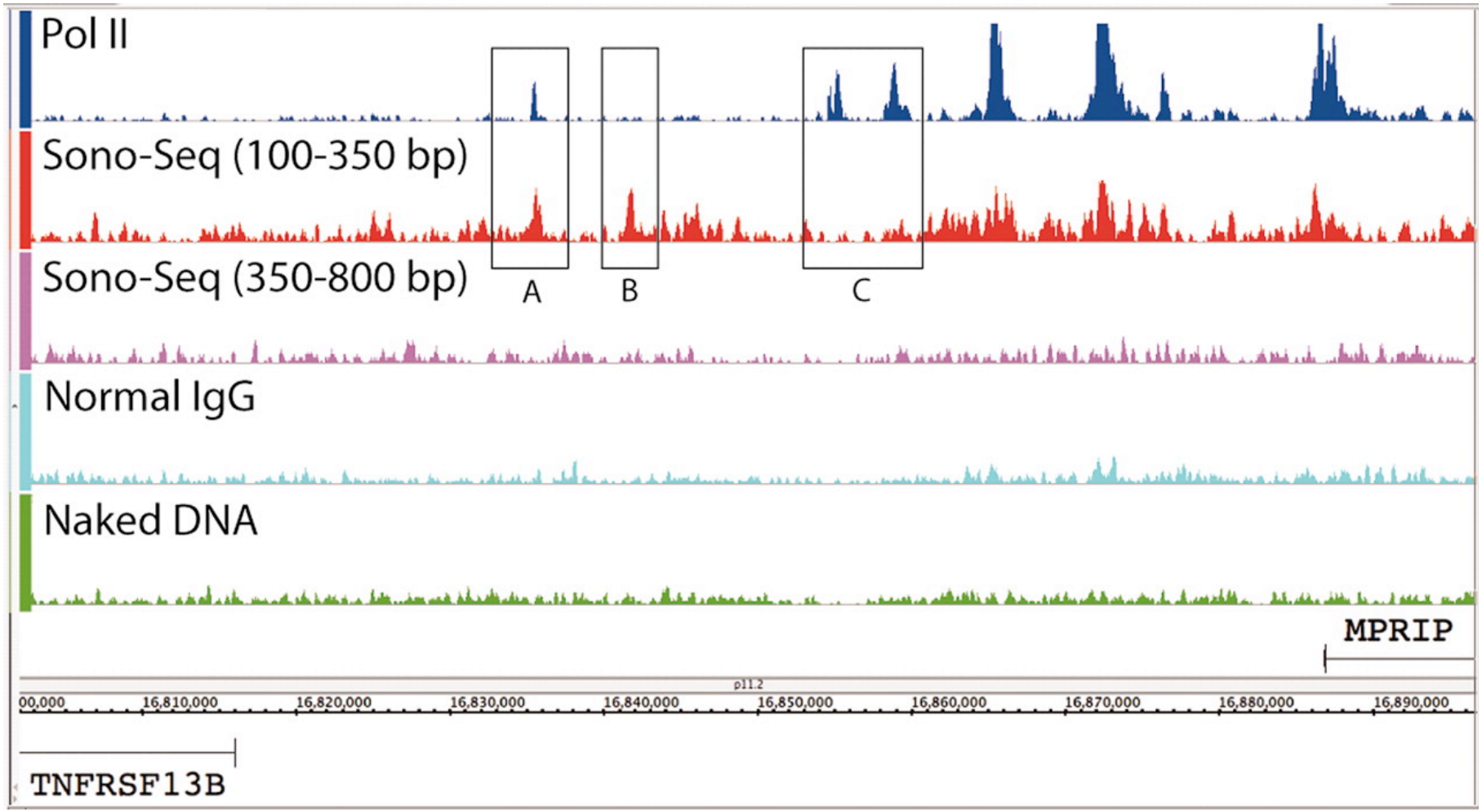
Controls

Sequencing is such high-sensitivity that signals invisible in any other assay are now apparent. Need rigorous controls to be confident of enrichment.

- Input DNA has non-random pattern (open chromatin shears more easily) - Sono-seq is an actual assay.
- Mock-IP controls for more steps in ChIP protocol than input DNA but not antibody cross-reactivity.
- Different antibodies (to different epitopes) in separate experiments, or ChIP after target protein has been depleted (or in cell-line without tagged protein), help control for cross-reactivity.

- To characterize new antibody, IP and mass-spec everything that comes down to verify only expected binding partners are seen.

Background rate is non-uniform



Auerbach R K et al. PNAS 2009;106:14926-14931

Identifying enriched regions

Identifying narrow ChIP peaks is very different from:

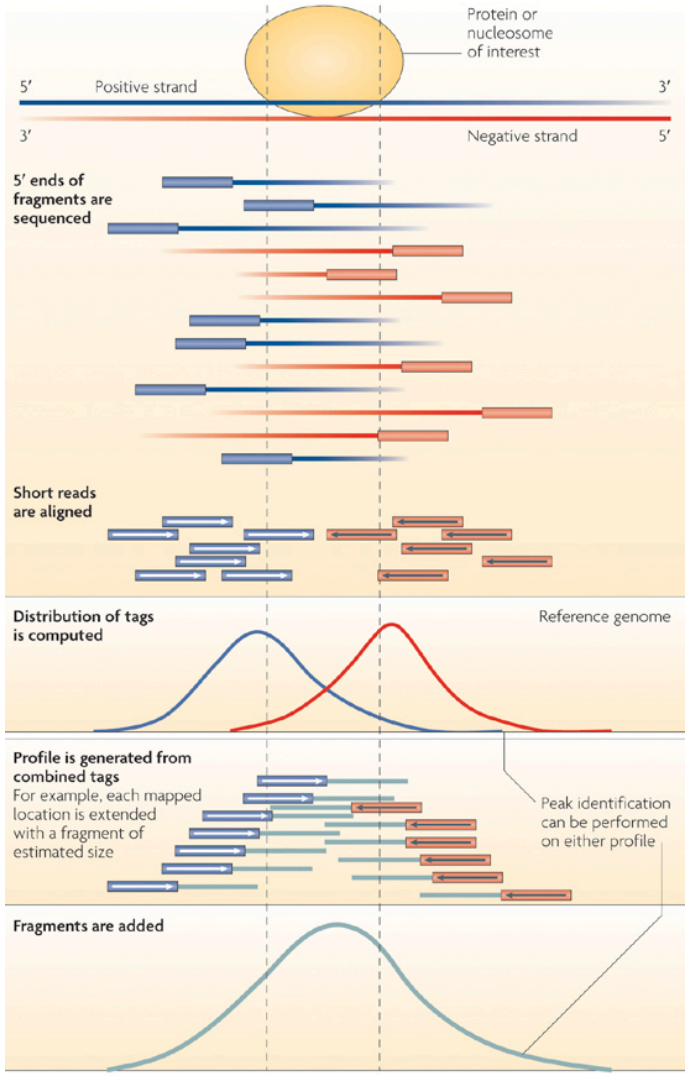
identifying broadly enriched regions in ChIP.

identifying narrow peaks from other assays.

- Many well-worked out programs for identifying narrow ChIP peaks (e.g. from sequence specific binding factors).

  The best programs can exploit strand-specific patterns, local versus global background levels, mappability, etc.
- Other assays also result in localized peaks but without same strand pattern.
- Most work on broad regions (e.g. particular histone modifications or RNA polymerase) is based on sliding windows (either of fixed length or fixed read count).

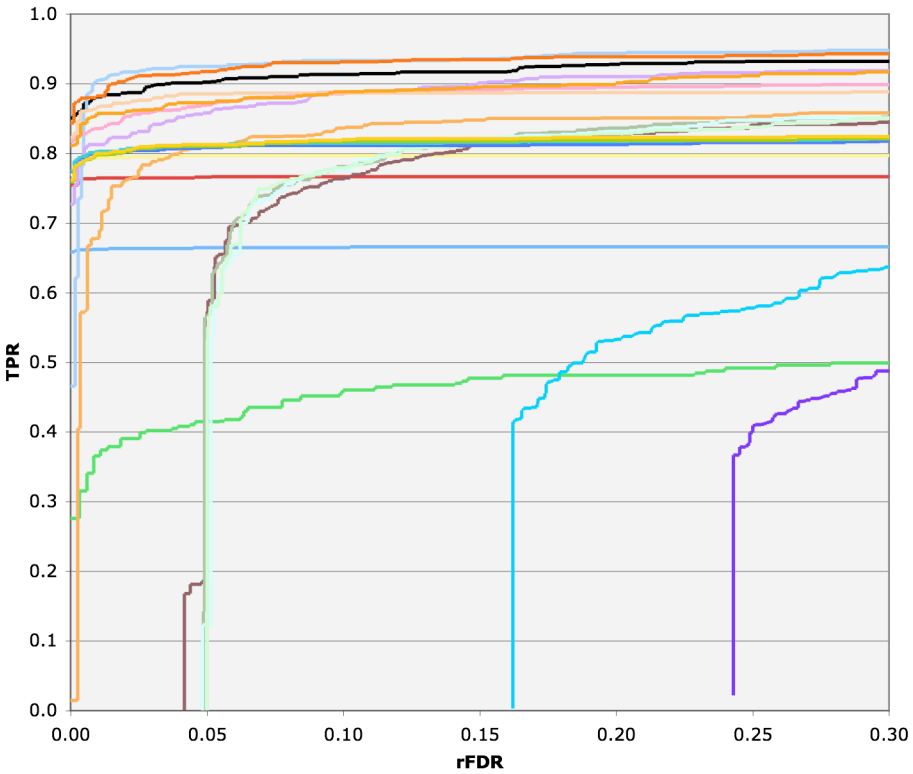
Strand specific pattern for localized peaks in ChIP-seq



Park, P.J., Nat. Genetics (2009) 10:669-680

Peak Detection Software

Is strand information used?
If fragment size is used, is it defined by user or estimated from data?
(How) is control data incorporated?
Is background defined locally or globally?
(How) are unmappable regions treated?



