Underlying Statistical Models of Functional Genomics Data Analysis Tools

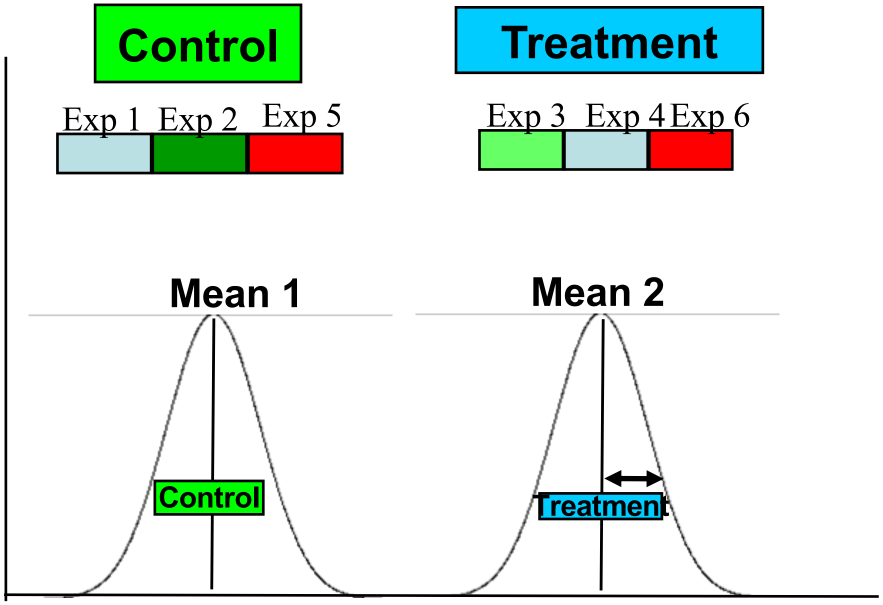
Stefan Bekiranov

BIOC 8145

Outline

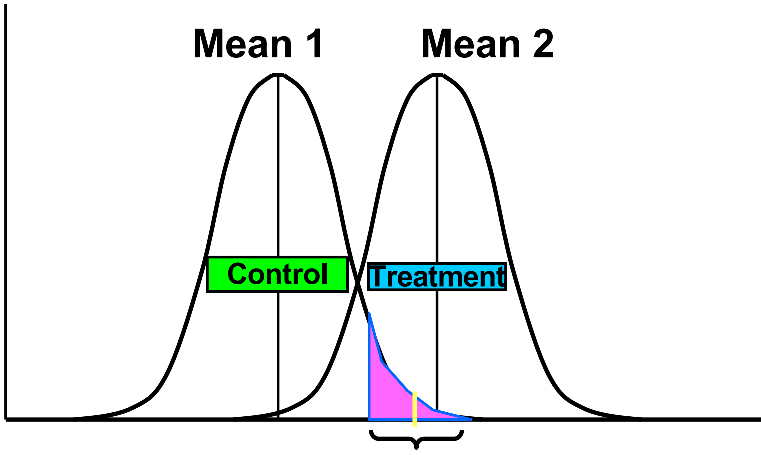
- Classic two sample test:
  - t-test
- Multiple Hypothesis Testing:
  - Bonferroni p-value Correction
  - Benjamini and Hochberg (False Discovery Rate) p-value Correction
- Single Sample Data
  - Binomial Distribution
  - Poisson Distribution
- ChIP-seq Peak Calling: MACS2
  - Poisson Distribution
  - Negative Binomial Distribution
- RNA-seq Differential Expression Analysis: DESeq2
  - Normalization
  - Statistical Inference/Null Model

Two Sample Data: t-test



t-statistic

$$t = \frac{(m1 - m2)}{\mathrm{stddev}(m1-m2)}$$



Calculating Significance of Differential Expression (p-value):

1. Randomly permute control and treatment data and compute t-statistic
2. Generate “null” t-distribution
3. p-value = fraction of random t > true t

In R: t.test()

