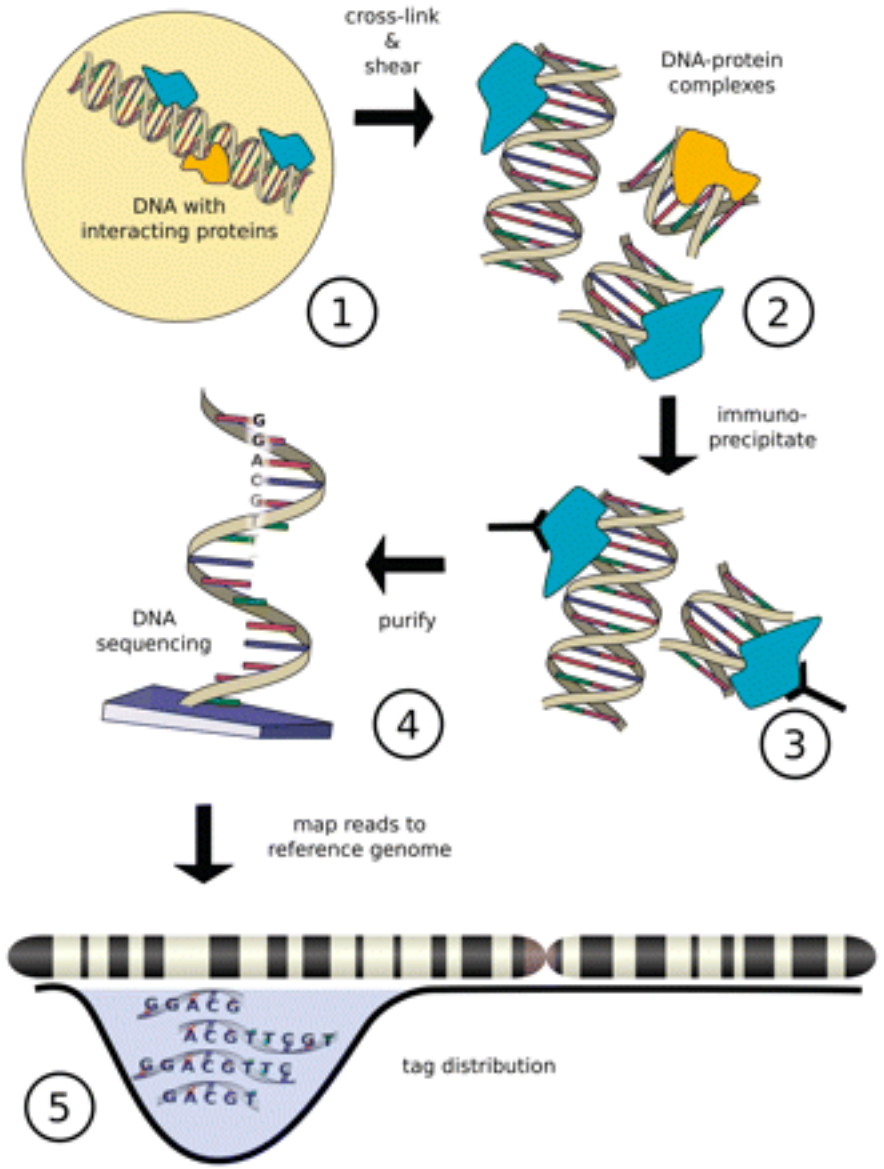
Statistical Methods for Sequence Data

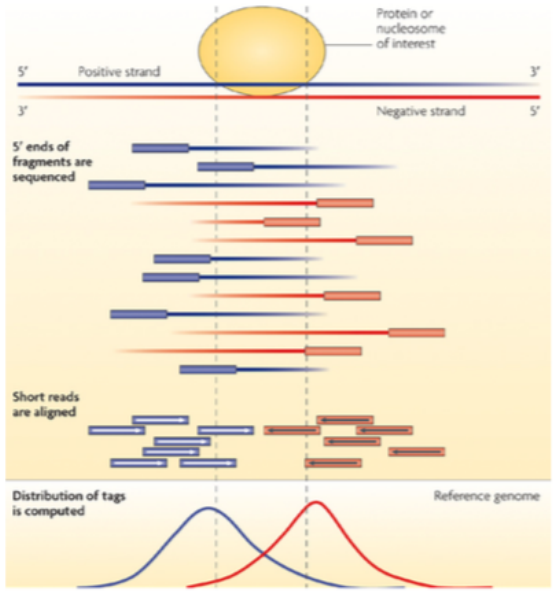
ChIPseq overview

- the protein of interest (POI) is cross-linked with the DNA with formaldehyde fixation that is reversible with heat.
- complexes are filtered out of the set of DNA fragments, using an antibody specific to the POI
- generate fragments by cutting or shearing the DNA
- Reverse cross-linking
- Purify, select for size, and (possibly) **amplify** the fragments



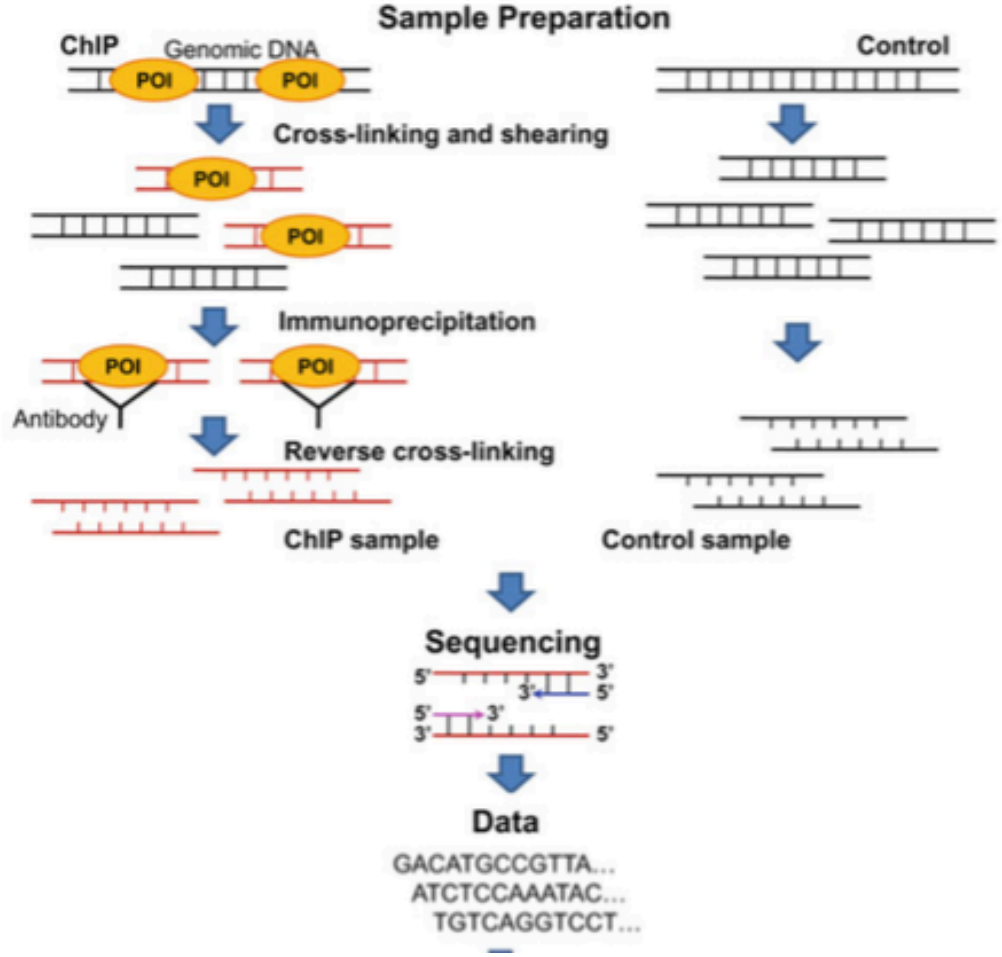
ChiPseq read distribution

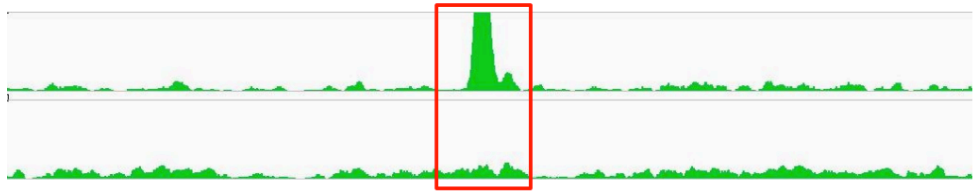
- Bound protein of interest is attached to a fragment that is several hundred nucleotides
- With ChIP-seq, the alignment of the reads to the genome results in two peaks (one on each strand) that flank the binding location of the protein or nucleosome of interest.