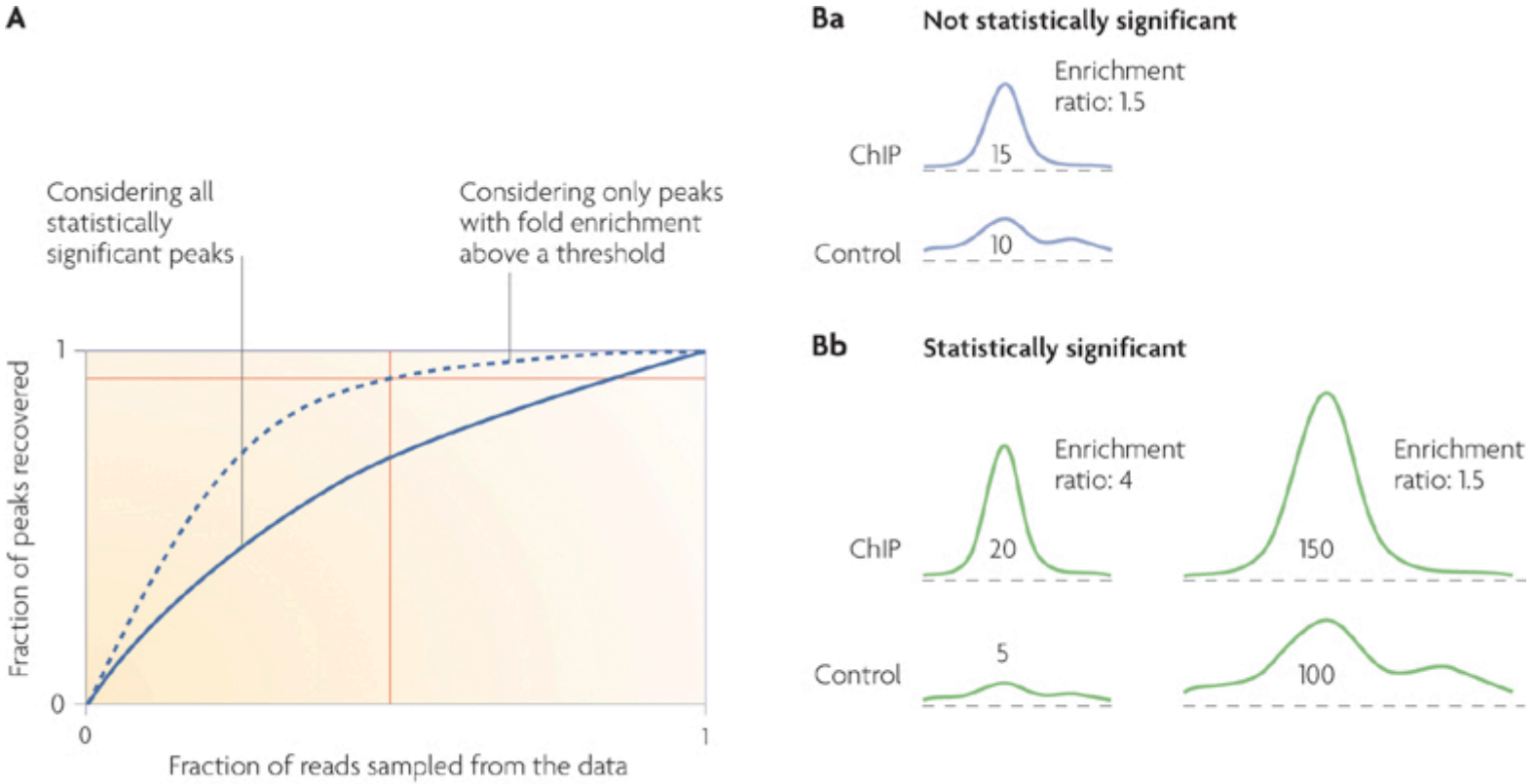
ChIP-seq peak calling software contest on http://seqanswers.com

Table 1 | Publicly available ChIP-seq software packages discussed in this review

	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: $\frac{\#\ control}{\#\ ChIP}$	Optional peak height, ratio to background	Yes / No	4,18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	14
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\#\ control}{\#\ ChIP}$	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\ control}{\#\ ChIP}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	*q* value	1: NA 2: $\frac{\#\ control}{\#\ ChIP}$ as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes	9
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value (no control)	No / Yes	15
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance	Used to compute fold-enrichment distribution	*P* value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling	*P* value	1: Monte Carlo simulation 2: $\frac{\#\ control}{\#\ ChIP}$	Ratio to background	Yes / No	12
USeq v4.2	Window scan	Binomial *P* value	Estimated or user specified	Subtracted before peak calling	*q* value	1, 2: binomial 2: $\frac{\#\ control}{\#\ ChIP}$	Target FDR	No / Yes	20

Pepke, S., Wold, B., Mortazavi, A. (2009) Nat. Methods 6:S22-S32

Assessing saturation and significance

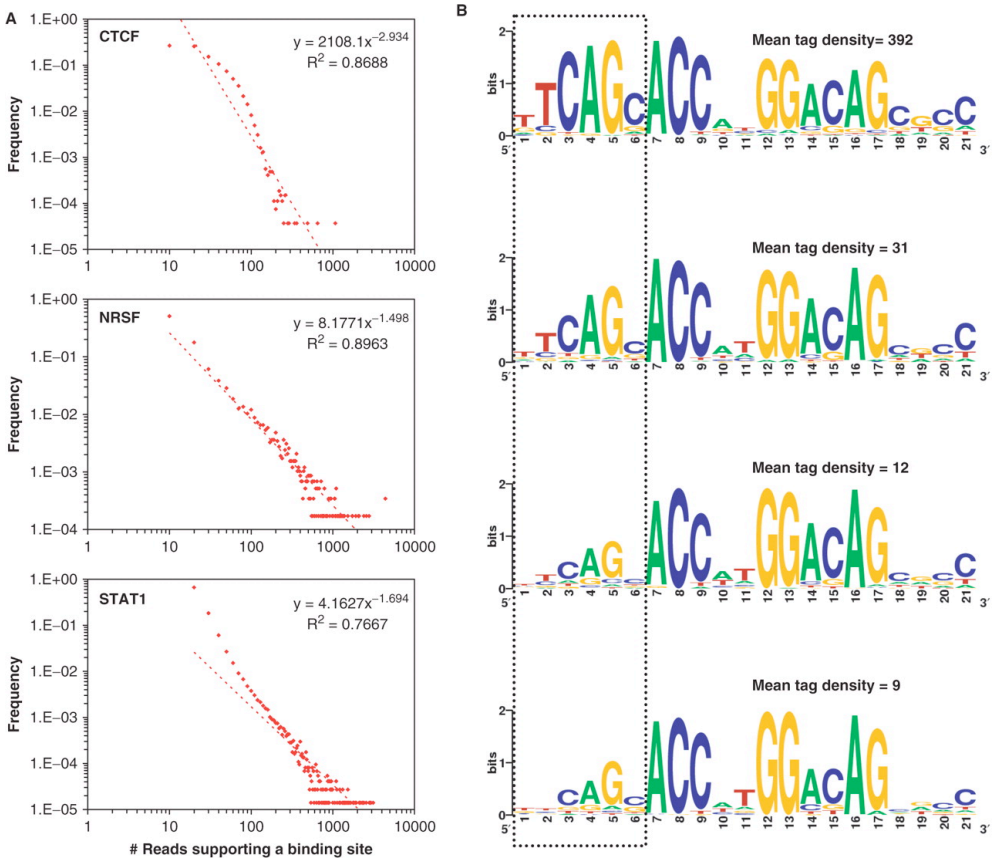


Saturation is dependent on both ChIP efficiency (globally) and factor:DNA affinity (locally)

Significance depends on both ChIP and control read counts

Park, P.J., Nat. Genetics (2009) 10:669-680

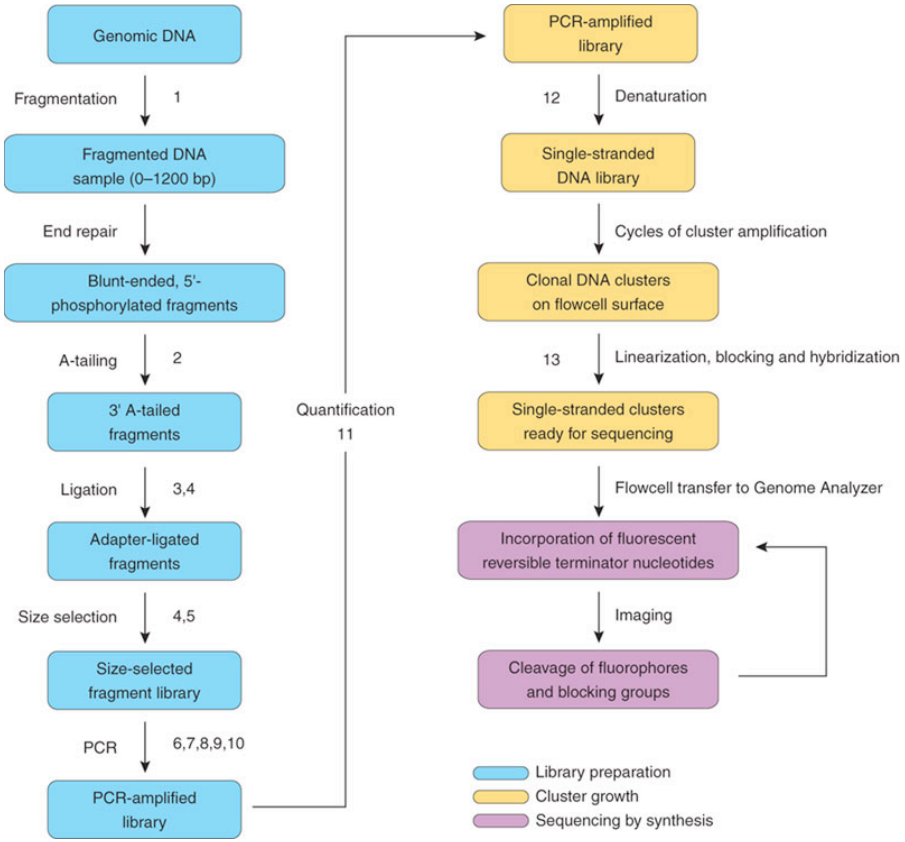
ChIP-seq read density is quantitative measure of binding level