Multiple Hypothesis Testing

- Assume we randomly split the same population of RNA into two samples and performed 10000 t-tests. Applying a cutoff of p-value < 0.05 would give 500 genes!!
- Fact: Applying t-tests to two populations with the same distributions generates a uniform distribution of p-values.
- Bonferroni correction: p-value* = # tests x p-value
  - Crudely/Naively, p-value* is the expected number of false positives
  - Apply p-value* as your new p-value cutoff
  - Very conservative for functional genomics but not statistical genetics

Multiple Hypothesis Testing (cont’d)

- Assume g genes.
- Two main approaches to defining false positive rates:
  - Family Wise Error Rate (FWER): *Probability* of having one (or more) false positives in the predicted set of genes
  - False Discovery Rate (FDR): *Expected* proportion of the predicted set which consists of false predictions
- If FWER or FDR < α, a procedure controls the FDR or FWER error rate to level α
- Many set α = 0.05
- FWER (5% chance of having >=1 false positive)
  - Single Step Method: All p-values given the same correction
    - Bonferroni correction:
      - Select genes whose p-values $p_i < \alpha/g$
      - Overly stringent for functional genomics but not statistical genetics!
  - Step Down Method: p-values given different correction
    - Westfall and Young
      - Sort p-values in increasing order: $p_{(1)},\ldots, p_{(g)}$
      - If $p_{(1)} < \alpha/g$, then null hypothesis (1) is rejected, go to step 2
      - If $p_{(2)} < \alpha/(g-1)$, then null hypothesis (2) is rejected, go to step 3, continue until
      - If $p_{(j)} > \alpha/(g-j+1)$, then all null hypotheses i=(j),...,(g) are accepted
      - Less stringent than Bonferroni correction
      - Still too conservative for gene expression studies

Multiple Hypothesis Testing (cont’d)

- FDR (5% of the predicted genes are false positives)
  - Benjamini and Hochberg Step-Up Methods
  - Q = V/R; R = # rejected null hypotheses; V = # rejected that are true nulls (false positives); assume g_0 out of g genes are true nulls
  - Theorem: $E(Q) = (g_0/g) * \alpha$
  - Assumes g different tests are independent
  - Sort p-values in increasing order: $p_{(1)},\ldots,p_{(i)},\ldots,p_{(g)}$
  - Let $q_i = i*\alpha/g$, $i=1,\ldots,g$
  - α is the desired FDR
  - Let k be a maximum i such that $p_{(i)} <= q_i$
  - If k >= 1, reject null hypotheses i=(1),…,(k) and accept all others
  - Note: There may be i’ < k such that $p_{(i')} > q_{i'}$
  - Appropriate for functional genomics studies which are followed by relatively easy experimental validation
  - Default multi-test correction in many Bioconductor packages.

Single Sample Data

- The t-test described above was developed for treatment versus control (2 sample).
- Functional Genomics:
  - In a discovery/characterization phase.
  - Costly and labor/time intensive,
  - High throughput sequencing data (e.g., ChIP-Seq, RNA-Seq, MeDIP-Seq,…) sometimes comes in the form of single samples:
    - No replicates.
    - No comparisons between treatment and control or multiple groups.
- What are the analysis goals?
  - Identify significantly enriched sites above background/noise.
  - Compare sites to annotations.
- How do we identify the significantly enriched sites?
  - Generate a null distribution from the background/noise in the data.
  - Calculate p-values using this distribution.
  - FDR correct the p-values.
  - Apply 0.05 (or 5%) FDR cutoff.