Background sample

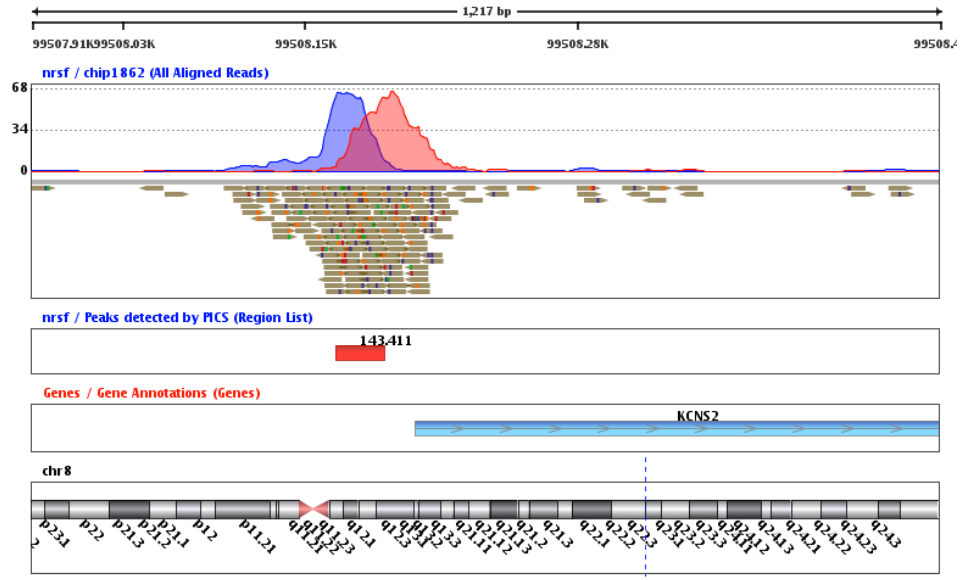
- The distribution of reads along the genome is non random even in the absence of any selection
  - General accessibility—more euchromatin
  - Sequence specific library prep effects-GC bias effects
  - Mappability-being able to align reads to the genome uniquely
- Most experimental protocols involve a control sample that is processed the same way as the test sample except that no specific antibody is used to enrich the bound protein. This serves to be able to calculate the background distribution.





Peak finding

- We would like to find regions that were enriched with the IP procedure—these are the regions that are likely to be bound by our protein of interest
- Model-based Analysis of ChIP-Seq data (MACS)
- one of the most commonly used peak finders.
- explicitly models fragment size
- the larger the fragment size, the higher the average coverage of the genome is, which has a direct influence on the calculation of the estimated significance threshold



MACS-statistical peak calling

- Use the input (background) sample to estimate local background read mapping rate (λ) at different windows
- The null hypothesis is then $\lambda_{local} = max(\lambda_{BG}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k})$
- λ_{BG} is calculated over the entire genome, and $\lambda_{1k}, \lambda_{5k}, \lambda_{10k}$ are calculated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample.
- This value is used to calculate enrichment significance using a **Poisson distribution**
- The ratio between the ChIP-Seq tag count and λ_{local} is reported as the fold enrichment

Poisson distribution

- Single parameter distribution: λ or rate
- Derivation– start with a binomial distribution (number of k success in n trials of probability p) and take the limit

If

$$n \to \infty, p \to 0, \text{ such that } np \to \lambda$$

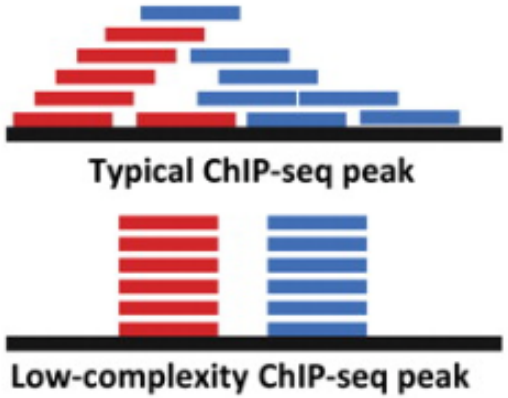
then

$$\frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}.$$

Library complexity

- Read duplication: many identical reads
- Often result from low material and high level of amplification
- One strategy is to remove duplicated reads before further processing
- Obviously, the deeper one sequences, the more likely it is to obtain duplicate reads for biological reasons
- NRF = #unique start positions of uniquely mappable reads #uniquely mappable reads

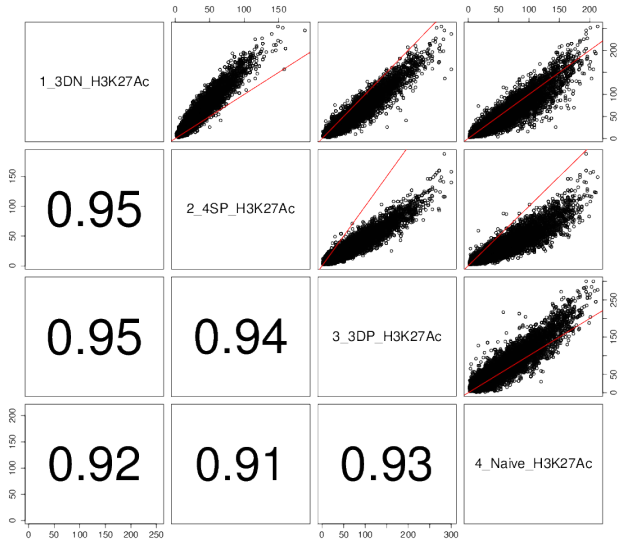
$$\mathrm{NRF} = \frac{\#\text{unique start positions of uniquely mappable reads}}{\#\text{uniquely mappable reads}}$$



- Note that NRF decreases with sequencing depth,
- ENCODE recommends target of NRF ≥ 0.8 for 10 million uniquely mapped reads

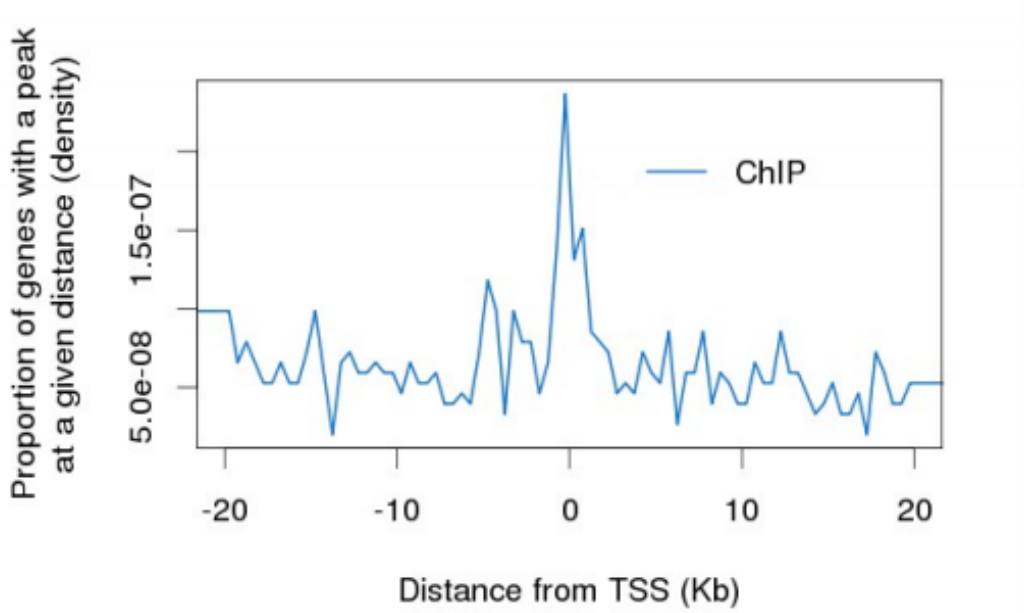
Normalization and differential enrichment

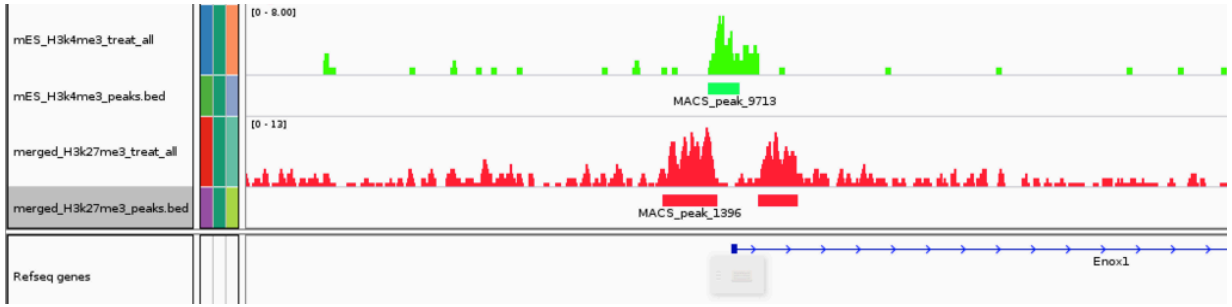
- So far we have identified enriched regions- what about looking for differences between samples
- Fold enrichment values can have very different distributions
- ChIPDiff (Xu. et al., Bioinformatics 2008)
  - statistical model : HMM (1 = not enriched; 2 = enriched in sampleA; 3 = enriched in sample B)
  - does not need pre-defined peaks/regions
- ChIPnorm (Nair et al., PLOS One 2012)
  - quantile normalization of enriched-significant bins in both samples
  - requires signal and input datasets for both samples
- MAnorm (Shao et al., Genome Biology 2012)
  - MA based normalization of regions containing common peaks
  - requires a priori defined peaks for each library