Jothi, R., et. al. (2008) Nucleic Acids Res. 36:5221-31

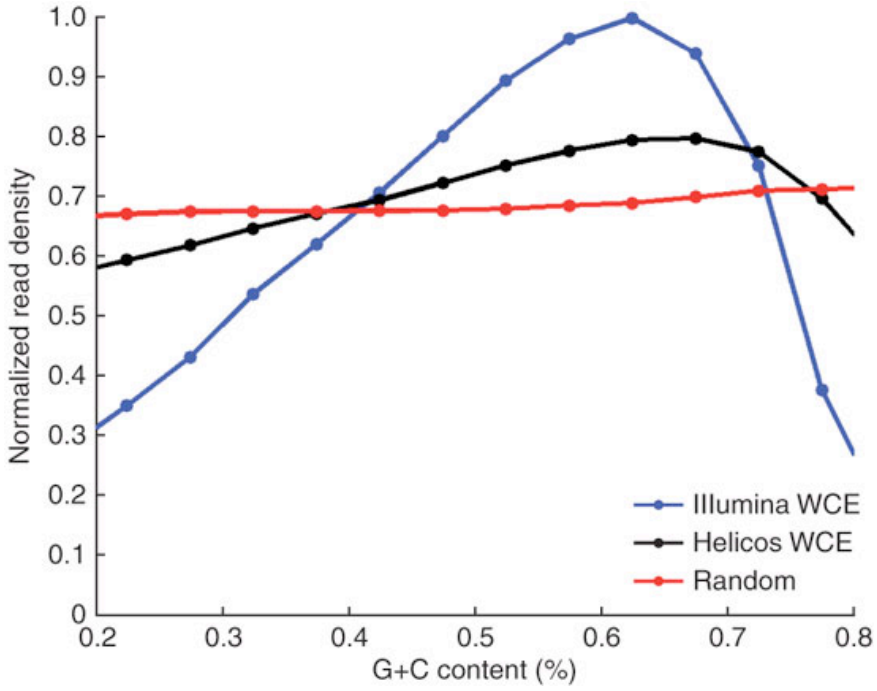
Solutions to GC bias

Improve Illumina protocol (especially gel purification)



Quail, MA, et. al. (2008) Nat. Methods 5:1005-1010

Use single molecule sequencing technology



Goren, A, et. al. (2009) Nat. Methods 7:47-49

Fragmenting DNA

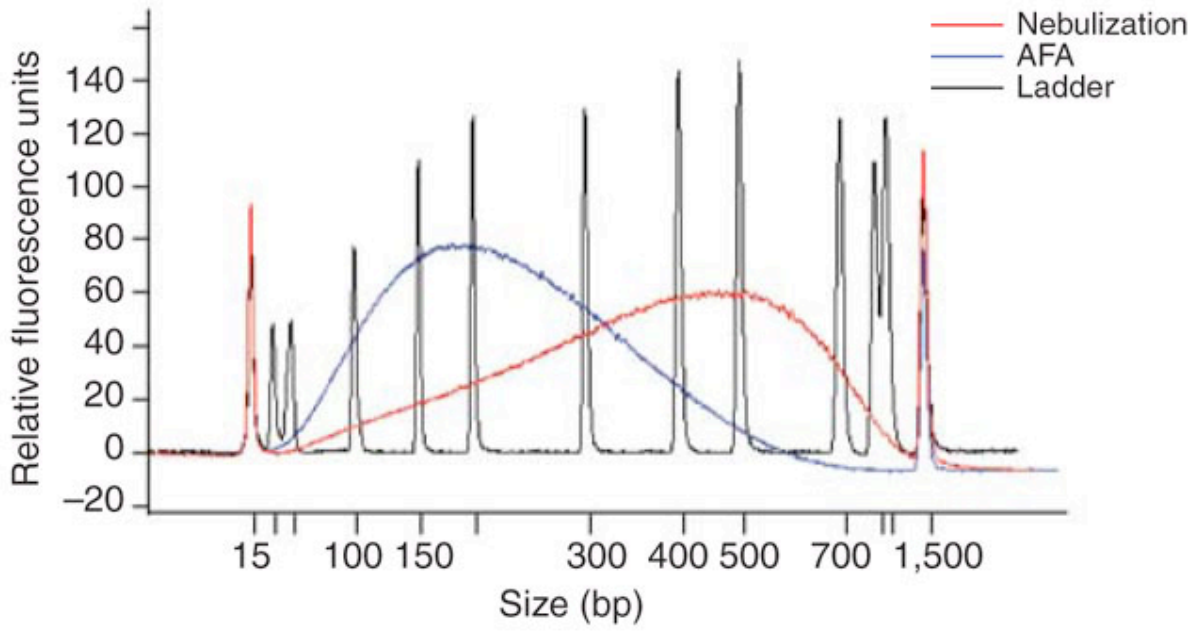
AFA (Adaptive Focused Acoustics)/Covaris

enzyme (MNase, NEB Fragmentase)

Nebulization

Sonication

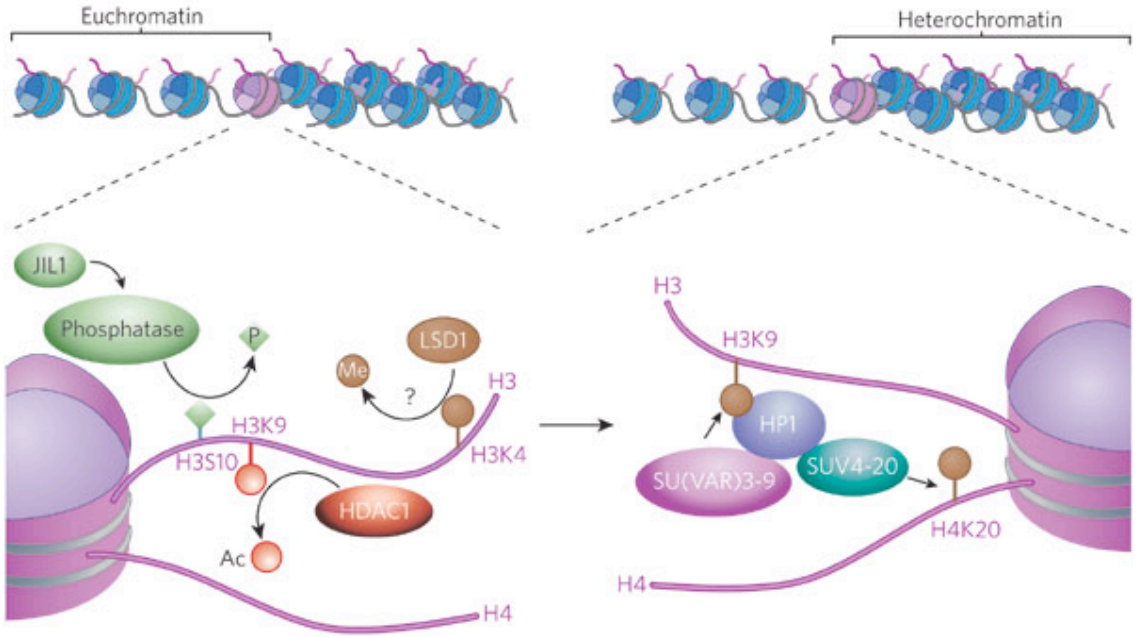
Quality Evaluation with Agilent Bioanalyzer



Quail, MA, et. al. (2008) Nat. Methods 5:1005-1010

Fragmenting DNA

Open chromatin fragments more easily than closed chromatin - ChIP for factors associated with transcriptional repression or heterochromatin can be more difficult than for factors associated with transcriptional activation.



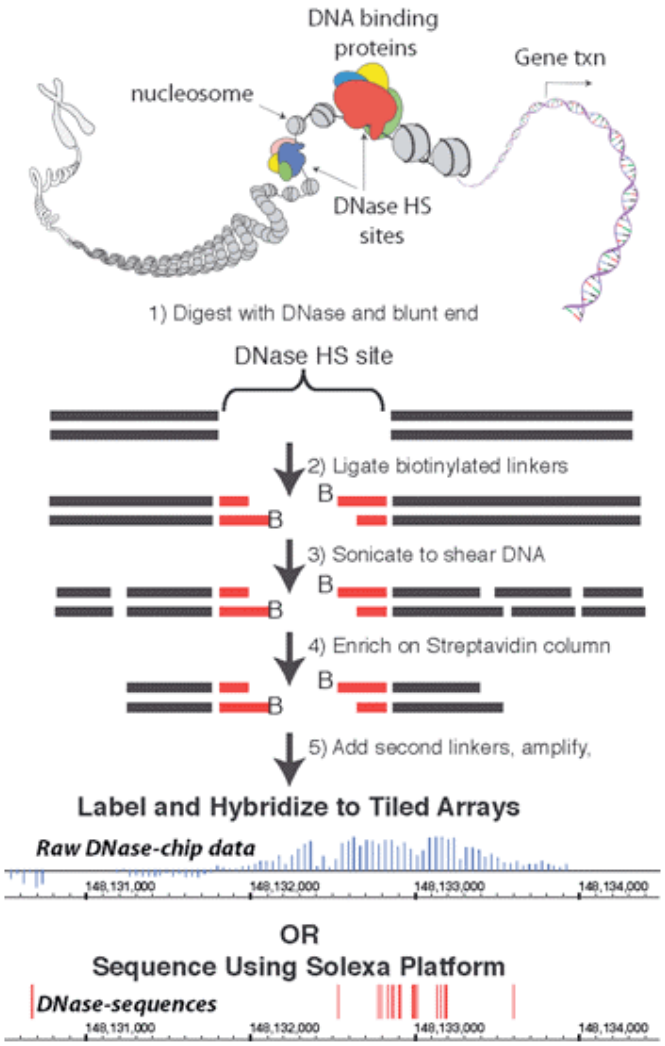
Grewal, SI, Elgin, SCR (2007) Nature 447:399-406