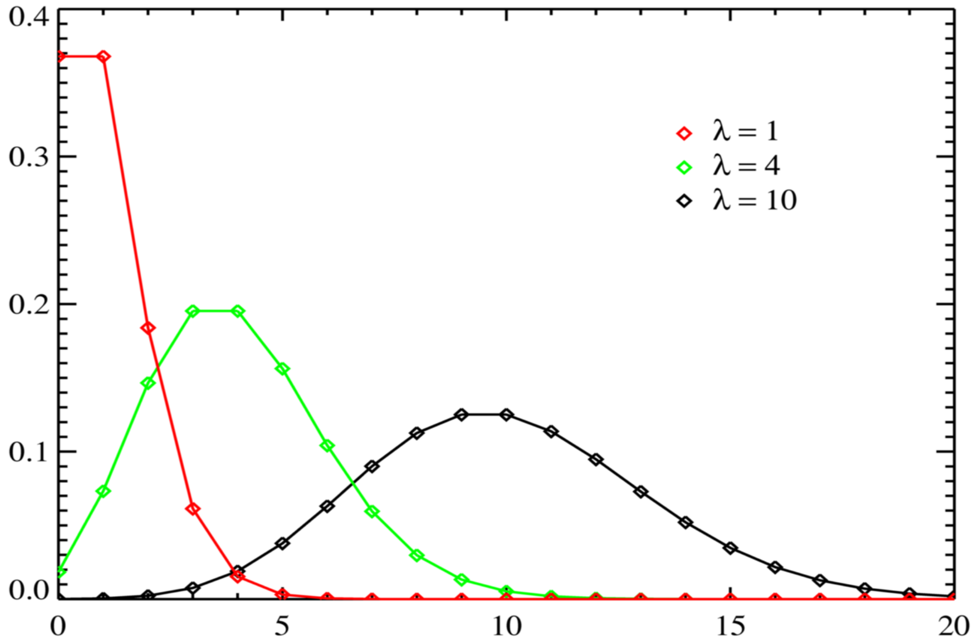
Binomial Distribution

- Suppose that we divide the genome into non-overlapping windows of size w.
- Assume n reads could map to window w_i given our sample.
- Assume the probability of sequencing and mapping a read in w_i is p and not sequencing/mapping it within w_i is 1-p.
- The total number of reads mapping within w_i, X, is a binomial random variable with parameters n and p.
- Note: n! = n * (n-1) * (n-2) * (n-3) * … * 3 * 2 * 1
- The probability that k reads map in w_i or X = k can be found as follows:
  - Any *specific* set of k reads out of n mapping within w_i (one instance out of many!) is given by the probability $p^k(1-p)^{n-k}$ assuming read sequencing/mapping events are independent (multiplication principle of probabilities).
  - The *total number of ways* that k reads out of n (that could be sequenced and mapped to w_i) are sequenced and mapped to w_i is given by n!/((n-k)!k!).
- Thus, the probability of X = k reads being sequenced and mapped to w_i is