From peaks to function

- where do peaks localize ?
- proximal to TSS
- distal (= enhancer) regions ?
- what are the closest genes (potential targets) ?
- is there a functional enrichment (e.g. GO categories) in genes/regions bound ?
- Statistical peak calling can only tell you so much: it is always a good idea to look at the data in a genome browser
  - Do the peaks have the shape and distribution that you expect





Motif finding

- If our ChIP protein of interest has sequence binding specificity we would like to know what it is
- HOMER (Hypergeometric Optimization of Motif EnRichment) is a set of tools for DNA motif discovery and over-representation analysis.
- Finds motifs enriched in sequences relative to genomic DNA background control.
- Controls for CpG content and low complexity regions
- *De novo* motifs are then reduced to non-redundant set and compared against known motif
- Home also does many other steps of ChIPseq analysis

Information for motif1

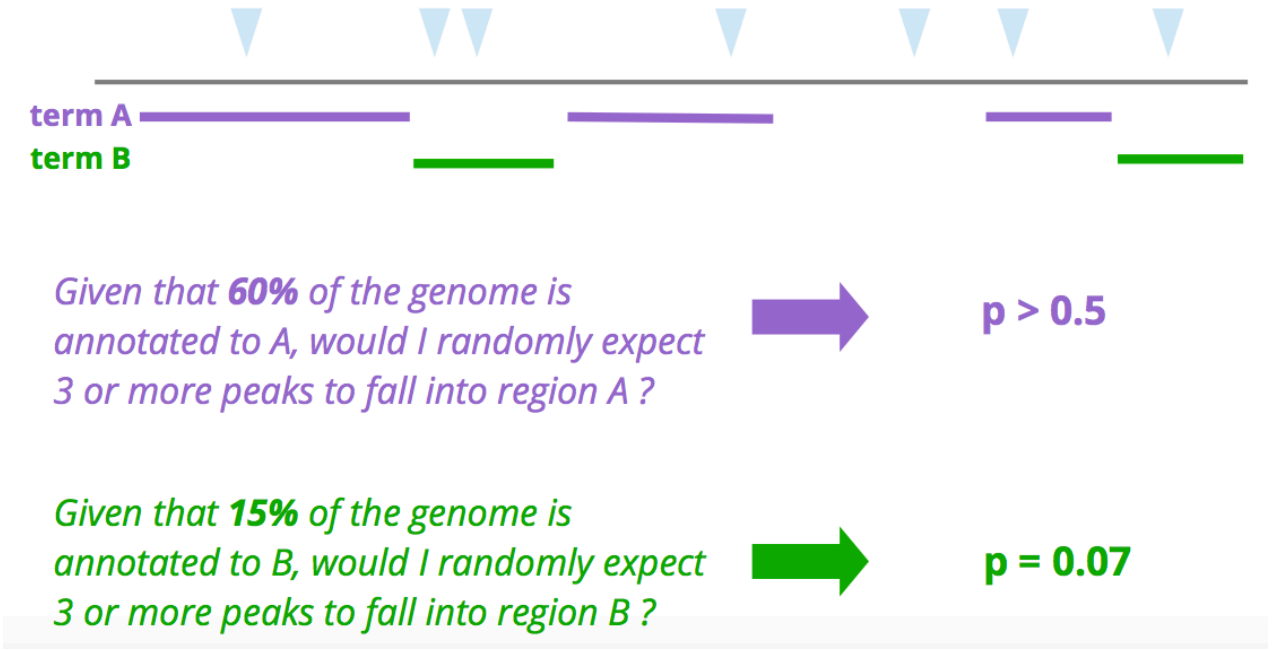
Reverse Opposite:

p-value:	4.941e-324
log p-value:	-7.441e+02
Total Number of Sequences:	49999.0
Total Number of Target Sequences:	2876.0
Total Instances of Motif:	2072.4
Total Instances of Motif in Targets:	685.0
Motif File:	file (matrix) reverse opposite
PDF Format Logos:	forward logo reverse opposite

From peaks to function-gene centric analysis

- Problem: we know little about function (phenotype/disease association) of arbitrary genomic regions that may contain peaks
- We know quite a bit about genes
- Solution: gene-centric analysis
  - Search for nearest gene/genes within a window
  - Perform enrichment analysis based on the genesets
  - restricting to proximal regions discards a large number of binding events
  - "nearest gene" approach introduces bias towards genes with large intergenic regions—these may differ in function from genomic background

Gene/genome centric analysis



GREAT improves functional interpretation of cis-regulatory regions" McLean et al. Nat. Biotech. (2010)

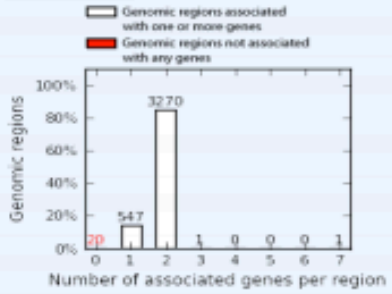
GREAT output

Job description

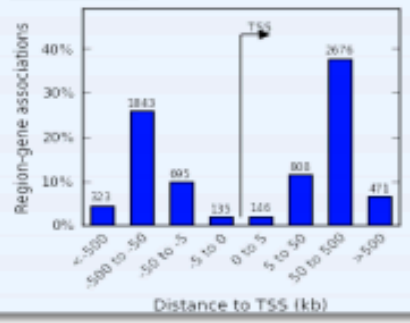
Job ID:	20111012-public-uXBYVG
Display name:	GSM348066_limb_p300_peaks.NEW
Test set:	GSM348066_limb_p300_peaks.NEW.bed (3,839 genomic regions)
	Show in UCSC genome browser. *What is this?*
Background:	Whole genome background
Assembly:	Mouse: NCBI build 37 (UCSC mm9, Jul 2007) *What gene set does GREAT use?*
Associated genomic regions:	Basal+extension (constitutive 5.0 kb upstream and 1.0 kb downstream, up to 1000.0 kb max extension). Curated regulatory domains are included.
	20 of all 3,839 genomic regions (0.5%) are not associated with any genes.
	View genomic region-gene associations. *What is this?*
	Revise the region-gene association rule. *What is a region-gene association rule?*

Region-gene association graphs:

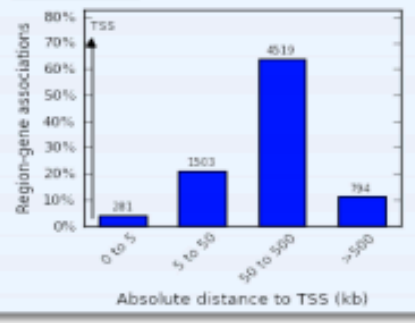
Number of associated genes per region — Download as PDF.



Binned by orientation and distance to TSS — Download as PDF.



Binned by absolute distance to TSS — Download as PDF.



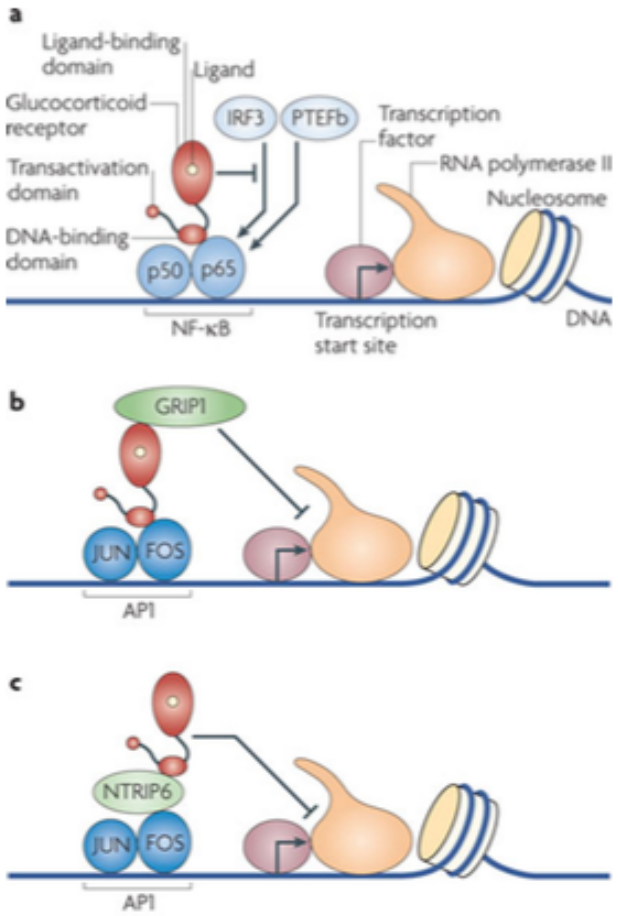
Mouse Phenotype — Global Controls

Table controls: Export | Shown top rows in this table: 20 Set | Term annotation count: Min: 1 Max: Inf Set

Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
abnormal limbs/digits/tail morphology	2	**2.0559e-91**	**6.6837e-88**	2.1465	780	20.32%	6	**2.5295e-40**	2.2020	278	681	8.31%
abnormal craniofacial morphology	3	**9.3822e-91**	**2.0334e-87**	2.0082	887	23.10%	10	**8.9231e-36**	2.0382	297	786	8.88%
abnormal limb morphology	5	**2.4990e-80**	**3.2497e-77**	2.3077	604	15.73%	9	**7.4787e-37**	2.4541	202	444	6.04%
abnormal appendicular skeleton morphology	10	**3.0255e-70**	**1.9672e-67**	2.3450	517	13.47%	17	**3.9549e-30**	2.4098	172	385	5.14%
abnormal skeleton extremities morphology	12	**3.2687e-69**	**1.7711e-66**	2.3724	499	13.00%	21	**7.0557e-29**	2.4222	163	363	4.87%
abnormal paw/hand/foot morphology	13	**4.0300e-69**	**2.0156e-66**	2.6813	404	10.52%	23	**5.4918e-28**	2.7186	126	250	3.77%
abnormal head morphology	14	**6.4657e-67**	**3.0029e-64**	2.0134	672	17.50%	25	**2.9042e-27**	2.0562	223	585	6.67%
abnormal digit morphology	18	**1.0543e-61**	**3.8084e-59**	2.6982	358	9.33%	36	**1.2033e-25**	2.7998	109	210	3.26%
abnormal cartilage morphology	23	**7.3728e-58**	**2.0843e-55**	2.3432	430	11.20%	29	**1.1337e-26**	2.5089	140	301	4.19%
abnormal skeleton development	24	**3.5769e-56**	**9.6904e-54**	2.0833	530	13.81%	38	**5.2377e-25**	2.1414	185	466	5.53%
abnormal long bone morphology	25	**4.6593e-56**	**1.2118e-53**	2.3374	419	10.91%	43	**4.9983e-24**	2.3823	140	317	4.19%

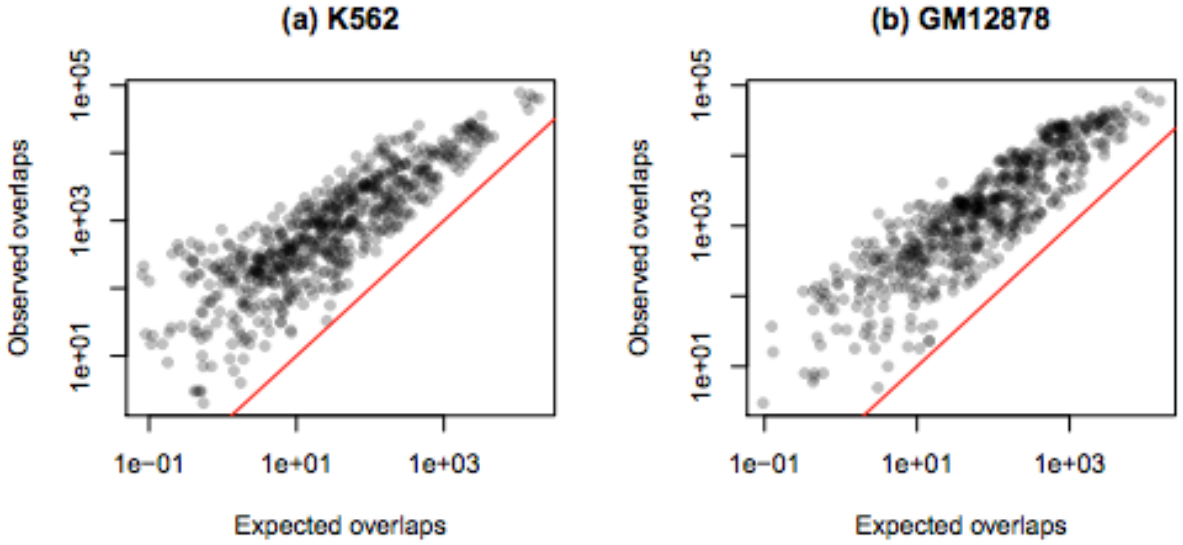
Cross dataset analysis

- Transcription factors have the ability to regulate gene expression by binding directly to DNA at sequence-specific response elements or by tethering to other response elements through protein-protein interactions with other DNA-bound factors
- The combinatorial usage of these response elements drives the regulation of target genes and ultimately determines stimulus and tissue specificity.



Comparing peak locations across different datasets

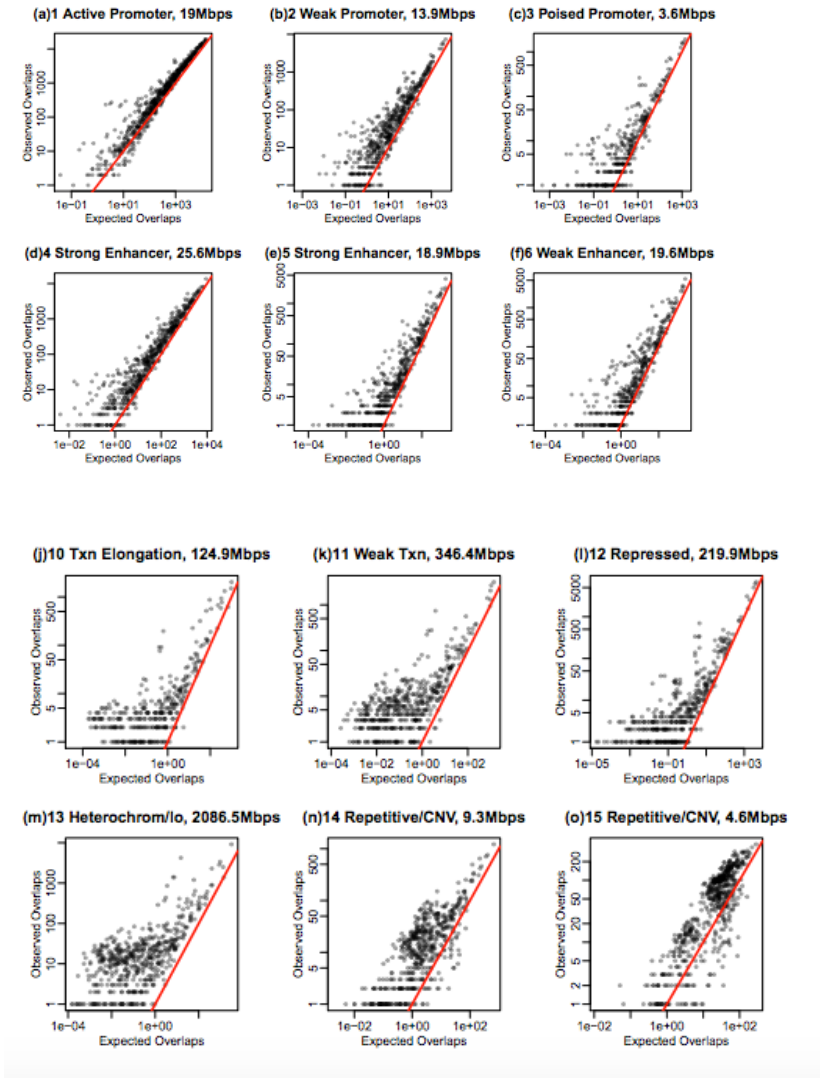
- Simple idea:
  - Bin the genome
  - Compute a 2 way contingency table for peak presence/absence for 2 different ChIPseq experiments
- Big problem: genome wide everything looks non-randomly associated: why?
  - Transcription factors tend to bind to promoters, open chromatin regions etc
  - Non-random binding makes all pairs look non-randomly associated even in the absence of biologically relevant interaction



Each point is a comparison of 2 ChIPseq experiments

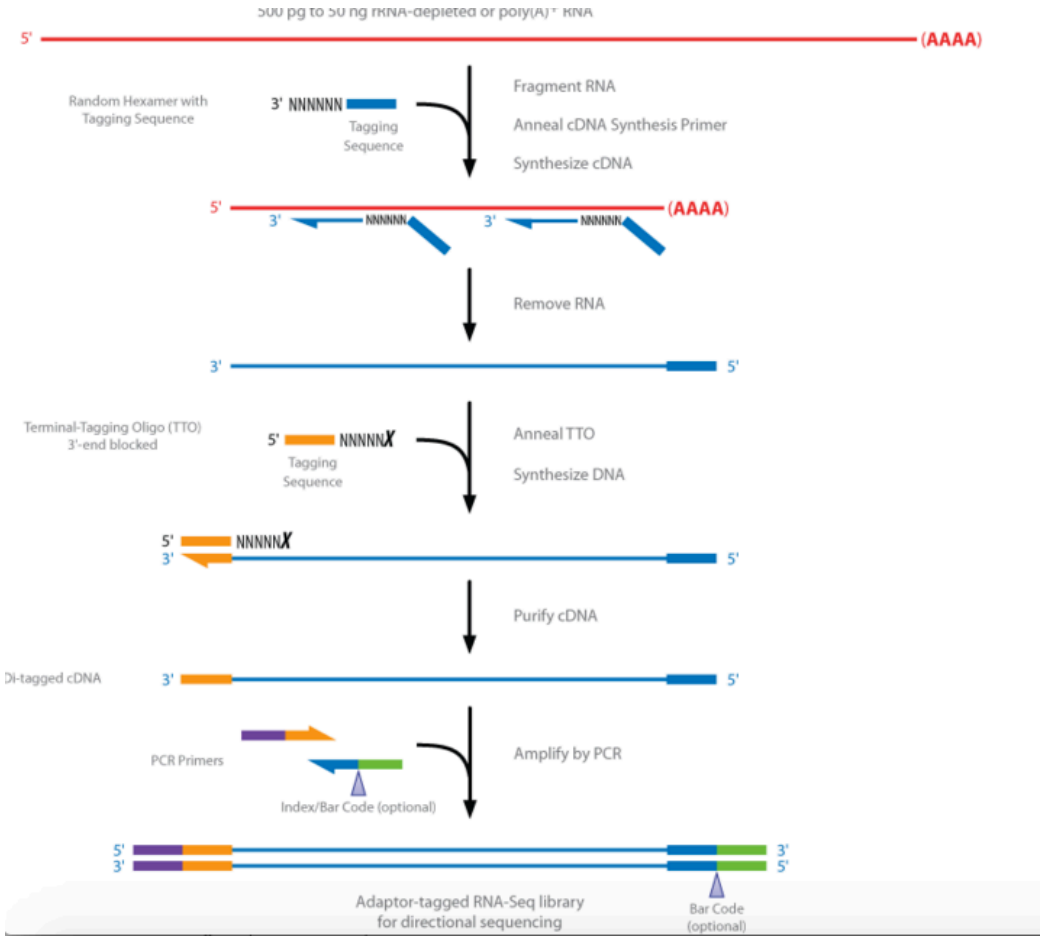
Solution: correct for global genome structure