Related Functional Genomic Assays

- Chromatin ImmunoPrecipitation (ChIP)
- DNAse I Hypersensitivity
- Formaldehyde Assisted Isolation of Regulatory Elements
- DNA Methylation (bisulfite, affinity, or restriction enzyme based)
- Cap Analysis of Gene Expression (CAGE)
- Genomic Run On (GRO-seq)
- Many more…

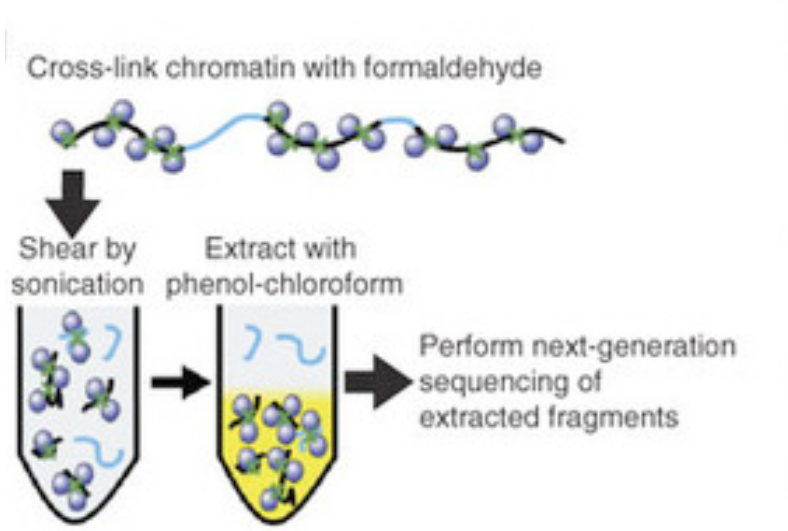
Each can couple to next-gen sequencing and entails analysis/identification of enriched regions

High-throughput methods to identify DNaseI HS sites.



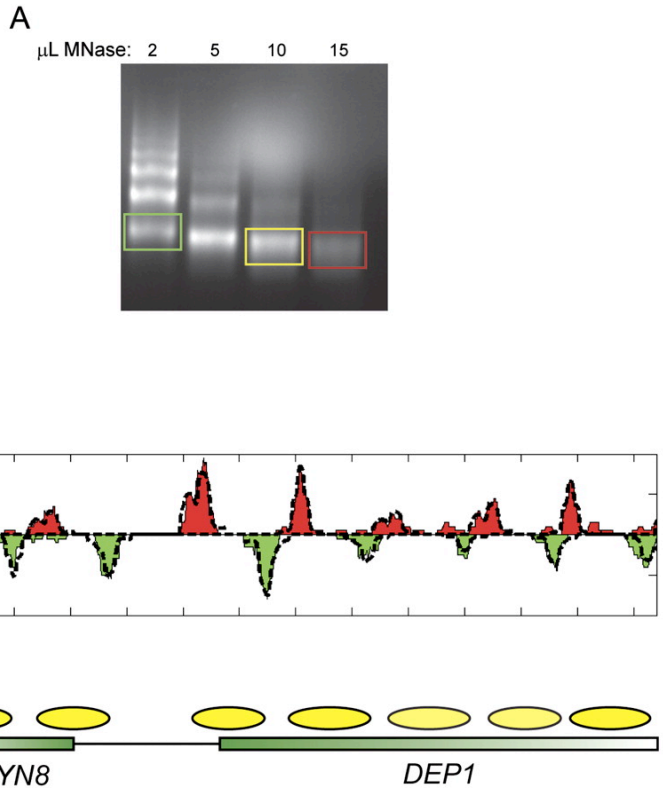
http://www.genome.duke.edu/people/faculty/crawford/research/

Formaldehyde Assisted Isolation of Regulatory Elements (FAIRE)



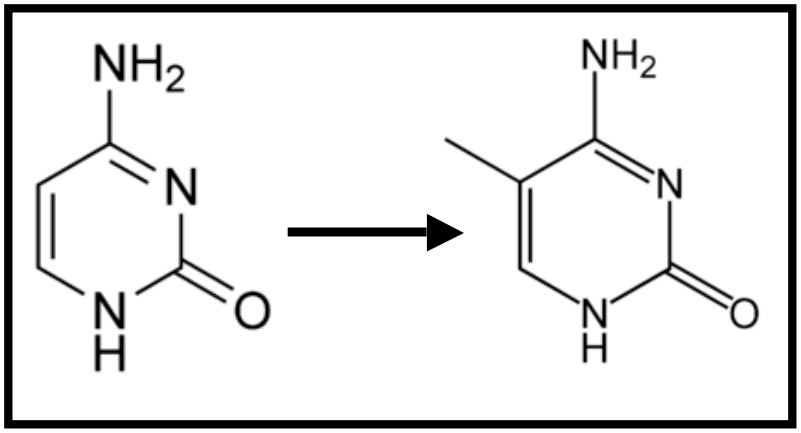
Gaulton, K.J., et. al. (2010) Nat. Genetics 42:255-259

Micrococcal Nuclease digestion (MNase-seq)



Weiner A et al. Genome Res. 2010;20:90-100

DNA methylation



3 broad classes of assays:

Enzyme based

methylation sensitive restriction enzymes

Affinity based

antibodies or other meDNA binding proteins used in ChIP-like experiment

Bisulfite based

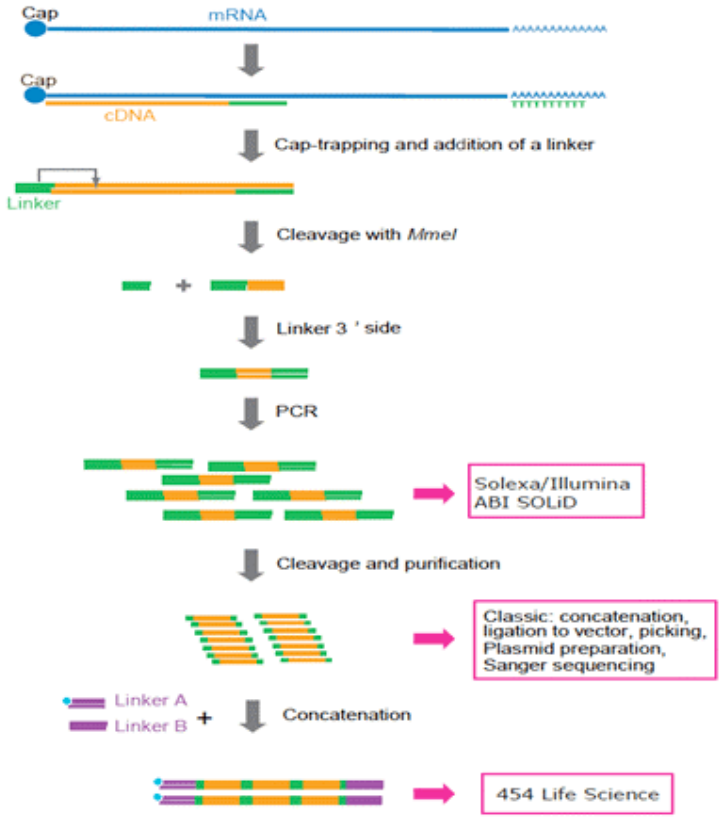
$NaHSO_3$ deaminates unmethylated cytosines to uracils but does not affect 5-methylcytosine.

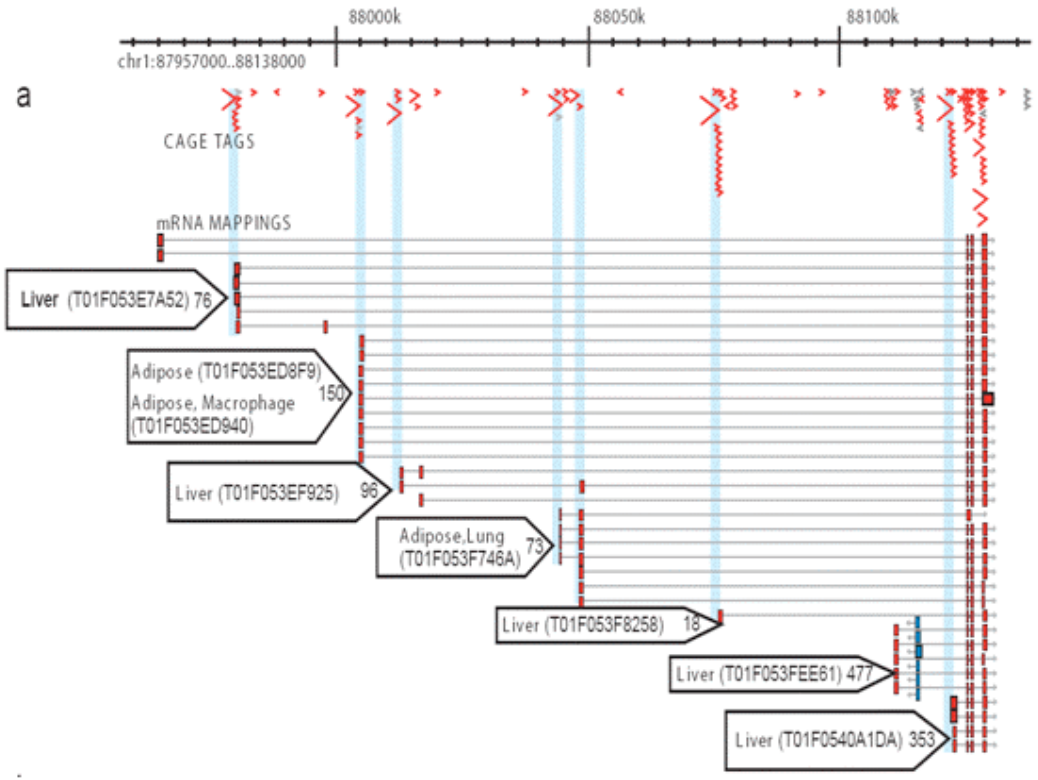
Reduced Representation based on enzymatic digest or hybridization enrichment common for cost efficiency in large, sparsely methylated (e.g. mammalian) genomes.