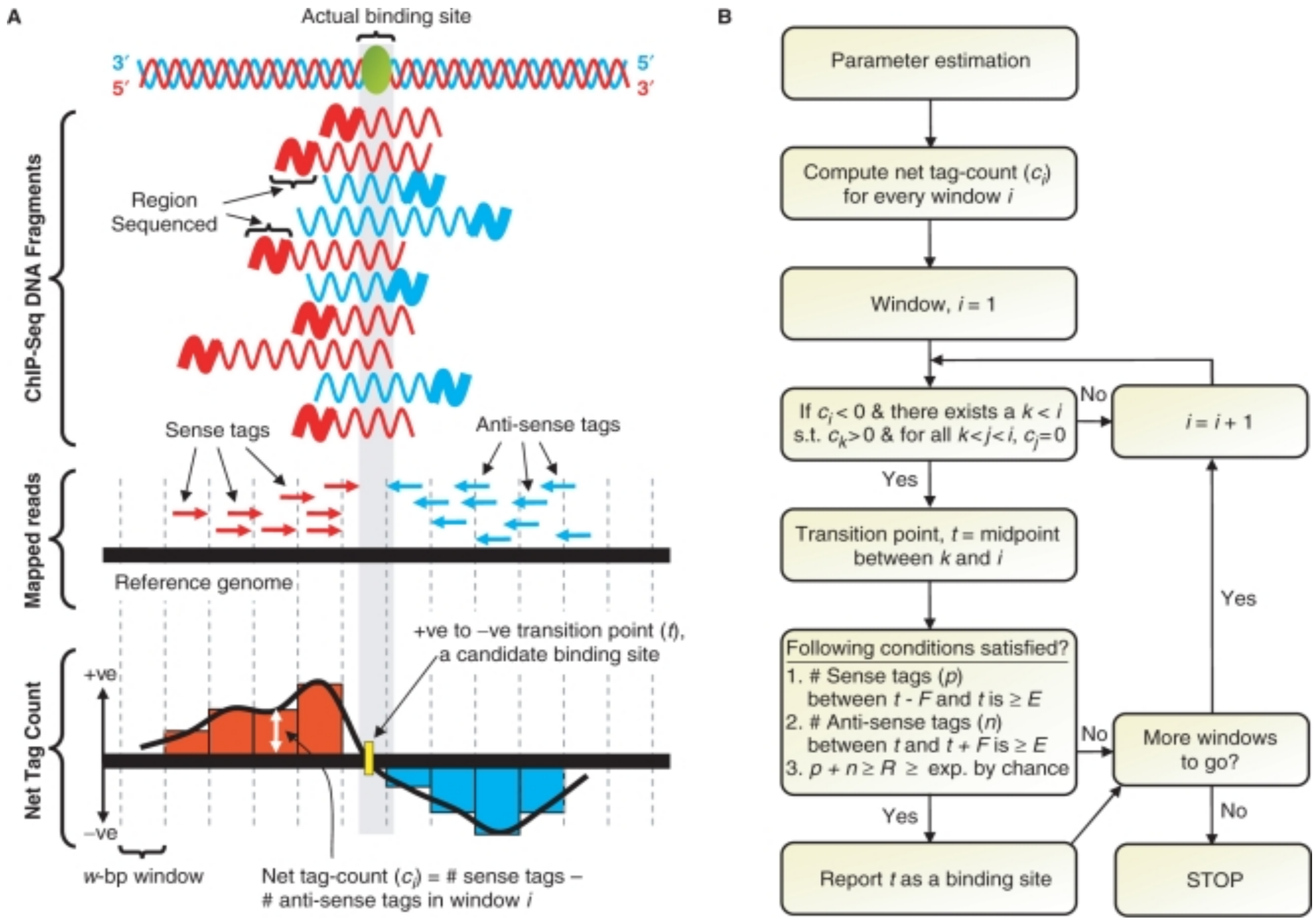
  $P(X = k) = n!/((n-k)!k!)\ p^k(1-p)^{n-k}$.
- Expectation of X is E(X) = n * p
- Variance of X is Var(X) = n * p * (1 – p)
- For n large and p far from 0 or 1; X ~ N(np, np(1-p)) is a good approximation.