- Restrict to a subset of the genome: such as promoters
- Segment the genome into different functional regions
- Measuring the spatial correlations of protein binding sites Yingying Wei, Hao Wu
  - Segment the genome based on chromoHMM



RNAseq—quantifying gene expression by sequencing

- Most cellular RNA is rRNA not mRNA
  - **polyA enrichment**
  - **Total RNA** with rRNA depletion—if we are interested in other non-polyA transcripts such as some lncRNAs
- **Multiplexing**—a bar code is added to each fragment so samples can be pooled together for sequencing
- **Single end** -sequence one end of the fragment
- **Paired end** -sequence both ends of the fragment-can be useful for building transcript structure as the distance between fragments is approximately known
- Even though genes are transcribed from specific strand information is typically lost
- **Stranded protocols** are used if strand information is needed—can be used to disambiguate overlapping genes from opposite strands



Advantages of quantification by sequencing

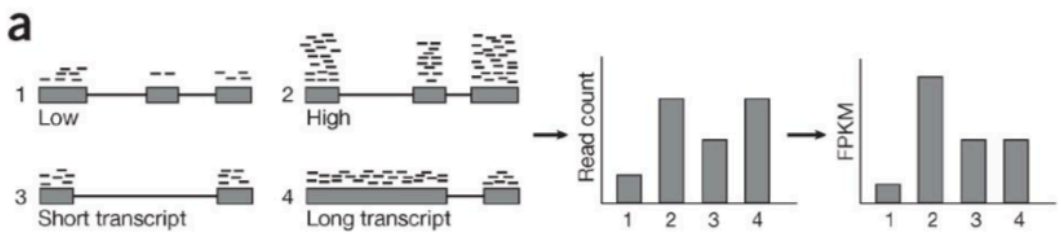
- **Unbiased detection of novel transcripts-** does not require species- or transcript-specific probes. It can detect novel transcripts, gene fusions, single nucleotide variants, etc
- **Broader dynamic range:** With array hybridization technology, gene expression measurement is limited by background at the low end and signal saturation at the high end.

- Most of sequencing real-estate is taken up by very high abundance housekeeping transcripts
  - Ribosomal protein genes
  - Histones
- For low expression genes it is difficult to make statements about biological consequences
  - Are the transcripts really expressed or is it genomic contamination?
  - What about pervasive transcription?
  - Meaningful difference: is 10 reads in one sample different from 2 reads in another sample?
- Increased cost, processing time, statistical complexity

Basic normalization: RPKM

- Number of reads from a transcript depends on:
  - Sequencing depth: total number of mapped reads
  - Length of the transcript
- RPKM (Reads Per Kilobase per Million mapped reads)

$$\text{RPKM} = \frac{\dfrac{\text{number of reads in region}}{\text{region length} \times 10^3}}{\text{total reads} \times 10^6}$$

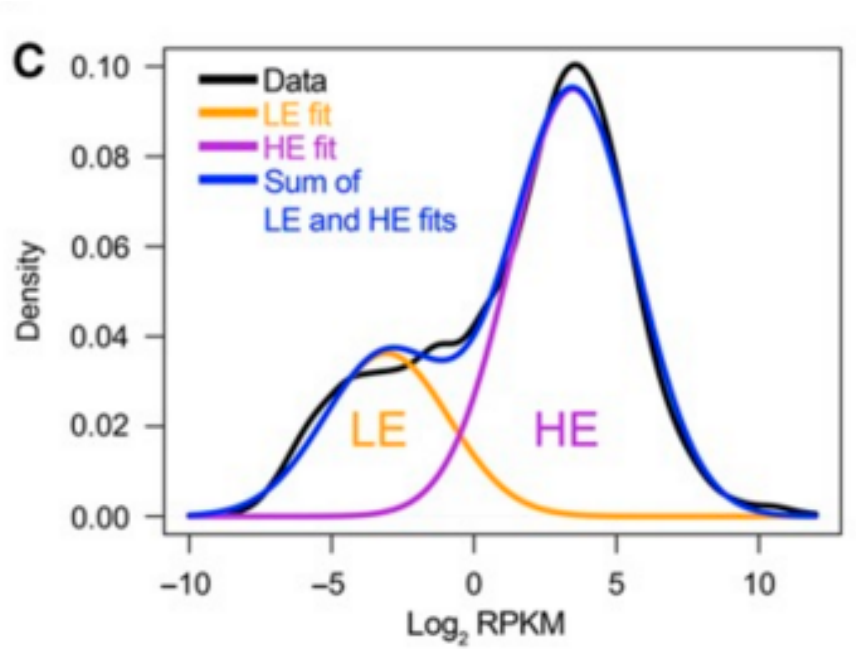


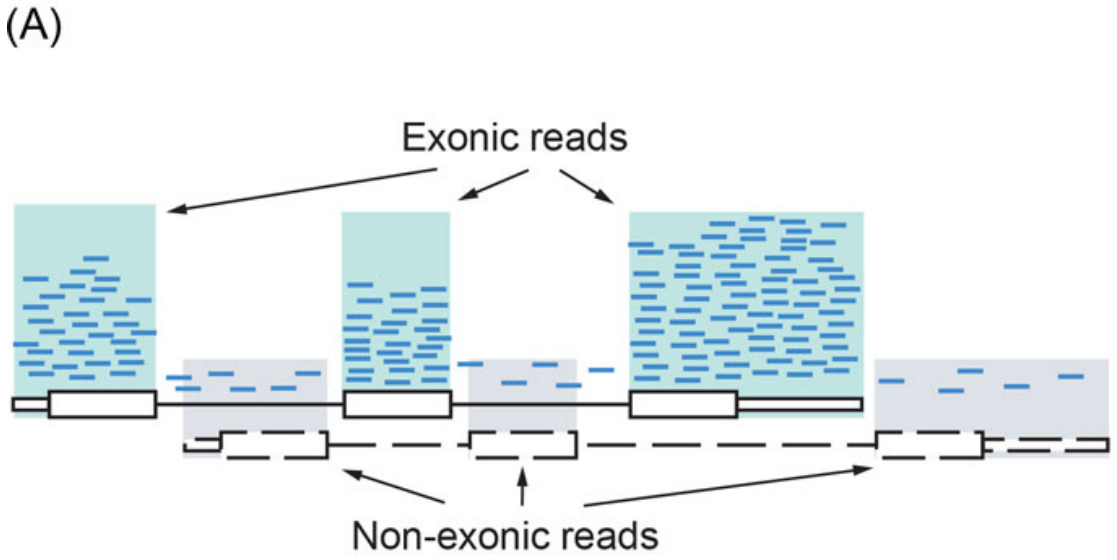
Garber *et al.* (2011) Nature Methods 8:469-477

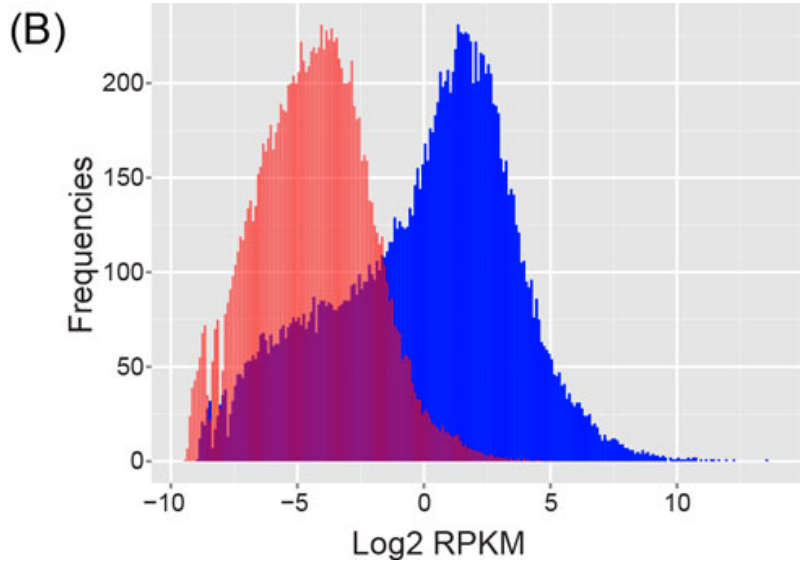
- FPKM (Fragments Per Kilobase per Million mapped reads)
  - Paired end– we have 2 reads per fragment