Read alignments done to *in silico* bisulfite-converted genome.

Direct sequencing of 5th base (future technologies)

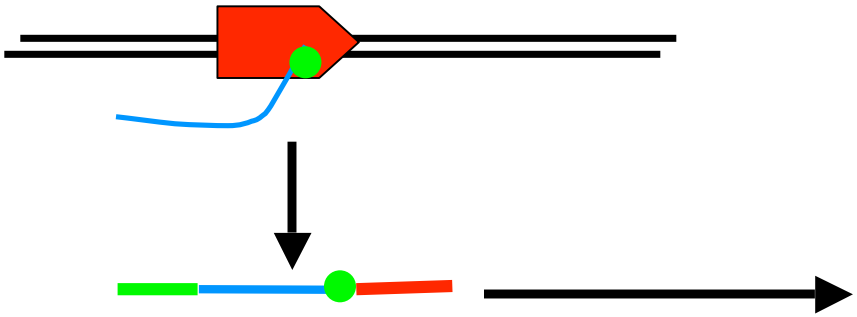
Cap Analysis of Gene Expression (CAGE)

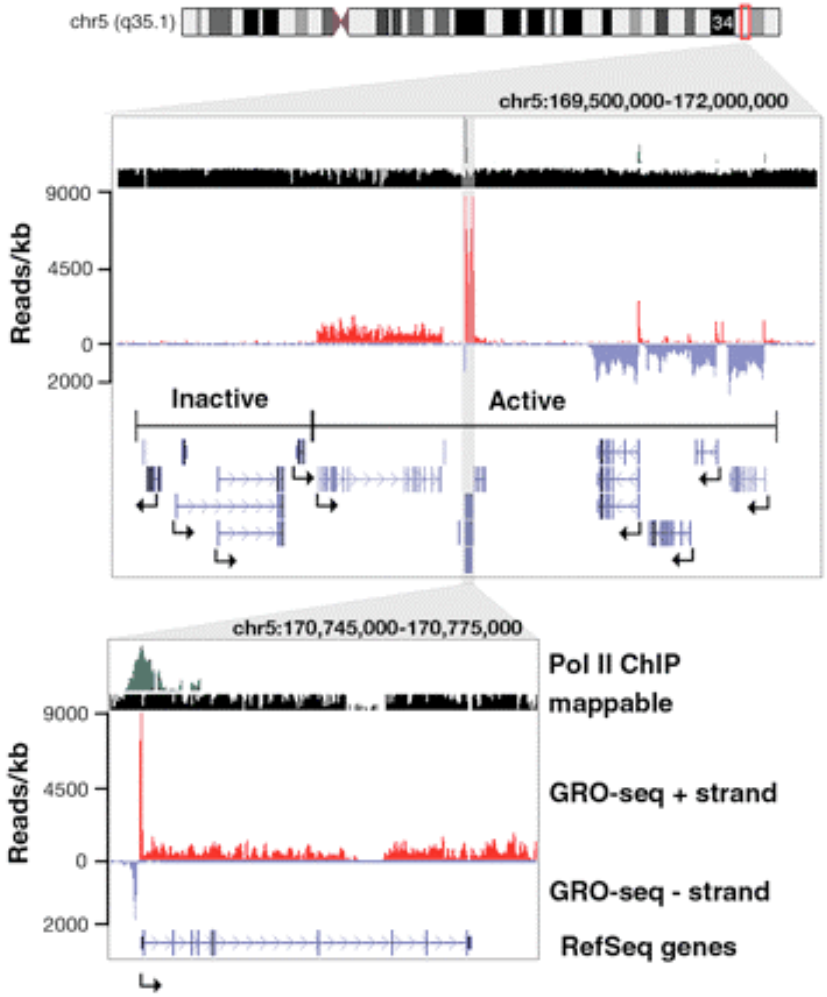




http://www.osc.riken.jp/english/activity/cage/basic/

Genomic Run On (GRO-seq)