Poisson Distribution

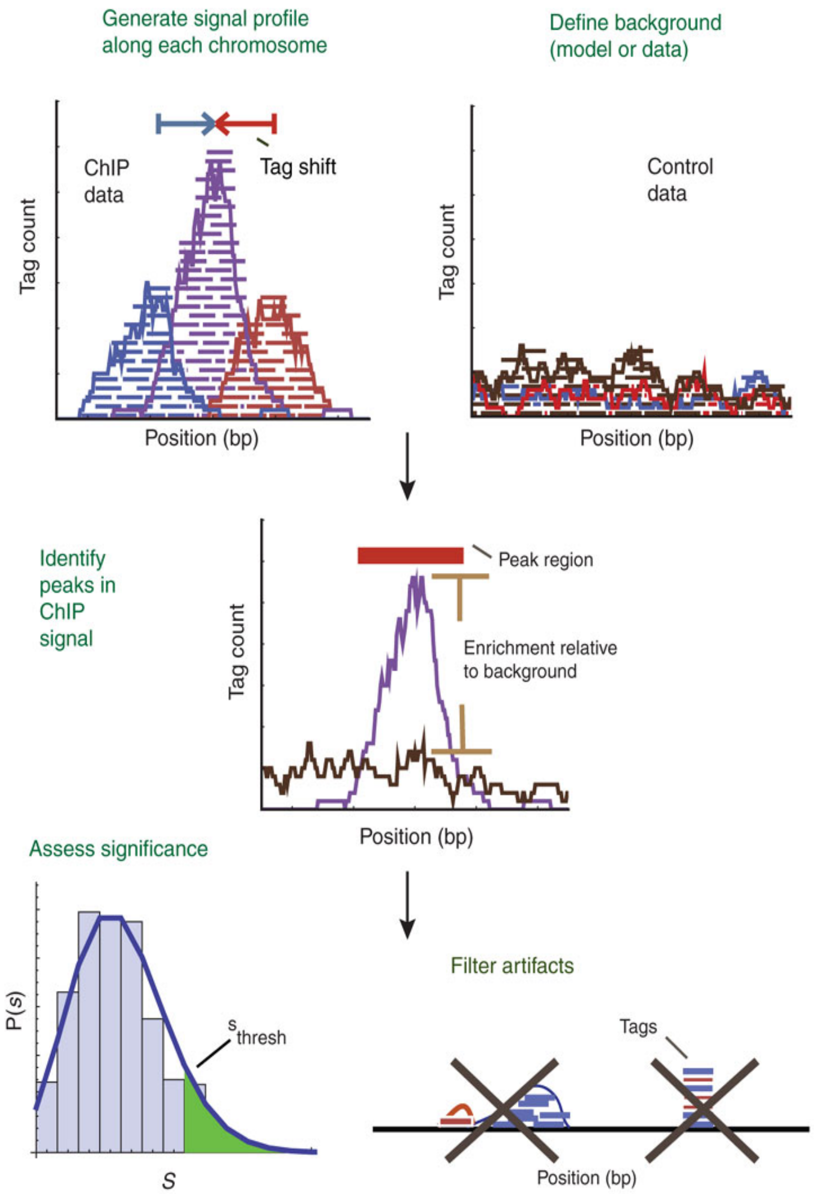
- Derived from the binomial distribution in the limit where
  - n (the number of possible mapped reads) is large
  - p (the probability that a read maps in w) is small
  - $n*p = \lambda$
- $P(X = k) = (\lambda^k/k!)e^{-\lambda}$;
- $E(X) = \lambda$;
- $Var(X) = \lambda$
- In R: dpois(); ppois(); qpois(); rpois()



Identifying Sites from ChIP-seq Data (MACS2)


A
Actual binding site
3′
5′
5′
3′
ChIP-Seq DNA Fragments
Region Sequenced
Sense tags
Anti-sense tags
Mapped reads
Reference genome
Net Tag Count
+ve
–ve
+ve to –ve transition point (t), a candidate binding site
w-bp window
Net tag-count (c_i) = # sense tags – # anti-sense tags in window i
B
Parameter estimation
Compute net tag-count (c_i) for every window i
Window, $i = 1$
If $c_i < 0$ & there exists a $k < i$ s.t. $c_k > 0$ & for all $k < j < i$, $c_j = 0$
No
$i = i + 1$
Yes
Transition point, t = midpoint between k and i
Following conditions satisfied?
1. # Sense tags (p) between $t - F$ and t is $\geq E$
2. # Anti-sense tags (n) between t and $t + F$ is $\geq E$
3. $p + n \geq R \geq$ exp. by chance
No
More windows to go?
Yes
Yes
No
Report t as a binding site
STOP

How do current algorithms identify sites? Available Tools: Pepke, S. et al. Nat. Meth. Supp. 6, pp. S22-S32 (2009)