RPKM distribution

- Often looks bimodal
- Low expression transcripts: what are they?
  - Different model of transcription
  - Genomic contamination
- Often low expression is often filtered at some arbitrary cutoff
- Can be filtered based on spike-ins of know concentration





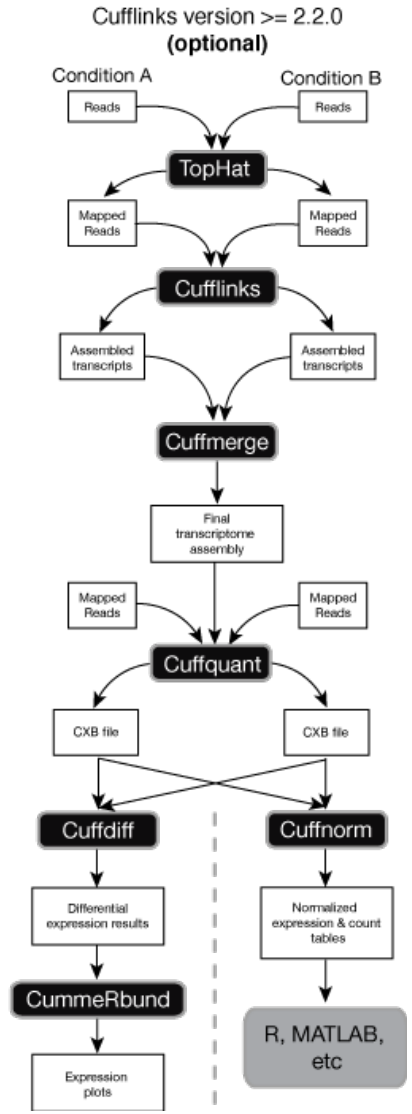


RPKM issues

- Hypothetical example: One sample: we have 10 million reads but 50% of the reads come from a single high-abundance transcript not expressed in the other samples
  - Should the library size be 10 million or 5 million?
- Read depth is affected not only by expression (and length), but also expression levels of other genes
- Especially relevant for samples from different tissues
- Possible solution: use median counts as apposed to total counts

Tuxedo suite for single stop RNAseq processing

- One of the earliest RNAseq pipelines
- Still actively developed
- Very good documentation and support
- Performs
  - Alignment (Tophat – based on bowtie)
  - Transcript assembly (Cufflinks)
  - Quantification
  - Differential expression testing
    - Inference is based on FPKM value
- Advantages: can use information from pervious steps such as ambiguous read mapping
  Disadvantages
  - Limited statistical framework
  - May have have lower power according to some benchmark studies

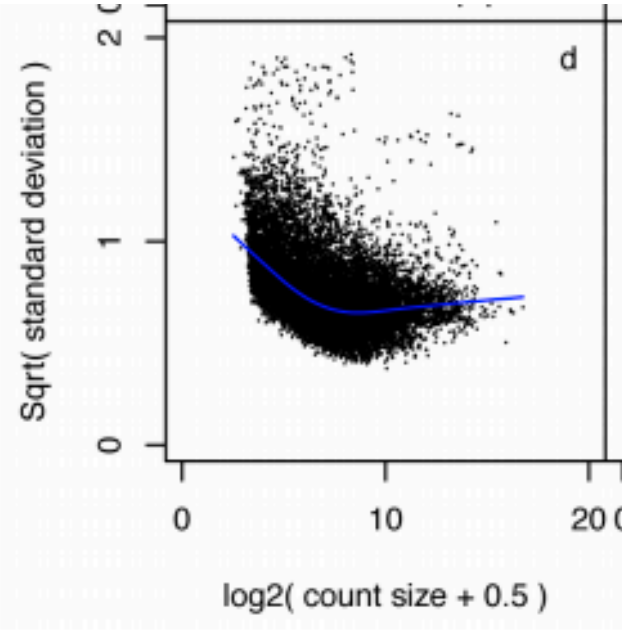


Advanced normalization methods

- RPKM would work perfectly if there were no experimental biases
- Goals of advanced normalization methods
  - Adjust for the possibly arbitrary though monotone function between true abundance and read counts
  - Adjust for systematic biases
- Example: cqn (Hansen et al., 2011): Conditional quantile normalization (CQN) algorithm combining robust generalized regression (corrects for library size, gene length, GC-content)
- Reads are drawn from a Poisson distribution
- Where the rate depends on
  - Library size
  - Systematic effects (GC specific effects, length effects, …)
  - Non decreasing arbitrary function –this is the same as the assumption for normal quantile normalization

Testing for differential expression

- The goal of a DE analysis is to find genes that have changed significantly in abundance across experimental conditions.
- Simple strategy: use log2 RPKM values with standard differential expression approaches such as T-tests or linear models
  - Criticisms: variance is not independent of the mean
- Alternative: use count-based statistics
  - Observation: when comparing the same transcript across different samples there is no need to normalize for transcript length



Modeling count distributions

- Poisson distribution is not appropriate

$$E(x) = Var(X) = \lambda$$

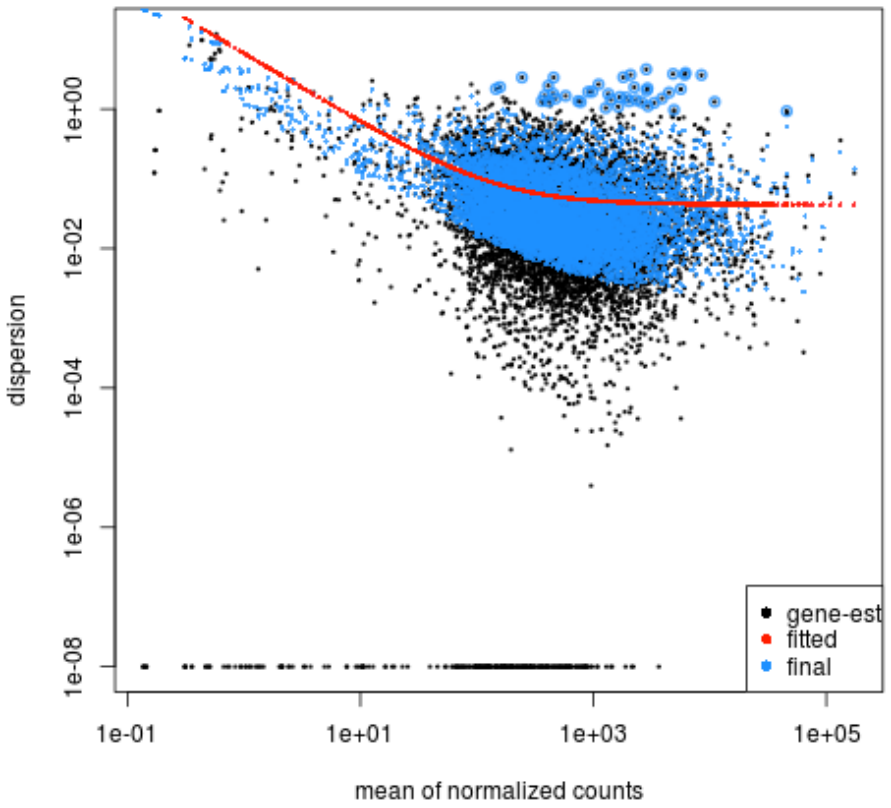
- Mean and variance are not equal: typically we have variance greater than the mean: **overdispersion**
- Negative Binomial(NB) Distribution: number of successes (denoted k) before a specified (non-random) number of failures (denoted *r*) occurs
  - 2-parmeter distribution

$$\Pr(X = k) = \binom{k+r-1}{k} p^k (1-p)^r$$

- Equivalent to Poisson where λ is itself a random variable drawn from the Gamma distribution—also a useful model for simulating random NB data
  - Gamma describes the biological variation—the transcript occurs at some sample dependent rate in the total transcript pool
  - Poisson describes the technical variation- i.e. variation induced by sampling

Estimating NB parameters

- Estimating the parameters is hard
- Parameters estimation is the main difference between methods based on NB distribution
- **baySeq** is based on estimating posterior likelihoods of differential expression via empirical Bayesian methods, assuming negative binomially distributed data.
- **edgeR** determines differential expression using empirical Bayes estimation and exact tests based on a negative binomial model. The dispersion is moderated across genes
- **DESeq(2)** uses similar negative binomial model as edgeR but models the observed relationship between the mean and variance when estimating dispersion, allowing a more general, data-driven parameter estimation. Version 2 uses moderated dispersion



What about complicated designs

- Simple example: we have two groups that were done in batches
- Continuous data: apply linear models
- Count data—linear models are not appropriate
  - Basic linear models: minimizing the squared error is equivalent to maximum likelihood estimation assuming independent Gaussian error
  - Count data (modeled as Poisson or NB) –variance depends on the mean
  - We apply a generalized linear model

$$Y_i \sim \mathrm{NB}(\mu_i, \phi)$$
$$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$$