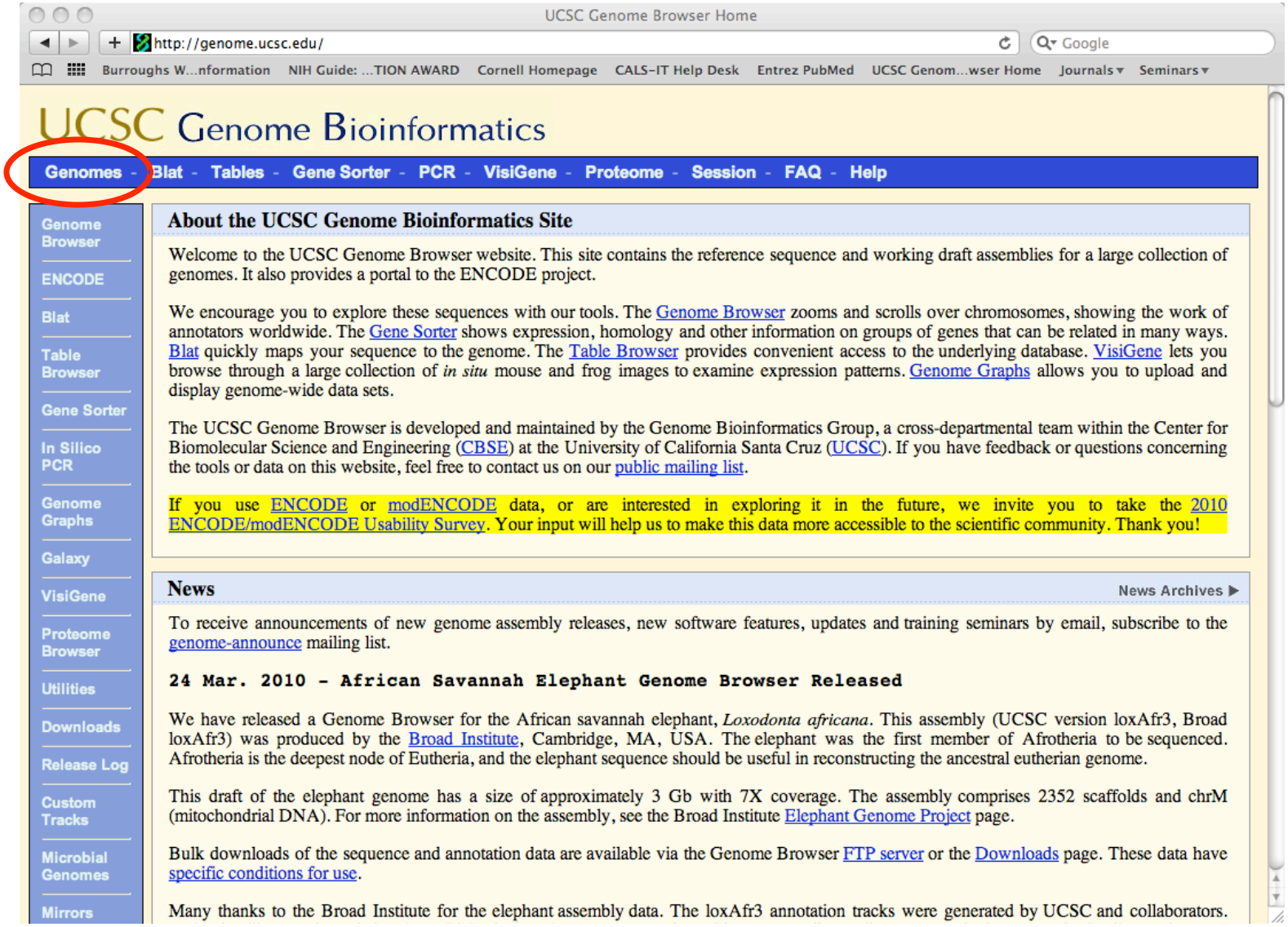


Core, LJ, Waterfall, JJ, Lis, JT (2008) Science 322:1845-8

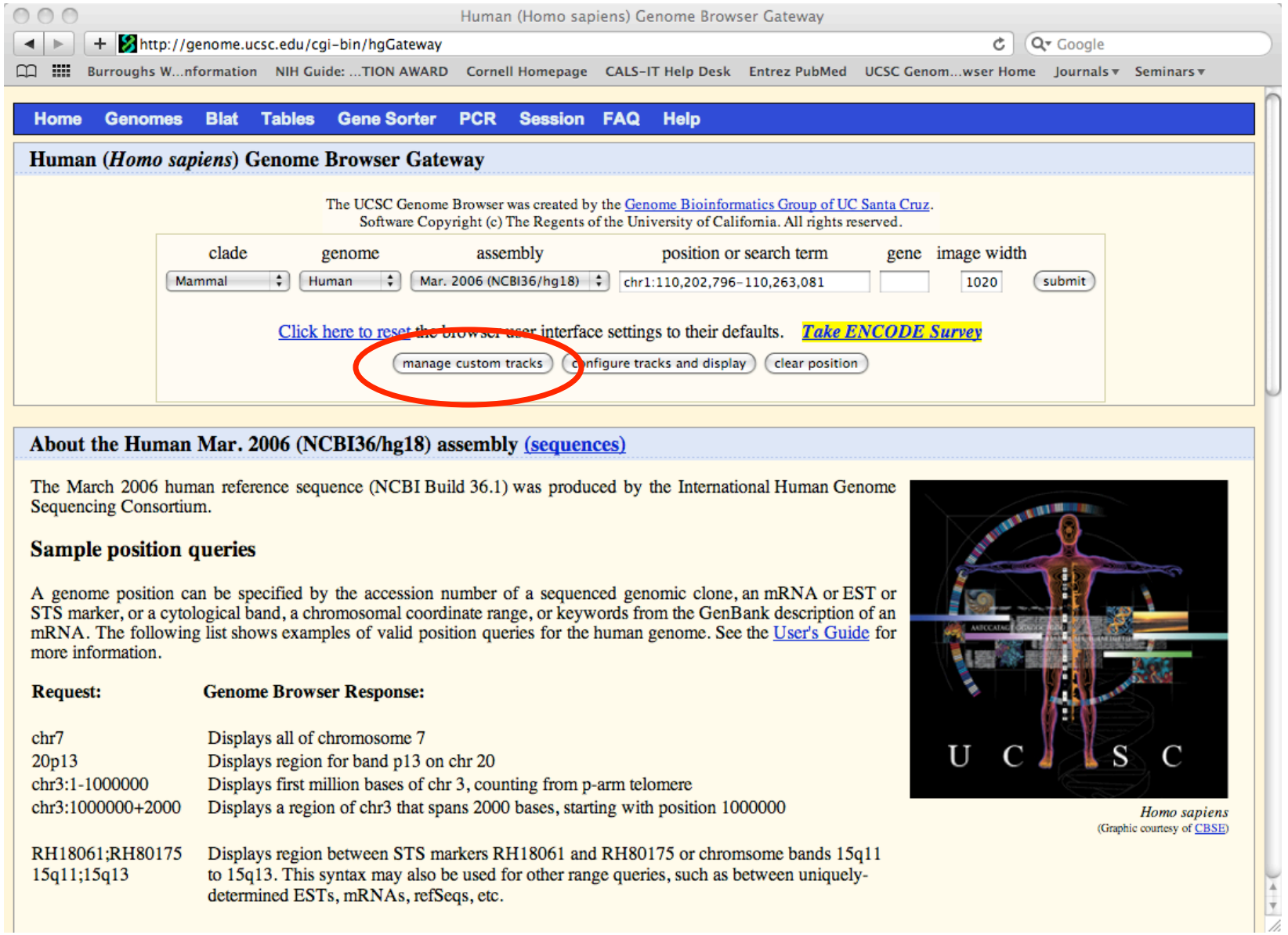
Web resources for analyzing, viewing, sharing, and collecting genomics data

UCSC Genome Browser (http://genome.ucsc.edu)
Galaxy (http://main.g2.bx.psu.edu)
NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/)

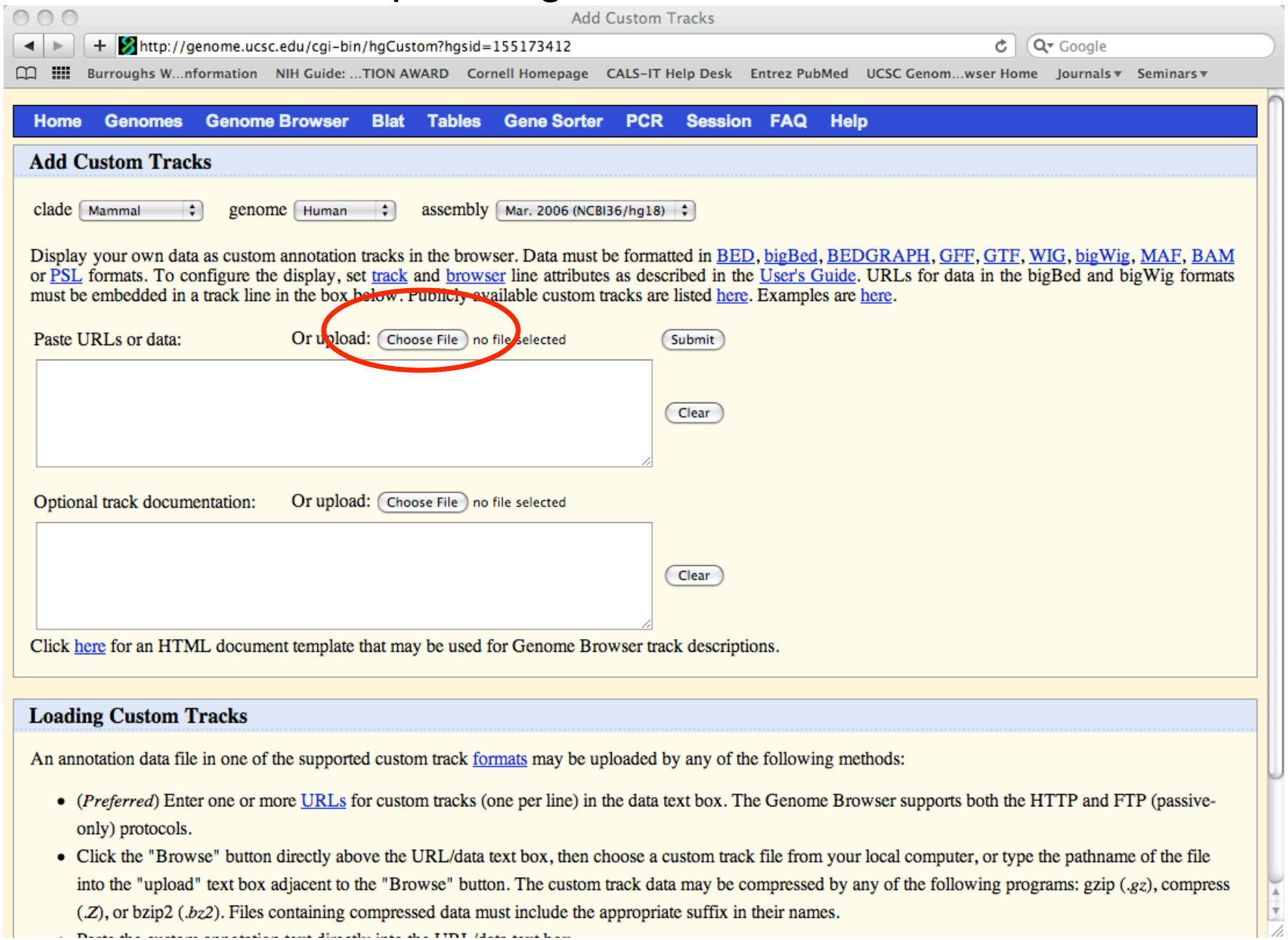
The UCSC Genome Browser
http://genome.ucsc.edu