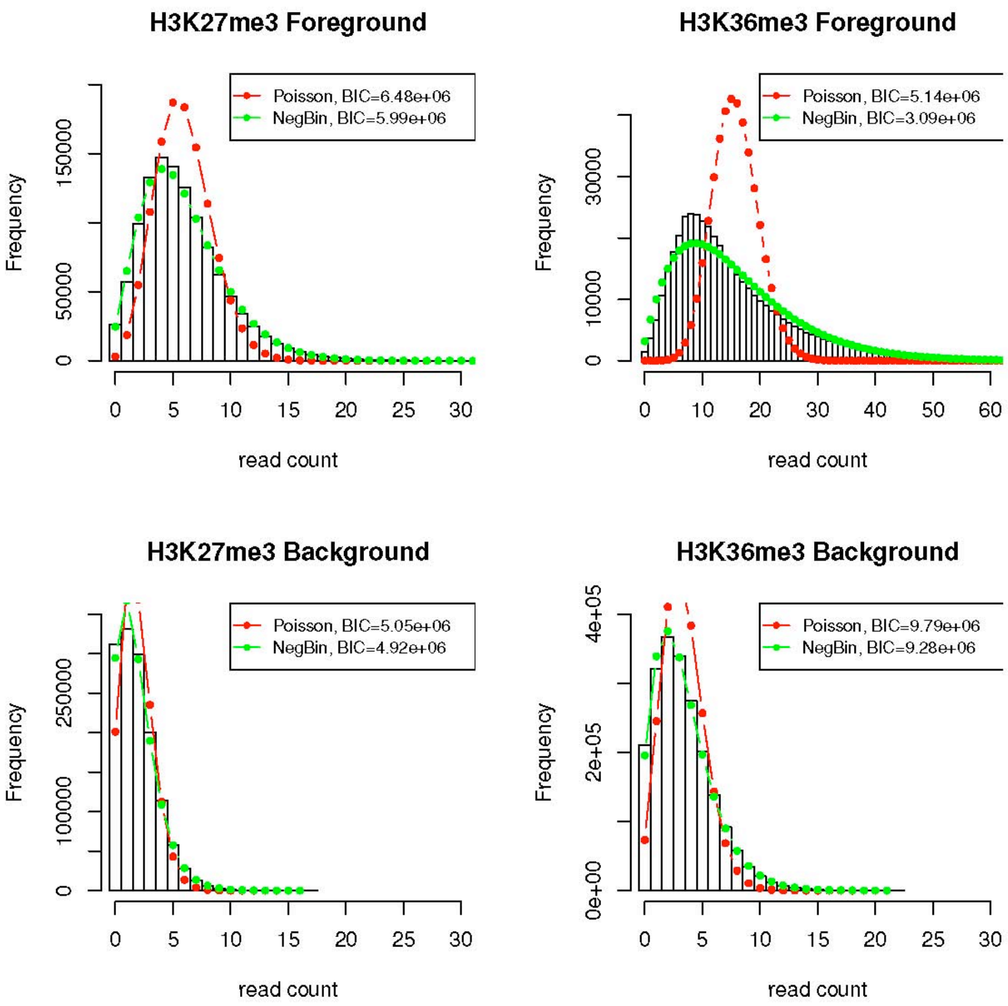


Poisson Background Model

- Bin the genome into windows of size w.
- λ = Expected (or average) number of reads in a window calculated from control or low read count data.
- Calculate p-value for each window.
  - p-value = $\Sigma_{k=R,\infty} e^{-\lambda} \lambda^{k}/k!$
  - R is the number of reads in a given window.
- FDR correct p-values and apply FDR cutoff.

Negative Binomial Distribution Background Model

The Negative Binomial Distribution is a mixture of a Poisson Distribution and a gamma distribution. It can viewed as a Poisson distribution with a variable λ (or expected number of reads in a window) that itself varies as a gamma distribution.



Identify Differentially Expressed Genes from RNA-seq Data using DESeq2

- Assumes read count K_{ij} for gene i in sample j is described by a Generalized Linear Model where:
  - K_{ij} is distributed as a Negative Binomial with mean = μ_{ij} and variance = σ_{ij}^2; Accounts for shot/sampling noise plus additional experimental and biological variation
  - Normalized read counts = $q_{ij} = \mu_{ij}/s_j$
  - Assume no global changes in gene expression
    - $s_j = \text{median}_i(K_{ij}/K_i^R)$; $K_i^R = (\Pi_{j=1..m}K_{ij})^{1/m}$
  - $\log(q_{ij}) = \Sigma_r x_{jr}\beta_{ir}$ (fit normalized read counts to line of user supplied covariates or contrasts/sample comparisons)
- Perform Wald test: compare $\beta_{ir}/SE(\beta_{ir})$ to standard normal distribution N(0,1).
- Calculate p-values by summing/integrating tail of normal distribution: two tailed test.
- Filter low expressed genes whose average across samples is below a threshold which is optimized to maximize the number of genes found at a user specified false discover rate (FDR).
- Adjust for multiple hypothesis testing by calculating the FDR from the p-values using the procedure of Benjamini and Hochberg.