X – design matrix
ln() – link function
β – parameters

  - ML estimate: Iteratively reweighted least squares method
    - Iterative method—takes longer
    - May not converge

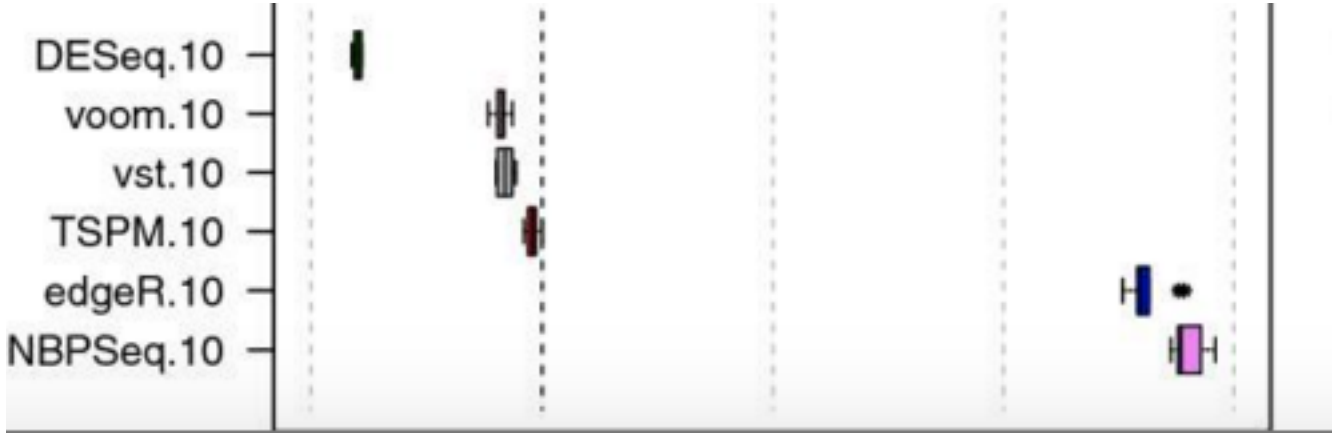
Technical variation with RNAseq-beyond batch effects

PF_BASES	The total number of PF bases including non-aligned reads.
PF_ALIGNED_BASES	The total number of aligned PF bases. Non-primary alignments are not counted. Bases in aligned reads that do not correspond to reference (e.g. soft clips, insertions) are not counted.
RIBOSOMAL_BASES	Number of bases in primary aligments that align to ribosomal sequence.
CODING_BASES	Number of bases in primary aligments that align to a non-UTR coding base for some gene, and not ribosomal sequence.
UTR_BASES	Number of bases in primary aligments that align to a UTR base for some gene, and not a coding base.
INTRONIC_BASES	Number of bases in primary aligments that align to an intronic base for some gene, and not a coding or UTR base.
INTERGENIC_BASES	Number of bases in primary aligments that do not align to any gene.
IGNORED_READS	Number of primary alignments that map to a sequence specified on command-line as IGNORED_SEQUENCE. These are not counted in PF_ALIGNED_BASES, CORRECT_STRAND_READS, INCORRECT_STRAND_READS, or any of the base-counting metrics. These reads are counted in PF_BASES.
CORRECT_STRAND_READS	Number of aligned reads that map to the correct strand. 0 if library is not strand-specific.
INCORRECT_STRAND_READS	Number of aligned reads that map to the incorrect strand. 0 if library is not strand-specific.
PCT_RIBOSOMAL_BASES	RIBOSOMAL_BASES / PF_ALIGNED_BASES
PCT_CODING_BASES	CODING_BASES / PF_ALIGNED_BASES
PCT_UTR_BASES	UTR_BASES / PF_ALIGNED_BASES
PCT_INTRONIC_BASES	INTRONIC_BASES / PF_ALIGNED_BASES
PCT_INTERGENIC_BASES	INTERGENIC_BASES / PF_ALIGNED_BASES
PCT_MRNA_BASES	PCT_UTR_BASES + PCT_CODING_BASES
PCT_USABLE_BASES	The percentage of bases mapping to mRNA divided by the total number of PF bases.
PCT_CORRECT_STRAND_READS	CORRECT_STRAND_READS/(CORRECT_STRAND_READS + INCORRECT_STRAND_READS). 0 if library is not strand-specific.
MEDIAN_CV_COVERAGE	The median CV of coverage of the 1000 most highly expressed transcripts. Ideal value = 0.
MEDIAN_5PRIME_BIAS	The median 5 prime bias of the 1000 most highly expressed transcripts, where 5 prime bias is calculated per transcript as: mean coverage of the 5' most 100 bases divided by the mean coverage of the whole transcript.
MEDIAN_3PRIME_BIAS	The median 3 prime bias of the 1000 most highly expressed transcripts, where 3 prime bias is calculated per transcript as: mean coverage of the 3' most 100 bases divided by the mean coverage of the whole transcript.
MEDIAN_5PRIME_TO_3PRIME_BIAS	The ratio of coverage at the 5' end of to the 3' end based on the 1000 most highly expressed transcripts.

- Various quality control metrics can be collected from the sequence and the alignment
- Tools
  - Picard RNAseq metrics
  - RNA-SeQC
- Often these will correlate with “batch effects”
- Example: Different batches may have different 3’ bias
- We may wish to model the effects of these on gene expression estimates
- For very large datasets we can fit complex models that remove all known technical variation

Problems with discrete distributions

- Mathematically less tractable
- Rely on many assumptions
- Statistical tests may be
  - Only asymptotically valid
  - Theoretically correct when the overdispersion is small
- A comparison of methods for differential expression analysis of RNA-seq data Charlotte Soneson and Mauro Delorenzi
  - Simulation with no differentially expressed genes (all genes are drawn from the same distribution)
  - Fraction of genes with a p-value of <0.05 (we expect 0.05 if the tests are are correct)
- Many simulation studies find that count-based models are overly permissive in calling differential expression

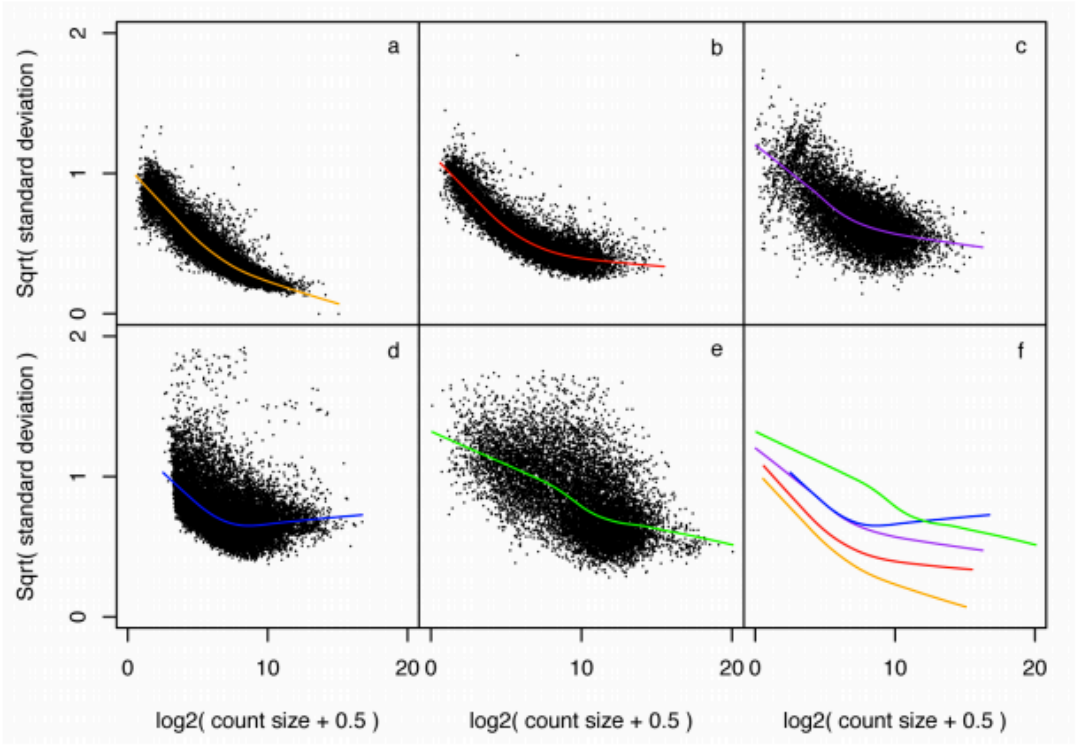


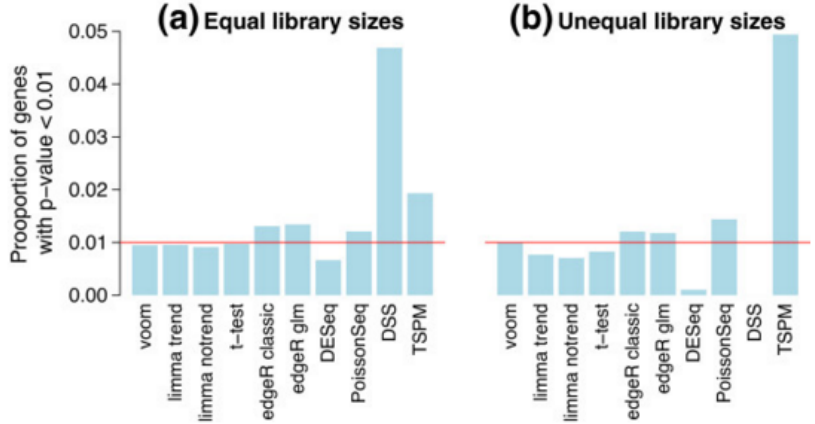
Voom method

- Converts discrete counts to (log-cpm)
- Fits trend in the variance of counts
- Observation: for RNAseq there is no gene-specific variance-depend on the observed counts which depend on sequencing depth of each sample
- Estimate “variances” of individual measurements which are then used as to compute weights for linear models using moderated T statistics
- Much faster than GLM for complex designs

$$\hat{\boldsymbol{\beta}}_g = (X^T \Sigma_g^{-1} X)^{-1} X^T \Sigma_g^{-1} \mathbf{y}_g,$$

where $\Sigma_g = diag(w_{g1}, \ldots, w_{gJ})$ is the diagonal matrix of prior weights.