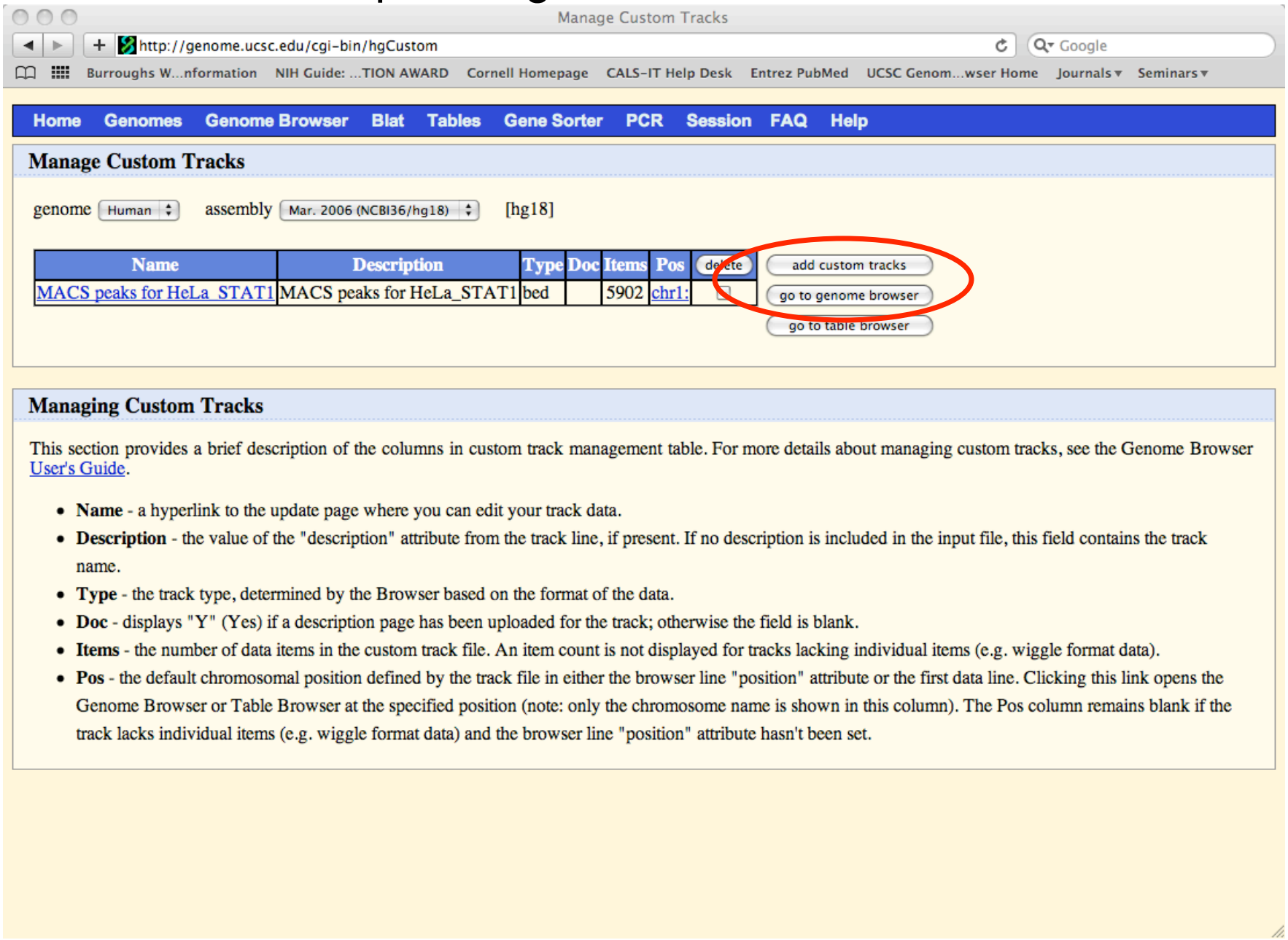
Uploading data to UCSC



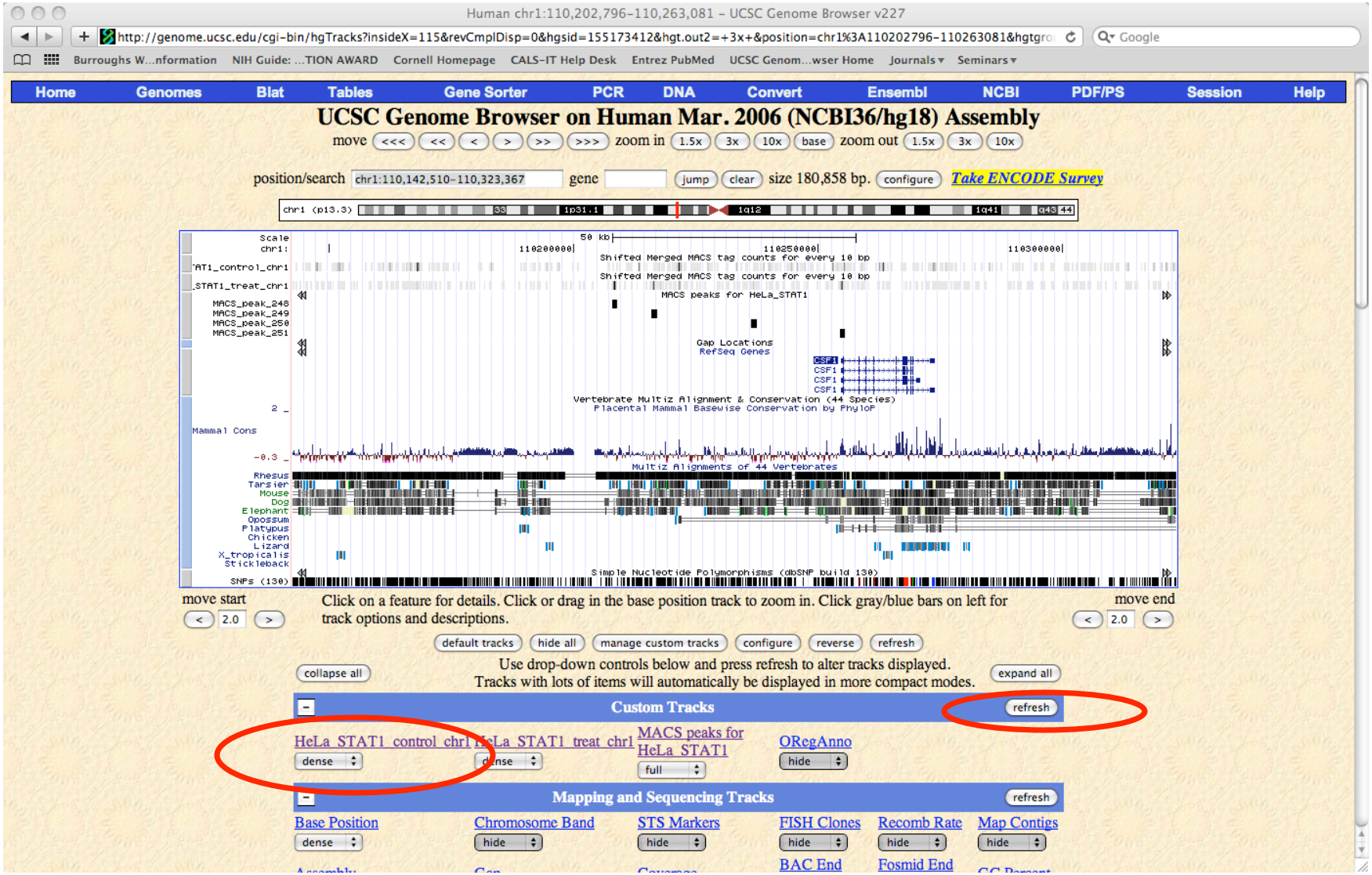
Uploading data to UCSC



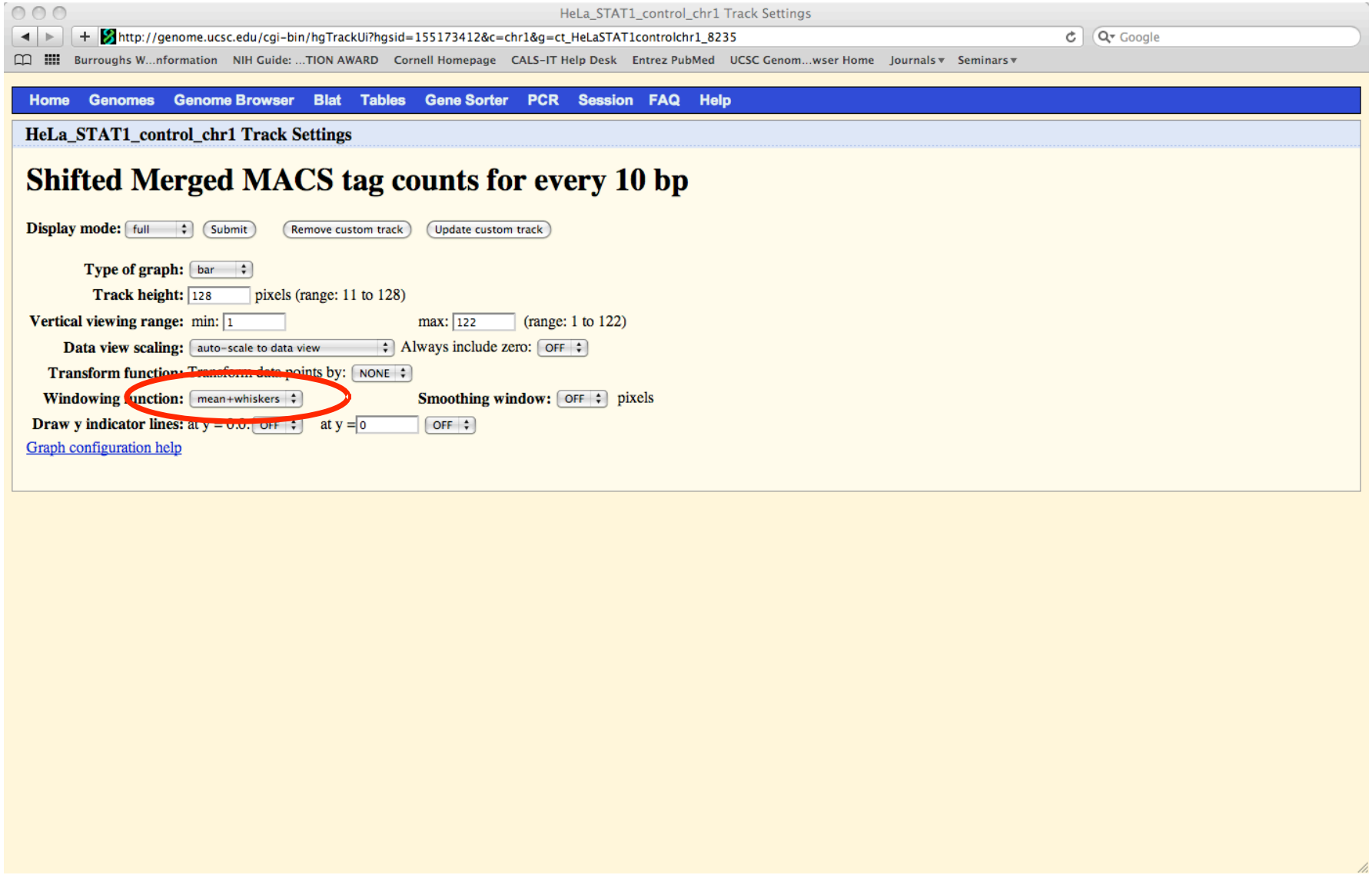
Uploading data to UCSC



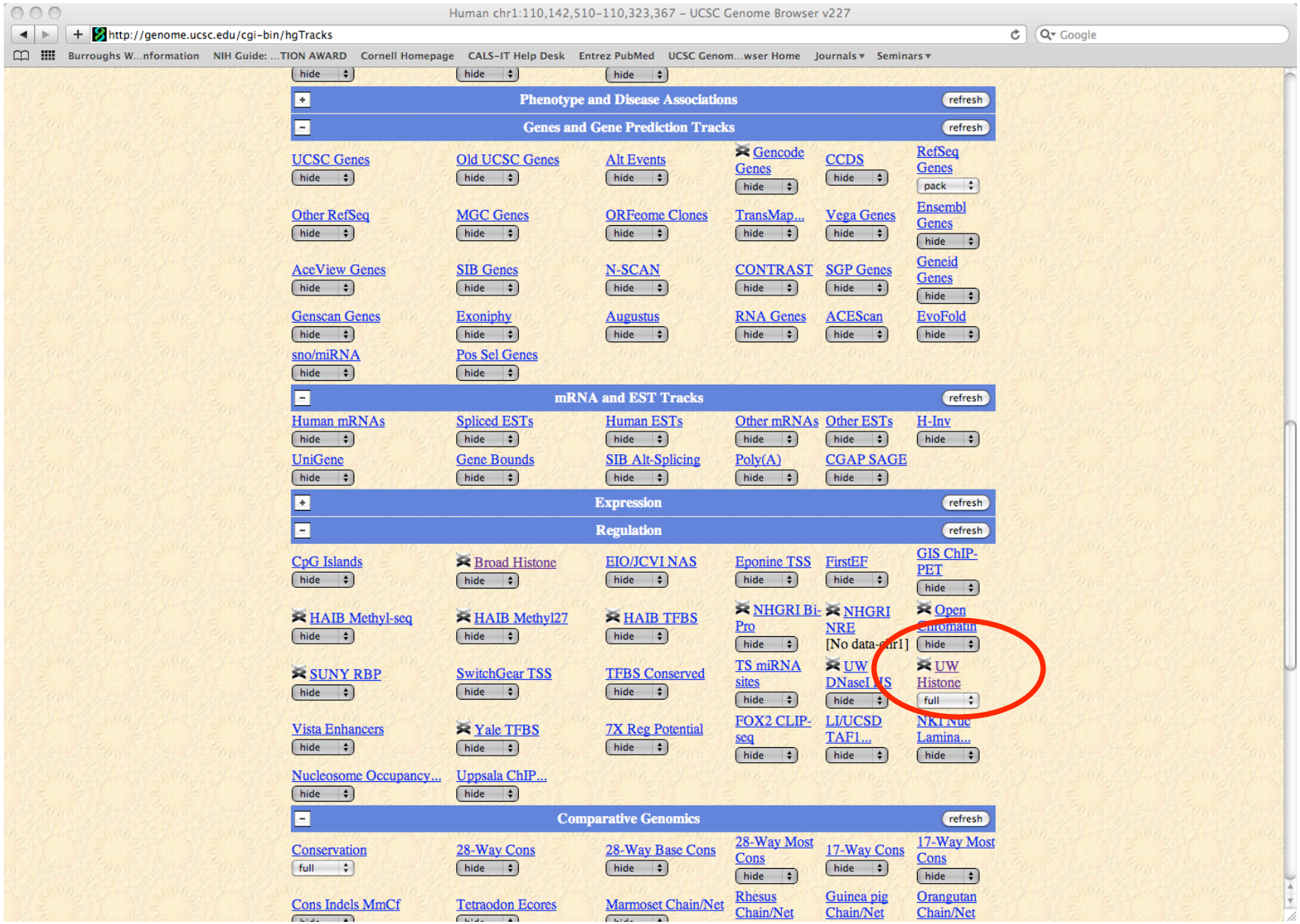
Modifying display settings



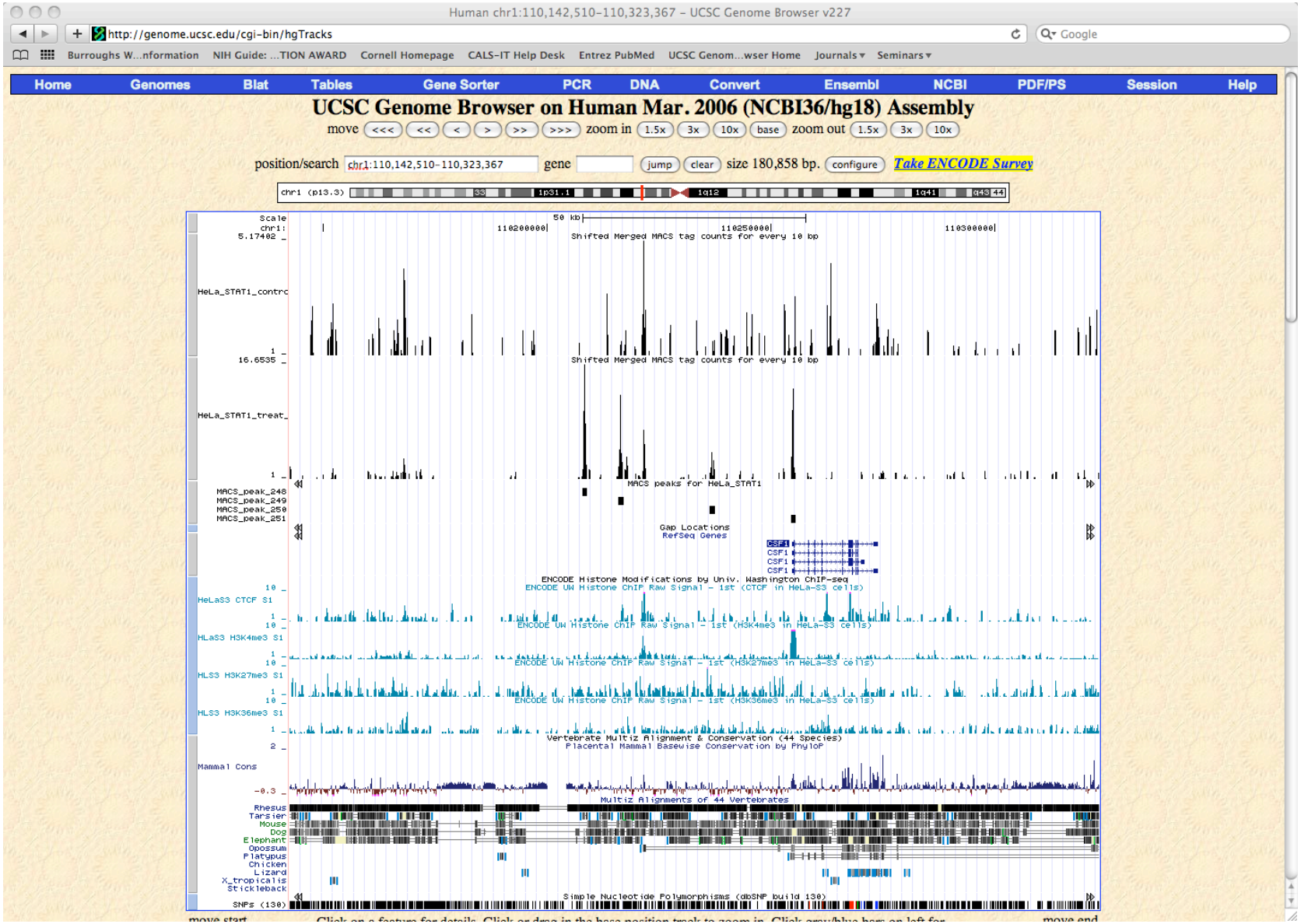
Modifying display settings



Using publicly available tracks



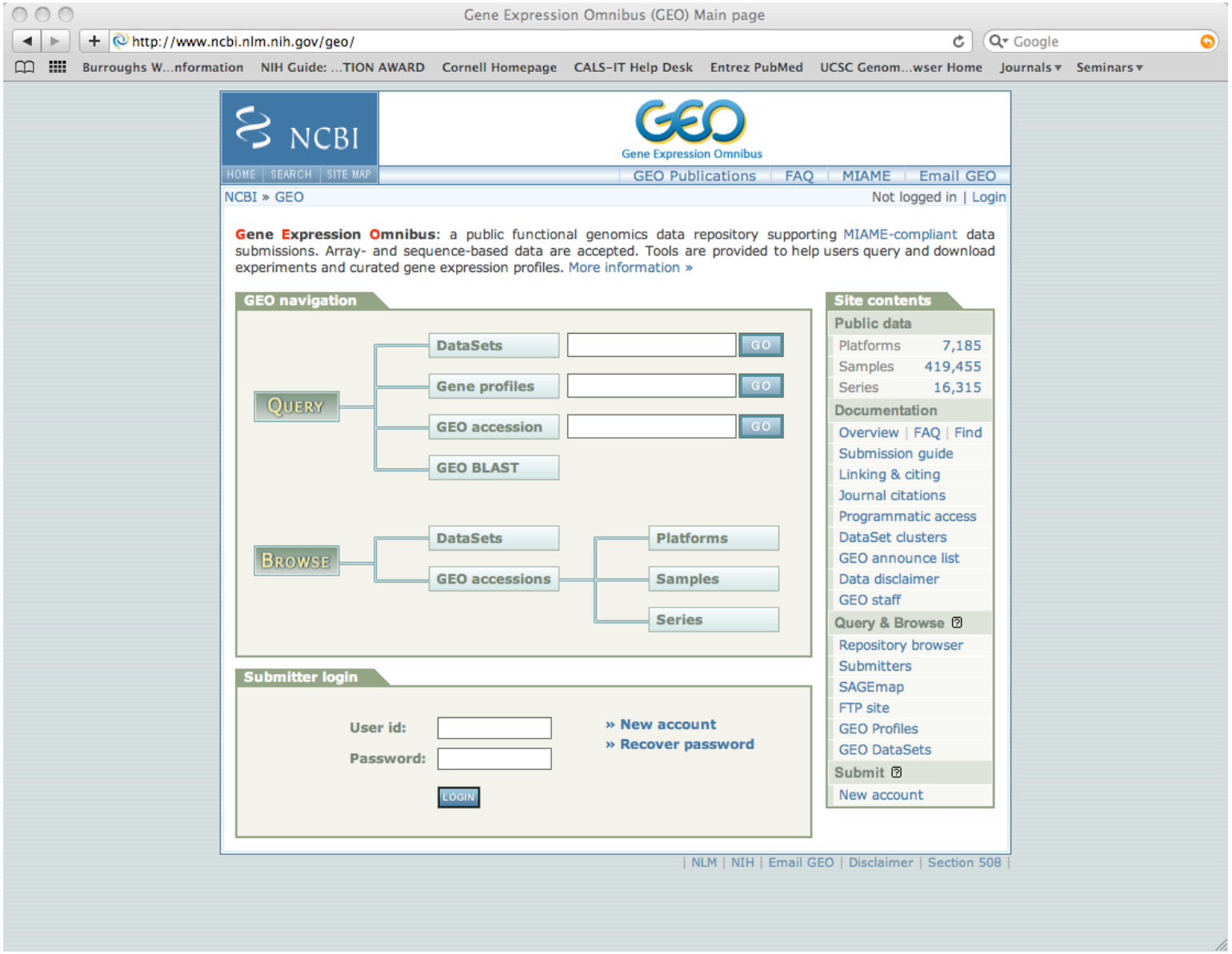
Using publicly available tracks



Galaxy

http://main.g2.mx.psu.edu

NCBI Gene Expression Omnibus (GEO)
http://www.ncbi.nlm.nih.gov/geo/



ChIP-seq exercise

Use MACS software to call peaks in STAT1 ChIP-seq experiment in human HeLA S3 cells after inteferon-γ stimulation. Analyze diagnostics of run and upload data to UCSC genome browser to look at results.

Office hour:

Friday, April 2

1pm - 2pm

102 Weill conference room