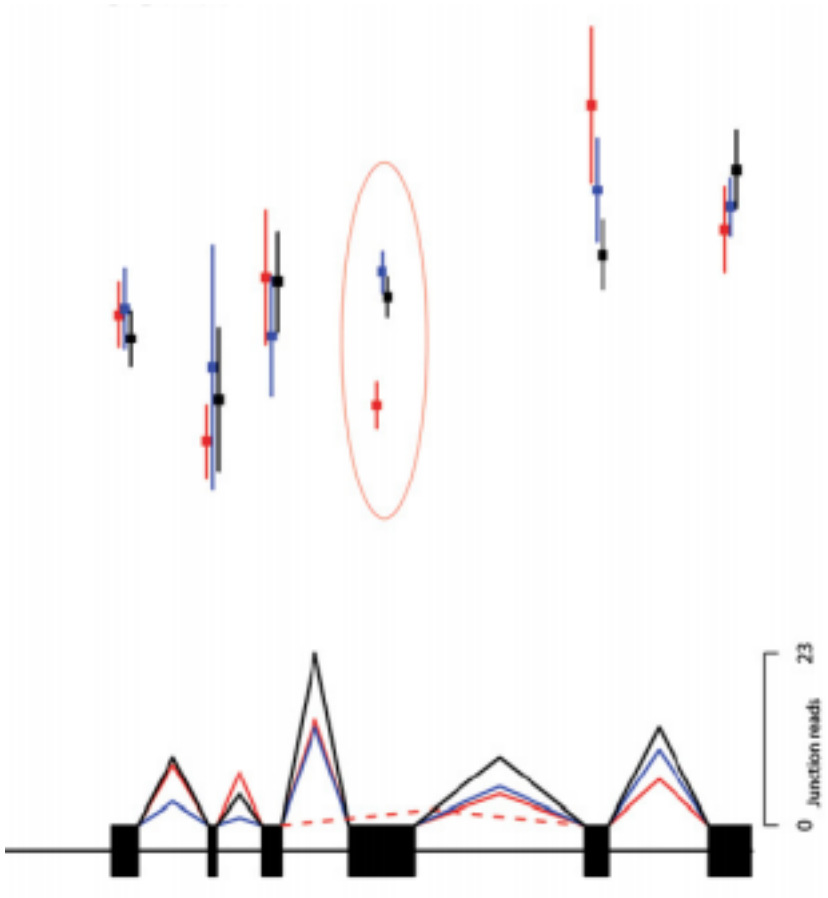




Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Law et al. 2014. Genome Biology.2014, 15:R2

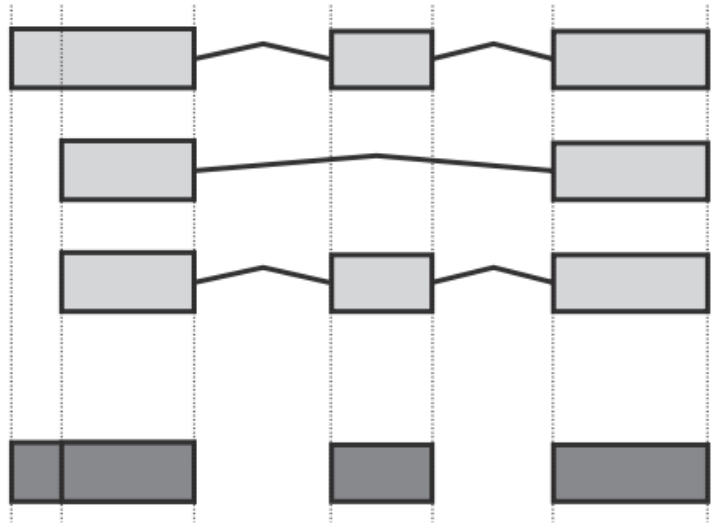
Count based methods for differential splicing

- Looking for genes that have an altered splicing pattern across the samples
- The gene may or may not be differentially expressed on the whole transcript level
- Simple idea: compare the fraction of transcript reads that map to a particular exon



DEXseq—differential splicing detection

- Flatten and bin the gene model
- Model the read counts for an exon bin as an NB distribution that depends on the total number of transcript counts
- GLM

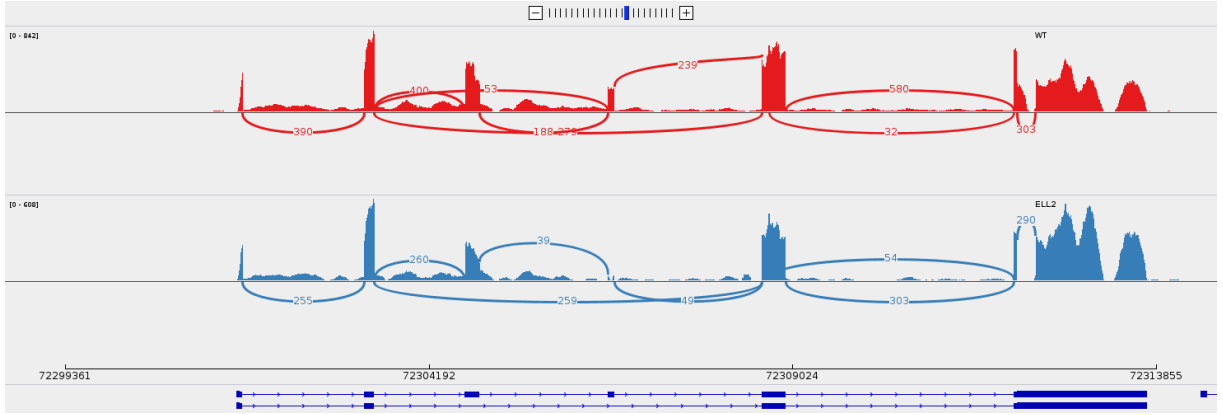


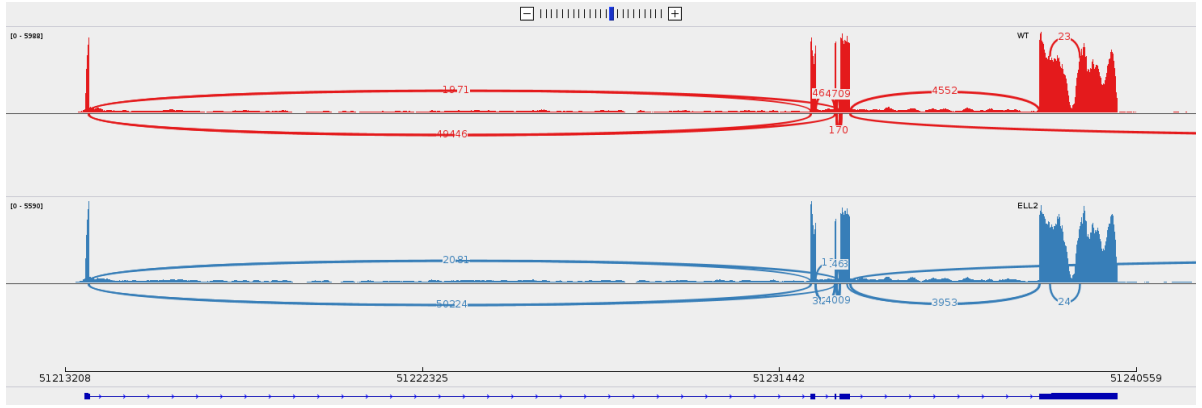
$$\log \mu_{ijl} = \beta_i^{G} + \beta_{il}^{E} + \beta_{i\rho_j}^{C} + \beta_{i\rho_j l}^{EC}.$$

β^G – baseline "expression strength"
β^E – "exon" (bin) effect
β^C – condition effect
β^{EC} – condition x "exon" interaction

Example

- These genes were both significant for differential exon usage with DEXseq
- Which exons are differentially spliced?





Sequence quantification summary

- The field is still developing rapidly
- Methods are constantly refined
- Many comparison studies but may have inconsistent conclusions
- Technology is both improving and becoming cheaper
  - Less noise
  - More samples
  - Will likely need different modelling strategies
- Never hurts to check results in the genome browser